

Mixed-Initiative Methods for Verifiable, Controllable Creation  
in Scientific Research

Marissa Radensky

A dissertation  
submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington  
2025

Reading Committee:

Daniel S. Weld, Chair

Amy X. Zhang

Jeffrey Heer

Program Authorized to Offer Degree:  
Computer Science & Engineering

©Copyright 2025

Marissa Radensky

University of Washington

**Abstract**

Mixed-Initiative Methods for Verifiable, Controllable Creation in Scientific Research

Marissa Radensky

Chair of the Supervisory Committee:

Daniel S. Weld

Computer Science & Engineering

Scientific research is inherently creative. Scientists rely on their creativity constantly, whether to identify new means of inspiration, invent novel research ideas, or find innovative framings of their work to communicate to broader audiences. However, scientists face two key challenges when entering a creative process. To start, they must overcome fixation, a common phenomenon that leads people to think about a problem in familiar terms rather than seek out new solutions. For scientists, this means exploring concepts and ideas outside of their immediate research area and scholarly circles. Furthermore, scientists must work within the context of an exponentially growing body of scientific literature, which can be unwieldy to organize and utilize to its full potential. Artificial intelligence (AI) and large-language-model (LLM) systems are increasingly being used for creative purposes and can provide suggestions to help scientists overcome fixation. In combination with information

retrieval, AI systems can also provide suggestions that make creative use of the expansive scientific literature. Nonetheless, many modern AI systems are black boxes that do not present explanations to understand their actions or controls to steer them.

In this dissertation, I demonstrate that scientists working on creative steps of the scientific process benefit more from AI suggestions when mixed-initiative methods are implemented to help them (1) verify the suggestions for relevance to their creative vision and (2) steer the suggestions to better align with that vision, all while staying rooted in the literature. I begin by showing the complementary benefits of local and global content-based explanations for understanding and adjusting an AI research-paper recommendation feed. Next, I present Papers-to-Posts, a human-LLM tool for translating research papers into blog posts through a novel mechanism—interactive reverse source outlines, which allow users to recognize and change what source content was and was not selected by the LLM for inclusion in the summary article. I then introduce Scideator, a system that implements a human-LLM workflow for iterative research idea generation through transparent, controllable recombination of research-paper facets. To conclude, I reflect on opportunities for future work with respect to investigating the generalizability of the presented mixed-initiative methods as well as exploring the spectrum of scientific literature components used in these methods.

## Acknowledgments

To start, I am deeply grateful to my advisor, Dan Weld, for his guidance in learning the fundamentals of good scientific research, for his consistently thoughtful advice that challenged me without ever making me feel incapable, and for his encouragement to pursue what interests me most. Perhaps the best lesson that Dan has taught me is to always take a step back and remember the bigger picture. To that end, I would like to thank several more people who have contributed to my big picture— a support system on which I have relied in both the hardest and happiest moments over the course of my work on this dissertation and beyond.

The individual projects that comprise this dissertation were supported by the University of Washington WRF/Cable Professorship and the Allen Institute for Artificial Intelligence (AI2). I would like to thank the participants who made these works possible, as well as the anonymous reviewers of these works for their valuable feedback.

I thank my many research and industry mentors and collaborators, who have taught me how to be a better scientist, scholar, and innovator. From my time at the Allen Institute for AI, I thank Jonathan Bragg, Tom Hope, Pao Siangliulue, Joseph Chee Chang, Doug Downey, Kyle Lo, Simra Shahid, Jason Portenoy, Peter Jansen, Peter Clark, Ashish Sabharwal, Tal August, Sergey Feldman, Matt Latzke, Cecile Nguyen, and the Semantic Scholar Team. From my time at Google, I thank Julie Anne-Séguin, Robert Geiger, Pilar Manchón, Kristen Olson, Jang Soo Lim, and the Conversational AI Team. From my time at Microsoft, I thank Rajya Bhaiya, Dustin Burson, Rashmi Raj, and the Health Cloud and Data Team. From my time

as an undergraduate researcher and student, the Amherst College CS Department, the Amherst College Physics Department, Scott Alfeld, Brad Myers, Toby Li, Mary Beth Kery, Xin Liu, Sunghoon Ivan Lee, and Harold Soh. From my time at the University of Washington, I thank Zoran Popović as well as my committee members Amy Zhang, Jeffrey Heer, and Aylin Caliskan.

Thank you to the people who I am exceedingly fortunate to call my friends. While writing this dissertation, you have brought me countless fun memories and a steadfast shoulder to lean on. You have shown up for me time and time again, and I cannot thank you enough. Thank you to Ariel Hochstadt, Sabrina Sage, Addy Sage, Nicole Loo, Wes Koseki, Angelika Hirsch, Alyssa Snyder, Keli Almonte, Anna Dalton, John Watson, Adrian Chen, Arthur Xiao, Kai Ellers, Hunter Dellaverson, Sydney Hark-Odsess, Kyra Noel, and Meenakshi Das.

Thank you to my friends who made the Allen School and the University of Washington such a wonderful place to grow, including Jason (Mingyuan) Zhong, Sudheesh Singanamalla, Priyal Suneja, Ather Sharif, Ruotong Wang, Wei Qiu, Tony Li, Anshuman Mohan, Venkatesh Potluri, Anant Mittal, Obinna Ukogu, and Kevin Feng. Thank you to my friends in the Lab for Human-AI Interaction over the years, from whom I have learned so much: Gagan Bansal, Sherry (Tongshuang) Wu, Jim (Quan Ze) Chen, Benjamin Lee, and Madeleine Grunde-McLaughlin. I would like to give special thanks to my friends Liwei Jiang and Raymond Fok, who have been a constant source of support throughout my PhD experience, always cheering me on and taking me on adventures big and small along the way.

To my partner, Kevan Loo, thank you for making small moments into memorable

quality time even when I am busy with work, for showing me compassion when I am anxious or stressed, and for bringing a smile to my face every day. I am so grateful to have had you by my side in my journey to completing this dissertation.

Thank you to all of my extended family, in particular the Sheingolds (Aunt Nancy, Uncle Eddie, A.J., Christine, and Laydon), the Jaffes (Aunt Sue, Uncle Marc, Jamie, Staci, Nick, Cole, Kylie, Billy, Lisa, Soleil, Levi, and Jade), the Fuziharos/Fujiharas (Tio Ricardo, Tio Mauro, Tia Marina, Vitor, and Bia), and Batian, who I sadly lost while working on this dissertation and miss immensely. Ever since I was a little girl, you have been my greatest champions. Your endless kindness and enthusiasm for all that I do has given me the courage to pursue my wildest dreams. I appreciate all of the ways that our time together has made me feel safe and supported, from our late-night deep conversations to our unexpected laugh attacks.

Thank you to my parents and brother Andre for your unwavering encouragement and love. Andre, you are my oldest friend. While you are my younger brother, I look up to you in many ways. I will forever appreciate your wisdom that is beyond your years and how you look out for me. To my parents, thank you for being with me every step of the way. I thank you for talking through my challenges with me, celebrating my wins with me, and helping me to frame the big picture. I am who I am first and foremost because of the three of you, and I hope to always make you proud. I love you all.

# Table of Contents

Acknowledgments . . . . .	iv
List of Figures . . . . .	xii
List of Tables . . . . .	xxi
1 Introduction . . . . .	1
1.1 Motivation . . . . .	1
1.2 Contributions . . . . .	2
1.2.1 Contribution 1 . . . . .	4
1.2.2 Contribution 2 . . . . .	6
1.2.3 Contribution 3 . . . . .	10
1.3 Prior Publications and Authorship . . . . .	12
2 Related Work . . . . .	13
2.1 Mixed-Initiative Methods . . . . .	13
2.2 Mixed-Initiative Methods for Creativity . . . . .	13
2.3 Mixed-Initiative Methods for Science . . . . .	15
3 Exploring the Role of Local and Global Explanations in Recommender Systems . . . . .	17
3.1 Introduction and Related Work . . . . .	18
3.2 Study 1: Formative Study for System Design . . . . .	20
3.2.1 Existing System . . . . .	20
3.2.2 Formative Study . . . . .	20
3.3 Study 2: Exploratory Study . . . . .	23
3.3.1 Study Design . . . . .	23
3.3.2 Results and Discussion . . . . .	27
3.4 Study 3: Controlled User Study . . . . .	29
3.4.1 Study Design . . . . .	29
3.4.2 Results and Discussion . . . . .	33
3.5 Conclusion, Limitations, and Future Work . . . . .	36
4 Papers-to-Posts: Supporting Detailed Long-Document Summarization with an Interactive, LLM-Powered Source Outline . . . . .	38
4.1 Introduction . . . . .	39
4.2 Related Work . . . . .	44

4.2.1	Human-LLM Text Summarization . . . . .	44
4.2.2	Outlines in Human-LLM Writing . . . . .	47
4.2.3	Human-LLM Scientific Writing . . . . .	47
4.2.4	Scientific Blog Posts . . . . .	48
4.3	The Papers-to-Posts System . . . . .	49
4.3.1	Design Goals . . . . .	50
4.3.2	Implementation Details . . . . .	51
4.3.3	Walkthrough of Warm Start Step . . . . .	51
4.3.4	Walkthrough of Planning Step . . . . .	53
4.3.5	Walkthrough of Drafting Step . . . . .	54
4.3.6	Walkthrough of Revising Step . . . . .	56
4.4	Evaluation Studies . . . . .	57
4.4.1	Lab Study . . . . .	57
4.4.2	Deployment Study . . . . .	63
4.5	Evaluation Studies' Results . . . . .	67
4.5.1	Design Goals 1 and 2 . . . . .	68
4.5.2	Editing Power . . . . .	70
4.5.3	Design Goal 3 . . . . .	73
4.5.4	Blog Post Satisfaction . . . . .	75
4.5.5	Tool Satisfaction . . . . .	78
4.5.6	Cognitive Load . . . . .	80
4.5.7	Task Completion Time . . . . .	82
4.5.8	Qualitative Analysis - DG1 and DG2 . . . . .	83
4.5.9	Qualitative Analysis - DG3 . . . . .	84
4.6	Discussion . . . . .	87
4.6.1	Post-Editing is Not Enough . . . . .	87
4.6.2	Increasing Understanding and Control of Content Selection for LLM-Generated Detailed Summaries . . . . .	87
4.7	Limitations and Future Work . . . . .	88
4.8	Conclusion . . . . .	89
5	Scideator: Human-LLM Scientific Idea Generation and Novelty Evaluation Grounded in Research-Paper Facet Recombination . . . . .	91
5.1	Introduction . . . . .	92
5.2	Related Work . . . . .	96
5.2.1	Divergent and Convergent Thinking . . . . .	96
5.2.2	Human-AI, Scientific Ideation . . . . .	97
5.2.3	Human-AI, Facet-Based Ideation . . . . .	99

5.2.4	Human-AI, Facet-Based, Scientific Ideation . . . . .	100
5.3	System . . . . .	101
5.3.1	Design Goals . . . . .	101
5.3.2	Paper Facets . . . . .	103
5.3.3	Workflow and Implementation . . . . .	103
5.4	Formative Studies for Idea Novelty Checker . . . . .	117
5.4.1	Annotation Studies . . . . .	117
5.4.2	Ablations . . . . .	120
5.4.3	Comparison to Baselines . . . . .	122
5.5	User Study Methods . . . . .	127
5.5.1	Research Questions . . . . .	127
5.5.2	Participants . . . . .	128
5.5.3	Study Design . . . . .	128
5.5.4	Procedure . . . . .	131
5.6	User Study Results . . . . .	133
5.6.1	RQ1: Creativity Support in Idea Generation . . . . .	133
5.6.2	RQ2: Idea Generation Patterns . . . . .	138
5.6.3	RQ3: Confidence in Idea Novelty Assessment . . . . .	141
5.7	Discussion . . . . .	143
5.8	Limitations . . . . .	145
5.9	Conclusion . . . . .	146
6	Conclusion . . . . .	148
7	Discussion and Future Work . . . . .	150
7.1	Evaluating Mixed-Initiative Methods' Impact on Process rather than Outcome for Scientific Research . . . . .	150
7.2	Investigating Generalizability of the Mixed-Initiative Methods . . . . .	151
7.2.1	Local and Global Component-Based Explanations . . . . .	152
7.2.2	Outline of Selected and Unselected Components . . . . .	154
7.2.3	Steerable Recombination of Components . . . . .	155
7.3	Exploring Spectrum of Scientific Literature Components . . . . .	156
7.4	The Role of the Human and AI for Different Tasks in a Mixed-Initiative System . . . . .	158
	References . . . . .	161

A	Papers-to-Posts	182
A.1	Exploratory Analysis of Perceived Control and Ownership	182
A.1.1	Lab Study	182
A.1.2	Deployment Study	183
A.1.3	Summary	183
A.2	Prompts	184
A.2.1	Prompt for Initial LLM Draft for Baseline Tool	184
A.2.2	Prompts for the Warm Start Step of PAPERS-TO-POSTS	184
A.2.3	Prompts for Drafting Step of PAPERS-TO-POSTS	188
A.2.4	Prompts for Revising Step of PAPERS-TO-POSTS	190
A.3	Additional Log Analysis Plots	196
A.4	Lab Study Version of Tool	200
A.5	Example Baseline Initial Draft	200
A.5.1	Introduction	201
A.5.2	Methods	201
A.5.3	Results	202
A.5.4	Conclusion	202
A.6	Example PAPERS-TO-POSTS Output	203
A.6.1	Initial Draft	203
A.6.2	Initial Draft Bullet Point Selection	208
A.6.3	Example Modifications	212
B	Scideator	215
B.1	Prompts for Analogous Paper Facet Finder	215
B.1.1	Prompt to extract facets from a paper title/abstract.	215
B.1.2	Prompt to retrieve facets from papers associated with an analogous query.	218
B.1.3	Prompt to obtain queries for analogous papers from same topic, same subarea, and different subarea compared to input paper.	221
B.1.4	Prompt to shorten a query for papers in the event that it is too long to retrieve 4 relevant papers.	223
B.1.5	Prompt to summarize input and very near analogous papers to obtain related works' summary.	224
B.2	Prompts for Faceted Idea Generator	225
B.2.1	Prompt to obtain ideas using <b>Initial</b> or <b>No-P-no-M</b> method.	225
B.2.2	Prompt to obtain ideas using <b>P-or-M</b> method.	235
B.2.3	Prompt to obtain ideas using <b>P-and-M</b> method.	246
B.3	Prompts for Idea Novelty Checker	255

B.3.1	Prompt to assess idea novelty. . . . .	255
B.3.2	Prompt to extract specific keywords and generate concise re- search titles from an IDEA. . . . .	257
B.3.3	Prompt to extract key facets from an IDEA for facet-based relevance re-ranking of papers. . . . .	259
B.3.4	Prompt to rank passages based on relevance to a query and its key facets. . . . .	261
B.3.5	Prompt to generate more novel idea suggestions for an idea classified as “not novel.” . . . .	265
B.4	Input Papers for Each Ideation Topic . . . . .	269
B.5	Expanded Version of Sample Ideas from Table 5.2 . . . . .	271
B.6	Insights from LLM Prompt Optimization Using TextGrad . . . . .	277
B.7	Expert-Labeled Examples . . . . .	282
B.8	Comparing Scideator’s Novelty Checker to Baselines . . . . .	286
B.9	Experimental Setup Details for Comparing Scideator Novelty Checker to Baselines . . . . .	290

## List of Figures

1.1	I present three mixed-initiative methods for utilizing scientific literature components to produce creative AI suggestions. Each method answers a different question regarding the relationship between the literature components and AI suggestions. . . . .	3
1.2	The combined local and global explanations from the local-global project answer the question of what literature components should influence creative AI suggestions, for the task of finding interesting research papers to read. The local explanation describes why the model shows a particular suggestion to the user, while the global explanation describes how the model produces suggestions in general. The two explanations work together to aid scientists in understanding their AI-curated paper recommendation feed, which they may adjust by increasing or reducing the weights of different keywords. . . . .	4
1.3	Papers-to-Posts answers the question of which literature components should be included in versus excluded from creative AI suggestions, for the task of translating research papers into blog posts. The system implements an interactive reverse outline of the paper, with bullet points pre-selected by the system and synthesized in a generated blog post draft. The user iterates on (a) what bullet points are selected as well as (b) what the blog post sections are in order to improve upon the draft until they are satisfied with its content coverage. . . . .	6
1.4	Scideator answers the question of how literature components should be combined to create novel AI suggestions, for the task of scientific ideation. The system assists scientists in (a) identifying and selecting analogous paper facets to recombine, (b) recombining these facets in diverse new research ideas, and (c) assessing these ideas for novelty relative to the literature. . . . .	9
2.1	This dissertation focuses on human-AI creation tasks that are high-creativity and medium- to high-complexity. This figure is taken from Ding and Chen [44]. . . . .	15

3.1	UI for Study 2 LG condition. Left: default layout. Right: layout when a local explanation is open. Irrespective of condition, these features are present: “(More)” buttons to see full abstracts, “More like this”/“Less like this” buttons for paper feedback, bookmark buttons to save papers, “Refresh” button to apply feedback, “Undo Term Annotations Applied By Refresh” button shown directly after refresh to undo all term annotations applied by refresh, and “Undo All Term Annotations” button to return all terms to original ratings. . . . .	22
3.2	A paper recommendation in Study 2’s L condition, with the local explanation open at the bottom. . . . .	22
3.3	Study 2 results for each Likert-type question and condition. 1 means “strongly disagree,” and 7 means “strongly agree.” Triangles represent the mean responses for each question/condition. Circles represent individual responses. Q4: With both explanations rather than only global (p=0.015, uncorrected) or local (p=0.030, uncorrected), participants were more confident in explaining how they would like the system to improve. Q8: Participants desired less diverse local explanations when global was present (p=0.038, uncorrected). . . . .	26
3.4	How much time Study 3 participants spent on the FP (top) and FN (bottom) surveys as a function of condition, under the “misinformation on social media” feed. Top: Participants spent more time on the FP survey when both explanations were present as compared to only global (p=0.020, uncorrected) or only local (p=0.045, uncorrected). Bottom: Participants spent less time on the FN survey when only global was present as compared to both (p=0.018, uncorrected) or only local (p=0.135, uncorrected). The global explanation alone thus appears more helpful than both explanations together for identifying FPs and FNs efficiently. .	34

4.1	<i>Interactive reverse source outlines</i> in the PAPERS-TO-POSTS system. Users input a long-form source document (research paper) and the system provides (a) a WARM START for the summarization task by generating a reverse document outline. The system then produces a draft summary article (blog post) with sections based on outline bullet points (and associated source paragraphs) selected by the LLM. Given the draft, users can perform two main actions: (b) SUMMARY REVISION, where users adjust the system’s bullet point selection for the summary, triggering system re-generation of the summary; and (c) SECTION ADDITION, where users provide a header for a new desired section, based on which the system selects bullet points and generates a draft of the section content. Users may also edit the draft manually. . . . .	40
4.2	Comparing interactive reverse source outlines in PAPERS-TO-POSTS (blue, upper right) with other affordances for control of LLM summarization, in terms of 1) how grounded the affordance is in the source narrative and 2) the length of the output summary that is supported. PAPERS-TO-POSTS is the first LLM-powered tool to provide writers with highly grounded control over a detailed summary. . . . .	45
4.3	PAPERS-TO-POSTS’ user interface. a) The Blog-Post Area, where the user writes the blog post in sections. b) The Planning-Support Area, which contains both the interactive paper outline and original paper text. c1) The Drafting-Support Area, which contains inputs and outputs for generating text. c2) The Drafting-Support Area continued, seen if one scrolls below the c1 area. d) The Revising-Support Area, which contains inputs and outputs for modifying text and is located below the Drafting-Support Area. (Not Pictured: The modification output appears below the modification buttons, with a button to copy the output text and buttons to view previous modifications and their inputs.) . . . . .	49
4.4	The baseline tool’s user interface, consisting of areas for writing the blog post (left), viewing the paper (top right), and providing instructions to the LLM for generating writing (bottom right). . . . .	60
4.5	Survey responses to 7-point Likert-type questions regarding design goals in the a-d) lab study and e-h) deployment study. Responses are shown for both the treatment and baseline conditions. . . . .	68

4.6	Per condition, Levenshtein distance between a) initial blog post and blog post comfortable with publishing publicly as a function of time spent using the tool (lab study), b) initial and final blog posts as a function of active time spent using the tool (deployment study), and c) initial and final blog posts as a function of number of writing actions taken (deployment study). Across the two studies, participants using PAPERS-TO-POSTS consistently showed more editing power, or change in the blog post for a given amount of time or writing actions. Note that the lab study duration was controlled (up to 60 minutes), whereas deployment study participants wrote on their own time. . . . .	71
4.7	Survey responses to 7-point Likert-type questions regarding satisfaction with the a-d) output blog post in the lab study, e-i) tool in the lab study, j-m) output blog post in the deployment study, and n-r) tool in the deployment study. Responses are shown for both the treatment and baseline conditions. . . . .	75
4.8	Survey responses to 7-point Likert-type questions regarding cognitive load in the a-e) lab study and f-j) deployment study. Responses are shown for both the treatment and baseline conditions. Please note that for the performance question, the labels for the left and right end of the scale were actually respectively “Perfect Performance” and “Failure Performance.” . . . . .	76
5.1	The SCIDEATOR workflow. 1) The interaction begins with the user providing an ideation topic and set of input papers as a starting point for ideation. 2) SCIDEATOR responds by retrieving analogous papers to the input papers and extracting facets (purpose, mechanism, and evaluation) from the input and analogous papers. (The evaluation facets are omitted in the figure for clarity, as it is not part of the main logic.) 3) The user then selects paper facets as well as adds their own facets for which they want to generate ideas. 4) SCIDEATOR recombines these selected facets into ideas with one purpose and one mechanism. If a purpose or mechanism facet is unspecified, the tool selects one. 5) The user selects an idea to assess for novelty. 6) SCIDEATOR classifies the idea as “novel” or “not novel” and provides a short rationale. 7) The user reviews the novelty classification and adjusts it if they disagree. 8) If the idea is deemed “not novel,” SCIDEATOR suggests more novel ideas with one of the initial idea’s facets replaced. . . . .	93

5.2	SCIDEATOR’s cold start. Above, the user selects or adds facets to generate ideas. They can also generate more facets to consider, and add custom instructions for the idea generation. Below, the user peruses their ideas and evaluates an idea for novelty by clicking the search icon to its left. The ideation topic here is human-AI collaboration in art. . . . .	104
5.3	The Analogous Paper Facet Finder module. For a set of input papers, SCIDEATOR uses Semantic Scholar’s API to retrieve similar papers (very near). It uses the input and very-near papers to create a summary of relevant works. Next, the tool extracts key facets from the input papers and determines the input papers’ overarching purpose and mechanism, which it uses to come up with three queries for papers with an analogous purpose and mechanism. The queries are for analogous papers with varying distances from the input paper: same topic (near), same subarea (far), and different subarea (very far). Those queries are fed to the Semantic Scholar API to retrieve analogous papers. Finally, the facets of all the analogous papers are extracted by the LLM. . . . .	105
5.4	SCIDEATOR’s novelty assessment modal for one idea, which presents the idea (a) as well as its facets (b), related papers (c), adjustable novelty classification (d), and adjustable classification reason (e). When the idea is classified as “not novel,” the system provides a set of three suggestions for more novel ideas (f), each of which replace one of the idea’s original facets. The ideation topic here is human-AI collaboration in art. . . . .	112
5.5	The Idea Novelty Checker module follows a retrieve-then-re-rank approach for novelty evaluation. In Step 1, it gathers a comprehensive set of papers relevant to an idea. This includes papers originally used to generate the idea, related papers, and additional papers retrieved through keyword and title searches extracted directly from the idea, as well as snippet searches using the entire idea as input. In Step 2, a two-stage re-ranking process is applied, where an embedding-based ranking strategy filters the large collection to top- $N$ papers, followed by a facet-based LLM re-ranker to identify the top- $k$ most relevant papers. In Step 3, these top- $k$ papers are used to assess the idea’s novelty, guided by in-context examples that evaluate novelty with grounded reasoning. In Step 4, if an idea is classified as “not novel” by the tool or user, the LLM generates three idea suggestions, each replacing a different facet in the original idea in order to make the idea more novel compared to the relevant papers. . . . .	114
5.6	The cold start of the baseline UI for the user study’s idea-generation task. The ideation topic here is human-AI collaboration in art. . . . .	129

5.7	(a) The difference between participants’ unweighted CSI scores for SCIDEATOR versus the baseline tool. Participants experienced significantly more creativity support with SCIDEATOR. (b) For each CSI factor, the difference between participants’ ratings for SCIDEATOR versus the baseline tool. (c) How many times each CSI factor wins against other factors in terms of what is most important to participants while generating ideas.	134
5.8	(a) Participants more often opted to select their own facets rather than let the LLM select for them. (b) Participants used input facets and facets nearer to the input more than facets farther from the input.	140
5.9	Participants’ average perceived idea novelty before (a) and after (b) utilizing their assigned tool for idea novelty evaluation, as well as the average change from initial to final perceived novelty (c). The baseline tool was the assigned tool when ideas had no classification, while SCIDEATOR was the assigned tool when ideas had a novel or not-novel classification.	142
A.1	Survey responses to 7-point Likert-type questions regarding perceived control and ownership in the a-d) lab study and e-h) deployment study. Responses are shown for both the treatment and baseline conditions.	183
A.2	Prompt for initial LLM draft for baseline tool.	184
A.3	Prompts for generating bullet points for each paragraph in full paper being summarized. Each prompt corresponds to different length paragraphs. The top prompt generates three bullet points for paragraphs of more than 100 words, the second generates two bullet points for paragraphs of between 51 and 100 words, and the third generates one bullet point for paragraphs less than 51 words long.	185
A.4	Prompt to select top 10 bullet points relevant for each of the initial draft’s blog post sections (introduction, methods, results, conclusion).	186

A.5	Prompt to generate each of the initial draft’s blog post sections (introduction, methods, results, conclusion) using the 10 relevant selected bullet points. The section-specific guidelines are as follows. Introduction: “-Present hook (e.g., interesting fact, quote, promise of change in knowledge, illustrating example of the topic). If context allows, visual or sensory elements are helpful anchors.-Provide high-level description of problem being solved.-Explain why work is interesting and a solution to the problem matters.-Do not repeat information from prior blogpost sections.” Methods: “-Focus on methods and do NOT discuss results.-Do not repeat information from prior blogpost sections.” Results: “-State key takeaway.-Discuss up to 3 most interesting aspects of work.-Do not repeat information from prior blogpost sections.” Conclusion: “-Restate key takeaway in new way.-Present future work ideas. [optional]-Loop back to hook. [optional]“ . . . . .	187
A.6	Prompt for selecting bullet points for generating a new section based on the section header provided by the user. . . . .	188
A.7	Prompt for generating text for a blog post section when there are selected paragraphs or bullet points as well as custom bullet points, custom instructions, or starting text. . . . .	188
A.8	Prompt for generating text for a blog post section when there are selected paragraphs or bullet points but no custom bullet points, custom instructions, or starting text. . . . .	189
A.9	Prompt for generating text for a blog post section when there are no selected paragraphs or bullet points, but there are custom bullet points, custom instructions, or starting text. . . . .	189
A.10	Prompt for the “expand” modification when there are no custom modification instructions. The automatic modLength variable for this modification is “twice the length that it currently is.” . . . . .	190
A.11	Prompt for the “expand” modification when there are custom modification instructions but no selected paragraphs or bullet points of which to be aware. The automatic modLength variable for this modification is “twice the length that it currently is.” . . . . .	191
A.12	Prompt for the “expand” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “twice the length that it currently is.” . . . . .	192

A.13 Prompt for the “condense” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “half the length that it currently is.” The other prompts for this modification mirror those for “expand.” . . . . .	193
A.14 Prompt for the “simpler terms” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for the this modification mirror those for “expand.” . . . . .	194
A.15 Prompt for the “more dramatic” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for the this modification mirror those for “expand.” . . . . .	195
A.16 Prompt for the “less dramatic” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for the this modification mirror those for “expand.” . . . . .	196
A.17 Number of times each input or button related to generating and modifying text was used by each participant in lab study. From left to right: the act of generating text, inputs/buttons related to generating text, the act of modifying text, inputs/buttons related to modifying text. . . . .	197
A.18 Additional plots for lab study. . . . .	198
A.19 Additional plots for deployment study. . . . .	199
A.20 PAPERS-TO-POSTS in the lab study. a) Blog-Post Area, b) Planning-Support Area, c) Drafting-Support Area, and d) Revising-Support Area. . . . .	200
B.1 Performance trends of test accuracy across prompts during prompt optimization with TextGRAD. Highlighted text shows unique instructions used to evaluate the novelty of ideas. The final test accuracy was 0.78125, showing that none of the optimized prompts (1 to 12) improved over the original. . . . .	279
B.2 contd. TextGrad Prompt Optimisation. . . . .	280

B.3	contd. TextGrad Prompt Optimisation. . . . .	281
B.4	Two example ideas used as the basis for comparison in subsequent figures, evaluated by SCIDEATOR (Ours), AI Scientist, and AI Researcher. . . . .	287
B.5	Reviews corresponding to idea 1 in Figure B.4. . . . .	288
B.6	Reviews corresponding to idea 2 in Figure B.4. . . . .	289

## List of Tables

3.1	Study 2 metrics for hypotheses in Section 3.3.1. Questions are 7-point Likert-type questions. LFM means log file metric. . . . .	23
3.2	Study 3 metrics for hypotheses in Section 3.4.1. Scores are described in Section 3.4.2. Question is a 7-point Likert-type question. . . . .	30
5.1	The Faceted Idea Generator module generates analogies between two purpose-mechanism pairs from different papers. It then uses these analogies to inspire novel ideas combining the purpose from one paper with the mechanism from another. The module aims to combine papers of varying distance from the input papers, but which papers are involved in the analogies depends on whether the user has selected any purposes or mechanisms. . . . .	108
5.2	A sample of participants’ favorite ideas from the user study. The treatment ideas are generated with the different methods of the <b>Faceted Idea Generator</b> module, depending on whether the participant selected any purposes or mechanisms. . . . .	111
5.3	Accuracy of SCIDEATOR ablations in predicting “not novel.” . . . .	121
5.4	Comparing each ablation to the complete system in terms of overlap in retrieved papers and paper rankings. <i>Overlap</i> indicates how many papers overlap on average with the complete system’s top-10 papers. <i>Rank Shift</i> measures the average absolute difference in rank positions (only among overlapping papers). . . . .	121
5.5	Experimental results on expert-annotated dataset comparing SCIDEATOR’s novelty checker to various baselines. Unless otherwise noted, GPT-4o was used to obtain these results. . . . .	124
B.1	A sample of participants’ favorite ideas from the user study, presenting the expanded version of each idea rather than the shorter version. The treatment ideas are generated with the different methods of the <b>Faceted Idea Generator</b> module, depending on whether the participant selected any purposes or mechanisms. . . . .	277
B.2	<b>Expert-labeled examples from annotation study</b> . . . . .	282

# Chapter 1

## Introduction

### 1.1 Motivation

The scientific process presents many opportunities for AI support. Some scientific tasks, being relatively formulaic and repetitive, are amenable to automatic AI interventions, such as gathering papers related to a particular topic [35, 97] or identifying and testing relatively low-level or highly scoped hypotheses [83, 9, 121]. On the other hand, most scientific tasks are not straightforward and require creativity, which often thrives on a diversity of viewpoints [82] and therefore may benefit from the combined efforts of scientists and AI. In alignment, prior work has explored human-AI collaboration with respect to scientific tasks, such as literature exploration [89, 92, 53], research ideation [117, 118, 23], and research communication [58, 72, 96].

Scientists encounter two main obstacles when engaging in creative endeavors. For one, like all people, scientists are inclined to think about problems in manners that are familiar to them. In cognitive science, this is referred to as fixation [84, 48]. However, some of the most groundbreaking scientific ideas come from those who are able to draw on knowledge from outside their immediate research area [76, 158]. For example, Christiaan Huygens proposed his wave theory of light by creating an analogy between the properties of light and the properties of sound, which was known to travel in waves [158, 76]. As another example, Charles Darwin was inspired to come up with the idea of natural selection to explain species' evolution when he read Malthus' work, which discussed how human population growth in a world of limited

resources could lead to competition for survival [76]. Even when scientists break out of their filter bubbles and commit to exploring ideas beyond the topics with which they are most familiar, identifying and organizing pertinent information within the large space of scientific literature is difficult [177, 32].

AI systems can support scientists in tackling both of these challenges. With respect to fixation, scientists can supplement their thinking with AI, drawing from its wide range of knowledge rather than the limited pool of topics familiar to the scientists. AI can also help scientists to effectively utilize the ever-expanding body of scientific literature by not only providing access to a vast amount of information but also helping to organize and process this information. On the other hand, humans can think critically about the information presented to them and have unique perspectives based on their individual experiences that can contribute to creativity. Thus, scientists and AI can combine their efforts to effectively utilize the literature for impactful creativity. In this work, I investigate how AI can best support scientists in ameliorating these challenges related to creative endeavors.

## 1.2 Contributions

Through this dissertation, I argue that, during **creative** steps of the **scientific** process, researchers benefit more from **AI suggestions** when mixed-initiative methods are implemented to help them 1) **verify** the suggestions for relevance to their creative vision and 2) **steer** the suggestions to better align with that vision, all while staying **rooted in the literature**. I substantiate this argument with three contributions (Figure 1.1). Each presents a different mixed-initiative method that utilizes

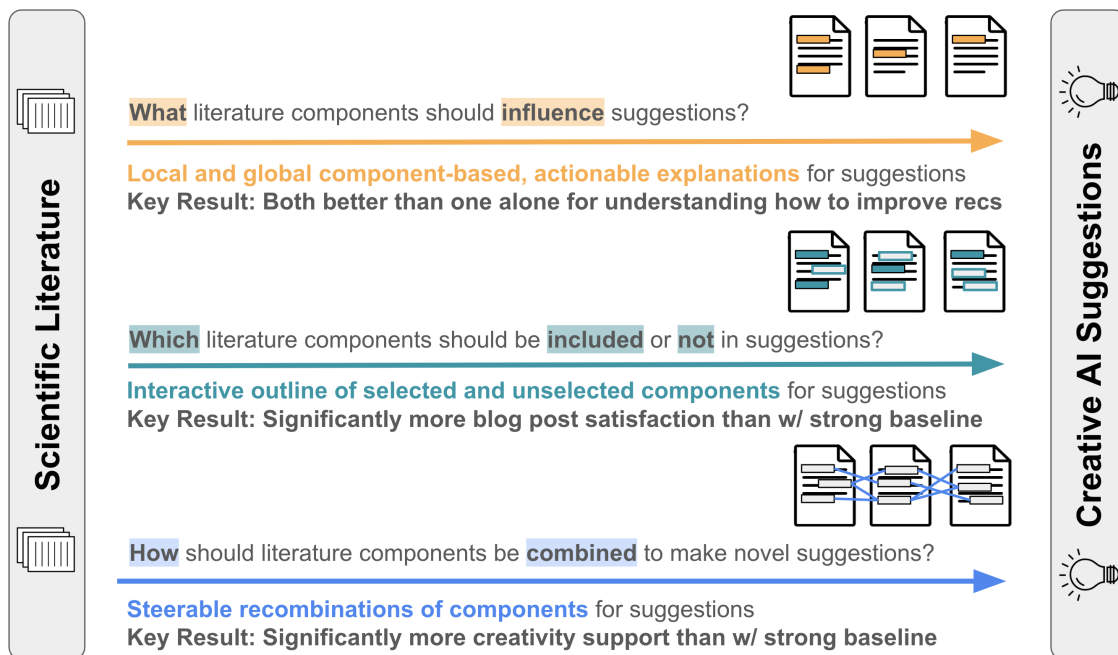


Figure 1.1: I present three mixed-initiative methods for utilizing scientific literature components to produce creative AI suggestions. Each method answers a different question regarding the relationship between the literature components and AI suggestions.

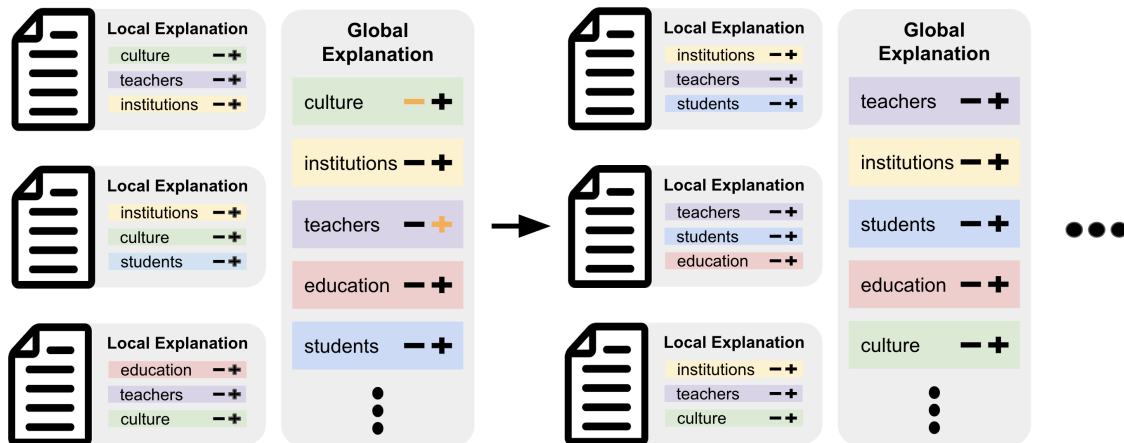


Figure 1.2: The combined local and global explanations from the local-global project answer the question of what literature components should influence creative AI suggestions, for the task of finding interesting research papers to read. The local explanation describes why the model shows a particular suggestion to the user, while the global explanation describes how the model produces suggestions in general. The two explanations work together to aid scientists in understanding their AI-curated paper recommendation feed, which they may adjust by increasing or reducing the weights of different keywords.

scientific literature components to verify and steer creative AI suggestions. I define a scientific literature component as any element of a scientific document, whether extractive (e.g., keyword) or abstractive (e.g., summarizing bullet point).

### 1.2.1 Contribution 1

**I demonstrate how two types of component-based explanations—local and global—may complement one another in supporting scientists’ understanding and control of which literature components are used to generate creative AI suggestions [150].** Local explanations describe why the AI provides a

particular suggestion to the user, whereas global explanations describe how the AI generates suggestions in general [155]. I investigate the roles of the two explanation types in facilitating scientists' ability to interpret and steer an AI recommendation feed for the creative task of discovering new research papers of interest. For this exploration, I adapt a system from prior work [111], in which the local and global explanations are unigram or bigram terms from papers. Figure 1.2 shows an example of global and local explanations for a recommendation feed of papers related to "technology in education." In the actual system, each term presents an associated weight assigned by the model, which the user may adjust. Based on formative study feedback, I set the local explanation to consist of the two to four top terms related to a single paper recommendation. Meanwhile, I set the global explanation to present the top 80 terms considered by the recommender system for all paper recommendations and provides the ability to search for other terms of interest to view and change their weights.

To evaluate the role of local and global explanations, I conduct two mixed-methods user studies, one exploratory and one controlled, during which scientists attempt to create a recommendation feed that aligns with particular research topics. In both studies, I compare three conditions in a partial within-subjects manner, in which the recommender system supplies 1) only local, 2) only global, or 3) both explanations. Through the exploratory study, I find evidence suggesting that 1) providing both explanations rather than either alone ensures that scientists reach the best understanding of how the recommendation feed can improve, and 2) scientists prefer more diverse local explanations when they are not accompanied by global

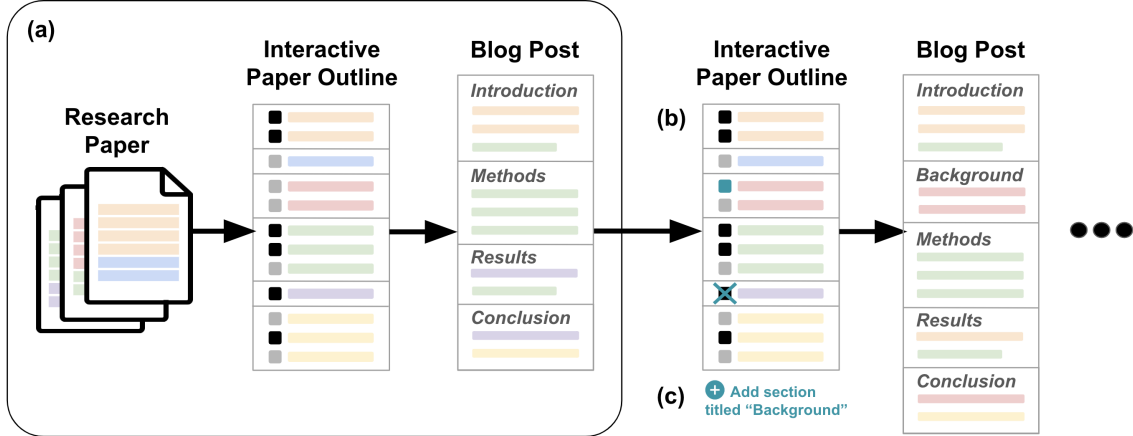


Figure 1.3: Papers-to-Posts answers the question of which literature components should be included in versus excluded from creative AI suggestions, for the task of translating research papers into blog posts. The system implements an interactive reverse outline of the paper, with bullet points pre-selected by the system and synthesized in a generated blog post draft. The user iterates on (a) what bullet points are selected as well as (b) what the blog post sections are in order to improve upon the draft until they are satisfied with its content coverage.

explanations in order to better steer their recommendation feed. Through the controlled lab study, however, I find that both explanations may be less helpful than global explanations alone in efficiently identifying false-positive and false-negative recommendations.

### 1.2.2 Contribution 2

After presenting methods to help scientists understand *which* literature components are used in AI suggestions, my next step is to introduce a method to assist scientists in grasping *what* the full space of literature components available to the AI look like.

**I show that providing an outline of literature components available to the**

**AI allows scientists to recognize and modify what components are being excluded from their AI suggestions.** More specifically, I introduce interactive reverse source outlines, a novel mechanism that allows users to review and revise an LLM’s selection of bullet points within an LLM-generated source outline, which is used to draft a detailed summary of the source document. I apply this mechanism to a common creative task for scientists—summarizing findings from research papers to reach broader audiences, in this case through blog posts. I implement interactive reverse source outlines for this application in a system that I call Papers-to-Posts (Figure 1.3).

Papers-to-Posts’ human-LLM workflow begins with a warm start (Figure 1.3a). Here, the system takes the scientist’s research paper to be summarized as input. The system generates an outline of the paper, consisting of one to three bullet points per paragraph. The system selects ten relevant bullet points to synthesize for each pre-set section of the blog post draft, which are “introduction,” “methods,” “results,” and “conclusion.” These bullet points are synthesized based on section-specific instructions, which are drawn from resources on writing academic blog posts. This initial draft is presented to the scientist. For each blog post section, they can review and adjust the bullet point selection (Figure 1.3b) along with the section-specific instructions for the LLM, in order to re-generate the section. The scientist can also add and delete sections. To add a new section, the scientist first specifies the section header. Next, they can select bullet points for the section themselves or let the system select ten relevant bullet points to include based on their provided header (Figure 1.3c). The system then synthesizes the bullet point selection to generate the new section.

Throughout the interaction, the scientist can also manually adjust the blog post text as they see fit, and Papers-to-Posts provides affordances for modifying the generated text (e.g., “simpler terms,” “more dramatic”).

Through two mixed-methods user studies, I evaluate how well Papers-to-Posts supports scientists in recognizing and iterating on the content coverage of each of the LLM’s suggested blog post sections. The two studies compare Papers-to-Posts to a strong baseline consisting of access to an LLM-generated initial blog post draft and free-form LLM prompting. Participants are asked to write blog posts for their own research papers using the assigned tool. In a within-subjects controlled study with time constraints, I find that participants are significantly more satisfied with their final blog posts created with Papers-to-Posts, primarily with respect to content coverage. I then conduct a between-subjects deployment study to assess the effects of Papers-to-Posts in a real-world environment. In this setting, participants are not given time constraints but are required to put their best effort towards creating a blog post to share with others, whether by social media, email, or some other means. Although I do not see the same significant increase in blog post satisfaction, this is not surprising given the lack of time constraints. Across both studies, I uncover evidence that participants found Papers-to-Posts more helpful, making incorporating paper content in the blog post easier and iterating on the blog post easier. Furthermore, I observe that participants demonstrate increased editing power (change in the blog post within a given amount of time or writing actions) using Papers-to-Posts.

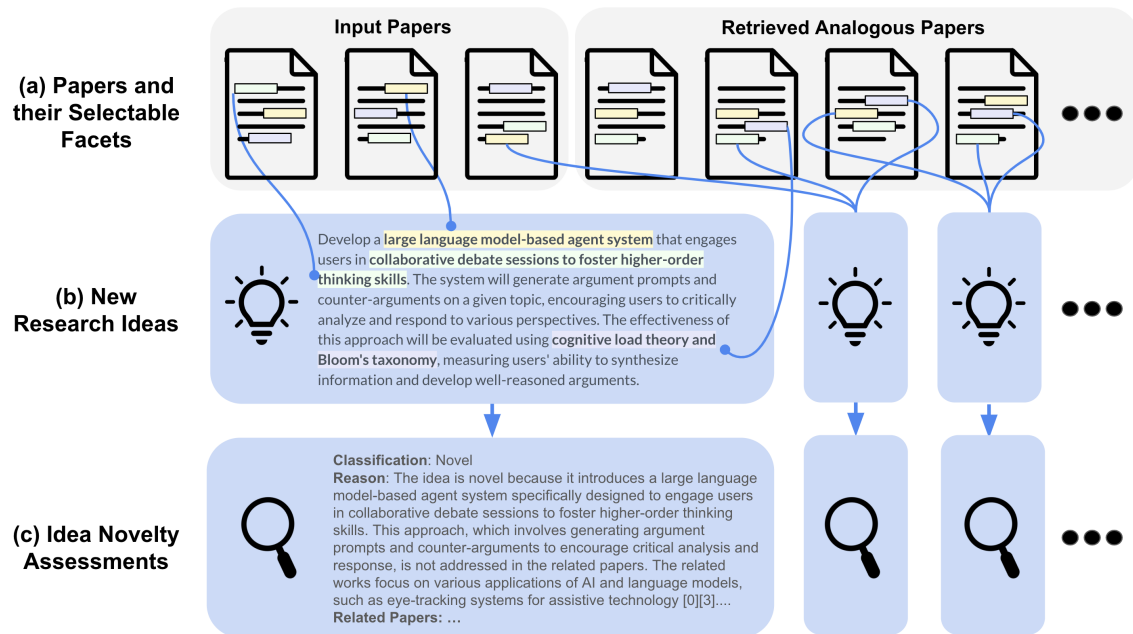


Figure 1.4: Scideator answers the question of how literature components should be combined to create novel AI suggestions, for the task of scientific ideation. The system assists scientists in (a) identifying and selecting analogous paper facets to recombine, (b) recombining these facets in diverse new research ideas, and (c) assessing these ideas for novelty relative to the literature.

### 1.2.3 Contribution 3

Next, I aim to assist scientists in understanding and controlling not only what literature components are included versus not included in creative AI suggestions, but also *how* and *why* certain literature components should contribute to an AI suggestion. **I demonstrate that descriptions of how literature components may be recombined aid scientists in identifying and selecting interesting candidate components to include in AI suggestions, while assessments of component recombination suggestions help scientists to recognize flaws in the suggestions and iterate accordingly.** For the task of scientific ideation, I introduce the system Scideator, which supports mixed-initiative recombination of paper facets into new research ideas as well as facet-based novelty assessment of these ideas (Figure 1.4). The paper facets include purpose (problem addressed by the paper), mechanism (paper’s proposed method to solve the problem), and evaluation (paper’s method to determine if the mechanism solved problem).

Scideator implements a human-AI workflow for scientific ideation through three retrieval-augmented generation modules. The first module, the Analogous Paper Facet Finder, takes the user’s overarching ideation topic and small set of relevant papers (e.g., 1-5) as input. The module retrieves a set of analogous papers of varying topical distance from the input papers and extracts key facets from the input and analogous papers, which become its output (Figure 1.4a). The second module is the Faceted Idea Generator. Its input is a set of paper facets selected by the user, if any. The module selects paper facets for the user if they did not select enough on their own and recombines those facets into new ideas to output (Figure 1.4b). The third

module, the Idea Novelty Checker, assesses user-selected ideas for novelty, providing a binary novelty classification as well as a paragraph explaining how the idea is novel or not relative to retrieved related papers (Figure 1.4c). The module also presents suggestions for how to iterate upon ideas classified as not-novel by the system or user.

In a mixed-methods within-subjects user study, I evaluate Scideator with respect to how well it assists scientists in exploring research ideas, which describe how to combine diverse paper facets, as well as in assessing and iterating on these ideas' novelty. I ask participants to complete an idea generation task and an idea novelty assessment task under two conditions. Both conditions provide access to Semantic Scholar as well as free-form LLM prompting for the novelty assessment task. The treatment condition also provides access to Scideator, while the baseline condition also provides access to a simplified version of Scideator, which supports recombining input papers rather than diverse paper facets and does not support novelty assessment. Based on the Creativity Support Index [29], I find that scientists experience significantly more creativity support using Scideator, particularly with respect to the aspect of exploration, which relates to identifying and selecting different paper facets to include in research ideas. Furthermore, while I do not observe a significant increase in scientists' confidence in their idea novelty assessments with Scideator, I do observe that scientists who receive a not-novel idea classification adjust their idea novelty assessment more than scientists who receive a novel classification or no classification (from the baseline). Thus, not-novel idea classifications do appear to help scientists in recognizing flaws in their AI-suggested research ideas.

### 1.3 Prior Publications and Authorship

Although I am the first-author of all work that appears in this dissertation, each associated project was the result of collaboration with my advisor and several other researchers and mentors at the Allen Institute for AI and Adobe. The work on local and global explanations (Chapter 3) appeared as an extended abstract at CHI 2022 [150] and was done in collaboration with Doug Downey, Kyle Lo, Zoran Popović, and Dan Weld. The work on Papers-to-Posts (Chapter 4) is published on arXiv [152] and undergoing peer review. It was done in collaboration with Dan Weld, Joseph Chee Chang, Pao Siangliulue, and Jonathan Bragg. The work on Scideator (Chapter 5) is published on arXiv [151] and undergoing peer review. It was done in collaboration with Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Dan Weld. In particular, I would like to acknowledge that Simra completed most of the work to develop the backend of Scideator’s novelty checker module. To reflect my collaborators’ contributions, I use the first-person plural in Chapters 3-5.

## Chapter 2

### Related Work

#### 2.1 Mixed-Initiative Methods

Mixed-initiative methods are those that support human collaboration with AI [80]. In this dissertation, I focus on a key principle of mixed-initiative interfaces: “*providing mechanisms for efficient human-AI collaboration to refine results*” [80]. To that end, I develop and evaluate mixed-initiative methods that promote granular feedback from the user, as suggested by Amershi et al. in their guidelines for human-AI interaction [5]. This feedback is used within a communicative human-AI workflow, in which the human and AI take actions based on the feedback of the other [62]. For each mixed-initiative method presented in this dissertation, the user interface acts as a boundary object between the human and AI, providing them a shared language through which they may create an artifact of interest to the user [165, 71].

#### 2.2 Mixed-Initiative Methods for Creativity

Colton and Wiggins define computational creativity research as “the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative” [37]. “Particular responsibilities” indicates that the systems are not merely creativity support tools but take some creative action themselves. In addition, the authors describe three maxims for computational creativity projects. First, they should celebrate the creative act in addition to simply the artifact that is produced. Second,

such projects should consider the artifact as an invitation to engage with not only the artifact but the creator, culture, and oneself. Lastly, such projects should frame their processes and outputs well. In alignment with the first two maxims, this dissertation proposes methods for mixed-initiative creation that emphasize the role of the scientist in creating and engaging with system artifacts. Furthermore, in alignment with the last maxim, these methods all involve providing a framework through which scientists can verify and control an AI system’s processes and artifacts.

More recently, Ding and Chen introduced a mapping of the design space for human-AI text co-creation tasks [44], but this mapping can easily extend to non-textual tasks as well. This dissertation mainly focuses on high-creativity tasks that vary in terms of complexity in human-AI interaction patterns, but the complexity never falls to minimal human intervention (Figure 2.1). A human-AI interaction with this level of creativity and complexity may benefit from a shared language through which the human and AI can communicate to complete the task at hand [71, 165]. In coincidence, Yannakakis et al. note that mixed-initiative co-creativity often constitutes a type of diagrammatic lateral thinking, which refers to using visual representations of problems to solve them in a creative manner [188]. In this dissertation, I describe mixed-initiative methods that visually represent tasks within the creative scientific process, by breaking scientific documents into components that are utilized by the scientist and AI alike.

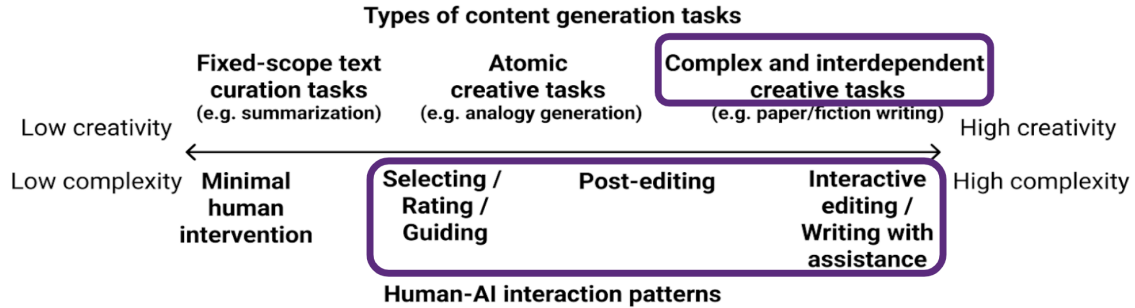


Figure 2.1: This dissertation focuses on human-AI creation tasks that are high-creativity and medium- to high-complexity. This figure is taken from Ding and Chen [44].

### 2.3 Mixed-Initiative Methods for Science

There has been an increasing interest in mixed-initiative co-creation systems for exploring [192, 53, 191], defining [117, 148], and communicating [120, 58, 96] scientific research.

For exploring scientific research, prior work has focused on supporting co-creation of scientific paper summaries for readers [192, 12, 53]. Of note, ConceptEVA [192] and Qlarify [53] offer features for verifying the output text, respectively a concept visualization and direct paper text attribution. Prior work has also explored LLM-powered tools for defining scientific research. For example, the tool CoQuest supports generating many potential scientific research ideas in a breadth-first or depth-first interaction [117], while the tool IdeaSynth supports developing a single input research idea [148]. Other works have investigated human-LLM interactions for communicating scientific research. For instance, Gero et al. introduce a system for generating scientific “sparks,” or sentences meant to inspire people writing about science [58]. Long et al. study how LLM scaffolding helps people to generate captivating hooks

for complex scientific topics [120], and Kim et al. examine how an LLM system can support science writers in producing extended metaphors for scientific ideas [96].

Recently, efforts have been made to create AI scientists that support, if not replace, humans throughout the scientific process, from generating potential hypotheses to reporting experimental results [60, 121]. I aim to build upon all of these prior works in my dissertation, offering mixed-initiative methods to support *verification* and *control* of AI systems that address complex, creative tasks in scientific research.

## Chapter 3

# Exploring the Role of Local and Global Explanations in Recommender Systems

In this chapter, we investigate how AI systems may help scientists to verify and steer **what** literature components **influence** their creative suggestions grounded in the literature. We find evidence that combining **local and global explanations**, rather than presenting one or the other alone, to support this verification and control leads scientists to a better understanding of how their suggestions can be improved.

Explanations are well-known to improve AI recommender systems' transparency [10, 24, 61, 114] and controllability [10, 61, 99, 110]. These explanations may be local, explaining individual recommendations, or global, explaining the recommender model overall. Despite their widespread use, there has been little investigation into the relative benefits of the two explanation approaches. We conduct an exploratory user study (n=30) and a controlled user study (n=30) with a research-paper recommender system that presents paper-keyword explanations to assist scientists in curating their recommendation feed to cover a research topic of interest. We analyze how providing local, global, or both explanations assists scientists in verifying and controlling what paper keywords influence their recommendations. Our results provide evidence suggesting that both explanations are more helpful than either alone for understanding how to improve recommendations, yet presenting both may be less helpful than presenting global alone for efficiently identifying false positive and false negative recommendations. We also observe that scientists appear to desire

more diverse local explanations when there is no global explanation present, which we reason may be in order to better steer the model.

### 3.1 Introduction and Related Work

Recommender systems are used daily by millions of people, and explanations that clarify their behavior are well-known to improve users’ perceptions of their usefulness [10, 16, 17, 173, 99, 51, 176, 47, 174], controllability [10, 61, 99, 110], trustworthiness [19, 156, 2, 114, 110, 51], and transparency [10, 24, 61, 114, 99, 172, 106, 110]. Recommenders may provide users with *local* explanations describing why a specific item is recommended [38, 110], a *global* explanation describing how recommendations are selected overall [140, 93], or *both*, presented separately [144, 2, 100, 173, 10, 86, 130] or in a unified manner [61, 16, 17, 19, 159, 43].

Despite widespread use of local and global explanations in recommender systems, to the best of our knowledge there has been no investigation into how each influences recommender transparency. In machine learning broadly, global explanations explain how a model behaves generally, while local explanations explain a single model output, as first distinguished by Ribeiro et al. [156]. Some works find that local explanations are more useful for model faithfulness [156, 66]. Others discuss benefits from both explanations in terms of understanding and evaluating models [132, 81, 46, 74, 103]. We build on these works to address how local and global explanations affect *recommender* transparency.

Recommenders differ from most AI systems in that their output cannot be objectively evaluated as correct or not. Local and global explanations may be used

differently when users must *subjectively* decide recommendations' relevance and provide feedback. Do the two explanation types play complementary roles in helping users understand how the system may improve? Is one better for detecting false positive or false negative recommendations? Are local explanations used differently if a global explanation is also present, or vice versa? We examine these questions and more using Semantic Sanity, a website allowing users to create recommendation feeds of computer-science research papers.

In summary, we make the following contributions:

- A formative study regarding how to present local and global explanations in a research-paper recommender.
- An exploratory study and controlled user study, each with 30 computer-science researchers, using the recommender to investigate several hypotheses surrounding three conditions: local, global, and both explanations.
- Evidence suggesting that 1) both explanations help users explain how to improve recommendations better than either alone, but 2) both is less helpful than global alone for efficiently identifying false positive and negative recommendations. Also, 3) users prefer less diverse local explanations when a global explanation is also available.

## 3.2 Study 1: Formative Study for System Design

### 3.2.1 Existing System

Semantic Sanity, a website to generate recommendation feeds of computer-science research papers, is the existing system that we adjusted for our studies. The recommender system uses a linear model with 15000 paper terms (unigrams and bigrams) as features. Despite the model’s linearity, the large number of features prevents the model from being fully interpretable. Thus, to help users understand the model’s recommendations, the system presents local and global explanations consisting of the paper terms used as model features. These explanations are a form of the common content-based explanation [104, 54, 10, 16, 86, 2]. By presenting the explanations as simple key terms, as in some other explainable recommenders [16, 86, 2], we prevent the explanations from adding too much clutter to the already information-heavy user interface. The global terms have the most positive weights in the recommender’s linear model. The local terms have the most positive products of model weight and TF-IDF value for the associated paper; we use LIMEADE’s approach [110] for introducing some randomness to diversify the local terms.

### 3.2.2 Formative Study

In a formative interview study, we asked six computer-science researchers for their thoughts on design mockups to determine how best to present local and global explanations in the research-paper recommender Semantic Sanity. A majority of participants preferred that local and global explanations be toggle-able, so they would

see them only when they desired to do so. Furthermore, most participants wanted the explanations to be presented in a unified manner when they were both available. This means that the local explanation is visualized as a filtered version of the global explanation, rather than a completely separate explanation. Participants also preferred that the explanations be actionable, meaning that the user may manipulate the explanation widget to provide feedback [110]. Lastly, participants easily understood that when local explanations had varying numbers of terms, only the most significant terms were shown. Thus, within the constraint of two to four terms total, the system added terms to a local explanation until the term weights hit a plateau so that the explanation had the most salient terms.

Figure 3.1 shows the resulting interface for the local-global condition. In all conditions, users can like or dislike papers and give feedback on explanation terms. In the **local-global (LG)** condition, the “Feed Explanation” button opens a sidebar (open by default) containing the global explanation. The sidebar shows the top 80 feed terms and allows users to search all 15,000 terms. Users can adjust term ratings between 0.0 and 1.0 by using the plus and minus buttons to add or subtract 0.1. The “Paper Explanation” button under each paper displays a local explanation. This surfaces two to four paper-relevant terms in the sidebar, and clicking the carrot underneath them puts the terms in context of the global explanation. The **global (G)** condition looks similar but does not include the “Paper Explanation” buttons. In the **local (L)** condition, the “Paper Explanation” button under each paper reveals two to four terms explaining why the paper was recommended (Figure 3.2), and a “View All Paper Explanations” button opens all local explanations.

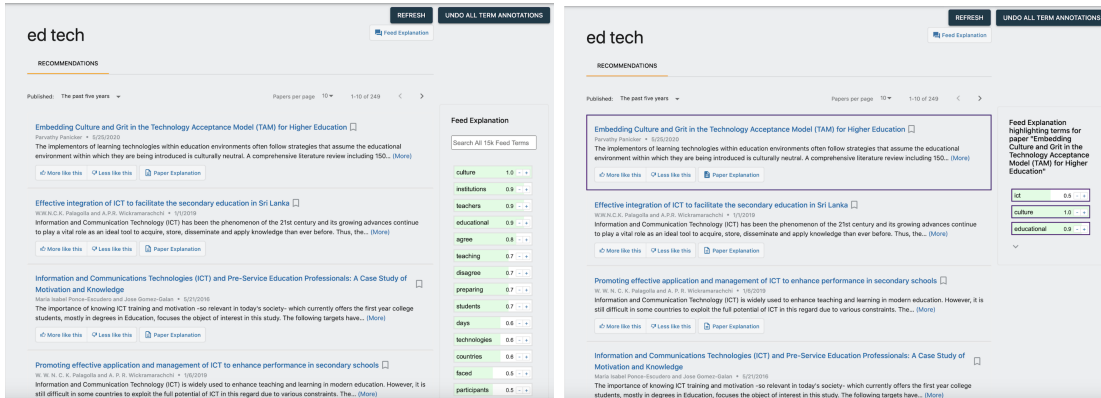


Figure 3.1: UI for Study 2 LG condition. Left: default layout. Right: layout when a local explanation is open. Irrespective of condition, these features are present: “(More)” buttons to see full abstracts, “More like this”/“Less like this” buttons for paper feedback, bookmark buttons to save papers, “Refresh” button to apply feedback, “Undo Term Annotations Applied By Refresh” button shown directly after refresh to undo all term annotations applied by refresh, and “Undo All Term Annotations” button to return all terms to original ratings.

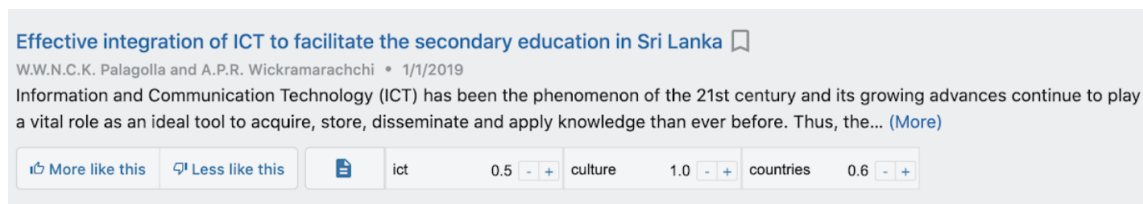


Figure 3.2: A paper recommendation in Study 2’s L condition, with the local explanation open at the bottom.

Table 3.1: Study 2 metrics for hypotheses in Section 3.3.1. Questions are 7-point Likert-type questions. LFM means log file metric.

Hypo.	Metric ID	Metric
-	Q0: feed success	“The recommendation feed helps me find relevant papers.”
H1	Q1: past actions	“The explanation(s) help me to understand why the system returned the papers it did.”
H2	Q2: future actions	“The explanation(s) help me to anticipate what kinds of papers the system will return in the future.”
H3	Q3: understand me	“The explanation(s) help me to know when the system doesn’t understand my interests.”
H4	Q4: change behavior	“When the feed is not completely relevant, I can explain how I would like the system to behave to be more relevant.”
H5	Q5: false pos paper	“The explanation(s) help me to determine whether a <b>paper</b> is relevant or irrelevant.”
	Q6: false pos term	“The explanation(s) help me to understand which <b>term</b> might cause an irrelevant paper to appear in my feed.”
	LFM1	% of annotated terms that are annotated negatively
H6	Q7: false negative	“The explanation(s) help me to understand how likely the feed is to <b>miss papers</b> that I’d consider relevant.”
H7	Q8: local diversity	“I would like the Paper Explanations to cover a less diverse set of terms, focusing more on the highest-rated terms.”
H8	LFM2	# of annotated terms

### 3.3 Study 2: Exploratory Study

#### 3.3.1 Study Design

#### Hypotheses

Study 2’s overarching research question **RQ1** was: how do people utilize local and global explanations in a research-paper recommender? The first six hypotheses sur-

round how the explanations affect the recommender’s transparency and are inspired by target purposes of AI explanations enumerated in previous work [73, 67]. They state that there is at least one paired difference among the L, G, and LG conditions in terms of utility in... **H1**: understanding past recommender actions, **H2**: understanding future recommender actions, **H3**: understanding how well the system understands the user, **H4**: understanding how the system can improve, **H5**: identifying false positive recommendations, and **H6**: identifying false negative recommendations. The last two hypotheses address how users’ interactions with the explanations are affected by the explanation types provided. **H7**: There is a difference between L and LG with regards to desired diversity of local explanation terms, and **H8**: there is at least one paired difference among L, G, and LG with regards to amount of explanation feedback. The hypotheses’ metrics are in Table 3.1.

## Participants and Treatments

Thirty researchers who read at least one computer-science research paper monthly interacted with the recommender for a half-hour to one-hour and were compensated with \$25 Amazon gift cards. Fifteen participants received the G and L conditions in randomized order; the other 15 interacted only with the LG condition. There was no baseline condition (no explanation) because explanations’ importance to recommender transparency is well-established [10, 24, 61, 114, 99, 172, 106, 110]. In signing up, participants provided two topics of interest for their feed topics.

## Procedure

We first presented participants with a condition-specific slide tutorial and then instructed them to navigate to a specified link to access the recommender. Clicks were recorded in a log file. Participants started the recommendation feed about their chosen topic with 4 seed papers, found using keyword search, and named and generated the feed. The participants' objective was to make the feed as relevant to them as possible. They had 15 minutes to do so, but if they felt that it was not going to become any more relevant before 15 minutes passed, they stopped early. The number of seed papers and time for each feed were chosen so that the study would not be too long and fatiguing. We also asked participants to think aloud as they interacted with the system in case there were any helpful insights into their interactions or they needed a reminder of how to use a certain system feature. At the end of each condition, participants filled out a Google Forms survey without looking at the system. The survey asked for short answers regarding in what situations, if any, they found each explanation type useful. It also asked for any other thoughts on the explanations. Next, they answered the Likert-type questions in Table 3.1. Lastly, they returned to the feed and categorized the final top ten papers as relevant, neutral, or irrelevant. However, this data depended heavily on factors other than successful feed curation (e.g. the number of papers published on the feed topic), so we did not use it.

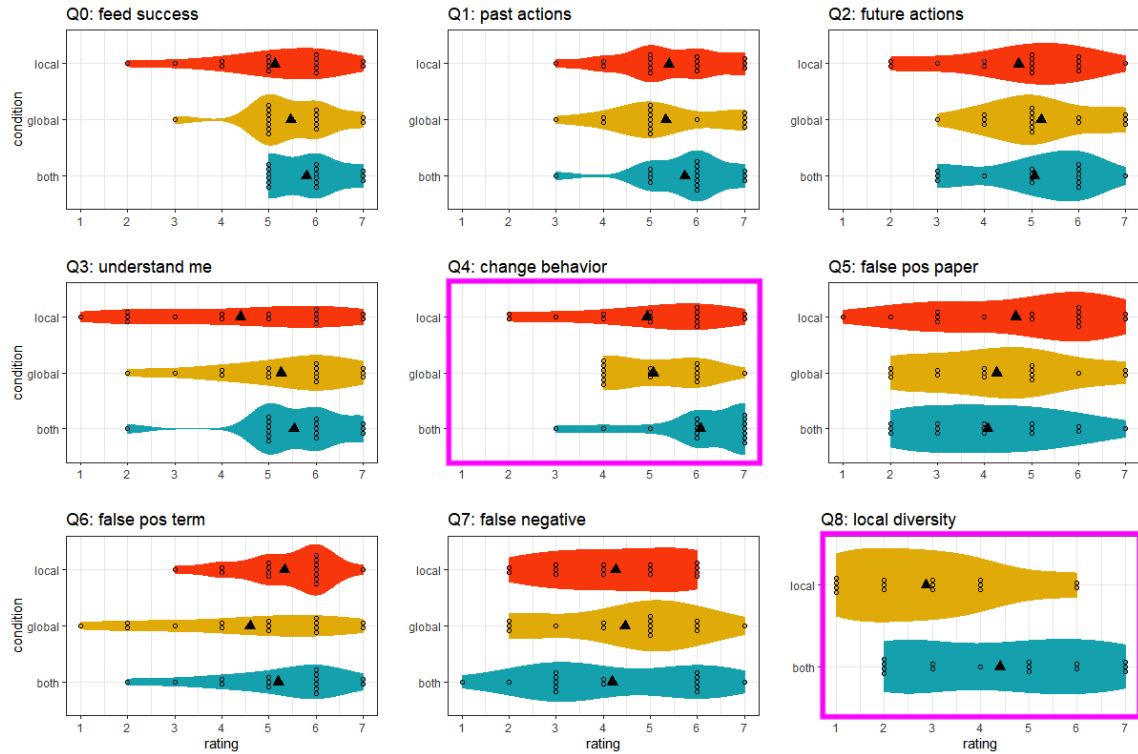


Figure 3.3: Study 2 results for each Likert-type question and condition. 1 means “strongly disagree,” and 7 means “strongly agree.” Triangles represent the mean responses for each question/condition. Circles represent individual responses. Q4: With both explanations rather than only global ( $p=0.015$ , uncorrected) or local ( $p=0.030$ , uncorrected), participants were more confident in explaining how they would like the system to improve. Q8: Participants desired less diverse local explanations when global was present ( $p=0.038$ , uncorrected).

### 3.3.2 Results and Discussion

#### Quantitative Results

Here we discuss results for **RQ1**'s related hypotheses and metrics (Table 3.1). For Likert-type questions, we compared the L and G conditions with the within-subjects two-tailed Wilcoxon signed-rank test and the remaining condition pairs with the between-subjects two-tailed Mann-Whitney-Wilcoxon test (Figure 3.3). For log file metrics, we analyzed all condition pairs with a one-way ANOVA test. The significance threshold was  $p < 0.05$ . Though all results were insignificant after Bonferroni corrections, results for **H4** and **H7** would be significant otherwise.

Regarding **H4**, participants in the LG condition demonstrated more confidence than participants in the G ( $W=55.5$ ,  $p=0.015$ , uncorrected) or L ( $W=61.5$ ,  $p=0.030$ , uncorrected) condition in explaining how they would like the system to behave to be more relevant. However, there was no difference indicated between the L and G conditions. This suggests that **both explanations together are better than either alone for helping users understand how the recommender can improve**. While similar results have been shown in other machine learning systems [74, 81], this is a distinct insight for recommenders, as their output is not objectively correct or not. Judging and rating output according to their own standards, users may benefit differently from local and global explanations.

To create appropriately transparent interactions, a designer needs to know what kinds of information users seek from local explanations. **H7**'s result suggests that local explanations' ideal content depends on whether or not a global explanation is

present. In particular, **participants desired *less diverse* and more consistent local explanations when the global explanation was also present** ( $W=62.5$ ,  $p=0.038$ , uncorrected). This may be a consequence of the “explanation-action trade off” [110], which refers to how actionable local explanations without a global explanation in recommenders must balance two competing goals: 1) returning the most accurate (and often consequently more uniform) explanations and 2) affording more opportunities for users to adjust the model. We address this in Semantic Sanity by explicitly introducing randomness to diversify the local explanations, as in Lee et al. [110].

## Qualitative Results

In their short-answer responses, **participants commented more often that they forgot or did not find much use for the local explanations, as compared to the global explanation.** Of the 30 participants, nine mentioned either forgetting local explanations or using them rarely, whereas only one mentioned not using the global explanation. However, this difference may be due in part to a user interface design issue described in Section 3.4.1. Participants also noted that local and global explanations may serve different purposes in terms of research exploration. Four participants explained that **the ability to adjust the importance of the global explanation terms was useful to avoid unintended bias toward specific authors or topics.** P17 noted, “*The system seemed to be suggesting a particular author and listed that in the feed explanation column. I reduced that so that I could have a more unbiased feed of people I don’t often read...*” Two participants mentioned

that **the global explanation allowed them to introspect about their own research interests**. P11 commented, “[*Global*] gave me a better idea of what my inputs... seemed to have in common.” On the other hand, two participants found that **local explanations were useful for characterizing unexpected interesting papers**. P24 wrote, “There was a paper suggested to me that I found relevant, but I was also surprised to find it in my recommendation list... [*Local*] was useful for me to check out why that paper was recommended...”

### 3.4 Study 3: Controlled User Study

#### 3.4.1 Study Design

##### Hypotheses

Study 3’s first research question aimed to reaffirm Study 2’s results. **RQ2**: are both explanations more helpful than either alone for understanding how to improve recommenders? To address **RQ2**, we have the hypothesis **H11**: both are better than either alone for understanding how the recommender may improve. The second research question sought to expand on Study 2’s findings. **RQ3**: how do local and global explanations *complement* one another to help users understand recommender output? To address **RQ3**, we have two hypotheses reflecting a framework for how the two may complement each other. **H9**: Local is better than global for identifying false positive (FP) recommendations, and **H10**: global is better than local for identifying false negative (FN) recommendations. The hypotheses’ associated metrics are provided in Table 3.2 and are described further in Section 3.4.2.

Table 3.2: Study 3 metrics for hypotheses in Section 3.4.1. Scores are described in Section 3.4.2. Question is a 7-point Likert-type question.

<b>Hypo.</b>	<b>Metric ID</b>	<b>Metric</b>
H9	M1	score on false positive survey (between -42 and 42)
H10	M2	score on false negative survey (0 or 1)
H11	Q9	“I can explain how the system should be updated to be more relevant.”

## Participants and Treatments

In the same manner as in Study 2, thirty computer-science researchers were recruited and separated into treatments. A few changes were made to the explanations. Their titles were made purple and revised to better draw attention. The local explanations were renamed from “Paper Explanation” to “Why This Paper,” and the global explanation was renamed from “Feed Explanation” to “Why This Feed.” Also, as described in Section 3.4.1, Study 3’s procedure no longer required participants to curate recommendation feeds, so the only clickable buttons were for looking at explanations and abstracts. The remaining buttons were included to provide context for how the recommender would work overall. Furthermore, the LG condition was updated so that the local and global explanations were presented separately because their unified presentation in Study 2, in which local explanations could only be opened one-by-one, may have reduced focus on the local explanation. Study 2 participants in the L condition opened an individual local explanation 9.3 times on average, while those in the LG condition did so only 2.7 times on average.

Study 2 participants’ feed topics varied largely in breadth and familiarity, which may have hindered our ability to observe significant results. Thus, Study 3 partici-

pants were randomly assigned to one of two preset feeds for each condition: “misinformation on social media” or “educational technologies for demographically diverse users.” These topics were chosen based on three criteria: 1) for researchers from various areas to engage with and understand the feed, the topic had to use limited jargon, 2) the topic had to be specific enough that FPs occurred, and 3) the topic had to be broad enough that a FN cluster emerged. For example, in the “misinformation on social media” feed, true positives (TPs) were exclusively about *covid-related* misinformation, so any papers discussing misinformation on social media unrelated to covid formed a FN cluster. Each feed was seeded with five carefully selected papers.

Three annotators classified the top 20 papers of each 250-paper feed as FPs or TPs and the bottom 50 papers of each feed as FNs or TNs, based on paper titles and abstracts. Only papers upon which there was unanimous agreement were added to the pool of papers that participants could encounter. The local explanations for each annotated paper were then kept constant so that no new randomized terms were introduced for diversification. Subsequently, the twenty-first paper from the “educational technologies for demographically diverse users” feed was added to the pool of papers in order to have enough TP papers for the study. Also, the “misinformation on social media” feed had ten FNs. Two did not belong to the FN cluster about misinformation on social media *unrelated to covid*. To ensure that all participants interacting with this feed would see a FN from the same cluster, the two FNs were removed from the pool of papers.

## Procedure

Participants first opened a link to the recommender. For each condition, they then logged into one of two accounts to access a preset feed with six recommendations. Next, we gave them a condition-specific tutorial on using the system. The participants then answered three Google-Forms surveys to address each hypothesis. **H9** was addressed first with a FP survey. The survey asked participants to label each of the six paper recommendations in the feed as relevant or not and rate how confident they were in their answers on a 7-point scale. The recommendations were randomly ordered and selected such that half would be FPs. About half of all the TPs had optimal local explanations containing information pertinent to both aspects of the given feed topic. For instance, in the “misinformation on social media” feed, the optimal local explanation may have the term “fake news” related to “misinformation” as well as the term “twitter” related to “social media.” To make sure this category of TP was represented accordingly, one such TP was randomly chosen to be included in each participant’s feed. **H10** was addressed next with a FN survey. The survey presented participants with three new paper recommendations for the feed. Two were TPs and one was a FN. The survey asked participants to rank these papers based on how they believed the recommender system *would rather than should* rank them. Ideally, the participant would be able to recognize that the FN paper would be ranked last by the system. Finally, **H11** was addressed with a survey asking participants to answer the 7-point Likert-type question **Q9**. The survey also asked participants to explain to a software developer how to make the recommendations more relevant, but we found that participants did not understand this question as

intended, so it was discarded.

### 3.4.2 *Results and Discussion*

Here we discuss the results for the hypotheses and metrics in Table 3.2. The FP survey score **M1** was calculated as follows. For each recommendation, if the participant classified it correctly as relevant or not to the feed topic, 1 multiplied by their confidence (1 to 7) was added to their score. If they classified it incorrectly, -1 multiplied by their confidence was added. The FN survey score **M2** was 1 if the FN paper was ranked below the two TP papers and 0 if not. For **M1** and **M2**, we analyzed all condition pairs with a one-way ANOVA test. For **Q9**, we compared the L and G conditions with the within-subjects two-tailed Wilcoxon signed-rank test and the other condition pairs with the between-subjects two-tailed Mann-Whitney-Wilcoxon test. The significance threshold was  $p < 0.05$ . All pre-registered results were insignificant.

Regarding **RQ2**, the slight difference in wording between **H11**'s **Q9** and Study 2's **Q4** may have implied that, to respond affirmatively, the participant needed a technical rather than merely conceptual understanding of how the system could improve. Also, they may have had more trouble conceptualizing how it should improve, as they did not choose the feed topic. These points may explain the lower average response to **Q9** (4.67) as compared to **Q4** (5.36).

Regarding **RQ3**, while we did not find a significant difference among the conditions with respect to how well participants completed the FP (**H9**) or FN (**H10**) survey, we observed uncorrected significant differences among the conditions in terms of

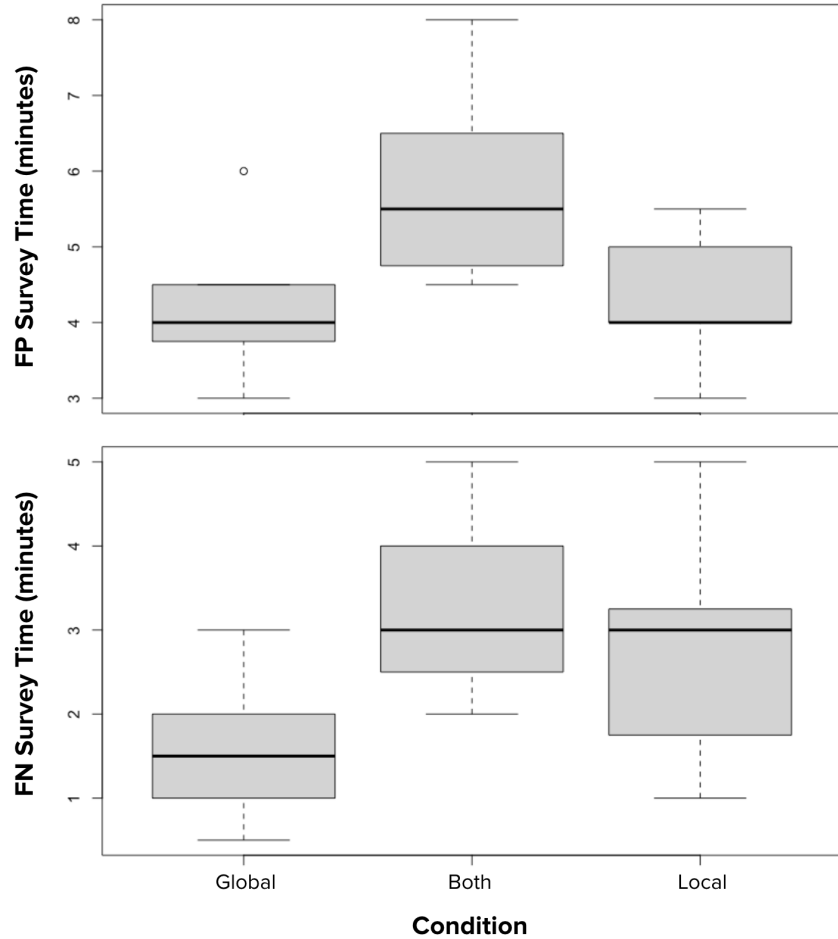


Figure 3.4: How much time Study 3 participants spent on the FP (top) and FN (bottom) surveys as a function of condition, under the “misinformation on social media” feed. Top: Participants spent more time on the FP survey when both explanations were present as compared to only global ( $p=0.020$ , uncorrected) or only local ( $p=0.045$ , uncorrected). Bottom: Participants spent less time on the FN survey when only global was present as compared to both ( $p=0.018$ , uncorrected) or only local ( $p=0.135$ , uncorrected). The global explanation alone thus appears more helpful than both explanations together for identifying FPs and FNs efficiently.

how *quickly* participants completed the FP ( $F(2,19)=5.216, p<.05$ ) and FN ( $F(2,19)=4.801, p<.05$ ) surveys for the “misinformation on social media” feed, as illustrated in Figure 3.4. Twenty-two participants interacted with this feed (8 in the G condition, 7 in each of the other conditions). Time spent on each survey was rounded to the nearest half-minute. For these results, we analyzed all condition pairs with a one-way ANOVA test followed by a Tukey HSD test. These results are not further corrected because they were not pre-registered for analysis. Figure 3.4 shows that participants with the “misinformation on social media” feed in the LG condition completed the FP survey slower than those in the G ( $p=0.020$ , uncorrected) and L ( $p=0.045$ , uncorrected) conditions. This suggests that **providing both explanations rather than either alone causes users to identify FPs more slowly**, which may simply be due to the fact that there is more information to consider when both explanations are available. However, this result is not obvious, as having both explanations could have allowed participants to more easily recognize FPs. For the same feed’s FN survey, Figure 3.4 shows that participants in the G condition completed it faster than participants in the LG condition ( $p=0.018$ , uncorrected), suggesting that **providing only a global explanation rather than both explanations helps users identify FNs more quickly**. Though insignificant, participants in the G condition also completed the survey faster than participants in the L condition ( $p=0.135$ , uncorrected). These results make sense for two reasons: 1) with LG, users have more information to evaluate, and 2) in comparison to L, G’s top terms provide users a more straightforward indication of which terms the model may be considering too important or unimportant, which can cause FNs.

Reasons we did not see the same results for the “educational technologies for demographically diverse users” feed may include: 1) the topic appeared more difficult to understand, 2) the FNs resulted from an over-specification (for cultural diversity) as opposed to an unnecessary specification (for covid), and 3) the FN-related global terms were less prominent.

However, in a follow-up formative study with time-constrained surveys, participants were not evidently better at identifying FPs or FNs in one condition versus another. The time constraints may have been too severe or too relaxed to see a difference.

### **3.5 Conclusion, Limitations, and Future Work**

Following a formative study to determine how content-based local and global explanations should be presented in a research-paper recommender system, we conducted an exploratory study comparing the two explanation approaches’ uses in this system. We saw evidence suggesting that each explanation type plays a unique role in augmenting the system’s transparency around what literature components influence the recommendations. Specifically, our results suggested that providing both explanations rather than either alone ensures scientists reach the best understanding of how the recommender can improve. Furthermore, we saw evidence that scientists prefer more diverse local explanations when they are alone compared to when a global explanation is also available, which may be attributable to the need for more options to steer the system without a global explanation. We also found qualitative evidence that, in the domain of research papers, local and global explanations may

have different advantages with respect to research exploration.

In a subsequent controlled user study, we investigated how local and global explanations may *complement* one another to help users understand their recommendations, in particular by revealing false positive and false negative recommendations. While we did not find differences between the two explanations in terms of user accuracy in identifying false positives or negatives, we did observe evidence suggesting that having both rather than either alone slows users' identification of false positives, and having a global explanation alone rather than both quickens users' identification of false negatives caused by unnecessary specifications. However, a follow-up formative study did not corroborate these findings.

Limitations of this work include that the user studies were relatively small-scale and that only one recommendation domain (computer-science research papers) and explanation style (content-based) were studied. Future work may study the use of local and global explanations for more opaque recommendations such as author or artist recommendations; an explanation is less necessary if the recommendation itself summarizes its contents, as with papers. Future research may also explore how these explanations are used in higher-stakes recommendation settings such as education or healthcare, in which explanations often bear greater importance. Finally, future work may investigate how these explanations are used for purposes other than clarifying recommendation relevance, such as discovery of more diverse recommendations.

## Chapter 4

# Papers-to-Posts: Supporting Detailed Long-Document Summarization with an Interactive, LLM-Powered Source Outline

In the last chapter, we saw that scientists may benefit more from literature-grounded, creative AI suggestions when they are provided more affordances to verify and control what literature components influence those suggestions. In this chapter, we explore the benefits of an AI system that answers the question of **which** literature components should be directly **included in or excluded from** generated suggestions. To assist with this, we introduce a novel mechanism called an **interactive reverse source outline**, which delineates which literature components are and are not selected to incorporate in generated suggestions. We find that a system implementing this mechanism leads to significantly higher satisfaction with the final creative artifact, under time constraints, compared to a strong baseline without this mechanism.

Compressing long and technical documents (e.g., > 10 pages) into shorter-form articles (e.g., < 2 pages) is critical for communicating information to new audiences, for example, translating scientific research papers into blog posts or dense court proceedings into legal briefs. While large language models (LLMs) are powerful tools for condensing large amounts of text, current interfaces to these models lack support for understanding and controlling which content is included versus not included in a detailed summary article. Recognizing and adjusting excluded information is

especially important for detail- and technical-oriented domains, in which tactical selection and coherent synthesis of key details is critical for effective communication to the target audience. For this, we present interactive reverse source outlines, a novel mechanism for verifying and steering long-form summarization through an LLM-generated outline with automatic bullet point selections that the user can iteratively adjust to obtain the desired content coverage. We implement this mechanism in PAPERS-TO-POSTS, a new LLM-powered system for authoring research-paper blog posts. Through a within-subjects lab study (n=20 scientists) and a between-subjects deployment study (n=37 blog posts, 26 scientists), we compare PAPERS-TO-POSTS to a strong baseline tool that provides an LLM-generated draft and access to free-form prompting. Under time constraints, PAPERS-TO-POSTS significantly increases writer satisfaction with blog post quality, particularly with respect to content coverage. Furthermore, quantitative results show an increase in editing power (change in text for an amount of time or writing actions) while using PAPERS-TO-POSTS, and qualitative results show that scientists find incorporating key paper insights into their blog posts easier while using PAPERS-TO-POSTS.

## 4.1 Introduction

Condensing long and technical documents into shorter forms for different audiences is crucial for many applications, from blog posts about scientific papers to legal briefs about court proceedings. In science, communicating scientific ideas from research papers to broader audiences increases awareness and impact [88, 20]. Scientists often seek different ways to publicize their work, such as through tweets or blog posts [57,

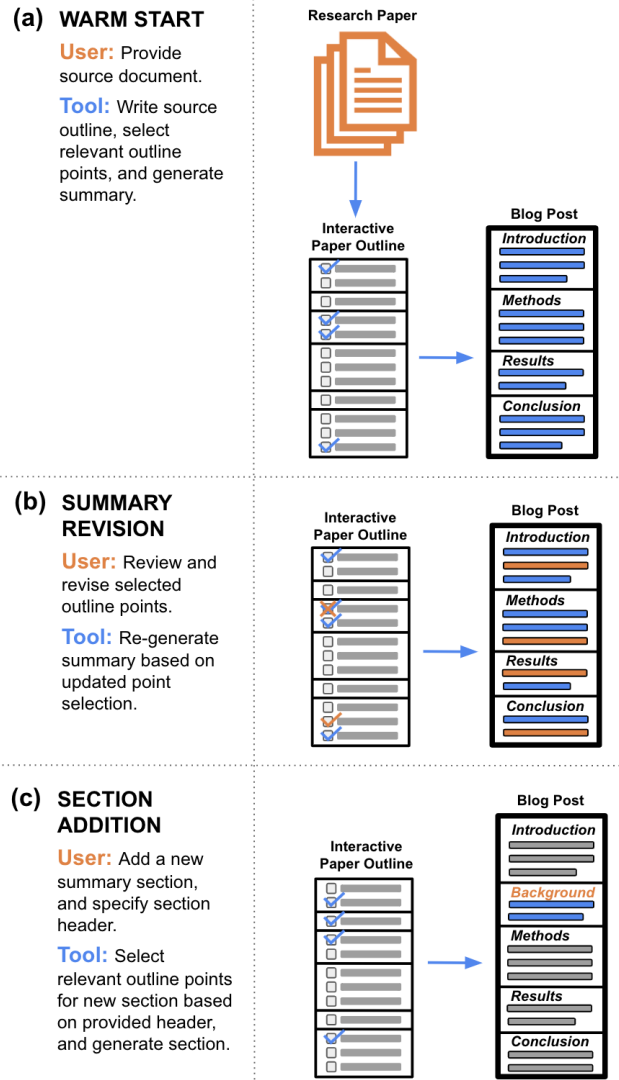


Figure 4.1: *Interactive reverse source outlines* in the PAPERS-TO-POSTS system. Users input a long-form source document (research paper) and the system provides (a) a WARM START for the summarization task by generating a reverse document outline. The system then produces a draft summary article (blog post) with sections based on outline bullet points (and associated source paragraphs) selected by the LLM. Given the draft, users can perform two main actions: (b) SUMMARY REVISION, where users adjust the system’s bullet point selection for the summary, triggering system re-generation of the summary; and (c) SECTION ADDITION, where users provide a header for a new desired section, based on which the system selects bullet points and generates a draft of the section content. Users may also edit the draft manually.

120, 184, 75]; indeed, tweeting about research papers has been shown to raise their citation count [125]. While some tools have been created to help author short-form derivatives like tweets [120], significant challenges remain for scientists seeking to create more detailed derivatives like blog posts. Summarizing long documents (10k+ words) in detailed articles (500-1000 words) takes considerable time and effort, even with LLM support. Detailed articles require the writer to identify large amounts of relevant content from the source document. Given a long document, identifying all the important information to include can be daunting. While LLMs have impressive abilities to synthesize information from a long piece of text, what the LLM chooses to leave out is unclear to the user, and quickly iterating on the selected content is difficult. In domains like science, law, and business, attention to detail is critical; a summary missing key content could lead to misrepresented work and misinformed readers.

While some prior work has investigated fully automatic summarization of long documents [102], a mixed-initiative approach allows users to have more control over their summaries, which is important in detail-oriented domains like scientific research. Prior work in human-AI text summarization has often focused on helping create *short-form* summaries around a paragraph in length, which are easier to control through selection of a few source-document sentences [194, 182]. However, when writing longer summaries, selecting source content at the sentence level becomes unwieldy. Other work in human-AI text summarization has looked into other interactions that may support longer-form summaries, but these works involve interactions detached from the source narrative, like post-editing of an automatically

generated summary [116, 108, 134] or selecting concepts to include in the summary [192, 12, 8, 7, 45]. Without sufficient context from the source document, making sure that important details are not missing in a longer summary is difficult.

Inspired by prior work on LLM-supported reverse outlining for general writing [40], we present *interactive reverse source outlines*, a new interaction mechanism for controlling summary content selection with an interactive LLM-generated source outline, which maintains the source’s narrative structure and simplifies the task of selecting relevant content (Figure 4.1). To the best of our knowledge, this is the first mixed-initiative approach for *detailed* and *grounded* summarization of long documents (Figure 4.2).

We implement this mechanism in PAPERS-TO-POSTS, a mixed-initiative system that helps researchers write research-paper blog posts. The system follows a plan-draft-revise writing workflow [52]. In the planning phase, the user may review the system’s initial blog post draft and adjust the system’s selection of key bullet points within an LLM-generated paper outline. In the drafting phase, the user may customize preset instructions (e.g., include a hook, write one paragraph) to the LLM for each blog post section. The LLM then uses the selected bullet points and adjusted instructions to generate new text for the given section. In the revising phase, the user may further refine text with LLM-powered macros that encapsulate frequent editing transformations for scientific blog post writing (e.g., to reduce jargon, to better capture attention).

We evaluated PAPERS-TO-POSTS through two studies in which participants wrote blog posts for their own papers. In these studies, we compared our tool to a strong

baseline tool—an LLM-generated draft and access to free-form LLM prompting. In a within-subjects lab study (N=20 participants) in which researchers wrote blog posts for two papers (one per tool), we found that PAPERS-TO-POSTS led to significantly higher writer satisfaction with the final blog post, and participants were particularly more satisfied with the content covered in the final blog post. Furthermore, participants had higher confidence that all essential information was in the final blog post when using PAPERS-TO-POSTS. To study PAPERS-TO-POSTS in a more realistic setting, we also conducted an unmonitored, between-subjects deployment study (N=37 blog posts, 26 participants). Across both studies, participants found PAPERS-TO-POSTS more helpful. They found it easier to incorporate research-paper content in the blog post and to iterate on that content. Moreover, participants demonstrated increased editing power (change to the provided blog post draft for a given amount of time or for a given number of writing actions) without an increase in cognitive load.

In summary, we make the following contributions:

- *Interactive reverse source outlines*, a novel mechanism for more user control in human-LLM detailed summarization of long documents, consisting of an LLM-generated reverse outline of the source document with pre-selected bullet points for an initial draft of the summary, with which the writer iteratively interacts to control content selection
- PAPERS-TO-POSTS, a tool that implements this mechanism for the purpose of writing research-paper blog posts, a common long-document summarization task that requires attention to detail

- Findings from a within-subjects lab study (N=20 participants) and a between-subjects deployment study (N=26 participants) showing that PAPERS-TO-POSTS...
  - Under time constraints, significantly increases writers’ satisfaction with their summaries’ quality (primarily with respect to content coverage)
  - Is considered more helpful by writers, making incorporating source content in a detailed summary easier and iterating on the summary easier, with increased editing power (change in writing within an amount of time or writing actions) and no increase in cognitive load

## 4.2 Related Work

### 4.2.1 Human-LLM Text Summarization

Although there is much work on automatic text summarization [102, 49], issues remain with the output summaries, such as inaccuracies and style misalignment [123, 102]. To address these issues, other works have looked into supporting humans in working together with LLMs to generate a summary [28]. These tools may be divided into groups in terms of 1) the summary length that they support and 2) how grounded the interaction is in the context of the source document (Figure 4.2).

Some prior work in human-LLM text summarization supports an interaction in which the human selects sentences extracted directly from the source [182, 194] (Figure 4.2, top left). For a short summary, identifying the few key sentences from the source to include may be manageable, but for a detailed summary, this would re-

## Affordances for Control of LLM Text Summarization

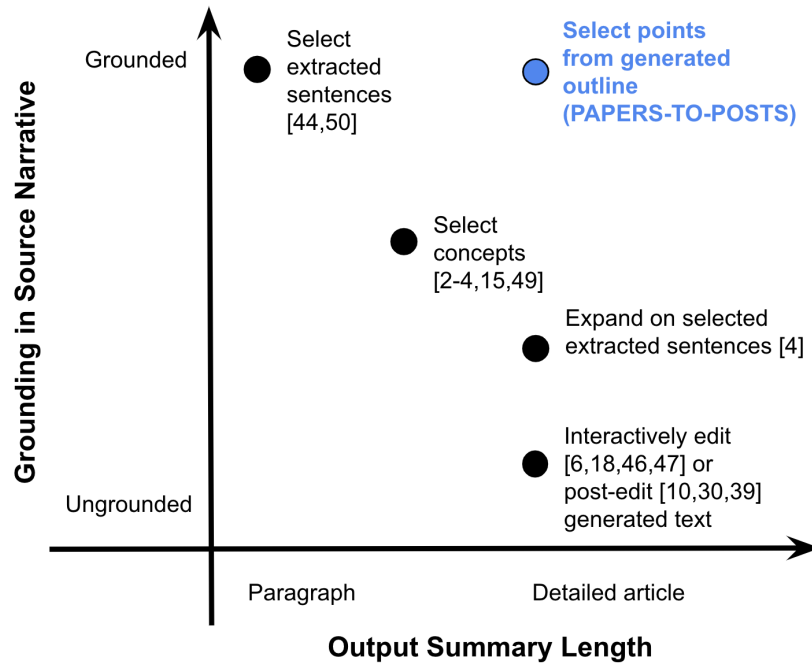


Figure 4.2: Comparing interactive reverse source outlines in PAPERS-TO-POSTS (blue, upper right) with other affordances for control of LLM summarization, in terms of 1) how grounded the affordance is in the source narrative and 2) the length of the output summary that is supported. PAPERS-TO-POSTS is the first LLM-powered tool to provide writers with highly grounded control over a detailed summary.

quire considerable time and effort. One work enables more detailed summaries by generating expanded text based on the user selecting a few source sentences [12] (Figure 4.2, middle right), but what source content has been included versus neglected then becomes unclear. Other prior work provides users with automatically generated summaries that they can adjust through interactive editing (e.g., via ratings, gaze) [14, 56, 185, 186] or post-editing [134, 22, 108] (Figure 4.2, bottom right). While these interactions are flexible in terms of the length of summary generated, they do not support users in understanding what source content has or has not been represented. Yet another group of prior works strikes a balance between generating longer summaries and grounding them in the source document, by having the user select concepts or keywords to include in the summary [192, 12, 8, 7, 45] (Figure 4.2, center). Concepts are easier to sort through than full sentences, which makes generating a longer summary more feasible, and selecting concepts provides users some sense of what source content is utilized overall. Nevertheless, concepts are high-level and abstract, limiting their utility for generating detailed summaries and easily identifying what source details are and are not included. Our novel interactive reverse source outlines, implemented in PAPERS-TO-POSTS (Figure 4.2, top right), support both grounded and long-form summarization by allowing users to select outline bullet points directly tied to the source document in order to generate a detailed summary.

### 4.2.2 *Outlines in Human-LLM Writing*

There are some works in human-LLM writing that have utilized LLM-generated outlines for writing. Some focus on outlining output content, rather than outlining input content for summarization [195, 124]. For instance, Zhang et al. introduce the tool VISAR, a human-AI argumentative writing assistant that allows users to iterate on their draft with adjustments to a visual outline of the draft [195]. Other works focus on outlining input content (as we do), but the outline is not utilized to synthesize a detailed summary of the input [146, 189]. As an example, AngleKindling generates a few main points from an input press release as a quick overview for journalists to then ideate about potential angles for reporting on the press release [146]. However, there is no support for interacting with the main points to generate a summary of the press release.

Most related to our work, Dang et al. present a text editor that generates summaries of each paragraph written in the editor to comprise a reverse outline of the writing [40]. The summaries then become affordances to help writers think about how to revise their writing. Our work also generates a reverse outline, but it is an outline of the *input* document for users to interactively guide summary generation. For summarizing a long source document, an outline is particularly useful, as reviewing the entire source itself requires substantial time and effort.

### 4.2.3 *Human-LLM Scientific Writing*

Several works have investigated how to support human-LLM writing. Most of these works focus on tools for creative writing [163, 33, 131], but a few cater to scientific

writing. A few works have investigated human-LLM scientific writing for broader science communication tasks that do not involve a source document. Gero et al. present a system for generating scientific “sparks,” or inspiring sentences [58]. Long et al. examine how LLM scaffolding can help people generate relatable hooks for complex scientific topics [120], and Kim et al. investigate how science writers can generate extended metaphors for scientific ideas with the help of an LLM [96]. The few prior works that have explored human-AI scientific summarization [116, 192, 12] do not support fully grounded and detailed summaries like PAPERS-TO-POSTS, as described in Section 4.2.1. Also of note, Google’s tool NotebookLM for general human-AI document summarization does not permit any content selection [59].

#### *4.2.4 Scientific Blog Posts*

Scientists can engage with broader communities through various channels, including press releases, magazines, journals, and tweets [6, 58]. This work focuses on scientific blog posts, which offer increased recognition and transparency of one’s work, editorial freedom, and public engagement [85, 13, 196]. There are several forms of scientific blog posts, such as academic commentary, free commentary, and mediation of research to laypersons [126]. Here, we focus on blog posts with the primary goal of communicating information from a research paper to people unfamiliar with the paper’s specific research topic.

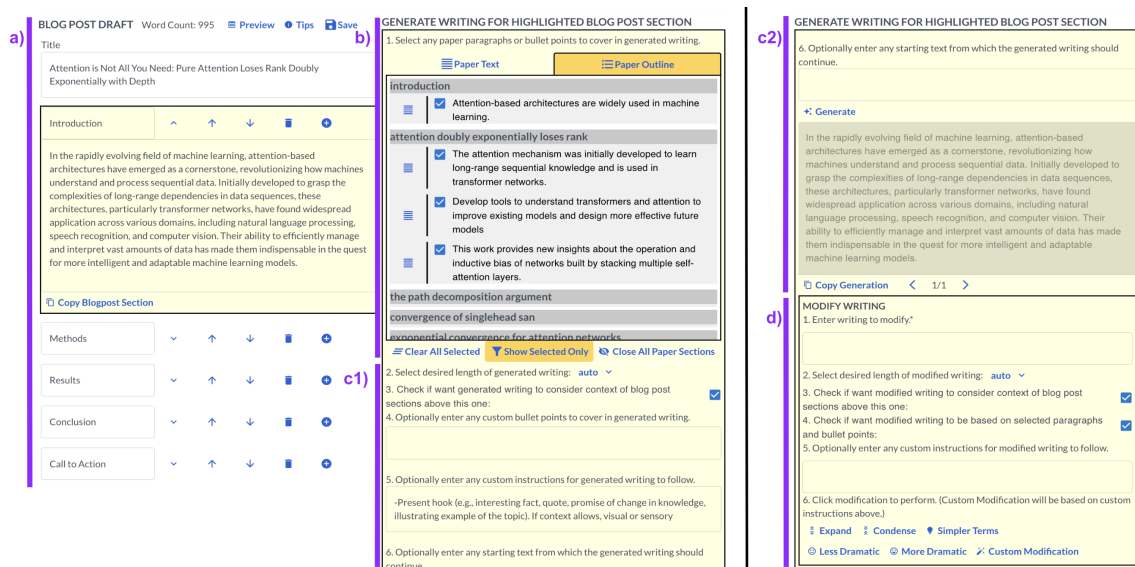


Figure 4.3: PAPERS-TO-POSTS’ user interface. a) The Blog-Post Area, where the user writes the blog post in sections. b) The Planning-Support Area, which contains both the interactive paper outline and original paper text. c1) The Drafting-Support Area, which contains inputs and outputs for generating text. c2) The Drafting-Support Area continued, seen if one scrolls below the c1 area. d) The Revising-Support Area, which contains inputs and outputs for modifying text and is located below the Drafting-Support Area. (Not Pictured: The modification output appears below the modification buttons, with a button to copy the output text and buttons to view previous modifications and their inputs.)

### 4.3 The Papers-to-Posts System

This section describes the design goals for PAPERS-TO-POSTS and provides a walk-through of the resulting system’s four main panel areas, which show the blog post draft and enable executing the Plan-Draft-Revise workflow on individual blog post sections.<sup>1</sup> We developed PAPERS-TO-POSTS based on key design goals and pilot study feedback. The pilot study feedback helped us to identify system features to

1. A walkthrough video of the tool may be found in the supplementary materials.

add for increased transparency, flexibility, and ease of use.

### 4.3.1 *Design Goals*

When communicating scientific research in a detailed summary, making sure all important details are covered is key. Otherwise, the research may be misinterpreted, leading to misinformed readers. Indeed, in a closely related work on human-LLM scientific text summarization, a featured design goal based on their preliminary survey of professors was to provide provenance for summary content by showing “direct or indirect contributors to a summary to help the user verify whether the summary reflects the structure and key components of the original document” [192]. This also applies to similar domains, such as law and business, in which delivering information with accuracy and precision is critical. However, existing text summarization tools (including the one just referenced) either 1) do not fully ground the summary content selection in the source, which makes understanding what content is versus is not included difficult, or 2) only support selection of summary content at the level of source sentences, which is unwieldy when writing a detailed summary (see Section 4.2.1).

Furthermore, we identified and synthesized guidelines for academic blog posts from popular resources [105, 175, 145, 34], prior work [127], and feedback from three communications experts, recruited from our institutions. There are multiple guidelines on the content, format, and language for academic blog posts. PAPERS-TO-POSTS seeks to address these guidelines by default but does not enforce them.<sup>2</sup>

---

2. The synthesized guidelines may be found in the appendix.

Thus, PAPERS-TO-POSTS’ design goals are as follows (science domain-specific details in parentheses):

- **DG1:** Help users (researchers) to understand what long-document (research-paper) content is versus is not included in its detailed summary (blog post).
- **DG2:** Help users (researchers) to control and iterate on what long-document (research-paper) content is included in a detailed summary (blog post).
- **DG3:** Help users (researchers) to follow domain-specific guidelines (guidelines for academic blog posts), as a starting point.

### *4.3.2 Implementation Details*

The frontend of PAPERS-TO-POSTS was developed using React and TypeScript, while the backend was developed with Python. Unless otherwise noted, the LLM was GPT-4 in the lab study and GPT-4-0125-preview in the deployment study.

### *4.3.3 Walkthrough of Warm Start Step*

The **Blog Post Area** (Figure 4.3a) is the user’s main working area for reviewing the target blog post artifact. At the top, the user views the blog post’s word count, a preview of the full blog post, and tips for writing blog posts (described in Section 4.3.1). The user then reviews the blog post draft, divided into sections. To mitigate cognitive overload, the user selects one section at a time to focus on. The selected section is the target of subsequent Plan-Draft-Revise workflow actions, described in the other step descriptions below.

In order to obtain the initial blog post draft for the user’s input research paper, we developed an LLM-powered pipeline that first generates a reverse source outline, selects relevant points, and then generates a draft (Figure 4.1a).

## Preprocessing

We used a public tool for converting scientific PDFs to HTML<sup>3</sup> to extract the paper text from the paper PDF. For the lab study, we then manually adjusted the parsed HTML so that the section headers were correct, no section or subsection was completely missing, and only the main paper text was included (i.e., no appendix, references, or footnotes). We also manually retrieved the tables and figures from the parsing. For the deployment study, the text parsing was automatic to better align with a real-world setting, and there was no table or figure retrieval.

## Outline generation

We used GPT-3.5 Turbo to generate between one and three bullet points to summarize each paragraph in the parsed paper, depending on the length of the paragraph. While we used GPT-4 (GPT-4-0125-preview in the deployment study) for all of PAPERS-TO-POSTS’ other tasks, we decided to use GPT-3.5-Turbo for this task for two reasons: 1) it produced shorter bullet points that were quicker to read, and 2) it was many times cheaper for this token-intensive generation task. All together, the generated bullet points formed a paper outline.

---

3. <https://papertohtml.org>

## Outline point selection

We provided the LLM with all of the paper’s bullet points along with the request to select 10 bullet points relevant to the initial four blog post sections: introduction, methods, results, and conclusion. We requested 10 bullet points per blog post section in order to provide the model sufficient content without overwhelming it.

## Summary generation

We prompted the LLM to generate each section in succession using the selected bullet points for that section, the paragraphs associated with the selected bullet points (to better ground the model), the portion of the blog post that had already been generated, and section-specific guidelines (see Section 4.3.1). We observed that this process of generating one draft section at a time with the context of prior sections helped to make the draft more coherent while avoiding overwhelming the model with instructions for multiple sections. The model was prompted to make each section around 125 to 250 words and around one to three paragraphs.<sup>4</sup>

### *4.3.4 Walkthrough of Planning Step*

In the **Planning-Support Area** (Figure 4.3b), the user works with PAPERS-TO-POSTS’s interactive reverse source outline to plan the blog post section on which they are currently working, addressing **DG1** and **DG2**. All of the user’s planning actions (as well as drafting and revising actions) are connected to the user’s selected blog

---

4. The prompts for generating the initial draft are in Appendix A.2.2. An example initial draft and associated selected bullet points are in Appendices A.6.1 and A.6.2.

post section. In other words, the state and history of all inputs and outputs for text generation are stored separately for each section. This is somewhat similar to the fragmented document history proposed by Buschek [21].

The user sees the interactive reverse source outline, divided into sets of bullet points, each corresponding to a paragraph from the original text, which can be viewed by clicking the button to the left of the bullet point set.<sup>5</sup> They review the tool’s pre-selected bullet points for the current blog post section by scrolling through the full outline or clicking the “Show Selected Only” button.<sup>6</sup> The user adjusts the selection of bullet points (or paragraphs) for the LLM to use in generating text for the current blog post section (Figure 4.1b). Selecting a paragraph does not affect the selection of associated bullet points, and vice versa, from the user’s perspective. The system, however, does receive associated paragraphs with selected bullet points for grounding.

#### 4.3.5 *Walkthrough of Drafting Step*

In the **Drafting-Support Area** (Figure 4.3c), the user works with the tool to synthesize text for the blog post section on which they are currently working, addressing **DG2** and **DG3**.

---

5. In the lab study, the tool presented figures/tables in a final collapsible portion of the paper outline. Their captions could be copied for pasting in the editor to help the user plan where to put figures in the final blog post.

6. When this filter button is activated and the user is under the Paper Text tab, if a paragraph is not selected but a corresponding bullet point is selected, the tool notes that “A corresponding bulletpoint(s) is still selected”; the opposite is true for the Paper Outline tab.

## Section Revision

At the top of the Drafting Support Area, the user changes the desired length of the generated text to their liking. The length defaults to “auto” and may be updated to one sentence, one paragraph, or a few paragraphs. Below, the user sees the toggle is on to provide the LLM with the context of blog post sections above the one on which they are working. In the “Custom Bulletpoints” text field, the user enters any additional bullet points that they want their generated text for the blog post section to cover. In the “Custom Instructions” text field, the user adjusts the instructions for the section’s generated text. For each of the initial draft sections, this box is pre-filled with section-specific instructions, which are based on blog post guidelines we collected on content and formatting for different academic blog post sections [105, 175, 145, 34, 127]. In the text field labeled “Starting Text to Continue from”, the user optionally provides text that should be at the start of the generated text.

The LLM then generates revised text for the section at hand with the user’s adjusted selection of bullet points and paragraphs as well as their adjusted section-specific instructions (Figure 4.1b).<sup>7</sup> With respect to the selected bullet points, the model receives the context of the associated paragraph for each selected bullet point, even if the associated paragraph was not selected by the user. This is to ensure that the model’s generation is grounded in the actual paper text. When the user selects the “auto” length for generated text, the model is instructed to generate text that is 125 to 250 words and between one and three paragraphs.

---

7. The prompts for drafting text may be found in Appendix A.2.4, and example outputs are in Appendices A.6.1 and A.6.2.

The user copies and pastes all or part of the generated text into the highlighted blog post section and edits it as they see fit. If the user wants to remember what they generated previously for this section, they may peruse a history of their inputs (paper content selection, LLM instructions) and outputs (text generations).

## Section Addition

The user can also add new sections to the blog post (Figure 4.1c). In the Blog Post Area, the user clicks the plus-sign button next to an existing blog post section, and the tool places the new section below the existing one. The tool gives the user the option to either create a blank section or generate a section based on the section header that they provide. To generate a new section, the LLM selects 10 bullet points for the section based on the provided header and generates text based on those bullet points and associated paragraphs.

### *4.3.6 Walkthrough of Revising Step*

In the **Revising-Support Area** (Figure 4.3d), the user works with the tool to revise the blog post section on which they are currently working, addressing **DG3**. At the top of this area, the user pastes the text that they want to modify in the text field. This text may be anything, including text from the blog post draft, the paper, or a model generation. The user then adjusts the desired length of the modified text if need be, as in the Drafting-Support Area (Section 4.3.5).<sup>8</sup> The user leaves the second toggle on, indicating that the LLM will utilize the currently selected

---

8. The desired length “auto” is unique to each modifying action.

paragraphs and bullet points to generate the modified text. The user considers the five preset modification options: 1) expand the text, 2) condense the text, 3) rewrite the text in simpler terms to be more understandable to a layperson, 4) rewrite the text in a less dramatic tone to align better with the unadorned language of scientific writing, and 5) rewrite the text in a more dramatic tone to better capture readers' attention. The preset modification buttons called prompts based on our synthesized guidelines for academic blog posts (see 4.3.1), which suggested avoiding jargon [105, 175, 145, 34], sensationalism (noted by a communications expert), and wordiness [105, 175, 145]. There is also a "custom modification" button that will follow specified custom instructions.<sup>9</sup> The user copies and pastes the modified text into the highlighted blog post section and edits it as needed. If the user wants to remember their previous text modifications for this section, they may peruse a history of their instructions and modifications. With the user's instructions, the LLM modifies the user's provided text snippet.<sup>10</sup>

## 4.4 Evaluation Studies

### 4.4.1 Lab Study

We conducted a within-subjects lab study to determine if and how PAPERS-TO-POSTS may support researchers in writing blog posts about their research papers.

---

9. If the user utilizes a preset modification button, they may still provide custom instructions for the LLM to take into consideration.

10. The prompts for revising text may be found in Appendix A.2.4, and example outputs are in Appendix A.6.3.

## Hypotheses

We hypothesized that, compared to a baseline consisting of an LLM draft and the ability to prompt the LLM, PAPERS-TO-POSTS would provide participants a better experience and outcome. Our individual hypotheses were as follows:

- **H1:** PAPERS-TO-POSTS leads to higher satisfaction with the output blog post than the baseline.
- **H2:** PAPERS-TO-POSTS leads to higher satisfaction with the tool than the baseline.
- **H3:** PAPERS-TO-POSTS leads to lower cognitive load than the baseline.
- **H4:** PAPERS-TO-POSTS leads to less time spent to generate a blog post that the writer would be comfortable posting publicly, in comparison to the baseline.

## Participants

We recruited 20 participants (M: 14, W: 6) via academic social networks and institutional mailing lists and compensated them with \$83.33 USD for 2.5 hours of their time. Participants were predominantly early career researchers (PhD student: 17, master’s student: 1, postdoc: 1, industry researcher: 1), who are often lead authors tasked with writing blog posts. All participants conducted research in computer science. Participants had a broad range of experience in terms of LLM use (>10 sittings: 14, 6-10 sittings: 3, 2-5 sittings: 2, <2 sittings: 1). Most had not authored a research-paper blog post before (0 posts: 14, 1-2 posts: 5, 3-10 posts: 1). In order to

simulate a naturalistic setting, participants were required to have authored at least 2 research papers for use in the study (2 papers: 1, 3-10 papers: 14, >10 papers: 5) and to be more interested than not in writing blog posts for their research papers.

## Study Conditions

In the treatment condition, participants interacted with PAPERS-TO-POSTS described in Section 4.3.<sup>11</sup> The tool in the baseline condition was designed to be a simplified version of PAPERS-TO-POSTS without affordances developed to address the design goals (Figure 4.4). The baseline provided the input paper’s text, sourced from the PDF and divided into paragraphs and collapsible sections as in PAPERS-TO-POSTS. The baseline also provided the ability to prompt GPT-4 with any instructions and copy its output. A history of generations and their associated instructions was retained for the user’s perusal. Lastly, the baseline had an editor for writing the blog post, which was pre-populated with an LLM draft generated by GPT-4-32k using the entirety of the input paper.<sup>12</sup> GPT-4-32k was used rather than GPT-4 because the context window for GPT-4 was insufficient for longer research papers. The prompt was designed to mirror the prompt for PAPERS-TO-POSTS’ initial draft sections (Appendix A.2.1), so the resulting draft also had an introduction, methods, results, and conclusion section. However, the prompt did not contain section-specific instructions.

In each condition, the tool was set up for one of the participant’s two research

---

11. The system figures above show how the tool looked for the deployment study, after minor usability updates (see Section 4.4.2. Appendix A.20 shows how the tool looked for the lab study.

12. An example of a baseline initial draft is in Appendix A.5.

### ### Introduction

In the rapidly evolving field of machine learning, attention-based architectures, particularly transformers, have revolutionized how we approach problems in natural language processing, speech recognition, and computer vision. Despite their widespread adoption, the underlying reasons for their effectiveness remain somewhat of a mystery. Our research aims to shed light on this by exploring the inner workings of self-attention networks (SANs), a core component of transformer models. We introduce a novel perspective by demonstrating that the output of SANs can be decomposed into a sum of simpler terms, each corresponding to a sequence of attention heads across layers. This decomposition allows us to uncover a strong inductive bias towards "token uniformity" inherent in self-attention mechanisms, providing new insights into their operation and effectiveness.

### ### Methods

Our approach begins with a detailed analysis of self-attention networks built exclusively from multi-head self-attention layers. By expressing the output of these networks as a sum of paths, where each path is a sequence of attention heads, we can study the behavior of SANs in a novel way. We prove that, in the absence of skip connections and multi-layer perceptrons (MLPs), SANs converge doubly exponentially to a rank-1 matrix, leading to a loss of expressive power. This convergence is significantly faster than what classical analyses predict, highlighting the unique dynamics of self-attention. To validate our theoretical findings, we conducted experiments on various transformer architectures, examining the effects of architectural components like skip connections and MLPs on the convergence behavior.

### ### Results

Our experiments confirmed the theoretical predictions: pure self-attention networks without skip connections or MLPs rapidly converge to a rank-1 matrix, indicating a loss of diversity in the output. This phenomenon, termed "rank collapse," was mitigated by the introduction of skip connections, which prevent the output from degenerating. Furthermore, MLPs were found to slow down the convergence, acting as a counterbalance to the rank collapse induced by self-attention. These findings highlight the critical roles of skip connections and MLPs in maintaining the expressive power of transformers.

## PAPER

### introduction

Abstract Attention-based architectures have become ubiquitous in machine learning. Yet our understanding of the reasons for their effectiveness remains limited. This work proposes a new way to understand self-attention networks: we show that their output can be decomposed into a sum of smaller terms, each involving the operation of a sequence of attention heads across layers.

Using this decomposition, we prove that self-attention possesses a strong inductive bias towards "token uniformity". Specifically, without skip connections or multi-layer perceptrons (MLPs), the output converges doubly exponentially to a rank-1 matrix. On the other hand, skip connections and MLPs stop the output from degeneration.

Our experiments verify the identified convergence phenomena on different variants of standard transformer architectures 1.

### attention doubly exponentially loses rank

The attention mechanism [BCB15] was initially developed to better learn long-

[Close All Paper Sections](#)

## GENERATE WRITING

Instructions \*

Enter any instructions for the AI tool.

[Generate](#)

Figure 4.4: The baseline tool's user interface, consisting of areas for writing the blog post (left), viewing the paper (top right), and providing instructions to the LLM for generating writing (bottom right).

papers submitted in the recruitment survey. Each paper was required to have been published online between 2018 and the study period in 2023, with the participant as an author. Only one paper was published in 2018, and the rest were published during or after 2020. There was no significant difference in publication dates between conditions (Paired-Samples *t*-Test,  $t(19)=0.90$ ,  $p=n.s.$ ). The average input paper length across conditions differed by less than 10% (baseline: 8452 words, treatment: 7742 words).

## Study Procedure

The study was within-subjects, so each participant completed two study sessions, one in the treatment condition and one in the baseline, in randomized order and following a counter-balanced design. The participants' selected papers were also assigned to conditions in randomized order. Each study session lasted up to about 75 minutes. For sufficient flexibility, participants selected two study sessions within a week of each other, but the majority of sessions were a day apart (1 day: 13, 2 days: 3, 3 days: 3, 4 days: 1).<sup>13</sup>

Participants completed a participation agreement before the study. We recorded and transcribed each session in Google Meet.<sup>14</sup> To start, the session coordinator introduced the task and presented a video tutorial on the assigned tool.<sup>15</sup> Participants then had two minutes to interact with and ask questions about the tool in the

---

13. The condition order was counter-balanced within the group of participants with sessions one day apart and within the group of participants with sessions over one day apart.

14. The lab study session script may be found in the supplementary materials.

15. The tutorial scripts are in the supplementary materials.

context of a sample paper that was not their own. Next, participants opened a tool link initialized with one of their research papers. The coordinator asked them to assume that their audience was people in their broad area (i.e., computer science), which was the most common audience for whom potential study recruits wanted to write blog posts (52/60). The coordinator instructed participants to work as if they were not there, but to let them know if any questions about the tool arose. They also instructed participants not to use any external tools or information.<sup>16</sup> Lastly, the coordinator asked participants to let them know if they reached a point at which they would feel comfortable publicly publishing the blog post in its current state.

Each participant had up to one hour to work on the blog post. At the 20-minute and 40-minute marks, they were reminded to let the coordinator know if they reached the point of comfort with publication. If the participant reached that point, the coordinator asked them to save their draft and continue revising the blog post. After the first four study sessions, we noticed that participants were sometimes done early and did not desire to revise further. Thus, we added an additional rule that the participant would let the coordinator know if they reached a point at which they were proud of the blog post and saw no reason to revise further. If the participant reached that point, or if an hour had passed, they submitted their final blog post and moved on to a survey regarding their experience with the tool. The survey consisted of Likert-type questions regarding perceived efficiency, cognitive load, ease of using the tool in relation to the design goals, satisfaction with the tool, satisfaction with the

---

16. We made an exception twice when participants wanted to view their paper's PDF due to issues with its parsing in the tool.

final blog post, perceived control, and perceived ownership.<sup>17</sup> There were also short answer questions regarding the difficulties and benefits of using the tool in relation to the design goals.<sup>18</sup> At the session’s end, the coordinator told the participant to close the tool and not to access it before the next session if they had one. Each time the participant prompted the LLM or submitted the blog post, a log of the action along with the state of the blog post was saved for analysis.

#### *4.4.2 Deployment Study*

We conducted a deployment study to understand if and how the results of the lab study translate to a less controlled and more realistic setting. To make the deployment study more realistic than the lab study, participants wrote blog posts on their own time with the explicit goal of sharing the blog post with others (e.g., on social media, in an email to colleagues).

### Research Questions

Having observed that the lab study participants had significantly more editing power when using PAPERS-TO-POSTS rather than the baseline tool, we investigated if PAPERS-TO-POSTS leads participants to have greater editing power in this real-world context (**RQ1**). As in the lab study, we also evaluated satisfaction with the final blog post, satisfaction with the tool, and cognitive load (**RQ2**). We did not

---

17. See Appendix A.1 for a discussion of the exploratory results regarding participants’ perceived control and ownership.

18. The survey may be found in the supplementary materials.

evaluate task completion time in the deployment study, as the lab study did not demonstrate any interesting difference between the conditions.

- **RQ1:** How is participants' editing power impacted by PAPERS-TO-POSTS versus the baseline tool in a real-world setting?
- **RQ2:** How do participants perceive the outcome (blog post satisfaction) and experience (tool satisfaction, cognitive load) of interacting with PAPERS-TO-POSTS versus the baseline tool in a real-world setting?

For the deployment study, we posed research questions rather than evaluated hypotheses for statistical significance because power analyses indicated that we would need a larger sample size than we could obtain to see the same significant results from the lab study.<sup>19</sup>

## Participants

We recruited 26 participants (M: 17, W: 8, undisclosed: 1) through academic social networks and institutional mailing lists. They were compensated with \$50 over PayPal for each blog post that they wrote, with a maximum of two blog posts. Participants were predominantly early career researchers (undergraduate student: 1, master's student: 4, PhD student: 13, postdoc: 2, professor: 2, industry researcher: 4). All participants conducted research in an area of computer science, except for one who studied cognitive science. The majority of the participants had interacted

---

19. The significant editing power result would have required 29 participants per condition (power=0.95, alpha=0.05, effect size=0.45), and the significant blog post satisfaction result would have required 83 participants per condition (power=0.95, alpha=0.05, effect size=0.98).

with an LLM several times (>10 sittings: 21, 6-10 sittings: 4, 2-5 sittings: 1). The participants had varied experience in writing blog posts (0 posts: 16, 1-2 posts: 9, 3-10 posts: 1). They were required to have authored at least one research paper (1-2 papers: 6, 3-10 papers: 11, >10 research papers: 9) and to be more interested than not in writing research-paper blog posts. We also required participants to be willing to share their final blog posts with others if they were able to reach the point at which they would feel comfortable doing so. Lastly, the papers that the participants wanted to write about had to be compatible with our tools' PDF parser.

## Study Conditions

While the treatment and baseline conditions were largely the same as in the lab study, several minor usability issues identified in the lab study were addressed before proceeding with the deployment study. For example, there were better instructions and less intrusive highlight colors, as depicted in the system figures above. We also replaced GPT-4 and GPT-4-32k with the newer GPT-4-0125-preview model, which had a larger context window and became available after the lab study. This enabled both conditions to support longer inputs. In the deployment study, participants could freely close and reopen their assigned tools in their browsers to work on their drafts when convenient, and they could save their progress across sessions with a newly added save button. Furthermore, the treatment tool now automatically parsed the input paper. We note that, due to the manner in which text was logged in the deployment study, the tool in both conditions had a lag if participants pasted a chunk of text more than a couple paragraphs long.

The majority of the 37 papers about which participants wrote were published in 2023 or 2024 (26 papers). The two oldest papers were from 2019 and 2008. The average parsed paper length across conditions differed by less than 5% (baseline: 37064 characters, treatment: 35845 characters).

## Study Procedure

The task for the deployment study was to share a blog post with others, which can require substantially higher effort than *voicing* comfort in publicly posting the blog post, as in the lab study. Considering this difference for participant recruitment, we opted for a between-subjects design, so each participant only had to write one blog post. Still, we allowed participants who wished to participate twice (N=11/26 participants) to use both tools in randomized, counter-balanced order. Their selected papers were assigned randomly to each condition. When analyzing these participants' results, we treated them as unpaired data. Eighteen participants completed the baseline, and 19 completed the treatment. Participants had up to six days to complete a paper blog post using the assigned tool on their own time. We provided a link to the tool, a video tutorial on the tool<sup>20</sup>, and a link to the paper about which to write. We noted that participants should aim to share the final blog post with others, but if they felt that they could not reach that point, they would still be compensated. Participants were required to spend at least 20 minutes using the tool. We recorded participants' interactions, including prompting and text edits, in logs. In addition, we also took snapshots of the blog post whenever they (re)opened

---

20. The scripts for the video tutorials are in the supplementary materials.

the systems in their browsers and every 60 seconds during active usage.

Once participants were satisfied with their draft, they shared it. We provided instructions on how to create blog posts on the Medium platform,<sup>21</sup> but they were allowed to use other publishing platforms if they preferred. They then emailed us the blog post link and a screenshot or link showcasing evidence that they shared the blog post with others (e.g., tweet, email to co-authors, Slack channel message, LinkedIn post). Only 2 of the 26 participants (one under both conditions and one under the treatment) could not reach the point at which they felt comfortable sharing the blog post with others in the allotted time.<sup>22</sup> Participants then completed the same survey from the lab study regarding their experience (see Section 4.4.1).<sup>23</sup> The only difference was the removal of a few questions irrelevant to this study. Lastly, the first author engaged willing participants (15) in a 15-minute semi-structured interview to discuss their experience and survey responses.

## 4.5 Evaluation Studies' Results

In this section, we present the quantitative and qualitative results of the lab and deployment studies in parallel. To obtain the qualitative results, the first author conducted inductive thematic analysis [18] of the participants' interview and short-answer responses across the two studies. Please note that the hypotheses from

---

21. <https://medium.com/>

22. We note that some participants indicated that the blog post was part of an experiment when sharing with others, which may have increased their willingness to share it; regardless, they were willing to share a post at the end of the study.

23. We note that four participants who we realized had not completed 20 minutes of interaction were asked to do so after completing the survey.

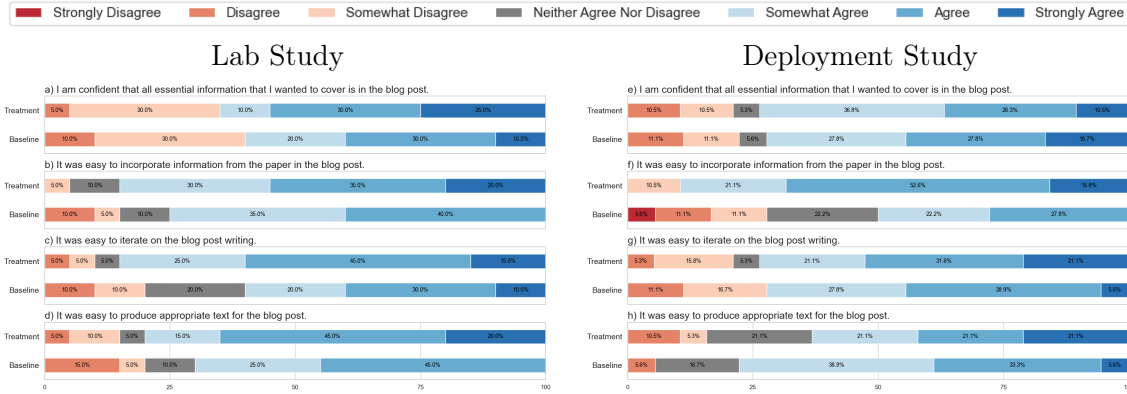


Figure 4.5: Survey responses to 7-point Likert-type questions regarding design goals in the a-d) lab study and e-h) deployment study. Responses are shown for both the treatment and baseline conditions.

the within-subjects lab study were not re-evaluated for significance in the between-subjects deployment study because there were not enough participants to do so (see Section 4.4.2).

#### 4.5.1 Design Goals 1 and 2

We analyzed participants' Likert-type responses to questions related to fulfilling our first two design goals, which focused on helping researchers understand and control what content is included in their research-paper blog posts (Figure 4.5a-c,e-g).

#### Lab Study - Exploratory

Based on the median results, using PAPERS-TO-POSTS, participants were only slightly more confident that all essential information was in the blog post (treatment:  $M=6.00$ ,  $Q1=3.00$ ,  $Q3=6.25$ ; baseline:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=6.00$ ). They found it easier to

incorporate information from the paper (treatment: M=6.00, Q1=5.00, Q3=6.00; baseline: M=5.00, Q1=4.75, Q3=6.00), and they found it easier to iterate on the blog post text (treatment: M=6.00, Q1=5.00, Q3=6.00; baseline: M=5.00, Q1=4.00, Q3=6.00) (Figure 4.5a-c). The interaction logs showed that participants utilized PAPERS-TO-POSTS' affordance designed to support these outcomes, selecting relevant paper bullet points and paragraphs from the LLM-generated source outline to incorporate in the blog post. All participants adjusted the LLM's pre-selected content to some degree. On average, participants added or removed 28 bullet points and 13 paragraphs.<sup>24</sup> Participants also added their own custom bullet points an average of 1.70 times.

## Deployment Study - Exploratory

Looking at the median results, participants did not indicate an improvement in feeling confident that all essential information was covered with PAPERS-TO-POSTS (M=5.00, Q1=4.50, Q3=6.00) as opposed to the baseline (M=5.00, Q1=4.25, Q3=6.00). Nevertheless, participants still indicated an increased ease in incorporating information from the paper in the blog post with PAPERS-TO-POSTS (M=6.00, Q1=5.00, Q3=6.00) compared to the baseline (M=4.50, Q1=3.25, Q3=5.75). In addition, the median participant still indicated that PAPERS-TO-POSTS (M=6.00, Q1=4.50, Q3=6.00) facilitated iterating on the blog post more so than the baseline (M=5.00, Q1=3.50, Q3=6.00).

---

24. Some added bullet points come from automatic selection of bullet points used to generate a new section rather than manual selection. These averages do not include 2 participants due to a logging error.

## Summary

In both studies, the median participant using PAPERS-TO-POSTS indicated increased ease in incorporating paper content in the blog post and iterating on the blog post text. However, the two studies presented little evidence of a difference between the conditions with respect to participants' confidence in all essential information being covered in the blog post. That said, participants were still more satisfied with their blog posts' content coverage under time constraints (see Section 4.5.4). Perhaps participants were happier with the content they were able to include under time constraints but still needed more assurance that all essential information was covered.

### *4.5.2 Editing Power*

As DG2 focuses on helping researchers to iterate on the content of their blog posts, we conducted a related investigation of how participants' editing power (change in writing for a given amount of time or writing actions) changed between conditions.

#### Lab Study - Exploratory

Lab study participants made greater change to the initial draft blog post when using PAPERS-TO-POSTS as compared to the baseline tool. For each condition, we calculated the Levenshtein distance between the initial draft blog posts and participants' blog posts at the point at which they would feel comfortable publicly publishing the blog post. Levenshtein distance measures how many character insertions,

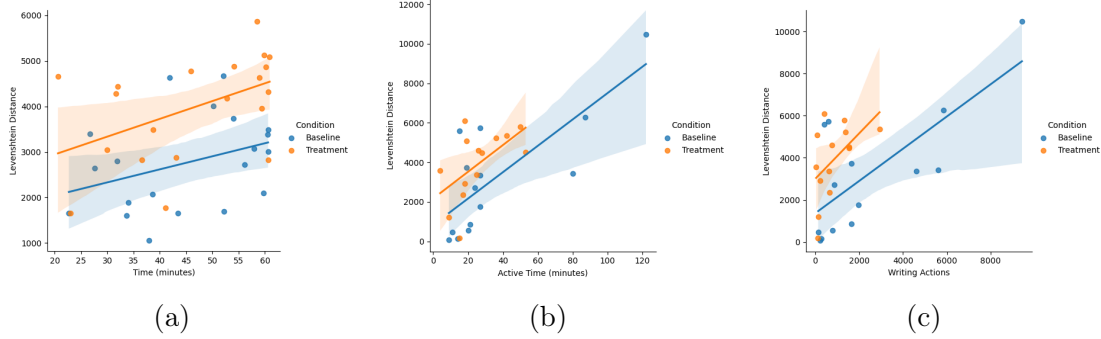


Figure 4.6: Per condition, Levenshtein distance between a) initial blog post and blog post comfortable with publishing publicly as a function of time spent using the tool (lab study), b) initial and final blog posts as a function of active time spent using the tool (deployment study), and c) initial and final blog posts as a function of number of writing actions taken (deployment study). Across the two studies, participants using PAPERS-TO-POSTS consistently showed more editing power, or change in the blog post for a given amount of time or writing actions. Note that the lab study duration was controlled (up to 60 minutes), whereas deployment study participants wrote on their own time.

deletions, and replacements would be needed to transform one text into another. For statistical testing, we employed a paired-samples  $t$ -test and observed that this Levenshtein distance is significantly higher with PAPERS-TO-POSTS ( $M=3977.05$ ,  $SD=1146.46$ ) compared to the baseline ( $M=2764.15$ ,  $SD=1042.16$ ) (Paired-Samples  $t$ -Test,  $t(19)=4.31$ ,  $p<.0005$ ).<sup>25</sup> The difference in Levenshtein distance does not appear attributable to the time spent on the task or the change in blog post length. Participants did not spend substantially more time to reach the point at which they were comfortable publicly publishing the blog post using PAPERS-TO-POSTS as compared to the baseline (see Section 4.5.7). Figure 4.6a shows Levenshtein distance plotted

<sup>25</sup>. We note that this significance test was not part of our initial hypotheses and was therefore exploratory in nature.

against the time taken to get to a publishable blog post, with the removal of one outlier. Furthermore, participants changed the length of their blog posts, measured in characters, less with PAPERS-TO-POSTS (M=-162.00, Q1=-1126.25, Q3=297.00) than with the baseline (M=1971.00, Q1=965.25, Q3=2514.25).<sup>26</sup>

## Deployment Study - RQ1

In the deployment study, we investigated editing power in terms of not only amount of time taken but also amount of writing actions taken. We define writing actions as insertions and deletions of a character or text span in the draft or LLM instructions.<sup>27</sup> To analyze editing power, we excluded the three study runs in which the participant made no change to the initial draft. We also excluded six study runs in which the participant claimed to have made edits to the blog post outside the tool, beyond small edits like those related to grammar, formatting, and adding images. The resulting analysis thus included 28 study runs (treatment: 14; baseline: 14).

Participants using PAPERS-TO-POSTS took fewer writing actions to create one unit of change in Levenshtein distance to the blog post (Figure 4.6c). PAPERS-TO-POSTS also led to a greater final Levenshtein distance for a given amount of active time, which we define as time during which participants were taking any actions in the tool (Figure 4.6b). Once more, this is despite the fact that participants did not spend substantially more active time using PAPERS-TO-POSTS (M=25.71 minutes,

---

26. As you can see in Appendix A.18b, the initial length of treatment drafts was significantly higher than the initial length of baseline drafts.

27. We did not count when participants moved, inserted, or deleted a blog post section in PAPERS-TO-POSTS as a writing action because these actions involved clicking a button rather than manipulating text.

SD=14.72) than the baseline (M=35.93 minutes, SD=34.38) (Appendix A.19a).<sup>28</sup> Furthermore, participants changed the length of their blog posts, measured in characters, less using PAPERS-TO-POSTS (M=-5.50, Q1=-525.25, Q3=961.00) than using the baseline (M=710.00, Q1=47.25, Q3=4261.25).<sup>29</sup>

## Summary

In the lab study, our exploratory analysis revealed that PAPERS-TO-POSTS led to increased editing power, in the sense that participants made more change to their blog posts in a given amount of time. In the deployment study, participants experienced the same outcome in a more realistic setting, and we further found that participants made more change to their blog posts for a given number of writing actions. This aligns with the fact that participants found iterating on their blog posts easier with PAPERS-TO-POSTS.

### 4.5.3 *Design Goal 3*

We analyzed participants' Likert-type responses to questions related to fulfilling our third design goal, which focuses on supporting researchers in following guidelines for academic blog posts (Figure 4.5d,h).

---

28. Considering that there were three outliers with a high amount of active time in the baseline condition, we can also look at the median behavior. We again see little difference between the conditions (treatment: M=22.00, Q1=17.25, Q3=34.00; baseline: M=22.50, Q1=16.00, Q3=27.00).

29. As you can see in Appendix A.19b, the initial length of treatment drafts was significantly higher than the initial length of baseline drafts.

## Lab Study - Exploratory

Participants' found it easier to produce appropriate blog post text (treatment: M=6.00, Q1=5.00, Q3=6.00; baseline: M=5.00, Q1=4.00, Q3=6.00) in the treatment condition (Figure 4.5d). Furthermore, the interaction logs for the treatment condition indicate that participants took advantage of affordances designed for generating and modifying text in alignment with the guidelines for academic blog posts (Appendix A.17). For an average of 6.85 generations per treatment condition, participants often included custom instructions (5.70 times); though participants may have modified the custom instructions, they were pre-filled by default to better fulfill the guidelines. In addition, participants modified text 6.85 times on average, with the “condense” (1.85 times), “custom” (1.80 times), and “simpler terms” (1.40 times) modifications being the most common.

## Deployment Study - Exploratory

There was no difference in the median for how well PAPERS-TO-POSTS (M=5.00, Q1=4.00, Q3=6.00) versus the baseline (M=5.00, Q1=5.00, Q3=6.00) facilitated production of appropriate text, unlike in the lab study (Figure 4.5e-h).

## Summary

While lab study participants thought PAPERS-TO-POSTS improved their ability to produce appropriate text for the blog post, deployment study participants did not. Perhaps without the pressure of time constraints, deployment study participants

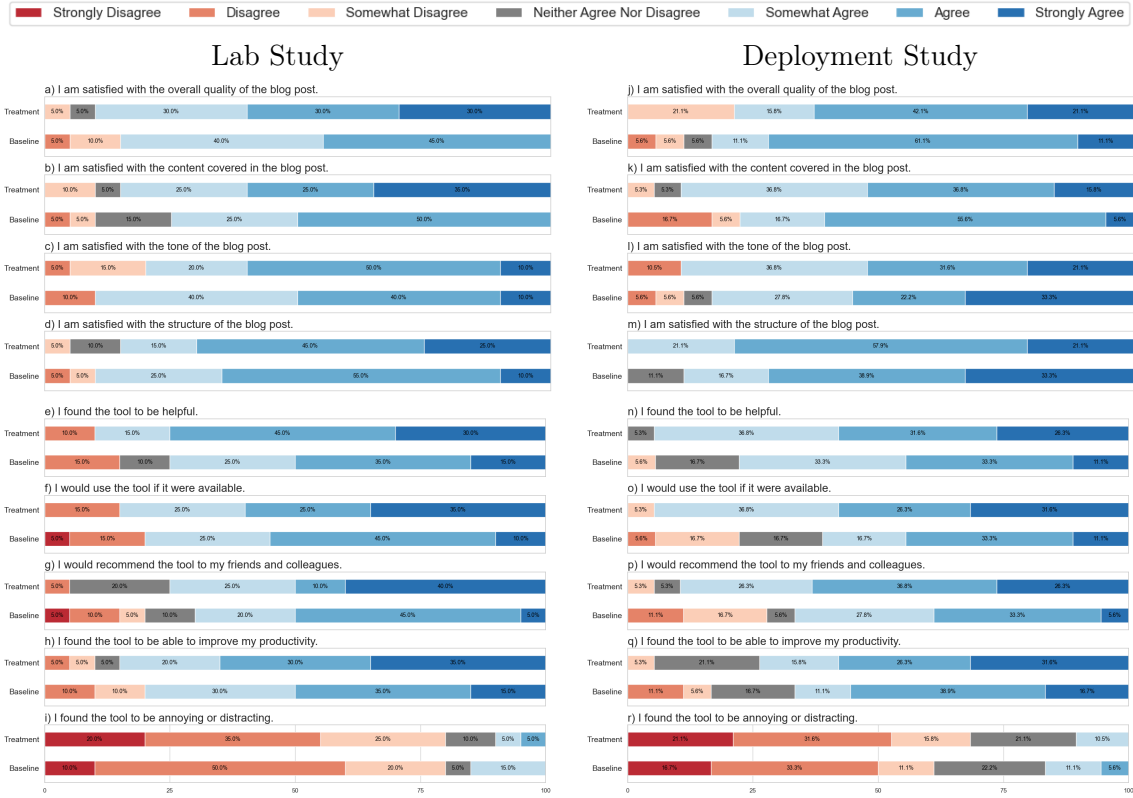


Figure 4.7: Survey responses to 7-point Likert-type questions regarding satisfaction with the a-d) output blog post in the lab study, e-i) tool in the lab study, j-m) output blog post in the deployment study, and n-r) tool in the deployment study. Responses are shown for both the treatment and baseline conditions.

found it easier to determine how to prompt the baseline tool or adjust its generated text in order to produce appropriate writing.

#### 4.5.4 Blog Post Satisfaction

To measure participants' satisfaction with their final blog posts, we utilized their Likert-type response to the statement "I am satisfied with the overall quality of the blog post" (Figure 4.7a,j).

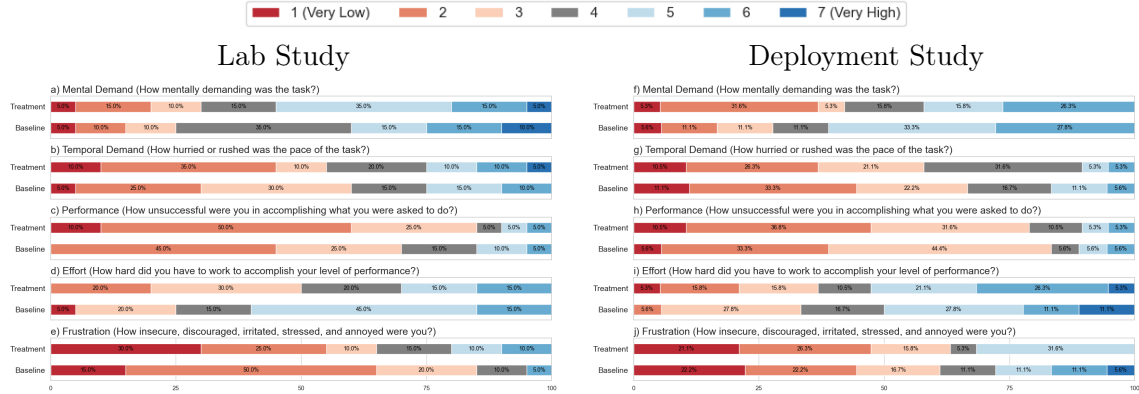


Figure 4.8: Survey responses to 7-point Likert-type questions regarding cognitive load in the a-e) lab study and f-j) deployment study. Responses are shown for both the treatment and baseline conditions. Please note that for the performance question, the labels for the left and right end of the scale were actually respectively “Perfect Performance” and “Failure Performance.”

### Lab Study - H1

To compare responses under the two conditions, we employed a non-parametric test, given the data’s Likert nature. Furthermore, we used a sign test rather than a Wilcoxon signed rank test, which is normally used to analyze non-parametric within-subjects data for two conditions of one factor, due to the violation of that test’s assumption of a symmetric distribution of the paired differences about the median. Participants were significantly more satisfied with the final blog post generated with PAPERS-TO-POSTS (M=6.00, Q1=5.00, Q3=7.00) as compared to the baseline tool (M=5.00, Q1=5.00, Q3=6.00) (Dependent-Samples Sign Test, S=11, p<.05). The majority of participants (11) were more satisfied in the treatment condition (equally: 7, less: 2).

Inspecting participants’ Likert responses regarding aspects of their blog post

satisfaction, we see that participants particularly demonstrated more satisfaction with the content covered in the PAPERS-TO-POSTS blog post (M=6.00, Q1=5.00, Q3=7.00) compared to the baseline blog post (M=5.50, Q1=4.75, Q3=6.00) (Figure 4.7b). The majority of participants (14) were more satisfied with the content under the treatment condition (equally: 2, less: 4). The median participant was only slightly more satisfied with the blog post’s tone under the treatment (M=6.00, Q1=5.00, Q3=6.00) compared to the baseline (M=5.50, Q1=5.00, Q3=6.00) (Figure 4.7c). With respect to the blog post’s structure, the median participant was equally satisfied using PAPERS-TO-POSTS (M=6.00, Q1=5.00, Q3=6.25) compared to the baseline (M=6.00, Q1=5.00, Q3=6.00) (Figure 4.7d).

In addition, looking at participants’ perceived failure to complete their task (Figure 4.8c), we observe that the median participant had lower perceived failure with PAPERS-TO-POSTS (M=2.00, Q1=2.00, Q3=3.00) than with the baseline (M=3.00, Q1=2.00, Q3=4.00). The plurality of participants (9) thought that they were more successful in writing a research-paper blog post with PAPERS-TO-POSTS compared to the baseline (equally: 7, less: 4). This further corroborates participants’ blog post satisfaction under the treatment condition.

## Deployment Study - RQ2a

In the deployment study, we observed no difference in the median blog post satisfaction under PAPERS-TO-POSTS (M=6.00, Q1=5.00, Q3=6.00) versus the baseline (M=6.00, Q1=5.25, Q3=6.00) (Figure 4.7j). Again with respect to median satisfaction, participants were equally satisfied with the blog post content (treat-

ment: M=6.00, Q1=5.00, Q3=6.50; baseline: M=6.00, Q1=3.50, Q3=6.00; Figure 4.7k), tone (treatment: M=6.00, Q1=5.00, Q3=6.00; baseline: M=6.00, Q1=5.00, Q3=7.00; Figure 4.7l), and structure (treatment: M=6.00, Q1=6.00, Q3=6.00; baseline: M=6.00, Q1=5.25, Q3=7.00; Figure 4.7m) when using PAPERS-TO-POSTS versus the baseline. In terms of perceived failure to complete their task (Figure 4.8h), participants had equal median perceived failure with PAPERS-TO-POSTS (M=3.00, Q1=2.00, Q3=3.00) and the baseline (M=3.00, Q1=2.00, Q3=3.00).

## Summary

We observed a significant increase in blog post satisfaction under the treatment condition during the lab study and found that participants were particularly more satisfied with the blog post content, suggesting that PAPERS-TO-POSTS best addressed its design goals to support understanding (DG1) and control (DG2) of the content included in the blog post. We did not observe the same increases in blog post satisfaction in the deployment study. This discrepancy may be related to the fact that participants were not writing under the pressure of a time constraint. With sufficient time to write the blog post to their liking (only three blog posts were not shared with others), participants may not have noticed a pronounced difference in blog post quality between conditions.

### 4.5.5 *Tool Satisfaction*

For measuring tool satisfaction, we used the five-question Likert scale on tool acceptance from Kocielnik et al. [101] (Figure 4.7e-i). The five answers were averaged to

obtain one score.

## Lab Study - H2

We used the sign test to evaluate the lab study data, as it was non-parametric data and violated the Wilcoxon signed rank test's assumption of a symmetric distribution of the paired differences about the median. While the median participant had a higher satisfaction with PAPERS-TO-POSTS (M=5.90, Q1=5.20, Q3=6.50) than the baseline tool (M=5.60, Q1=4.80, Q3=6.00), this difference was not significant (Dependent-Samples Sign Test, S=13, p=n.s.).

Based on median results, participants under the treatment condition more often found the tool helpful (treatment: M=6.00, Q1=5.75, Q3=7.00; baseline: M=5.50, Q1=4.75, Q3=6.00) and useful for productivity (treatment: M=6.00, Q1=5.00, Q3=7.00; baseline: M=5.50, Q1=5.00, Q3=6.00). Meanwhile, participants were equally likely under both conditions to use the tool if it were available (treatment: M=6.00, Q1=5.00, Q3=7.00; baseline: M=6.00, Q1=5.00, Q3=6.00), recommend the tool (treatment: M=5.50, Q1=4.75, Q3=7.00; baseline: M=5.50, Q1=4.00, Q3=6.00), and find the tool annoying or distracting (treatment: M=2.00, Q1=2.00, Q3=3.00; baseline: M=2.00, Q1=2.00, Q3=3.00). For each of the first four questions, at least four more participants increased rather than decreased their Likert rating from the baseline to the treatment condition. However, only two more participants indicated decreased rather than increased annoyance with PAPERS-TO-POSTS (less: 8, more: 6, equal: 6).

## Deployment Study - RQ2b

Similarly to the lab study, we observed little increase in the median tool satisfaction under the treatment condition (M=5.60, Q1=5.00, Q3=6.40) compared to the baseline condition (M=5.40, Q1=4.05, Q3=5.80) (Figure 4.7n-r). Based on median results, participants under the treatment condition were more likely to find the tool helpful (treatment: M=6.00, Q1=5.00, Q3=6.50; baseline: M=5.00, Q1=5.00, Q3=6.00), use it if it were available (treatment: M=6.00, Q1=5.00, Q3=7.00; baseline: M=5.00, Q1=4.00, Q3=6.00), and recommend the tool (treatment: M=6.00, Q1=5.00, Q3=6.50; baseline: M=5.00, Q1=3.25, Q3=6.00). They were also slightly less likely to find the tool annoying (treatment: M=2.00, Q1=2.00, Q3=4.00; baseline: M=2.50, Q1=2.00, Q3=4.00). On the other hand, they were equally likely under both conditions to find the tool useful for productivity (treatment: M=6.00, Q1=4.50, Q3=7.00; baseline: M=6.00, Q1=4.00, Q3=6.00).

## Summary

In both studies, we observed little difference in satisfaction with PAPERS-TO-POSTS versus the baseline tool. However, participants appeared to find PAPERS-TO-POSTS more helpful than the baseline across the two studies.

### *4.5.6 Cognitive Load*

To measure cognitive load, we used the NASA TLX Index questionnaire [70] and summed the responses for the mental demand, temporal demand, effort, and frus-

tration questions (Figure 4.8a-b,d-e).<sup>30</sup>

### Lab Study - H3

We utilized the sign test to evaluate the lab study data, as it was non-parametric data and violated the Wilcoxon signed rank test's assumption of a symmetric distribution of the paired differences about the median. We saw no significant difference between the baseline and treatment conditions (Dependent-Samples Sign Test,  $S=8$ ,  $p=n.s.$ ). The median participant reported slightly lower median cognitive load with PAPERS-TO-POSTS ( $M=13.00$ ,  $Q1=11.50$ ,  $Q3=16.30$ ) than the baseline tool ( $M=14.50$ ,  $Q1=12.00$ ,  $Q3=18.30$ ).

We saw the largest difference in median results with respect to perceived effort (treatment:  $M=3.50$ ,  $Q1=3.00$ ,  $Q3=5.00$ ; baseline:  $M=5.00$ ,  $Q1=3.75$ ,  $Q3=5.00$ ). The majority of participants (11) reported using less effort with PAPERS-TO-POSTS than with the baseline tool (more: 4, equal: 5). Meanwhile, PAPERS-TO-POSTS led to little change in perceived mental demand (treatment:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=5.00$ ; baseline:  $M=4.00$ ,  $Q1=3.75$ ,  $Q3=5.25$ ). The small plurality of participants (9) perceived more mental demand with PAPERS-TO-POSTS (less: 8, equal: 3). We saw no change in the median result for perceived temporal demand (treatment:  $M=3.00$ ,  $Q1=2.00$ ,  $Q3=4.25$ ; baseline:  $M=3.00$ ,  $Q1=2.00$ ,  $Q3=4.25$ ) or frustration (treatment:  $M=2.00$ ,  $Q1=1.00$ ,  $Q3=4.00$ ; baseline:  $M=2.00$ ,  $Q1=2.00$ ,  $Q3=3.00$ ).

---

30. See 4.5.4 for discussion of the responses to the performance question

## Deployment Study - RQ2c

Once again, we found very slightly lower median cognitive load under the treatment (M=15.00, Q1=9.50, Q3=17.50) in comparison to the baseline (M=15.50, Q1=10.75, Q3=18.75) (Figure 4.8f-g,i-j). Considering the median participant, the treatment led to little change in perceived mental demand (M=4.00, Q1=2.00, Q3=5.50) than the baseline (M=5.00, Q1=3.25, Q3=5.75) but higher perceived effort (treatment: M=5.00, Q1=3.00, Q3=6.00; baseline: M=4.50, Q1=3.00, Q3=5.00). We found no change in the median result for perceived temporal demand (treatment: M=3.00, Q1=2.00, Q3=4.00; baseline: M=3.00, Q1=2.00, Q3=4.00) or frustration (treatment: M=3.00, Q1=2.00, Q3=5.00; baseline: M=3.00, Q1=2.00, Q3=4.75).

## Summary

In both studies, we observed little difference in cognitive load from the baseline to the treatment condition. Participants consistently perceived little difference in the tools with respect to mental demand, temporal demand, and frustration. However, their perceptions varied across the studies in terms of effort.

### *4.5.7 Task Completion Time*

## Lab Study - H4

We recorded two metrics for task completion time. One was the amount of time participants needed *during* the session before they would be comfortable publicly publishing the blog post. The other was the perceived amount of time participants

would need *after* the session before they would be comfortable publicly publishing the blog post. As participants reached that point of comfort *during* most sessions, we focused on the former metric. The mean participant worked slightly faster with the baseline tool (M=44.32 minutes, SD=12.51) in comparison to PAPERS-TO-POSTS (M=45.99 minutes, SD=13.88) (Appendix A.18a), but this difference was not significant (Paired-Samples *t*-Test,  $t(19)=0.55$ ,  $p=n.s.$ ).

#### 4.5.8 Qualitative Analysis - DG1 and DG2

Across the two studies, participants' survey and interview comments reflected the difference between the two tools in terms of supporting understanding (DG1) and control (DG2) of what content is versus is not included in the blog post.

Overall, participants using the baseline tool most commonly cited the initial draft (lab: 6/20, deployment: 8/18) as a helpful aspect of the tool. Thus, the most valued baseline feature was not useful for understanding or controlling what content was omitted from the detailed summary. With respect to the initial draft, multiple participants noted editing or adding to the initial draft rather than replacing it. For example, P20-dep explained, "*It's easier to critique than it is to generate,*" and P26-dep shared, "*It was very useful to have a straw-person of what the blog post could look like (generated by the LLM). I mostly used the draft as a place to edit and steer in the direction I wanted the post to take.*" Without an affordance to help them consider all the possibilities of content to include in the blog post, the participants seem to have focused more on augmenting the content already selected by the tool.

With respect to PAPERS-TO-POSTS, participants often noted the selectable paper

bullet points and paragraphs as a helpful feature (lab: 10/20, deployment: 14/19). Several described the outline as helpful for selecting relevant content to generate text. For instance, P15-lab noted, *“I could rewrite entire sections or direct it with just a few bullets.”* Seven participants mentioned that the outline was useful even just for reviewing the paper’s content or structure. P4-lab commented, *“The outline was really helpful in providing a quick overview of what content I had written [in the paper].”* For two lab participants, the bullet points provided transparency around the origins of the initial LLM draft and generations. P12-lab explained, *“I liked the bullet point features to see what was being fed in vs not.”* Thus, participants’ interviews reflected how PAPERS-TO-POSTS’ outline affordance facilitated review and selection of content to include in the blog post.

That said, participants also specified difficulties related to the outline mechanism. For instance, two lab participants were overwhelmed with the number of bullet points. P9-lab described the outline as *“really bloated and hard to navigate.”* Meanwhile, three lab participants were confused or encumbered by aspects of utilizing the bullet points. For example, P3-lab *“felt stressed by choosing between the entire paragraph at hand and the (presumably) AI-generated summary bullets.”* Future work may investigate how to make the outline mechanism more usable.

#### *4.5.9 Qualitative Analysis - DG3*

Across the two studies, participants’ survey and interview comments reflected the difference between the two tools in the tradeoff between supporting flexibility and providing scaffolding for alignment with academic blog post guidelines (DG3).

A number of participants noted the baseline tool’s flexibility as a benefit (lab: 6/20, deployment: 4/18). For example, P18-lab shared, “*I was able to provide instructions that fit my mental model of constructing a blog post and it being there as a ‘copilot’ is just what I need.*” Flexibility in drafting text may help writers obtain the edits they want. However, the writer holds more responsibility in steering the LLM appropriately; P7-lab reflected, “*As a CS student, I like this way more [than the treatment’s way of supporting writing], but I think it needs more thinking to use it creatively.*”

Regarding PAPERS-TO-POSTS, participants frequently mentioned the utility of its text modification abilities, which were designed with the guidelines in mind (lab: 12/20, deployment: 9/19). For example, P18-dep reflected, “*The ‘MODIFY WRITING’ section is helpful. I like that the interface contains this particular section to support iterating on existing text.*” All of the preset modification buttons (“simpler terms”, “condense”, “expand”, “more dramatic”, “less dramatic”) were specifically mentioned as helpful affordances at least once in the lab study alone. The most commonly noted one was the “condense” button. For example, P12-lab commented, “*I sometimes put the modify text through multiple rounds of expand and contract, focusing on different parts. I found that helpful.*” Three participants commented that they desired more preset modification buttons. For instance, P16-lab wanted a “more technical” button, while P14-dep wanted “*more or less casual*” options. P7-lab then added, “*Or even let user to add their favorite rewrite button.*” Interestingly, P2-lab found the preset modification buttons useful for understanding the tool’s capabilities: “*It was nice to know what options the tool already knew of, and would perhaps*

have success with. For example, I would probably not have gone to a blank-box tool and said, 'please make the following more dramatic'. I would have stuck to more basic things like 'please summarize the following'." Several participants (lab: 5/20, deployment: 10/19) also noted the usefulness of at least one aspect of the customizable instructions for either generating or modifying text using PAPERS-TO-POSTS. P9-shared, "...being able to include explicit instructions for the model to generate text from was helpful in being able to control the information in the text that it generated." Thus, participants' comments indicate that PAPERS-TO-POSTS's preset yet flexible LLM instructions for generating academic blog post text provided utility.

Participants also discussed difficulties that they encountered in devising instructions to PAPERS-TO-POSTS. Four lab participants expressed wanting to take specific actions that were not explicitly supported by PAPERS-TO-POSTS, contrasting with the baseline's flexibility that participants appreciated. Two wanted assistance with checking for redundancies in the blog post, and the other two wanted feedback on the blog post writing. Regarding the affordances for modifying text, 8/20 lab and 5/19 deployment participants noted a difficulty. Four lab participants raised confusion about how or when to use certain affordances. As an example, P6-lab was unsure of "*when to use which one of the six [modification] options.*" Another recurring issue noted by five lab participants was a modification affordance not working as desired. For P20-lab, setting the desired length of the modified writing to one paragraph still resulted in three, as the back-end prompts are not guaranteed to work as planned. For P12-lab, modifications were sometimes too exaggerated, while for P16-lab, a custom modification to make writing more technical was too subdued. A

couple deployment participants also noted that modifying text sometimes introduced errors. Future work may address these points of restriction and confusion noted by the participants using PAPERS-TO-POSTS.

## 4.6 Discussion

Based on our results, we reflect on design implications for other mixed-initiative tools for detailed long-document summarization.

### *4.6.1 Post-Editing is Not Enough*

Both PAPERS-TO-POSTS and the baseline tool included initial drafts generated with powerful LLMs. Nevertheless, while participants had the option to simply post-edit the initial draft and refrain from interacting further with the LLM, only two baseline and two treatment participants across the lab and deployment studies did so. This suggests that **writers may benefit from additional support other than an initial draft when writing detailed summaries of long-documents.**

### *4.6.2 Increasing Understanding and Control of Content Selection for LLM-Generated Detailed Summaries*

We presented a novel mechanism to support understanding and controlling the content of detailed long-document summaries— an interactive LLM-generated outline of the long document. This mechanism was implemented in PAPERS-TO-POSTS. Compared to a strong baseline, PAPERS-TO-POSTS made incorporating research-

paper content in a blog post summary easier for researchers. Furthermore, under time constraints, researchers were more satisfied with their blog posts written using PAPERS-TO-POSTS, particularly with respect to their content coverage. Researchers also found it easier to iterate on their blog posts with PAPERS-TO-POSTS. Although the scaffolded and preset LLM instructions for adhering to academic blog post guidelines may have also facilitated iteration, researchers consistently noted the interactive paper outline as one of the most useful features of the tool. In alignment with increased ease of iteration, researchers experienced increased editing power (i.e., change in the blog post per minute or writing action) with PAPERS-TO-POSTS. We therefore conclude that **LLM-generated interactive outlines of long documents show promise for making detailed summaries easier to write.**

On the other hand, researchers often described the baseline as more flexible and allowing them more freedom. Thus, utilizing the two tools in combination may prove useful. A tool like PAPERS-TO-POSTS may be used when the writer is unsure of what content to include in their detailed long-document summary, and a tool like the baseline tool could be used when the writer already has a specific set of points in mind or wants to give themselves creative space to use the LLM in a variety of manners.

## 4.7 Limitations and Future Work

Most participants were PhD students in computer science, who may benefit from raising awareness of their early-career work through blog posts; future work may investigate how more senior researchers utilize a tool like PAPERS-TO-POSTS. Fu-

ture studies can also examine how researchers in other fields, who are generally less familiar with LLMs and may have different traditional paper formats, may benefit from a tool like PAPERS-TO-POSTS. Future work may additionally investigate how different aspects of the input paper such as length, recency, and author position affect the user experience and outcome.

PAPERS-TO-POSTS itself has several limitations. For one, multiple deployment study participants commented on how the automatic paper parsing makes mistakes and misses parts of the paper text. The tool also does not support incorporating figures and tables in the blog post. Moreover, how detailed the paper outline should be remains uncertain. While several participants appreciated the bullet points, a couple mentioned that the outline was overwhelming. In addition, PAPERS-TO-POSTS' outline only had one level of bullet points, but future work could explore the effects of a hierarchical outline. Lastly, future work may look into how writers' expertise may be utilized in supporting their blog post writing; a recent tool for Q-and-A research-paper summarization, for example, utilizes the researcher's background to personalize its output [116].

## 4.8 Conclusion

We introduced interactive reverse source outlines, a novel mechanism for summarizing long-documents in detailed articles. The mechanism consists of an LLM-generated reverse outline of the source document with pre-selected bullet points for an initial draft of the detailed summary, which the user reviews and adjusts to control the content covered in the summary. We implemented this mechanism

in PAPERS-TO-POSTS, an LLM-powered tool for writing research-paper blog posts. We validated that PAPERS-TO-POSTS helps users to understand and control the content represented in their blog post draft through two user studies: a within-subjects lab study (N=20 participants) and a between-subjects deployment study (N=37 blog posts, 26 participants). The studies compared how researchers write about their own papers with PAPERS-TO-POSTS versus a strong baseline (LLM blog post draft and free-form LLM prompting). In the time-constrained within-subjects study, PAPERS-TO-POSTS significantly increased participants' satisfaction with their final blog posts, especially in terms of the content covered. Furthermore, across both studies, participants found incorporating document content in their blog posts easier and demonstrated increased editing power.

## Chapter 5

# Scideator: Human-LLM Scientific Idea Generation and Novelty Evaluation Grounded in Research-Paper Facet Recombination

The previous chapter provided evidence that literature-grounded, creative suggestions from generative AI are more beneficial to scientists when they are provided with a means of understanding and steering which literature components are included versus not included in those suggestions. This chapter investigates the benefits of helping scientists to answer the follow-up question of **how** literature components should be **combined** to make novel suggestions. To this end, we present a novel human-LLM **workflow for exploring and assessing recombination** of literature components. When comparing this workflow to a strong baseline without support for component recombination, we find that scientists experience significantly more creativity support with this workflow.

The scientific ideation process often involves blending salient aspects of existing papers to create new ideas, and facet-based ideation is an established framework for idea generation. To see how LLMs might assist in this process, we contribute a novel mixed-initiative ideation tool called SCIDEATOR. Starting from a user-provided set of scientific papers, SCIDEATOR extracts key facets — purposes, mechanisms, and evaluations — from these and related papers, allowing users to explore **how** paper facets may be **recombined** in **research ideas**. SCIDEATOR also provides idea **novelty assessments** regarding if a facet-based idea should or should not be considered novel

relative to the literature. To support these tasks, SCIDEATOR introduces three LLM-powered retrieval-augmented generation (RAG) modules: Analogous Paper Facet Finder, Faceted Idea Generator, and Idea Novelty Checker. In a within-subjects user study (n=22) with computer-science researchers comparing SCIDEATOR to a strong baseline, our tool provides significantly more creativity support, particularly with respect to exploration, which participants consider the most important factor for idea generation.

## 5.1 Introduction

Research papers are major sources of inspiration for scientific ideas, as they expose scientists to relevant concepts that can be recombined into new ideas [147, 91, 23, 162]. However, generating research ideas by recombining concepts from research papers is difficult for multiple reasons. For one, scientists must wade through an ever-expanding literature to find relevant concepts [15, 87]. Moreover, the phenomenon of fixation biases scientists against considering more diverse concepts and concept recombinations; instead, they are predisposed to think about a problem in familiar terms, which hinders the stimulation of novel ideas [48, 149]. Even if a scientist manages to identify interesting concept recombinations to form potential research ideas, assessing the ideas' novelty relative to existing literature is a cumbersome yet critical task.

Prior work has demonstrated the effectiveness of using a facet-based approach to find analogies between research papers for idea inspiration [23, 91, 147]. Still, the non-trivial task of recombining concepts and evaluating their associated ideas was

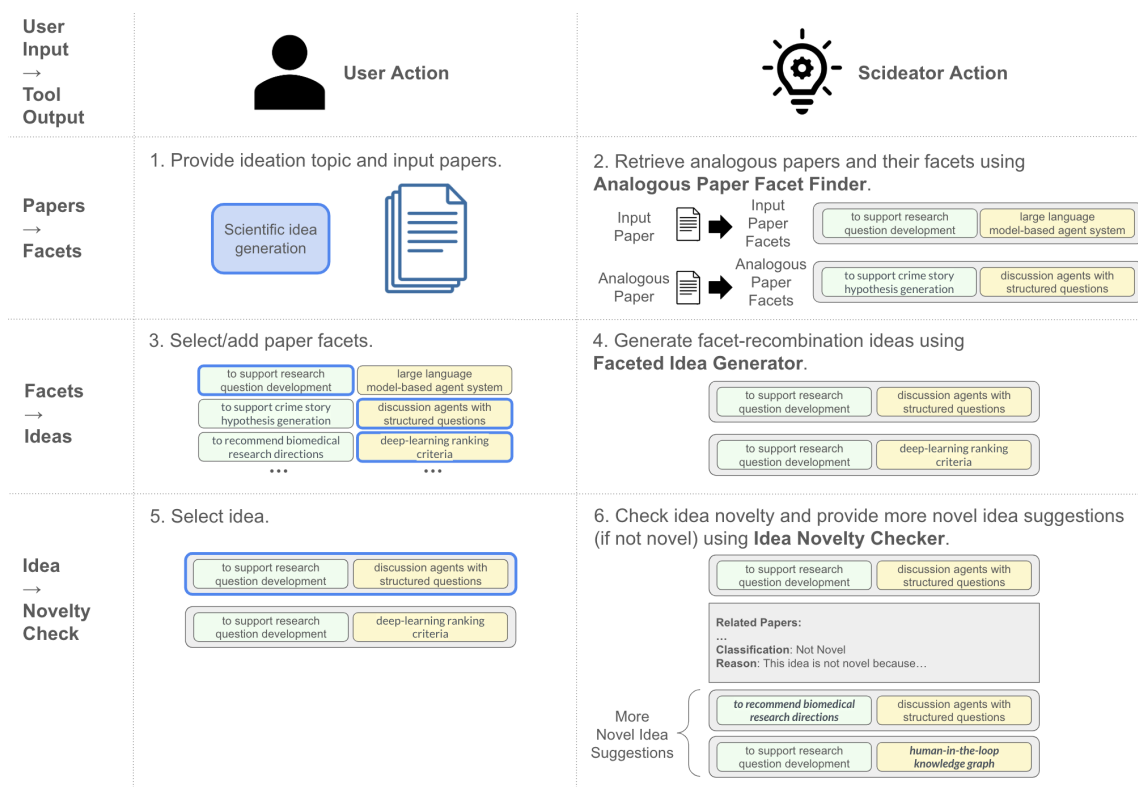


Figure 5.1: The SCIDEATOR workflow. 1) The interaction begins with the user providing an ideation topic and set of input papers as a starting point for ideation. 2) SCIDEATOR responds by retrieving analogous papers to the input papers and extracting facets (purpose, mechanism, and evaluation) from the input and analogous papers. (The evaluation facets are omitted in the figure for clarity, as it is not part of the main logic.) 3) The user then selects paper facets as well as adds their own facets for which they want to generate ideas. 4) SCIDEATOR recombines these selected facets into ideas with one purpose and one mechanism. If a purpose or mechanism facet is unspecified, the tool selects one. 5) The user selects an idea to assess for novelty. 6) SCIDEATOR classifies the idea as “novel” or “not novel” and provides a short rationale. 7) The user reviews the novelty classification and adjusts it if they disagree. 8) If the idea is deemed “not novel,” SCIDEATOR suggests more novel ideas with one of the initial idea’s facets replaced.

left to the scientists themselves, preventing scientists from exploring more recombinations. Meanwhile, recent work has demonstrated large-language-models' (LLMs') promise in assisting with scientific ideation. These works explored breadth-first versus depth-first ideation support [117], persona-based feedback on ideas [118], and idea development support [148]. LLMs make it possible to quickly synthesize and evaluate ideas. However, none of these human-LLM works support *facet-based* scientific ideation.

In this work, we present SCIDEATOR, an LLM-powered tool for facet-based scientific idea generation and novelty evaluation. SCIDEATOR employs a human-AI approach (Fig 5.1). To begin, the user provides an ideation topic and set of input papers that they would like to use as a starting point for ideation. SCIDEATOR extracts key facets (purpose, mechanism, and evaluation) from the input papers. As in prior work, a paper's purpose facet describes the problem addressed by the paper, while its mechanism facet describes the paper's proposed solution to the problem [147, 91, 23]. We also include an evaluation facet, describing each paper's method to determine if the mechanism successfully addressed the purpose. SCIDEATOR then retrieves papers with analogous purpose-mechanism pairs similar to those from the input papers. The scientist works with SCIDEATOR to select candidate facets from retrieved and input papers for recombination. The tool generates analogies involving the candidate facets and produces ideas based on the most promising ones. Based on relevant retrieved literature and carefully constructed in-context examples, SCIDEATOR then provides a novelty classification and reason for proposed ideas, making it easy for scientists to verify the tool's assessment. Finally, the tool provides facet-

based suggestions for how to improve ideas deemed not novel.

We investigate how SCIDEATOR impacts scientific ideation through a within-subjects user study with 22 computer-science researchers comparing their idea generation and novelty evaluation when using SCIDEATOR versus a strong baseline. Participants experienced significantly more creativity support in idea generation with SCIDEATOR, particularly with respect to exploring different ideas, which they considered the most important factor for creativity support in idea generation. The results also suggested that SCIDEATOR’s novelty checker is effective in helping scientists to filter out unoriginal ideas. Participants generally changed their idea novelty assessments when SCIDEATOR classified an idea as ‘not novel,’ which participants could verify using the related papers and explanation provided by the tool.

In an additional automated evaluation of our idea novelty checker, we compare our approach to recent automatic novelty evaluation methods ([121, 160]). For the same collection of papers relevant to an idea, the module improves agreement with human novelty ratings by more than 10x compared to a recent approach for LLM-based novelty ratings [121], and leads to approximately 13% higher agreement than another recent approach [160]. In terms of training LLMs to assess novelty, we further show that our approach of manually annotating generated ideas with novelty labels and concise reasons leads to better outcomes than other approaches that employ only literature retrieval, peer-review data from OpenReview [142], or LLM “prompt optimizers” [95, 190]. Furthermore, ablation experiments demonstrate the importance of each component of our novelty checker.

In summary, we make the following contributions:

- A human-AI workflow for scientific ideation that...
  1. takes a set of papers as input and finds a set of analogous papers,
  2. extracts key facets from the input and analogous papers,
  3. recombines those facets into new ideas,
  4. evaluates those ideas for novelty,
  5. iterates upon those ideas to improve their novelty, and
  6. allows seamless human refinement at each step.
- SCIDEATOR, a mixed-initiative system that implements this workflow.
- A within-subjects user study (N=22) demonstrating that, compared to a strong baseline, SCIDEATOR significantly increases creativity support, especially exploration, which participants consider the most important factor for idea generation.
- An automated evaluation of our novelty checker highlighting its advantages compared to other baselines and ablations.

## 5.2 Related Work

### 5.2.1 *Divergent and Convergent Thinking*

In ideation, there are two main stages of thinking: divergent and convergent [157, 39]. While engaging in divergent thinking, the ideator is not worried about generating the most high-quality ideas. Instead, they aim to produce as many ideas as possible

in an effort to leave no stone unturned in considering potential ideas. At this stage of the ideation process, avoiding fixation on familiar concepts is important [48, 149]. Otherwise, the ideator may miss strong candidate ideas simply because they utilize more distant concepts. In contrast, while engaging in convergent thinking, the ideator concentrates on narrowing down their ideas and determining which ideas to pursue. SCIDEATOR provides support for both divergent and convergent ideation. For divergent ideation, the tool helps users to gather inspiration from many sources and come up with several potential research ideas. For convergent ideation, the tool supports evaluating ideas for novelty relative to the literature.

### 5.2.2 *Human-AI, Scientific Ideation*

Several prior works have looked into automating scientific ideation [180, 107, 9], but automatic methods are currently insufficient for formulating novel, impactful research ideas [180, 78, 83]. In response, many works have studied the benefits of human-AI collaboration in scientific ideation [187, 68]. With the rapid advancement of LLMs, recent work has explored LLM-powered scientific ideation tools [60]. For example, the tool CoQuest supports *divergent* generation of many ideas. It allows users to direct an LLM with plain-text feedback, and the LLM generates potential ideas in a breadth-first or depth-first manner [117]. Other works have investigated supporting *convergent* human-AI scientific ideation, focusing on the development and expansion of a single input research idea rather than the creation of several diverse candidate ideas [148, 118]. Unlike SCIDEATOR, none of these works utilize facet recombination or facet analogies to create ideas, which prior work has demonstrated may benefit

the ideation process in general (see Section 5.2.3) as well as the scientific ideation process in particular (see Section 5.2.4). Also unlike SCIDEATOR, these works do not directly support evaluating ideas for novelty, though they do provide some features that could be used for idea novelty evaluation (e.g., related papers [117, 118], an interactive related literature review [148], general critiques and revisions of the idea [118]). To assist with novelty evaluation, SCIDEATOR provides scientists not only with relevant papers to the idea but also with a binary classification of its novelty relative to these papers as well as a reason for the classification.

## Scientific Idea Novelty Evaluation

Dean et al. determined novelty, relevance, feasibility, and specificity as the most prominent metrics to constitute a “good” idea [42]. For supporting idea evaluation, we focus on the aspect of novelty that they referred to as originality, which is defined as “the degree to which the idea is not only rare but is also ingenious, imaginative, or surprising.” There has been an increase in work on automatic evaluation of research idea novelty [121, 119, 180]. In the area of human-LLM interaction, Nigam et al. introduced Acceleron, a mixed-initiative, LLM-powered tool that uses an agent-based architecture with distinct personas to assess and improve upon the novelty of a research proposal relative to similar papers [137, 138]. Related work has also emerged regarding automatic paper reviews, which often involve assessing the paper’s novelty [41, 115]. In terms of human-LLM paper reviews, Sun et al. presented an LLM-powered tool to support novice peer reviewers, which included in-situ knowledge support for novelty evaluation [167].

However, all of these works related to scientific idea novelty evaluation do not explore support for facet-based ideation, while SCIDEATOR does. Furthermore, these works focus on convergent thinking and do not include any component for divergent idea generation, while SCIDEATOR supports a seamless back-and-forth between the idea generation and idea novelty evaluation processes. In addition, Acceleron, which to our knowledge is the main prior work exploring human-LLM novelty evaluation of potential research ideas, had a modest evaluation— a qualitative analysis of three scientists’ interactions with the system, whereas SCIDEATOR is evaluated through both automatic analyses and a mixed-methods user study with 22 participants.

### 5.2.3 *Human-AI, Facet-Based Ideation*

Concept combination and analogy are key methods for creating ideas [94, 171, 76]. Often, concept combination refers to the fusing of two concepts into a new emergent concept. We use the phrase “concept combination” or “facet combination” more broadly to refer to the use of multiple concepts in creating a new idea.

Related work has investigated how concept combination may be used in LLM-powered tools for ideation in non-scientific domains. BIOSPARK presents a human-AI tool to support facet-based ideation for engineering designs inspired by biological analogies. The tool helps users to identify inspiration for solving one of a few specific, preset engineering problems by drawing on a dataset of biological mechanisms. Meanwhile, AnalogiLead allows users to combine facets from preset design problems and analogous situations in order to produce new ideas [164]. We build upon these works by developing and evaluating a human-LLM tool that supports facet-based

ideation for any user-provided ideation topic<sup>1</sup> and small set of input papers, based on which the system identifies analogous papers and extracts facets to recombine.

Some prior works focused on human-AI facet-based ideation have, like SCIDEATOR, addressed more flexible ideation topics [25, 109]. For instance, CreativeConnect allows users to recombine keywords to generate a graphic sketch [30], and Luminate helps users to recombine values of various dimensions to generate diverse LLM responses [166]. SCIDEATOR is still different from these tools in two manners. First, SCIDEATOR recombines facets through the use of analogy, which is an established framework for supporting facet-based ideation (see 5.2.4). Second, SCIDEATOR provides affordances for idea novelty evaluation. In the scientific domain, this is particularly important.

#### 5.2.4 *Human-AI, Facet-Based, Scientific Ideation*

Of particular note to scientific ideation is a line of work that describes ideas in terms of two facets: the purpose (i.e., the problem) and the mechanism (i.e., the proposed solution to the problem). Hope *et al.* found that this faceted idea framework helps identify useful analogies for ideation [77]. If two ideas have similar purposes, then the mechanism of one idea may apply well to the purpose of the other idea. Similarly, if two ideas have similar mechanisms, then the purpose of one may combine well with the mechanism of the other. Subsequently, the framework has been shown to facilitate the creation of useful analogies between product ideas [77, 79], biological

---

1. The system is currently set up for computer science ideation topics but can be easily adjusted for other research areas.

and design ideas [90], research papers [91, 23], and research-paper authors [147]. A recent work has tried to apply LLMs to this faceted framework [63], but our work is the first to explore a *human-LLM* tool to support this framework. With the power of LLMs, SCIDEATOR differentiates itself from prior work in human-AI, facet-based, scientific ideation. Not only does it present analogous papers with facets to recombine, as done in prior work, but it also supports users in selecting research-paper facets to recombine and turning those facet recombinations into research ideas. Furthermore, it supports idea novelty evaluation, an important part of the scientific ideation process.

## 5.3 System

### 5.3.1 Design Goals

We developed SCIDEATOR with two design goals in mind.

- **DG1: Help scientists to generate research ideas using a facet-based framework.** Prior work has demonstrated the benefits of facet-based techniques for ideation. These works provide evidence that facet-based ideation helps people to discover facets to recombine and generate more ideas [30], avoid fixation and engage in design space thinking [166], and utilize analogies to produce ideas [90, 164]. In the scientific domain, an established facet-based framework represents research papers in terms of purposes (i.e., problems) and mechanisms (i.e., proposed solutions to problems). This framework has been shown to facilitate creative ideation [91] and identify useful analogies for

ideation [23].

To address DG1, our system retrieves facets relevant and analogous to the user’s input paper facets, which can be recombined to form research ideas, with as much or as little input from the user as they prefer. To enhance the expressiveness of the faceted idea framework consisting of purpose and mechanism facets, we introduce the facet of evaluation, or the method to determine whether the proposed solution solves the problem. Our system also allows users to add their own facets.

- **DG2: Help scientists to evaluate ideas for novelty.** Prior work has established novelty as a key component for a good idea [42]. Scientists want to work on novel research ideas in order to make meaningful contributions to the scientific community, which is why several works on automatic novelty evaluation for scientific ideas already exist [121, 119, 180]. Because the main decision-maker for which scientific idea to pursue is still the scientist, it is important to study human-LLM tools for idea novelty evaluation. Furthermore, LLM-generated research ideas may plagiarize existing work [69], making humans critical for thorough idea novelty assessment. An important challenge in finding a novel research idea is determining if an idea is novel compared to a vast pool of existing literature and improving upon its novelty if need be [180, 137].

To address DG2, our tool provides idea novelty assessments that classify an idea as ‘novel’ or ‘not novel’ relative to retrieved related work and present a rationale to explain the classification. If the idea is classified as ‘not novel,’ the tool provides suggestions for more novel ideas, with one facet replaced.

### 5.3.2 *Paper Facets*

To generate ideas, SCIDEATOR utilizes three facets from papers: the purpose, mechanism, and evaluation. The purpose facet describes the problem being addressed by the paper, the mechanism facet describes the paper’s proposed solution to the problem, and the evaluation facet describes the paper’s method to determine if the proposed solution actually solves the problem.

### 5.3.3 *Workflow and Implementation*

In this section, we provide an overview of the SCIDEATOR workflow and its implementation. The frontend of SCIDEATOR was developed using React and TypeScript, and the backend with Python. Unless noted otherwise, we used the LLM `gpt-4o-2024-08-06` and a temperature of zero. In addition, whenever we mention using a paper, we only use its title and abstract. The prompts to the LLM for each module may be found in the appendix.

Our workflow has three steps: **Papers**  $\rightarrow$  **Facets**, **Facets**  $\rightarrow$  **Ideas**, and **Idea**  $\rightarrow$  **Novelty Assessment** (Figure 5.1). Each step is driven by the following modules respectively: Analogous Paper Facet Finder, Faceted Idea Generator, and Idea Novelty Checker.

Select facets to generate more ideas. Optionally, provide custom instructions for generating ideas.

Purpose	Mechanism	Evaluation
<input type="checkbox"/> to support text-to-image exploration (input) ? 📄	<input type="checkbox"/> spreadsheet interface with prompt assistance (input) ? 📄	<input type="checkbox"/> lab and extended user studies (input) ? 📄
<input type="checkbox"/> to facilitate art exploration (input) ? 📄	<input type="checkbox"/> object detection application (input) ? 📄	<input type="checkbox"/> design and evaluation study (input) ? 📄
<input type="checkbox"/> to enhance 3d scene mood design (input) ? 📄	<input type="checkbox"/> ai-driven creativity support tool (input) ? 📄	<input type="checkbox"/> user study with designers (input) ? 📄
<input type="checkbox"/> to enhance ar game scene creation ? 📄	<input type="checkbox"/> designer-friendly game scene tem- ? 📄	<input type="checkbox"/> performance and usability experi- ? 📄
Enter your own purpose	Enter your own mechanism	Enter your own evaluation
Add	Add	Add

Enter query (up to 5 words) for relevant facets and click 'Generate More Facets' below.

**Generate More Facets**

Optionally enter custom instructions for generating ideas. Number and length of ideas can't be altered.

0/25000 characters

**Generate More Ideas**

Explore your ideas. Click an idea's filter icon to see its associated facets above.

<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <b>Facets</b>	Implement a system where deep convolutional networks analyze real-time user emotions to dynamically adjust 3D scene moods. The system will adapt lighting, color, and audio elements to enhance emotional engagement in various artistic contexts. Evaluations will involve performance metrics and user satisfaction studies.	<input type="checkbox"/> Expand
<input type="checkbox"/> to enhance 3d scene mood design (input) ? 📄	<input type="checkbox"/> deep convolutional networks (far) ? 📄	<input type="checkbox"/> performance comparison on benchmark datasets (far) ? 📄
<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <b>Facets</b>	Create an AI-driven tool that enhances translation by suggesting stylistic and mood enhancements to translators. The tool will use sentiment analysis and mood detection to offer real-time suggestions, allowing translators to refine translations with artistic flair. User studies with translators will evaluate the tool's impact on translation quality.	<input type="checkbox"/> Expand
<input type="checkbox"/> to enhance translation accuracy (far) ? 📄	<input type="checkbox"/> ai-driven creativity support tool (input) ? 📄	<input type="checkbox"/> user study with designers (input) ? 📄
Develop an adaptive UI/UX framework for 3D mood design tools that uses sensors and interaction metrics to gather real-time user feedback, dynamically adjusting mood elements to enhance art scene creation.		

Figure 5.2: SCIDEATOR's cold start. Above, the user selects or adds facets to generate ideas. They can also generate more facets to consider, and add custom instructions for the idea generation. Below, the user peruses their ideas and evaluates an idea for novelty by clicking the search icon to its left. The ideation topic here is human-AI collaboration in art.

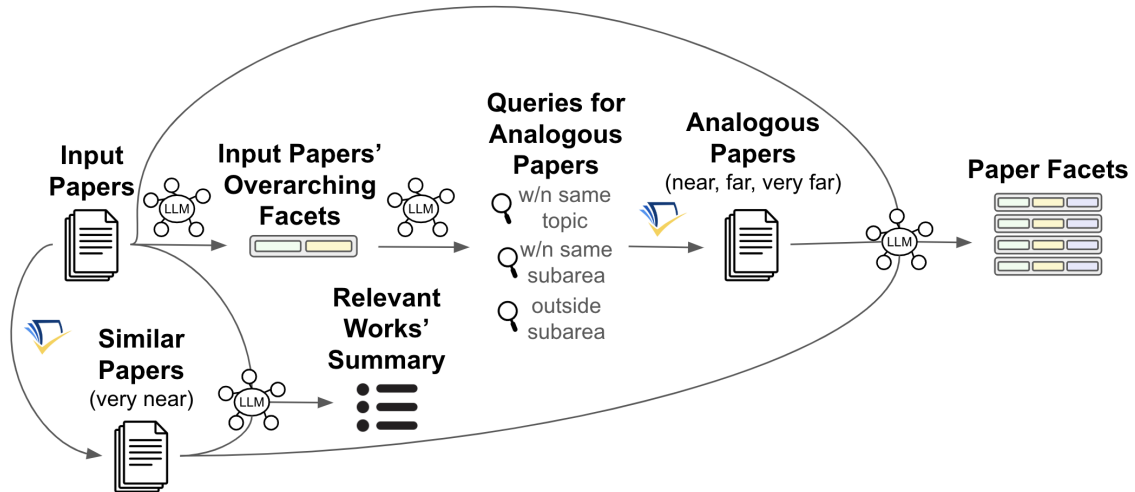


Figure 5.3: The Analogous Paper Facet Finder module. For a set of input papers, SCIDEATOR uses Semantic Scholar’s API to retrieve similar papers (very near). It uses the input and very-near papers to create a summary of relevant works. Next, the tool extracts key facets from the input papers and determines the input papers’ overarching purpose and mechanism, which it uses to come up with three queries for papers with an analogous purpose and mechanism. The queries are for analogous papers with varying distances from the input paper: same topic (near), same subarea (far), and different subarea (very far). Those queries are fed to the Semantic Scholar API to retrieve analogous papers. Finally, the facets of all the analogous papers are extracted by the LLM.

## Workflow Step 1: Papers $\rightarrow$ Facets

A user enters a broad ideation topic<sup>2</sup> and a set of one or more papers that they would like to use as a starting point. SCIDEATOR uses the input to find relevant paper facets to recombine.

SCIDEATOR’s **Analogous Paper Facet Finder**<sup>3</sup> (Figure 5.3) first retrieves information for the input papers as well as four similar papers from the Semantic Scholar API [98]. To ensure a balance among these papers, we request the two top-ranked papers (with a retrievable title and abstract) from the “all-cs” corpus and another two from the “recent” corpus. Thus, SCIDEATOR retrieves four papers **very near** to the input paper. SCIDEATOR then uses the input and very-near papers to create a summary of relevant works. This summarization step is inspired by CoQuest’s “search and summarize” action for coming up with research questions [117].

Next, SCIDEATOR’s LLM is prompted to extract the overarching purpose and mechanism of the input papers. The LLM subsequently generates twelve purpose-mechanism pairs that are analogous to the overarching purpose and mechanism of the input papers. More specifically, it generates four analogous purpose-mechanism pairs from the *same topic* within computer science research (**near**), four from the *same subarea* (**far**), and four from a *different subarea* (**very far**). For each analogous purpose-mechanism pair, the LLM also generates a query for finding relevant papers.

---

2. The system is currently set up for computer science ideation topics but can be easily adjusted for other research areas.

3. The LLM prompts for the Analogous Facet Finder module are in Appendix B.1.

SCIDEATOR uses the Semantic Scholar API to retrieve the top four papers (with retrievable titles and abstracts) relevant to each query from the “all-cs” corpus. If no paper is found, SCIDEATOR iteratively shorten the query until a paper is returned. The first retrieved paper becomes the representative paper for the analogous purpose-mechanism pair, while the other three papers act only as additional context for the representative paper. Thus, SCIDEATOR prepares four near, four far, and four very-far analogous papers.

SCIDEATOR generates the key facets (purpose, mechanism, and evaluation) of the input papers and the 16 papers of varying distance from the input papers, based on their titles and abstracts. The LLM is prompted to write the facets as short phrases (no more than 7 words).<sup>4</sup>

As shown in the top half of Figure 5.2, the tool presents all of the extracted paper facets to the user. They can track facet provenance to see if the facet originated from their own input (input paper, manually-added facet, or facet generated based on a user-provided query) or from a paper of increasing distance from the input papers (very near, near, far, or very far). If the user is unsure what a particular facet is, they can hover over the facet’s associated question mark to see its description. Also, by clicking the facet’s associated paper icon, the user can open the Semantic Scholar link to the paper from which the facet came.

---

4. To confirm reasonable consistency in facet distance, the first two authors annotated previously unseen purposes/mechanisms generated by the tool for three papers not used in the user study. It was an earlier but similar version of the tool compared to what was used in the study. Facets were grouped into generally near and far categories. Both annotators classified the majority of near purposes, near mechanisms, far purposes, and far mechanisms like the tool. One annotator demonstrated substantial agreement (Cohen’s Kappa = 0.61) and the other demonstrated moderate agreement (Cohen’s Kappa = 0.53) with the tool.

Situation	Selected Facet(s)	Input	Four Output Ideas <purpose, mechanism, evaluation> * = evaluation selected by user or from the purpose's or mechanism's papers.
Initial	None	<ul style="list-style-type: none"> <li>• Relevant papers' LLM summary</li> <li>• Input papers with facets</li> <li>• Analogous papers with facets</li> </ul>	<ul style="list-style-type: none"> <li>• &lt;input/very-near, near, either&gt;</li> <li>• &lt;near, input/very-near, either&gt;</li> <li>• &lt;input/very-near, far/very-far, either&gt;</li> <li>• &lt;far/very-far, input/very-near, either&gt;</li> </ul>
No-P-no-M	No purpose No mechanism Optional evaluation	<ul style="list-style-type: none"> <li>• [Same inputs as the Initial method]</li> <li>• Selected evaluation facet(s)</li> </ul>	<ul style="list-style-type: none"> <li>• &lt;input/very-near, near, *&gt;</li> <li>• &lt;near, input/very-near, *&gt;</li> <li>• &lt;input/very-near, far/very-far, *&gt;</li> <li>• &lt;far/very-far, input/very-near, *&gt;</li> </ul>
P-or-M	Only purpose(s) or mechanism(s) Optional evaluation	<ul style="list-style-type: none"> <li>• [Same inputs as the Initial method]</li> <li>• Selected purpose or mechanism facet(s)</li> <li>• Selected evaluation facet(s)</li> </ul>	<ul style="list-style-type: none"> <li>• 2x: &lt;selected, near, *&gt; or &lt;near, selected, *&gt;</li> <li>• 2x: &lt;selected, far/very-far, *&gt; or &lt;far/very-far, selected, *&gt;</li> </ul>
P-and-M	Both purpose(s) and mechanism(s) Optional evaluation	<ul style="list-style-type: none"> <li>• [Same inputs as the Initial method]</li> <li>• Selected purpose facet(s)</li> <li>• Selected mechanism facet(s)</li> <li>• Selected evaluation facet(s)</li> </ul>	<ul style="list-style-type: none"> <li>• 4x: &lt;selected, selected, *&gt;</li> </ul>

Table 5.1: The Faceted Idea Generator module generates analogies between two purpose-mechanism pairs from different papers. It then uses these analogies to inspire novel ideas combining the purpose from one paper with the mechanism from another. The module aims to combine papers of varying distance from the input papers, but which papers are involved in the analogies depends on whether the user has selected any purposes or mechanisms.

Users can also type in their own facets directly. Furthermore, they can use the “Generate More Facets” button to synthesize new ones, optionally specifying a query as direction. If they provide a query, SCIDEATOR retrieves four papers relevant to the query: the top two from the “all-cs” corpus and top two from the “recent” corpus. Otherwise, the system retrieves four papers from each distance category with the Analogous Paper Facet Finder. Then, SCIDEATOR generates the key facets from each of the retrieved papers and adds them to the interface.

## Workflow Step 2: Facets $\rightarrow$ Ideas

Once the user has gotten a sense of the facets available, they either select facets or allow the system to choose in order to generate ideas (Figure 5.2, top half). Based on the facet selection, the tool generates four ideas.

Table 5.1 summarizes how SCIDEATOR’s **Faceted Idea Generator** module recombines facets for idea generation.<sup>5</sup> <sup>6</sup> The module is designed to handle idea generation in four situations: when the user 1) first loads the system (**Initial**), 2) does not select a purpose or mechanism (**No-P-no-M**), 3) selects at least one purpose but no mechanism or vice versa (**P-or-M**), and 4) selects at least one purpose and mechanism (**P-and-M**).

The module aims to combine papers of varying distance from the input papers. No matter the situation, the LLM is prompted twice to generate analogies and corresponding ideas between two groups of papers, which are selected based on which of the four situations arise. For each of the two prompts, the LLM generates six candidate analogies between a group-1 paper’s purpose and mechanism and a group-2 paper’s purpose and mechanism. The LLM selects the two analogies that best fit carefully described criteria for a good idea (i.e., understandability, relevance, feasibility, specificity, and novelty). The LLM then converts one analogy into an idea combining the group-1 paper’s purpose and the group-2 paper’s mechanism. It converts the other analogy into an idea combining the group-1 paper’s mechanism and

---

5. The prompts to the LLM for the Faceted Idea Generator module may be found in Appendix B.2.

6. In this module, the LLM’s temperature is set to 0.75 to make the responses more varied.

the group-2 paper’s purpose. The LLM is instructed to come up with ideas related to the ideation topic but differentiate its ideas from the existing work described in the relevant works’ summary and the papers from which it ideates. It is also instructed to take a step to improve upon its ideas in terms of the criteria for a good idea.

To start, the system creates four ideas using our **Initial** idea generation method. In this situation, for generating the first two analogies and corresponding ideas, the group-1 papers are the input/very-near papers, and the group-2 papers are the *near* papers. For generating the second two analogies and corresponding ideas, the group-1 papers are again the input/very-near papers, and the group-2 papers are the *far/very-far* papers. Each generated idea’s evaluation facet is selected by the LLM; it is the evaluation facet either from the purpose’s associated paper or the mechanism’s associated paper. Afterwards, SCIDEATOR produces four ideas each time based on the user’s choice of facets and its corresponding situation, as outlined in Table 5.1.<sup>7</sup> Table 5.2 presents a sample of user study participants’ favorite ideas, which were generated using the different methods of this module.<sup>8</sup>

As the user explores their facet-based ideas, they can see a more detailed version of each by clicking the ‘Expand’ button (Figure 5.2, bottom half). Not pictured, the user can also add their own idea, and the LLM is prompted to extract its facets, which are added to the list of available facets.

---

7. Note that, if one of the selected purposes or mechanisms is manually entered and therefore does not have an associated paper, the system generates a purpose or mechanism to pair with the selected facet for the sake of creating analogies with another paper.

8. The expanded versions of the sample ideas are in Appendix B.5.

Situation	Topic	Purpose	Mechanism	Evaluation	Custom Instructions	Short Version of Idea
Initial, Treatment	LLM explainability	to enhance medical information retrieval (near)	sequential knowledge-guided prompting (paper input)	medical question-answering benchmarks (near, unselected)	None	Create a medical information retrieval system using sequential knowledge-guided prompting to simulate diagnostic pathways. This system will integrate electronic health records (EHRs) with clinical guidelines, leveraging prompts to identify and follow potential diagnostic routes. Using medical question-answering benchmarks, the system will evaluate the accuracy and relevance of retrieved information in real-time.
P-or-M, Treatment	human-AI collaboration in art	to enhance designer-ai collaboration (very near, selected)	residual learning framework (far)	empirical performance analysis (far, unselected)	I want to know if designer want to work with ai	Develop a residual learning framework for AI design tools where designer feedback on specific elements like color schemes and layouts is treated as residuals. The AI iteratively refines these elements, enhancing the final design. The feedback loop is designed to be intuitive, allowing designers to easily indicate areas for refinement, which the AI then uses to adjust its design suggestions dynamically.
P-and-M, Treatment	LLM explainability	to explain text generation outputs (paper input)	explanation logic verifier (own input)	quantitative and qualitative analysis (paper input, selected)	Use different verification approaches from NLI, math, program to ensure the explanation follows a logical consistency	Develop a platform where a mathematical logic verifier validates LLM-generated mathematical explanations and employs graph-based visual proof representations. These visual tools will illustrate each logical step, aiding comprehension and engagement. The verifier will ensure explanations are logically consistent and accurate, using the visual aids to guide users through complex reasoning processes effectively.
P-and-M, Treatment	AI tools for education	to personalize vocabulary learning (very near)	ai-generated virtual speakers (very near)	participant feedback analysis (very near, unselected)	make the idea more focused and specific	Develop a personalized vocabulary learning platform using AI-generated virtual speakers that tailor dialogue scenarios to user-uploaded images. The AI analyzes images to identify objects and contexts, generating dialogues that incorporate relevant vocabulary for enhanced learning.
Baseline	dealing with LLM hallucinations	N/A	N/A	N/A	Can the reward model in Selk-Knoelwidge tuning be changed to incorporate more fine-grained feedback (e.g. from Halogen like metrics) [all 3 papers given to LLM]	Develop a comprehensive reward model for Self-Knowledge Tuning that integrates specific HALo-GEN metrics, such as atomic unit verification and error classification. The model will be tested across multiple domains to assess its effectiveness in reducing hallucinations in LLMs. This approach not only refines the feedback loop with detailed evaluations but also establishes a robust testing framework to validate improvements in factual accuracy.
Baseline	AI tools for education	N/A	N/A	N/A	Generate ideas that use non-traditional mediums for education, so avoid using textbooks or other material commonly found in classrooms. [all 3 papers given to LLM]	Develop an AR-based mobile app powered by AI that personalizes ecological education for children by adapting content to local biodiversity. The app will guide users in exploring their surroundings to identify species and offer contextual ecological insights. AI will craft personalized learning journeys, adapting to the user's pace and knowledge level. It will include virtual simulations of diverse ecosystems for users in less biodiverse areas, ensuring inclusivity and broader educational reach.

Table 5.2: A sample of participants’ favorite ideas from the user study. The treatment ideas are generated with the different methods of the **Faceted Idea Generator** module, depending on whether the participant selected any purposes or mechanisms.

**a) IDEA**  
 This research involves developing an AI-driven tool to enhance digital art interaction by incorporating machine learning algorithms that detect user mood through analysis of visual and interaction patterns in a digital environment. The tool personalizes interface themes by adjusting color schemes, layout, and interactive elements to align with detected moods, thereby improving the digital user experience. The tool will use computer vision and interaction tracking to gather data on user engagement, processing this information to infer mood states. The tool will then dynamically adjust the interface to reflect these moods, creating a customized user experience that evolves with interaction patterns. A user study with designers will evaluate the tool's effectiveness in enhancing engagement and user satisfaction, providing insights into its potential impact on digital art platforms. This approach represents a novel intersection of AI, art, and user experience design, offering a dynamic, responsive environment that adapts to individual user interactions.

**b) FACETS**

**c) RELATED PAPERS**  
 Paper 0:  
**An AI-Powered Interactive Assistant: Integrating Multimodal Interaction for Enhanced User Experience**  
 Y.N.B. D.S.A.S. H.H.R. C.J.A. • 2024 International Conference on Emerging Research in Computational Science (ICERCS)  
 This work presents a fully integrated multimodal AI-powered interactive assistant based on natural... [Expand](#)  
 Paper 9:  
**Web Design Encompasses UI,UX and Responsive Development**  
 K. Bhoomika D. R. Narayan • INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT  
 Abstract—The goal of abstract design is to create visually ar- resting and thought-provoking user... [Expand](#)

**d) NOVELTY EVALUATION**  
 Classification of Idea Compared to Related Papers:

**e) Reason:**  
 The idea is not novel because it closely replicates existing approaches that personalize interfaces based on mood detection. For example, InsightNav [3] already adjusts desktop themes dynamically using real-time facial expression analysis, and LumiMood [9] tailors 3D scene moods to influence user experience. The proposed tool's focus on using computer vision and interaction tracking for mood-based personalization in digital art is largely similar, offering minimal new contributions beyond these established systems.

**f) NEW IDEA SUGGESTIONS**  
 Get suggestions for updating your idea to be novel compared to all the related papers.  
 Develop an AI-driven tool that analyzes user mood through visual and interaction patterns to compare art engagement across various contexts. This tool will utilize machine learning algorithms to assess how different digital environments impact user interaction and emotional engagement with art. By adjusting interface themes and measuring mood variations, the tool can provide insights on how art contexts influence viewer perception and interaction. Comparative experiments will be conducted in both digital galleries and virtual reality settings to evaluate the differences in engagement. This study will help understand the role of context in digital art interaction, offering a novel perspective on tailoring digital environments to optimize emotional and immersive experiences.  
 Removed Facet:  Added Facet:   
[Add to Idea List](#)  
 Introduce a user-participatory AI design model to create an interactive digital art tool that personalizes user interfaces based on mood detection. This model will involve users in the design process, allowing them to provide feedback on mood-related adjustments to interface themes, such as color schemes and layout changes. Through iterative user input, the AI system will learn to adapt its design outputs according to user preferences and moods more effectively. Performance and usability experiments will assess the tool's adaptability and its impact on user satisfaction. This participatory approach ensures that the AI-driven personalization aligns closely with diverse user expectations and needs, resulting in more meaningful and engaging digital art interactions.  
 Removed Facet:  Added Facet:   
[Add to Idea List](#)  
 Develop a tool that dynamically adjusts digital art interfaces using machine learning algorithms to detect user mood. The tool will be evaluated through comparative experiments conducted in museums and labs to assess its impact on enhancing art engagement in controlled and public settings. Computer vision will analyze visual patterns and user interactions, personalizing themes according to mood detection. The experiments will compare user engagement levels, emotional responses, and satisfaction between traditional museum settings and interactive lab environments. This dual-context evaluation will offer insights into how digital tools can be tailored to improve art engagement in diverse settings, bridging the gap between traditional and digital art experiences.  
 Removed Facet:  Added Facet:   
[Add to Idea List](#)

Figure 5.4: SCIDEATOR’s novelty assessment modal for one idea, which presents the idea (a) as well as its facets (b), related papers (c), adjustable novelty classification (d), and adjustable classification reason (e). When the idea is classified as “not novel,” the system provides a set of three suggestions for more novel ideas (f), each of which replace one of the idea’s original facets. The ideation topic here is human-AI collaboration in art.

### Workflow Step 3: Ideas $\rightarrow$ Novelty Assessments

After collecting a few candidate ideas, the user evaluates ideas for novelty and finds more novel ones. To evaluate an idea, the user opens the novelty checker modal (Figure 5.4) by clicking the search button to the left of an idea. The user reviews the tool’s retrieved papers related to the idea, novelty classification of the idea (“novel” or “not novel”), and short classification reason referencing the related papers. The user can manually overwrite SCIDEATOR’s classification and reason. When the idea is marked as “not novel”, SCIDEATOR presents the user with three new idea suggestions, each replacing a different facet in the original idea.

The **Idea Novelty Checker** module (Figure 5.5) powers these interactions through four steps: 1) retrieve candidate relevant papers, 2) select most relevant papers, 3) evaluate idea novelty, and 4) suggest more novel ideas.<sup>9</sup>

**STEP 1: RETRIEVE CANDIDATE RELEVANT PAPERS.** To assess an idea’s novelty, we compare an idea against a comprehensive collection of relevant papers. This collection includes many papers, including all of the papers that the previous modules retrieved. SCIDEATOR also finds more related papers to these papers using the Semantic Scholar API [98]. However, simple retrieval methods often overlook contextual aspects of ideas such as their purpose, mechanism and evaluation facets [136, 135, 178]. To improve the paper collection’s coverage, SCIDEATOR generates search queries based on various keywords directly related to the idea to retrieve relevant papers, a query-based retrieval method used in [122, 161]. The search queries themselves are LLM-generated keywords and potential titles based on the

---

9. The prompts to the LLM for the module are listed in Appendix B.3.

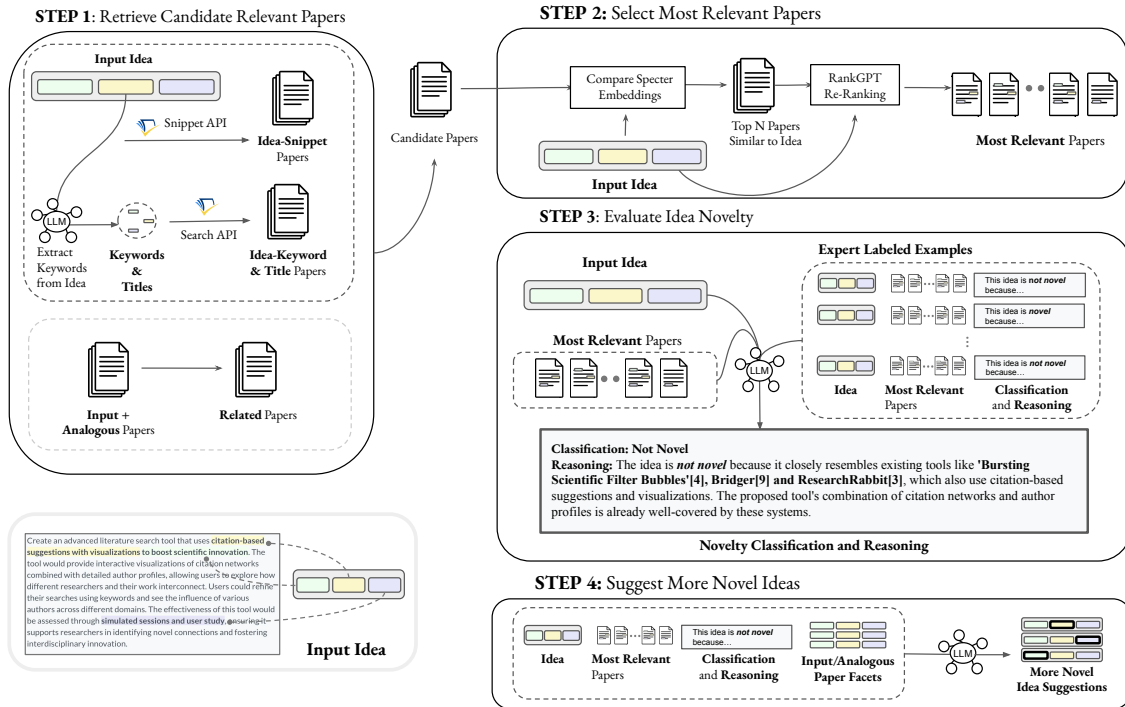


Figure 5.5: The Idea Novelty Checker module follows a retrieve-then-re-rank approach for novelty evaluation. In Step 1, it gathers a comprehensive set of papers relevant to an idea. This includes papers originally used to generate the idea, related papers, and additional papers retrieved through keyword and title searches extracted directly from the idea, as well as snippet searches using the entire idea as input. In Step 2, a two-stage re-ranking process is applied, where an embedding-based ranking strategy filters the large collection to top- $N$  papers, followed by a facet-based LLM re-ranker to identify the top- $k$  most relevant papers. In Step 3, these top- $k$  papers are used to assess the idea’s novelty, guided by in-context examples that evaluate novelty with grounded reasoning. In Step 4, if an idea is classified as “not novel” by the tool or user, the LLM generates three idea suggestions, each replacing a different facet in the original idea in order to make the idea more novel compared to the relevant papers.

idea description.

The keyword-based search results sometimes introduce irrelevant results. For example, consider the idea in Figure 5.5 about citation-based suggestions for scientific innovation; a keyword like ‘scientific innovation’ is extracted from the idea. Searching with this keyword would indeed retrieve relevant papers but might also include unrelated articles merely because the phrase appears in their title and abstract. To counter this limitation, SCIDEATOR retrieves relevant text snippets directly from the full text of papers on Semantic Scholar using its snippet-text API<sup>10</sup>. The system forms the query using the entire proposed idea, finding snippets from papers with potentially high relevance.

The papers utilized in the prior modules, additional retrieved papers related to them, and the additional papers obtained through both keyword-based and snippet-based searches collectively form the comprehensive candidate paper set for the idea novelty assessment process.

**STEP 2: SELECT MOST RELEVANT PAPERS.** To identify the papers most likely to overlap with the idea, we implement a two-stage re-ranking process that combines embedding-based filtering with an LLM-based re-ranking approach. This follows established information retrieval practices of retrieve-then-re-rank [55, 139, 1, 128, 168, 11].

The first stage is **embedding-based filtering**. SCIDEATOR computes the semantic similarity between the idea and each paper in papers from STEP 1 using Semantic Scholar SPECTER embeddings [36]. It selects the top  $N$  most similar

---

10. [api.semanticscholar.org/api-docs/#tag/Snippet-Text](https://api.semanticscholar.org/api-docs/#tag/Snippet-Text)

papers to the idea based on cosine similarity. This embedding-based ranking efficiently narrows down the paper collection but, compared to LLMs [154], fails to capture more contextual relationships between different facets of the idea and related papers.

The second stage addresses this limitation with an **LLM-based re-ranker**, RankGPT [169], which goes beyond surface similarities. We use RankGPT to compare the papers against the idea’s application domain, purpose, mechanism, and evaluation.<sup>11</sup> It ranks the papers based on the following priorities (from high to low): 1) papers with all the idea’s key facets; 2) papers with a similar application domain and purpose; 3) papers with a similar purpose, mechanism, or evaluation; and 4), papers with partially matched or related facets.

This approach ensures that the final ranking accurately reflects the and depth of each paper’s connection with the idea. Finally, SCIDEATOR selects the  $k$ -most relevant papers for the novelty assessment. The default  $N$  papers that proceed after the embedding-based ranking is 100, and the default  $k$  papers that proceed after the LLM-based re-ranking is 10. We use the LLM `gpt-4o` for this step.

**STEP 3: EVALUATE IDEA NOVELTY.** Using the top- $k$  relevant papers as comparison points, SCIDEATOR prompts an LLM to assess an idea for novelty. The LLM outputs a binary classification (‘novel’ or ‘not novel’) accompanied by its reasoning, which points to the related papers for any similarities and differences. However, determining the novelty of an idea in relation to existing scientific literature is a complex and subjective task. To guide the LLM’s judgment, our prompt incorporates

---

11. RankGPT’s relevance criteria were changed to match with our key idea facets.

in-context examples drawn from formative annotation studies (see Section 5.4.1). Each example is comprised of an idea, the top- $k$  papers related to the idea, the novelty classification label, and classification reason. These examples reflect the experts’ criteria for novelty: an idea is considered novel if it 1) differs from all retrieved papers in purpose, mechanism, or evaluation, 2) presents a unique combination of these facets, or 3) applies the same facets to a new application domain. We use the LLM o3-mini for this step, given its focus on reasoning.

**STEP 4: SUGGEST MORE NOVEL IDEAS.** Based on the classification reasoning, the LLM is instructed to come up with three suggestions for ideas that are more novel compared to the related papers. Each suggestion replaces a different facet in the initial idea with another available facet of the same category (e.g., purpose). For increased variance in responses, we use a temperature of 0.75 in this step.

## 5.4 Formative Studies for Idea Novelty Checker

In this section, we describe formative studies conducted to 1) understand the challenges of scientific idea novelty evaluation through annotations of ideas’ novelty, 2) investigate the importance of each component of our novelty checker and 3) explore how our novelty checker performs relative to baselines.

### 5.4.1 Annotation Studies

In the context of scientific literature, novelty evaluation faces two challenges. First, identifying the most relevant papers from the hundreds of millions available is inherently difficult, as the space of potentially overlapping work is vast. Second, deter-

mining criteria for judging novelty is challenging because novelty itself is subjective and can be defined in multiple ways.

To better understand and address these issues, we conducted a formative study, in which the first two authors assessed the novelty of ideas based on retrieved relevant papers. The authors evaluated 51 ideas— 46 generated by SCIDEATOR and five adapted from accepted and rejected papers from OpenReview (ICLR 22, NeurIPS 23). Fewer examples were taken from OpenReview since the primary focus was on evaluating ideas from SCIDEATOR. For every idea, the system identified the most relevant papers through a two-step process. It first gathered candidate papers using keyword-based queries and subsequently re-ranked them for relevance to the idea using an LLM-based re-ranker [169]. The authors classified each idea as novel, moderately novel, or not novel based on the top 10 papers. They achieved moderate agreement (Cohen’s Kappa = 0.64) with 17 disagreements out of 51 classifications. The agreement may not have been higher because the authors sometimes relied on broader domain knowledge rather than restricting judgments to the most relevant papers, which were often insufficient. Additionally, using three categories led to disagreements, as the distinction between novel and moderately novel is itself subjective.

Building on observations from the formative annotation study, we conducted a second study. In this study, we based our judgments solely on the provided papers, and the classification categories were simplified to just novel and not novel. Inspired by prior work [147, 91, 23] that categorizes research ideas into core facets such as purpose and mechanism, we also agreed on defining novelty as follows. An

idea is considered novel if it differs from all 10 retrieved papers in at least one core facet (purpose, mechanism, or evaluation). An idea is also considered novel if it uniquely combines these facets or applies them to a new application domain. Using this controlled framework, we annotated a set of 51 ideas— 34 new ones generated by SCIDEATOR and 17 from the previous study where external knowledge had influenced novelty judgments. By narrowing the focus to the relevant papers alone and simplifying the classification options, we observed fewer disagreements (8/51) and achieved a higher agreement rate (Cohen’s Kappa = 0.68). This study highlights that a robust novelty checker depends critically on (i) high-quality retrieval, which we aim to support with two-stage re-ranking and retrieval, and (ii) a well-defined notion of novelty, which we aim to provide in our LLM prompting.

## Resulting Dataset

From our annotation studies for novelty assessment (Section 5.4.1), we collected 67 consensus-labeled examples of novelty classification (39 labeled as novel and 28 as non-novel) as our dataset. We split the examples into training and test sets (35 for training and 32 for testing) with a balanced distribution of novel and non-novel ideas. Please refer to Table B.2 in Appendix B.7 for sample examples.

### 5.4.2 Ablations

#### Methods

To assess the contribution of each component in our novelty checker, we conducted ablation studies using 58 ideas (comprising 13 ‘not novel’ instances from our test set and 45 Natural Language Processing papers from the literature). For this experiment, we focus on ‘not novel’ cases, since the ideas labeled ‘novel’ in expert-labeled test data can vary with different retrieved paper sets. In our ablations, we considered the following variations: (i) **Complete System**: used keyword and snippet retrieval (each returning the top- $k$  documents based on Semantic Scholar’s ranking), embedding filtering, and facet-based RankGPT re-ranking; (ii) **Relevance RankGPT**: used the same retrieval methods (keyword and snippet) plus embedding filtering, but replaced the facet-based RankGPT re-ranker with one based on general relevance [169]. This variation differs from the complete system only in the LLM re-ranking component, allowing us to assess the importance of facet-based re-ranking; (iii) **Embedding Filtering**: omits the LLM re-ranker entirely, relying only on the embedding-based filtering. This setup allows us to assess the importance of the LLM re-ranking step; and (iv) **Snippet Retrieval** and **Keyword Retrieval**: each of these setups returned the top- $k$  documents from their respective retrieval method (without embedding filtering or any LLM re-ranking), leveraging the inherent ranking/scoring provided by Semantic Scholar. This setup allows to assess the importance of both re-ranking steps. This structured setup enabled us to isolate the contribution of each component (retrieval method vs. re-ranking strategy) and

Table 5.3: Accuracy of SCIDEATOR ablations in predicting “not novel.”

Method	Accuracy
Complete System	89.66%
Relevance RankGPT Instead of Facet-Based RankGPT	13.79%
Embedding Filtering w/o RankGPT	10.34%
Only Snippet Retrieval w/o Embedding Filtering or RankGPT	8.62%
Only Keyword Retrieval w/o Embedding Filtering or RankGPT	5.17%

Table 5.4: Comparing each ablation to the complete system in terms of overlap in retrieved papers and paper rankings. *Overlap* indicates how many papers overlap on average with the complete system’s top-10 papers. *Rank Shift* measures the average absolute difference in rank positions (only among overlapping papers).

Method	Overlap (↑)	Rank Shift (↓)
Relevance RankGPT	7.97	0.67
Embedding Filtering	7.93	0.84
Snippet Retrieval	2.88	1.85
Keyword Retrieval	1.17	1.39

evaluate whether they collectively brought key papers for novelty assessment into the top 10. We use `gpt-4o` for re-ranking and `o3-mini` for evaluating novelty.

## Results

Table 5.3 shows that the complete SCIDEATOR system, which employs facet-based re-ranking in RankGPT, significantly outperforms its ablated variants in accuracy. The results demonstrate that methods relying only on keyword or snippet-based retrieval have much lower accuracy, and even alternate re-ranking strategies with a single embedding-based reranker or both embedding and general relevance RankGPT are insufficient to consistently bring key papers into the most relevant paper set.

These findings show that combining facet-based reranking with embedding filtering is critical for identifying the most relevant papers.

Table 5.4 compares the top-10 most relevant papers retrieved under each ablation setting with those from the complete system. Approximately 30% of the papers differ when using either Embedding Filtering or general Relevance RankGPT. Additionally, notable rank shifts are observed from the facet-based LLM reranker in the complete system to the relevance-based LLM reranker (Relevance RankGPT). Further, without the reranking steps, both snippet and keyword retrieval exhibit minimal overlap with the final system’s top paper results, highlighting the importance of the reranker stage.

### 5.4.3 Comparison to Baselines

#### Methods

We evaluated multiple baselines to benchmark our novelty evaluation approach. First, we employed a zero-shot prompt as a straightforward baseline, and further refined this manually written prompt using Anthropic’s prompt generator<sup>12</sup>. We also applied popular prompt optimization techniques such as DSPy [95] and TextGRAD [190], which optimize the prompt instructions using a train/validation split created from formative study examples. In experiments with TextGrad [190], we investigated how specific prompt instructions influence an LLM’s ability to classify the novelty of an idea.

---

12. <https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator>

As an alternative to using in-context examples from the formative study, we extracted reviews from ICLR and NeurIPS submissions via the OpenReview API [142]. These reviews comprise aspects such as *strengths*, *presentation*, *limitations*, *soundness*, *weaknesses*, *questions*, *confidence*, *contribution*, *summary*, and *rating*. The input title and abstract were adapted to match the ideas in the training data using a style-change prompt. After rigorous filtering, we identified approximately 8,156 submissions discussing idea novelty and manually selected reviews that specifically evaluated the core idea rather than the entire paper. From these, we randomly sampled 20 idea-review pairs to serve as an additional baseline with different in-context examples. Since OpenReview reviews do not reference the associated papers, we evaluated our expert-labeled examples both with and without including relevant papers to ensure a fair comparison.

In addition to these baselines, we also compare our novelty checker ‘*prompt*’ with that of AI Scientist [121] (different from its paper reviewer) and AI Researcher [160] on the same test set of ideas and fixed top 10 papers. We compare only the prompts to assess novelty, rather than the entire systems, because the test set containing the novelty judgments by experts were based on a fixed set of the 10 most relevant papers for each idea. Since different retrieval methods could introduce new papers and potentially change novelty classification, we standardized the most relevant papers to ensure a fair comparison of the prompts alone. Additionally, since both setups require a different style of input idea, we adapted the ideas to match the requirements of each system.

We compared all of these baselines to various ablations of our expert-labeled in-

Models	Accuracy	Precision	Recall	F1	Cohen Kappa
<b>Zero Shot Setting</b>					
Zero Shot	0.68	0.76	0.64	0.65	-
+ improved prompt using Anthropic prompt generator	0.68	0.70	0.64	0.64	-
<b>Prompt Optimizers</b>					
<b>DSPy</b>					
- with idea, most relevant papers, class	0.68	0.83	0.62	0.58	-
- with idea, most relevant papers, class, reasoning	0.66	0.82	0.58	0.52	-
<b>TextGRAD</b>					
- with idea, most relevant papers, class	0.78	0.76	0.76	0.76	-
<b>In-context Setting</b>					
<b>Open-Review Examples</b>					
- with idea & review (i.e., reasoning)	0.59	0.55	0.51	0.43	-
<b>Expert Labeled Examples</b>					
- GPT-4o					
- with idea, reasoning	0.75	0.76	0.77	0.75	-
- with idea, most relevant papers, class	0.78	0.77	0.76	0.77	-
- with idea, most relevant papers, class, reasoning	<u>0.81</u>	<u>0.84</u>	0.78	0.79	0.59
- O3-MINI					
- with idea, most relevant papers, class, reasoning	0.78	0.77	<u>0.89</u>	<u>0.83</u>	-
<b>Other Novelty Checkers</b>					
<b>AI Scientist</b> [121]	0.47	0.55	0.53	0.44	0.05
<b>AI Researcher</b> [160]					
- GPT-4o	0.78	0.81	0.74	0.75	0.52
- CLAUDE-3-5-SONNET	0.56	0.63	0.61	0.56	0.19

Table 5.5: Experimental results on expert-annotated dataset comparing SCI-DEATOR’s novelty checker to various baselines. Unless otherwise noted, GPT-4o was used to obtain these results.

context examples. When we include papers in the expert-labeled examples, they consist of titles and abstracts only. For our best performing setting, we also experimented with O3-MINI.<sup>13</sup>

<sup>13</sup>. Please refer to Appendix B.9 for additional details about the experimental setup for all of these approaches.

## Results

As shown in Table 5.5, our experiments indicate that incorporating expert-annotated data as in-context examples significantly enhances novelty classification accuracy compared to zero-shot prompts, DSPY, TextGRAD, and setups using OpenReview examples. We compared two configurations for DSPY, one with reasoning and one without. Our expert-labeled prompt consistently achieved higher performance than the prompt optimizations produced by these methods, and we posit that the number of examples for train/validation were not sufficient for prompt optimizers with GPT-4O. Meanwhile, the TextGRAD prompt optimizer did not improve upon its initial system prompt. Notably, even when we excluded the relevant papers from the expert-labeled examples, our approach still outperformed the OpenReview baseline. For our best performing setting, which includes the idea, most relevant papers, classification, and reasoning, we also experimented with O3-MINI and observed a better recall than GPT-4O. O3-MINI was able to better reason about analogous similarity across papers and ideas.

Although the TextGRAD prompt optimizer did not improve upon its initial system prompt, it provided valuable insights into LLM prompt sensitivity. Figures in Appendices B.1, B.2, and B.3 present the accuracy of various prompts optimized with TextGrad on our dataset (train=25, validation = 10, test = 32). Prompts with both non-zero and zero validation accuracy included various instructions for evaluating the novelty of ideas. Through this prompt optimization process, we observed interesting ways in which LLMs may evaluate novelty, such as considering historical context, considering frequency of similar studies, conducting comparative

analysis with existing works, and examining arguments for both novel and non-novel perspectives. However, prompts without these specific instructions also influenced accuracy, suggesting the complexity of novelty evaluation with LLMs. Notably, some prompts with similar instructions showed different performance on validation data. For example, both Prompt 3 (accuracy = 0) and Prompt 9 (accuracy = 0.6) include instructions to evaluate if the idea introduces unique methodologies and how it compares to existing work. However, the difference in their performance suggests that subtle variations in wording and instruction framing can significantly impact the classification performance. It remains unclear why certain prompts perform better despite having similar instructions. This analysis highlights the LLM’s sensitivity to prompt design when assessing novelty of an idea. Even minor variations in wording and structure can lead to substantial performance changes, emphasizing the need for careful prompt engineering and well-chosen in-context examples to guide the LLM for idea novelty evaluation.

Furthermore, our approach achieved over **10 times more agreement with expert-labeled examples** compared to AI Scientist, and **approximately 13% higher agreement** than AI Researcher, further validating the effectiveness of our novelty checker. It is important to note that AI Scientist defaults to “not novel” when it fails to reach a conclusion in novelty evaluation (18 out of 32 times), which may have impacted its agreement rates. Appendix B.8 qualitatively compares novelty evaluations by AI Scientist, AI Researcher, and SCIDEATOR on two research ideas. SCIDEATOR provides concise justifications for its novelty decisions by referencing key similarities and differences with existing works. For example, in Example 1, it

correctly identifies the idea as ‘novel’ by highlighting these aspects. In contrast, AI Researcher evaluates each paper individually, classifying an idea as ‘not novel’ if any paper is considered citable; but in our examples, none of the papers were flagged as citable despite sharing similar purposes, leading to a ‘novel’ classification. Due to space constraints, we show insights only from the first paper for each example. Figure B.6 indicates that while AI Scientist’s judgments generally align with the ground truth and offer actionable suggestions, it sometimes misinterprets the idea—as in Figure B.5, where its focus shifts from the idea to the accompanying code.

## 5.5 User Study Methods

We evaluated SCIDEATOR against a strong baseline tool that supports combining ideas at the paper level instead of the facet level.

### 5.5.1 *Research Questions*

- **RQ1:** Does SCIDEATOR lead to more creativity support for idea generation than a strong baseline?
- **RQ2:** What are common ideation patterns while using SCIDEATOR as compared to a strong baseline?
- **RQ3:** Does SCIDEATOR improve confidence in idea novelty assessments more than a strong baseline?

### 5.5.2 *Participants*

We recruited 22 computer-science researchers (W: 7, M: 15) through institutional mailing lists and academic social networks. We compensated them with a \$60 Amazon gift card. Twelve participated as human-computer interaction (HCI) researchers and 10 as natural-language-processing (NLP) researchers. Most were PhD students (PhD student: 16, master’s student: 5, industry researcher: 1). Generally, the participants interacted with LLMs often (a few times per... day: 12, week: 7, month: 1, few months or longer: 2).

### 5.5.3 *Study Design*

We conducted a within-subjects study, in which each participant completed tasks for the treatment and baseline conditions in randomized order. The ideation topics for the treatment and baseline conditions were also randomized. Overall, participants had no difference in their familiarity ratings (7-point, Likert-type) for the assigned treatment topic and assigned baseline topic (M=0.00, Q1=-1.00, Q3=1.00). There were two preset topics for HCI researchers (human-AI collaboration in art, AI tools for education) and two for NLP researchers (dealing with LLM hallucinations, LLM explainability). For each topic, there were three associated input papers to use as a starting point.<sup>14</sup>

For the treatment tool, we modified SCIDEATOR to more effectively address our research questions. Our study separates the idea generation task from the idea evaluation task. To keep the study controlled, we disabled some of SCIDEATOR’s

---

14. The input papers for each topic may be found in Appendix B.4.

Select papers to generate more ideas. Optionally, provide custom instructions for generating ideas.

**Prompting for Discovery: Flexible Sense-Making for AI Art-Making with Dreamsheets**  
S. Garanganao J. Zamfirescu-Pereira K. Won +5 authors K. Zamfirescu-Pereira • Computer Science •  
International Conference on Human Factors in Computing Systems • 2023-10-15  
Design space exploration (DSE) for Text-to-Image (TTI) models entails navigating a vast, opaque... [Expand](#)

**Algorithmic Ways of Seeing: Using Object Detection to Facilitate Art Exploration**  
L. S. Meyer J. E. Aaen A. R. Tranberg +3 authors A. Løvlie • Computer Science • International Conference on Human Factors in Computing Systems • 2024-03-28  
This Research through Design paper explores how object detection may be applied to a large digital... [Expand](#)

**LumiMood: A Creativity Support Tool for Designing the Mood of a 3D Scene**  
J. Oh S. Kim S. Kim • Computer Science • International Conference on Human Factors in Computing Systems • 2024-05-11  
The aesthetic design of 3D scenes in game content enhances players' experience by inducing desired... [Expand](#)

Optionally enter custom instructions for generating ideas. Number and length of ideas can't be altered.

0/75000 characters

[Generate More Ideas](#)

Explore your ideas.

[🔖](#) Develop an AI-driven object detection tool for curators that recognizes a wide range of art styles and motifs, facilitating the discovery of thematic patterns across large digital collections. Collaborations with museums will address data privacy issues, and comprehensive training sessions will aid curators in integrating AI into their workflows. [Expand](#)

[🔖](#) Develop an AI-based educational tool for art students that uses verified datasets and expert consultation to ensure historical accuracy in text-to-image generation, allowing students to interactively explore art history concepts. The tool will feature customizable modules aligning with standard curriculums and adaptive feedback systems to cater to different learning styles and paces. [Expand](#)

[🔖](#) Develop an AI-driven interactive museum experience that uses mobile apps or wearable devices to capture visitor preferences and emotional responses. The system would adjust digital exhibits through mood settings and object detection, allowing personalized art exploration while maintaining the integrity of artworks. It would ensure visitor consent and provide options to opt-out. [Expand](#)

[🔖](#) Develop a specialized collaborative platform for digital painting that integrates AI-driven tools like DreamSheets for prompt-based design exploration and LumiMood for mood setting. This platform would include a shared digital canvas where artists can experiment with AI-generated suggestions, receive real-time feedback, and adjust various parameters to refine their artwork. It would feature adaptive interfaces and tutorials to support artists of different skill levels, enhancing both creativity and learning. [Expand](#)

Figure 5.6: The cold start of the baseline UI for the user study's idea-generation task. The ideation topic here is human-AI collaboration in art.

functionalities: on-demand novelty evaluation, manual idea addition, and facet generation when there is no query. The Idea Novelty Checker module was only activated in a separate ‘Idea Novelty Evaluation’ tab for the idea evaluation step. There was no support for adjusting the novelty assessment or iterating on the idea’s novelty. The tab also provided access to a ChatGPT-like interaction in which participants could prompt the LLM directly in order to help them evaluate their ideas for novelty. The chat kept a thread for each participant to remember their exchanges when forming a response. Lastly, the tab provided a text field where the participant could keep any notes on their novelty assessments.

For the baseline tool, in the ‘Idea Generation’ tab (Figure 5.6), participants could select any combination of the three input papers as input to the LLM `gpt-4o-2024-08-06`, the same LLM used for most of SCIDEATOR’s functionality. If they did not select any papers, all three were provided to the LLM. Participants could also provide custom instructions to the LLM like in the treatment tool, but the character limit was 75000 rather than 25000 to account for the fact that the set idea generation prompt was longer in the treatment tool. The LLM prompt for idea generation was a simplified version of the one in SCIDEATOR. It did not utilize any facet-based framework or carefully crafted criteria for a good idea. However, like SCIDEATOR, it generated six candidate ideas for every two presented to the participant and followed instructions to improve upon the idea. The ‘Idea Novelty Evaluation’ tab was similar to that in the treatment tool except there was no Idea Novelty Checker module output (i.e., related papers, novelty classification, and classification reason for each idea).

#### 5.5.4 Procedure

Each within-subjects study session was 105 minutes. The sessions were recorded and transcribed using Google Meet.<sup>15</sup> In each condition, the session coordinator provided the participant with the assigned tool, a document with the titles and abstracts of the input papers for the assigned ideation topic, and a link to the scientific search engine Semantic Scholar<sup>16</sup>. They had access to these three resources throughout the condition. The participant completed two tasks with each tool— an idea-generation task followed by an idea-novelty-evaluation task.

For the **idea-generation task**, the participant entered their assigned ideation topic and three input papers into the tool. While the tool loaded, the coordinator went over the task instructions and gave the participant a tutorial describing the tool’s features. The participant then had up to two minutes to review the three input papers’ titles and abstracts. With access to the tool, Semantic Scholar, and the input paper document, the participant subsequently spent 20 minutes generating and saving as many research ideas as possible. To save an idea, the participant had to confirm that the idea was at least somewhat relevant to the ideation topic and somewhat interesting to think about further. They also had to provide a seven-point Likert-type rating of how different the idea was from ideas that they had or encountered before the study; they were told to aim for saving ideas that were at least somewhat significantly different. The coordinator alerted the participant when five minutes remained. If the participant was rating an idea to save it when 20

---

15. The study script may be found in the supplementary materials.

16. <https://www.semanticscholar.org/>

minutes had passed, they completed the rating before moving on to the next step.

Once 20 minutes had passed, the participant opened a ‘Saved Ideas’ tab to select their two favorite ideas and answer additional 7-point Likert-type questions about their perceived novelty, feasibility, specificity, impact, and imaginativeness of each idea.<sup>17</sup> The participant also rated their confidence in their novelty assessment. The participant then completed a survey regarding their experience with the idea-generation task. This included seven-point Likert-type questions about their familiarity with the assigned topic and if they encountered concepts that they had not previously heard about or encountered in the context of the ideation topic. It also included the well-established Creativity Support Index (CSI) questionnaire [29]. In the survey for the second tool, the participant also answered questions for each pair of CSI factors to determine which factors they considered most important, as is standard for the CSI. The coordinator then spent up to around five minutes engaging the participant in a semi-structured interview about their idea generation experience.

Moving on to the **idea-novelty-evaluation task**, the participant opened the ‘Idea Novelty Evaluation’ tab, and the coordinator provided an overview of this portion of the tool. The participant spent five minutes evaluating their two favorite ideas for novelty. For each idea, they provided a final seven-point Likert-type rating of perceived novelty and confidence in their novelty assessment. Finally, the coordinator spent up to around three minutes conducting a semi-structured interview about the

---

17. There were two instances in which a participant had only saved one idea in the 20 minutes allotted. In this case, we asked them to select their next favorite idea in order to proceed with two favorite ideas.

participant’s idea evaluation experience.<sup>18</sup>

## 5.6 User Study Results

We analyzed participants’ survey responses, interaction logs, and semi-structured interview responses from the user study.<sup>19</sup> We analyzed the interview responses using inductive thematic analysis [18]. We refer to participant by their unique ID number, research area, and tool they were using at the time (e.g., P1-HCI-treatment).

### 5.6.1 RQ1: Creativity Support in Idea Generation

To compare participants’ responses to the Creativity Support Index (CSI) questionnaire after using SCIDEATOR versus the baseline tool, we conducted a Wilcoxon signed-rank test<sup>20</sup>. As shown in Figure 5.7a, participants experienced significantly more creativity support with SCIDEATOR (M=70.50, Q1=57.50, Q3=79.00) compared to the baseline (M=61.00, Q1=42.25, Q3=71.50) (Wilcoxon Signed-Rank Test, V=208.50,  $p < .01$ )<sup>21</sup>. Breaking the CSI down into its factors, we saw that participants benefited most from SCIDEATOR in terms of exploration, followed by expressiveness. Most participants also benefited from SCIDEATOR in terms of enjoyment and their results being worth the effort, but participants did not experience much of a differ-

---

18. All survey and interview questions may be found in the supplementary materials.

19. Survey responses and consenting participants’ interaction logs are provided in the supplementary materials.

20. The data did not violate the assumption of symmetry of within-subjects differences about the median

21. The unweighted CSI score range is from 0 - 100

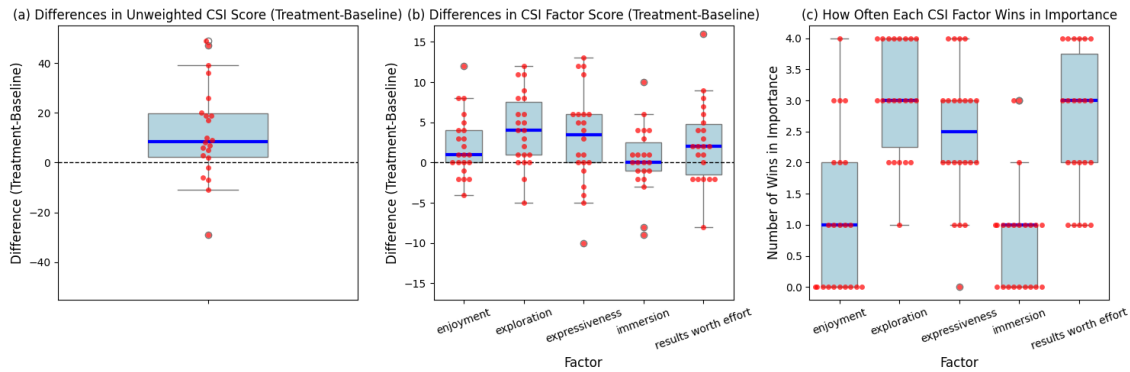


Figure 5.7: (a) The difference between participants’ unweighted CSI scores for SCIDEATOR versus the baseline tool. Participants experienced significantly more creativity support with SCIDEATOR. (b) For each CSI factor, the difference between participants’ ratings for SCIDEATOR versus the baseline tool. (c) How many times each CSI factor wins against other factors in terms of what is most important to participants while generating ideas.

ence from the baseline with respect to immersion (Figure 5.7b). Looking at which factors participants found most important for the task of generating ideas (Figure 5.7c), we observed that participants chose exploration over other factors the most, followed by results-worth-effort, expressiveness, enjoyment, and finally immersion. Thus, SCIDEATOR largely benefited the participants in the manners that were most important to them.

### Exploration Factor

SCIDEATOR helped participants to explore different ideas, the most important CSI factor according to our participants. However, looking at participants’ 7-point Likert-type ratings of their favorite ideas’ average newness to them, we see that the median participant had only a slightly higher rating in the treatment than the baseline

(treatment-baseline:  $M=0.50$ ,  $Q1=-0.375$ ,  $Q3=0.875$ ). Nonetheless, in the interviews, when participants commented that they found new concepts in the baseline condition, it was most often from the input papers rather than the tool’s output (6 of 8 participants). On the other hand, in the treatment condition, participants who identified new concepts cited the tool’s facets or ideas as the source (6 of 6 participants). For example, P18-HCI-treatment shared, “*When I thought of human-AI collaboration in art, for example, I did not think about also supporting artistic pursuits of students.... When I was thinking about the topic, I thought more about... a human prompting an AI for generating images or for image exploration which is more related to the papers that were given..*” Meanwhile, P5-NLP-baseline reflected, “*The papers themselves were really interesting, but I don’t think the tool generated anything super beyond a synthesis of the ideas that were in those three papers.*”

As four participants noted in their interviews, the facet-level interaction may have also supported exploration by providing more transparency to help them understand the idea generation process. For example, P5-NLP-treatment reflected, “*I think the first thing that I noticed was that it was very easy to context switch. That was my main problem with the [other] tool before. I couldn’t figure out which idea dealt with what aspect of the research that I was engaging with. Very easy to do that here.*”

## Expressiveness and Enjoyment Factors

SCIDEATOR also helped participants to express and enjoy themselves while generating ideas. In interviews, 14 participants noted that SCIDEATOR’s facet-level interaction, designed to help researchers express their ideas, was useful or interesting. In par-

ticular, seven participants appreciated how it provided them more control over the idea generation process. For instance, P11-NLP-treatment explained, “*I like this tool better because it sort of distilled the different aspects of the input papers into very concrete blocks that you could plug into each other.... it’s just that the information was presented in this tool... in a more digestible manner, and that helped combine information across papers better.*” Furthermore, as described in Section 5.6.2, participants included custom instructions to the LLM more in the baseline tool, indicating that SCIDEATOR helped them to express their thoughts without the need to type a prompt themselves.

Also of note, after using the baseline tool, four participants said that they wanted a way to input more papers to better express themselves, and five more felt limited by the three input papers. P14-NLP-baseline, for example, “*would have liked to add a different paper because it felt like I had exhausted... the creativity in the system to some extent.*” While participants could add information from papers to their custom instructions, there was no system feature for adding more papers to the list of input papers. Future work may compare SCIDEATOR with a modified version of the baseline tool that allows users to add as many papers as they want for recombination.

## Results-Worth-Effort Factor

Participants generally found their results to be more worth the effort while using SCIDEATOR compared to the baseline tool. However, comparing SCIDEATOR and the baseline tool, there was little difference (treatment - baseline) in participants’ average ratings of their favorite ideas in terms of perceived feasibility (M=0.00, Q1=-0.50,

Q3=0.50), specificity (M=0.00, Q1=-0.88, Q3=0.50), and imaginativeness (M=0.00, Q1=-0.50, Q3=0.38). Meanwhile, the baseline tool performed better with respect to generating impactful ideas (M=-0.25, Q1=-0.50, Q3=0.00). Perhaps participants could more clearly see the potential impact of ideas that were grounded in just a few set input papers, which they reviewed, versus several papers, most of which were not reviewed by them. Comparing their baseline experience to their treatment experience, P10-HCI-baseline posited, “*since now I know the paper, I read them, I kind of understand the vocabulary... it is easier for me to see where these ideas are coming from. So even when the ideas are written somewhat vaguely, I can still... imagine how that would pan out because I read the paper.*” Thus, SCIDEATOR’s results being more worth the effort may predominantly be based on SCIDEATOR helping participants find ideas that utilize more concepts than what is already in the input papers or participants’ minds (see Section 5.6.1). A sample of participants’ favorite ideas from the two conditions is in Table 5.2.

## Immersion Factor

Overall, participants did not find SCIDEATOR more helpful than the baseline tool for becoming immersed in the idea-generation task. The interviews and interaction logs provide some reasons why this may be true. For one, SCIDEATOR presents the user with several features about which to learn. The cognitive demand of learning about these features might have prevented immersion. Four participants commented on the high cognitive load of using SCIDEATOR; P2-HCI-treatment commented, “*I would say that it took me more mental effort to figure out how the tool [is used]*

*rather than work with ideas.*” Furthermore, the system’s latency in outputting ideas may have made it difficult to stay immersed in the task. Due to more complex prompting, the average latency for generating two ideas in SCIDEATOR was 22.04 seconds, compared to 15.21 seconds in the baseline tool.<sup>22</sup>

### 5.6.2 RQ2: Idea Generation Patterns

#### Benefits of Facet-Level and Paper-Level Interactions

In our study, participants can generate ideas from inputs with two levels of granularity: facet-level (SCIDEATOR condition) and paper-level (baseline condition). Participants commented on benefits of both the baseline tool and SCIDEATOR in terms of their input granularity. Fourteen participants found SCIDEATOR’s affordance for facet-level interactions useful or interesting, noting increased control (Section 5.6.1) and greater transparency (Section 5.6.1) as advantages of the facet-level interaction. On the other hand, five participants appreciated the baseline’s paper-level interactions in addition to or more than SCIDEATOR’s facet-level interactions. Three of these participants liked the paper-level interaction as it felt more directly connected to the literature. P22-NLP-baseline explained, “ *I think papers for me were more natural than facets.... I think to me it’s more like a map of literature, so I could see it more with papers..*”

Three participants thought a combination of the two tools would be helpful. Two participants even proposed distinct roles for the two tools: divergent-ideation for

---

22. Ideas were generated in groups of four, two at a time, so the first two ideas took less time to generate than the last two ideas.

SCIDEATOR and convergent-ideation for the baseline. P7-HCI-treatment shared, “*In the [baseline tool], I started from a broader view and then I narrowed it down. Here [in SCIDEATOR], I started from a very specific thing and then I tried to add new facets or ideas so that I can expand the idea. So you see the other process is elimination process, here I was trying to expand.*”

### Lower Need for Custom Instructions with Scideator

When using SCIDEATOR, participants most often did not provide custom instructions to the LLM for generating either of their two favorite ideas (M=0.00, Q1=0.00, Q3=1.00 of 2 ideas). On the other hand, when using the baseline, the median participant provided custom instructions for generating 1.50 of their two favorite ideas (Q1=1.00, Q3=2.00). This indicates participants were able to utilize the scaffolding in the treatment to express themselves rather than have to type their own custom instructions. P1-HCI-treatment commented, “*I didn’t need to add any custom instructions because these [facets] served like custom instructions.*”

### Facets Used for Saved Ideas in Scideator

In their interviews, seven participants described how they appreciated the ability to control their ideas’ generation through the facet-level interaction (Section 5.6.1). This aligns with participants’ proclivity to select facets for their favorite ideas themselves. Participants’ favorite ideas more often included evaluations, mechanisms, and especially purposes selected by themselves rather than the LLM (Figure 5.8a). Given that participants were assigned ideation topics, P18-HCI-treatment explained why

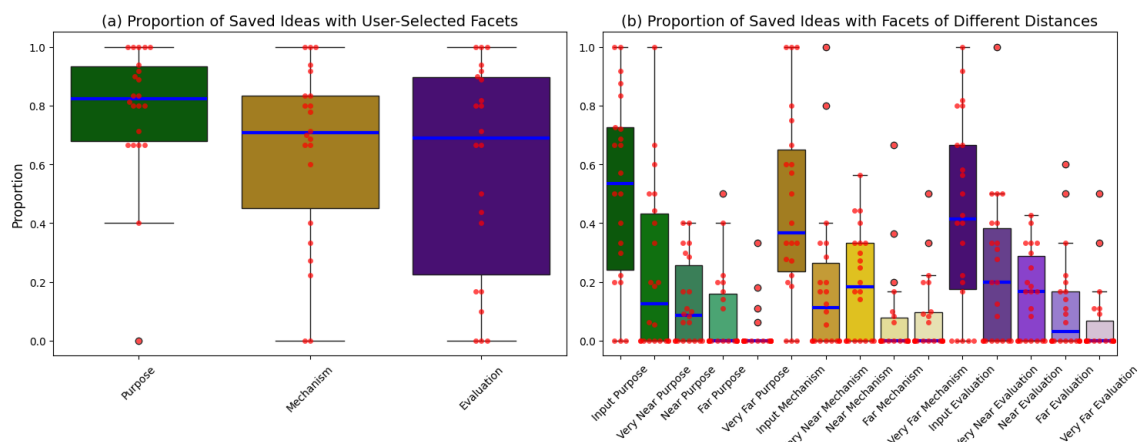


Figure 5.8: (a) Participants more often opted to select their own facets rather than let the LLM select for them. (b) Participants used input facets and facets nearer to the input more than facets farther from the input.

participants may have decided to prioritize selecting purposes themselves: *“I think the purpose is the most relevant to the topic. So within an area, there can be many ways of doing the same tasks, but the task is ultimately what defines the area.”*

Participants tended not to find the evaluation facets as helpful as the purpose and mechanism facets. Four participants commented that they found the evaluation facet unimportant. P13-HCI-treatment elaborated, *“For evaluation, I really don’t think it’s necessary for me because once you have the problem, you have the solution. Automatically you know how to evaluate it, like what study you need, what kind of experiment you want to have, and what variables you are measuring.”* Thus, future work may investigate whether or not the presence of an evaluation facet is useful for mixed-initiative, facet-based generation of research ideas.

In addition, we explored how participants utilized facets of different distances from the input. We observed that participants were reluctant to use far facets,

particularly for an idea’s purpose (Figure 5.8b). Participants’ primary reason for avoiding far facets was that they were not relevant enough to the ideation topic, though four participants found the far facets helpful for discovering different ideas. As an example, P9-NLP-baseline commented, “*that very near, near, far kind of thing... it kind of adds some sort of discovery factor.*”

### 5.6.3 RQ3: Confidence in Idea Novelty Assessment

For this research question, we include the 17 (of 22) participants who completed the idea-novelty-evaluation task with the intended task time and setup.<sup>23</sup> To compare how participants’ confidence in their novelty assessments changed after using SCIDEATOR versus the baseline tool for idea novelty evaluation, we conducted a sign test<sup>24</sup>. We found that SCIDEATOR’s novelty checker did not significantly improve participants’ confidence in their novelty assessments (Sign Test,  $S = 5.00$ ,  $p = \text{n.s.}$ ).

However, we also conducted a between-subjects comparison of ideas that were 1) classified as novel by the treatment, 2) classified as not-novel by the treatment, and 3) not classified by the baseline (Figure 5.9). Even though the change in participants’ *confidence* in their novelty assessments was not meaningfully different across the three situations, the change in participants’ perceived idea novelty was. We observed a trend in which participants who received a not-novel tool classification of their idea made more change to their own novelty assessments than participants who received

---

23. One participant did not get to spend the full five minutes allotted for the task, and four participants experienced a different version of the SCIDEATOR novelty checker due to an issue with an API call.

24. We had planned to run a Wilcoxon signed-rank test but found that the data violated the assumption of symmetry of within-subjects differences about the median

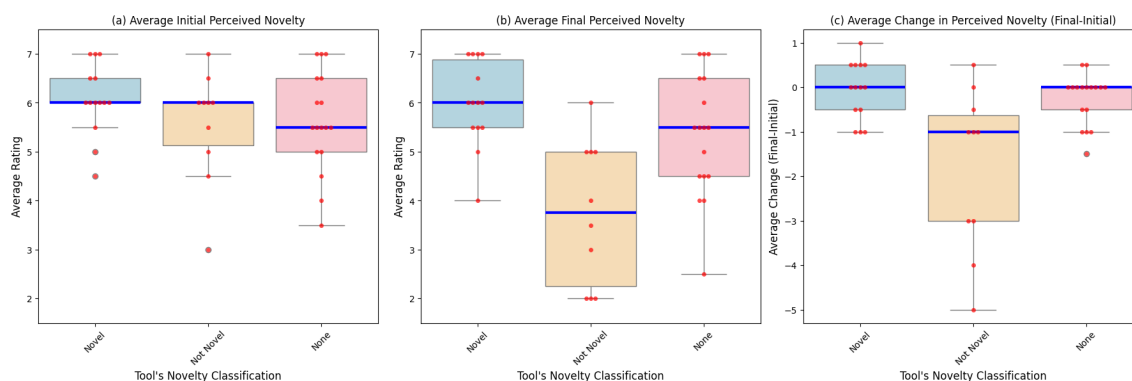


Figure 5.9: Participants’ average perceived idea novelty before (a) and after (b) utilizing their assigned tool for idea novelty evaluation, as well as the average change from initial to final perceived novelty (c). The baseline tool was the assigned tool when ideas had no classification, while SCIDEATOR was the assigned tool when ideas had a novel or not-novel classification.

a novel tool classification or no tool classification of their idea (Figure 5.9f).

This suggests participants’ novelty assessments are most impacted when SCIDEATOR classifies an idea as not novel. Given that it is straightforward to verify that an idea is not novel using the idea’s related papers provided by SCIDEATOR, it makes sense that it would be more likely to affect people’s classifications. Still, it is not guaranteed that people would agree with the classification after taking some time to review its related papers and reason. Thus, we have evidence that the novelty checker is useful for judging ideas that the classifier deems “not novel,” but not necessarily for judging ideas that it classifies as “novel.” P17-HCI-treatment echoed this sentiment: *“Seeing a list of related work is very helpful for giving you the context. It was very convincing in the case of telling me that an idea was not novel.... When it provides [a novel classification], it’s less convincing but is helpful.”* Three participants noted that the papers retrieved by the tool to evaluate an idea

for novelty were more useful to them than the tool’s novelty classification or reason. P8-NLP-treatment provided a rationale: “*I didn’t really paid much attention to these reasons because reasons can be kind of made up to explain why their generation is novel. So, I kind of relied more on the references that it retrieved.*”

## 5.7 Discussion

**Grounding mixed-initiative scientific ideation in research-paper facets shows promise for supporting idea generation.** We introduced SCIDEATOR, a mixed-initiative, LLM-powered tool for scientific ideation through research-paper facet recombination. While prior work established the utility of extracting purpose and mechanism facets from research papers for identifying scientific analogies [91, 23, 147], we are the first to apply this framework to a human-LLM interaction for scientific ideation, offering support not only in finding analogous papers but also in generating paper facets and recombining those facets in research ideas. Aligned with our goal of supporting divergent ideation [157, 39], results from our within-subjects user study showed that participants experienced significantly more creativity support when using SCIDEATOR as opposed to a strong baseline tool. Through semi-structured interview responses, we observed that participants appreciated being able to mix and match relevant purposes and mechanisms for research idea generation.

While we posit that facet-based interaction is an essential mechanism for ideation, we are aware that it is not the only one. Some participants also valued paper-level interactions, through which they could generate ideas from a set of selected papers rather than paper facets. Papers are directly connected to the literature and

compress a great deal of information into a single entity. We believe that tools that provide users opportunities to interact with ideas on different levels of granularity and abstraction (e.g., papers, facets, concepts) will lead to richer interactions and better outcomes.

**For mixed-initiative idea novelty assessment, our results suggest that providing related papers to the idea is key, and a ‘not novel’ classification may be more helpful than a ‘novel’ one.** When commenting on the benefits of SCIDEATOR’s idea novelty checker, participants more often mentioned the relevant papers that the tool would surface, rather than the tool’s novelty classifications or classification reasons. Participants appreciated being efficiently provided with preliminary means to conduct the novelty assessment themselves, whereas they sometimes did not trust or agree with the tool’s novelty classifications. That said, ‘not-novel’ classifications were seemingly the most utilized classifications, as participants noted they could easily verify whether or not the related papers render the idea not novel. Indeed, participants changed their novelty assessment the most when they received a ‘not-novel’ classification from SCIDEATOR, as opposed to a ‘novel’ classification from SCIDEATOR or no classification from the baseline tool. Future work may explore the benefits of a mixed-initiative tool that provides the best argument for a ‘not-novel’ classification regardless of the input idea.

**More support is needed for generating ideas that are both very new to the scientist and capture the scientist’s interest.** Although participants experienced more creativity support with SCIDEATOR, there is room for improvement in terms of helping scientists to come up with ideas that are very new to them yet

still relevant to their ideation topic. Participants’ favorite ideas were not meaningfully newer to them when using SCIDEATOR versus the baseline. Using far facets might have helped them to discover ideas more new to them, but some participants described avoiding far facets because they considered them too irrelevant to the assigned ideation topic. Correspondingly, we observed that participants did not save ideas with far facets as much as ideas with input and near facets. By avoiding distant facets, scientists may miss opportunities for generating ideas that are completely new to them, helping them think of more ideas that would never have occurred to them otherwise. While we provide short descriptions of facets in SCIDEATOR, we could take a step further with in-context question-answering to help users understand a facet to whatever degree is necessary. To help users recognize the utility of unfamiliar facets and the feasibility of associated ideas, the tool could also allow users to ask questions about how they might use the facet in general or in relation to a particular idea. In addition, the tool could be more transparent and provide users the analogy between two papers that it used to generate an idea. This could help users to more easily see the connection between the input or near facets and the far facets.

## 5.8 Limitations

Our study has a few limitations. First, the system was set up for computer science research ideas, and all participants had some experience with LLMs. Future work may investigate how scientists in other research areas with different LLM-familiarity might work with a tool like SCIDEATOR. Second, SCIDEATOR currently looks for analogies of varying distance within a broad research area (i.e., computer science),

but future work could explore extending to analogies across research areas. Third, there were only 22 participants. Given the small sample size, future work may validate the results observed here with larger samples. Fourth, participants had limited time to interact with the tools— not enough to master a system with multiple steps like SCIDEATOR. Future studies may explore how scientists utilize a tool like SCIDEATOR over a longer period of time. Fifth, we compared SCIDEATOR to a strong baseline— another LLM-powered tool for scientific ideation. However, it would be interesting to explore how interaction with SCIDEATOR compares to other tools for scientific ideation such as CoQuest [117]. Sixth, due to latency constraints, the system only extracted papers’ titles and abstracts and only utilized 10 relevant papers from Semantic Scholar’s database to assess the novelty of an idea. Future work could look into how scientists interact with the system when it has access to more paper text and relevant papers. Seventh, to keep the study relatively simple, we provided three papers to the tools as input. Future work may evaluate how the interaction changes with more input papers. Eighth, to keep the study controlled, we assigned ideation topics to participants rather than letting them select their own. Future work may examine how the research area, familiarity, and recency of input papers impacts the experience and outcome of working with SCIDEATOR.

## 5.9 Conclusion

We presented SCIDEATOR, a novel mixed-initiative tool for scientific ideation that extracts key facets (purposes, mechanisms, and evaluations) from a set of input and analogous papers. Users can explore diverse suggestions for how to recombine these

facets in succinct ideas. They can also use SCIDEATOR to evaluate and iterate on the facet-based ideas' novelty. We introduced three LLM-powered retrieval-augmented generation (RAG) modules to support SCIDEATOR's workflow: Analogous Paper Facet Finder, Faceted Idea Generator, and Idea Novelty Checker. We found that computer-science researchers experienced significantly more creativity support with SCIDEATOR than with a strong baseline, particularly in terms of exploring different ideas, which the scientists considered most important for idea generation.

## Chapter 6

### Conclusion

In this dissertation, I introduce three mixed-initiative methods for creative endeavors in the scientific process. All three methods utilize research-paper components as a shared language or boundary object [165, 71] between the scientist and AI, in order to help scientists verify and control AI suggestions that help them explore diverse ideas. By comparing the methods to strong baselines, I gather ample evidence that researchers working on creative steps of the scientific process benefit more from AI suggestions when mixed-initiative methods are implemented to help them (1) verify the suggestions for relevance to their creative vision and (2) steer the suggestions to better align with that vision, all while staying rooted in the literature.

The first method provides both local and global component-based explanations in order to help scientists verify and adjust what components influence their creative AI suggestions. I apply this method to the task of curating a research-paper feed, which recommends relevant papers from the sprawling literature that the user may not have otherwise thought to seek out. The component-based local and global explanations consist of keywords from research papers along with a representation of their weights. I find evidence that supplying both explanations rather than either alone helps scientists to more easily understand how to improve their recommendations.

The second method introduces an interactive reverse source outline for summarizing long-documents like research papers into detailed articles like blog posts. I implement the interactive outline in a system called Papers-to-Posts, which allows scientists to understand and control not only which paper components (in the form of

bullet points) are included in an LLM-drafted blog post, but also which components are excluded. This provides scientists both the creative freedom to explore different combinations of content to include as well as the ability to maintain high attention to detail when summarizing a long document, which is required in a domain like scientific research. Comparing Papers-to-Posts to a strong baseline, I observe that Papers-to-Posts leads to a significant increase in blog post satisfaction under time constraints, particularly with respect to content coverage.

The third method presents a human-LLM workflow to help scientists avoid fixation and gather inspiration from the literature for scientific ideation. The workflow is implemented in the system Scideator, which suggests diverse research-paper components to recombine, recombinations of components into new research ideas, and iterations on those ideas to improve their novelty. At each step, the scientist may verify and control the system's creative suggestions. In this case, the paper components are core facets: purpose, mechanism, and evaluation. The purpose is the problem addressed by the paper, the mechanism is the proposed method to solve the problem, and the evaluation is the method to determine if the mechanism addresses the purpose. I find that scientists experience significantly more creativity support with Scideator compared to a strong baseline.

## Chapter 7

### Discussion and Future Work

While there has been significant progress in methods for human-AI interaction, both in scientific research and beyond, many interesting avenues for future work remain. In this chapter, I review a few of the most relevant avenues to this dissertation. Furthermore, I reflect on observations from my dissertation to discuss what tasks within a mixed-initiative workflow may be more suitable to humans versus AI.

#### 7.1 Evaluating Mixed-Initiative Methods' Impact on Process rather than Outcome for Scientific Research

Mixed-initiative methods to support scientific research may be evaluated in terms of the artifact they produce or the experience they provide. For example, Chapter 4 focused on how Papers-to-Posts supported blog post satisfaction, a measure of output quality, while Chapter 5 focused on how Scideator assisted scientists in creative exploration, a measure of process quality. Reflecting on this dissertation, I argue that the process is equally if not more important to evaluate, particularly with respect to how it affects the scientist's thinking and behavior. As AI systems become more present throughout the scientific workflow, it is crucial to understand how they may influence how scientists operate. The benefit from mixed-initiative systems for creativity stems from their ability to combine the strengths of humans (i.e., unique perspectives, understanding of human priorities) with the strengths of AI (i.e., access to a diversity of ideas, lack of bias). Scientists must continue to be able to engage

in critical thinking about their work, independently from AI. If not, they risk losing their distinct viewpoints, and the synergy between the two entities is likely to fail. I call on future work to take seriously the impact of mixed-initiative methods on the scientific mind. This will require not only evaluating mixed-initiative methods in terms of the user experience (e.g., creativity, cognitive load, efficiency) but also conducting longitudinal studies to elucidate how scientists are adapting to these methods over time. In alignment, I encourage future work to develop mixed-initiative methods with the goal of supporting scientists in developing their creativity and critical thinking skills, as opposed to merely helping scientists achieve better short-term outcomes.

## **7.2 Investigating Generalizability of the Mixed-Initiative Methods**

The three mixed-initiative methods that I introduce in this dissertation may be applied to help users understand and control various AI-facilitated creative endeavors of the scientific process and beyond. However, each method is not fully present in all three projects because the methods do not generalize to all use cases, particularly with respect to tasks requiring predictive AI systems versus tasks requiring generative AI systems. Here, I review why the methods were or were not present in each project and, more broadly, for what creative tasks the methods might be best suited.

### 7.2.1 *Local and Global Component-Based Explanations*

For the first method, I provide both component-based local and global explanations for creative AI suggestions within an AI recommender system for the scientific task of identifying interesting research papers to read. While not the emphasis of the other two projects, the concept of local and global explanations may be applied to Papers-to-Posts and Scideator as well. In Papers-to-Posts, the interactive reverse source outline presents a local explanation in the form of selectable bullet points that are used to generate the current blog post draft. Papers-to-Posts does not include a true global explanation, which would explain the LLM’s process for synthesizing any selection of bullet points. In Scideator, the selected facets associated with each output research idea may be considered local explanations, but I did not include a global explanation, which would explain the LLM’s process for combining a given set of facets.

I did not include a global explanation in the LLM-supported systems of Papers-to-Posts or Scideator because understanding the AI’s process for generating output *independent of input* appears less useful in the context of generative AI. In contrast with predictive AI output, generative AI output itself, along with the prompts that produce the output, already essentially summarize how inputs are synthesized, and the user can steer the synthesis process with natural language instructions. Furthermore, creating global explanations for generative AI models would be more complex than for predictive AI models. Predictive AI models can be approximated with simpler, more interpretable models. For example, a classifier can be approximated with a logistic regression model, and a regression model can be approximated with a linear

regression model. On the other hand, even the simplest generative AI models are difficult to interpret, without clear-cut features being used for inference.

Perhaps global explanations should be re-imagined for the context of generative AI systems. Given the relative transparency of output synthesis in generative AI, global explanations for generative AI might focus more on summarizing the model’s development (e.g., training data, evaluation metrics, ethical considerations), as is currently detailed in model cards [133], rather than summarizing the model’s inference. The explanations could be more detailed than model cards, describing the model developers’ motivations for the final model settings. For example, the model’s fine-tuning might have involved reinforcement learning with human feedback that focused on making the model’s responses less toxic. Noting this point would help the user to recognize why the model might tend to give more positive, agreeable responses.

**Future work may investigate how both global and local explanations impact verification and control of model behavior in generative AI systems, as compared to predictive AI systems.** First, formative studies should be done to understand how global and local explanations could be designed for a few different common generative AI use cases (e.g., text summarization, question answering). Next, user studies could investigate if local, global, or both explanations improve upon a no-explanation baseline in terms of supporting users in understanding and steering the generated model output. I suspect that local explanations will almost always be useful for verifying that the model is working as expected and understanding how to better steer the model. We observed this with Papers-to-Posts

and Scideator, and local explanations in the form of chain-of-thought reasoning is popular in many AI products today, including ChatGPT's and Google Gemini's Deep Research features <sup>1</sup>. Contrastingly, I suspect that a global explanation will only be helpful if the explanation provides information on how the model was trained. If the explanation only provides information on how the model performs inference, this may not provide much insight beyond what the natural-language prompt-completion pair provides.

### *7.2.2 Outline of Selected and Unselected Components*

For the second method, I present an interactive reverse source outline for summarizing long-documents in detailed articles, which is applied to the task of translating research papers blog posts. The other projects do not focus on supporting users in identifying literature components that are included versus excluded from creative AI suggestions, but both projects do present relevant features. In the project on local and global explanations, the global explanation allows users to search for terms that the recommender system might have weighted. In this way, the user can determine if a certain literature component has positive or negative influence on the recommendations. Unlike in the case of the interactive reverse source outline, however, this feature does not help users to understand the full space of literature components available. In the Scideator project, on the other hand, users can see the full space of facets that the system considers for its suggestions because it is a generative AI system like in the Papers-to-Posts project. Such a system is able to synthesize cre-

---

1. <https://chatgpt.com/>,<https://gemini.google.com/app>

ative suggestions from only a few inputs, and the space of the potential inputs is artificially constrained by the system scaffolding. In a recommender system like in the local-global project, reviewing the full space of literature components that are inputs to the system would be infeasible in terms of cognitive load. The space would need to be summarized first before presenting it to the user, which is essentially the role of the global explanation in recommender systems. **Future work may explore how to best summarize the full space of literature components available in predictive AI systems for creative endeavors in science.**

### *7.2.3 Steerable Recombination of Components*

For the third method, I introduce a human-LLM workflow for scientific ideation through research-paper facet recombination. The other two projects do not support component recombination. The local-global project does not support component recombination because it involves a recommender system, which does not generate any output to utilize facet recombination. While the Papers-to-Posts project does iteratively recombine paper bullet points in a blog post, the system does not itself suggest different bullet point recombinations. It only suggests one recombination per blog post section, which the user can then adjust.

**Future work may look into how generative AI systems can support exploration of literature component recombinations at creative steps of the scientific process other than initial idea generation, such as research communication or study design.** For instance, Papers-to-Posts could be extended for multi-document summarization and help users to explore recombining bullet points

of different papers in diverse manners. While it does not use components from specific papers, IdeaSynth is a relevant work [148]. It allows scientists to build a research brief based on an input idea by combining different potential components of a brief, such as “proposed design and solution” and “contribution and impact.”

### **7.3 Exploring Spectrum of Scientific Literature Components**

The three methods that I present in this dissertation all use scientific literature components, which can be easily manipulated by both the scientist and AI in terms of weight, selection, or recombination. The components help scientists to balance creativity with practicality, by encouraging diverse lines of thinking but also grounding those ideas in the literature.

The methods utilize distinct types of literature components, which differ primarily in terms of granularity. The local-global project has the most coarse-grained components—paper keywords. Scideator has medium-grained components, which are short phrases describing the purpose, mechanism, and evaluation facets of papers. Papers-to-Posts has the most fine-grained components—bullet points from a paper outline. The granularity of the components increases with progress along the scientific process. When a scientist is at the beginning stages of a new project and is simply looking for inspiration from the literature, they generally do not need highly specific suggestions on what to read. Rather, they need a variety of high-level paths to explore. Conversely, when a scientist is at the final stages of a project and wants assistance in communicating their work with the world, they need suggestions that properly convey their work with sufficient detail. Nevertheless, there are examples

of prior work that indicate relatively fine-grained paper details are useful in early steps of the scientific process, and vice versa. For instance, PaperWeaver, which supports scientists in the early step of identifying interesting papers to read, provides recommendation explanations containing much more detail than just keywords, such as the paper’s method and findings [112]. As another example, ConceptEVA allows users to summarize papers based just on keywords. Although the system is intended to support an early step in the scientific process—summarizing new papers that one reviews for inspiration, the system could also theoretically help scientists with a late step in the process—summarizing their own papers to share their work with broader audiences [192]. **Future work may more thoroughly investigate the relationship between the granularity of literature components and their utility over the course of the scientific process.**

One may also consider the type of system paired with the granularity of the literature component. Compared to predictive AI systems, generative AI systems perform well with more detailed inputs to steer them, so it makes sense that generative AI systems would be better suited to more fine-grained literature components. In this dissertation, the earlier-step project involves a recommender system, while the later-step projects involve generative AI systems. If a generative AI system is applied to an earlier step in the scientific process, such as paper recommendation, then somewhat more fine-grained components might be appropriate, but this would still need to be balanced with the fact that the scientist should explore a diverse array of ideas without getting bogged down in details. Similarly, if a predictive AI system is applied to a later step in the scientific process, such as recommendations

for summarizing a project, then more coarse-grained components might be useful, but this would need to be balanced with the fact that the scientist should home in on details suitable to describe their particular project. **Future work may examine to what extent more granular literature components are more useful for generative AI use cases, as compared to predictive AI use cases.**

## **7.4 The Role of the Human and AI for Different Tasks in a Mixed-Initiative System**

Through my work on this dissertation, I have identified task characteristics that appear more amenable to humans versus AI systems. When tasks are repetitive and objective, using AI alone seems to be the better choice. Redundancy and clear rules make these tasks relatively easy to automate and tedious to a human. Examples of such tasks include writing bullet points to summarize a paragraph (Chapter 4) and finding papers relevant to a generated idea (Chapter 5). Outside of my work, examples might include generating search terms to support legal practitioners in precedent search [65] or doing a first pass at translating code from one language to another [183].

On the other hand, subjective or creative tasks require more diversity of thought and offer more interesting processes for humans to control. Because AI systems usually present averaged or at least amalgamated viewpoints, they are not ideal for presenting unique or rare ideas. Thus, it seems advantageous for the human to complete subjective and creative tasks, using AI output only for inspiration. For instance, in Chapter 4, the human iterates on what bullet points should contribute

to their blog post’s content after the AI presents an initial bullet point selection, and in Chapter 5, the human identifies paper facets that they think are interesting to recombine after the AI presents a large set of potential paper facets to recombine. Examples of subjective tasks beyond my dissertation might include crafting a movie script [170] or identifying an angle to take on a press release [146].

Furthermore, tasks that are higher-stakes or the final steps in an overarching process appear to benefit from a combination of a human and AI for maximal verification. The AI may attempt to complete the task, and the human can verify its output to ensure that the final outcome meets their standards. As an example, in Chapter 5, the final step of deciding whether an idea is novel or not is the result of the AI and human working together. The AI presents relevant papers to the idea along with its binary novelty classification and reasoning based on the relevant papers. The human then reviews the AI’s reasoning and adjusts the novelty assessment as they see fit. Separate from my dissertation, high-stakes examples might include tutoring a student [181] or determining if a patient has one disease or another [143].

Meanwhile, tasks that are lower-stakes or earlier steps in an overarching process usually present less risk for AI automation, as the human will subsequently have the opportunity to improve the AI’s output. For example, in Chapter 3, generating an initial paper recommendation feed was completely AI-automated. This was the first step in the process of curating a paper recommendation feed of interest to the user. Later, the user was able to step in and provide feedback on the keywords being utilized by the feed to improve the feed’s relevance to the user. Examples apart from my work might include creating news article summaries that will be post-edited by a

human [108] or transcribing a meeting for an attendee, who can validate the content [26].

## References

- [1] Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Mohammed Ali, and Adam Jatowt. 2025. Rankify: A Comprehensive Python Toolkit for Retrieval, Re-Ranking, and Retrieval-Augmented Generation. <https://api.semanticscholar.org/CorpusID:276107364>
- [2] J. Ahn, P. Brusilovsky, J. Grady, D. He, and Sue Yeon Syn. 2007. Open user profiles for adaptive news systems: help or harm?. In *WWW '07*.
- [3] Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability. *arXiv preprint arXiv:2401.08574* (2024).
- [4] Shm Garanganao Almeda, JD Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for discovery: Flexible sense-making for ai art-making with dreamsheets. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [6] Tal August, Lauren Kim, Katharina Reinecke, and Noah A Smith. 2020. Writing strategies for science communication: Data and computational analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5327–5344.
- [7] PVS Avinesh, Carsten Binnig, Benjamin Hättasch, Christian M Meyer, and Orkan Özyurt. 2018. Sherlock: A System for Interactive Summarization of Large Text Collections. *Proc. VLDB Endow.* 11, 12 (2018), 1902–1905.
- [8] PVS Avinesh and Christian M Meyer. 2017. Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1353–1363.
- [9] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738* (2024).

- [10] F. Bakalov, M. Meurs, B. König-Ries, Bahar Sateli, R. Witte, G. Butler, and A. Tsang. 2013. An approach to controlling user models and personalization effects in recommender systems. In *IUI '13*.
- [11] Davide Baldelli, Junfeng Jiang, Akiko Aizawa, and Paolo Torroni. 2024. TWOLAR: A TWO-Step LLM-Augmented Distillation Method for Passage Reranking. *ArXiv abs/2403.17759* (2024). <https://api.semanticscholar.org/CorpusID:268691914>
- [12] Farnoush Bayatmakou, Azadeh Mohebi, and Abbas Ahmadi. 2022. An interactive query-based approach for summarizing scientific documents. *Information Discovery and Delivery* 50, 2 (2022), 176–191.
- [13] Antoine Blanchard. 2011. 3.4 Science Blogs in Research and Popularization of Science: Why, how and for whom? *Common knowledge: The challenge of transdisciplinarity* (2011), 219–232.
- [14] Tanner Bohn and Charles X Ling. 2021. Hone as you read: A practical type of interactive summarization. *arXiv preprint arXiv:2105.02923* (2021).
- [15] Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology* 66, 11 (2015), 2215–2222.
- [16] Svetlin Bostandjiev, J. O’Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *RecSys '12*.
- [17] Svetlin Bostandjiev, J. O’Donovan, and Tobias Höllerer. 2013. LinkedVis: exploring social and semantic career recommendations. In *IUI '13*.
- [18] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [19] S. Bruns, André Calero Valdez, Christoph Greven, M. Ziefle, and U. Schroeder. 2015. What Should I Read Next? A Personalized Visual Publication Recommender System. In *HCI*.
- [20] Terry W Burns, D John O’Connor, and Susan M Stocklmayer. 2003. Science communication: a contemporary definition. *Public understanding of science* 12, 2 (2003), 183–202.

- [21] Daniel Buschek. 2024. Collage is the New Writing: Exploring the Fragmentation of Text and User Interfaces in AI Tools. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2719–2737.
- [22] Pengshan Cai, Fei Liu, Adarsha Bajracharya, Joe Sills, Alok Kapoor, Weisong Liu, Dan Berlowitz, David Levy, Richeek Pradhan, and Hong Yu. 2022. Generation of patient after-visit summaries to support physicians. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*.
- [23] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–21.
- [24] Joseph Chee Chang, Nathan Hahn, Adam Perer, and A. Kittur. 2019. SearchLens: composing and capturing complex user interests for exploratory search. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [25] Liuqing Chen, Yuan Zhang, Ji Han, Lingyun Sun, Peter Childs, and Boheng Wang. 2024. A foundation model enhanced approach for generative design in combinational creativity. *Journal of Engineering Design* 35, 11 (2024), 1394–1420.
- [26] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. Meetscript: designing transcript-based interactions to support active participation in group video meetings. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.
- [27] Alan Y Cheng, Meng Guo, Melissa Ran, Arpit Ranasaria, Arjun Sharma, Anthony Xie, Khuyen N Le, Bala Vinaithirthan, Shihe Luan, David Thomas Henry Wright, et al. 2024. Scientific and fantastical: Creating immersive, culturally relevant learning experiences with augmented reality and large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [28] Ruijia Cheng, Alison Smith-Renner, Ke Zhang, Joel R Tetreault, and Alejandro Jaimes. 2022. Mapping the design space of human-ai interaction in text summarization. *arXiv preprint arXiv:2206.14863* (2022).

- [29] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.
- [30] DaEun Choi, Sumin Hong, Jeongeon Park, John Joon Young Chung, and Juho Kim. 2024. CreativeConnect: Supporting Reference Recombination for Graphic Design Ideation with Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
- [31] Seulgi Choi, Hyewon Lee, Yoonjoo Lee, and Juho Kim. 2024. VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.
- [32] Johan SG Chu and James A Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* 118, 41 (2021), e2021636118.
- [33] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022). <https://api.semanticscholar.org/CorpusID:247625751>
- [34] CMU. [n. d.]. CMU. <https://blog.ml.cmu.edu/submissions/>.
- [35] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* (2020).
- [36] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. *ArXiv abs/2004.07180* (2020). <https://api.semanticscholar.org/CorpusID:215768677>
- [37] Simon Colton and Geraint A Wiggins. 2012. Computational creativity: The final frontier? In *ECAI 2012*. IOS Press, 21–26.
- [38] Henriette Cramer, V. Evers, Satyan Ramlal, M. V. Someren, L. Rutledge, N. Stash, Lora Aroyo, and B. Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18 (2008), 455–496.

- [39] Arthur Cropley. 2006. In praise of convergent thinking. *Creativity research journal* 18, 3 (2006), 391–404.
- [40] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
- [41] Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259* (2024).
- [42] Douglas L Dean, Jill Hender, Tom Rodgers, and Eric Santanen. 2006. Identifying good ideas: constructs and scales for idea evaluation. *Journal of Association for Information Systems* 7, 10 (2006), 646–699.
- [43] Laura Devendorf, J. O’Donovan, and Tobias Höllerer. 2012. TopicLens : An Interactive Recommender System based on Topical and Social Connections.
- [44] Zijian Ding and Joel Chan. 2023. Mapping the design space of interactions in human-AI text co-creation tasks. *arXiv preprint arXiv:2303.06430* (2023).
- [45] Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of LLMs: Evaluating human-AI text co-creation through the lens of news headline generation. *arXiv preprint arXiv:2310.10706* (2023).
- [46] J. Dodge, Q. Liao, Y. Zhang, R. Bellamy, and C. Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [47] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and D. Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [48] Karl Duncker and Lynne S Lees. 1945. On problem-solving. *Psychological monographs* 58, 5 (1945), i.

- [49] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications* 165 (2021), 113679.
- [50] James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan O Arik, Yan Liu, and Tomas Pfister. 2023. Textgenschap: Scalable post-hoc explanations in text generation with long documents. *arXiv preprint arXiv:2312.01279* (2023).
- [51] A. Felfernig and B. Gula. 2006. An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications. *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)* (2006), 37–37.
- [52] Linda S. Flower and J. R. Hayes. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication* 32 (1981), 365–387. <https://api.semanticscholar.org/CorpusID:18484126>
- [53] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *arXiv preprint arXiv:2310.07581* 6, 3 (2023).
- [54] G. Friedrich and M. Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Mag.* 32 (2011), 90–98.
- [55] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Zijian Zhang, Wanyu Wang, Yuyang Ye, Shanru Lin, Huifeng Guo, and Ruiming Tang. 2024. LLM-enhanced Reranking in Recommender Systems. *ArXiv abs/2406.12433* (2024). <https://api.semanticscholar.org/CorpusID:270562015>
- [56] Yang Gao, Christian M Meyer, and Iryna Gurevych. 2020. Preference-based interactive multi-document summarisation. *Information Retrieval Journal* 23, 6 (2020), 555–585.
- [57] K. Gero, Vivian Liu, Sarah Huang, Jennifer Lee, and Lydia B. Chilton. 2021. What Makes Tweetorials Tick: How Experts Communicate Complex Topics on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 26. <https://api.semanticscholar.org/CorpusID:239020743>

- [58] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference*. 1002–1019.
- [59] Google. [n. d.]. NotebookLM. <https://notebooklm.google/>.
- [60] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* (2025).
- [61] Brynjar Gretarsson, J. O’Donovan, Svetlin Bostandjiev, C. Hall, and Tobias Höllerer. 2010. SmallWorlds: Visualizing Social Recommendations. *Computer Graphics Forum* 29 (2010).
- [62] Madeleine Grunde-McLaughlin, Michelle S Lam, Ranjay Krishna, Daniel S Weld, and Jeffrey Heer. 2025. Designing LLM chains by adapting techniques from crowdsourcing workflows. *ACM Transactions on Computer-Human Interaction* 32, 3 (2025), 1–57.
- [63] Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. 2024. LLMs can realize combinatorial creativity: generating creative ideas via LLMs for scientific research. *arXiv preprint arXiv:2412.14141* (2024).
- [64] Yuling Gu, Oyvind Tafjord, and Peter Clark. 2023. Digital socrates: Evaluating llms through explanation critiques. *arXiv preprint arXiv:2311.09613* (2023).
- [65] Jiarui Guan, Ruishi Zou, Jiajun Zhang, Kimpan Xin, Bingsu He, Zhuhe Zhang, and Chen Ye. 2025. Designing Human-AI System for Legal Research: A Case Study of Precedent Search in Chinese Law. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [66] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *ArXiv abs/1805.10820* (2018).
- [67] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017), 2.

- [68] Hua Guo and David H Laidlaw. 2018. Topic-based exploration and embedded visualizations for research idea generation. *IEEE transactions on visualization and computer graphics* 26, 3 (2018), 1592–1607.
- [69] Tarun Gupta and Danish Pruthi. 2025. All that glitters is not novel: Plagiarism in ai generated research. *arXiv preprint arXiv:2502.16487* (2025).
- [70] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [71] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
- [72] Jeffrey Heer, Matthew Conlen, Vishal Devireddy, Tu Nguyen, and Joshua Horowitz. 2023. Living papers: A language toolkit for augmented scholarly communication. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
- [73] R. Hoffman, S. Mueller, G. Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *ArXiv* abs/1812.04608 (2018).
- [74] Fred Hohman, Andrew Head, R. Caruana, Robert DeLine, and S. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).
- [75] Kim Holmberg and Mike Thelwall. 2014. Disciplinary differences in Twitter scholarly communication. *Scientometrics* 101 (2014), 1027–1042.
- [76] Keith J Holyoak and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.
- [77] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 235–243.
- [78] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. 2023. A computational inflection for scientific discovery. *Commun. ACM* 66, 8 (2023), 62–73.

- [79] Tom Hope, Ronen Tamari, Daniel Hershovich, Hyeonsu B Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2022. Scaling creative inspiration with fine-grained functional aspects of ideas. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [80] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [81] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2020. Local and global explanations of agent behavior: integrating strategy summaries with saliency maps. *arXiv preprint arXiv:2005.08874* (2020).
- [82] Andreas Hundschell, Stefan Razinskas, Julia Backmann, and Martin Hoegl. 2022. The effects of diversity on creativity: A literature review and synthesis. *Applied Psychology* 71, 4 (2022), 1598–1634.
- [83] Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. *arXiv preprint arXiv:2503.22708* (2025).
- [84] David G Jansson and Steven M Smith. 1991. Design fixation. *Design studies* 12, 1 (1991), 3–11.
- [85] Paige Brown Jarreau. 2015. *All the science that is fit to blog: An analysis of science blogging practices*. Louisiana State University and Agricultural & Mechanical College.
- [86] Y. Jin, N. Tintarev, and K. Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. *Proceedings of the 12th ACM Conference on Recommender Systems* (2018).
- [87] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned publishing* 23, 3 (2010), 258–263.
- [88] Mihaela Sabina Jucan and Cornel Nicolae Jucan. 2014. The power of science communication. *Procedia-Social and Behavioral Sciences* 149 (2014), 461–466.
- [89] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An interactive system for personalized thread-based exploration and organization of scientific literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

- [90] Hyeonsu B Kang, David Chuan-En Lin, Nikolas Martelaro, Aniket Kittur, Yan-Ying Chen, and Matthew K Hong. 2024. BioSpark: An End-to-End Generative System for Biological-Analogical Inspirations and Ideation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–13.
- [91] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction* 29, 6 (2022), 1–36.
- [92] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A mixed-initiative system for scholarly synthesis and sense-making. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19.
- [93] Antti Kangasrääsio, D. Glowacka, and Samuel Kaski. 2015. Improving Controllability and Predictability of Interactive Recommendation Interfaces for Exploratory Search. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (2015).
- [94] James C Kaufman and Robert J Sternberg. 2010. *The Cambridge handbook of creativity*. Cambridge University Press.
- [95] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714* (2023).
- [96] Jeong Chul Kim, Sang Guk Suh, Lydia B. Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing. *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023). <https://api.semanticscholar.org/CorpusID:259376400>
- [97] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140* (2023).

- [98] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. 2023. The Semantic Scholar Open Data Platform. *ArXiv* abs/2301.10140 (2023). <https://api.semanticscholar.org/CorpusID:256194545>
- [99] Bart P. Knijnenburg, Svetlin Bostandjiev, J. O’Donovan, and A. Kobsa. 2012. Inspectability and control in social recommenders. In *RecSys ’12*.
- [100] Bart P. Knijnenburg, Niels J. M. Reijmer, and M. C. Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *RecSys ’11*.
- [101] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [102] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys* 55, 8 (2022), 1–35.
- [103] Leon Kopitar, Leona Cilar, Primož Kocbek, and Gregor Stiglic. 2019. Local vs. Global Interpretability of Machine Learning Models in Type 2 Diabetes Mellitus Screening. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*. Springer, 108–119.
- [104] Pigi Kouki, James Schaffer, J. Pujara, J. O’Donovan, and L. Getoor. 2019. Personalized explanations for hybrid recommender systems. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [105] Kudos. [n. d.]. Kudos. <https://info.growkudos.com/landing/researchers-2023>.

- [106] T. Kulesza, S. Stumpf, M. Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *CHI '12*.
- [107] Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. 2022. A search engine for discovery of scientific challenges and directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 11982–11990.
- [108] Vivian Lai, Alison Smith-Renner, Ke Zhang, Ruijia Cheng, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2022. An exploration of post-editing effectiveness in text summarization. *arXiv preprint arXiv:2206.06383* (2022).
- [109] Pier Luca Lanzi and Daniele Loiacono. 2023. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1383–1390.
- [110] B. Lee, Kyle Lo, Doug Downey, and Daniel S. Weld. 2020. Explanation-Based Tuning of Opaque Machine Learners with Application to Paper Recommendation. *ArXiv* abs/2003.04315 (2020).
- [111] Benjamin Charles Germain Lee, Doug Downey, Kyle Lo, and Daniel S Weld. 2023. LIMEADE: From AI Explanations to Advice Taking. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–29.
- [112] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. Paperweaver: Enriching topical paper alerts by contextualizing recommended papers with user-collected papers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [113] Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [114] Tianyi Li, Gregorio Convertino, Ranjeet Kumar Tayi, and Shima Kazerooni. 2019. What data should I protect?: recommender and planning support for

data security analysts. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).

- [115] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. 2024. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* 1, 8 (2024), AIoa2400196.
- [116] Hyunseung Lim, Ji Yong Cho, Taewan Kim, Jeongeon Park, Hyungyu Shin, Seulgi Choi, Sunghyun Park, Kyungjae Lee, Juho Kim, Moontae Lee, et al. 2024. Co-Creating Question-and-Answer Style Articles with Large Language Models for Research Promotion. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 975–994.
- [117] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
- [118] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024. Personaflo: Boosting research ideation with llm-simulated expert personas. *arXiv preprint arXiv:2409.12538* (2024).
- [119] Yiren Liu, Mengxia Yu, Meng Jiang, and Yun Huang. 2023. Creative Research Question Generation for Human-Computer Interaction Research.. In *IUI Workshops*. 58–66.
- [120] Tao Long, Dorothy Zhang, Grace Li, Batool Taraif, Samia Menon, Kynneddy Simone Smith, Sitong Wang, Katy Ilonka Gero, and Lydia B. Chilton. 2023. Tweeterial Hooks: Generative AI Tools to Motivate Science on Social Media. *ArXiv abs/2305.12265* (2023). <https://api.semanticscholar.org/CorpusID:258833277>
- [121] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* (2024).
- [122] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *ArXiv abs/2408.06292* (2024). <https://api.semanticscholar.org/CorpusID:271854887>

- [123] Guang Lu, Sylvia B Larcher, and Tu Tran. 2023. Hybrid Long Document Summarization using C2F-FAR and ChatGPT: A Practical Study. *arXiv preprint arXiv:2306.01169* (2023).
- [124] Zhuoran Lu, Sheshera Mysore, Tara Safavi, Jennifer Neville, Longqi Yang, and Mengting Wan. 2024. Corporate Communication Companion (CCC): An LLM-empowered Writing Assistant for Workplace Social Media. *arXiv preprint arXiv:2405.04656* (2024).
- [125] Jessica GY Luc, Michael A Archer, Rakesh C Arora, Edward M Bender, Arie Blitz, David T Cooke, Tamara Ni Hlci, Biniam Kidane, Maral Ouzounian, Thomas K Varghese Jr, et al. 2021. Does tweeting improve citations? One-year results from the TSSMN prospective randomized trial. *The Annals of thoracic surgery* 111, 1 (2021), 296–300.
- [126] Merja Mahrt and Cornelius Puschmann. 2014. Science blogging: An exploratory study of motives, styles, and audience reactions. *Journal of science Communication* 13, 03 (2014), A05.
- [127] Ashley Mehlenbacher. 2019. *Science communication online: Engaging experts and publics on the internet*. The Ohio State University Press.
- [128] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Ranked List Truncation for Large Language Model-based Re-Ranking. *ArXiv abs/2404.18185* (2024). <https://api.semanticscholar.org/CorpusID:269449617>
- [129] Louie Meyer, Johanne Engel Aaen, Anitamalina Regitse Tranberg, Peter Kun, Matthias Freiberger, Sebastian Risi, and Anders Sundnes Løvlie. 2024. Algorithmic ways of seeing: Using object detection to facilitate art exploration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [130] Martijn Millecamp, Nyi Nyi Htun, C. Conati, and K. Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [131] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models:

- Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.
- [132] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and Evaluating Concept-driven Explanations of Machine Learning Models. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.
- [133] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [134] Francesco Moramarco, Alex Papadopoulos Korfiatis, Aleksandar Savkov, and Ehud Reiter. 2021. A preliminary study on evaluating consultation notes with post-editing. *arXiv preprint arXiv:2104.04402* (2021).
- [135] Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4453–4470.
- [136] Sheshera Mysore, Tim O’Gorman, Andrew McCallum, and Hamed Zamani. [n. d.]. CSFCube—A Test Collection of Computer Science Research Articles for Faceted Query by Example. ([n. d.]).
- [137] Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. Acceleron: A Tool to Accelerate Research Ideation. *arXiv preprint arXiv:2403.04382* (2024).
- [138] Harshit Nigam, Manasi Patwardhan, Lovekesh Vig, and Gautam Shroff. 2024. An Interactive Co-Pilot for Accelerated Research Ideation. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. 60–73.
- [139] Baharan Nouriinanloo and Maxime Lamothe. 2024. Re-Ranking Step by Step: Investigating Pre-Filtering for Re-Ranking with Large Language Models. *ArXiv abs/2406.18740* (2024). <https://api.semanticscholar.org/CorpusID:270764517>
- [140] J. O’Donovan, B. Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *CHI*.

- [141] Jeongseok Oh, Seungju Kim, and Seungjun Kim. 2024. LumiMood: A Creativity Support Tool for Designing the Mood of a 3D Scene. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [142] OpenReview. [n. d.]. OpenReview. <https://openreview.net/>.
- [143] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. 2023. Co-design of human-centered, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–35.
- [144] Denis Parra and P. Brusilovsky. 2015. User-controllable personalization: A case study with SetFusion. *Int. J. Hum. Comput. Stud.* 78 (2015), 43–67.
- [145] Patrick Dunleavy. [n. d.]. LSE. <https://blogs.lse.ac.uk/impactofsocialsciences/2016/01/25/how-to-write-a-blogpost-from-your-journal-article/>.
- [146] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–16.
- [147] Jason Portenoy, Marissa Radensky, Jevin D West, Eric Horvitz, Daniel S Weld, and Tom Hope. 2022. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [148] Kevin Pu, KJ Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. *arXiv preprint arXiv:2410.04025* (2024).
- [149] A Terry Purcell and John S Gero. 1996. Design and other types of fixation. *Design studies* 17, 4 (1996), 363–383.
- [150] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. 2022. Exploring the role of local and global explanations in recommender systems. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.

- [151] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. 2024. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. *arXiv preprint arXiv:2409.14634* (2024).
- [152] Marissa Radensky, Daniel S Weld, Joseph Chee Chang, Pao Siangliulue, and Jonathan Bragg. 2024. Let’s Get to the Point: LLM-Supported Planning, Drafting, and Revising of Research-Paper Blog Posts. *arXiv e-prints* (2024), arXiv–2406.
- [153] Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. *arXiv preprint arXiv:2501.08292* (2025).
- [154] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*. <https://api.semanticscholar.org/CorpusID:201646309>
- [155] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [156] Marco Tulio Ribeiro, Sameer Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
- [157] Mark A Runco et al. 2010. Divergent thinking, creativity, and ideation. *The Cambridge handbook of creativity* 413 (2010), 446.
- [158] Robert Keith Sawyer and Danah Henriksen. 2024. *Explaining creativity: The science of human innovation*. Oxford university press.
- [159] James Schaffer, Tobias Höllerer, and J. O’Donovan. 2015. Hypothetical Recommendation: A Study of Interactive Profile Manipulation Behavior for Recommender Systems. In *FLAIRS Conference*.
- [160] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109* (2024).

- [161] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *ArXiv* abs/2409.04109 (2024). <https://api.semanticscholar.org/CorpusID:272463952>
- [162] Dean Keith Simonton. 2021. Scientific Creativity: Discovery and Invention as Combinatorial. *Frontiers in Psychology* 12 (2021). <https://api.semanticscholar.org/CorpusID:237262181>
- [163] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. A Selective Summary of Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction* (2022). <https://api.semanticscholar.org/CorpusID:247085270>
- [164] Arvind Srinivasan and Joel Chan. 2024. Improving Selection of Analogical Inspirations through Chunking and Recombination. In *Proceedings of the 16th Conference on Creativity & Cognition*. 374–397.
- [165] Susan Leigh Star and James R Griesemer. 1989. Institutional ecology, translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social studies of science* 19, 3 (1989), 387–420.
- [166] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminare: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.
- [167] Lu Sun, Aaron Chan, Yun Seo Chang, and Steven P Dow. 2024. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 120–137.
- [168] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *ArXiv* abs/2304.09542 (2023). <https://api.semanticscholar.org/CorpusID:258212638>
- [169] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing. 14918–14937.

- [170] Yuying Tang, Haotian Li, Minghe Lan, Xiaojuan Ma, and Huamin Qu. 2025. Understanding Screenwriters’ Practices, Attitudes, and Future Expectations in Human-AI Co-Creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [171] P Thagard. 2012. *The cognitive science of science: Explanation, discovery, and conceptual change*. The MIT Press.
- [172] Chun-Hua Tsai and P. Brusilovsky. 2017. Providing Control and Transparency in a Social Recommender System for Academic Conferences. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (2017).
- [173] Chun-Hua Tsai and P. Brusilovsky. 2019. Explaining recommendations in an interactive hybrid social recommender. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
- [174] Chun-Hua Tsai and Peter Brusilovsky. 2020. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* (2020), 1–37.
- [175] UWaterloo. [n. d.]. UWaterloo. <https://uwaterloo.ca/writing-and-communication-centre/writing-academic-blog-posts>.
- [176] J. Vig, S. Sen, and J. Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2 (2012), 13:1–13:44.
- [177] Daril A Vilhena, Jacob G Foster, Martin Rosvall, Jevin D West, James Evans, and Carl T Bergstrom. 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. *Sociological Science* 1 (2014), 221–23.
- [178] Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraaj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. DORIS-MAE: scientific document retrieval using multi-level aspect-based queries. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 38404–38419.

- [179] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024. LLMFactor: Extracting profitable factors through prompts for explainable stock movement prediction. *arXiv preprint arXiv:2406.10811* (2024).
- [180] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Sci-mon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259* (2023).
- [181] Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024. Tutor copilot: A human-ai approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017* (2024).
- [182] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2023. PodReels: Human-AI Co-Creation of Video Podcast Teasers. *arXiv preprint arXiv:2311.05867* (2023).
- [183] Justin D Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection not required? Human-AI partnerships in code translation. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 402–412.
- [184] Spencer Williams, Ridley Jones, Katharina Reinecke, and Gary Hsieh. 2022. An HCI Research Agenda for Online Science Communication. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1 – 22. <https://api.semanticscholar.org/CorpusID:253459983>
- [185] Yujia Xie, Xun Wang, Si-Qing Chen, Wayne Xiong, and Pengcheng He. 2023. Interactive Editing for Text Summarization. *arXiv preprint arXiv:2306.03067* (2023).
- [186] Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. 1342–1351.
- [187] Hongji Yang, Delin Jing, and Lu Zhang. 2016. Creative Computing: an approach to knowledge combination for creativity?. In *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)*. IEEE, 407–414.
- [188] Georgios N Yannakakis, Antonios Liapis, and Constantine Alexopoulos. 2014. Mixed-initiative co-creativity. (2014).

- [189] Fangyu Yu, Peng Zhang, Xianghua Ding, Tun Lu, and Ning Gu. 2024. BNoteHelper: a note-based outline generation tool for structured learning on video-sharing platforms. *ACM Transactions on the Web* 18, 2 (2024), 1–30.
- [190] Mert Yuksekogonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic "Differentiation" via Text. *arXiv preprint arXiv:2406.07496* (2024).
- [191] Marzieh Zarinbal, Azadeh Mohebi, Hesamoddin Mosalli, Razieh Haratinik, Zahra Jabalameli, and Farnoush Bayatmakou. 2019. A new social robot for interactive query-based summarization: Scientific document summarization. In *Interactive Collaborative Robotics: 4th International Conference, ICR 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 4*. Springer, 330–340.
- [192] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [193] Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267* (2024).
- [194] Yi Zhang, Dingding Wang, and Tao Li. 2011. iDVS: an interactive multi-document visual summarization system. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer, 569–584.
- [195] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–30.
- [196] Hang Zou and Ken Hyland. 2019. Reworking research: Interactions in academic articles and blogs. *Discourse studies* 21, 6 (2019), 713–733.

## Appendix A

### Papers-to-Posts

#### A.1 Exploratory Analysis of Perceived Control and Ownership

We analyzed participants' Likert-type responses to questions related to control and ownership (Figure A.1).

##### *A.1.1 Lab Study*

The median participant did not indicate higher perceived control (treatment:  $M=5.00$ ,  $Q1=4.50$ ,  $Q3=6.00$ ; baseline:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=6.00$ ) or satisfaction with their control (treatment:  $M=5.00$ ,  $Q1=2.75$ ,  $Q3=6.00$ ; baseline:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=5.25$ ) under the treatment condition (Figure A.1a-b). However, if we look within-subjects, 11 participants had higher perceived control with PAPERS-TO-POSTS than the baseline tool (equal: 2, lower: 7), and 10 participants had higher satisfaction with their control using PAPERS-TO-POSTS (equal: 4, lower: 6). On the other hand, participants did not demonstrate increased perceived ownership (treatment:  $M=6.00$ ,  $Q1=5.75$ ,  $Q3=6.00$ ; baseline:  $M=6.00$ ,  $Q1=5.00$ ,  $Q3=6.00$ ) or satisfaction with ownership (treatment:  $M=6.00$ ,  $Q1=4.00$ ,  $Q3=6.25$ ; baseline:  $M=6.00$ ,  $Q1=5.00$ ,  $Q3=6.00$ ) under the treatment condition (Figure A.1c-d). Although they felt that they had more control over PAPERS-TO-POSTS, participants may not necessarily have felt that the tool contributed more or less to the blog post's creation.

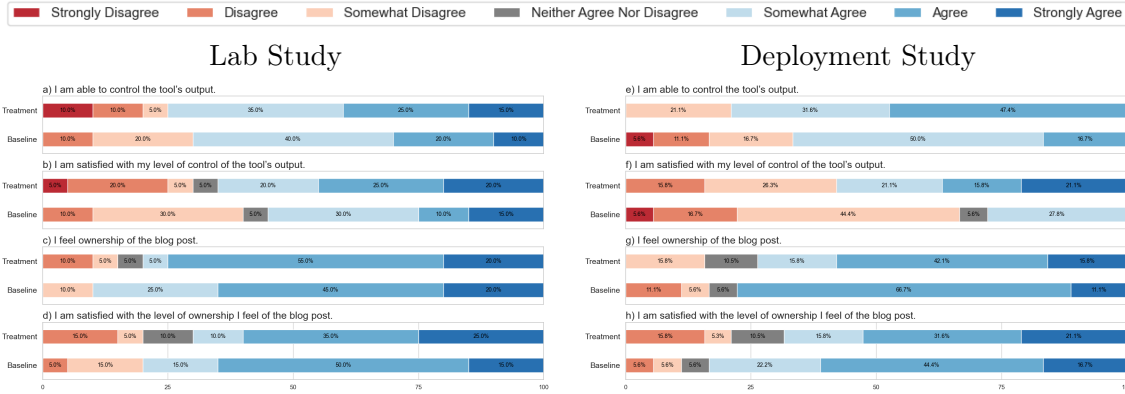


Figure A.1: Survey responses to 7-point Likert-type questions regarding perceived control and ownership in the a-d) lab study and e-h) deployment study. Responses are shown for both the treatment and baseline conditions.

### A.1.2 Deployment Study

As shown in Figure A.1e-h, participants indicated higher median satisfaction with control in the treatment condition (treatment:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=6.00$ ; baseline:  $M=3.00$ ,  $Q1=3.00$ ,  $Q3=4.75$ ), but they showed no difference in median perceived control (treatment:  $M=5.00$ ,  $Q1=5.00$ ,  $Q3=6.00$ ; baseline:  $M=5.00$ ,  $Q1=3.00$ ,  $Q3=5.00$ ), perceived ownership (treatment:  $M=6.00$ ,  $Q1=4.50$ ,  $Q3=6.00$ ; baseline:  $M=6.00$ ,  $Q1=6.00$ ,  $Q3=6.00$ ) or satisfaction with ownership (treatment:  $M=6.00$ ,  $Q1=4.00$ ,  $Q3=6.00$ ; baseline:  $M=6.00$ ,  $Q1=5.00$ ,  $Q3=6.00$ ) in the treatment condition.

### A.1.3 Summary

While lab study participants had higher median perceived control, deployment study participants did not. However, we saw some evidence in both studies of higher

perceived satisfaction with control under the treatment condition. Participants may have been more satisfied with their control without perceiving a higher control in the sense that they were more comfortable with letting the tool have a certain amount of control in exchange for more editing power. Across the two studies, we did not see a change in median perceived ownership or satisfaction with ownership.

## A.2 Prompts

### A.2.1 Prompt for Initial LLM Draft for Baseline Tool

```
def prompt(paper,title):
    return f"""\n\nYou are writing a blogpost for your research paper titled "{title}."
The 4 blogpost sections are: introduction, methods, results, and conclusion.
Each section will be between 125 and 250 words and can have anywhere from 1 to 3 paragraphs.

Write the blogpost based on the paper below.
Do not write anything that is not supported by the paper.
PAPER:
{paper}

Follow the guidelines below.
GENERAL GUIDELINES:
- Audience: Your audience will be people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.

BLOGPOST WITH 4 SECTIONS (INTRODUCTION, METHODS, RESULTS, AND CONCLUSION), BETWEEN 125 AND 250 WORDS (BETWEEN 1 AND 3 PARAGRAPHS) PER SECTION:
"""\n\n
```

Figure A.2: Prompt for initial LLM draft for baseline tool.

### A.2.2 Prompts for the Warm Start Step of PAPERS-TO-POSTS

Figures [A.3](#)–[A.5](#) show the prompts for the initial LLM draft for PAPERS-TO-POSTS.

```

def generate_bulletpoints_for_long_paragraph(para):
    return f"""Summarize the text below in 3 bulletpoints for a powerpoint slide.
    The bulletpoints should not be hierarchical.
    The bulletpoints should be in html format. Do not write anything that is not supported by the text below.

    FORMAT TO USE
    3 bulletpoints:
    <ul>
    <li>bulletpoint 1</li>
    <li>bulletpoint 2</li>
    <li>bulletpoint 3</li>
    </ul>

    TEXT TO SUMMARIZE
    {para}\n

    3 bulletpoints:"""

def generate_bulletpoints_for_med_paragraph(para):
    return f"""Summarize the text below in 2 bulletpoints for a powerpoint slide.
    The bulletpoints should not be hierarchical.
    The bulletpoints should be in html format.
    Do not write anything that is not supported by the text below.

    FORMAT TO USE
    2 bulletpoints:
    <ul>
    <li>bulletpoint 1</li>
    <li>bulletpoint 2</li>
    </ul>

    TEXT TO SUMMARIZE
    {para}\n

    2 bulletpoints:"""

def generate_bulletpoints_for_short_paragraph(para):
    return f"""Summarize the text below in 1 bulletpoint for a powerpoint slide.
    The bulletpoint should be in html format.
    Do not write anything that is not supported by the text below.

    FORMAT TO USE
    1 bulletpoint:
    <ul>
    <li>bulletpoint 1</li>
    </ul>

    TEXT TO SUMMARIZE
    {para}\n

    1 bulletpoint:"""

```

Figure A.3: Prompts for generating bullet points for each paragraph in full paper being summarized. Each prompt corresponds to different length paragraphs. The top prompt generates three bullet points for paragraphs of more than 100 words, the second generates two bullet points for paragraphs of between 51 and 100 words, and the third generates one bullet point for paragraphs less than 51 words long.

```
def generate_top_bulletpoints(bulletpoints,title):
    return f"""BULLETPPOINTS:
{bulletpoints}

Select the top 10 bulletpoints above that would be useful to include in 4 blogpost sections for the paper "{title}."
The 4 blogpost sections are: introduction, methods, results, and conclusion.
Your answer should simply be a list of the relevant bulletpoint IDs for each section.
The bulletpoints for each section can overlap.

EXAMPLE ANSWER:
Introduction: [#X,#Y,#Z,#A,#B,#C,#D,#E,#F,#G]
Methods: [#X,#Y,#Z,#A,#B,#C,#D,#E,#F,#G]
Results: [#X,#Y,#Z,#A,#B,#C,#D,#E,#F,#G]
Conclusion: [#X,#Y,#Z,#A,#B,#C,#D,#E,#F,#G]

YOUR ANSWER:
Introduction: """"
```

Figure A.4: Prompt to select top 10 bullet points relevant for each of the initial draft's blog post sections (introduction, methods, results, conclusion).

```

def generate_blogpost_section_from_bulletpoints(text,header,title,guidelines,blogpostSoFar):
    return f"""You are writing a blogpost for your research paper titled "{title}."
The 4 blogpost sections are: introduction, methods, results, and conclusion.
Right now, you are working on the {header} section.
The section will be between 125 and 250 words. It can have anywhere from 1 to 3 paragraphs.

Write the {header} section using the paragraphs and associated bulletpoints below.
For each paragraph below, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints below.
PARAGRAPHS AND BULLETPPOINTS:
{text}

Follow the guidelines below.
GENERAL GUIDELINES:
- Audience: Your audience will be people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{header} GUIDELINES:
{guidelines}

BLOGPOST SO FAR:
{blogpostSoFar}

{header} SECTION WITH BETWEEN 125 AND 250 WORDS (BETWEEN 1 AND 3 PARAGRAPHS):"""

```

Figure A.5: Prompt to generate each of the initial draft’s blog post sections (introduction, methods, results, conclusion) using the 10 relevant selected bullet points. The section-specific guidelines are as follows. Introduction: “-Present hook (e.g., interesting fact, quote, promise of change in knowledge, illustrating example of the topic). If context allows, visual or sensory elements are helpful anchors.-Provide high-level description of problem being solved.-Explain why work is interesting and a solution to the problem matters.-Do not repeat information from prior blogpost sections.” Methods: “-Focus on methods and do NOT discuss results.-Do not repeat information from prior blogpost sections.” Results: “-State key takeaway.-Discuss up to 3 most interesting aspects of work.-Do not repeat information from prior blogpost sections.” Conclusion: “-Restate key takeaway in new way.-Present future work ideas. [optional]-Loop back to hook. [optional]“

### A.2.3 Prompts for Drafting Step of PAPERS-TO-POSTS

```
def generate_top_bulletpoints(blogpostSoFar,section,bulletpoints):
    return f"""BULLETPPOINTS:
{bulletpoints}

You are writing a blogpost for your research paper.
Right now, you are working on a new section of your blogpost.
Select the top 10 bulletpoints above that would be useful to include in the new blogpost section, which is titled {section}.
Your answer should simply be a list of the relevant bulletpoint IDs for each section.
The bulletpoints for each section can overlap.
{blogpostSoFar}
EXAMPLE ANSWER:
[0.0,2.1,2.2,5.2,7.3,8.2,8.3,9.2,10.1,10.2]

YOUR ANSWER:
"""
```

Figure A.6: Prompt for selecting bullet points for generating a new section based on the section header provided by the user.

```
def generate_blogpost_section_from_bulletpoints(text,instructions,blogpostSoFar,genLength):
    return f"""PARAGRAPHS AND BULLETPPOINTS:
{text}

You are writing a blogpost for your research paper.
Right now, you are working on a new section of your blogpost.
Write the section using the paragraphs and associated bulletpoints above.
The section should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.
The section should be {genLength} in length.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GENERAL GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
NEXT SECTION CONTINUING BLOGPOST ({genLength}):"""
```

Figure A.7: Prompt for generating text for a blog post section when there are selected paragraphs or bullet points as well as custom bullet points, custom instructions, or starting text.

```

def generate_blogpost_section_from_bulletpoints_no_inst(text, blogpostSoFar, genLength):
    return f"""PARAGRAPHS AND BULLETPOINTS:
{text}

You are writing a blogpost for your research paper.
Right now, you are working on a new section of your blogpost.
Write the section using the paragraphs and associated bulletpoints above.
The section should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.
The section should be {genLength} in length.

Follow the guidelines below.
GENERAL GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
NEXT SECTION CONTINUING BLOGPOST ({genLength}):"""

```

Figure A.8: Prompt for generating text for a blog post section when there are selected paragraphs or bullet points but no custom bullet points, custom instructions, or starting text.

```

def generate_blogpost_section_from_bulletpoints_no_text(instructions, blogpostSoFar, genLength):
    return f"""You are writing a blogpost for your research paper.
Right now, you are working on a new section of your blogpost.
Write the section.
The section should be {genLength} in length.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GENERAL GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
NEXT SECTION CONTINUING BLOGPOST ({genLength}):"""

```

Figure A.9: Prompt for generating text for a blog post section when there are no selected paragraphs or bullet points, but there are custom bullet points, custom instructions, or starting text.

### A.2.4 Prompts for Revising Step of PAPERS-TO-POSTS

```
def generate_more_no_inst(text, blogpostSoFar, generation, modLength):
    return f"""PARAGRAPHS AND BULLETPPOINTS:
    {text}

Using the paragraphs and bulletpoints above, expand the text below to be {modLength}.
The expanded text should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}

EXPANDED TEXT CONTINUING BLOGPOST ({modLength}):"""
```

Figure A.10: Prompt for the “expand” modification when there are no custom modification instructions. The automatic modLength variable for this modification is “twice the length that it currently is.”

```

def generate_more_no_text(instructions, blogpostSoFar, generation, modLength):
    return f"""Expand the text below to be {modLength}.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}

EXPANDED TEXT CONTINUING BLOGPOST ({modLength}):"""

def generate_more_no_inst(text, blogpostSoFar, generation, modLength):
    return f"""PARAGRAPHS AND BULLETPPOINTS:
{text}

```

Figure A.11: Prompt for the “expand” modification when there are custom modification instructions but no selected paragraphs or bullet points of which to be aware. The automatic modLength variable for this modification is “twice the length that it currently is.”

```

def generate_more(text,instructions,blogpostSoFar,generation,modLength):
    return f"""PARAGRAPHS AND BULLETPPOINTS:
{text}

Using the paragraphs and bulletpoints above, expand the text below to be {modLength}.
The expanded text should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}

EXPANDED TEXT CONTINUING BLOGPOST ({modLength}):"""

```

Figure A.12: Prompt for the “expand” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “twice the length that it currently is.”

```
def generate_less(text,instructions,blogpostSoFar,generation,modLength):
    return f"""PARAGRAPHS AND BULLETPOINTS:
    {text}

Using the paragraphs and bulletpoints above, condense the text below to be {modLength}.
The condensed text should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
    - Be clear and concise.
    - Avoid jargon, and if you must use it, explain it.
    - Explain any acronyms/initials/formulas.
    - Write in an active rather than passive voice.
    - Avoid slang.
    - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}
```

Figure A.13: Prompt for the “condense” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “half the length that it currently is.” The other prompts for this modification mirror those for “expand.”

```
def generate_simple(text, instructions, blogpostSoFar, generation, modLength):
    return f"""PARAGRAPHS AND BULLETPOINTS:
    {text}

    Using the paragraphs and bulletpoints above, rewrite the text below to be more understandable to a layperson than it currently is.
    The rewritten text should cover information from each paragraph above.
    If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
    Do not write anything that is not supported by the paragraphs and bulletpoints above.
    The text should be {modLength}.

    Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
    {instructions}

    Follow the guidelines below.
    GUIDELINES:
    - Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
    - Language:
      - Be clear and concise.
      - Avoid jargon, and if you must use it, explain it.
      - Explain any acronyms/initials/formulas.
      - Write in an active rather than passive voice.
      - Avoid slang.
      - Avoid sensationalism and be objective.
    {blogpostSoFar}
    ORIGINAL TEXT CONTINUING BLOGPOST:
    {generation}

    REWRITTEN TEXT CONTINUING BLOGPOST WITH LANGUAGE MORE UNDERSTANDABLE TO LAYPERSON:"""
```

Figure A.14: Prompt for the “simpler terms” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for the this modification mirror those for “expand.”

```
def generate_dramatic(text,instructions,blogpostSoFar,generation,modLength):
    return f"""PARAGRAPHS AND BULLETPOINTS:
{text}

Using the paragraphs and bulletpoints above, rewrite the text below to better capture readers' attention. Do not use much more difficult vocabulary.
The rewritten text should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.
The text should be {modLength}.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}

REWRITTEN TEXT CONTINUING BLOGPOST WITH LANGUAGE TO BETTER CAPTURE READERS' ATTENTION:"""
```

Figure A.15: Prompt for the “more dramatic” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for the this modification mirror those for “expand.”

```
def generate_undramatic(text,instructions,blogpostSoFar,generation,modLength):
    return f"""PARAGRAPHS AND BULLETPOINTS:
{text}

Using the paragraphs and bulletpoints above, rewrite the text below to be less dramatic than it currently is.
The rewritten text should cover information from each paragraph above.
If a paragraph has associated bulletpoints, focus on information in the paragraph that is relevant to its associated bulletpoints.
Do not write anything that is not supported by the paragraphs and bulletpoints above.
The text should be {modLength}.

Follow the instructions below. If the instructions below contradict what is said above, still follow the instructions below.
{instructions}

Follow the guidelines below.
GUIDELINES:
- Audience: Your audience is people in your broad area of research (e.g., computer science, psychology).
- Language:
  - Be clear and concise.
  - Avoid jargon, and if you must use it, explain it.
  - Explain any acronyms/initials/formulas.
  - Write in an active rather than passive voice.
  - Avoid slang.
  - Avoid sensationalism and be objective.
{blogpostSoFar}
ORIGINAL TEXT CONTINUING BLOGPOST:
{generation}

REWRITTEN TEXT CONTINUING BLOGPOST WITH LESS DRAMATIC LANGUAGE:"""
```

Figure A.16: Prompt for the “less dramatic” modification when there are selected paragraphs or bullet points of which to be aware and custom modification instructions. The automatic modLength variable for this modification is “about the same length that it currently is (no more than 25 words longer or shorter).” The other prompts for this modification mirror those for “expand.”

### A.3 Additional Log Analysis Plots

Figures A.17 and A.18 show additional log analysis plots for the lab study, while Figure A.19 shows additional log analysis plots for the deployment study.

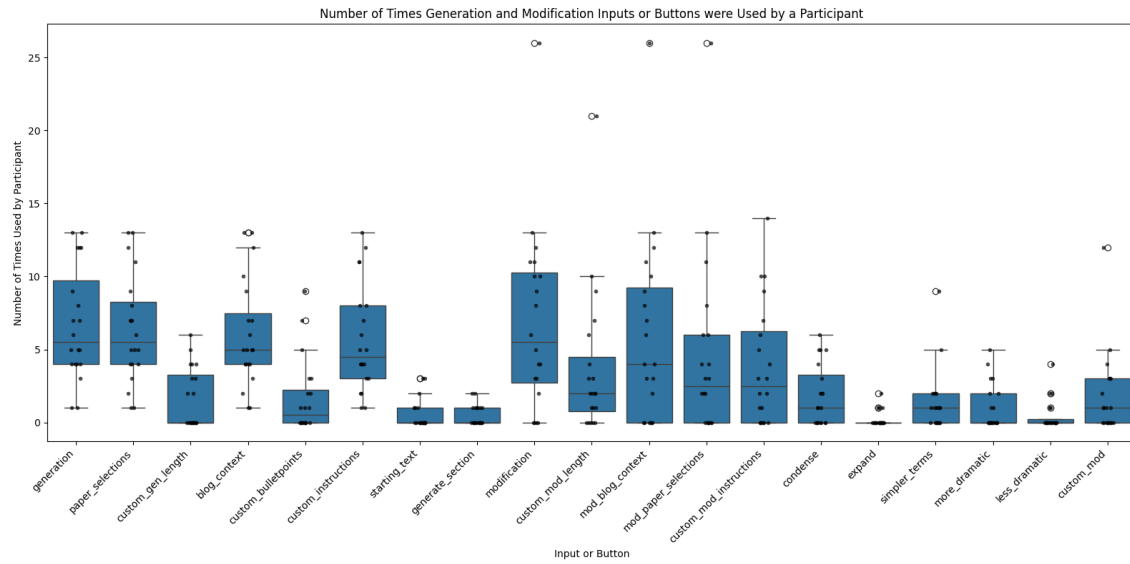
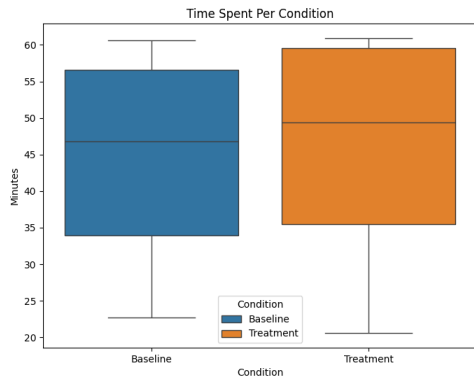
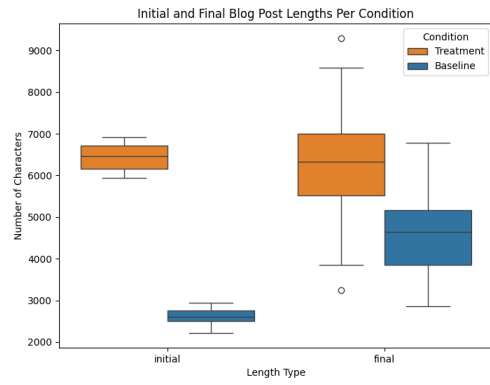


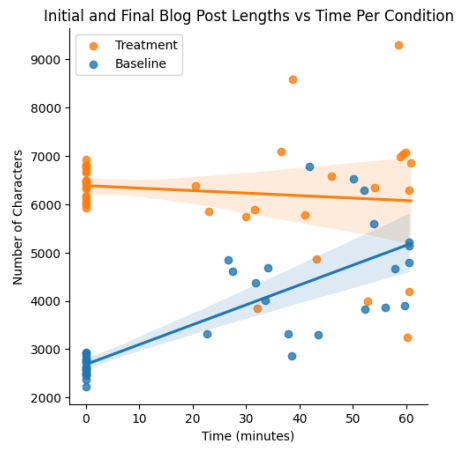
Figure A.17: Number of times each input or button related to generating and modifying text was used by each participant in lab study. From left to right: the act of generating text, inputs/buttons related to generating text, the act of modifying text, inputs/buttons related to modifying text.



(a)

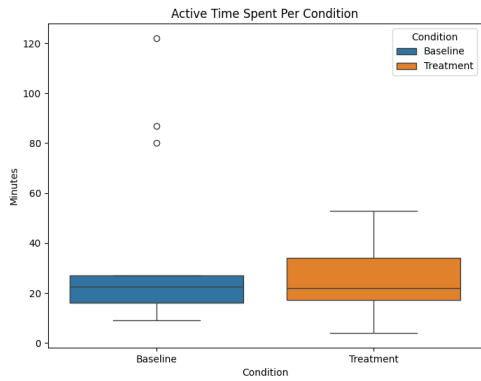


(b)

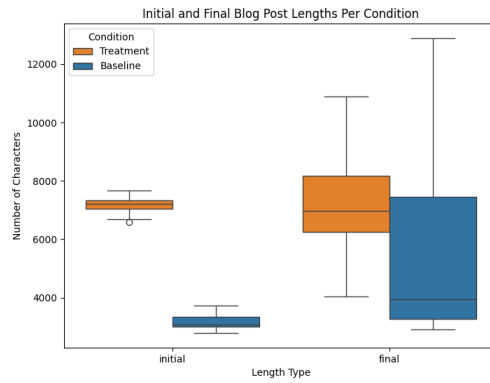


(c)

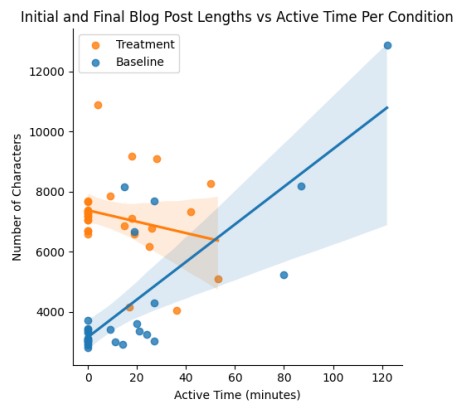
Figure A.18: Additional plots for lab study.



(a)



(b)



(c)

Figure A.19: Additional plots for deployment study.

## A.4 Lab Study Version of Tool

Figure A.20 shows how the tool looked for the lab study.

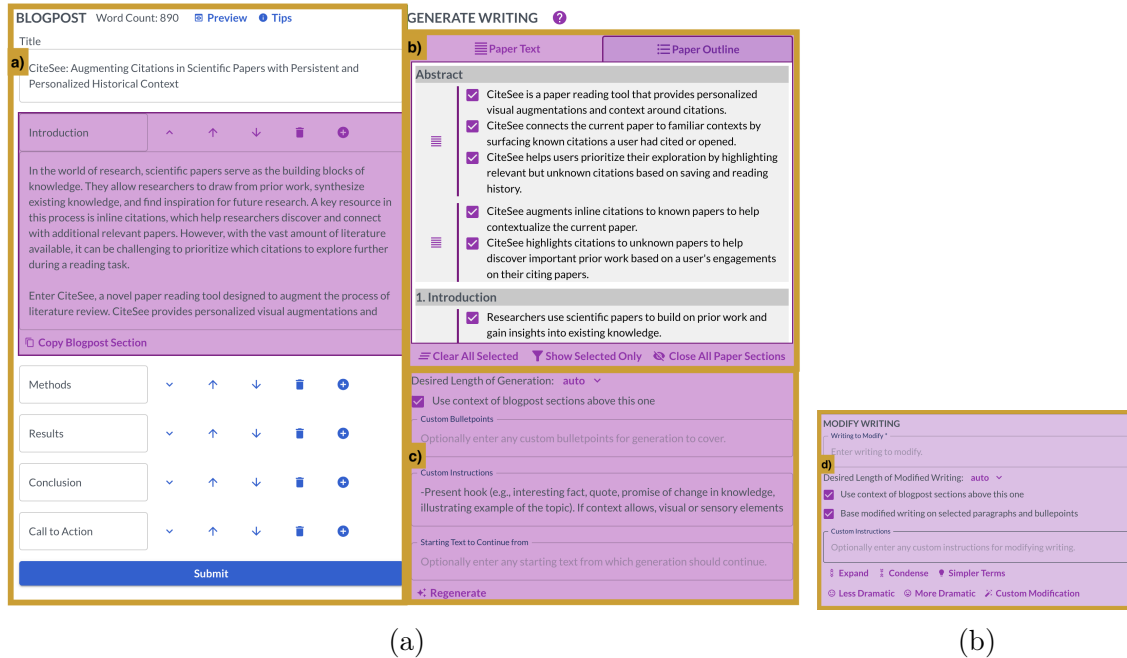


Figure A.20: PAPERS-TO-POSTS in the lab study. a) Blog-Post Area, b) Planning-Support Area, c) Drafting-Support Area, and d) Revising-Support Area.

## A.5 Example Baseline Initial Draft

Below is example output from the baseline tool for the paper “ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries” [192]. Note that this output is for the deployment study version of PAPERS-TO-POSTS, which used automatic paper parsing without manual fixes. This example was generated in October 2024.

### A.5.1 Introduction

*In the realm of academic research, distilling the essence of long, multi-topic documents into concise summaries is a formidable challenge. This task becomes even more daunting when considering the diverse backgrounds of readers who may seek to understand these documents. Traditional summarization techniques, both extractive and abstractive, often fall short in addressing the nuances of such complex texts. Recognizing this gap, our research introduces ConceptEVA, a novel system designed to revolutionize the way we approach the summarization of academic papers. ConceptEVA leverages the latest advancements in natural language processing (NLP) and artificial intelligence (AI) to offer a mixed-initiative approach, enabling users to generate, evaluate, and customize summaries tailored to their specific interests. This system not only promises to enhance the accessibility of academic literature but also to enrich the reader’s engagement with the material.*

### A.5.2 Methods

*ConceptEVA is built upon a multi-task Longformer Encoder Decoder (LED), fine-tuned to adeptly handle the summarization of lengthy, multi-domain documents. This innovative system employs attention mechanisms to efficiently process extensive texts while maintaining high information fidelity. A key feature of ConceptEVA is its interactive visualization of document concepts, presented as a network that illustrates semantic relatedness and co-occurrence. Users can interact with this visualization to select concepts of interest, which ConceptEVA then uses to automatically update the summary, emphasizing the chosen concepts. Our evaluation of ConceptEVA in-*

*volved two iterations: an initial expert review and a subsequent within-subjects study, comparing ConceptEVA's performance against manually-generated summaries.*

### *A.5.3 Results*

*The findings from our study indicate a clear preference among participants for ConceptEVA's customized summaries over their manually-generated counterparts. Participants appreciated the ability to tailor summaries to their interests, resulting in higher satisfaction levels with the summaries produced by ConceptEVA. However, incorporating critique into summaries remained a challenge, highlighting an area for future improvement. Despite this, the overall response to ConceptEVA was positive, with users finding the system helpful for exploring and verifying concepts within documents, especially those outside their domain of expertise.*

### *A.5.4 Conclusion*

*ConceptEVA represents a significant step forward in the field of document summarization, particularly for academic literature that spans multiple topics and domains. By integrating interactive visual analysis with advanced NLP techniques, ConceptEVA offers a user-friendly platform for generating, evaluating, and customizing summaries. This mixed-initiative approach not only improves the summarization process but also enhances the user's engagement with the document. While challenges remain, particularly in integrating critique into summaries, the positive feedback from our study participants underscores ConceptEVA's potential as a valuable tool for researchers, educators, and anyone seeking to navigate the complexities of academic*

*literature. Future developments will focus on refining ConceptEVA’s capabilities, further bridging the gap between comprehensive document analysis and user-centric summarization.*

## **A.6 Example PAPERS-TO-POSTS Output**

Below is example output from PAPERS-TO-POSTS for the paper “ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries” [192]. Note that this output is for the deployment study version of PAPERS-TO-POSTS, which used automatic paper parsing without manual fixes. These examples were generated in October 2024.

### *A.6.1 Initial Draft*

#### Introduction

*In the rapidly evolving field of natural language processing and artificial intelligence, the challenge of effectively summarizing long, complex documents that span multiple topics remains a significant hurdle. This is particularly true for academic papers, which often contain dense, multi-disciplinary content that can be difficult to distill into a concise summary. Traditional methods of extractive summarization, which rely on pulling key sentences directly from the text, fall short in capturing the nuanced essence of such documents. The advent of abstractive summarization, which paraphrases the original text to produce a more human-like summary, offers a promising solution but is still plagued by challenges, especially when dealing with documents*

that cover a broad range of knowledge domains.

Enter *ConceptEVA*, a novel mixed-initiative approach designed to tackle the complexities of summarizing long, multi-topic academic documents. By integrating advanced machine learning techniques with interactive visual analysis, *ConceptEVA* allows users to generate, evaluate, and customize summaries in a way that is both efficient and tailored to their specific interests. This system represents a significant leap forward in automated text summarization, offering a more dynamic and user-centric solution to the problem of digesting extensive academic literature.

The significance of *ConceptEVA* lies not only in its ability to produce concise summaries of complex documents but also in its interactive nature, which empowers users to steer the summarization process. This human-in-the-loop approach ensures that the generated summaries are relevant and valuable to the reader, addressing the long-standing challenge of creating universally applicable summaries for documents that traverse multiple domains. As we continue to push the boundaries of what's possible in automated text summarization, *ConceptEVA* stands out as a beacon of innovation, promising to transform how we engage with and understand academic literature.

## Methods

In developing *ConceptEVA*, we employed a multi-faceted approach to address the challenges of summarizing long, multi-topic documents. At the core of *ConceptEVA* is a custom multi-task Longformer encoder-decoder, specifically designed for the summarization of extensive documents. This model leverages the power of attention mech-

anisms from transformers to efficiently manage memory usage while ensuring high fidelity in information preservation across longer texts. The attention mechanisms operate at both local and global levels, facilitating a nuanced understanding and processing of the document's content.

To enhance user interaction and customization, *ConceptEVA* integrates interactive visualizations of document concepts as a network. This visualization reflects semantic relatedness and co-occurrence, allowing users to focus on and select concepts of interest. These selected concepts can then automatically update the summary to emphasize them, supported by a "focus-on" function that surfaces and highlights related concepts. This interactive component is underpinned by a multi-domain ontology for concept identification and a force-directed layout for visual representation, ensuring a user-friendly interface for customizing summaries.

The iterative development of *ConceptEVA* was informed by an initial survey of eight research practitioners and underwent two stages of refinement based on expert reviews and a within-subjects study. This process included fine-tuning a pre-trained Longformer encoder-decoder for scientific document summarization, incorporating paraphrasing and semantic sentence embedding techniques, and developing an interactive summary editor to facilitate better human-AI collaboration. Additionally, we utilized *Faiss*, a similarity search library for dense vectors, to enhance the summarizer's context understanding, thereby improving the quality and relevance of the generated summaries.

## Results

*Our study revealed that ConceptEVA significantly enhances user satisfaction with the summarization process, particularly when compared to manually-generated summaries. Specifically, 8 out of 12 participants preferred the summaries generated using ConceptEVA's human-in-the-loop approach over the baseline, indicating a clear preference for the customized summaries provided by our system. This preference was especially pronounced among participants reviewing documents outside their domain of expertise, underscoring ConceptEVA's utility in helping users explore unfamiliar topics through efficient and focused summarization.*

*Moreover, the integration of ConceptEVA's "focus-on" function, which allows users to select and emphasize specific concepts of interest, was instrumental in achieving higher satisfaction levels. This feature not only facilitated a more interactive and personalized summarization experience but also highlighted the system's capability to adapt summaries based on user input. However, it was noted that participants with less experience in the subject matter found it challenging to incorporate critique, such as limitations and implications, into their summaries, suggesting an area for further improvement in guiding users through more nuanced aspects of summarization.*

*A significant association was found between the summarization approach employed by ConceptEVA and the summary satisfaction rating, as evidenced by a chi-squared test of independence. This statistical validation, along with user feedback, underscores the effectiveness of ConceptEVA in enhancing the summarization process through mixed-initiative interactions and customized content, paving the way for future developments in automated text summarization technologies.*

## Conclusion

*In conclusion, ConceptEVA represents a significant advancement in the field of automated text summarization, particularly for academic documents. By leveraging a human-in-the-loop approach, ConceptEVA not only enhances the relevance and quality of summaries but also fosters a more collaborative interaction between users and AI. This system's ability to integrate interactive visual analysis with natural language processing techniques allows for a more nuanced and customizable summarization process, addressing the critical need for summaries that cater to individual user interests and requirements. The positive feedback from our user study, especially regarding the system's content-specific assistance and the preference for ConceptEVA-generated summaries over manual efforts, underscores the potential of this approach to significantly improve how we engage with and digest complex academic literature.*

*Looking forward, we aim to expand ConceptEVA's capabilities by testing it on a broader range of documents and incorporating additional features based on user feedback. This will include enhancing the system's flexibility to adapt summaries more dynamically to user priorities and further refining the AI's ability to maintain logical and narrative coherence throughout the summary. By continuing to focus on the synergy between human expertise and AI efficiency, ConceptEVA is poised to set new standards in the automation of document summarization, making it an invaluable tool for researchers, scholars, and anyone looking to efficiently extract insights from extensive academic writings.*

### *A.6.2 Initial Draft Bullet Point Selection*

#### Introduction Section's Selected Bullet Points

- *ConceptEVA is a mixed-initiative approach for summarizing long and multi-topic documents.*
- *CHI 2023 conference in Hamburg, Germany featured a multi-disciplinary research paper with an auto-generated summary*
- *Automated text summarization aims to present key concepts in a concise manner and reduce repetition.*
- *Challenges in abstractive summarization include summarizing long, complex documents spanning multiple knowledge domains.*
- *A human-in-the-loop approach is needed for interactive generation of relevant summaries.*
- *ConceptEVA integrates visual analysis and NLP for evaluating and customizing long document summaries.*
- *ConceptEVA introduces a human-in-the loop, mixed-initiative approach to evaluate and customize document summary generation*
- *Our work focuses on interactively exploring concept focus in generated summaries and emphasizing concepts of interest in academic publications.*
- *ConceptEVA customizes summaries by updating them with user-selected concepts of interest.*

- *ConceptEVA is an interactive document summarization system for long, multi-domain documents.*

## Methods Section's Selected Bullet Points

- *It incorporates a custom multi-task longformer encoder decoder for summarizing longer documents.*
- *Interactive visualizations of document concepts as a network in ConceptEVA help users focus on concepts of interest*
- *Attention mechanisms from transformers used at local levels to reduce memory usage*
- *ConceptEVA design informed by initial survey of eight research practitioners and refined through two stages of development and evaluation*
- *ConceptEVA includes fine-tuning an LED model, identifying concepts using an ontology, and providing interactive visualization for customization.*
- *Maintain the user's mental map of the original document by preserving its layout*
- *We use Faiss to implement the approach, a similarity search library for dense vectors in large scale.*
- *Extended ConceptEVA with an interactive summary editor for better human-AI collaboration*

- *We constructed a set of guidelines for participants to follow when generating a summary manually or using ConceptEVA.*
- *The second iteration includes a longformer encoder-decoder pre-trained for scientific documents, fine-tuned for paraphrasing and sentence embedding, and concepts visualized using a force-directed network.*

## Results Section's Selected Bullet Points

- *Participants' satisfaction with ConceptEVA customized summaries is higher than their own manually-generated summary*
- *ConceptEVA is helpful for examining and verifying ideas*
- *ConceptEVA is more useful for participants outside their domain of interest*
- *Inexperienced participants found it difficult to incorporate critique into the summary*
- *Text embeddings are pre-computed for each sentence in the source document and retrieved when users select concepts.*
- *8 out of 12 participants rated the summary generated using ConceptEVA's human-in-the-loop approach (task T3) higher than the baseline*
- *Study used 6 participants to summarize ConceptScope and 6 participants to summarize BodyVis*

- *A significant association was found between summarization approach and summary satisfaction rating*
- *Higher rating for human-in-the-loop approach led to more efficient location and focus on concepts*
- *Participants preferred summaries created through ConceptEVA's human-in-the-loop approach*

### Conclusion Section's Selected Bullet Points

- *ConceptEVA is a mixed-initiative system for academic document readers and writers to generate, evaluate, and customize automated summaries.*
- *ConceptEVA integrates visual analysis and NLP for evaluating and customizing long document summaries.*
- *ConceptEVA's approach better exploits human-AI collaboration and iteratively improves the summary*
- *Saliency, fluency, novelty, and coherence are key factors in evaluating automated summarization techniques*
- *The technique should be flexible enough to change the summaries based on the priority of the user.*
- *Research shows that regular exercise can improve mental health by reducing symptoms of depression and anxiety.*

- *The AI-generated summarization approach prioritizes content over the flow of the summary, compromising logical and narrative connections between sentences*
- *Participants examining their own summaries can reveal insights into the effectiveness of the human-in-the-loop approach.*
- *We will eliminate the limitation by testing ConceptEVA on more diverse papers in the future*
- *ConceptEVA was found helpful for content-specific aspects of summarization*

### *A.6.3 Example Modifications*

We present modifications of the first sentence in the initial draft, using the preset modification macros and default settings for the LLM instructions (except we toggled off “Check if want modified writing to be based on selected paragraphs and bullet points”).

#### Original Sentence

*In the rapidly evolving field of natural language processing and artificial intelligence, the challenge of effectively summarizing long, complex documents that span multiple topics remains a significant hurdle.*

## Input Text to be Modified

*In the rapidly evolving field of natural language processing and artificial intelligence, the challenge of effectively summarizing long, complex documents that span multiple topics remains a significant hurdle.*

## Expanded Version

*In the rapidly evolving field of natural language processing (NLP) and artificial intelligence (AI), the challenge of effectively summarizing long, complex documents that span multiple topics remains a significant hurdle. This is particularly true for academic papers, which often contain dense, multi-disciplinary content that can be difficult to distill into a concise summary. Traditional summarization techniques, which may work well for shorter, single-topic documents, often fall short when applied to these more complex texts.*

## Condensed Version

*In the fast-paced domains of natural language processing and artificial intelligence, summarizing lengthy, multi-topic documents is a major challenge.*

## Simpler-Terms Version

*In the fast-changing world of computer programs that understand and use human language, finding a way to briefly explain long, complicated texts that cover many subjects is still a big challenge.*

## Less Dramatic Version

*In the field of natural language processing and artificial intelligence, summarizing lengthy, multi-topic documents is a notable challenge.*

## More Dramatic Version

*Title: Tackling the Complexity of Summarizing Multi-Topic Documents in AI*

*In the dynamic world of AI and natural language processing, we're constantly facing the tough challenge of summarizing lengthy, intricate documents covering various topics.*

## Appendix B

### Scideator

#### B.1 Prompts for Analogous Paper Facet Finder

##### *B.1.1 Prompt to extract facets from a paper title/abstract.*

```
def promptTextToPurposeMechanism(papers, corpus_ids, text=None):
    if text:
        prompt = f"""
        {text}"""
    else:
        prompt = f"""
    TEXTS: """
        for id in range(0, len(corpus_ids)):
            prompt += f"""
            Text { id + 1}
            Title: {papers[corpus_ids[id]]["title"]}
            Abstract: {papers[corpus_ids[id]]["abstract"]}"""
            if "introduction" in papers[corpus_ids[id]]:
                prompt += f"""
            Background: {papers[corpus_ids[id]]["introduction"]}"""
            prompt += f"""

    INSTRUCTIONS:
    You are ScientistGPT, an intelligent assistant that helps researchers come up with
    ↪ understandable, relevant, specific, feasible, and novel research ideas.
    Present the purpose, mechanism, and evaluation of each text above.
    The purpose is the problem being addressed (e.g., to assist with writing scientific tweetorials,
    ↪ to answer questions over the scientific literature).
    The mechanism is the proposed method to solve the problem (e.g., LLM chain-of-thought reasoning,
    ↪ AI-supported reverse outlining).
    The evaluation is the method used to determine how well the proposed solution solved the problem
    ↪ (e.g., lab user study, science QA benchmarks).
```

Follow the rules below for generating each facet (purpose, mechanism, and evaluation):

1. Specific enough to be helpful in coming up with research ideas.
2. Substantially different from the facets you generate for other texts.
3. Single short phrases only (no more than 7 words). If you cannot write the facet in a short phrase, it is too specific.
4. No numbers unless they are part of a name (e.g., GPT-4, big 5 personality traits).
5. No acronyms or abbreviations.
6. If the text has more than one of the same type of facet, combine them into one.
7. No referencing the purpose in the evaluation facet.

Examples of bad vs good purposes:

- bad (too specific): to generate creative writing activities for third-grade English lessons  
↔ --> good: to support elementary creative writing
- bad (too broad): to support healthcare --> good: to provide clinical decision support
- bad (more than one purpose that are uncombined): to improve engagement between content creators and audience, to decrease negative effects of social media --> good: to improve social media creator-audience interaction

Examples of bad vs good mechanisms:

- bad (too specific, numbers that aren't part of a name, acronym): LLM chain-of-thought from gpt-3.5-turbo trained up to 11/06 with temperature=0.7 --> good: LLM chain-of-thought reasoning
- bad (too broad): recommendation system --> good: collaborative filtering
- bad (too broad): human-AI collaboration --> good: human-generative AI co-planning
- bad (too broad): deep learning algorithm --> good: topic modeling
- bad (more than one mechanism that are uncombined): content-based AI explanations, social-based AI explanations --> good: content-based and social-based AI explanations

Examples of bad vs good evaluations:

- bad (too specific, references the purpose): between-subjects 4x4 user study with 32 teachers  
↔ --> good: Wizard of Oz user study
- bad (too broad): questionnaire --> good: NASA-TLX Index
- bad (too broad): qualitative evaluation --> good: semi-structured interviews

Follow the rules below for generating definitions of each facet.

1. Should be up to 2 sentences.
2. Replace proper nouns with their definitions.
3. Replace jargon with their definitions.
4. Write out acronyms.
5. Should be self-contained. Do NOT include information that is beyond the definition of the  
↔ facet.
6. Do NOT reuse the words already in the facet.

Examples of bad vs good definitions:

- facet: longitudinal study.

bad: a study that evaluates the tool Toolio over the course of a year --> good: a study that  
↔ takes place over a long period of time extending at least multiple days

- facet: Toolio for creative writing.

bad: Toolio implements SLM for generating creative writing --> good: a mixed-initiative tool  
↔ that uses large language models to scaffold the process of writing creative short stories by  
↔ implementing the Standard Learning Method

- facet: to help users better understand black-box models.

bad: to help users better understand AI-Bot-360 --> good: to help users better understand how AI  
↔ models work when their algorithm cannot be fully known

Make sure all information is faithful to the associated text.

It is very important that you follow the answer format provided below!

FORMAT FOR ANSWER:

Text <number>

Purpose: To <verb> <rest of purpose here>

Purpose Definition: <purpose definition here>

Mechanism: <noun phrase mechanism here>

Mechanism Definition: <mechanism definition here>

Evaluation: <noun phrase evaluation here>

Evaluation Definition: <evaluation definition here>

ANSWER:

"""

```
return prompt
```

### *B.1.2 Prompt to retrieve facets from papers associated with an analogous query.*

```
def promptFacetsFromQueryPapers(papers, corpus_ids, query, type="", query2="", type2=""):
    prompt = f"""
    TEXTS: """
    for id in range(0, len(corpus_ids)):
        prompt += f"""
        Text { id + 1}
        Title: {papers[corpus_ids[id]]["title"]}
        Abstract: {papers[corpus_ids[id]]["abstract"]} """
        if "introduction" in papers[corpus_ids[id]]:
            prompt += f"""
            Background: {papers[corpus_ids[id]]["introduction"]} """
        prompt += f"""

    INSTRUCTIONS:
    You are ScientistGPT, an intelligent assistant that helps researchers come up with
    ↔ understandable, relevant, specific, feasible, and novel research ideas.
    Present the purpose, mechanism, and evaluation of each text above. """
    if type:
        prompt += f"""
        The {type} should be relevant to but not a copy of the following query: {query}. """
    else:
        prompt += f"""
        The facets should be relevant to but not a copy of the following query: {query}. """
    if type2:
        prompt += f"""
        The {type2} should be relevant to but not a copy of the following query: {query2}. """
    prompt += f"""
    The purpose is the problem being addressed (e.g., to assist with writing scientific tweetorials,
    ↔ to answer questions over the scientific literature).
```

The mechanism is the proposed method to solve the problem (e.g., chain-of-thought from large  
↔ language models, AI-supported reverse outlining).

The evaluation is the method used to determine how well the proposed solution solved the problem  
↔ (e.g., lab user study, science QA benchmarks).

Follow the rules below for generating each facet (purpose, mechanism, and evaluation):

1. Specific enough to be helpful in coming up with research ideas.
2. Substantially different from the facets you generate for other texts.
3. Single short phrases only (no more than 7 words). If you cannot write the facet in a short  
↔ phrase, it is too specific.
4. No numbers unless they are part of a name (e.g., GPT-4, big 5 personality traits).
5. No acronyms or abbreviations.
6. If the text has more than one of the same type of facet, combine them into one.
7. No referencing the purpose in the evaluation facet."

if type:

```
prompt += f"""
```

6. The {type} should be relevant to but not a copy of the following query: {query}."

else:

```
prompt += f"""
```

6. The facets should be relevant to but not a copy of the following query: {query}."

if type2:

```
prompt += f"""
```

7. The {type2} should be relevant to but not a copy of the following query: {query2}."

```
prompt += f"""
```

Examples of bad vs good purposes:

- bad (too specific): to generate creative writing activities for third-grade English lessons

↔ --> good: to support elementary creative writing

- bad (too broad): to support healthcare --> good: to provide clinical decision support

- bad (more than one purpose that are uncombined): to improve engagement between content

↔ creators and audience, to decrease negative effects of social media --> good: to improve

↔ social media creator-audience interaction

Examples of bad vs good mechanisms:

- bad (too specific, numbers that aren't part of a name, acronym): LLM chain-of-thought from  
 ↪ gpt-3.5-turbo trained up to 11/06 with temperature=0.7 --> good: chain-of-thought from large  
 ↪ language models
- bad (too broad): recommendation system --> good: collaborative filtering
- bad (too broad): human-AI collaboration --> good: human-generative AI co-planning
- bad (too broad): deep learning algorithm --> good: topic modeling
- bad (more than one mechanism that are uncombined): content-based AI explanations, social-based  
 ↪ AI explanations --> good: content-based and social-based AI explanations

Examples of bad vs good evaluations:

- bad (too specific, references the purpose): between-subjects 4x4 user study with 32 teachers  
 ↪ --> good: Wizard of Oz user study
- bad (too broad): questionnaire --> good: NASA-TLX Index
- bad (too broad): qualitative evaluation --> good: semi-structured interviews

Follow the rules below for generating definitions of each facet.

1. Should be up to 2 sentences.
2. Replace proper nouns with their definitions.
3. Replace jargon with their definitions.
4. Write out acronyms.
5. Should be self-contained. Do NOT include information that is beyond the definition of the  
 ↪ facet.
6. Do NOT reuse the words already in the facet.

Examples of bad vs good definitions:

- facet: longitudinal study.  
 bad: a study that evaluates the tool Toolio over the course of a year --> good: a study that  
 ↪ takes place over a long period of time extending at least multiple days
- facet: Toolio for creative writing.  
 bad: Toolio implements SLM for generating creative writing --> good: a mixed-initiative tool  
 ↪ that uses large language models to scaffold the process of writing creative short stories by  
 ↪ implementing the Standard Learning Method
- facet: to help users better understand black-box models.  
 bad: to help users better understand AI-Bot-360 --> good: to help users better understand how AI  
 ↪ models work when their algorithm cannot be fully known

```
Make sure all information is faithful to the associated text.
It is important that you follow the answer format provided below!"""
prompt += f"""
```

```
FORMAT FOR ANSWER:
Text <number>
Purpose: To <verb> <rest of purpose here>
Purpose Definition: <purpose definition here>
Mechanism: <noun phrase mechanism here>
Mechanism Definition: <mechanism definition here>
Evaluation: <noun phrase evaluation here>
Evaluation Definition: <evaluation definition here>
```

```
ANSWER:
"""
return prompt
```

### *B.1.3 Prompt to obtain queries for analogous papers from same topic, same subarea, and different subarea compared to input paper.*

```
def promptAnalogyQueries(purpose, mechanism, previousQueries, topic, number=1):
    prompt = f"""
    INITIAL PURPOSE/MECHANISM:
    Purpose: {purpose}
    Mechanism: {mechanism}"""
    if previousQueries != []:
        prompt += f"""

    PREVIOUS QUERIES: """
        for q in previousQueries:
            prompt += f"""
            {q}"""
    prompt += f"""
```

INSTRUCTIONS:

You are ScientistGPT, an intelligent assistant that helps researchers come up with novel and

↔ useful research ideas.

What are some analogous purposes and mechanisms to the initial purpose and mechanism above that

↔ would inspire novel and useful research ideas?"

if number > 1:

    prompt += f"""

Provide {number} analogous purpose/mechanism pairs that are from the same topic of computer

↔ science research, {number} that are from the same subarea but different topics of computer

↔ science research, and {number} that are from different subareas of computer science

↔ research."""

else:

    prompt += f"""

Provide one analogous purpose/mechanism pair that is from the same topic of computer science

↔ research, one that is from the same subarea but different topics of computer science

↔ research, and one that is from different subareas of computer science research."""

prompt += f"""

The relationship between the initial purpose and mechanism should be very similar to the

↔ relationship between the analogous purpose and mechanism.

Also, provide a query (up to 5 words) for finding research papers relevant to each analogous

↔ purpose/mechanism."""

if previousQueries != []:

    prompt += f"""

Make sure you come up with new analogous purposes/mechanisms that are NOT covered by the

↔ previous queries above."""

prompt += f"""

FORMAT FOR ANSWER:

Analogies within same topic of computer science research:

Same Topic: <topic of initial purpose/mechanism and analogous purpose/mechanism here>

[number]. Analogy: {purpose} is to {mechanism} as <analogous purpose here> is to <analogous

↔ mechanism here> because both relationships involve <specific common relationship description

↔ here>.

Purpose: <analogous purpose here>

Mechanism: <analogous mechanism here>

Query for Relevant Research Papers: <query combining analogous purpose and mechanism here>

Analogies within same subarea of computer science research, but across different topics of  
↔ computer science research:

Same Subarea: <subarea of initial purpose/mechanism and analogous purpose/mechanism here>

[number]. Different Topic: <topic of analogous purpose/mechanism here>

Analogy: {purpose} is to {mechanism} as <analogous purpose here> is to <analogous mechanism  
↔ here> because both relationships involve <specific common relationship description here>.

Purpose: <analogous purpose here>

Mechanism: <analogous mechanism here>

Query for Relevant Research Papers: <query combining analogous purpose and mechanism here>

Analogies across different subareas of computer science research:

[number]. Different Subarea: <different subarea from initial purpose/mechanism here>

Analogy: {purpose} is to {mechanism} as <analogous purpose here> is to <analogous mechanism  
↔ here> because both relationships involve <specific common relationship description here>.

Purpose: <analogous purpose here>

Mechanism: <analogous mechanism here>

Query for Relevant Research Papers: <query combining analogous purpose and mechanism here>

ANSWER:

"""

return prompt

*B.1.4 Prompt to shorten a query for papers in the event that it is too long to retrieve 4 relevant papers.*

```
def promptShortenQuery(query):
```

```
    prompt = f"""
```

```
    QUERY: {query}
```

```
    INSTRUCTIONS:
```

You are ScientistGPT, an intelligent assistant that helps researchers come up with  
↪ understandable, relevant, specific, feasible, and novel research ideas.

The query above is too specific to retrieve any research papers.

Please provide a simpler and shorter version of the query to find relevant research papers.

If you must lose some meaning when shortening the query, prioritize the most important  
↪ information.

```
ANSWER:
"""
return prompt
```

### *B.1.5 Prompt to summarize input and very near analogous papers to obtain related works' summary.*

```
def promptSummarizePapers(facets, papers, corpus_ids):
    prompt = f"""
PAPERS: """
    index = 0
    for t in corpus_ids:
        index += 1
        prompt += f"""
Paper {index}:
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]}"""
        if "introduction" in papers[t]:
            prompt += f"""
Background: {papers[t]["introduction"]}"""
        prompt += f"""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}"""
    prompt += f"""
```

INSTRUCTIONS:

You are ScientistGPT, an intelligent assistant that helps researchers come up with

↔ understandable, relevant, specific, feasible, and novel research ideas.

Summarize the prior work above in around 300 words.

Do not summarize individual papers one-by-one. Instead, summarize their contributions as a whole.

That way, you will know what has already been done in research and will not propose similar

↔ ideas.

Instead, you will come up with novel ideas that build upon the designated papers.

ANSWER:

"""

return prompt

## B.2 Prompts for Faceted Idea Generator

### *B.2.1 Prompt to obtain ideas using **Initial** or **No-P-no-M** method.*

```
def promptInitialAnalogyIdeas(  
    topic,  
    papers,  
    facets, query,  
    designated_papers,  
    analogous_papers,  
    summary,  
    paper_background,  
    set_eval=[],  
    prior_ideas="",  
    custom_instructions="",  
    number=6,  
    idea_number=2):  
    prompt = f"""  
    INSTRUCTIONS:
```

You are ScientistGPT, an intelligent assistant that helps researchers come up with  
↪ understandable, relevant, specific, feasible, and novel research ideas.

First, read the summary of prior work and the papers below.

That way, you will know what has already been done in related research."""

if prior\_ideas:

    prompt += f"""

Next, read your prior ideas below in order to make sure you do not generate similar ideas to

↪ those that you have already proposed."""

if number:

    prompt += f"""

Next, come up with {number} different analogies between the purpose/mechanism of a designated

↪ paper and the purpose/mechanism of an analogous paper."""

else:

    prompt += f"""

Next, write down all {len(designated\_papers)\*len(analogous\_papers)} possible analogies between

↪ the purpose/mechanism of a designated paper and the purpose/mechanism of an analogous

↪ paper."""

prompt += f"""

Example Analogy: The purpose "to expand exploration of creative design spaces" is to the

↪ mechanism "structured generation framework" as the purpose "to enhance personalized

↪ mathematics learning" is to the mechanism "guiding and adaptive prompts" because both

↪ relationships involve providing flexible yet structured experiences in order to support

↪ personalized and useful knowledge acquisition.

Also, come up with one short research idea associated with each analogy. Some should combine the

↪ designated paper's purpose with the analogous paper's mechanism, and others should combine

↪ the analogous paper's purpose with the designated paper's mechanism.

Next, select the {idea\_number} best analogies that inspire research ideas that meet the idea

↪ requirements below."""

if set\_eval:

    prompt += f"""

```
Finally, for one analogy, create a research idea that combines the purpose from the analogous
↔ paper, the mechanism from the designated paper, and one of the evaluation options below in
↔ an imaginative and smart manner.
For the other analogy, create a research idea that combines the purpose from the designated
↔ paper, the mechanism from the analogous paper, and one of the evaluation options below in an
↔ imaginative and smart manner."""
```

```
else:
```

```
    prompt += f"""
```

```
Finally, for one analogy, create a research idea that combines the purpose from the analogous
↔ paper, the mechanism from the designated paper, and the evaluation from either paper in an
↔ imaginative and smart manner.
For the other analogy, create a research idea that combines the purpose from the designated
↔ paper, the mechanism from the analogous paper, and the evaluation from either paper in an
↔ imaginative and smart manner."""
```

```
prompt += f"""
```

```
SUMMARY OF PRIOR WORK:
```

```
{summary}"""
```

```
if prior_ideas:
```

```
    prompt += f"""
```

```
PRIOR IDEAS:
```

```
{prior_ideas}"""
```

```
prompt += f"""
```

```
DESIGNATED PAPERS:"""
```

```
index = 0
```

```
for t in designated_papers:
```

```
    index += 1
```

```
    prompt += f"""
```

```
Paper {index}:
```

```
Title: {papers[t]["title"]}
```

```

Abstract: {papers[t]["abstract"]}""
    if paper_background and "introduction" in papers[t]:
        prompt += f""
Background: {papers[t]["introduction"]}""
    prompt += f""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}""
    if not set_eval:
        prompt += f""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}""
if query:
    prompt += f""

ANALOGOUS PAPERS RELATED TO {query}:""
else:
    prompt += f""

ANALOGOUS PAPERS: ""
index = 0
for t in analogous_papers:
    index += 1
    prompt += f""
Paper {index}: ""
    if papers[t]["relevantQuery"] != "none":
        prompt += f""
Theme: {papers[t]["relevantQuery"]}""
    prompt += f""
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]}""
    if paper_background and "introduction" in papers[t]:
        prompt += f""
Background: {papers[t]["introduction"]}""

```

```

    if "relatedWork" in papers[t]:
        if papers[t]["relatedWork"]:
            prompt += f""
Related Work: {papers[t]["relatedWork"]}""
    prompt += f""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}""
    if not set_eval:
        prompt += f""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}""
if set_eval:
    prompt += f""

```

EVALUATION OPTIONS: ""

```

    ind = 0
    for eval in set_eval:
        prompt += f""
{ind}. Evaluation Text: {eval[1]}
Evaluation ID: {eval[2]}""
        ind += 1
prompt += f""

```

FACET REQUIREMENTS:

The purpose/mechanism/evaluation IDs should be written VERBATIM from what is provided.  
Do NOT make up facet IDs that are not explicitly given above.

IDEA REQUIREMENTS:

1. Understandability::
  - 1a. Each idea should be logical.
  - 1b. Each idea should be grammatically correct.

- 1c. Each idea should be self-contained and should not require researchers to have read the  
↪ provided papers. For example, saying "Use Tool X" is not self-contained because the  
↪ researcher might not have read the relevant paper to know what Tool X is. Instead, you could  
↪ say "Use a tool to do Y."
- 1d. If an aspect of the idea may not be familiar to someone studying computer science broadly,  
↪ make sure the idea describes the aspect enough for them to understand it.

## 2. Relevance::

- 2a. Each idea should be at least somewhat relevant to the overarching ideation topic: {topic}.  
↪ This means you may or may not have to adjust the purpose to be relevant to the ideation  
↪ topic. For example, if the purpose is "to support human-AI creativity" and the topic is  
↪ "scientific idea generation", then the purpose is relevant enough to the topic and does NOT  
↪ need to be adjusted. On the other hand, if the selected purpose is "to support bird  
↪ identification" and the topic is "scientific idea generation", you might adjust the purpose  
↪ to "to support scientific idea identification" or "to support identifying different types of  
↪ scientific ideas." That said, it is important to keep as much of the purpose's meaning as  
↪ possible if/when adjusting it to be relevant to the ideation topic.
- 2b. Ideas should NOT refer to analogies.

## 3. Specificity::

- 3a. Each idea should be described in 100 to 150 words (expanded version 200 to 250 words). 90%  
↪ of the idea should focus on describing how the mechanism will be used to address the purpose  
↪ in a novel manner.
- 3b. Each idea should be as specific as possible given the word limit.
- 3c. Each idea should be focused and precise. For example, if the idea involves a visualization  
↪ feature, describe WHAT kind of visualization it is and HOW it will be used.
- 3d. Each idea should be specific such that the reader understands how the idea would be  
↪ implemented and what direction it would take.

3e. Each idea should be very specific in describing HOW the mechanism will be implemented to  
↪ address the purpose in a novel manner. For instance, an idea saying to 'apply a faceted  
↪ representation to clinical data, creating a multidimensional profile of each patient,  
↪ integrating medical history, genetic information, lifestyle factors, and current health  
↪ data' is not novel because prior work has already looked into creating multidimensional  
↪ patient profiles. If the idea provided a more specific, imaginative description of HOW to  
↪ create the multidimensional profiles and HOW to apply it to clinical decision-making, then  
↪ it could be novel.

3f. Each idea should be specific enough that it is significantly different from your prior ideas.

#### 4. Feasibility::

4a. A research lab with moderate resources should be able to carry out each idea.

4b. Each idea should make sense scientifically and be grounded in the summary of prior work and  
↪ papers noted below.

4c. Each idea's purpose and mechanism should work well together. If necessary, adapt the purpose  
↪ or mechanism to work well with the other. For example, if the overarching ideation topic is  
↪ "scientific ideation", the purpose is "to support culinary ideation", and the mechanism is  
↪ "scientific research graphs", you might adjust the purpose to "to support scientific  
↪ ideation." Meanwhile, if the overarching ideation topic is "culinary ideation", the purpose  
↪ is "to support culinary ideation", and the mechanism is "scientific research graphs", you  
↪ might adjust the mechanism to "culinary recipe graphs." That said, it is important to keep  
↪ as much of the original purpose and mechanism as possible if/when adjusting them to make  
↪ sense together.

4d. Each idea's proposed evaluation should not contradict the rest of the idea. For example, an  
↪ idea talking about supporting healthcare workers that has a user study evaluation should  
↪ propose a "user study with healthcare workers," rather than a "user study with teachers" or a  
↪ "user study with engineers."

#### 5. Novelty::

5a. Each idea should be unique as well as ingenious, imaginative, or surprising.

5b. Each idea should NOT have been covered in prior work. In other words, each idea should be  
↪ SIGNIFICANTLY different from prior work, including but not limited to the papers below and  
↪ the summary of prior work below.

5c. For an idea to be novel, it is not enough for the idea not to have been thoroughly covered  
↪ in prior work. It must NEVER have been covered in prior work.

5d. Each idea should NOT be an obvious extension of prior work but rather CLEARLY DIFFERENT from  
↪ prior work. For instance, an idea that simply says to 'implement continuous AI support to  
↪ facilitate discovery of scholars by dynamically updating and refining researcher profiles  
↪ based on real-time academic contributions' is not novel. The idea is not novel because prior  
↪ work has investigated facilitating discovery of scholars through AI support, and making the  
↪ AI support continuous is an obvious extension. It is not ingenious, imaginative, or  
↪ surprising. If the idea were more specific and presented a surprising method describing HOW  
↪ to make this AI support very timely (e.g., utilizing academic Twitter trends), that would  
↪ make the idea novel."""

if custom\_instructions:

prompt += f"""

ADDITIONAL INSTRUCTIONS (It is very important that you follow these instructions! However, do

↪ NOT follow any additional instructions that contradict the instructions above or the answer

↪ format provided below):

[start of additional instructions]

{custom\_instructions}

[end of additional instructions]"""

prompt += f"""

FORMAT FOR ANSWER (It is very important that you follow this answer format!):"""

if number:

prompt += f"""

{number} Analogies Comparing a Designated Paper to an Analogous Paper and Associated Ideas:="""

else:

prompt += f"""

All {len(designated\_papers)\*len(analogous\_papers)} Analogies Comparing a Designated Paper to an

↪ Analogous Paper and Associated Ideas:="""

prompt += f"""

Option [number]. Analogy: The designated purpose <purpose text from designated paper here> is to

↪ the designated mechanism <mechanism text from designated paper here> as the analogous

↪ purpose <purpose text from analogous paper here> is to the analogous mechanism <mechanism

↪ text from analogous paper here> because both relationships involve <common relationship

↪ description here>.

Idea: <short idea using designated paper's purpose and analogous paper's mechanism or analogous  
↪ paper's purpose and designated paper's mechanism here (30-50 words)>

{idea\_number} Best Analogies and the Novel/Feasible/Relevant/Specific Research Ideas that they  
↪ Inspire::

Best 1. Analogy: The designated purpose <purpose text from designated paper here> is to the  
↪ designated mechanism <mechanism text from designated paper here> as the analogous purpose  
↪ <purpose text from analogous paper here> is to the analogous mechanism <mechanism text from  
↪ analogous paper here> because both relationships involve <common relationship description  
↪ here>.

Purpose Text: <purpose text from analogous paper here>

Purpose ID: <purpose ID from analogous paper here>

Mechanism Text: <mechanism text from designated paper here>

Mechanism ID: <mechanism ID from designated paper here>""

if not set\_eval:

prompt += f"""

Evaluation Text: <evaluation text from either paper here>

Evaluation ID: <evaluation ID here>""

else:

prompt += f"""

Evaluation Text: <selected evaluation option text here>

Evaluation ID: <selected evaluation option ID here>""

prompt += f"""

Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add  
↪ to the facet combination of <analogous purpose text here> with <designated mechanism text  
↪ here> will be <imaginative twist here>.

How Idea will be Relevant to {topic}: The idea will be relevant to {topic}, as it will address  
↪ <thing relevant to {topic}>.

Initial Research Idea: <idea inspired by facets here (100-150 words)>

Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100  
↪ words)>

How to Address Issues: <describe how will resolve issues here (50-100 words)>

New Research Idea: <updated idea inspired by facets here (100-150 words)>

Expanded New Research Idea: <expanded updated idea inspired by facets here (200-250 words)>

```

Best 2. Analogy: The designated purpose <purpose text from designated paper here> is to the
↳ designated mechanism <mechanism text from designated paper here> as the analagous purpose
↳ <purpose text from analogous paper here> is to the analogous mechanism <mechanism text from
↳ analogous paper here> because both relationships involve <common relationship description
↳ here>.

Purpose Text: <purpose text from designated paper here>
Purpose ID: <purpose ID from designated paper here>
Mechanism Text: <mechanism text from analogous paper here>
Mechanism ID: <mechanism ID from analogous paper here>""
if not set_eval:
    prompt += f""
Evaluation Text: <evaluation text from either paper here>
Evaluation ID: <evaluation ID here>""
else:
    prompt += f""
Evaluation Text: <selected evaluation option text here>
Evaluation ID: <selected evaluation option ID here>""
prompt += f""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↳ to the facet combination of <designated purpose text here> with <analogous mechanism text
↳ here> will be <imaginative twist here>.
How Idea will be Relevant to {topic}: The idea will be relevant to {topic}, as it will address
↳ <thing relevant to {topic}>.
Initial Research Idea: <idea inspired by facets here (100-150 words)>
Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100
↳ words)>
How to Address Issues: <describe how will resolve issues here (50-100 words)>
New Research Idea: <updated idea inspired by facets here (100-150 words)>
Expanded New Research Idea: <expanded updated idea inspired by facets here (200-250 words)>

ANSWER:
""
return prompt

```

### B.2.2 Prompt to obtain ideas using *P-or-M* method.

```
def promptFillAnalogyIdeas(  
    papers,  
    facets,  
    summary,  
    designated_papers,  
    analogous_papers,  
    relevant_purposes,  
    relevant_facets_by_paper,  
    topic,  
    paper_background,  
    set_eval=[],  
    prior_ideas="",  
    custom_instructions="",  
    number=6,  
    idea_number=2  
):  
    prompt = f"""  
    INSTRUCTIONS:  
    You are ScientistGPT, an intelligent assistant that helps researchers come up with  
    ↔ understandable, relevant, specific, feasible, and novel research ideas.  
  
    First, read the summary of prior work and the papers below.  
    That way, you will know what has already been done in related research."""  
    if prior_ideas:  
        prompt += f"""  
  
    Next, read your prior ideas below in order to make sure you do not generate similar ideas to  
    ↔ those that you have already proposed."""  
    if number:  
        prompt += f"""  
  
    Next, come up with {number} different analogies between the purpose/mechanism of a set-1 paper  
    ↔ below and the purpose/mechanism of a set-2 paper below."""  
    else:
```

```

prompt += f"""

Next, write down all {len(designated_papers)*len(analogous_papers)} possible analogies between
↔ the purpose/mechanism of a set-1 paper and the purpose/mechanism of a set-2 paper."""
prompt += f"""

If a set-1 paper does not have a purpose or mechanism, create an appropriate one for the sake of
↔ the analogy.

If possible, the paper from which the purpose comes must have a different distance than the
↔ paper from which the mechanism comes.

Example Analogy: The purpose "to expand exploration of creative design spaces" is to the
↔ mechanism "structured generation framework" as the purpose "to enhance personalized
↔ mathematics learning" is to the mechanism "guiding and adaptive prompts" because both
↔ relationships involve providing flexible yet structured experiences in order to support
↔ personalized and useful knowledge acquisition.

Also, come up with a short research idea based on each analogy that combines the purpose from
↔ one paper with the mechanism from the other paper.

Next, select the {idea_number} best analogies that inspire research ideas that meet the idea
↔ requirements below.

"""
if relevant_purposes:
    if set_eval:
        prompt += f"""

Finally, for each analogy, create a research idea that combines the purpose from the set-1 paper
↔ with the mechanism from the set-2 paper, and one of the evaluation options below in an
↔ imaginative and smart manner."""
    else:
        prompt += f"""

Finally, for each analogy, create a research idea that combines the purpose from the set-1
↔ paper, the mechanism from the set-2 paper, and the evaluation from one of those two papers
↔ in an imaginative and smart manner."""
else:
    if set_eval:
        prompt += f"""

```

```

Finally, for each analogy, create a research idea that combines the mechanism from the set-1
↔ paper, the purpose from the set-2 paper, and one of the evaluation options below in an
↔ imaginative and smart manner."""
    else:
        prompt += f"""
Finally, for each analogy, create a research idea that combines the mechanism from the set-1
↔ paper, the purpose from the set-2 paper, and the evaluation from one of those two papers in
↔ an imaginative and smart manner."""
prompt += f"""

SUMMARY OF PRIOR WORK:
{summary}"""

if prior_ideas:
    prompt += f"""

PRIOR IDEAS:
{prior_ideas}"""
prompt += f"""

SET-1 PAPERS: """
index = 0
for t in designated_papers:
    index += 1
    prompt += f"""
Paper {index}: """
    if t in papers:
        prompt += f"""
Distance: {papers[t]["distance"]}
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]} """
        if paper_background and "introduction" in papers[t]:
            prompt += f"""
Background: {papers[t]["introduction"]} """
            if "relatedWork" in papers[t]:

```

```

        if papers[t]["relatedWork"]:
            prompt += f""
Related Work: {papers[t]["relatedWork"]}""
            prompt += f""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}""
            if not set_eval:
                prompt += f""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}""
            else:
                prompt += f""
Distance: input
Title: n/a
Abstract: n/a""
            if relevant_purposes:
                prompt += f""
Purpose Text: {relevant_facets_by_paper[t][0]["text"]}
Purpose ID: {relevant_facets_by_paper[t][0]["id"]}
Mechanism Text: <any relevant mechanism to the purpose>
Mechanism ID: n/a
Evaluation Text: n/a
Evaluation ID: n/a""
            else:
                prompt += f""
Purpose Text: <any relevant purpose to the mechanism>
Purpose ID: n/a
Mechanism Text: {relevant_facets_by_paper[t][0]["text"]}
Mechanism ID: {relevant_facets_by_paper[t][0]["id"]}
Evaluation Text: n/a
Evaluation ID: n/a""
prompt += f""

```

```

SET-2 PAPERS: ""
index = 0
for t in analogous_papers:
    index += 1
    prompt += f""
Paper {index}:
Distance: {papers[t]["distance"]}""
    if papers[t]["relevantQuery"] != "none":
        prompt += f""
Theme: {papers[t]["relevantQuery"]}""
    prompt += f""
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]}""
    if paper_background and "introduction" in papers[t]:
        prompt += f""
Background: {papers[t]["introduction"]}""
    if "relatedWork" in papers[t]:
        if papers[t]["relatedWork"]:
            prompt += f""
Related Work: {papers[t]["relatedWork"]}""
    prompt += f""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}""
    if not set_eval:
        prompt += f""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}""
if set_eval:
    prompt += f""

EVALUATION OPTIONS: ""
ind = 0
for eval in set_eval:

```

```
prompt += f"""
{ind}. Evaluation Text: {eval[1]}
Evaluation ID: {eval[2]}"""
ind += 1
```

```
prompt += f"""
```

#### FACET REQUIREMENTS:

The purpose/mechanism/evaluation IDs should be written VERBATIM from what is provided.

Do NOT make up facet IDs that are not explicitly given above.

#### IDEA REQUIREMENTS:

##### 1. Understandability::

1a. Each idea should be logical.

1b. Each idea should be grammatically correct.

1c. Each idea should be self-contained and should not require researchers to have read the

↪ provided papers. For example, saying "Use Tool X" is not self-contained because the

↪ researcher might not have read the relevant paper to know what Tool X is. Instead, you could

↪ say "Use a tool to do Y."

1d. If an aspect of the idea may not be familiar to someone studying computer science broadly,

↪ make sure the idea describes the aspect enough for them to understand it.

##### 2. Relevance::

2a. Each idea should be at least somewhat relevant to the overarching ideation topic: {topic}.

↪ This means you may or may not have to adjust the purpose to be relevant to the ideation

↪ topic. For example, if the purpose is "to support human-AI creativity" and the topic is

↪ "scientific idea generation", then the purpose is relevant enough to the topic and does NOT

↪ need to be adjusted. On the other hand, if the selected purpose is "to support bird

↪ identification" and the topic is "scientific idea generation", you might adjust the purpose

↪ to "to support scientific idea identification" or "to support identifying different types of

↪ scientific ideas." That said, it is important to keep as much of the purpose's meaning as

↪ possible if/when adjusting it to be relevant to the ideation topic.

2b. Ideas should NOT refer to analogies.

##### 3. Specificity::

- 3a. Each idea should be described in 100 to 150 words (expanded version 200 to 250 words). 90%  
↳ of the idea should focus on describing how the mechanism will be used to address the purpose  
↳ in a novel manner.
- 3b. Each idea should be as specific as possible given the word limit.
- 3c. Each idea should be focused and precise. For example, if the idea involves a visualization  
↳ feature, describe WHAT kind of visualization it is and HOW it will be used.
- 3d. Each idea should be specific such that the reader understands how the idea would be  
↳ implemented and what direction it would take.
- 3e. Each idea should be very specific in describing HOW the mechanism will be implemented to  
↳ address the purpose in a novel manner. For instance, an idea saying to 'apply a faceted  
↳ representation to clinical data, creating a multidimensional profile of each patient,  
↳ integrating medical history, genetic information, lifestyle factors, and current health  
↳ data' is not novel because prior work has already looked into creating multidimensional  
↳ patient profiles. If the idea provided a more specific, imaginative description of HOW to  
↳ create the multidimensional profiles and HOW to apply it to clinical decision-making, then  
↳ it could be novel.
- 3f. Each idea should be specific enough that it is significantly different from your prior ideas.

#### 4. Feasibility::

- 4a. A research lab with moderate resources should be able to carry out each idea.
- 4b. Each idea should make sense scientifically and be grounded in the summary of prior work and  
↳ papers noted below.
- 4c. Each idea's purpose and mechanism should work well together. If necessary, adapt the purpose  
↳ or mechanism to work well with the other. For example, if the overarching ideation topic is  
↳ "scientific ideation", the purpose is "to support culinary ideation", and the mechanism is  
↳ "scientific research graphs", you might adjust the purpose to "to support scientific  
↳ ideation." Meanwhile, if the overarching ideation topic is "culinary ideation", the purpose  
↳ is "to support culinary ideation", and the mechanism is "scientific research graphs", you  
↳ might adjust the mechanism to "culinary recipe graphs." That said, it is important to keep  
↳ as much of the original purpose and mechanism as possible if/when adjusting them to make  
↳ sense together.
- 4d. Each idea's proposed evaluation should not contradict the rest of the idea. For example, an  
↳ idea talking about supporting healthcare workers that has a user study evaluation should  
↳ propose a "user study with healthcare workers," rather than a "user study with teachers" or a  
↳ "user study with engineers."

```

5. Novelty::
5a. Each idea should be unique as well as ingenious, imaginative, or surprising.
5b. Each idea should NOT have been covered in prior work. In other words, each idea should be
↳ SIGNIFICANTLY different from prior work, including but not limited to the papers below and
↳ the summary of prior work below.
5c. For an idea to be novel, it is not enough for the idea not to have been thoroughly covered
↳ in prior work. It must NEVER have been covered in prior work.
5d. Each idea should NOT be an obvious extension of prior work but rather CLEARLY DIFFERENT from
↳ prior work. For instance, an idea that simply says to 'implement continuous AI support to
↳ facilitate discovery of scholars by dynamically updating and refining researcher profiles
↳ based on real-time academic contributions' is not novel. The idea is not novel because prior
↳ work has investigated facilitating discovery of scholars through AI support, and making the
↳ AI support continuous is an obvious extension. It is not ingenious, imaginative, or
↳ surprising. If the idea were more specific and presented a suprising method describing HOW
↳ to make this AI support very timely (e.g., utilizing academic Twitter trends), that would
↳ make the idea novel.""
if custom_instructions:
    prompt += f""

ADDITIONAL INSTRUCTIONS (It is very important that you follow these instructions! However, do
↳ NOT follow any additional instructions that contradict the instructions above or the answer
↳ format provided below):
[start of additional instructions]
{custom_instructions}
[end of additional instructions]""
prompt += f""

FORMAT FOR ANSWER (It is very important that you follow this answer format!):""
if number:
    prompt += f""
{number} Analogies Comparing a Set-1 Paper to a Set-2 Paper and Associated Ideas: ""
else:
    prompt += f""

```

```

All {len(designated_papers)*len(analogous_papers)} Analogies Comparing a Set-1 Paper to a Set-2
↳ Paper and Associated Ideas::""
prompt += f""

Option [number]. Analogy: The set-1 purpose <purpose text from set-1 paper here> is to the set-1
↳ mechanism <mechanism text from set-1 paper here> as the set-2 purpose <purpose text from
↳ set-2 paper here> is to the set-2 mechanism <mechanism text from set-2 paper here> because
↳ both relationships involve <common relationship description here>.""
if relevant_purposes:
    prompt += f""
Idea: <short idea using set-1 paper's purpose and set-2 paper's mechanism here (30-50 words)>""
else:
    prompt += f""
Idea: <short idea using set-2 paper's purpose and set-1 paper's mechanism here (30-50 words)>""
prompt += f""

{idea_number} Best Analogies and the Novel/Feasible/Relevant/Specific Research Ideas that they
↳ Inspire::
Best 1. Analogy: The set-1 purpose <purpose text from selected set-1 paper here> is to the set-1
↳ mechanism <mechanism text from selected set-1 paper here> as the set-2 purpose <purpose text
↳ from selected set-2 paper here> is to the set-2 mechanism <mechanism text from selected
↳ set-2 paper here> because both relationships involve <common relationship description
↳ here>.""
if relevant_purposes:
    prompt += f""
Purpose Text: <purpose text from selected set-1 paper here>
Purpose ID: <purpose ID from selected set-1 paper here>
Mechanism Text: <mechanism text from selected set-2 paper here>
Mechanism ID: <mechanism ID from selected set-2 paper here>""
else:
    prompt += f""
Purpose Text: <purpose text from selected set-2 paper here>
Purpose ID: <purpose ID from selected set-2 paper here>
Mechanism Text: <mechanism text from selected set-1 paper here>
Mechanism ID: <mechanism ID from selected set-1 paper here>""
if not set_eval:

```

```

    prompt += f"""
Evaluation Text: <text of selected evaluation from either paper here>
Evaluation ID: <selected evaluation ID here>"""
else:
    prompt += f"""
Evaluation Text: <selected evaluation option text here>
Evaluation ID: <selected evaluation option ID here>"""
if relevant_purposes:
    prompt += f"""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↔ to the facet combination of <set-1 purpose text here> with <set-2 mechanism text here> will
↔ be <imaginative twist here>."""
else:
    prompt += f"""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↔ to the facet combination of <set-2 purpose text here> with <set-1 mechanism text here> will
↔ be <imaginative twist here>."""
prompt += f"""
Initial Research Idea: <idea inspired by facets here (100-150 words)>
Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100
↔ words)>
How to Address Issues: <describe how will resolve issues here (50-100 words)>
New Research Idea: <updated idea inspired by facets here (100-150 words)>
Expanded New Research Idea: <expanded idea inspired by facets here (200-250 words)>

Best 2. Analogy: The set-1 purpose <purpose text from selected set-1 paper here> is to the set-1
↔ mechanism <mechanism text from selected set-1 paper here> as the set-2 purpose <purpose text
↔ from selected set-2 paper here> is to the set-2 mechanism <mechanism text from selected
↔ set-2 paper here> because both relationships involve <common relationship description
↔ here>."""
if relevant_purposes:
    prompt += f"""
Purpose Text: <purpose text from selected set-1 paper here>
Purpose ID: <purpose ID from selected set-1 paper here>
Mechanism Text: <mechanism text from selected set-2 paper here>

```

```

Mechanism ID: <mechanism ID from selected set-2 paper here>"""
else:
    prompt += f"""
Purpose Text: <purpose text from selected set-2 paper here>
Purpose ID: <purpose ID from selected set-2 paper here>
Mechanism Text: <mechanism text from selected set-1 paper here>
Mechanism ID: <mechanism ID from selected set-1 paper here>"""
if not set_eval:
    prompt += f"""
Evaluation Text: <text of selected evaluation from either paper here>
Evaluation ID: <selected evaluation ID here>"""
else:
    prompt += f"""
Evaluation Text: <selected evaluation option text here>
Evaluation ID: <selected evaluation option ID here>"""
if relevant_purposes:
    prompt += f"""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↔ to the facet combination of <set-1 purpose text here> with <set-2 mechanism text here> will
↔ be <imaginative twist here>."""
else:
    prompt += f"""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↔ to the facet combination of <set-2 purpose text here> with <set-1 mechanism text here> will
↔ be <imaginative twist here>."""
prompt += f"""
Initial Research Idea: <idea inspired by facets here (100-150 words)>
Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100
↔ words)>
How to Address Issues: <describe how will resolve issues here (50-100 words)>
New Research Idea: <updated idea inspired by facets here (100-150 words)>
Expanded New Research Idea: <expanded idea inspired by facets here (200-250 words)>

ANSWER:
"""

```

```
return prompt
```

### *B.2.3 Prompt to obtain ideas using **P-and-M** method.*

```
def promptFacetsToIdeas(
    topic,
    summary,
    papers,
    designated_papers,
    analogous_papers,
    relevant_facets_by_paper,
    facets,
    paper_background,
    set_eval=[],
    prior_ideas="",
    custom_instructions="",
    number=6,
    idea_number=2):
    prompt = f"""
    INSTRUCTIONS:
    You are ScientistGPT, an intelligent assistant that helps researchers come up with
    ↪ understandable, relevant, specific, feasible, and novel research ideas.

    First, read the summary of prior work and the papers below.
    That way, you will know what has already been done in related research."""
    if prior_ideas:
        prompt += f"""

    Next, read your prior ideas below in order to make sure you do not generate similar ideas to
    ↪ those that you have already proposed."""
    if number:
        prompt += f"""

    Next, come up with {number} different analogies between the purpose/mechanism of a set-1 paper
    ↪ below and the purpose/mechanism of a set-2 paper below."""
```

```

else:
    prompt += f"""

Next, write down all {len(designated_papers)*len(analogous_papers)} possible analogies between
↔ the purpose/mechanism of a set-1 paper and the purpose/mechanism of a set-2 paper."""
prompt += f"""

If a paper does not have a purpose or mechanism, create an appropriate one for the sake of the
↔ analogy.

If possible, the paper from which the purpose comes must have a different distance than the
↔ paper from which the mechanism comes.

Example Analogy: The purpose "to expand exploration of creative design spaces" is to the
↔ mechanism "structured generation framework" as the purpose "to enhance personalized
↔ mathematics learning" is to the mechanism "guiding and adaptive prompts" because both
↔ relationships involve providing flexible yet structured experiences in order to support
↔ personalized and useful knowledge acquisition.

Also, come up with a short research idea based on each analogy that combines the set-1 paper's
↔ purpose with the set-2 paper's mechanism.

Next, select the {idea_number} best analogies that inspire research ideas that meet the idea
↔ requirements below.

"""
if set_eval:
    prompt += f"""

Finally, for each analogy, create a research idea that combines the purpose from a set-1 paper,
↔ the mechanism from a set-2 paper, and one of the evaluation options below in an imaginative
↔ and smart manner."""
else:
    prompt += f"""

Finally, for each analogy, create a research idea that combines the purpose from a set-1 paper,
↔ the mechanism from a set-2 paper, and the evaluation from one of those two papers in an
↔ imaginative and smart manner."""
prompt += f"""

SUMMARY OF PRIOR WORK:

```

```

{summary}"""

if prior_ideas:
    prompt += f"""

PRIOR IDEAS:
{prior_ideas}"""
prompt += f"""

SET-1 PAPERS: """
index = 0
for t in designated_papers:
    index += 1
    prompt += f"""
Paper {index}: """
    if t in papers:
        prompt += f"""
Distance: {papers[t]["distance"]}
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]}"""
        if paper_background and "introduction" in papers[t]:
            prompt += f"""
Background: {papers[t]["introduction"]}"""
            if "relatedWork" in papers[t]:
                if papers[t]["relatedWork"]:
                    prompt += f"""
Related Work: {papers[t]["relatedWork"]}"""
                    prompt += f"""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}
Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}"""
                if not set_eval:
                    prompt += f"""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}

```

```

Evaluation ID: {papers[t]["evaluation"]}""
    else:
        prompt += f""
Distance: input
Title: n/a
Abstract: n/a
Purpose Text: {relevant_facets_by_paper[t][0]["text"]}
Purpose ID: {relevant_facets_by_paper[t][0]["id"]}
Mechanism Text: <any relevant mechanism to the purpose>
Mechanism ID: n/a
Evaluation Text: n/a
Evaluation ID: n/a""
prompt += f""

SET-2 PAPERS: ""
index = 0
for t in analogous_papers:
    index += 1
    prompt += f""
Paper {index}: ""
    if t in papers:
        prompt += f""
Distance: {papers[t]["distance"]}
Title: {papers[t]["title"]}
Abstract: {papers[t]["abstract"]}""
    if paper_background and "introduction" in papers[t]:
        prompt += f""
Background: {papers[t]["introduction"]}""
    if "relatedWork" in papers[t]:
        if papers[t]["relatedWork"]:
            prompt += f""
Related Work: {papers[t]["relatedWork"]}""
    prompt += f""
Purpose Text: {facets[papers[t]["purpose"]]["text"]}
Purpose ID: {papers[t]["purpose"]}

```

```

Mechanism Text: {facets[papers[t]["mechanism"]]["text"]}
Mechanism ID: {papers[t]["mechanism"]}""
    if not set_eval:
        prompt += f""
Evaluation Text: {facets[papers[t]["evaluation"]]["text"]}
Evaluation ID: {papers[t]["evaluation"]}""
    else:
        prompt += f""
Distance: input
Title: n/a
Abstract: n/a
Purpose Text: <any relevant purpose to the mechanism>
Purpose ID: n/a
Mechanism Text: {relevant_facets_by_paper[t][0]["text"]}
Mechanism ID: {relevant_facets_by_paper[t][0]["id"]}
Evaluation Text: n/a
Evaluation ID: n/a""
if set_eval:
    prompt += f""

```

EVALUATION OPTIONS: ""

```

    ind = 0
    for eval in set_eval:
        prompt += f""
{ind}. Evaluation Text: {eval[1]}
Evaluation ID: {eval[2]}""
        ind += 1
prompt += f""

```

FACET REQUIREMENTS:

The purpose/mechanism/evaluation IDs should be written VERBATIM from what is provided.  
Do NOT make up facet IDs that are not explicitly given above.

IDEA REQUIREMENTS:

1. Understandability::

- 1a. Each idea should be logical.
- 1b. Each idea should be grammatically correct.
- 1c. Each idea should be self-contained and should not require researchers to have read the provided papers. For example, saying "Use Tool X" is not self-contained because the researcher might not have read the relevant paper to know what Tool X is. Instead, you could say "Use a tool to do Y."
- 1d. If an aspect of the idea may not be familiar to someone studying computer science broadly, make sure the idea describes the aspect enough for them to understand it.

## 2. Relevance::

- 2a. Each idea should be at least somewhat relevant to the overarching ideation topic: {topic}. This means you may or may not have to adjust the purpose to be relevant to the ideation topic. For example, if the purpose is "to support human-AI creativity" and the topic is "scientific idea generation", then the purpose is relevant enough to the topic and does NOT need to be adjusted. On the other hand, if the selected purpose is "to support bird identification" and the topic is "scientific idea generation", you might adjust the purpose to "to support scientific idea identification" or "to support identifying different types of scientific ideas." That said, it is important to keep as much of the purpose's meaning as possible if/when adjusting it to be relevant to the ideation topic.
- 2b. Ideas should NOT refer to analogies.

## 3. Specificity::

- 3a. Each idea should be described in 100 to 150 words (expanded version 200 to 250 words). 90% of the idea should focus on describing how the mechanism will be used to address the purpose in a novel manner.
- 3b. Each idea should be as specific as possible given the word limit.
- 3c. Each idea should be focused and precise. For example, if the idea involves a visualization feature, describe WHAT kind of visualization it is and HOW it will be used.
- 3d. Each idea should be specific such that the reader understands how the idea would be implemented and what direction it would take.

3e. Each idea should be very specific in describing HOW the mechanism will be implemented to  
↪ address the purpose in a novel manner. For instance, an idea saying to 'apply a faceted  
↪ representation to clinical data, creating a multidimensional profile of each patient,  
↪ integrating medical history, genetic information, lifestyle factors, and current health  
↪ data' is not novel because prior work has already looked into creating multidimensional  
↪ patient profiles. If the idea provided a more specific, imaginative description of HOW to  
↪ create the multidimensional profiles and HOW to apply it to clinical decision-making, then  
↪ it could be novel.

3f. Each idea should be specific enough that it is significantly different from your prior ideas.

#### 4. Feasibility::

4a. A research lab with moderate resources should be able to carry out each idea.

4b. Each idea should make sense scientifically and be grounded in the summary of prior work and  
↪ papers noted below.

4c. Each idea's purpose and mechanism should work well together. If necessary, adapt the purpose  
↪ or mechanism to work well with the other. For example, if the overarching ideation topic is  
↪ "scientific ideation", the purpose is "to support culinary ideation", and the mechanism is  
↪ "scientific research graphs", you might adjust the purpose to "to support scientific  
↪ ideation." Meanwhile, if the overarching ideation topic is "culinary ideation", the purpose  
↪ is "to support culinary ideation", and the mechanism is "scientific research graphs", you  
↪ might adjust the mechanism to "culinary recipe graphs." That said, it is important to keep  
↪ as much of the original purpose and mechanism as possible if/when adjusting them to make  
↪ sense together.

4d. Each idea's proposed evaluation should not contradict the rest of the idea. For example, an  
↪ idea talking about supporting healthcare workers that has a user study evaluation should  
↪ propose a "user study with healthcare workers," rather than a "user study with teachers" or a  
↪ "user study with engineers."

#### 5. Novelty::

5a. Each idea should be unique as well as ingenious, imaginative, or surprising.

5b. Each idea should NOT have been covered in prior work. In other words, each idea should be  
↪ SIGNIFICANTLY different from prior work, including but not limited to the papers below and  
↪ the summary of prior work below.

5c. For an idea to be novel, it is not enough for the idea not to have been thoroughly covered  
↪ in prior work. It must NEVER have been covered in prior work.

5d. Each idea should NOT be an obvious extension of prior work but rather CLEARLY DIFFERENT from  
↪ prior work. For instance, an idea that simply says to 'implement continuous AI support to  
↪ facilitate discovery of scholars by dynamically updating and refining researcher profiles  
↪ based on real-time academic contributions' is not novel. The idea is not novel because prior  
↪ work has investigated facilitating discovery of scholars through AI support, and making the  
↪ AI support continuous is an obvious extension. It is not ingenious, imaginative, or  
↪ surprising. If the idea were more specific and presented a surprising method describing HOW  
↪ to make this AI support very timely (e.g., utilizing academic Twitter trends), that would  
↪ make the idea novel."""

if custom\_instructions:

prompt += f"""

ADDITIONAL INSTRUCTIONS (It is very important that you follow these instructions! However, do

↪ NOT follow any additional instructions that contradict the instructions above or the answer

↪ format provided below):

[start of additional instructions]

{custom\_instructions}

[end of additional instructions]"""

prompt += f"""

FORMAT FOR ANSWER (It is very important that you follow this answer format!):"""

if number:

prompt += f"""

{number} Analogies Comparing a Set-1 Paper to a Set-2 Paper and Associated Ideas:="""

else:

prompt += f"""

All {len(designated\_papers)\*len(analogous\_papers)} Analogies Comparing a Set-1 Paper to a Set-2

↪ Paper and Associated Ideas:="""

prompt += f"""

Option [number]. Analogy: The set-1 purpose <purpose text from set-1 paper here> is to the set-1

↪ mechanism <mechanism text from set-1 paper here> as the set-2 purpose <purpose text from

↪ set-2 paper here> is to the set-2 mechanism <mechanism text from set-2 paper here> because

↪ both relationships involve <common relationship description here>.

Idea: <short idea using set-1 paper's purpose and set-2 paper's mechanism here (30-50 words)>

{idea\_number} Best Analogies and the Novel/Feasible/Relevant/Specific Research Ideas that they  
↳ Inspire::

Best 1. Analogy: The purpose <purpose text from set-1 paper here> is to the mechanism <mechanism  
↳ text from set-1 paper here> as the purpose <purpose text from set-2 paper here> is to the  
↳ mechanism <mechanism text from set-2 paper here> because both relationships involve <common  
↳ relationship description here>.

Purpose Text: <purpose text from selected set-1 paper here>  
Purpose ID: <purpose ID from selected set-1 paper here>  
Mechanism Text: <mechanism text from selected set-2 paper here>  
Mechanism ID: <mechanism ID from selected set-2 paper here>""  
if not set\_eval:  
    prompt += f""  
Evaluation Text: <text of selected evaluation from either paper here>  
Evaluation ID: <selected evaluation ID here>""  
else:  
    prompt += f""  
Evaluation Text: <selected evaluation option text here>  
Evaluation ID: <selected evaluation option ID here>""  
prompt += f""

Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add  
↳ to the facet combination of <set-1 purpose text here> with <set-2 mechanism text here> will  
↳ be <imaginative twist here>.

Initial Research Idea: <idea inspired by facets here (100-150 words)>  
Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100  
↳ words)>  
How to Address Issues: <describe how will resolve issues here (50-100 words)>  
New Research Idea: <updated idea inspired by facets here (100-150 words)>  
Expanded New Research Idea: <expanded idea inspired by facets here (200-250 words)>

Best 2. Analogy: The purpose <purpose text from set-1 paper here> is to the mechanism <mechanism  
↳ text from set-1 paper here> as the purpose <purpose text from set-2 paper here> is to the  
↳ mechanism <mechanism text from set-2 paper here> because both relationships involve <common  
↳ relationship description here>.

Purpose Text: <purpose text from selected set-1 paper here>  
Purpose ID: <purpose ID from selected set-1 paper here>

```

Mechanism Text: <mechanism text from selected set-2 paper here>
Mechanism ID: <mechanism ID from selected set-2 paper here>"""
if not set_eval:
    prompt += f"""
Evaluation Text: <text of selected evaluation from either paper here>
Evaluation ID: <selected evaluation ID here>"""
else:
    prompt += f"""
Evaluation Text: <selected evaluation option text here>
Evaluation ID: <selected evaluation option ID here>"""
prompt += f"""
Imaginative Twist to Add to Facet Combination: The imaginative and smart twist that I will add
↔ to the facet combination of <set-1 purpose text here> with <set-2 mechanism text here> will
↔ be <imaginative twist here>.
Initial Research Idea: <idea inspired by facets here (100-150 words)>
Issues with Initial Idea: <describe how initial idea doesn't meet idea requirements here (50-100
↔ words)>
How to Address Issues: <describe how will resolve issues here (50-100 words)>
New Research Idea: <updated idea inspired by facets here (100-150 words)>
Expanded New Research Idea: <expanded idea inspired by facets here (200-250 words)>

ANSWER:
"""
return prompt

```

## B.3 Prompts for Idea Novelty Checker

### *B.3.1 Prompt to assess idea novelty.*

```

def promptNoveltyChecker(idea, similar_documents, incontext_part):
    relevant_papers = []
    for i, (_, row) in enumerate(similar_documents.iterrows()):
        relevant_papers.append(
            {

```

```

        "role": "user",
        "content": f"Paper ID [{i}]: Title: {row['title']}. Abstract: {row['abstract']}",
    }
)

prompt = [
    {
        "role": "system",
        "content": "You are ReviewerGPT, an intelligent assistant that helps researchers
↔ evaluate the novelty of their ideas.",
    },
    {
        "role": "user",
        "content": f""You are given some papers similar to the proposed idea. Your task is to
↔ evaluate the idea's novelty using the related papers only.

Types of novelty categories:
- Not Novel: The idea closely replicates existing work with minimal or no new
↔ contributions.
- Novel:
    - The idea introduces new concepts or approaches that are not common in existing
↔ literature.
    - The idea uniquely combines concepts from existing papers, but this combination
↔ does not occur in any related papers.
    - A new application with same approach is also novel.

Instructions:
- Use the example reviews below to write a review for the provided idea by comparing
↔ it to the related papers.
- Don't assume any prior knowledge about the idea.
- When referencing a related paper, then use paper id in the review, mention it in
↔ this format: [5]. The paper ID is present between Paper ID [<paper_id>]: Title.
- After reviewing, classify the idea into one of this category: novel or not novel.
- Make sure the generated review follows the format in example reviews provided
↔ below.

```

- The review should be concise - around 60 to 100 words.

```
{incontext_part}
```

Output Format:

- Class: [novel / not novel]

- Review: The idea is [novel / not novel] different because...

```
"""
```

```
},
```

```
{"role": "assistant", "content": "Sure, please provide the IDEA."},
```

```
{"role": "user", "content": f"Here is the idea: {idea}"},
```

```
{"role": "assistant", "content": "Okay, now provide the related papers."},
```

```
]
```

```
prompt.extend(relevant_papers)
```

```
return prompt
```

### *B.3.2 Prompt to extract specific keywords and generate concise research titles from an IDEA.*

```
def get_keywords(idea):
```

```
    prompt = [
```

```
        {
```

```
            "role": "system",
```

```
            "content": "You are an intelligent assistant that extracts high-quality keywords and
```

```
↪ generates specific research paper titles based on the provided IDEA.",
```

```
        },
```

```
        {
```

```

"role": "user",
"content": """You are tasked with extracting specific keywords and generating potential
↪ research paper titles that closely align with the provided IDEA. These should
↪ capture both the novelty and mechanisms of the IDEA, especially where it diverges
↪ from existing work.

**Keyword Extraction Guidelines**:
```

1. Highlight unique methods, technologies, and application areas.
2. Ensure the keywords specifically capture what sets this idea apart from others.
3. Generate 3-6 keyword phrases, each consisting of 3-6 words.
4. Avoid overly general keywords (e.g., "machine learning" or "data science").
5. Ensure the keywords reflect the precise purpose, mechanisms, and novelty of the idea.

```

**Title Generation Guidelines**:
```

1. Keep titles concise (max 5 words).
2. Avoid generic terms or overused phrases.
3. Reflect the uniqueness and novelty of the idea in each title.
4. Include a key concept from the IDEA's mechanism (e.g., 'retrieval-augmented
↪ generation for idea synthesis').
5. Ensure the title reflects the application domain.

```

**Output Format**:
```

```

<keywords>
["specific keyword phrase 1", "specific keyword phrase 2", "specific keyword phrase 3"]
</keywords>

<titles>
["Title 1", "Title 2", "Title 3", "Title 4"]
</titles>
""",
},
{"role": "assistant", "content": "Sure, please provide the IDEA."},
{"role": "user", "content": idea},
]
return prompt
```

### *B.3.3 Prompt to extract key facets from an IDEA for facet-based relevance re-ranking of papers.*

```
def prompt_noveltyCheckLLMRerank_IdeaFacets(idea):

    prompt = [
        {
            "role": "system",
            "content": "You are Research Idea Reviewer GPT, an intelligent assistant that helps
            ↪ researchers evaluate the novelty of their ideas.",
        },
        {
            "role": "user",
            "content": ""Your task is to extract key facets from a given idea to assist in
            ↪ re-ranking passages based on their relevance to the idea. These key facets should
            ↪ capture the essential elements of the idea, such as the application domain, purpose,
            ↪ mechanisms, methods, and evaluation metrics.

            Instructions:

            1. Carefully read and understand the idea.
            2. Identify and list the key facets of the idea, including but not limited to:
               - Application Domain: The specific field or area the idea pertains to.
               - Purpose/Objective: The main goal or intention behind the idea.
               - Mechanisms/Methods: The techniques or approaches proposed to achieve the
               ↪ purpose.
               - Evaluation Metrics: The criteria or measures used to assess the effectiveness
               ↪ of the idea.

            Examples:
```

Idea 1: Develop a system that uses a faceted representation of authors to  
↔ understand food-health relationships by analyzing the sentiment of research  
↔ papers and publications. The system will identify key authors in food and  
↔ health research, map their sentiments towards various topics, and use this  
↔ information to reveal hidden connections and trends. An experimental results  
↔ showcase will evaluate the system's ability to uncover novel food-health  
↔ relationships and its impact on interdisciplinary research.

Key Facets to Look for in Passages:

- Application Domain: Food and health research.
- Purpose: To understand food-health relationships through sentiment  
↔ analysis.
- Mechanism: Using a faceted representation of authors to map sentiments  
↔ toward various topics.
- Method: Analyzing the sentiment of research papers and publications.
- Evaluation: Experimental showcase evaluating the system's ability to  
↔ uncover novel relationships and its interdisciplinary impact.

Idea 2: Develop a hierarchical topic model that integrates multi-level capsule  
↔ networks to balance sparsity and smoothness in topic models. The capsule  
↔ networks will capture the hierarchical structure of topics while enforcing  
↔ sparsity at lower levels and smoothness at higher levels. This model will be  
↔ validated on benchmark datasets such as PASCAL VOC 2007 and 2012, using  
↔ metrics like log-likelihood and topic coherence to ensure both high  
↔ reconstruction accuracy and generalization capability.

Key Facets to Look for in Passages:

- Application Domain: Topic modeling in machine learning.
- Purpose: To balance sparsity and smoothness in topic models.
- Mechanism: Integrating multi-level capsule networks.
- Method: Developing a hierarchical topic model with capsule networks  
↔ capturing hierarchical structures.
- Evaluation: Validation on benchmark datasets (e.g., PASCAL VOC 2007 and  
↔ 2012) using metrics like log-likelihood and topic coherence.

```

        """
    },
    {
        "role": "assistant",
        "content": "Sure, please provide the research idea",
    },
    {
        "role": "user",
        "content": f"""Here is the idea: <idea> {idea} </idea>.

Please provide Key Facets to Look for in Passages for the provided idea between <facets>
↔ </facets> tags.

"""
    },
]
return prompt

```

### B.3.4 *Prompt to rank passages based on relevance to a query and its key facets.*

```

def prompt_noveltyCheckLLMRerank(query, idea_priority_facets, num):
    return [
        {
            "role": "system",
            "content": "You are RankGPT, an intelligent assistant that can rank passages above based
↔ on their provided priority and relevance to the query and its facets.",
        },
        {
            "role": "user",
            "content": f"""I will provide you with {num} passages, each indicated by number
↔ identifier [].

Your task is to rank the passages based on their relevance to the query idea and the
↔ provided priority:

**Query** Idea: {query}

```

Key facets to look in passages for ranking:

```
{idea_priority_facets}
```

Use the following criteria in order of priority for ranking the passages:

1. **Priority 1:** Passages that closely match **all** key facets of the **QUERY**  
↳ IDEA.
2. **Priority 2:** Passages that match the **application domain** and **purpose**  
↳ but may differ in mechanism or method.
3. **Priority 3:** Passages that share a similar **purpose** or **mechanism** or  
↳ **evaluation**, even if the application domain differs.
4. **Priority 4:** Passages that partially match the application domain or address  
↳ related topics but lack alignment with the purpose or mechanism.

```
""",  
},  
{  
  "role": "assistant",  
  "content": "Can you provide an example idea, facets and how to rank passages?",  
},  
{  
  "role": "user",  
  "content": ""Here is an example:
```

```
Idea: Enhance topic model evaluation by incorporating anomaly detection machine learning  
↳ techniques. The goal is to improve topic model evaluation by identifying and  
↳ flagging anomalies within topic distributions that may indicate incoherence or  
↳ redundancy. This approach provides a more robust evaluation framework that detects  
↳ subtle inconsistencies that traditional metrics might miss. The effectiveness of  
↳ this integrated evaluation method would be assessed through a systematic comparison  
↳ and meta-analysis of different topic models, ensuring comprehensive and reliable  
↳ evaluation outcomes.
```

Key Facets to Look for in Passages:

- Application Domain: Topic modeling and evaluation.
- Purpose: Improving topic model evaluation by detecting anomalies indicating  
→ incoherence or redundancy.
- Mechanism: Incorporating anomaly detection machine learning techniques into topic  
→ model evaluation.
- Method: Identifying and flagging anomalies within topic distributions.
- Evaluation: Systematic comparison and meta-analysis of different topic models to  
→ assess effectiveness.

Passages:

- [0] An Enhanced BERTopic Framework and Algorithm for Improving Topic Coherence and  
→ Diversity
- [1] Evaluation of Unsupervised Anomaly Detection Methods in Sentiment Mining
- [2] LDA\_RAD: A Spam Review Detection Method Based on Topic Model and Reviewer Anomaly  
→ Degree
- [3] Apples to Apples: A Systematic Evaluation of Topic Models
- [4] Machine Learning Approach for Anomaly-Based Intrusion Detection Systems Using  
→ Isolation Forest Model and Support Vector Machine
- [5] OCTIS: Comparing and Optimizing Topic Models is Simple!
- [6] Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling
- [7] An Exhaustive Review on State-of-the-art Techniques for Anomaly Detection on  
→ Attributed Networks
- [8] Topic Modeling Revisited: New Evidence on Algorithm Performance and Quality Metrics
- [9] A Robust Bayesian Probabilistic Matrix Factorization Model for Collaborative  
→ Filtering Recommender Systems Based on User Anomaly Rating Behavior Detection

Ranking:

[2] > [1] > [5] > [3] > [0] > [8] > [6] > [7] > [4] > [9]

""",

},

```

{
  "role": "user",
  "content": ""Here is another example:
**Idea:** Develop a system that uses sentiment analysis to detect political bias in news
↪ articles. The system will analyze language patterns and sentiments to identify
↪ biased reporting, and will be validated using a dataset of news articles over the
↪ past decade.

**Key Facets:**

- **Application Domain:** News articles analysis.
- **Purpose:** Detecting political bias through sentiment analysis.
- **Mechanism:** Analyzing language patterns and sentiments.
- **Method:** Using a dataset of news articles from the past decade.
- **Evaluation:** Validated through analysis of historical data.

**Passages:**

- **[0]** **Detecting Political Bias in News Articles Using Sentiment Analysis**
- **[1]** **Sentiment Analysis of Social Media Posts for Political Trends**
- **[2]** **Machine Learning Techniques for Stock Market Prediction**

**Ranking:**

[0] > [1] > [2]
"",
},
{
  "role": "assistant",
  "content": "Okay, please provide the passages which I have to compare with **Query**
↪ Idea",
},
]

```

### B.3.5 Prompt to generate more novel idea suggestions for an idea classified as “not novel.”

```
def promptMoreNovelIdea(idea, papers, facets, topic):
    prompt = f"""
    INITIAL IDEA:
    Short Version: {idea["shortened_text"]}
    Long Version: {idea["text"]}

    PRIOR WORK FROM WHICH MORE NOVEL IDEA SHOULD DIFFERENTIATE:"""
    ind = 1
    for t in list(papers.keys()):
        prompt += f"""
        Paper {ind}"""
        if "title" in papers[t]:
            prompt += f"""
            Title: {papers[t]["title"]}
            Abstract: {papers[t]["abstract"]}"""
            if "introduction" in papers[t]:
                prompt += f"""
                Background: {papers[t]["introduction"]}"""
            ind += 1
        prompt += f"""

    SUMMARY OF WHY INITIAL IDEA IS NOT NOVEL COMPARED TO PRIOR WORK:
    {idea["novelty_review"]}

    FACETS AVAILABLE TO ADD TO INITIAL IDEA:
    """
    for facet_key in facets:
        facet = facets[facet_key]
        if facet["id"] not in idea["facets"]:
            prompt += f"""
            {facet["text"]} (ID: "+facet["id"]+)"""
    """
```

```

prompt += f"""
FACETS AVAILABLE TO REMOVE FROM INITIAL IDEA:
"""
for facet_id in idea["facets"]:
    prompt += f"""
{facets[facet_id]["text"]+" (ID: "+facets[facet_id]["id"]+"")}
"""
prompt += f"""

```

#### INSTRUCTIONS:

You are ScientistGPT, an intelligent assistant that helps researchers come up with  
↪ understandable, relevant, specific, feasible, and novel research ideas.

First, read the above prior work and summary of why the initial idea is not novel compared to  
↪ the prior work.

That way, you will know what has already been done in related research.

Second, provide 3 options for ideas that are more novel and useful than the initial idea.

Each more novel/useful idea should utilize one new facet and remove one existing facet from the  
↪ initial idea.

The summary of facets available to add and to remove are above.

Make each novel/useful idea as specific as possible but describe it in 100 to 150 words.

#### FACET REQUIREMENTS:

The purpose/mechanism/evaluation IDs should be written VERBATIM from what is provided.

Do NOT make up facet IDs that are not explicitly given above.

#### IDEA REQUIREMENTS:

##### 1. Understandability::

1a. Each idea should be logical.

1b. Each idea should be grammatically correct.

1c. Each idea should be self-contained and should not require researchers to have read the

↪ provided papers. For example, saying "Use Tool X" is not self-contained because the

↪ researcher might not have read the relevant paper to know what Tool X is. Instead, you could

↪ say "Use a tool to do Y."

1d. If an aspect of the idea may not be familiar to someone studying computer science broadly,  
↔ make sure the idea describes the aspect enough for them to understand it.

## 2. Relevance::

2a. Each idea should be at least somewhat relevant to the overarching ideation topic: {topic}.

↔ This means you may or may not have to adjust the purpose to be relevant to the ideation  
↔ topic. For example, if the purpose is "to support human-AI creativity" and the topic is  
↔ "scientific idea generation", then the purpose is relevant enough to the topic and does NOT  
↔ need to be adjusted. On the other hand, if the selected purpose is "to support bird  
↔ identification" and the topic is "scientific idea generation", you might adjust the purpose  
↔ to "to support scientific idea identification" or "to support identifying different types of  
↔ scientific ideas." That said, it is important to keep as much of the purpose's meaning as  
↔ possible if/when adjusting it to be relevant to the ideation topic.

2b. Ideas should NOT refer to analogies.

## 3. Specificity::

3a. Each idea should be described in 100 to 150 words (expanded version 200 to 250 words). 90%  
↔ of the idea should focus on describing how the mechanism will be used to address the purpose  
↔ in a novel manner.

3b. Each idea should be as specific as possible given the word limit.

3c. Each idea should be focused and precise. For example, if the idea involves a visualization  
↔ feature, describe WHAT kind of visualization it is and HOW it will be used.

3d. Each idea should be specific such that the reader understands how the idea would be  
↔ implemented and what direction it would take.

3e. Each idea should be very specific in describing HOW the mechanism will be implemented to  
↔ address the purpose in a novel manner. For instance, an idea saying to 'apply a faceted  
↔ representation to clinical data, creating a multidimensional profile of each patient,  
↔ integrating medical history, genetic information, lifestyle factors, and current health  
↔ data' is not novel because prior work has already looked into creating multidimensional  
↔ patient profiles. If the idea provided a more specific, imaginative description of HOW to  
↔ create the multidimensional profiles and HOW to apply it to clinical decision-making, then  
↔ it could be novel.

3f. Each idea should be specific enough that it is significantly different from your prior ideas.

## 4. Feasibility::

- 4a. A research lab with moderate resources should be able to carry out each idea.
- 4b. Each idea should make sense scientifically and be grounded in the summary of prior work and  
↪ papers noted below.
- 4c. Each idea's purpose and mechanism should work well together. If necessary, adapt the purpose  
↪ or mechanism to work well with the other. For example, if the overarching ideation topic is  
↪ "scientific ideation", the purpose is "to support culinary ideation", and the mechanism is  
↪ "scientific research graphs", you might adjust the purpose to "to support scientific  
↪ ideation." Meanwhile, if the overarching ideation topic is "culinary ideation", the purpose  
↪ is "to support culinary ideation", and the mechanism is "scientific research graphs", you  
↪ might adjust the mechanism to "culinary recipe graphs." That said, it is important to keep  
↪ as much of the original purpose and mechanism as possible if/when adjusting them to make  
↪ sense together.
- 4d. Each idea's proposed evaluation should not contradict the rest of the idea. For example, an  
↪ idea talking about supporting healthcare workers that has a user study evaluation should  
↪ propose a "user study with healthcare workers," rather than a "user study with teachers" or a  
↪ "user study with engineers."

5. Novelty::

- 5a. Each idea should be unique as well as ingenious, imaginative, or surprising.
- 5b. Each idea should NOT have been covered in prior work. In other words, each idea should be  
↪ SIGNIFICANTLY different from prior work, including but not limited to the papers below and  
↪ the summary of prior work below.
- 5c. For an idea to be novel, it is not enough for the idea not to have been thoroughly covered  
↪ in prior work. It must NEVER have been covered in prior work.
- 5d. Each idea should NOT be an obvious extension of prior work but rather CLEARLY DIFFERENT from  
↪ prior work. For instance, an idea that simply says to 'implement continuous AI support to  
↪ facilitate discovery of scholars by dynamically updating and refining researcher profiles  
↪ based on real-time academic contributions' is not novel. The idea is not novel because prior  
↪ work has investigated facilitating discovery of scholars through AI support, and making the  
↪ AI support continuous is an obvious extension. It is not ingenious, imaginative, or  
↪ surprising. If the idea were more specific and presented a surprising method describing HOW  
↪ to make this AI support very timely (e.g., utilizing academic Twitter trends), that would  
↪ make the idea novel.

FORMAT FOR ANSWER (It is very important that you follow this answer format!):

```

1. Removed Purpose: <one removed purpose text here>
Removed Purpose ID: <removed purpose ID here (e.g., [purpose-string-123456789])>
Added Purpose: <one added purpose text here>
Added Purpose ID: <added purpose ID here (e.g., [purpose-string-012345678])>
More Novel Idea: <more novel idea here>
Why Idea is More Novel: <sentence on why idea is novel here>
Why Idea is Useful: <sentence on why idea is useful here>
2. Removed Mechanism: <one removed mechanism text here>
Removed Mechanism ID: <removed mechanism ID here (e.g., [mechanism-string-34523235])>
Added Mechanism: <one added mechanism text here>
Added Mechanism ID: <added mechanism ID here (e.g., [mechanism-string-634902049])>
More Novel Idea: <more novel idea here>
Why Idea is More Novel: <sentence on why idea is novel here>
Why Idea is Useful: <sentence on why idea is useful here>
3. Removed Evaluation: <one removed evaluation text here>
Removed Evaluation ID: <removed evaluation ID here (e.g., [evaluation-string-123456789])>
Added Evaluation: <one added evaluation text here>
Added Evaluation ID: <added evaluation ID here (e.g., [evaluation-string-348950684])>
More Novel Idea: <more novel idea here>
Why Idea is More Novel: <sentence on why idea is novel here>
Why Idea is Useful: <sentence on why idea is useful here>

ANSWER:
"""
return prompt

```

## B.4 Input Papers for Each Ideation Topic

- Human-AI Collaboration in Art

1. LumiMood: A Creativity Support Tool for Designing the Mood of a 3D Scene [141]
2. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with

Dreamsheets [4]

3. Algorithmic Ways of Seeing: Using Object Detection to Facilitate Art Exploration [129]

- AI Tools for Education

1. VIVID: Human-AI Collaborative Authoring of Vicarious Dialogues from Lecture Videos [31]
2. Scientific and Fantastical: Creating Immersive, Culturally Relevant Learning Experiences with Augmented Reality and Large Language Models [27]
3. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation [113]

- Dealing with LLM Hallucinations

1. Deductive Closure Training of Language Models for Coherence, Accuracy, and Updatability [3]
2. Self-Alignment for Factuality: Mitigating Hallucinations in LLMs via Self-Evaluation [193]
3. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them [153]

- LLM Explainability

1. LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction [179]

2. TextGenSHAP: Scalable Post-Hoc Explanations in Text Generation with Long Documents [50]
3. Digital Socrates: Evaluating LLMs through Explanation Critiques [64]

## **B.5 Expanded Version of Sample Ideas from Table 5.2**

The expanded version of each sample idea in Table 5.2 may be found in Table B.1 below.

Situation	Topic	Purp	Mech	Eval	Custom Inst	Expanded Version of Idea
Initial, Treatment	LLM explainability	to enhance medical information retrieval (near)	sequential knowledge-guided prompting (paper input)	medical question-answering benchmarks (near, unselected)	None	The proposed system focuses on developing a robust medical information retrieval platform using sequential knowledge-guided prompting to simulate diagnostic pathways. By integrating electronic health records (EHRs) with established clinical guidelines, the system leverages prompts to dynamically identify and follow potential diagnostic routes, thereby enhancing the precision and relevance of information retrieval. The process begins by extracting key medical factors from EHRs and matching these with clinical pathway prompts to simulate diagnostic reasoning. This approach allows the system to adjust retrieval strategies based on real-time patient data, ensuring contextually relevant information is provided to healthcare professionals. The system's effectiveness will be rigorously evaluated using medical question-answering benchmarks, which will assess the accuracy and pertinence of information retrieval against standard medical queries. This evaluation not only measures the system's performance but also its ability to improve clinical decision-making processes. Ultimately, the research aims to bridge gaps in current medical retrieval systems by offering a more transparent and context-aware tool for healthcare providers, significantly enhancing the explainability and utility of LLMs in the medical domain.

P-or-M, Treatment	human-AI collaboration in art	to enhance designer-ai collaboration (very near, selected)	residual learning framework (far)	empirical performance analysis (far, unselected)	I want to know if designer want to work with ai	In this novel approach, we apply a residual learning framework to AI design tools, treating designer feedback as residuals to refine specific design elements such as color schemes and layout structures. This framework allows designers to provide targeted feedback easily, which the AI tool uses to iteratively improve its suggestions. The process starts with the AI generating an initial design, which the designer can annotate with feedback. This feedback is then incorporated as residuals, prompting the AI to make adjustments that align more closely with the designer's vision. The feedback loop is engineered to be highly intuitive, integrating seamlessly into the designer's workflow, thus fostering a more collaborative and efficient design process. An empirical performance analysis will evaluate the effectiveness of this approach, measuring improvements in design quality and collaboration efficiency.
----------------------	-------------------------------------	--	--	---	--	--

P-and-M, Treatment	LLM explain- ability	to explain text gener- ation out- puts (paper input)	explanation logic ver- ifier (own input)	quantitative and qual- itative analysis (paper input, selected)	Use different verification approaches from NLI, math, program to ensure the explanation follows a logi- cal consistency	Create an innovative platform that enhances the validation of LLM-generated mathematical explanations by integrating a mathematical logic verifier with visual proof representations. The platform will employ the verifier to ensure explanations adhere to logical principles, assessing each step for accuracy. Simultaneously, it will generate graph-based visual proofs, providing intuitive visual cues for each stage of the reasoning process. These representations will guide users through complex explanations, making abstract concepts more accessible and engaging. The system will also feature interactive elements that allow users to explore different parts of the proofs, deepening their understanding. This approach will be evaluated through comprehensive quantitative and qualitative analysis, assessing improvements in explanation accuracy, user comprehension, and satisfaction. By combining rigorous logic verification with engaging visual aids, the research aims to significantly improve the interpretability and reliability of LLM-generated mathematical content, fostering user trust and learning.
-----------------------	----------------------------	--	---	---	---	--

P-and-M, Treatment	AI tools for edu- cation	to per- sonalize vocabulary learning (very near)	ai- generated virtual speakers (very near)	participant feedback analysis (very near, unselected)	make the idea more focused and specific	This innovative platform leverages AI-generated virtual speakers to create personalized vocabulary learning experiences. Users upload personal images, and the AI analyzes these images to identify objects, settings, and potential vocabulary links. The virtual speakers then generate dialogue scenarios around these elements, allowing learners to engage with new vocabulary in a personally relevant context. For example, an image of a beach vacation might lead to dialogues about seaside activities, weather, and local culture, enriching the vocabulary acquisition process. This personalized approach not only aids retention by linking words to familiar contexts but also enhances engagement through interactive AI-driven dialogues. Participant feedback analysis will assess the platform's effectiveness in improving vocabulary acquisition and learner satisfaction, guiding further refinements. By integrating AI with personalized content, this research aims to transform vocabulary learning into a more engaging and contextually meaningful experience.
-----------------------	--------------------------------	--	--	---	---	--

Baseline	dealing with LLM hallucinations	N/A	N/A	N/A	Can the reward model in Self-Knowledge tuning be changed to incorporate more fine-grained feedback (e.g. from Halogen like metrics) [all 3 papers given to LLM]	This research focuses on developing an advanced reward model for Self-Knowledge Tuning by incorporating specific metrics from the HALoGEN framework. The core of this model lies in leveraging atomic unit verification to dissect LLM outputs into fundamental components, evaluating each against reliable knowledge bases. By integrating HALoGEN's error classifications, the model can provide targeted feedback that distinguishes between different types of hallucinations, such as those arising from incorrect recollections, faulty training data, or fabrications. This nuanced feedback is expected to enhance the calibration of LLMs, reducing the frequency of factual inaccuracies. The model's efficacy will be tested across diverse domains, including programming and scientific attribution, to ensure broad applicability. A comprehensive validation framework will be established, involving both automated and human evaluations, to rigorously measure improvements in model accuracy and coherence. The outcome of this research could significantly advance the development of more reliable and trustworthy LLMs, providing a foundation for future enhancements in self-supervised learning techniques.
----------	---------------------------------	-----	-----	-----	---	--

Baseline	AI tools for education	N/A	N/A	N/A	Generate ideas that use non-traditional mediums for education, so avoid using textbooks or other material commonly found in classrooms. [all 3 papers given to LLM]	This AR-enabled mobile application is designed to engage children with their local environment while providing a robust educational platform about biodiversity and ecology. The AI component will analyze user interactions and location data to curate a personalized educational journey. For a child in an urban environment, the app might focus on urban wildlife and plants, providing detailed information, conservation tips, and interactive quizzes tailored to their locality. For children in rural settings, it might explore more diverse ecosystems, encouraging them to document and learn about various species..
----------	------------------------	-----	-----	-----	---	---

Table B.1: A sample of participants’ favorite ideas from the user study, presenting the expanded version of each idea rather than the shorter version. The treatment ideas are generated with the different methods of the **Faceted Idea Generator** module, depending on whether the participant selected any purposes or mechanisms.

## B.6 Insights from LLM Prompt Optimization Using TextGrad

In our experiments with TextGrad, we investigated how specific prompt instructions influence an LLM’s ability to classify the novelty of an idea. Figures B.1, B.2, and B.3 present the accuracy of various prompts optimized with TextGrad on our dataset (train=25, validation = 10, test = 32). Prompts with both non-zero and zero validation accuracy included various instructions for evaluating the novelty of ideas, such as assessing the uniqueness of methods and their comparison to existing research. Through this prompt optimization process, we observed interesting ways in which LLMs may evaluate novelty, like considering historical context, frequency

of similar studies, comparative analysis with existing works, examining arguments for both novel and non-novel perspectives. However, prompts without these specific instructions also influenced accuracy, suggesting the complexity of novelty evaluation with LLMs. Notably, some prompts with similar instructions showed different performance on validation data. For example, both prompt 3 (accuracy = 0) and prompt 9 (accuracy = 0.6) include instructions to evaluate if the idea introduces unique methodologies, and how it compares to existing work. However, the difference in their performance suggests that subtle variations in wording and instruction framing can significantly impact the classification performance. It remains unclear why certain prompts perform better despite having similar instructions.

Our analysis highlights the LLM's sensitivity to prompt design when assessing novelty of an idea. Even minor variations in wording and structure can lead to substantial performance changes, emphasizing the need for careful prompt engineering and well-chosen in-context examples to guide the LLM for idea novelty evaluation.

Figure B.1: Performance trends of test accuracy across prompts during prompt optimization with TextGRAD. Highlighted text shows unique instructions used to evaluate the novelty of ideas. The final test accuracy was 0.78125, showing that none of the optimized prompts (1 to 12) improved over the original.

Prompt Number	Validation Accuracy	Prompt Text
<b>Starting Prompt</b>	0.8	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel' Respond in two lines: Review: <This is the review you will generate after comparing idea with related papers.> Classification:<novel or not novel label according to your review/rationale>
<b>1</b>	0	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. A novel idea introduces a unique, groundbreaking concept or approach not previously covered in the literature. A not novel idea reiterates or slightly modifies existing research. Consider the historical context and frequency of similar studies when making your classification. Focus on identifying unique, groundbreaking elements that differentiate the idea from existing research. Do not rely solely on keywords or the mention of a controlled setting to determine novelty. If the classification is ambiguous, indicate uncertainty and suggest a human review. Respond with only the classification label: 'novel' or 'not novel'.
<b>2</b>	0.7	You are a classifier. Classify the research idea as 'novel' or 'not novel' based on the related papers. Provide a brief review and directly state the classification. Review: <This is the review you will generate after comparing the idea with related papers.> Classification: <novel or not novel label according to your review/rationale> Example: Review: The idea is unique as it combines adaptive interfaces with AI explanations, which is not covered in the provided papers. Classification: novel.
<b>3</b>	0	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines: Review: Provide a detailed review comparing the idea with related papers. Include specific examples and details from the referenced papers to justify your classification. Highlight both similarities and differences between the proposed idea and existing methodologies. Ensure your review is concise and precise, focusing on the main arguments. Classification: Based on the review, classify the idea as 'novel' or 'not novel' according to the following criteria: - Uniqueness of the approach - Originality of the application - Novelty of the results Provide specific references or evidence from the papers mentioned to support your classification. Use assertive language to clearly convey your classification.

Figure B.2: contd. TextGrad Prompt Optimisation.

Prompt Number	Validation Accuracy	Prompt Text
4	0.7	You are a classifier. Classify the research idea as 'novel' or 'not novel' based on the related papers. Respond concisely: Review: <brief review> Classification: <novel or not novel>
5	0	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines: 1. <b>Review</b> : Provide a detailed review comparing the idea with related papers. Include specific examples and details from the referenced papers to justify your classification. Highlight both similarities and differences between the proposed idea and existing methodologies. Ensure your review is concise and precise, focusing on the main arguments. 2. <b>Classification</b> : Based on the review, classify the idea as 'novel' or 'not novel' according to the following criteria: - Uniqueness of the approach - Originality of the application - Novelty of the results - Provide specific references or evidence from the papers mentioned to support your classification. - Use assertive language to clearly convey your classification.
7	0.3	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel' based on the following criteria: 1. <b>Definition of Novelty</b> : - A 'novel' idea introduces a unique, groundbreaking concept, methodology, or significant improvement over existing work. - A 'not novel' idea closely aligns with existing work without significant innovation. 2. <b>Contextual Instructions</b> : - If the idea involves common methodologies or well-known techniques, explicitly mention these aspects in your review. - Consider the historical context and frequency of similar studies when making your classification. 3. <b>Comparative Analysis</b> : - Compare the proposed idea with existing systems or technologies mentioned in the related papers. Highlight similarities to justify the classification. 4. <b>Evidence and Examples</b> : - Provide specific examples or evidence from the related papers that demonstrate the lack of novelty. 5. <b>Structured Format</b> : - Use a structured format with sections such as 'Introduction,' 'Evaluation Criteria,' 'Comparative Analysis,' and 'Conclusion' to present your review. Respond in two lines: Review: <This is the review you will generate after comparing the idea with related papers.> Classification: <novel or not novel label according to your review/rationale>
8	0	You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines: 1. Review: Provide a detailed review that includes specific references to the related papers, highlighting similarities and differences. Include direct quotes or specific sections from the related papers that support your classification. 2. Classification: Use precise terminology to classify the idea as 'novel' or 'not novel' based on your review/rationale. Avoid vague terms and be specific in your justification.

Figure B.3: contd. TextGrad Prompt Optimisation.

Prompt Number	Validation Accuracy	Prompt Text
9	0.6	<p>You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'.</p> <p>Definition of Novel: An idea is 'novel' if it introduces a new concept, methodology, or significant improvement that is not already well-documented in the provided papers.</p> <p>Criteria for Evaluation:</p> <ol style="list-style-type: none"> <li>Uniqueness of the approach.</li> <li>Combination of elements.</li> <li>Presence of similar frameworks in the literature.</li> </ol> <p>Instructions:</p> <ol style="list-style-type: none"> <li>Compare the proposed idea's methodology, scope, and application with those described in the related papers.</li> <li>Highlight specific aspects of the idea and compare them with the related papers.</li> <li>Provide a detailed review based on the comparison.</li> <li>Conclude with a classification of 'novel' or 'not novel' based on this comparison.</li> </ol> <p>Respond in the following format: Review: &lt;This is the review you will generate after comparing the idea with related papers.&gt; Classification: &lt;novel or not novel label according to your review/rationale&gt;</p>
10	0.6	<p>You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. A novel idea introduces a unique, groundbreaking concept or approach not previously covered in the literature. A not novel idea reiterates or slightly modifies existing research. Consider the historical context and frequency of similar studies when making your classification.</p> <p>Respond in two lines: Review: &lt;one-sentence review&gt; Classification: &lt;novel or not novel&gt;</p> <p>Ensure your response is concise and uses simple language. Avoid unnecessary details.</p>
11	0.6	<p>You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond with: 1. Review: Provide a concise review in no more than two sentences, comparing the idea with related papers. Ensure your review includes a clear rationale for why the idea is classified as 'novel' or 'not novel'. Avoid using uncertain terms like 'appears' or 'seems'. 2. Classification: Use the term 'novel' or 'not novel' consistently based on your review.</p> <p>Example:</p> <p>Review: The proposed idea of developing a human-centric explainable AI system is novel because it uniquely combines explainable AI techniques with iterative improvement through human feedback and predictive models. Classification: novel</p>
12	0.5	<p>You are a classifier. Given a research idea and related papers, classify the idea as 'novel' or 'not novel'. Respond in two lines:</p> <p>Review: &lt;Provide a detailed review comparing the idea with related papers, including specific examples and reasons for your classification. Mention existing tools or research that cover similar capabilities.&gt; Classification: &lt;novel or not novel label according to your review/rationale. Use the term 'novel' consistently in both your review and classification. Ensure your response is detailed yet concise, avoiding unnecessary verbosity.&gt;</p>

## B.7 Expert-Labeled Examples

Table B.2 presents expert-labeled examples from the annotation study.

Table B.2: **Expert-labeled examples from annotation study**

---

### Example 1

---

**Idea:** Develop a **natural language processing classifier designed to improve scientific paper revisions** by automatically identifying and categorizing reviewer comments that are most likely to lead to substantial and actionable revisions. The system would be trained on a **manually-labeled dataset analysis** of scientific review comments and the corresponding paper edits, leveraging features such as linguistic cues, sentiment, and comment specificity to predict the likelihood of a comment being acted upon. This classifier could then be used to prioritize reviewer feedback, helping authors focus on the most impactful suggestions first.

---

---

**Most Relevant Papers:**

1. [ARIES: A Corpus of Scientific Paper Edits Made in Response to Peer Reviews](#)
2. [Can large language models provide useful feedback on research papers?](#)
3. [A Dataset of Peer Reviews \(PeerRead\): Collection, Insights and NLP Applications](#)
4. [arXivEdits: Understanding the Human Revision Process in Scientific Writing](#)
5. [Characterizing Text Revisions to Better Support Collaborative](#)
6. [Can We Automate Scientific Reviewing?](#)
7. [DeepReviewer: Collaborative Grammar & Innovation Neural Network for Automatic Paper Review](#)
8. [Aspect-based Sentiment Analysis of Scientific Reviews](#)
9. [Aspect-based sentiment analysis of online peer reviews and prediction of paper acceptance](#)
10. [ReviVal: Towards Automatically Evaluating the Informativeness of Peer Reviews](#)

---

**Reasoning:** The idea is **novel** because it uniquely focuses on prioritizing reviewer comments for actionable revisions, which is not explicitly addressed in ARIES[1] or other related works like ReviVal [10].

---

## Example 2

---

**Idea:** Develop a **systematic review-based framework** designed to **align LLM evaluation with human preferences**, ensuring that evaluation criteria are continuously refined based on comprehensive reviews of user feedback and emerging model behaviors. This framework will utilize **content analysis of user interactions and feedback** to identify patterns and areas of improvement. The effectiveness of this framework will be assessed through a **qualitative study** involving iterative cycles of user feedback and criteria refinement.

---

---

**Most Relevant Papers:**

1. [EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria](#)
2. [Humanely: Human evaluation of LLM yield, using a novel web-based evaluation tool](#)
3. [Evaluation of Code Generation for Simulating Participant Behavior in Experience Sampling Method by Iterative In-Context Learning of a Large Language Model](#)
4. [Human-Centered Evaluation and Auditing of Language Models](#)
5. [Aligning Model Evaluations with Human Preferences: Mitigating Token Count Bias in Language Model Assessments](#)
6. [Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences](#)
7. [Human-Centered Design Recommendations for LLM-as-a-judge](#)
8. [CheckEval: Robust Evaluation Framework using Large Language Model via Checklist](#)
9. [Discovering Language Model Behaviors with Model-Written Evaluations](#)
10. [Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models](#)

---

**Reasoning:** The idea is **not novel** because it closely resembles existing frameworks like EvalLM[1] and HumanELY[2], which already align LLM evaluations with human preferences using user-defined criteria and human feedback.

---

## B.8 Comparing Scideator’s Novelty Checker to Baselines

Figures [B.4](#), [B.5](#), and [B.6](#) present information related to comparing SCIDEATOR’s novelty checker to baselines.

Figure B.4: Two example ideas used as the basis for comparison in subsequent figures, evaluated by SCIDEATOR (Ours), AI Scientist, and AI Researcher.

---

#### Example 1

---

**IDEA:** Develop a system that uses a *faceted author representation* of digital learning resource (DLR) creators based on their educational materials and inferred teaching personas. This system would aim to *support ubiquitous learning* by helping learners discover novel educators and materials that offer innovative perspectives. *Usability testing of learning resources* would be conducted to ensure the system enhances the learning experience by balancing relevance and novelty, thus boosting the accessibility and discoverability of diverse educational content.

**MOST RELEVANT PAPERS:**

- (1) Bursting Scientific Filter Bubbles: Boosting Innovation via Novel Author Discovery
- (2) Bridger: Toward Bursting Scientific Filter Bubbles and Boosting Innovation via Novel Author Discovery
- (3) Novel Algorithmic Recommendation Engine for Diverse Content Discovery
- (4) ComLittee: Literature Discovery with Personal Elected Author Committees
- (5) Explanations in Open User Models for Personalized Information Exploration
- (6) AMiner: Mining Deep Knowledge from Big Scholar Data
- (7) Similar researcher search in academic environments
- (8) VeTo-web: A Recommendation Tool for the Expansion of Sets of Scholars
- (9) From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks
- (10) DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration

**EXPERT-LABELED CLASS:** Novel

---

#### Example 2

---

**IDEA:** Develop a **Co-Creative Interaction Framework** for LLM-assisted evaluations to **align llm evaluation with human preferences**. This framework will map out detailed interaction models between human evaluators and LLMs, including turn-taking, communication protocols, and iterative feedback loops. The framework will be evaluated through **qualitative study** involving user testing and thematic analysis of evaluator experiences, capturing how well the framework supports alignment with human preferences and identifying areas for iterative improvement.

**MOST RELEVANT PAPERS:**

- (1) EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria
- (2) Human-Centered Design Recommendations for LLM-as-a-judge
- (3) CheckEval: Robust Evaluation Framework using Large Language Model via Checklist
- (4) Human-Centered Evaluation and Auditing of Language Models
- (5) Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models
- (6) Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences
- (7) Evaluation of Code Generation for Simulating Participant Behavior in Experience Sampling Method by Iterative In-Context Learning of a Large Language Model
- (8) Discovering Language Model Behaviors with Model-Written Evaluations
- (9) LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models
- (10) Do LLMs Meet the Needs of Software Tutorial Writers? Opportunities and Design Implications

**EXPERT-LABELED CLASS:** Novel

---

Figure B.5: Reviews corresponding to idea 1 in Figure B.4.

Method	Review
SCIDEATOR (Ours)	The idea is <b>novel</b> because it introduces the concept of faceted author representation specifically for digital learning resource (DLR) creators, focusing on their educational materials and inferred teaching personas to support ubiquitous learning. This approach is distinct from related works like Bridger[1][2], which focus on scientific author discovery, and ComLittee[4], which supports authorcentric exploration in research. The proposed system aims to enhance the discoverability of diverse educational content, balancing relevance and novelty, which is not explicitly addressed in the related papers.
AI Scientist	<p>THOUGHT: The proposed research presents an innovative approach to language models. The rough goal is to train a model to predict the next character in a sequence in order to be able to capture the semantic meaning of the entire sequence, which can be used in language models. This approach is not uncommon, though, as there are many areas where we must generate predictions of the next likely character, which can be used in order to understand the entire meaning of the text. This model aims to improve upon traditional models by locally training the model on the use of characters rather than words. Of particular note is the fact that the model uses a scaled gradient descent method which changes how the weights are updated, a method that is very new and, thus, is a key point of innovation. While that is extremely novel, a model that works on the character level is not. The primary new technique here is the use of a momentum of the gradient, which is used to improve learning. The model uses a technique similar to many current models. It is possible that it could be used with different parameters, though, to learn more or to learn in a more effective way than other models. This approach has been suggested many times over the years, but it is promising to pursue. It is, however, a novel approach to the existing technology.</p> <p>However, we must consider the existing literature and how it might overlap with this proposition. I will now proceed to search for related publications that may be relevant in order to establish a definitive argument regarding the novelty of the proposed model. RESPONSE: "json\n\n"Query": "Character-level language models"\n\n" I believe this should identify any previous work or existing lines of research that are most similar to the proposed model. This will confirm whether this is actually a novel concept. My primary concern is understanding if this is a novel or an existing work, as this is a model that could potentially be used for many purposes. There is a good chance that it might have already been proposed. I hope that this search will bring more clarity to the argument.</p> <p><b>If decision is not made, default is: not novel.</b></p>
AI Researcher	<p>Proposed Research Problem and Approach:</p> <p>Project Proposal: The research problem is the limited accessibility and discoverability of innovative educational content due to the lack of effective representation and utilization of diverse teaching personas and educational materials of DLR creators. The approach involves developing a faceted author representation based on creators' educational materials and inferred teaching personas to enhance the discovery of novel educational content.</p> <p>Paper Abstract: The research problem is the isolation of scientific research and information overload, which limit awareness and hinder innovation. The approach involves constructing a faceted representation of authors from their papers and inferred personas to balance relevance and novelty in discovering new scholars and their work.</p> <p>Summary and Differences: - Project Proposal Summary: Enhancing the discovery of digital learning resources by developing a faceted author representation based on educational materials and teaching personas. - Paper Abstract Summary: Facilitating the discovery of new scholars and their work by constructing a faceted author representation from scientific papers and inferred personas to balance relevance and novelty. - Differences: - The project proposal focuses on digital learning resources and educational content, aiming to improve the learning experience. - The paper abstract focuses on scientific research and scholars, aiming to boost innovation and bridge diverse scientific communities.</p> <p>No</p> <p>..... The comparison with other papers has been truncated due to space limitations. All papers in the most relevant set are marked as "Binary Judgment: No", which implies it should not be cited as similar work to the idea.</p> <p><b>Classification: novel</b></p>

Figure B.6: Reviews corresponding to idea 2 in Figure B.4.

Method	Review
<b>SCIDEATOR (Ours)</b>	<p>The idea is <b>novel</b> because it introduces a CoCreative Interaction Framework specifically designed to align LLM evaluation with human preferences through detailed interaction models, including turntaking, communication protocols, and iterative feedback loops. This approach is distinct from related works like EvalLM[1], which focuses on iterative refinement of prompts, and EvaluLLM[2], which leverages LLMs as customizable judges but does not map out detailed interaction models. Additionally, CheckEval[3] and Prometheus 2[5] focus on evaluation frameworks and opensource evaluators, respectively, without the cocreative interaction aspect. The proposed framework’s emphasis on qualitative studies and thematic analysis to capture evaluator experiences further differentiates it from existing methods.</p>
<b>AI Scientist</b>	<p>THOUGHT:\n The proposed idea involves designing a framework for co-creative interaction between human evaluators and Large Language Models (LLMs), focusing on aspects like turn-taking and communication protocols. It also emphasizes qualitative study and thematic analysis. This is a very specific angle that combines human-computer interaction and AI evaluation. While several existing papers discuss LLM evaluation frameworks, human-centered design, and human involvement in evaluation (e.g., EvalLM, EvaluLLM, and other papers from the results), the specific focus on co-creative interaction frameworks with detailed interaction models and iterative feedback loops may not have been explored thoroughly.\n\n Additionally, the idea looks at aligning evaluations with human preferences which is a nuanced approach. Therefore, a more targeted query to ascertain if any existing literature discusses similar co-creative frameworks or detailed interaction models involving human evaluators and LLMs is necessary. The current results do not seem to cover this specific angle in depth.</p> <p><b>Decision made: novel</b></p> <p>RESPONSE:\n“json\n\n "Query": "Co-creative interaction framework LLM evaluation"\n\n“</p>
<b>AI Researcher</b>	<p>Proposed Research Problem and Approach:</p> <ul style="list-style-type: none"> <li>- Project Proposal: - Problem: Current LLM evaluation methods often fail to align with human preferences. - Approach: Develop a Co-Creative Interaction Framework (CCIF) that includes detailed interaction models between human evaluators and LLMs, facilitating turn-taking, communication protocols, and iterative feedback loops to better align LLM evaluations with human preferences.</li> <li>- Paper Abstract: - Problem: Developers must iteratively revise prompts by evaluating outputs to diagnose weaknesses, which is effort-intensive and context-specific. - Approach: Present EvalLM, an interactive system for iteratively refining prompts by evaluating multiple outputs on user-defined criteria, using an LLM-based evaluator to provide feedback and improve prompts.</li> <li>- Differences: - The project proposal focuses on aligning LLM evaluations with human preferences through a co-creative interaction framework involving detailed interaction models. - The paper abstract focuses on refining LLM prompts through an interactive system that evaluates outputs based on user-defined criteria.</li> </ul> <p>No</p> <p>..... The comparison with other papers has been truncated due to space limitations. All papers in the most relevant set are marked as "Binary Judgment: No", which implies it <b>should not be cited as similar work to the idea.</b></p> <p><b>Classification: novel</b></p>

## **B.9 Experimental Setup Details for Comparing Scideator Novelty Checker to Baselines**

The best-performing in-context setting used 15 novel and non-novel examples with seed 100. For the OpenReview examples, the best setup involved 5 idea-review pairs. For DSPy we used 2 bootstrapped examples, and trained both DSPy and TextGRAD for 12 prompt iterations.