

Human-AI Mechanisms for Scholarly Knowledge Synthesis

Raymond Fok

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
James Fogarty

Program Authorized to Offer Degree:
Computer Science and Engineering

© Copyright 2025

Raymond Fok

University of Washington

Abstract

Human-AI Mechanisms for Scholarly Knowledge Synthesis

Raymond Fok

Chair of the Supervisory Committee:

Daniel S. Weld

Computer Science and Engineering

The rapid growth in research publication presents significant challenges for researchers trying to stay current with relevant literature. This information overload risks duplicated research efforts, limits innovation, and hinders scholarly growth as researchers struggle to identify gaps and contextualize their work within the existing body of knowledge. While recent advances in AI offer potential solutions to support researchers in engaging with the literature, effectively integrating these techniques into existing research processes requires a richer understanding of how these AI mechanisms can both enhance and threaten researchers' cognitive processes.

In this thesis, I examine how new human-AI tools can support researchers in reading, reviewing, and synthesizing research literature more effectively to inform and drive future scientific endeavors. I propose *structured mediation* as a design framework for human-AI collaboration in scholarly sensemaking that creates intermediate, interpretable representations to enhance researcher capabilities while preserving agency and enabling verification. This framework operates through three core mechanisms: *discovery* (AI-provided information scent that guides attention and exploration), *iteration* (collaborative refinement of data gathering and conceptual frameworks), and *verification* (maintaining clear provenance to source material through structured representations).

I present three systems that demonstrate this framework, each exploring interactive, verifiable AI-

powered mechanisms that serve as cognitive scaffolds to support researchers' existing literature review processes. First, I explore a new interaction paradigm in recursive, just-in-time expansion of paper abstracts that enables researchers to retrieve clarifying information from full texts of papers during triage with new information presented in-context and attributed to the source text. Then, within the context of a paper's full text, I demonstrate how AI-suggested faceted highlights can aid researchers in more rapidly gathering relevant information by judiciously directing their attention while reading. Finally, intermediate structured representations progressively transform large sets of unstructured research papers into familiar sensemaking schemas, such as tables, hierarchies, and textual summaries, that support steerable and transparent exploration, comparison, and synthesis of information from papers at scale.

Through lab and deployment user studies, I demonstrate how these interactive, human-AI systems offer valuable cognitive scaffolds that augment researchers' abilities to navigate across, drill down into, and make sense of dense, complex information within research papers. The result is more effective and efficient literature navigation and synthesis that enhances rather than replaces human judgment in the research process. This work offers guidance for designing mixed-initiative systems that foster effective human-AI collaboration across the review and synthesis of scientific knowledge.

Acknowledgments

I am truly grateful to the wonderful people I have met and worked with throughout my time in graduate school. To my advisor, Dan Weld, your mentorship allowed me to reach this point of my journey. Thank you for your patience and generosity, even when the wit and significance of your advice was often lost on me in the moment. To my committee members and mentors: Amy Zhang, for your feedback and advice that helped shape my perspective of human-centered research. Andrew Head, whose endless pool of thought-provoking questions helped me think more critically about my research and craft an attention to detail; James Fogarty, for being a great advisor while I was still figuring out what I wanted to do, and for our early conversations about life and research that bring me back to simpler times; Gary Hsieh, for your positivity throughout the final years. I also reflect fondly on my mentors at Google and Adobe Research, who helped open my eyes to the world beyond academia. Alexa Siu, your research guidance, friendship, and kindness over the past few years have been invaluable.

Thank you also to my collaborators and friends in the Lab for Human-AI Interaction: Marissa Radensky, Jim Chen, Sherry Wu, Gagan Bansal, Benjamin Charles Germain Lee, Joyce Zhou, and Madeleine Grunden-McLaughlin, and to my collaborators across UW and Ai2 who all contributed to shaping my PhD experience: Joseph Chee Chang, Jonathan Bragg, Pao Siangliulue, Tal August, Luca Soldaini, Kyle Lo, Matt Latzke, Lucy Lu Wang, Mingyuan Zhong, Anne Spencer Ross, Jacob Wobbrock, Benjamin Newman, Yoonjoo Lee, and Hita Kambhamettu.

To my wife and best friend, Liwei Jiang, this dissertation would not exist without you. Your quirkiness and kindness were constant sources of joy in times I found myself struggling. Lastly, to my parents and family—your support and trust throughout this long and uncertain journey were everything to me.

Contents

1	Introduction	19
1.1	Thesis Overview	24
2	Background and Related Work	29
2.1	Information Foraging and Sensemaking	29
2.2	Systems Supporting Scholarly Sensemaking	31
3	Qlarify: Enhancing In-Situ Information Retrieval with Expandable Abstracts	35
3.1	Introduction	36
3.2	Recursively Expandable Abstracts	38
3.3	System Design	44
3.4	Study 1: Interview Study	50
3.5	Study 2: Field Deployment Study	54
3.6	Study 3: Comparative Evaluation	55
3.7	Evaluation of Generated Expansions	60
3.8	Related Work	61
3.9	Discussion	64
3.10	Conclusion	67
4	Scim: Guiding Attention in Paper Skimming with Faceted Highlights	69
4.1	Introduction	70
4.2	Background and Related Work	72

4.3	Design Goals	76
4.4	SCIM User Interface	80
4.5	Technical Implementation	83
4.6	Study 1: In-Lab Usability Study	86
4.7	Study 2: Longitudinal Diary Study	90
4.8	Study 3: Deployment of SCIM at Production Scale	97
4.9	Discussion	104
4.10	Conclusion	108
5	Challenges and Opportunities for Human-AI Scholarly Synthesis	109
5.1	Introduction	110
5.2	Background	111
5.3	Methods	113
5.4	Findings	115
5.5	Discussion	125
5.6	Conclusion	128
6	DimInd: Scaffolding Synthesis with Intermediate Structured Representations	129
6.1	Introduction	130
6.2	Design Goals	132
6.3	System Design	134
6.4	Evaluation	142
6.5	Discussion	152
6.6	Conclusion	156
7	Discussion and Conclusion	157
7.1	Evolving Models of Human-Information Interaction in the Age of AI	157
7.2	Design Implications: Lessons from Structured Mediation	159
7.3	Designing for Metacognitive Support in AI-Assisted Knowledge Work	160
7.4	Concluding Remarks	163

A	Additional Details for Qlarify	201
A.1	Evaluation Details	201
A.2	Example Qlarify Outputs	203
A.3	LLM Prompts	204
B	Interview Protocol for Living Syntheses Study	209
B.1	Interview Protocol	209
C	Additional Details for DimInd	211
C.1	User Study Details	211
C.2	Creating Paper Collections	213
C.3	DIMIND LLM Prompts	214

List of Figures

1.1	We address information overload in scholarly knowledge workflows through three human-AI systems supporting different stages of literature engagement. The three structured mediation mechanisms of discovery, iteration, and verification are embodied in each system, demonstrating how AI can enhance scholarly sensemaking while preserving researcher agency.	21
1.2	Mapping of human-AI systems explored in this dissertation to structured mediation mechanisms of discovery, iteration, and verification.	22
1.3	The QLARIFY system explores an <i>expandable abstract</i> interaction paradigm. Static abstracts are transformed into fluid text artifacts through AI-suggested expandable entities and user-driven queries.	24
1.4	The SCIM system scaffolds paper reading with AI-powered faceted highlights organized by rhetorical function (e.g., Objective, Method, Finding).	25
1.5	The DIMIND system supports scholarly sensemaking through <i>multi-level information displays</i> in which large paper collections are transformed and progressively disclosed through intermediate structured representations—tables, taxonomies, and narrative syntheses.	26
3.1	<i>Recursively expandable abstracts</i> present a novel mixed-initiative interaction technique, leveraging large language models to enable a low-cost, on-demand, and fluid expansion of static abstracts with information retrieved from full papers.	36
3.2	Recursively expandable abstracts allows users to retrieve clarifying information from a broader <i>expansion context</i> (e.g., the full paper) in response to dynamic information needs, forming <i>expansions</i> that grow the abstract fluidly.	41

3.3	The design space for an expandable abstracts interaction paradigm, formed through an iterative design process and validation with external scholars. Alternatives we explored in QLARIFY are highlighted in gray, and those included in the final system are outlined in red.	43
3.4	Recursively expandable paper abstracts with attribution in QLARIFY. Expansions are created on-demand by highlighting text in the abstract or selecting an AI-suggested expandable entity (A), revealing a question palette (B). Selecting a question in the palette prompts an LLM to retrieve relevant clarifying information, presented as a fluid expansion within the abstract (C). Users can drill-down to see evidence for a response in a paper excerpt (D) and within the full paper context itself (E).	44
3.5	QLARIFY’s system architecture. Each paper is first processed (Document Processing) and initial expandable entities are extracted from the abstract (Expandable Entity Extraction). When a user asks a question for an expandable entity, QLARIFY uses a retrieval-augmented generation approach to form a response and retrieve attribution (Question Answering). QLARIFY then suggests expandable entities within the response, allowing recursive expansions.	47
3.6	Distribution of participants’ self-reported ratings within each condition in the comparative evaluation. Participants in the QLARIFY condition felt more satisfied with their exploration, more confident in retrieving relevant information from the full paper, more motivated to explore deeply, and a greater desire to use in the future. See Appendix A.1.2 for the precise wording used in the survey questions.	56
4.1	SCIM is an intelligent reading interface for skimming scientific papers. To help readers develop a broad overview of content in a paper, SCIM intelligently highlights passages (A). The passages are colorized to indicate the rhetorical role of the passage, i.e., whether it describes the research’s objectives, novelty, methods, and results. Highlights are distributed throughout the text to support a holistic skim. Readers can request additional (or fewer) highlights by using paragraph-local (B) and paper-wide (C) controls. To understand where to find information of a certain kind, readers can glance at highlight markers in the scroll bar (D). Readers can also collect an overview of the paper by reviewing highlighted passages in a sidebar (E).	70

4.2	Our formative research revealed that intelligent highlights need to do more than pointing readers to important content. They should also be <i>well-distributed</i> throughout a paper (D3, D5) and steer readers towards <i>diverse</i> content types (D2).	78
4.3	SCIM classifies and highlights four facets of information commonly found in papers: OBJECTIVE, NOVELTY, METHOD, and RESULT. These facets aim to surface specific kinds of paper content that align with common skimming goals identified in formative research, reflecting design guideline D2. Above, we show example passages matching each of the four facets. The passages appear in Wu et al.’s scientific paper, “Cross-document Misinformation Detection based on Event Graph Reasoning” [Wu et al., 2022b].	81
4.4	Overview of SCIM’s paper processing pipeline. SCIM takes as input a scientific paper in PDF format and then parses it into sentences with bounding boxes and other accompanying metadata. It then classifies sentences into one of four facets using a large language model fine-tuned via a data programming approach. SCIM chooses which highlights to show by reconciling prediction weights with heuristics controlling highlight distribution and readers’ preferences.	83
4.5	(Left) Time taken by participants to complete each information seeking question in Tasks 2 and 3 of the in-lab usability study. Overall, participants answered questions more quickly with SCIM than with a standard (baseline) document reader. (Right) Participants’ subjective responses regarding the ease of completing information seeking tasks with SCIM compared to a standard document reader, and whether they found SCIM’s highlights distracting.	89
4.6	We integrated SCIM’s interface augmentations into a publicly-available reading interface for scientific papers, Semantic Reader, refined into the following set of features: margin flags which can be turned on or off (A); multi-faceted highlights (Goal, Method, and Results), the density and visual characteristics of which can be customized (B); user-controlled settings for further customizability of the interface (C); screen-reader accessible snippets containing the extractive text from each highlight (D).	99

4.7	When users in the deployment study first encountered a paper with skimming features available, the capabilities were introduced with a popup window in the upper right, directing the user to controls for opening the sidebar with highlights and customizable settings.	103
6.1	We present an LLM-assisted workflow aimed at scaffolding literature review over large paper collections, and instantiate it in a prototype system, DIMIND. Users can interactively construct and explore four successive <i>structured representations</i> of literature information: a <i>paper collection</i> listing papers and metadata, a <i>literature review table</i> with columns that render relevant evidence snippets from paper full texts along defined facets, a <i>facet taxonomy</i> that organizes faceted information into a higher-level conceptual overview, and a <i>facet synthesis</i> that provides a controllable faceted summary across the collection.	130
6.2	Columns can be added to the literature review table in two ways: A) <i>User-defined</i> columns precisely specify a faceted information need and allow additional context for steering LLM assistance; B) <i>System-suggested</i> columns offer collection-aware recommendations for columns that can be added with a single click.	134
6.3	In DIMIND, users review large paper collections by navigating and analyzing information across various structured representations. Each cell in the literature review table is a snippet of faceted information from a paper (evidence snippet). Clicking on a snippet shows a popover with additional detail (evidence summary), with a button that can further open the paper PDF in an integrated paper reader with attributed paragraphs highlighted (evidence source). Faceted columns are transformed into distinct hierarchical taxonomies (facet taxonomy), which can be explored, refined, and used to controllably generate a narrative summary with citations (facet synthesis).	135
6.4	The facet taxonomy. Each category shows the number of included papers (A). Users can manually refine the taxonomy through drag-and-drop interactions (B) or add additional categories (C). If at least one category is selected, the taxonomy can be summarized into prose (D).	138

6.5	Selecting specific categories in the facet taxonomy: 1) highlights cells for the included papers in the literature review table, allowing users to quickly delineate between and browse the selected (and not selected) papers; 2) controls the structure and papers included in the generated summary.	139
6.6	Users can view additional detail while exploring the synthesized representations: 1) Clicking an evidence snippet in the facet taxonomy shows the full evidence summary; 2) Clicking a citation in the facet synthesis shows an in-situ citation card. From either, users can click <i>See in Table</i> to scroll directly to the corresponding row in the literature review table.	140
6.7	Participants' post-task ratings (7-point Likert scale) of literature review utility, control, information verifiability, and confidence across DIMIND and BASELINE.	145
A.1	Comparison of systems in the three conditions in Study 3: Comparative Evaluation.	202
A.2	Updated QLARIFY system interface for the comparative evaluation. Refinements included allowing users to edit the AI-suggested question (A), moving the button for viewing attributed evidence in the PDF to the right margin (B), and adding a button to remove a generated expansion (C).	203
A.3	Time-graph of user actions during task completion in the comparative evaluation.	204
A.4	QLARIFY's user interface with expandable entities underlined and no expansions created (abstract shown is from [Kang et al., 2023]).	205
C.1	Paper collections in DIMIND can be created interactively from a research question or search query.	213

List of Tables

1.1	AI-enabled intermediate representations supporting scholarly sensemaking.	23
3.1	Participants in our formative study and selected questions they asked while reading scientific paper abstracts.	39
4.1	A summary of usage of SCIM’s features during the diary study. Notably, most readers used most features at least once. Use of the highlight controls varied widely, with some readers using them heavily (P6, P10, P12), and others less often (P1–3, P6, P7–9). All readers used the highlight browser on multiple occasions. Faceted highlights are omitted because we could not collect log data as to when readers looked at highlights.	92
5.1	Participant biographies and self-reported research expertise.	113
5.2	Opportunities for supporting various updates to narrative literature reviews.	124
6.1	The survey topics and taxonomy dimensions used in the tasks, and their source papers. . . .	142
A.1	Examples of LLM-generated expansions for the three static questions in Qlarify (from [Kang et al., 2023]; abstract in Figure A.4).	206
A.2	Examples of Qlarify’s LLM-generated expandable entities, suggested questions, and expansions. All examples are actual system output for the paper <i>Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking</i> [Kang et al., 2023] (abstract in Figure A.4).	207

- C.1 User study participants. **YoE** refers to years of experience conducting scholarly research.
LLM-General refers to how frequently LLM applications are used for everyday tasks.
LLM-Research refers to how frequently LLM applications are used for research activities. . . 212
- C.2 Post-task survey questions and their corresponding labels (rated on 7-point Likert scale). . . 213

Chapter 1

Introduction

Scientific progress is fundamentally a collective and cumulative endeavor, built upon the synthesis of prior knowledge. However, the modern research landscape is increasingly strained by systemic pressures that have intensified the need to produce publications [Hanson et al., 2024; Rawat and Meena, 2014], posing challenges for comprehensive synthesis. Scientific output is growing at approximately 4% annually, with the total number of publications doubling every 12 years [Landhuis, 2016; Bornmann and Mutz, 2015]; in 2024 alone, arXiv received over 240,000 submissions, while major publishers like PLOS ONE have published hundreds of thousands of articles over the past decade. This rapid expansion has resulted in an era of unprecedented growth in scientific knowledge, and researchers increasingly report feeling overwhelmed by the volume of available knowledge—a classic information overload problem [Landhuis, 2016; Shahrzadi et al., 2024; Morris, 2023a; AAAS, 2016]. Over the past two decades, the development of scholarly search engines (e.g., Google Scholar, Semantic Scholar) and the shift toward digital libraries (e.g., ACM Digital Library, PubMed, IEEE Xplore), online preprint repositories (e.g., arXiv, bioRxiv), and open-access journals (e.g., PLOS ONE) have democratized access to millions of papers. Researchers today face more new publications in a month than their historical counterparts encountered in a year. However, the tools that support researchers in making sense of this unprecedented volume of information remain largely underdeveloped, focusing often on paper retrieval rather than scaffolding the comprehension of each paper and synthesis of knowledge across related papers. Researchers therefore face a growing burden of knowledge management and must allocate their limited cognitive resources judiciously.

To mitigate these challenges, research has explored how computational approaches, such as large language models (LLMs) from BERT [Devlin et al., 2019] to OpenAI’s GPT series of models [Brown et al., 2020a; Achiam et al., 2023] offer new opportunities to support and augment existing cognitive capabilities in literature engagement. These models can be trained and fine-tuned to answer questions, generate summaries, and identify patterns across vast corpora of text. However, despite their revolutionary potential as scholarly research aids, we find ourselves at a critical juncture in their application. The question is not whether AI can help researchers navigate the information landscape, but how these new tools can extend, rather than threaten, human scholarly reasoning. The challenge therefore lies not only in improving the performance of these models in automating our current scholarly activities, but also in developing the appropriate affordances and interactions that simultaneously preserve critical thinking and domain expertise while supporting researchers’ capabilities across core scholarly functions, e.g., to make sense of the literature, discover patterns, and derive insights. This tension between researchers’ management of information overload and their cognitive limitations, along with growth in AI capabilities, present both a challenge and an emergent opportunity to develop human-AI systems that improve how researchers engage with scientific knowledge. Toward this end, one key challenge lies in designing interactions that leverage the efficiency of AI integration while maintaining the serendipity, interpretability, and control required of scholarly work. I argue that moving beyond simple automation tools toward collaborative systems that scaffold sensemaking while preserving researcher agency requires a structured form of mediation between AI capabilities and human scholarly reasoning.

Thesis: Structured mediation—a design framework that uses AI-assisted intermediate, interactive, and transparent representations to scaffold scholarly sensemaking—can improve how researchers gather and synthesize knowledge across scientific literature. I develop three interactive systems that instantiate its core mechanisms and evaluate them through mixed-methods studies with researchers engaged in scholarly reading and synthesis. Together, these studies demonstrate how structured representations can increase the efficiency of information seeking, deepen comprehension, foster exploration, and improve perceived satisfaction while reading individual papers, and support the iterative refinement of conceptual structures across paper collections.

As a design framework, structured mediation provides guiding principles and mechanisms that shape

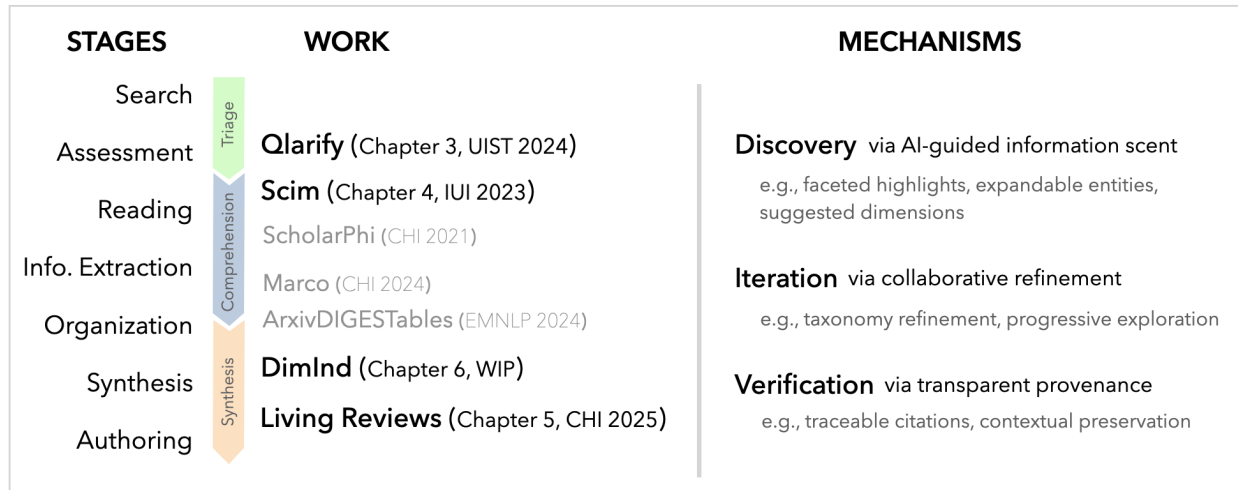


Figure 1.1: We address information overload in scholarly knowledge workflows through three human-AI systems supporting different stages of literature engagement. The three structured mediation mechanisms of discovery, iteration, and verification are embodied in each system, demonstrating how AI can enhance scholarly sensemaking while preserving researcher agency.

how systems may be designed and evaluated, building on information foraging and sensemaking theories by introducing intermediate representations that align AI assistance with natural processes of scholarly reasoning. Information foraging models describe how analysts follow information scent cues to efficiently navigate large knowledge spaces [Pirolli and Card, 1999], while classical sensemaking theories in HCI characterize the iterative process of building mental models from foraged information fragments [Pirolli and Card, 2005; Russell et al., 1993; Klein et al., 2007]. Reifying structured mediation involves designing technologies that enrich information scent, scaffold the development and refinement of sensemaking frames [Klein et al., 2007], and preserve researcher agency and control through efficient access to source evidence.

The framework addresses a fundamental tension in AI-assisted scholarly work: how to leverage computational support without undermining the critical evaluation and deep understanding that define scholarly expertise. Rather than replacing researcher judgment, the proposed approach uses malleable yet informative representations that enhance existing sensemaking capabilities while maintaining easy verification and refinement of these co-created representations. To operationalize structured mediation in system design, I identify three core mechanisms that reflect natural processes in literature-based scholarly work:

- The *Discovery* mechanism leverages AI to provide enriched information scent that guides attention and exploration across literature. Rather than dictating conclusions, AI suggests where to look and

Qlarify Chapter 3	AI-suggested expandable entities	Mixed-initiative, threaded expansions	Deep context linking
Scim Chapter 4	AI-suggested faceted highlights	Limited iteration; Highlight facet/density adjustments	Linked representations; in-context overlays
DimInd Chapter 6	AI-suggested dimensions and themes	User-steered, multi-level representations	Deep context linking; linked representations
	Discovery Guiding attention	Iteration Refinement and co-creation	Verification Traceability and oversight

Figure 1.2: Mapping of human-AI systems explored in this dissertation to structured mediation mechanisms of discovery, iteration, and verification.

what patterns might be worth investigating, enhancing traditional browsing behaviors while preserving researcher autonomy in interpretation.

- The *Iteration* mechanism enables collaborative refinement of both data gathering and conceptual frameworks. Researchers can progressively build understanding by refining AI-suggested structures, expanding content based on emerging hypotheses, and customizing organizational patterns to match their analytical goals. This mirrors the natural sensemaking cycle of questioning assumptions and seeking confirmatory evidence, but with AI handling tedious extraction and organization tasks.
- The *Verification* mechanism maintains provenance of source text through structured representations that preserve traceability. Users can trace AI-extracted information back to original evidence, enabling the verification behaviors that distinguish expertise-driven activities from simple information consumption. This preserves the evidential rigor critical to scholarly work.

To explore and validate this framework, I developed three human-AI systems that instantiate structured mediation across different stages of literature engagement: QLARIFY, which emphasizes discovery and iteration at the abstract level; SCIM, which highlights discovery and verification during full-text reading; and DIMIND, which integrates all three mechanisms to support collection-level synthesis. QLARIFY transforms

System	Representations
QLARIFY	Threaded fluid-text expansions with concise LLM-generated clarifying answers embedded in abstracts, each grounded in and linked to the full text
SCIM	Faceted highlight overlays with AI-suggested passages categorized by rhetorical function, presented in situ with sidebar and scrollbar overviews
DIMIND	Multi-level structured displays including literature review tables, hierarchical thematic taxonomies, and narrative syntheses, all linked to underlying evidence

Table 1.1: AI-enabled intermediate representations supporting scholarly sensemaking.

static abstracts into dynamic artifacts by enabling progressive exploration through AI-suggested expandable entities, with recursive expansions supporting iterative refinement and deep linking maintaining provenance to source material. SCIM augments full papers with faceted highlights that surface content by rhetorical function, guiding discovery while preserving reading context and enabling verification through traceable links. Finally, DIMIND restructures paper collections into multi-level representations—such as review tables, taxonomies, and narrative syntheses—while maintaining interactive links between abstraction levels and underlying evidence. Together, these systems demonstrate how structured mediation functions as both a conceptual framework and a practical design strategy across literature review stages, from individual paper comprehension to collection-level synthesis.

In sum, this dissertation makes three primary contributions:

- Structured mediation, a design framework for human-AI scholarly sensemaking, which leverages AI-generated intermediate representations to guide discovery, support iterative refinement, and facilitate verification of AI assistance.
- Three novel system designs and interaction techniques that demonstrate how AI can support researcher capabilities across different stages of literature engagement—from progressive abstract exploration to faceted paper reading to multi-level literature synthesis.
- Empirical findings from user evaluations that characterize how researchers engage with AI-assisted literature tools, including their benefits, limitations, and design implications.

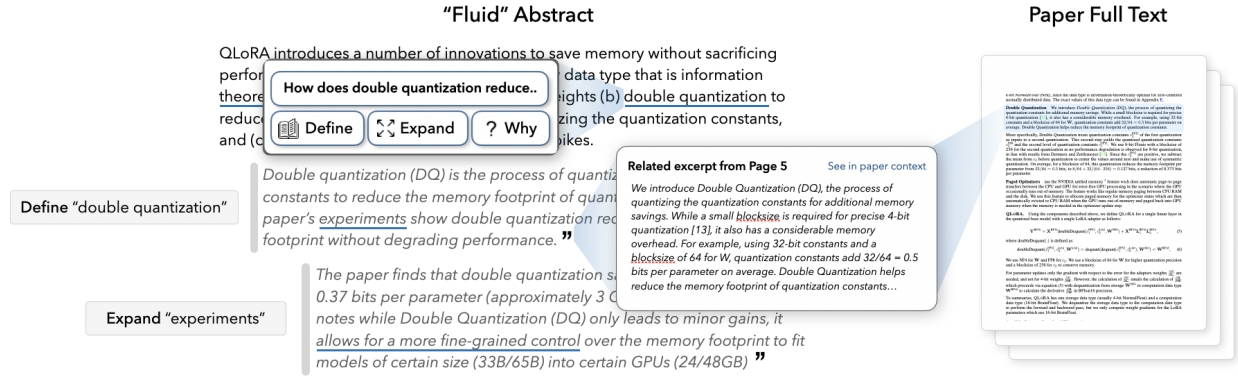


Figure 1.3: The QLARIFY system explores an *expandable abstract* interaction paradigm. Static abstracts are transformed into fluid text artifacts through AI-suggested expandable entities and user-driven queries.

1.1 Thesis Overview

Chapter 2 first situates this dissertation within the context of prior research across sensemaking, information foraging, and mixed-initiative interfaces, with a particular focus on the vast design space of human-AI tools and techniques introduced for enhancing scholarly engagement with the scientific literature. Chapters 3-6 then describe systems and studies that explore novel mixed-initiative methods for supporting scholarly information foraging, sensemaking, and synthesis. In these chapters, I present insights from formative studies with scholars, detailed descriptions of the design and implementation of each system, and findings from mixed-methods evaluative studies.

In Chapter 3, I present **QLARIFY**, an augmented reading interface for scientific paper abstracts that explores a fluid text interface paradigm for interacting with summaries in which users can progressively explore full papers by expanding paper abstracts in a just-in-time and in-situ manner with clarifying information (Figure 1.3). This mixed-initiative approach combines user-defined queries and AI-suggested expandable entities to bridge the informational gap between static abstracts and full papers, allowing scholars to retrieve clarifying information while linking to underlying evidence via interaction. This paradigm allows a natural, controllable, and verifiable exploration that reduces the cognitive cost of transitioning from abstract-level triage to deeper paper understanding.

In Chapter 4, I present **SCIM**, an augmented reading interface for full text scientific papers that introduces faceted highlights suggested by a fine-tuned language model to steer reader attention toward and

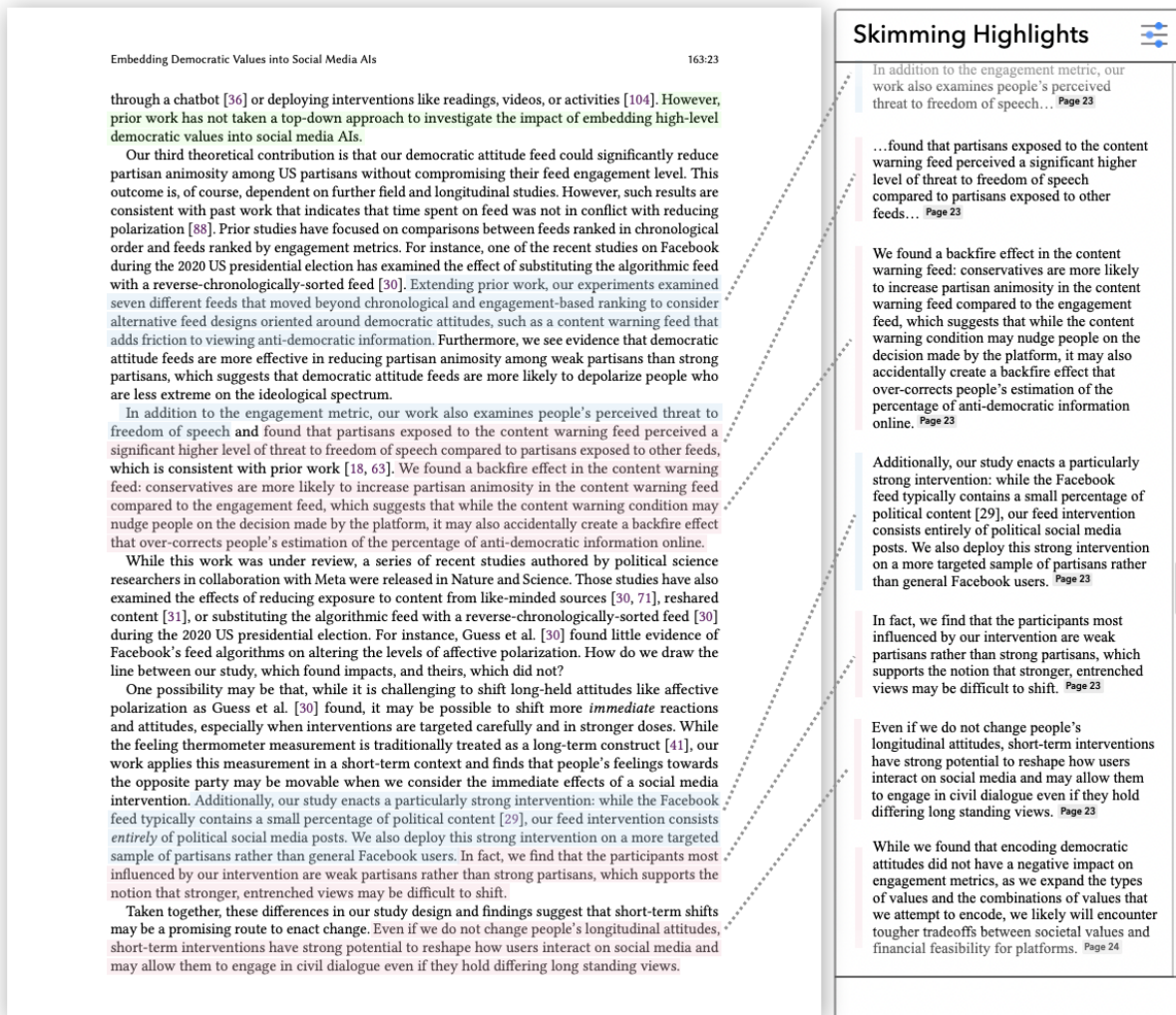


Figure 1.4: The SCIM system scaffolds paper reading with AI-powered faceted highlights organized by rhetorical function (e.g., Objective, Method, Finding).

reduce the cost of access to high-value content within papers (Figure 1.4). The design decision to augment papers with suggested highlights reflects a lightweight, embedded approach to AI assistance that minimizes disruption to native reading workflows. Compared to external summaries, this approach offers lower risk of semantic distortion, preserves context and structural richness, and enables users to easily “recover” from model errors by simply reading surrounding text. SCIM further provides mechanisms for gaining a rapid overview of model-suggested highlights, through a sidebar of faceted highlights that allow interactive linking to the respective content in the PDF and scrollbar annotations that offer a visual road-map of the distribu-

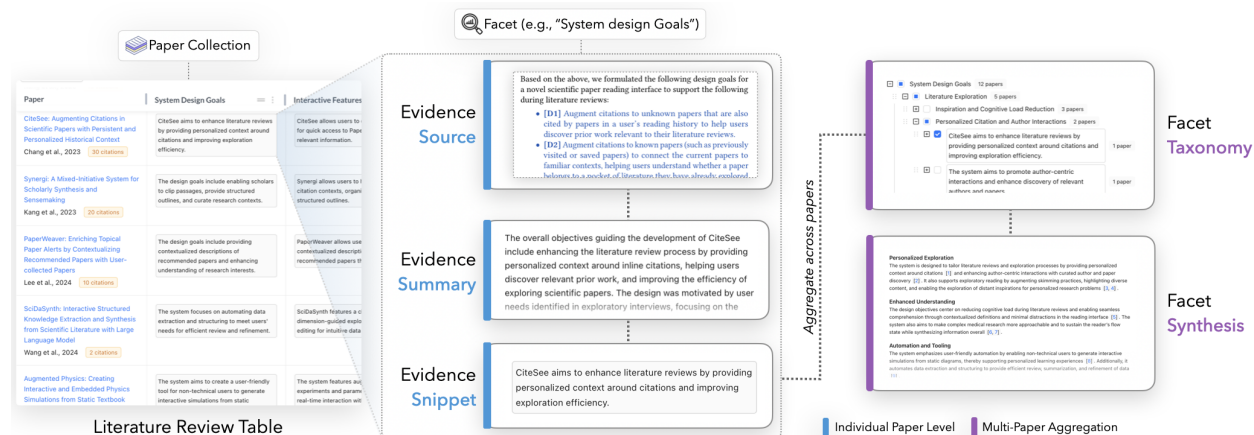


Figure 1.5: The DIMIND system supports scholarly sensemaking through *multi-level information displays* in which large paper collections are transformed and progressively disclosed through intermediate structured representations—tables, taxonomies, and narrative syntheses.

tion of salient content within a paper. Finally, controls over the density, presentation, and type of faceted highlights allow users to customize how these AI cognitive aids are integrated into their reading process.

In Chapter 5, I present findings from a retrospective interview study that examines the practices and challenges researchers face when authoring and updating semi-systematic literature reviews in computing research (i.e., survey papers). Through in-depth interviews with 11 survey paper authors, this study reveals the varied methodologies used in creating narrative reviews and identifies three key types of updates essential for maintaining current literature syntheses: empirical updates involving evidence and examples, structural updates to taxonomies and organization, and interpretive updates to synthesis and framing. The findings highlight the tension between the recognized value of keeping surveys current and the practical barriers that make continuous updating infeasible against current academic incentives. The analysis suggests specific leverage points where AI assistance can address the most burdensome aspects of literature synthesis—particularly the challenge of organizing large paper collections into meaningful structured representations such as tables and taxonomies, and maintaining comprehensive coverage as new research emerges—while preserving the expertise-driven interpretive work that researchers identified as the core scholarly value of their review processes.

In Chapter 6, guided by these findings, I present **DIMIND**, a system that implements a mixed-initiative literature review workflow enabled by language models. This work explores *multi-level information display*

mechanisms for AI-assisted literature review (Figure 1.5), transforming unstructured literature into structured representations for sensemaking. The system guides users through successive information transformations—from individual papers to faceted literature review tables, to hierarchical taxonomies of concepts, to controllable narrative syntheses—while maintaining clear provenance. This approach naturally engages users in defining custom facets and refining the AI-generated taxonomies, allowing researchers to steer the automatically suggested organization of information while preserving their ability to verify the validity of system outputs by diving into evidence within individual papers. By providing multiple levels of information abstraction that users can fluidly navigate between, the approach allows scalable literature understanding that balances improved information extraction efficiency with scholarly control, supporting both top-down, collection-level exploration and bottom-up, paper-level investigation.

Finally, Chapter 7 summarizes the contributions of this dissertation, the design lessons from the systems presented in the previous chapters and their evaluations, and broader implications for the design of mixed-initiative systems to facilitate effective scholar-AI collaboration in scholarly sensemaking and synthesis.

Chapter 2

Background and Related Work

This dissertation draws on research in information foraging, sensemaking, and human-AI systems supporting scholarly literature engagement. Here I provide a broad overview of these areas to situate the contributions of this work; additional related work is discussed in subsequent chapters where relevant (§3.8,§4.2).

2.1 Information Foraging and Sensemaking

A rich history of cognitive science and human-computer interaction (HCI) research examines how people seek, navigate, and make sense of complex information. Pirolli and Card [2005]’s notional model of sensemaking proposes two loops of activity that underlie processes such as reasoning, problem solving, and decision making. The first, a *foraging* loop, describes how people often engage in activities such as “seeking information, searching and filtering it, and reading and extracting information” from a broader information corpus relevant to their goals. The second, a *sensemaking* loop, captures how individuals iteratively develop and refine organizational structures, or *schemas*, both within mental models and externalized through intermediate representations or artifacts that support analysis and synthesis. Notably, this foraging loop is informed by an earlier information foraging theory, which conceptualizes information-seeking as a cost-benefit process: people aim to maximize the value of information gained relative to the effort required to obtain it. Moreover, people attend to and judge the reliability and utility of an information source through the detection of *information scent*, guided by cues such as visual markers, headers, or keywords. Foraging theory lends an understanding toward how users choose which resources to pursue, when to shift their

attention, and how they navigate across information-dense environments.

Klein et al. [2007] further propose a *data-frame theory of sensemaking* that postulates elements are explained when fitted into structures linking them to other elements. Specifically, a *frame* denotes an explanatory structure that defines entities by describing their relationship to other entities. For instance, a *story* explains the chronology of events and their causal relationships, a *map* describes distances, directions, landmarks, and routes, or a *plan* describes a sequence of actions; these demonstrate just a handful of abstractions that a frame may represent. This theory centers on the interplay between foraged data and existing mental frames, emphasizing how people use existing frames to interpret information, while also updating, elaborating, or reframing those structures in response to discoveries or surprises. There is a tight interdependence of frames and data: data help elicit and shape frames, while frames guide attention to relevant data. This iterative, exploratory nature of sensemaking points to a need for AI systems that offer useful abstractions which adapt to dynamic user understanding rather than providing static outputs.

Prior work has explored numerous systems to support these interleaved foraging and sensemaking processes by scaffolding the cognitive labor involved. Across these systems, one common thread involves transforming internal knowledge into externalized representations—structures that offload memory and cognitive processing demands onto visual or textual formats [Zhang, 1997; Norman, 1988]. Representations might include tables (e.g., decision matrices, literature review tables, qualitative coding spreadsheets), graphs (e.g., mind maps, argument maps, conceptual diagrams), taxonomies (e.g., ontologies, trees, hierarchies), timelines, or document-centric annotations (e.g., highlights, comments, tags). Such representations (or frames) not only reduce cognitive load, but also provide artifacts that can anchor subsequent interpretation, comparison, and synthesis.

The creation of a comparison table or thematic taxonomy can be seen as a concrete instantiation of a frame; similarly, highlights or facet tags embedded within a text serve to make information scent explicit and navigable, enabling more efficient foraging. Many interactive systems draw from these strategies, supporting sensemaking contexts across exploratory online search [Nguyen et al., 2016; Hearst and Deger, 2013; Hahn et al., 2018; Chang et al., 2019; Ramos et al., 2022; Suh et al., 2023], programming [Liu et al., 2019, 2021, 2022a], mobile information exploration [Swearngin et al., 2021], or scholarly literature review [Zhang et al., 2008; Kang et al., 2022, 2023; Palani et al., 2023]. Early examples such as *Scatter/*

Gather[Cutting et al., 1992], *SenseMaker*[Baldonado and Winograd, 1997], and faceted browsing [Hearst and Stoica, 2009; schraefel et al., 2006] aimed to assist information foraging over online web pages through better clustering and filtering. Subsequent work has extended support across the full sensemaking loop—helping users not only find and reduce costs of saving relevant information [Kittur et al., 2013; Ramos et al., 2022; schraefel et al., 2002; Dontcheva et al., 2007; Liu et al., 2022b] but also construct and externalize evolving schemas [Chang et al., 2020, 2019; Chen and Wang, 2017; Dontcheva et al., 2007; Chen et al., 2014; Newman et al., 2024].

In scholarly sensemaking, these interleaved processes of data foraging and frame construction point to a broad design space for technological interventions that can mediate between human sensemaking and the scale and complexity of scientific literature. As constructing these frames manually is costly, this creates an opportunity for AI assistance that can automatically generate structured representations while preserving user agency over their refinement and interpretation—an approach we call “structured mediation”. Specifically, in this dissertation we focus on three types of conceptual representations at different information granularities: *expandable abstracts*, which bridge static abstracts and full-text exploration, adapting to users’ interpretive frames; *faceted highlights* which enable frame-guided attention by making relevant evidence patches immediately visible and reducing the cost of scanning dense texts; and *multi-level information displays* that externalize conceptual frames and support iterative synthesis across large literature collections.

2.2 Systems Supporting Scholarly Sensemaking

The rapid growth in scientific publishing [Foundation, 2021; Chu and Evans, 2021], shift from print to digital publication [Liu, 2005; Tenopir et al., 2009], and growing popularity of preprint distributions has led to a problem of information overload [Eppler and Mengis, 2004] that has increasingly challenged how scholars can explore, attend to, and make sense of relevant literature [Sultanum et al., 2020; Mysore et al., 2023]. Reviewing the literature is therefore an opportunistic and dynamic sensemaking process, and one that is often tedious, time-consuming, and fragmented [Soufan et al., 2022; Mysore et al., 2023]. As a result, numerous academic and commercial systems have been proposed to aid scholars throughout this process.

One line of work has focused on fully automating the literature review process [Susnjak et al., 2025; Altmami and Menai, 2022; Erera et al., 2019; Lu et al., 2020]. Recent approaches relying on large lan-

guage models show promise but remain nascent [Qureshi et al., 2023; Bolanos et al., 2024], facing challenges in inconsistent quality of generated syntheses [Martin-Boyle et al., 2024; DeYoung et al., 2024], hallucination [George and Stuhlmüller, 2023; Belem et al., 2024], misalignment with researchers’ specific goals [Qureshi et al., 2023; Yang et al., 2023], and opaque reasoning processes that limit interpretability. Moreover, as researchers often want to “stay close to the data” [Jiang et al., 2021; Feuston and Brubaker, 2021], these approaches may fail to capture the nuanced, personalized, and iterative nature of sensemaking in literature review. These limitations motivate the design of interactive systems that enable researchers to leverage evolving LLM capabilities while maintaining control over the sensemaking process through transparent, verifiable representations.

Specifically, we review interactive human-AI systems that support foraging and synthesis, two phases of literature review most pertinent to this dissertation. Other aspects of the research process—such as hypothesis generation, experimental design, writing and revision, peer review, and interdisciplinary collaboration, among others—also present rich opportunities for human-AI interaction and have garnered recent interest since the advent of LLMs (e.g., see surveys on AI and LLMs for automating scientific discovery [Zhang et al., 2024b], peer review [Kousha and Thelwall, 2024], literature review generation [Altmami and Menai, 2022], and research activities more broadly [Luo et al., 2025; Zhou et al., 2025]). Complementary to these efforts that focus on automating scientific research, we believe structured mediation and its principles explored in this dissertation can help inspire the design of more AI-supported and researcher-centric systems for other stages of scholarly research.

2.2.1 Support for Reading and Foraging Within Papers

One line of work has explored augmented reading interfaces that aim to mitigate challenges scholars face when engaging with individual papers. For instance, they may encounter high cognitive costs while reading due to unfamiliar jargon [Head et al., 2021], complex mathematical notation [Head et al., 2022], or the need for domain-specific schemas for effective comprehension [Bazerman, 1985]. These include tools that aid comprehension by simplifying scholarly language [Head et al., 2021; August et al., 2023], improve efficiency by selectively guiding reader attention [Lee et al., 2016], and lower the costs of saving relevant information [Han et al., 2022; Kang et al., 2022]. Other tools address fragmentation by weaving external

context into the paper environment, for example by visually augmenting inline citations based on a user’s reading history [Chang et al., 2023], by linking tables and text through contextual visualizations or interaction [Kim et al., 2018a; Badam et al., 2019], or by embedding relevant follow-on work [Rachatasumrit et al., 2022] and presentation videos [Kim et al., 2023] as margin notes localized to relevant parts of a paper. An opportunity exists for more dynamic, user-driven augmentation that can adapt to individual information needs as they emerge during reading—supporting the progressive disclosure that characterizes effective information foraging.

2.2.2 Support for Synthesis and Organization Across Papers

Beyond reading support for individual papers, prior work has also explored how to help researchers more deeply understand a small set of papers. For example, systems such as PaperWeaver [Lee et al., 2024b] or ACCoRD [Murthy et al., 2022] generate comparative statements between *two papers or concepts* to help researchers better understand an unfamiliar paper by contextualizing it to a familiar paper. More closely related to our work, systems like Elicit [Elicit, 2023] and SciDaSynth [Wang et al., 2024a] allow users to compare small sets of papers in a fixed table-based representation. Tables are ubiquitous across diverse sensemaking domains, such as online search [Chang et al., 2019; Spence et al., 1996; Chang et al., 2020], developer support [Liu et al., 2019, 2022a], and business analysis [Fok et al., 2024b]; within scholarly sensemaking, researchers also often manually create literature review tables to organize, compare, and synthesize information across many papers [Hashimoto et al., 2017; Newman et al., 2024; Wang et al., 2024a]. User-driven systems also exist that facilitate a more top-down overview over the literature, aiding paper discovery and topic-level navigation across the knowledge space. For example, some approaches cluster a paper’s references in a reading environment into related research threads [Kang et al., 2023, 2022], visually group papers on a canvas based on citation-edges or semantic similarity [Narechania et al., 2022; Moss, 2021], organize papers into hierarchical structures [Hsu et al., 2024; Zhu et al., 2023], and create overarching narrative summaries [Susnjak et al., 2025; Altmami and Menai, 2022]. While these systems demonstrate the value of computational support for literature synthesis, most provide single-level representations (e.g., tables or clusters) instead of supporting navigation across multiple levels of abstraction. The challenge of sensemaking in comprehensive literature review requires researchers to fluidly move between detailed evidence

and high-level themes, suggesting a need for multi-level information displays that maintain clear provenance and linking across different representations.

Chapter 3

Qlarify: Enhancing In-Situ Information

Retrieval with Expandable Abstracts

While abstracts serve as entry points for researchers triaging relevant literature to explore in further depth, their static representation as summaries—condensing long, complex research papers into fewer than two hundred words—limits their ability to address researchers’ individualized and dynamic information needs. In this chapter, we explore a paradigm for augmenting researchers’ literature exploration by introducing QLARIFY, a system that offers enriched information scent within paper abstracts themselves, enabling researchers to progressively explore full papers through low-cost, mixed-initiative interaction features. Rather than replacing the core practice of reviewing abstracts, we surface retrieval-augmented, AI-generated passages as intermediate representations that scaffold the transition from cursory skimming to more detailed review. QLARIFY embodies the core mechanisms of structured mediation: *Discovery* through AI-suggested expandable entities that offer information scent and guide attention; *Iteration* where scholars can both broaden and deepen their understanding through recursive expansions; and *Verification* through clear interactive attribution linking AI-generated content back to source paragraphs in the full paper. Through this exploration of an expandable text interaction paradigm, we demonstrate how intermediate representations can transform static text into dynamic artifacts, preserving the natural flow of scholarly sensemaking through reviewing abstracts while enabling just-in-time access to clarifying information. This work was previously published at ACM UIST 2024 [Fok et al., 2024a] and has been adapted for this document.

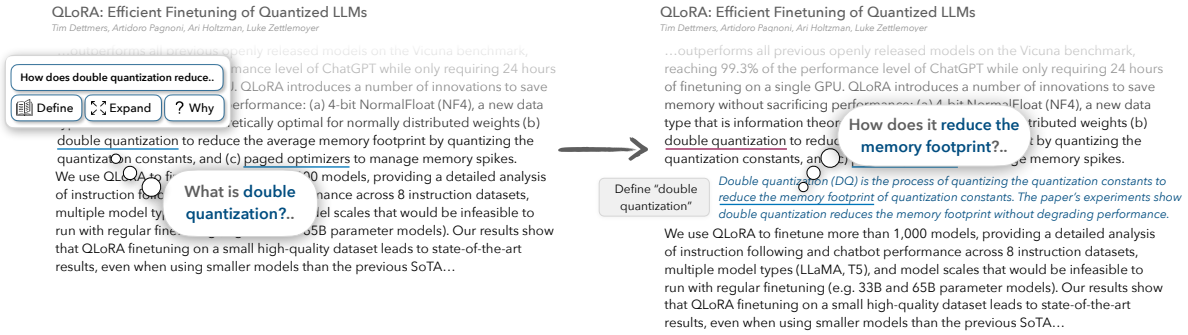


Figure 3.1: *Recursively expandable abstracts* present a novel mixed-initiative interaction technique, leveraging large language models to enable a low-cost, on-demand, and fluid expansion of static abstracts with information retrieved from full papers.

3.1 Introduction

Reviewing prior work is critical to scientific progress, helping scholars to survey what has already been done, identify open challenges, and spark inspiration [Knopf, 2006]. In response to the burgeoning task, technological interventions have sought to aid scholars in discovering and consuming the vast literature. Systems such as Google Scholar and Semantic Scholar help scholars *discover* relevant papers, while other tools assist along various aspects of *consumption*, such as improving comprehension [Head et al., 2021; August et al., 2023], efficiency [Lee et al., 2016; Fok et al., 2023], or sensemaking [Kang et al., 2023, 2022] of discovered works.

Scholars often need to browse and triage collections of papers, such as email digests of recent publications or conference proceedings, to assess relevance, identify future reading material, or satisfy an informational curiosity. However, reading papers can be cognitively costly due to barriers in full text, such as text density, jargon, and difficulty locating specific information [Head et al., 2021; August et al., 2023], and the overall declining readability of scientific texts [Plavén-Sigra et al., 2017]. To sidestep these issues, prior work has pointed to a common triaging workflow of moving from paper abstracts to full texts, for instance perusing each paper’s abstract to determine its relevance and legitimacy and only later examining a subset in more detail [Ishita et al., 2018; van Dinter et al., 2021]. Here, at the seam of literature discovery and consumption there exists another challenge which has received less attention—the initial, breadth-first

exploration of the literature. While literature discovery tools provide scholars with potential papers, and literature consumption tools assist scholars in reading selected papers, neither supports the preliminary exploration of papers through their abstracts.

Though often used for triage, abstracts also have several limitations. As static summaries that condense upwards of 10,000 words into one or two paragraphs, abstracts rarely address the entirety of scholars' information needs and may require clarification to fully comprehend. For instance, an abstract could mention the size of a study, but leave a scholar uncertain about the recruitment process, participant demographics, or analyses. This challenge of locating additional context is further complicated when scholars need to triage numerous papers. A scholar interested in understanding how prior work built on a specific dataset would need to manually open the PDFs for each candidate and search that paper for mentions of the dataset name in order to find passages that contained detailed descriptions about how the dataset was used.

Through a formative study, we first examined the common types of information needs that scholars held when reading abstracts for triage. We found that many of these needs were expressed as questions grounded in the abstract itself, and furthermore answerable given additional information from the paper's full text. To bridge these knowledge gaps, we propose a novel interaction technique, *recursively expandable abstracts*, that connects abstracts and full papers by allowing users to interactively ask clarifying questions and expand abstracts with information retrieved from the full paper texts. We reify expandable abstracts within QLARIFY, an LLM-powered augmented reading interface for scientific paper abstracts. In contrast to other chat-based systems that require users to formulate questions from scratch, QLARIFY showcases a dynamic text interaction paradigm where just-in-time questions are formed through direct, lightweight engagement with abstract itself.

By highlighting any part of an expandable abstract in QLARIFY, users can request an elaboration of the abstract and progressively expand the abstract with relevant context from the full paper. These expandable abstracts are also recursive, affording clarifying questions that probe deeper into details of the paper in a threaded manner. QLARIFY bootstraps the question-asking process by recommending potentially expandable entities within an abstract, providing information scent toward informative areas to expand (Figure 3.1). QLARIFY reduces the cost of asking an appropriate question to a single click by providing two types of question scaffolding: (1) a contextually-sensitive AI-suggested question which aims to infer a

user’s information-seeking intent, and (2) three static questions—*Define*, *Expand*, and *Why*—which mirror the common information needs scholars may desire in an expansion. Finally, QLARIFY enables efficient verification of generated expansions with attribution via deep linking to relevant passages in the paper.

Though an interview study, field deployment, and comparative user study, we evaluated the design and efficacy of QLARIFY and recursively expandable abstracts for abstract and paper exploration. In our studies, we found participants used the AI-suggested expandable entities as a guide for probing abstracts for details within papers, often creating threaded expansions to satisfy follow-up curiosities. Participants tended to ask more questions with QLARIFY than a question answering baseline and overall preferred expandable abstracts over baseline paper exploration approaches. Our findings also indicate LLMs can effectively infer scholars’ information-seeking intents within abstracts and add value by surfacing clarifying information generated from the full text of papers, with over 88% accuracy using a standard retrieval-augmented generation approach over full papers. We conclude with future opportunities for recursive expansion interactions and their implications for AI-infused scholarly support tools.

3.2 Recursively Expandable Abstracts

3.2.1 Formative Study

To understand the types of information needs that emerge when exploring scientific paper abstracts, we first conducted a formative study observing scholars reading abstracts in a familiar domain.

Participants

We recruited seven participants from an academic institution via social media and snowball sampling (Table 3.1). All participants actively conducted research across different areas of computer science, and reported familiarity with the research process, including triaging, reading, and organizing scientific papers.

Procedure

After introducing the study and obtaining consent, participants were asked to read 3–5 abstracts of their choice. Participants selected abstracts from various sources: many used results from a paper search engine

Table 3.1: Participants in our formative study and selected questions they asked while reading scientific paper abstracts.

Title (Research Area)	Representative Questions
P1 PhD student (Explainable AI)	What does “TAM” mean? What is a “path analysis”? What does “visual question answering task” mean? What’s an example of a “trustworthiness cue”?
P2 PhD student (Computational Biology)	What is the “two-stage algorithm”? What are the 12 challenging reasoning tasks? What’s an example of this task? What is “the zero hypothesis”? What does “outlying failure cases” mean?
P3 PhD student (Human-AI Interaction)	What does “comparable to SoTA phrase based systems” mean? What is a “sequence transduction model”? What is the “path-x challenge” and “path-256 task”? What does “retrieval set” mean?
P4 PhD student (NLP)	What loss function did they use? What’s the model architecture? What’s the metrics they use? Why do they call it “human-interpretable”?
P5 Post-doctoral scholar (HCI)	How do they define “trust” in human-AI teamwork; is this reliance? Is “human-agent teamwork” a defined sub-field of prior work, or is that just jargon the authors use? What are examples of “spatial crowdsourcing”? What are examples for the “two realistic task assignment settings”?
P6 PhD student (HCI)	What do the authors mean by “gigification of knowledge work”? What were the design recommendations? What are the key findings in a simplified sense? What papers are they building upon? What is their study design?
P7 PhD student (HCI)	What does “perceived valence” mean? What does “participatory foresight” mean? What is “early testing of AI-based features”? What does “practitioners with reports” look like?

seeded with a relevant prompt for their area of research, one used abstracts for papers they were currently reviewing, and one used papers they had previously saved for a later reading session. While reading each abstract, participants were asked to share aloud any thoughts, questions, or confusions they had about the information within the abstract or about the paper in general. All studies lasted about 45 minutes. Participants were thanked for their time.

Findings

Our study revealed four common types of information needs participants had while reading an abstract:

Definition Participants wanted to define jargon, unfamiliar language, or acronyms they encountered in the abstract. The definitions they sought were often not in the general sense, but rather specific to the context of

the paper they were currently reading.

Instantiation Participants sought examples to provide concrete context for under-specified language. For instance, in the sentence, “We find our approach outperforms three baselines on a common question-answering benchmark,” it is unclear which specific baselines or question-answering benchmark was used. When an abstract described an unfamiliar concept, e.g., a new task or dataset, participants also wanted to view an instance of the concept to help visualize its structure and compare it against familiar concepts.

Clarification Participants sought additional context to help explain technical or unfamiliar language in an abstract. Since abstracts are concise, self-contained summaries for a long paper, authors are compelled to withhold particular details and use dense language to convey information. As a result, scholars reading abstracts often have information needs expressed through implicit clarification questions, personalized by their own expertise and reading goals.

Motivation Participants expressed a desire to probe the authors’ motivations and justifications for aspects of the paper. For instance, some participants asked why the authors chose their particular method (e.g., model architecture, loss function, or task), why particular language was used to describe their system (e.g., “human-interpretable”), or the significance and novelty of their research problem. Addressing this need could help satisfy curiosities, expand their understanding, or evaluate a paper’s validity.

3.2.2 Recursively Expandable Abstracts Interaction and Design Space

Based on the information-seeking behaviors found in our formative study, we propose an interaction technique called *recursively expandable abstracts*, that augments abstracts with additional relevant information in response to dynamic user queries for clarification (Figure 3.2). This expandable text paradigm is inspired by *StretchText* (or transclusion) [Bernstein, 2009; Nelson, 1983], an early vision in Project Xanadu and hypertext design from the 1980s for structuring text on the web that allows users to choose the level of detail they want to see. When a specific area or keyword is selected, the originally concise text “stretches” to reveal additional details. While the original vision for *StretchText* requires carefully-authored, structured text and has not gained wide adoption, we revisit and build on this vision by leveraging LLMs to dynamically

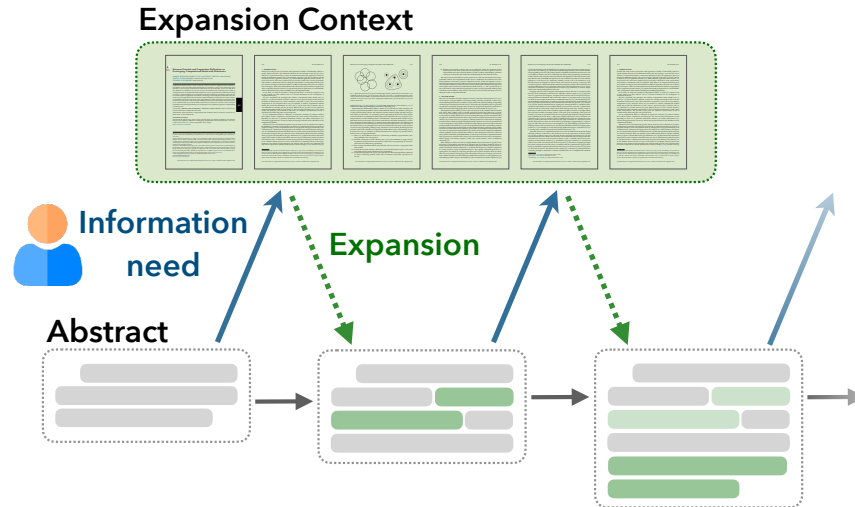


Figure 3.2: Recursively expandable abstracts allows users to retrieve clarifying information from a broader *expansion context* (e.g., the full paper) in response to dynamic information needs, forming *expansions* that grow the abstract fluidly.

generate on-demand summaries that support personalized and interactive expandable text.

For this, we focused on designing expandable summaries based on scientific abstracts and papers. Abstracts are concise, static, author-crafted text summaries for a long scientific document; as such, no single abstract can concisely capture the interests of every reader or address dynamic information needs that arise while reading [Zhang et al., 2023]. Expandable abstracts ameliorates the static limitations of summaries by allowing scholars to interactively expand abstracts with additional clarifying information retrieved from an *expansion context*. The expansion context for many clarification questions arising from an abstract is often the corresponding paper’s full text. However, expansions could also be drawn from information in a broader domain, such as related papers in the paper’s citation network or general information in an online resource (e.g., Wikipedia). To inform the interaction design, we first articulate the plausible dimensions and alternatives of its design space (Figure 3.3). This structure emerged organically through an iterative design process within the research team and validation with external scholars, and considered alternatives were inspired by established information foraging affordances.

Information needs

One set of dimensions of the design space concerns what and how information is selected for expansion. First, what *type* of information needs should be expanded? The four types of needs identified in the formative study are similar in that they all represent an information need *grounded* in language from the abstract. These may emerge as scholars read an abstract, anchored to specific words in the abstract. For instance, in the sentence, “We evaluated against three baseline approaches on a popular question-answering benchmark,” a grounded information need might be: “What were the three baseline approaches” or “What was the popular QA benchmark used?” Information needs may also extend beyond the content explicitly stated in an abstract. Scholars might approach abstracts with predefined questions *agnostic* to any particular abstract but relevant to their broader research goals. For example, they may seek to understand the methods, experiments, or findings across multiple papers as part of their exploration. Finally, *latent* information needs refer to details that scholars may not consciously recognize, but are relevant to their goals. These details, although valuable for expansion, are not explicitly mentioned in the abstract or at the forefront of the scholar’s consciousness, making them difficult to query for directly. Expansions could also be instantiated from different sources, for instance stemming from a user’s question, an AI’s suggested entity for expansion, or some mixed-initiative approach.

Expansion

A second set of dimensions considers the content and visualization of information within an expansion. Information used to expand an abstract could be retrieved from different *contexts*, e.g., the full text of the paper for an abstract, other relevant papers, or an external knowledge base such as Wikipedia. The *length* of the generated expansion is also an important consideration, with good designs aiming to balance addressing a scholars’ information needs yet being judicious in length to limit the added cognitive load. Generated expansions may range from a phrase or sentence, to a longer paragraph with details that could motivate further exploration.

The *placement* of expansions within a system interface is another important design choice. Possibilities include placing expansions within an adjacent pane (similar to many chat-based applications with documents), in a popup card (similar to citation cards in augmented paper reading interfaces or page previews

Dimension	Alternatives			
Information needs type	Agnostic	Grounded	Latent	
Information needs source	User-suggested	AI-suggested	Mixed-initiative	
Expansion context	Same doc	Related docs	Open-domain	
Expansion length	Short phrase	One sentence	Several sentences	
Expansion placement	Fluid	Inline	Appended	Popup Sidebar
Expansion delineation	Bold	Italicize	Colorize	Indent Quote
Attribution method	Embedded			Separate
Attribution granularity	Phrase	Sentence	Entire expansion	
Attribution length	Phrase	Sentence	Paragraph	Page

Figure 3.3: The design space for an expandable abstracts interaction paradigm, formed through an iterative design process and validation with external scholars. Alternatives we explored in QLARIFY are highlighted in gray, and those included in the final system are outlined in red.

in Wikipedia), appended at the end of a summary, or placed in-situ near an appropriate text anchor in the abstract. Selecting the optimal placement requires trading off the navigational effort between an expansion and the abstract, and the potential for visual distraction or clutter. For placements that interweave expansions and the original abstract, it is important to consider how a system may differentiate between these two sources of information. Unlike the original abstract, whose provenance is known and trusted, expansions are AI-generated, introducing concerns around hallucination or trustworthiness of the generated information. Approaches may include visually delineating between the expansion and abstract text with standard visual cues, e.g., bold, italics, color, or through positional displacement, helping users to visually identify and switch between the two sources of text.

Attribution

A third set of dimensions considers designing to convey information provenance for the generated expansions. Provenance for question-answering contexts is often achieved through attribution, i.e., retrieving evidence from the expansion context that support the generated answer. There are several *methods* of pre-

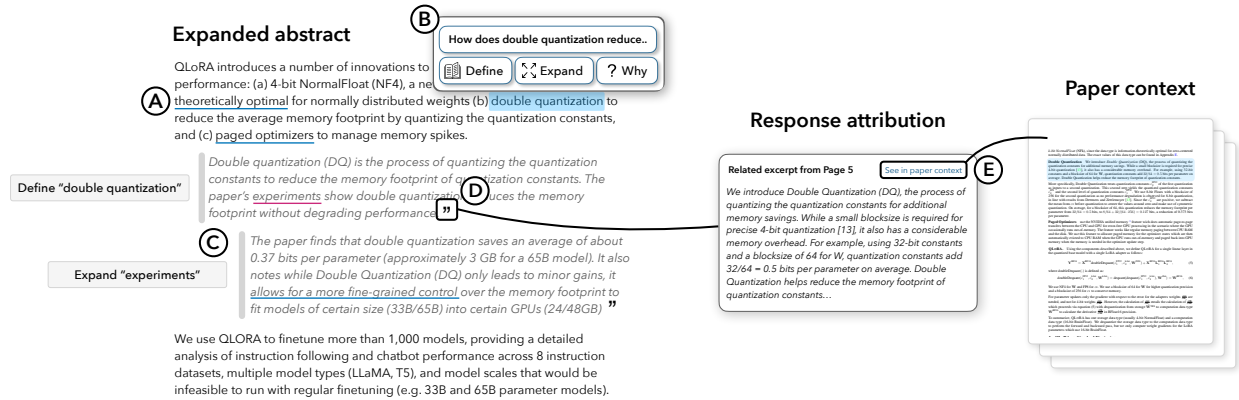


Figure 3.4: Recursively expandable paper abstracts with attribution in QLARIFY. Expansions are created on-demand by highlighting text in the abstract or selecting an AI-suggested expandable entity (A), revealing a question palette (B). Selecting a question in the palette prompts an LLM to retrieve relevant clarifying information, presented as a fluid expansion within the abstract (C). Users can drill-down to see evidence for a response in a paper excerpt (D) and within the full paper context itself (E).

senting attribution—evidence could be embedded directly in an expansion, akin to quotes, and use visual cues to demarcate abstractive and extractive text, or provided as separate excerpts surfaced verbatim from the expansion context. For longer or more complex expansions, multiple pieces of evidence may be necessary to support each claim in the generated text. In these cases, systems may consider the most appropriate *granularity* of attribution to determine the utility of providing evidence for each phrase, sentence, or entire expansion. Finally, selecting the appropriate *length* of attributed evidence can ensure sufficient validation of the generated expansion without introducing excessive cognitive burden.

3.3 System Design

Next, we describe the design (§3.3.1) and technical implementation (§3.3.2) of Qlarify, an augmented reading interface that implements recursively expandable abstracts.

3.3.1 User Interface

The design of Qlarify was guided by the four types of information needs observed in our formative study, and refined through an iterative design process in which alternatives of the design space were implemented and evaluated.

Eliciting Information Needs as Clarifying Questions

With Qlarify, users begin by reading an abstract as they typically would. As information needs arise, they can highlight any part of the abstract to request additional information. In response, Qlarify displays a *question palette* centered above the highlighted text (Figure 3.4B) that enables users to specify their information needs as clarifying questions anchored to a specific context in the abstract (i.e., the highlighted text). The overarching design goal of the question palette is to reduce the cost of asking these questions to a single button click. As a result, the question palette includes four buttons: three static buttons with the questions *Define*, *Expand*, and *Why*, and one dynamic button with an AI-suggested question.

The three static questions are fixed regardless of the selected text and reflect the common types of grounded information needs revealed in our formative study. Specifically, the *Define* question aims to address *Definition* and *Instantiation* needs, the *Expand* question aims to address *Clarification* and *Instantiation* needs, and the *Why* question aims to address *Motivation* needs. The *Expand* question is visually centralized, serving as a “catch-all” option to incrementally retrieve more details appropriate for most information-seeking circumstances. While increasing the number of static questions could provide more flexibility, based on feedback to initial prototypes of Qlarify we determined that providing more questions could clutter the interface, occlude more of the abstract, and cause decision paralysis in selecting an appropriate question. To complement the static questions, the palette includes an LLM-generated question that infers the most probable information need based on the user’s highlighted text. Only the top-ranked candidate question is displayed in the palette, again to avoid overload.

To complement the manual highlighting of text users want to expand, Qlarify also pre-selects several *expandable entities*. These entities capture spans of text in the abstract that the system believes could be further expanded from the expansion context; for instance, they could include under-specified language (e.g., “some”, “several”, “various”) or jargon (e.g., acronyms). Expandable entities are visually indicated with a blue underline (Figure 3.4A), and users can click on an entity to reveal the question palette. Altogether, Qlarify aims to reduce the cost of asking grounded information-seeking questions through these two lightweight interactions.

Expanding Abstracts with Clarifying Information

When users select a question from the question palette, Qlarify creates an *expansion* by fluidly expanding the abstract with in-situ information retrieved from a larger expansion context (Figure 3.4C). Each expansion is an abstractive, LLM-generated response to a user’s question, containing up to three sentences. In instances where a question cannot be answered, no expansion is created, and a toast alert is shown in the bottom right of the screen instead to indicate an expansion could not be created.

Key to the expansion’s design is ensuring visual delineation between the original text of the abstract and that of the generated expansion. Expansions are presented as indented blocks, appended below the sentence containing the selected expandable entity. A vertical bar and tag containing the question used to generate the expansion are shown to the left of each expansion to help users easily identify the visual boundary and purpose of each expansion. The text of an expansion is also colored blue when initially created, gradually transitioning into a light gray after a few seconds. This produces a smooth animation that visually cues attention to a new expansion, before fading into a color (gray) that is similar but not overly distracting from the color of the original abstract text (black). To indicate parts of the abstract previously expanded, text used to seed the expansion—either from a user highlight or an expandable entity—is underlined in purple.

For each generated expansion, Qlarify suggests additional entities that could be expanded. Similar to interactions with the original abstract text, users can select an AI-suggested entity or highlight any text in the new expansion to recursively expand further. Expansions created on other expansions form a threaded abstract reading experience, allowing users to easily dive deeper into aspects of interest from the abstract by gradually retrieving details from the full paper. As each expansion is limited to three sentences in length, users are less likely to be overwhelmed by information in any single expansion. They can continue expanding until their information needs are satisfied, at which point they can easily pop back up into the original abstract or expansions at any level. If a particular expansion is no longer needed, users can click on the question tag at the left of the expansion to collapse it into its parent expansion.

Drilling Into a Paper with Attributed Responses

Since each expansion is generated automatically by an LLM, there is a potential risk of generating content that is unfaithful to the original paper or otherwise factually incorrect, a problem referred to as hallucina-

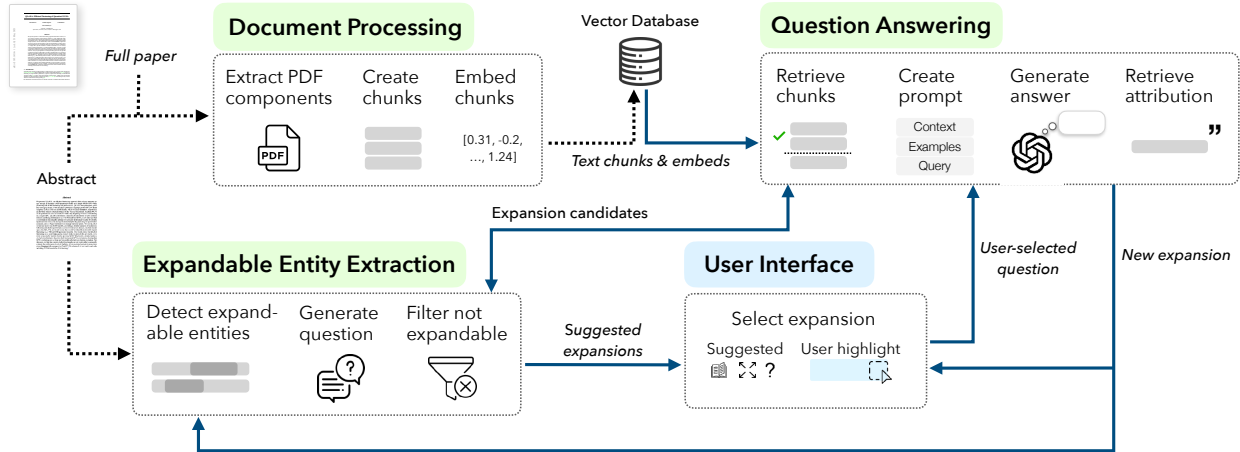


Figure 3.5: QLARIFY’s system architecture. Each paper is first processed (Document Processing) and initial expandable entities are extracted from the abstract (Expandable Entity Extraction). When a user asks a question for an expandable entity, QLARIFY uses a retrieval-augmented generation approach to form a response and retrieve attribution (Question Answering). QLARIFY then suggests expandable entities within the response, allowing recursive expansions.

tion [Maynez et al., 2020; Ji et al., 2023]. To help mitigate these risks, Qlarify provides *attribution*, or extractive supporting evidence, for each expansion. Users can click on a quote button at the end of each expansion to show a card with the most relevant paragraph from the full paper (Figure 3.4D). Within the card, users can further drill-down to open the paper in an integrated document viewer overlay, with the attributed paragraph navigated to and highlighted in the context of the paper (Figure 3.4E). Through these two levels of interactive attribution, Qlarify allows users to explore the surrounding paper context and verify the accuracy of an expansion.

3.3.2 System Architecture

To create an expandable abstract, Qlarify implements three natural language services (illustrated in Figure 3.5): (1) A *document processing* service extracts and encodes information for a paper’s full text; (2) A *question answering* service generates attributed answers to users’ document-centered questions; and (3) An *expandable entity extraction* service identifies expansion candidates within an abstract or generated expansion. We provide an overview of our implementation of these services, which may serve as a starting reference for others exploring similar expandable summary interactions.

Document Processing

Papers ingested by Qlarify are first processed to reduce latency when interactively generating expansions at query-time. Each PDF is parsed into its constituent tokens and bounding boxes, and sentences and paragraphs are constructed from the full text. Then, chunks are created with a chunk size of three sentences and a two sentence overlap. Each chunk is embedded using the `all-mpnet-base-v2` encoder from the SentenceTransformers framework [Reimers and Gurevych, 2019] and stored into a vector database. Embeddings of paragraphs are also created and stored in a separate index, which are used for retrieving attribution.

Attributed Question Answering

We use a retriever-reader architecture with LLMs for retrieval-augmented question answering. When users select a question from the question palette, Qlarify first embeds the question with the same encoder used during processing. It then retrieves the 12 most relevant paper chunks (with relevance determined by cosine similarity between chunk and question embeddings) to form a context. An LLM prompt is then formed by concatenating a general description of the question answering task, the context, a few examples of question and answer pairs, and the question. The prompt further instructs the LLM to be concise, use language from the provided paper context when appropriate, generate answers containing no more than three sentences, and return no answer if the question cannot be answered given the context. These specific prompt tuning adjustments were made over several iterations of inspecting Qlarify’s expansions. Finally, we use `gpt-3.5-turbo-1106` to generate an answer for the question using this few-shot prompt. The current Qlarify prototype answers questions using information from the full text of the source paper only; we leave consideration of other possible expansion contexts (e.g., other related papers) for future exploration.

Qlarify further provides attribution for each of its expansions to enable users to verify the accuracy of the generated answer and ease into the full paper. To generate attributions, Qlarify retrieves the most relevant paragraph to the generated text (by cosine similarity). We explored other attribution schemes in earlier iterations of Qlarify. For instance, we tried retrieving chunks for each individual sentence, but found chunks were less preferred than paragraphs since they sometimes lacked sufficient context. We also tried providing attribution for each sentence where expansions consisted of multiple sentences. However, we found the need to read and reconcile multiple attribution sources introduced confusion and made verification more

challenging.

Expandable Entity Extraction

To complement users in manually specifying their own expansions, Qlarify proactively suggests parts of an abstract or expansion that could benefit from additional context. To identify these regions within an abstract, Qlarify uses `gpt-4-1106-preview` with a few-shot prompting strategy. The model is instructed to identify short text spans (i.e., entities) which may be expanded to provide clarification for vague, dense, or jargon-rich language. The prompt also specifies that information required to expand each entity should not be already available in the abstract or expansion. For each entity, Qlarify performs a dry-run expansion (using the same *Expand* question in the question palette), and removes entities for which no answer is found.

For each remaining entity, `gpt-4-1106-preview` is prompted to generate a single question that users might ask in expanding that entity. For instance, given the sentence, “We propose a new framework to address the ACTA task,” Qlarify could identify the entities “a new framework” and “ACTA,” and generate the questions “What are the main characteristics of the proposed framework?” and “What is the ACTA task?” respectively. All of these *expandable entities* are underlined in reading interface and the suggested question is shown in the question palette. The same question generation prompt is executed on-the-fly to generate the suggested question within the question palette when users create an expansion by highlighting any text.

3.3.3 Implementation Details

Qlarify was implemented as a standalone web application using TypeScript, CSS, and the React framework [Source, 2023] for the user interface. Backend services and LLM-powered functions were implemented in Python and the Flask framework [Pallets, 2023]. GROBID [Grobid, 2008–2025] was used to parse paper PDFs into a structured JSON format, and the MMDA open-source library [for AI, 2023] was used to construct `Document` objects from the output, from which full text paragraphs and sentences could be retrieved. The PDF reader for viewing expansion attribution in context was adapted from an open-source PDF reader library [for AI, 2022]. The `gpt-3.5-turbo-1106` (with 16k token context window) and `gpt-4-1106-preview` (with 128k token context window) LLMs were accessed via OpenAI’s APIs with

a generation temperature of 0 and maximum length of 750 tokens. The specific LLM prompts we used for each service is provided in Appendix A.3.

3.4 Study 1: Interview Study

To evaluate the usability and utility of Qlarify, we first conducted a qualitative interview study using a think-aloud protocol. In this study, we aimed to answer two research questions:

RQ1. How do users engage with Qlarify for exploring abstracts?

RQ2. What are the advantages and limitations of AI-augmented abstracts?

3.4.1 Study Design

Participants

We recruited 9 researchers (6 male, 3 female; Age: $M = 27.8$, $SD = 9.3$) via university mailing lists and Slack channels. Eight participants were doctoral students within computer and information science and one participant was a research scientist.

Selected Papers

To incentivize engagement, for each participant we curated a personalized set of abstracts that aligned with their research expertise and interests. In a screening survey, we asked participants to list up to 5 “seed” papers representative of their research interests. We then used the Semantic Scholar Recommendations API¹ to obtain 25 additional recommended papers for each participant based on their seed papers. We created expandable abstracts within Qlarify for all seed and recommended papers, excluding those without a valid PDF file.

Procedure

Participants first completed a tutorial that introduced them to Qlarify’s features (5 min). They were then instructed to browse their personalized list of expandable abstracts, imagining the list was recommended to

¹<https://api.semanticscholar.org/api-docs/recommendations>

them from a colleague or search engine (25 min). During their exploration, participants were asked to think aloud, sharing any observations or questions as they emerged and explaining the interactions they chose to use. After the task, participants were engaged in semi-structured interviews that sought to elaborate on the advantages and limitations of Qlarify’s expandable abstracts (15 min). All think-aloud sessions and interviews were recorded, transcribed, and coded through a thematic analysis (additional details in Appendix A.1.1). Study sessions lasted 45 minutes and participants were compensated with \$25 USD. This and all following studies involving human subjects were approved by a university institutional review board.

3.4.2 Results

An analysis of interview transcripts and interaction logs uncovered various ways in which Qlarify supported the exploration of abstracts, such as using interactive expansions to retrieve additional information on-demand from full papers, threaded exploration to dive deeper into aspects of interest, and LLM-generated expandable entities and questions to guide attention. In the following results, we refer to participants with the pseudonyms P1–9.

Abstract expansions allowed an on-demand recursive extraction of high-quality information from papers

Participants were actively engaged with Qlarify during the study. On average, each participant explored 4.0 papers ($SD = 1.1$, $Mdn = 4.0$) and created a total of 20.8 expansions ($SD = 7.8$, $Mdn = 18.0$). Based on the think-aloud, participants liked how Qlarify allowed them to surface details from the paper using simple interactions with the abstracts over manually searching for them over the full papers. For example, one participant remarked, “*I was impressed by the things that I was able to pull from the paper and the amount of additional details I can get just by reading this abstract*” (P3). Participants also pointed to how abstracts have a familiar structure that served as natural entry-points to scaffold pulling in additional details from parts of the full paper when needed:

“I think that one of the key things about being a PhD student is being able to quickly break down information without having to spend too much time reading the entirety of the paper. And so the abstract allowed for me to do that without having to even read that. Usually the rule of

thumb is abstract, intro and conclusion. And with this, I feel I get a bit of the intro, conclusion, results, discussion, analysis, all that within the abstract breakdown.” – P7

Participants further noted how Qlarify’s generated expansions answered the questions they asked surprisingly well (P1–3, P5, P6, P8). One participant appreciated how the expansions “*didn’t just summarize, but seemed to extract meaning from the paper*” (P6), validating the LLM’s ability to form concise and useful answers drawn from complex text. We observed some participants beginning their exploration by browsing the abstract of a paper they were familiar with, attempting to gauge the accuracy and reliability of the generated expansions (P5, P8, P9). Others instead mentioned trusting the model’s responses, especially with the confidence that they could dive into the paper to verify the attributed evidence if needed (P4, P7, P8).

Participants found the AI-suggested question in the question palette often aligned well with their information-seeking intents and reduced the costs of forming a question to expand the abstract. P8 described the suggested question as, “*It seems to almost read my mind when I click on something or highlight something,*” and P5 remarked, “*Every time I think of what the question is, that’s pretty much what the question it already thought of is.*” A similar sentiment was reflected in Qlarify’s usage behaviors; within the question palette, participants selected the LLM-generated question most often (40.1% of clicks, *Define*: 23.5%, *Expand*: 22.5%, *Why*: 13.9%).

Participants also frequently utilized the recursive expansion feature of Qlarify—58% of expansions created by participants were threaded (i.e., formed by asking questions about text in another expansion rather than from the abstract). Based on their think-aloud, participants mentioned how the threaded expansions empowered them to dig deeper and ask follow-up questions if an initial expansion did not fully satisfy their information needs (P1, P3, P4, P6, P8). For example, P4 said, “*I really did love the way you could keep going and go branching into a tree.*” Similarly, P1 described how the continuous presence of underlined AI-suggested entities served to motivate further exploration:

“I can keep diving more because the highlighting feature is not provided only on the first level of the abstract, it’s also provided on the next level. When its generates a response in the second level, it also provides these underlining features, which if I have any more questions I could keep using these interactions to help me make sense of the abstract.” – P1

In summary, the threaded expansions not only supported ongoing engagement but also sensemaking effi-

ciency within the abstract.

AI-suggested expandable entities guided and motivated deeper exploration of abstracts.

The majority of the time, participants clicked on a pre-selected underlined entity to create expansions (77.5%) as opposed to manually selecting a custom text span (22.5%). While we initially designed this feature to lower the interaction costs, the think-aloud suggested that the pre-selected entities also served as information cues that can facilitate discovery. Many participants (P1–4, P6, P8) commented how the underlined entities served as visual cues to “*keywords that may be relevant*” and “*tell me what to focus on*” (P1).

Conversely, if the underlines did not precisely capture participants’ needs, the ability to highlight custom text spans to create an expansion was appreciated (P1, P2, P8). P6 summarized how the interactive features of Qlarify in concert could address all of her envisioned information needs over an abstract:

“I think the underlines were already really good. Highlighting something as a backup to the underlines already gets you to probably 99% coverage for the things you would want to ask. And then for the last 1%, I could click on the quote and go into the paper.” – P6

On the other hand, visually augmented reading interfaces such as Qlarify and its underlined entities can introduce distractions for some scholars. Since expanding the abstracts by interacting with the underlined entities required less effort than highlighting text, P4 noticed how she “*tended to default to whatever was already underlined,*” and it became “*hard to remember that I can just like pick anything out unless I was really curious about it.*” This behavior is not necessarily undesirable, but suggests careful consideration should be given to how augmented interface elements may inadvertently guide or constrain user interactions. Some suggested how the underlined entities could open up a rabbit-hole of exploration, derailing the reading of an abstract (P3, P7). P7 further noted how the seemingly limitless freedoms afforded by an abstract expansion interaction could be double-edged and inhibit a sense of completion:

“The endlessness of the underlines, as a completionist, my mindset, I want to click them all. And so I liked the fact that there were no bounds, but I could also feel overwhelmed knowing there are no bounds. So I feel like I could miss something the AI could uncover for me if I just kept clicking all the underlines.” – P7

These observations suggest that while the underlines may help guide an in-depth investigation of details in the paper, it can also potentially hinder the process of triage. As such, balancing interactivity and efficiency is critical for an expandable abstract interaction, ensuring users can engage deeply when needed but also efficiently navigate the content within the abstract and generated expansions.

3.5 Study 2: Field Deployment Study

To further investigate how scholars would interact with expandable abstracts in the wild, we deployed Qlarify during the 49th International Conference on Very Large Data Bases (VLDB 2023). For the duration of the conference, members of the research team invited conference attendees and other scholars to try out Qlarify via social media announcements including email, Twitter, Slack, and LinkedIn. We created a landing page within Qlarify to allow users to easily browse the 248 papers within the conference proceedings, with a paginated, scrollable list of paper metadata and a search bar for filtering papers. Clicking on a paper title navigated to a separate page with an expandable abstract for that paper.

During the week of the conference and two subsequent weeks, a total of 275 unique users interacted with 50 unique papers using Qlarify. Based on the interaction logs, each user created 3.3 expansions on average ($SD = 4.6$, $Mdn = 2.0$). We found users expanded abstracts using the pre-selected expandable entities rather than manually selecting custom text spans (80.4% vs 19.6% of interactions), similar to findings from the interview study.

We further found that users more often selected one of the three static questions (i.e., *Expand*, *Define*, and *Why*; 88.2% of interactions) than the more specific LLM-generated questions. Across the static questions, users created 41.6% of expansions with *Expand*, 31.5% with *Define*, and 15.1% with *Why*, a distribution that closely corroborates the frequency of question types we observed in our formative study. These results are in contrast to our interview study, where participants were more likely to select LLM-generated questions (40.1% of expansions). This difference might be due to the higher relevance of abstracts in our interview study compared to our deployment (i.e., paper abstracts were selected based on personalized recommendations for each participant).

While the majority of the time users created a single level of expansion from the original abstract, a significant portion (27.7%) of interactions were threaded, meaning users recursively asked follow on questions

by selecting additional entities in the expanded text. Some users recursively created up to 5 nested expansions. These results demonstrate the users' needs for recursive expansion of abstracts and Qlarify's ability to support this. In comparison to our interview study, fewer users in our deployment study created threaded expansions (58.0% vs. 27.7%, respectively), perhaps due to differences in user engagement or relevance of abstracts.

We also observed users actively engaging with the attributed evidence paragraphs and viewing the paper itself. In 14.8% of expansions, users viewed the attributed evidence paragraph for a generated expansion. About 60% of the time users were satisfied with the extracted evidence, while 40% of the time they further opened the PDF to view the highlighted evidence in the context of the paper. These behaviors suggest users were either interested in understanding the supporting information or sought to use the evidence as an efficient entry point into reading the paper.

Altogether, our deployment study suggests that an expandable abstract interaction presents a simple yet effective means to elicit clarification questions for abstract-grounded information needs, allowing users to retrieve attributed answers on-demand. Our findings complement the interview study and characterize real world usage behaviors of Qlarify within a natural context of user interaction where scholars are actively triaging papers.

3.6 Study 3: Comparative Evaluation

Finally, to understand how Qlarify compares to existing modes of triage, we conducted a within-subjects study where we compared Qlarify to two other paper exploration strategies. In this study, we aimed to answer the following research questions:

RQ3. How does Qlarify affect the quantity and types of questions users ask over abstracts and papers?

RQ4. How does Qlarify compare to and affect users' current navigational strategies across abstracts and papers?

Based on feedback from the previous interview and field deployment studies, we made small refinements to the design and usability of Qlarify. Specifically, we enabled users to ask any question they desired by editing the AI-suggested question in the question palette. We also removed the *Why* question as it was

the least frequently used, and the additional option for expansion added to users' cognitive load, especially given that users could now compose their own questions. Lastly, we moved the quote button for displaying attributed evidence from inline with the expansion text to the right margin adjacent to the expansion, and added a second button that hid a generated expansion, allowing users to easily view just the original abstract.

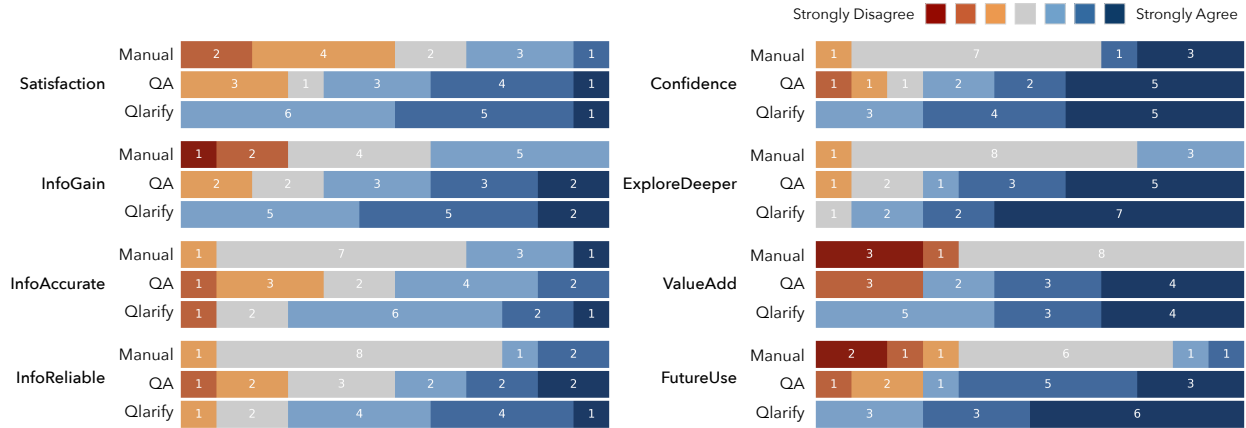


Figure 3.6: Distribution of participants' self-reported ratings within each condition in the comparative evaluation. Participants in the QLARIFY condition felt more satisfied with their exploration, more confident in retrieving relevant information from the full paper, more motivated to explore deeply, and a greater desire to use in the future. See Appendix A.1.2 for the precise wording used in the survey questions.

3.6.1 Study Design

Participants

We recruited 12 researchers (9 female, 2 male, 1 non-binary; Age: $M=25.5$, $SD=3.1$), who had previously read at least one HCI research paper, via Slack and snowball sampling. 10 were doctoral students, 1 was a Master's student, and 1 was an undergraduate student. The doctoral students reported having between 1–5 years of research experience. 7 of 12 participants identified their primary discipline as HCI or related sub-fields (e.g., human-centered AI, tech policy), 2 as AI/ML, and 2 as robotics. All participants reported reading research abstracts and papers at least once a week; a majority reported reading more than 10 abstracts and 1–2 full papers on average each week.

Conditions

The study included three system interface conditions: Manual, QA, and Qlarify. A comparison of the interfaces is provided in Figure A.1 in the Appendix. In the Manual condition, participants were given a list of paper titles and other metadata (authors, venue, publication year), abstracts, and PDFs, reflecting the manual process of browsing papers and abstracts. In the QA condition, participants were given the same paper elements as in Manual, but with an additional text field, below the paper abstract, that allowed users to sequentially ask questions about the paper. Responses to submitted questions were generated by an LLM (using the retrieval-augmented generation method in §3.3.2) and placed in a box below the question text field. The QA condition aimed to emulate a simple question answering service over full papers, and did not share context between multiple questions or provide attribution. In the Qlarify condition, participants were given the same paper elements as in Manual, but additionally had access to all features of Qlarify as described in §3.3.1.

Procedure

Participants were first provided with a tutorial introducing them to the three conditions (15 min). They then completed a task of reviewing papers from UIST 2023, selected based on the availability of PDFs (80 out of 123 papers). They were asked to envision themselves as TAs for an HCI seminar class and to create a shortlist of papers suitable for student reading, discussion, and presentation. This collection was selected to ensure the participants had relevant interest and expertise for triaging the papers. Participants completed the task using each of the three conditions (15 min each), the order of which was counterbalanced to mitigate the influence of any ordering effects. After the task, participants completed a post-task survey and a short interview describing their experience using the three systems. Study sessions were conducted remotely through an online video conferencing software and lasted 75 minutes. Participants were compensated with \$35 USD.

Measures

For quantitative data, we analyzed responses to the post-task survey in which participants rated their agreement with eight statements on seven-point Likert scale for each condition. These statements included par-

participants' self-reported measures of satisfaction and confidence completing the task, the quality and reliance of gained information, and desire for future system use. Detailed survey questions are provided in Appendix A.1.2. We also analyzed participants' interaction logs to measure the quantity and types of questions users asked and their patterns of navigation throughout the abstracts and papers (e.g., opening a paper's PDF). We used Wilcoxon signed-rank tests for two-condition comparisons and Friedman tests with Nemenyi post-hoc tests for three-condition comparisons. For qualitative data, we transcribed the post-task interviews and coded them through a thematic analysis. In this section, we refer to participants with the pseudonyms P1–12.

3.6.2 Results

Asking Questions (RQ3)

On average, participants asked more questions in the Qlarify condition than in the QA condition (Qlarify: 15.0 ($SD=5.4$), $Mdn=15.5$ ($IQR=11.3-19.0$), QA: 9.6 ($SD=3.9$), $Mdn=10.0$ ($IQR=6.8-12.5$), $W=4.0$, $p=.02$). Participants interacted with a similar number of abstracts in both conditions (Qlarify: 5.0 ($SD=1.9$), QA: 4.7 ($SD=2.0$)), but on average asked more questions per abstract using Qlarify (3.7 ($SD=2.9$) vs. 2.2 ($SD=0.7$)). These findings suggest that Qlarify motivated participants to ask more questions about the abstracts than unaided question answering, with participants describing how the underlined expandable entities and lightweight interactions in Qlarify helped to reduce friction in formulating and asking questions about papers. Specifically, we found that when using Qlarify, participants created an expansion by selecting an AI-suggested expandable entity 72% of the time (as opposed to manually highlighting an entity 28% of the time). Within the expansion palette, participants selected the LLM-generated question most often (38.9% of the time), in line with observations from the interview study. They also opted to use the newly added feature of editing the LLM-generated question with their own question within the question palette 11.1% of the time, and used the two static questions approximately equally (*Expand*: 27.2% of clicks, *Define*: 22.8%). Many of the questions that participants asked were further threaded, i.e., not on the original abstract (30.3%), suggesting the Qlarify encouraged participants to ask follow-up questions and dive deeper into specific parts of the paper through the abstract.

One disadvantage participants noted in the Qlarify condition was that the requirement of grounding in

the abstract for generating expansions made it harder to ask general questions about the paper. Although possible by selecting a span and overwriting the AI-suggested question, this process felt less natural than asking questions in a blank text box. For P5, Qlarify and QA met different needs, saying that QA “allows overall questions I had about the paper” while “Qlarify is useful for a deeper dive, aiding me in better understanding particular parts of the paper.” Other participants used the QA condition to emulate familiar LLM-based chat interfaces; for example, one prompted multiple abstracts with, “Explain this paper to me like I’m a 10 year old” (P1). Consequently, many participants mentioned an ideal system would combine both conditions, offering interactive expansions in the abstract for grounded questions and a text box for general questions.

Comparison with Current Navigation Strategies (RQ4)

Overall, participants reported feeling more satisfied with their exploration of abstracts and papers in the Qlarify condition (6.2 ($SD=0.8$), $Mdn=6.0$ ($IQR=5.8-7.0$)) than the Manual (4.8 ($SD=1.5$), $Mdn=4.0$ ($IQR=4.0-6.25$)) and QA (5.5 ($SD=1.7$), $Mdn=6.0$ ($IQR=4.8-7.0$)) conditions. A Friedman test yielded a difference between the three conditions ($\chi^2(2) = 8.4, p=.01$); post-hoc tests found a significant difference between Qlarify and Manual ($p=.02$). Participants also felt they were able to gain more information during their exploration in the Qlarify condition (5.8 ($SD=0.8$), $Mdn=6.0$, $IQR=(5.0-6.0)$) than in Manual (3.8 ($SD=1.4$), $Mdn=4.0$ ($IQR=3.5-5.0$)) and QA (5.1 ($SD=1.4$), $Mdn=5.0$ ($IQR=4.0-6.0$)); $\chi^2(2) = 9.6, p=.01$; post-hoc tests found a significant difference between Qlarify and Manual ($p=.02$).

Reinforcing observations of Qlarify scaffolding the question asking process, participants reported greater motivation to explore deeper into papers in the Qlarify condition (6.3 ($SD=1.1$), $Mdn=7.0$ ($IQR=5.8-7.0$)) than in Manual (4.2 ($SD=0.6$), $Mdn=4.0$ ($IQR=4.0-4.3$)) and QA (5.8 ($SD=1.4$), $Mdn=6.0$ ($IQR=4.8-7.0$)); $\chi^2(2) = 15.4, p=.0004$; post-hoc tests found a significant difference between Qlarify and Manual ($p=.002$). Participants mentioned how Qlarify enabled them to retrieve information that felt “detailed and useful,” allowing them to understand for instance, “what the authors had done, the authors’ use of the terms they developed, the things that they built on” (P8). In contrast, they found the QA condition sometimes delivered information that was “too high-level,” likely due to the more generic nature of questions they asked when unguided.

All participants appreciated the added value provided by Qlarify and reported a greater desire to use the Qlarify (6.3 ($SD=0.9$), $Mdn=6.5$ ($IQR=5.8-7.0$)) condition compared to Manual (3.5 ($SD=1.5$), $Mdn=4.0$ ($IQR=2.8-4.0$)) and QA (5.3 ($SD=1.7$), $Mdn=6$ ($IQR=4.5-6.3$)); $\chi^2(2) = 10.2, p=.006$; post-hoc tests found a significant difference between Qlarify and Manual ($p=.01$). Across all survey questions, participants reported measures that favored Qlarify over QA (Figure 3.6), though none of these differences were significant after post-hoc correction for the multiple statistical tests in our study.

On average, we found participants viewed a similar number of abstracts across the three conditions (Manual: 6.0 ($SD=3.4$), QA: 6.8 ($SD=3.5$), Qlarify: 6.3 ($SD=2.5$)), but opened fewer paper PDFs in the QA and Qlarify conditions (Manual: 5.3 ($SD=3.2$), QA: 4.0 ($SD=1.7$), Qlarify: 3.7 ($SD=1.8$)). Furthermore, participants spent less total time in the paper PDFs in the Qlarify condition (Manual: 538.8s ($SD=223.5s$), QA: 262.1s ($SD=97.2s$), Qlarify: 201.1s ($SD=96.7s$)). Together, these findings suggest participants were less inclined to view the full paper when provided with some affordances for information retrieval in the abstract, and spent less time skimming through the full PDF when an expandable abstracts were available.

Lastly, in the Qlarify condition, we found 10 of 12 participants drilled down to view attributed evidence for an expansion highlighted in the PDF at least once during the study. When participants opened the PDF via attribution, they spent on average 32.4s ($SD=21.0s$) in the PDF. In contrast, whenever participants opened the PDF from the title, they spent longer on average scrolling through the PDF, 55.6s ($SD=30.0s$). These findings suggest participants used the attributed evidence as a quick entry point into the full paper, often as a “guide back to the paper” to “cross-check whatever the AI generated” (P9).

3.7 Evaluation of Generated Expansions

We conducted a small-scale evaluation of the quality of expansions generated within Qlarify, using a subset of the data collected in the deployment study. Members of the research team annotated 120 randomly sampled expansions for which an answer was found (30 for each of the *Define*, *Expand*, *Why*, and AI-suggested questions).

Of the 120 expansions, 105 (87.5%) were entirely accurate (i.e., all statements were grounded in verifiable information from the paper) corroborating perceptions of high expansion quality by scholars in our interview study. While we did not explicitly assess relevance, our annotation and participant observations

from our interview study suggest that the LLM-generated expansions were largely relevant to the queried information. Two of the main sources of error within the analyzed expansions included:

Inaccurate details. Seven (5.8%) expansions included detail inaccuracies, often involving numerical or mathematical content. These included false navigational references (e.g., attributing statements to an incorrect section in the paper), numerical values in experimental results (e.g., fabricated numbers in “the additional mean overhead time of 0.47s is only 12.8% of the average episode duration of 3.67s”), and acronyms (e.g., describing FMs as “language guided models” rather than “foundation models”).

Missing content. Eight (6.7%) expansions contained phrases such as, “the paper does not provide explicit details for. . .” This error tended to occur when a portion of the relevant information was provided in a table rather than in the body text of the paper. Rather than a limitation of LLMs, this perhaps reflects how Qlarify preprocesses papers into a flat representation without delimitation for structural or visually salient content such as tables. In other cases, the context provided to the LLM may have lacked sufficient information to answer the question, suggesting the need to further investigate robust chunk retrieval techniques.

Overall, our analysis highlights the infrequent yet subtle errors within retrieval-augmented generation approaches, such as how plausible yet hallucinated details can become embedded into an otherwise accurate expansion. It is worth noting that no scholars in our interview study explicitly noticed any errors, and we identified errors only through extensive checking with the original paper. These hallucinations can thus be challenging to detect—especially within cognitive demanding or rapid exploratory processes such as triage—and potentially lead to harmful misinterpretations and an erosion of trust in the reliability of the generated expansions.

3.8 Related Work

3.8.1 Addressing Document-Centered Information Needs with Summarization

The aim of summarization is to condense long documents into short and concise texts, encapsulating the most important information required for comprehension. With the advent of neural architectures [Koh et al., 2022], significant efforts have been dedicated to improving the capabilities of automatic text summarization systems. Some work has focused especially on summarizing domain-specific and long-form documents,

such as scientific text [Cachola et al., 2020; Yasunaga et al., 2019; Ibrahim Altmami and El Bachir Menai, 2022]. Consuming a traditional summary requires little to no user effort, but the static medium of a summary makes it impossible to capture the personalized and nuanced information needs of every individual. Some research has explored ways to incorporate humans in the loop to generate more personalized summaries [Zhang et al., 2023; Shapira et al., 2022; Ghodratnama et al., 2021], but once generated, these summaries lack the ability for iterative refinement, for instance to reflect updated information needs.

Prior work has explored hierarchical approaches to summarization that enable a reader to interactively specify the degree of depth they wish to explore. These approaches require the authoring of summaries at differing levels of depth (e.g., for summarizing large-scale online discussions [Zhang et al., 2017], books [Wu et al., 2021], or web documents [Bernstein, 2009; Nelson, 1983]), forming a static summary tree artifact that enables structured navigation between the summaries and long documents. These summaries are typically constructed in a bottom-up fashion, working recursively starting from the full text, so as to break down the work into more manageable chunks. In contrast, we take a top-down approach; rather than condensing information to *generate* summaries, we instead leverage automated techniques to *expand* summaries—incrementally and recursively—to reveal relevant information from a long document on-demand. By constructing expandable summaries in the same direction that readers explore (i.e., by drilling down), the summary trees generated by our system are personalized to each reader’s exploration, as opposed to having to conform to a single rigid structure.

This concept of expanding text to incorporate more context and resolve ambiguities is also related to decontextualization, an NLP task studying automated approaches for rewriting extractive excerpts divorced from a longer document to be interpretable out of context while preserving meaning [Choi et al., 2021; Newman et al., 2023]. The expansion interaction we explore can be seen as a form of interactive decontextualization, sensitive to users’ personal information needs.

3.8.2 Querying Documents with QA Systems

People often express their information needs within documents as natural language questions [Jahanbakhsh et al., 2022; ter Hoeve et al., 2020; Ko et al., 2020], and ask follow-up questions when an initial answer is not sufficient [Meng et al., 2023]. Question answering is a long-standing problem in NLP, including con-

siderable efforts for answering questions over scientific documents [Dasigi et al., 2021; Rogers et al., 2023; Saad-Falcon et al., 2023]. Recent advances in LLMs have seen the proliferation of prototypes for conversational question answering over long documents (e.g., *ChatDoc* [ChatDOC, 2023], *ChatPDF* [ChatPDF, 2023]). The expansion interaction we propose in this work is related to these systems, but differs in two main ways: first, we aim to integrate answers in-situ by expanding the summary at the point where a question was asked to maintain the flow of reading, and second, we introduce mixed-initiative interactions that reduce the cost of forming and asking questions. To enable these question-driven expansions, we implement a retriever-reader architecture with a dense representation-based retriever and a generative, LLM-powered reader.

Attributed question answering [Bohnet et al., 2023], where AI models return evidence in addition to their answer, has gained recent interest due to the potential for hallucinations in LLM-generated answers [Maynez et al., 2020; Ji et al., 2023]. Some work has sought to more closely integrate attribution with generation (e.g., with post-editing [Gao et al., 2023]), while systems such as *GopherCite* [Menick et al., 2022], *We-bGPT* [Nakano et al., 2022], and *LaMDA* [Thoppilan et al., 2022] place the burden of fact verification on the user by providing URLs and excerpts as supporting evidence. We take a similar user-centric approach to attribution, providing simple affordances for users to drill down into excerpts and then the full paper from an LLM-generated response.

Prior conversational interfaces with documents often make two assumptions: that users know what to ask, and that the most natural interaction is for users to manually type their intended questions. Yet this is not always the case [August et al., 2023]. Instead, we propose that carefully crafted interactions and language understanding techniques can effectively infer the intent of users, reducing the cost of asking questions to a single click. This work explores the potential for intelligent support to not only answer users' questions, but also suggest questions prompting relevant entities to expand. Recent work has begun to explore the effectiveness of LLMs in this task of question generation, such as for generating follow-up and clarification questions [Meng et al., 2023; Ko et al., 2020; Kumar and Black, 2020]. As prompting LLMs can be challenging [Zamfirescu-Pereira et al., 2023], QLARIFY obviates the need for users to craft their own prompts by providing a selection of common questions. These questions are grounded in a taxonomy of document-centered information-seeking questions people ask while reading a document [Ko et al., 2020],

and refined for scientific documents through our formative study.

3.9 Discussion

In this chapter, we proposed *recursively expandable abstracts*, a novel document-centric interaction technique that dynamically elaborates on abstracts along directed information needs with details retrieved from the full texts of papers. Findings from our studies reveal how Qlarify can support the retrieval of information from a paper on-demand through one-click questions, bridging the informational gaps between an abstract and a paper. In this section, we discuss potential extensions of Qlarify and opportunities for future generalizations of LLM-powered, interactive expansions.

Our exploration of a fluid text medium within Qlarify leaves exciting design dimensions of the interaction for future research. As suggested in §3.2.2, abstracts could be similarly expanded for information needs not grounded in the abstract’s text. For instance, expanding abstracts with agnostic information needs comprising questions that exist divorced of any specific abstract (e.g., “What are their contributions”) could draw on prior work extracting key faceted information within papers [Fok et al., 2023; Chan et al., 2018]. Qlarify also currently supports users in verifying the accuracy of information within the LLM-generated expansions by surfacing attributed evidence and linking to relevant paper context. Unlike informal information retrieval settings such as online search, which are tailored to improve user satisfaction by providing *some* answer possibly at the cost of accuracy, the scholarly context necessitates a higher standard for the factuality and faithfulness of retrieved information [Morris, 2023a]. The utility of Qlarify and similar scholarly support tools could be improved by further embedding accuracy awareness into the generated expansions, for instance by eliciting and indicating the system’s overall accuracy or (well-calibrated) confidence in specific generated expansions [Tian et al., 2023; Xiong et al., 2024].

For information needs that were not grounded in specific parts of an abstract, study participants also acknowledged the value in freely asking any question to a paper, similar to the QA condition of the comparative evaluation. Complementing the AI-suggested expandable entities in Qlarify which offer structured guidance toward likely relevant information, allowing users to freely ask their own questions could encourage a more self-driven and critical exploration of papers. Future systems could offer both affordances to more flexibly address scholars’ various information needs.

Our studies also reveal opportunities to better support context switching between the reading of abstracts and full papers. For instance, systems could include additional entry points into a paper from the abstract. Qlarify provides one per expansion by allowing scholars to drill down into a paper’s context through attributed evidence, which we found helped participants validate and build trust in the AI-generated expansions. Participants noted how this interaction could go further by scaffolding the transition between triage and reading, for instance by guiding attention to multiple related passages throughout a paper for an expansion. An interesting question is then how these systems can help scholars retain their newly acquired knowledge as they return from the full paper back to the abstract triaging process.

Moreover, participants in our studies used the ability to drill down into a paper not only to validate the generated expansions, but also as a means to quickly open the paper to browse visual content such as a system diagram or a table of experimental results. Rather than the text-only modality of Qlarify, future expandable abstracts may learn to surface graphical content from papers, such as figures and tables, which participants often desire in their exploration. Expandable abstracts could benefit from personalization, with systems learning from a user’s expansion history to suggest tailored expandable entities for new abstracts to lower interaction costs and encourage exploration, or even regenerate abstracts tailored to a user’s interests.

Although Qlarify was designed and evaluated with papers largely within computer science, we believe the underlying interaction paradigm can transcend disciplines. For instance, prior work has shown that reading medical literature can be challenging and overwhelming due to barriers such as dense and unfamiliar terminology, not knowing what to read, and the inability to find answers to specific questions [August et al., 2023]. One avenue for future work could investigate how expandable abstracts with similar question asking and answering assistance could make medical papers more accessible by providing just-in-time responses to questions directly within an abstract, without requiring lay readers to interact with the full paper. A similar idea might allow policymakers to understand the implications of scientific developments, a critical need [Tyler et al., 2023]. Future studies could also examine the potential for recursive expansion interactions over summaries of documents in other domains (e.g., legal documents, medical notes, or discussions in online forums). Moving beyond comprehension of a single paper and into synthesis across papers, another direction could explore support for expansion contexts outside abstracts or with multiple documents. For example, could Qlarify be extended to allow users to expand related work sections and explore information

across many cited papers, synthesizing the information retrieved across these multiple documents?

3.9.1 Limitations

While many participants appreciated the AI-suggested expandable entities, some felt its visual salience could hinder reading of the underlying abstract. Participants also found the in-situ expansions could interrupt their reading flow, especially if the generated answers were verbose or inaccurate. To mitigate these issues, future systems could provide customizability for the presentation (e.g., fluid text, sidebar, popup) and quantity (e.g., number of expandable entities) of AI-enabled reading augmentations.

Beyond user control, the development of scholarly support tools also necessitates consideration of the implications of deploying such systems. While LLM-generated questions within Qlarify can help scaffold an active reading process, they also reduce a scholar's agency over their exploration. Interactions such as abstract expansions could discourage scholars from reading full papers, instead encouraging more superficial exploration through interactions with abstracts only. Several participants in our studies noted how over-reliance on the “path of least resistance” offered by such tools could undermine the knowledge and self-actualization gained from years of triaging and consuming research literature by themselves, and potentially harm new scholars in particular. Nonetheless, we believe these tools can significantly enhance the efficiency and efficacy of scholarly processes, though they should be developed mindful of these risks and pedagogical implications.

Finally, expandable abstracts (and many scholarly support tools) require access to a paper's full text or PDF. In this work, we had the privilege of institutional access to the full text of many papers, but we note a significant portion of science remains inaccessible behind paywalls enacted by academic publishers. While legal and institutional challenges remain, open access initiatives (e.g., the Open Access movement², arXiv, S2ORC [Lo et al., 2020]) have made notable strides in changing the landscape of publishing to encourage more accessible dissemination of scientific knowledge.

²See <https://www.doaj.org/> or <https://www.budapestopenaccessinitiative.org/read/>

3.10 Conclusion

In this chapter, we introduced *recursively expandable abstracts*, a novel text interaction paradigm enabling direct expansion of abstracts with clarifying information from paper full texts, and demonstrated it within an augmented reading interface system, QLARIFY. To support the creation of expansions, QLARIFY uses LLMs to identify entities within abstracts that are informative to expand, suggest intent-inferring questions that scholars can ask in a single click, and generate concise, attributed responses. Through an interview study, we found expandable abstracts helped scholars to rapidly and deeply address information-seeking needs during paper exploration. A field deployment further characterized how scholars would use expandable abstracts for paper triage in a real-world setting. Our final (comparative) evaluation showed that participants felt they could explore more deeply and were more satisfied with QLARIFY than with plain abstracts, even if question answering was available. We believe QLARIFY contributes a valuable step toward LLM-enabled systems that effectively empower an interactive, low-cost, and just-in-time exploration of long, complex texts.

Chapter 4

Scim: Guiding Attention in Paper Skimming with Faceted Highlights

While QLARIFY addressed progressive exploration during abstract-to-paper transition, once relevance is determined, researchers spend significant time and cognitive effort engaging with full paper texts. In this chapter, we present SCIM, an augmented reading interface system which demonstrates structured mediation through AI-assisted discovery via suggested faceted highlights that organize content by rhetorical function (Objective, Novelty, Method, Result) while preserving reading flow. In contrast to external summaries, SCIM provides intermediate representations through these color-coded highlights that offer information scent for attention guidance while naturally maintaining direct access to the underlying paper text by design, affording verification (and the complementary principle of easy recovery from AI errors [Amershi et al., 2019]). Through a series of lab and longitudinal user evaluations, we find that SCIM reduces information-seeking time, helps researchers focus attention on key content, and proves particularly valuable for dense texts and unfamiliar domains, demonstrating how AI-generated cognitive scaffolds via faceted highlighting can effectively augment rather than replace human reading strategies. This work was previously published as a conference paper at ACM IUI 2023 [Fok et al., 2023] and later as an extended version for ACM TiiS [Fok et al., 2024c], and has been adapted for this document.

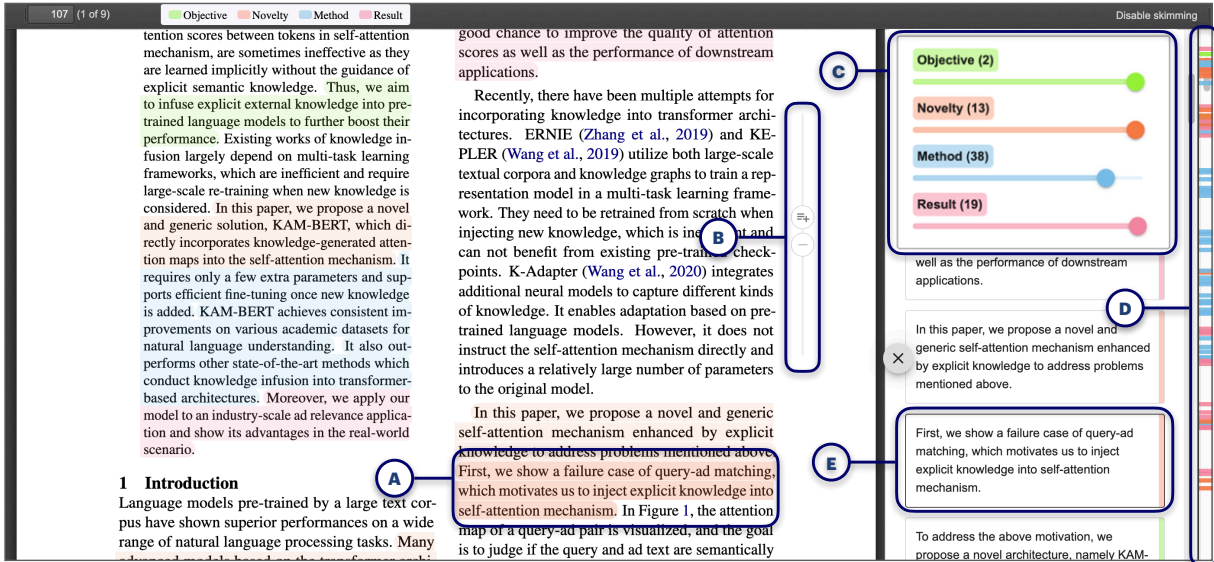


Figure 4.1: SCIM is an intelligent reading interface for skimming scientific papers. To help readers develop a broad overview of content in a paper, SCIM intelligently highlights passages (A). The passages are colorized to indicate the rhetorical role of the passage, i.e., whether it describes the research’s objectives, novelty, methods, and results. Highlights are distributed throughout the text to support a holistic skim. Readers can request additional (or fewer) highlights by using paragraph-local (B) and paper-wide (C) controls. To understand where to find information of a certain kind, readers can glance at highlight markers in the scroll bar (D). Readers can also collect an overview of the paper by reviewing highlighted passages in a sidebar (E).

4.1 Introduction

With the rise of knowledge work and a contemporaneous explosion of information, experts are expected to sift through and make sense of large volumes of rapidly evolving information. One domain where this trend is particularly pronounced is scientific research. Researchers spend a tremendous amount of effort staying up to date with the literature. They do so by regularly undertaking the tasks of foraging for papers, skimming or reading those deemed most relevant, and integrating knowledge gained from reading into their personal records.

Skimming is a critical task, and requires researchers to quickly review the contents of a paper to develop a cursory understanding of its contents. While faster than reading, skimming achieves a coarser view of papers’ contents. With the shift of scientific publishing from paper to digital online publications, the practice of skimming has become yet more widespread [Liu, 2005; Tenopir et al., 2009]. Despite the pervasiveness of skimming as a practice for reviewing papers [Rayner et al., 2016], skimming is not easy [Maxwell, 1972;

Duggan and Payne, 2009]. Skimming may devolve into reading should a reader find themselves drawn into the details of a passage. Even for experienced readers, skimming requires attention to make strategic choices of what to read, where, and when to stop reading.

In this chapter, we explore how an intelligent user interface can help both novice and expert researchers skim scientific papers more efficiently. Today, AI-powered techniques are increasingly used in tools for searching over the scholarly literature (e.g., [Beel and Gipp, 2009; Ammar et al., 2018]) and in scientific reading applications (e.g., [Lo et al., 2023; Head et al., 2021; August et al., 2023]). This chapter explores how intelligent tools can facilitate the task of skimming, the seam between searching and reading.

As a starting point, we ask how judicious use of automatic highlighting can be presented in a tool to help readers direct their attention while they skim. To gain inspiration for designing such a tool, we conducted formative studies with researchers, including interviews, observations, and pilot studies of prototype highlighting tools. In these studies, we found that readers desired highlights that cover diverse content, are evenly distributed across a paper, and capture important paper content. These studies reveal a tension between reader expectations and system design, because it is not always possible to highlight according to passage importance while achieving a desirable distribution of highlights. Readers also desired some influence over the quantity and distribution of highlights within a paper.

We incorporated these insights into the design of SCIM,¹ a prototype, intelligent user interface for skimming scientific papers (Figure 4.1) and subsequently into skimming features within Semantic Reader, a production-grade reading tool (Figure 4.6). To address readers' needs around highlighting, SCIM highlights passages in the following ways. First, passages are highlighted with distinct colors for each of four diverse kinds of content sought by readers: research objectives, novel aspects of the research, methodology, and results. Second, SCIM aims to support an evenly-distributed skim of a paper, highlighting passages in a way that most paragraphs contain at least one highlighted sentence. Finally, SCIM lets readers customize the number of highlights in a paper, both across an entire paper and within individual paragraphs.

We conducted a sequence of studies to evaluate SCIM's utility as an intelligent skimming tool. First, we performed a lab usability study to understand how SCIM affects readers' ability to search for specific kinds of information in a paper. When using SCIM, readers located the desired information in significantly

¹Code available at <https://github.com/rayfok/scim>

less time compared to a standard document reader, with comparable effort and accuracy. Second, to understand more realistic usage, we conducted a two-week-long diary study. In this study, readers found SCIM particularly useful when skimming text-dense passages with few visuals, or when skimming a paper that fell outside their area of expertise. SCIM became more usable over time as readers became accustomed to the highlights. The study also suggests how skimming assistants could be improved in the future, for instance by highlighting passages that provide background for later highlighted passages, and integrating highlights with the typographical emphases authors may have already provided, such as boldface font and text formatting.

Before deploying skimming capabilities to an even larger population of users, we conducted another round of iterative design with several aims. First, we sought to make the skimming features more inclusive for blind and low vision users with marginal flags, user-controllable color intensities, and a screen-reader compatible side bar. Second, we evaluated and improved the quality of the salient passage classifiers, while re-engineering them to scale to process millions of documents. We then conducted a large-scale randomized controlled trial to understand skimming usage and check if skimming might decrease important metrics such as the number of papers read or the number of return visits. We found that 19% of users exposed to skimming capabilities used the features, and with no significant effect on guardrail metrics. As a result, skimming was enabled for all users in Semantic Reader. The SCIM system and its accompanying studies contribute to a vision of how AI-enabled augmentations can help readers rapidly attend to and make sense of complex information in scientific papers through intelligent faceted highlights.

4.2 Background and Related Work

4.2.1 The Skimming Process

In the literature, skimming is characterized as a form of rapid reading in which the goal is to get a general idea of the text or visual content, typically by focusing on information relevant to one’s goals and skipping over irrelevant information [Masson, 1983; Rayner et al., 2016]. Skimming is a necessary and useful skill for researchers. As the number of published papers increases year over year and papers have moved from print into digital media, scholars have tended toward reading more papers and spending less time on each,

likely doing so by skimming [Tenopir et al., 2009; Liu, 2005].

The psychology literature describes skimming as a cognitively demanding task. In this task, readers incrementally build a mental model of the text and integrate information across sentences as they read [Rayner et al., 2016; Rapp and van den Broek, 2005; Tashman and Edwards, 2011]. Generally, readers are not accurate at identifying goal-relevant information within text. Skimming is also physically demanding—limitations in the oculomotor system, which is responsible for controlling eye movements, preclude rapid, accurate placements of eye gaze for extended periods of time, such as when a reader skims a long document [Masson, 1982, 1983].

Amidst the challenges of skimming, success is often determined by a reader’s ability to “satisfice” [Reader and Payne, 2007; Duggan and Payne, 2009, 2011]. Satisficing is a skim reading strategy where a reader sets a threshold of how useful information should be to deserve their attention, and if a unit of text falls below that threshold, they skip to the next unit of text. Studies have found that readers tend to spend more time at the beginning of paragraphs, the top of pages, and the beginning of documents [Duggan and Payne, 2009], perhaps in part because this information is often believed to have high relevance. When grounded in particular information needs, skimming for rapid high-level comprehension is also often complemented by scanning, comprised of saccadic movements through a text in search of specific pieces of information.

One study of skimming for scientific document triage found readers were hasty and incomplete, with readers scrolling through documents quickly and paying attention to highly visual content and section headers [Loizides and Buchanan, 2009]. Scientific documents are laden with visual content, typographical cues (e.g., italicized, bold, or colored text), and structural information. Readers draw on document features to support rapid comprehension via these macro- and micro-structures [Carrell, 1985; Lacroix, 1999; Machulla et al., 2018] and visual content [Yi, 2014; Keshav, 2007]. In this chapter, we explore how automated assistance may support skimming by cueing readers towards significant sentences that might otherwise be missed. SCIM’s use of highlighting lets readers continue to pay attention to traditional visual and structural landmarks, while also heeding the passages highlighted by the skimming assistant.

4.2.2 Tools for Reading and Skimming

Researchers have long sought to equip readers with tools to support and augment their cognition while reading documents. The nascent days of human-computer interaction saw the introduction of augmented reading interfaces to support the reading process, including fluid documents that provided contextual access to supplemental information between lines of text [Chang et al., 1998], fluid hypertext [Zellweger et al., 1998], visualizations for social annotations within papers [Hill et al., 1992], and affordances for annotating papers and jumping readers to passages of interest [Graham, 1999; Schilit et al., 1998]. Since then, several approaches have been proposed to support the various aspects of reading, such as document navigation and comprehension.

Modified Scrolling Interactions

One line of research sought to facilitate the rapid exploration of long documents by modifying the behavior of reading interfaces during scrolling. Applications of content-aware scrolling were used to redefine the presentation order of content within a document [Ishak and Feiner, 2006], provide pseudo-haptic feedback when scrolling past relevant information [Kim et al., 2014], and dynamically resize document headings within paper thumbnails in a document viewer [Buchanan and Owen, 2008]. The Spotlights project implemented an attention allocation technique which pinned headings and figures as static overlays to a document as it was continuously scrolled [Lee et al., 2016].

Typographical Cueing

Another approach involved augmenting reading interfaces with typographical cues, e.g., highlighting. Studies in cognitive psychology have found visual cueing mechanisms can be effective in focusing reader attention [Chi et al., 2007] and improving retention of material [Fowler and Barker, 1974; Rickards, 1980]. The Semantize system used highlights to visualize sentiment within a document, and underlined words with positive or negative sentiment in different colors [Wecker et al., 2014]. The ScentHighlights system used highlights to identify conceptually relevant text based on a user's query [Chi et al., 2005]. The HiText technique introduced dynamic graded highlighting of sentences within a document in accordance with their salience [Yang et al., 2017]. Modern reading interfaces also commonly support readers in marking regions

of interest with a document with highlights or free-text annotations. The pervasiveness of highlighting as a technique for drawing readers' attention can be attributed to the von Restorff isolation effect, which states an item isolated against a homogeneous background will be more likely to be attended to and remembered [von Restorff, 1933; Hunt, 1995]. Studies have since found evidence of this effect on the visual foraging behavior of readers, finding that highlights attract about half of the total number of fixations within a document, and readers' eyes are often drawn to them [Chi et al., 2007]. In this chapter, we explore how typographical cues such as highlighting can augment the paper skimming process. In contrast to existing systems, we introduce an approach to generate these highlights automatically, and further organize the suggested highlights into semantically meaningful facets to facilitate machine-supported navigation throughout a paper.

Document Augmentations

Beyond typographical cues, other reading interface augmentations exist to specifically support the reading of scientific papers. For instance, online paper providers like Springer, PubMed, and Semantic Scholar provide readers with in-context citation information. Experimental systems have linked document text to marks within charts [Kong et al., 2014] and cells within tables [Kim et al., 2018b], generated on-demand visualizations based on text within the paper [Badam et al., 2019], transformed static visualizations with animated [Grossman et al., 2015] or interactive [Masson et al., 2020] overlays, augmented citations with cited paper [Rachatasumrit et al., 2022] and personalized contexts [Chang et al., 2023], integrated segments of authors' talk videos into corresponding paper passages [Kim et al., 2023], and provided in-context definitions for nonce words [Head et al., 2021]. We design SCIM with inspiration from many of these prior reading interfaces, presenting a novel augmentation technique to support the process of skimming papers by visualizing information scent through the identification and highlighting of salient sentences.

Summarization

An alternative method to skimming a full paper is to read a shortened representation of the paper's content in the form of a summary. An author-provided summary is de facto included with each paper as an abstract, which researchers often read before continuing to the rest of the paper. Automated summarization has garnered significant interest from the natural language processing community, and many extractive and ab-

stractive methods for generating summaries from long-form documents have been developed over the years (e.g., see [Nenkova and McKeown, 2012; Sefid and Giles, 2022] for a survey). Some methods have even been proposed for generating extreme (single sentence) summaries, or TLDRs, from full papers [Cachola et al., 2020].

However, summaries are often unsatisfactory. Despite recent advances, automated summaries remain error-prone, susceptible to hallucination [Zhao et al., 2020], and unreliable as a standalone replacement for reading a paper itself. Furthermore, summaries do not provide readers with the ability to quickly interact with the full paper. As readers’ goals and interests change while reading, they may wish to explore certain sections in further detail. Unlike summaries, augmented reading interfaces like the one we explore in this chapter naturally retain the context of the paper.

4.3 Design Goals

To better understand how to design usable, intelligent skimming interfaces, we used an iterative design process that began with interviews and observations of academic researchers (referred to as *readers*), and continued into an evaluation of an early prototype of SCIM. In this section, we first describe that design process, and then distill the lessons learned from this formative research into a set of design goals (Section 4.3.2) that inform the design and implementation of intelligent, highlighting-based skimming support tools.

4.3.1 Design Methodology

Formative interviews and observations

We conducted formative study sessions with eight readers (F1–8) to better understand how they skim scientific papers. All readers belonged to the target user group for SCIM, and were either graduate students or academic faculty. Readers were first observed as they skimmed a paper of their choice. Then, readers were asked to describe their skimming process, including goals they held while skimming, strategies they employed, and any aspects of skimming they found difficult or tedious.

Prototype development and evaluation

A prototype of SCIM was iteratively designed and developed based on our formative interviews and observations. While many kinds of tools could support skimming, our design exploration focused specifically on skimming aids which incorporate intelligent highlights.

The prototype was similar to the version of the SCIM system later described in Section 4.4, with a few differences. First, the prototype’s highlighting policy was different, resulting in fewer highlighted passages, and a less uniform distribution of highlights. Second, the prototype had no paragraph-level or facet-specific controls for the number of highlights, but rather only global-level controls on the number of highlights and switches to turn on or off individual facets.

Two preliminary usability studies were conducted with this prototype. Thirteen readers (E1–13) were recruited from university mailing lists, and via direct outreach following purposive and snowball sampling approaches. Sessions in both studies were one hour in length and conducted on the Zoom platform. In both studies, readers skimmed papers with SCIM for a limited amount of time and completed a task demonstrating their understanding of the paper, for instance outlining the paper or answering questions about the paper. Afterward, readers were asked to comment on their interactions with SCIM and what aspects of the system required improvement.

Analysis

One author conducted analyzed data from the formative study and preliminary evaluations following a thematic analysis methodology [Blandford et al., 2016, Ch. 5]. Notes and transcripts from study sessions were analyzed for themes and supporting evidence. Themes were validated through discussion and review with a second author. Those themes that provided actionable guidance for design are reported in the next section.

4.3.2 Design Goals

We present seven design goals for intelligent highlight-based skimming interfaces, based our formative research.

DG1. Augment readers’ skimming practices. Readers described myriad strategies they already used to skim papers. One common strategy was to first read the abstract and introduction of a paper. Then, readers

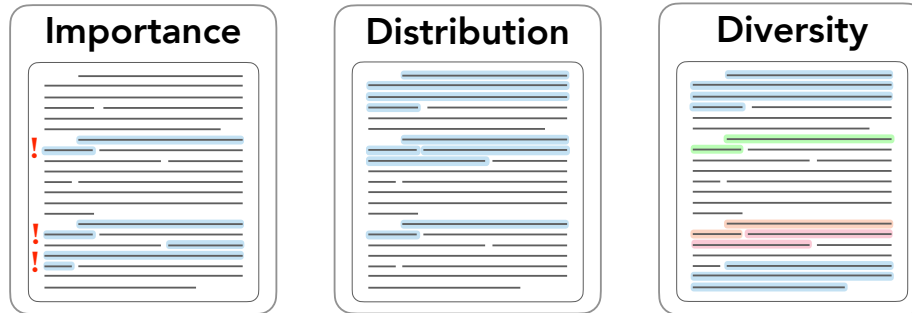


Figure 4.2: Our formative research revealed that intelligent highlights need to do more than pointing readers to important content. They should also be *well-distributed* throughout a paper (D3, D5) and steer readers towards *diverse* content types (D2).

consulted other key material in the paper, including bulleted lists of contributions (F1, F4, F6), summaries of results (F1–3), and conclusions (F1, F3, F6, F7). Readers also employed strategies particular to their goals, the paper, or their level of comfort with the paper. Readers relied on various visible cues in the text to help them identify important information, including typographical cues (e.g., italics, boldface) (F3, F6), structural cues (e.g., section headers) (F2, F6), visuals (e.g., figures and tables) (F1, F2, F4, F6, F8), and text position (e.g., inspecting the first sentences of paragraphs) (F2, F3, F6). We believe skimming interfaces should not impede or replace these reading strategies.

DG2. Highlight diverse kinds of content. Readers’ skimming goals were diverse. For instance, some readers sought to learn specific techniques introduced in a paper (F1), and others wished to understand a paper’s relationship to prior research, or discover new research directions (F2–4, F7). Some desired a high-level understanding suitable for discussing the paper with colleagues (F3, F7). These goals influenced readers’ skimming strategies, leading them to look for answers to different sorts of questions. We suggest skimming tools should support readers’ diverse goals by enabling review of varied aspects of paper contents.

DG3. Support skimming in the lengthy middle sections of the paper. Readers noted that while one recommended strategy for skimming is to read the beginning and ends of paragraphs, important content may reside in the middle of paragraphs. When asked to skim, we often observed readers transitioning into a deep read of some passages in the paper (F1, F3, F5). We propose that skimming tools should help readers identify important passages which conventional strategies do not reach, such as content in the middle of paragraphs and in the middle of the paper.

DG4. Minimize distraction. Without careful visual design, an augmented reading tool can occlude text

or misdirect readers' attention. Our early prototypes incorporated a variety of text highlighting techniques, including underlines, lowlighting unimportant paper contents (inspired by ScholarPhi [Head et al., 2021]), and highlighting text by setting its background color. Underlining was too subtle to consistently catch the reader's eye. Lowlighting tended to distract readers, requiring additional effort to read lowlighted content. Highlighting was chosen for its familiar use in documents, with the colors tuned to distinguish the categories of text and minimal contrast to avoid an unpleasant visual pop-out effect. We suggest that other designers similarly aim to minimize the visual distraction introduced by design interventions.

DG5. Supply enough highlights. In our preliminary usability studies, readers often felt uncomfortable when they saw long, unhighlighted passages where they thought important information likely could be found. Some readers wanted to see highlights distributed more uniformly throughout the paper (as opposed to highlights concentrated primarily in an introduction or conclusion). We suggest the rule of thumb that a highlight should be provided around once per paragraph, and that readers should be able to request additional highlights in particularly dense passages.

DG6. Provide accurate highlights. A side effect of introducing faceted highlights (where highlights are color-coded by their predicted rhetorical category) was that classification errors became obvious to readers, such as when a passage about results was labeled as being about methods. Readers found themselves distracted when the classification of a passage clashed with their expectations and became skeptical of the tool's capabilities (E11, E12). If skimming tools provide faceted highlights, it is especially important to classify these categories correctly.

DG7. Support user control and personalization. Readers desired more control over the amount of highlights shown by the prototype. Many suggested that the tool could help them fine-tune what was highlighted, either through manual adjustments, or with adaptive personalization of the highlights (i.e., responding to passages a reader has highlighted themselves or highlights they have deleted) (E5, E7, E8, E12).

A final takeaway from our formative research was that readers believed their comfort using intelligent highlights would change over time, as they became more familiar with the features, the colors associated with the highlights, and the accuracy of the highlights. One reader described this as the issue of "getting used to seeing highlights that aren't my own" (E13). This observation motivated our choice of a longitudinal

diary study (Section 4.7) and production deployment (Section 4.8.3) as summative evaluation methods for SCIM.

4.4 SCIM User Interface

We now describe the design of SCIM, an interface that provides intelligent support for skimming scientific papers, and explain how particular aspects of the system address our design goals (D1–7).

4.4.1 Overview

A reader interacts with SCIM as a tool that supports and augments their typical skimming process (D1). One common strategy for readers is to begin with a paper’s title and abstract, followed by a piecemeal review of the paper. A reader can employ this strategy, and at the same time follow the highlights offered by SCIM, which extend into parts of the paragraph that a reader may not notice otherwise (D3).

4.4.2 Faceted Highlights

SCIM intelligently highlights a paper to direct a reader’s attention to key passages (Figure 4.1.A). These highlights were tailored in three ways to support skimming.

Faceted. Because readers have different goals when skimming, SCIM colorizes highlights according to facets of information (D2). To promote memorability we limit the number of facets to four. The specific set of facets was selected to encompass the kinds of information participants described in the formative study, balanced by the requirement that we could detect them reliably (D6), as described in the implementation section.

Numerous schemes exist for sentence-level classification of scientific literature into facets. Coarse-grained schemes classify sentences according to common section names from scientific papers (e.g., [Hirohata et al., 2008; Cohan et al., 2019]) and consist of a small number of facets. Other fine-grained schemes rely on argumentative zones and conceptual structure (e.g., [Teufel and Moens, 2002; Teufel et al., 2009; Liakata et al., 2010, 2012]).

We derived a taxonomy of four facets by augmenting facets from of one coarse-grained schema for classifying scientific abstracts [Cohan et al., 2019] with the “NOV_ADV” category (i.e., corresponding to

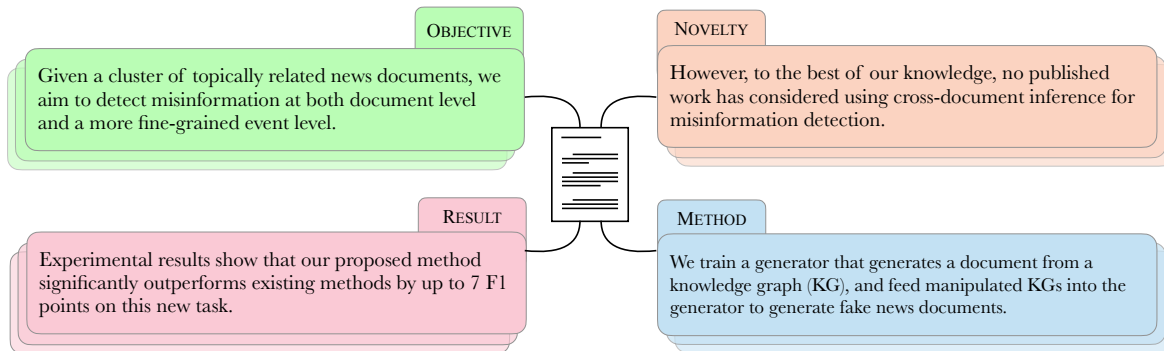


Figure 4.3: SCIM classifies and highlights four facets of information commonly found in papers: OBJECTIVE, NOVELTY, METHOD, and RESULT. These facets aim to surface specific kinds of paper content that align with common skimming goals identified in formative research, reflecting design guideline *D2*. Above, we show example passages matching each of the four facets. The passages appear in Wu et al.’s scientific paper, “Cross-document Misinformation Detection based on Event Graph Reasoning” [Wu et al., 2022b].

sentences describing the novelty of a paper) from Argumentative Zoning [Teufel et al., 2009]. As shown in Figure 4.3, SCIM’s four facets are: OBJECTIVE, NOVELTY, METHOD, and RESULT, each of which is represented in SCIM with its own color.

Low distraction. Text is highlighted using the familiar paradigm of a solid rectangular box behind the text, since this was observed in our evaluations of prototypes to be noticeable yet minimally distracting (*D4*). By using the same facet color mapping across papers, we hoped to foster a learned association device for each facet (*D4*). To help readers develop familiarity with highlight colors, the SCIM interface header includes a legend mapping colors to facets.

Distributed. Since users of initial prototypes were concerned when they saw passages without any highlights (*D5*), we post-processed model predictions to distribute highlights approximately evenly throughout the paper.

4.4.3 Controls

Different readers may have very different goals in skimming, and even a single reader’s goals may vary from one passage to the next. To provide flexibility in the skimming experience (*D7*), SCIM provides two kinds of controls:

Paper-level controls. If a reader wishes to perform a cursory high-level skim of a paper, they can reduce the density of highlights, or to inspect a paper more closely, they can increase the density. If a reader does

not wish to review a particular kind of content as they skim (e.g., they want to learn about the results of a study but not its methodology), the reader can disable highlights of a certain facet. Readers can control the density of highlights using facet-specific sliders found in SCIM’s side bar (Figure 4.1.C). As a reader drags a slider, they can see the effect on highlight density as highlights appear and disappear in the paper, markers appear and disappear in the scrollbar (Section 4.4.4), and a count of highlights change next to the slider.

Paragraph-level controls. If a reader desires additional highlights (e.g., if they have encountered a long paragraph of results they wish to skim more closely), SCIM provides paragraph-level controls allowing them to rapidly access additional highlights. Readers can request more or fewer highlights by hovering their mouse over a block of text, and then clicking on “+” and “-” buttons that appear in the margin (Figure 4.1.B). This feature provides quick and flexible control to complement paper-level controls, allowing a reader to request highlights precisely where they need them. For both paper- and paragraph-level controls, highlights are added and removed using a sentence prioritization score assigned during the document processing phase, as described in Section 4.5.

4.4.4 Scrollbar Annotations

A reader can discover where to skim in a paper by viewing highlight annotations in the scrollbar (Figure 4.1.D). This feature is inspired by edit wear and read wear affordances [Hill et al., 1992] and scrollbar annotations in code editors (e.g., [Microsoft, 2022]). When viewed together, these annotations can suggest paper structure, for instance implying if a paper has a particularly lengthy methods or results section, and where to find that information. The annotations also offer feedback to readers as they configure highlight density with SCIM’s controls.

4.4.5 Sidebar Display of Faceted Highlights

A reader may also review a paper’s key passages by opening a side bar, which shows a compact list of all highlighted passages in order, grouped by paper section (Figure 4.1.E). This display updates dynamically as a reader configures the highlight density. A vertical colored bar appears next to each passage, providing a subtle and compact indication of the passage’s facet. A reader desiring more context for a passage can click on it to scroll SCIM to the passage’s position in the paper, an interaction we call *context linking*.

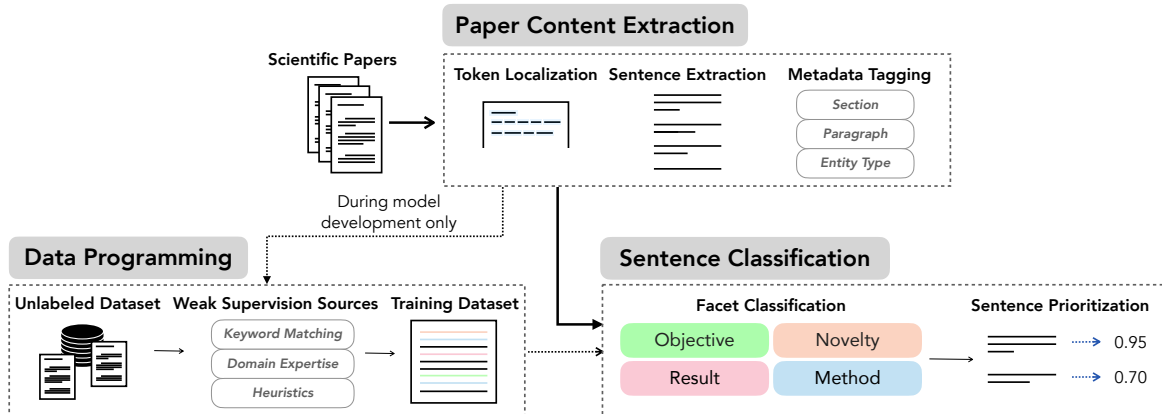


Figure 4.4: Overview of SCIM’s paper processing pipeline. SCIM takes as input a scientific paper in PDF format and then parses it into sentences with bounding boxes and other accompanying metadata. It then classifies sentences into one of four facets using a large language model fine-tuned via a data programming approach. SCIM chooses which highlights to show by reconciling prediction weights with heuristics controlling highlight distribution and readers’ preferences.

4.5 Technical Implementation

SCIM was developed with an end-to-end document processing pipeline that supports intelligent highlighting. The main component of this pipeline is a pretrained language model, fine-tuned via weak supervision to identify and classify salient sentences within papers. An overview of the pipeline is presented in Figure 4.4.

4.5.1 Paper Content Extraction

Given an input PDF document, SCIM uses VILA [Shen et al., 2022] and the open-source Multimodal Document Analysis (MMDA) library [for Artificial Intelligence, 2022] to extract textual tokens, mathematical symbols, section headers, and metadata. SCIM then segments the tokens into sentences, simultaneously merging bounding boxes for tokens into bounding boxes for sentences. Each sentence is labeled with its corresponding section header and paragraph index, attributes which are later used in the prioritization of sentences for which highlights should be shown.

4.5.2 Sentence Classification

To classify sentences into facets, we adapted the sequential sentence classification model from Cohan et al. [2019], replacing the base BERT model with a pretrained MiniLM model [Wang et al., 2020, 2021]. The

MiniLM model considers surrounding context—up to a combined sequence length of 512 words or 10 sentences—when classifying a target sentence. We fine-tuned the model with the CSABSTRACT dataset [Cohan et al., 2019], a corpus of abstracts from computer science papers with manually-curated “gold” labels. Since sentences only came from paper abstracts, we ultimately found the model insufficient for classifying sentences from the body of papers, so we pursued additional fine-tuning as we describe in the next section.

Data Programming

We initially attempted to create manually-curated datasets of “gold” facet labels for sentences from full papers. However, this task was difficult to define, time-consuming, and expensive to execute during our pilot runs of the data collection process. As a result, we decided to extend our dataset with weak supervision following a data programming approach [Ratner et al., 2016] to further fine-tune the model. Weak supervision provides a model-agnostic way to incorporate domain expertise into a model, and is sometimes a satisfying alternative to costly manual annotation. Weak supervision assumes access to a large unlabeled dataset and one or more labeling functions (e.g., heuristics encapsulating domain expertise, crowdsourcing, or knowledge bases), which are used to generate noisy labels for the dataset. While a collection of labeling functions can on their own serve as a classifier, we sought generalization beyond precise but potentially brittle labeling rules. We therefore employed a data programming paradigm to unify and de-noise the labeling functions, creating a weakly-labeled training set of sentences for downstream fine-tuning.

To build an unlabeled dataset for weak supervision, we extracted full paper sentences from the proceedings of NAACL 2018, 2019, and 2021, and ACL 2020–2022. In total, the dataset consisted of 3,051 papers with 606,400 unlabeled sentences. We then created weak supervision labeling functions consisting of heuristic rules and keyword matches to provide noisy facet labels for sentences in the dataset. For example, one rule-based supervision function detected sentence salience based on the presence of author intent via keywords such as “we”, “our”, “this paper,” and their aliases. Other labeling functions relied on keyword matches to perform facet labeling. For example, sentences were weakly labeled as NOVELTY if any relevant keywords (e.g., “novel”, “propose”, “differ,” and their aliases) could be found. We used Snorkel [Ratner et al., 2017] to unify these labeling functions and output a dataset of weakly labeled sentences.

The dataset was further improved by incorporating weakly labeled negative sentences, selected from

the full papers associated with the CSABSTRACT abstracts used during the first round of training. We selected novel sentences by using the `all-mpnet-base-v2` model [Song et al., 2020] from the Sentence Transformers library [Reimers and Gurevych, 2019] to score sentence similarity between full text and the abstract, and then labeling the most dissimilar sentences to the abstract, and which were not labeled with a facet in the prior phase, as not relevant for any facet (using an empirically chosen threshold cosine similarity of 0.25). Model fine-tuning was done on an NVIDIA A6000 GPU, using 0.1 dropout rate and Adam optimizer [Kingma and Ba, 2014] over 5 epochs, and $5e-5$ learning rate. All parameters were determined using the CSABSTRACT validation split.

Evaluation

We conducted a preliminary evaluation of SCIM’s sentence classification model over a set of 20 NLP papers. We recruited annotators from Upwork, an online freelancing marketplace. All hired annotators were required to have experience with NLP and scientific writing. Detailed instructions asked annotators to role play as a reviewer for a scientific communication magazine, tasked with creating abridged versions of scientific papers. Annotators were asked to identify significant, complete sentences within each of the 20 papers, and were paid \$20 USD/hr.

Each paper took on average 20 minutes to annotate, and was annotated by three Upworkers using the PAWLS PDF annotation tool [Neumann et al., 2021]. Sentences selected by at least two of the three annotators were considered as ground truth “significant sentences,” and collected into a test set. On this test set, our classification model achieved an F1 score of 0.533, compared to an annotator-annotator F1 score of 0.725 (which we consider as a gold-standard, i.e., a performance ceiling, since there is inherent variability in which sentences annotators believe are significant for skimming). While our goal with this preliminary evaluation was not to evaluate whether we necessarily advanced the state-of-the-art for this (new) task, our results helped to verify that the model could reliably identify meaningful highlights for use in SCIM.

4.5.3 Cleaning and Prioritizing Highlights

SCIM’s user interface selects which highlights to show using the predicted facet label, probability score, and other heuristics. One heuristic enforced consistency between facet labels and the section in which a sentence

appeared (e.g., if a highlight appeared within a methods section, it had to be tagged with the “METHOD” facet; similar constraints were imposed for the “NOVELTY” and “RESULTS” facets). Another heuristic prompted a more uniform distribution of highlights throughout a paper, prioritizing sentences within paragraphs which did not already contain other highlights.

4.5.4 User Interface Implementation

SCIM is implemented as a web application built atop the PDF rendering platform `pdf.js` [Mozilla, 2022]. The system retains text markup already present in the paper which may support skimming, such as hyperlinks, clickable citations, bold and italicized text, and other visual cues provided by the authors. SCIM’s features including highlights, side bars, and controls were implemented as interactive React components incorporating widgets from the Material UI library [Mui, 2025].

4.6 Study 1: In-Lab Usability Study

To assess SCIM’s usability and effectiveness for skimming, we conducted a three-phase sequence of studies. This section describes the first phase, an in-lab usability study aiming to evaluate how SCIM affects readers’ speed of skimming and their ability to identify relevant information throughout a paper, in comparison to a baseline of unaided paper navigation. In Section 4.7, we describe the second phase, a longitudinal diary study building on insights from the usability study, aimed at understanding the real-world value and applicability of SCIM over a two-week period. Finally, in Section 4.8.3, we detail the third phase, a deployment-scale evaluation with SCIM’s features integrated into the Semantic Reader, a publicly available augmented reading interface.

4.6.1 Study Design

For our usability study, participants were asked to complete a series of short tasks using both SCIM and a standard document reader. This study sought to answer two research questions:

RQ1. *Does SCIM enable readers to skim papers more quickly?*

RQ2. *How does SCIM affect readers’ ability to identify relevant information after a skim?*

Participants

We recruited 19 participants (8 male, 10 female, 1 non-binary) via university-affiliated mailing lists and Slack channels. We also conducted pilot studies with three additional participants, results of which we do not include in our analysis. Participants were required to have experience reading NLP papers, since they would be required to do so during the study. They ranged from 21 to 30 years of age, and included 11 PhD students, 5 master’s students, 2 software engineers, and 1 industry researcher. Participants self-reported an average of 3.78 (on a 5-point Likert scale) for comfort with reading NLP papers, suggesting they were generally familiar with the type of literature used in the study. Participants were compensated \$25 USD for their time.

Procedure

Participants first provided consent and then were led through a tutorial of SCIM’s features. The study used a within-subjects design, and consisted of three tasks, each with two sub-tasks, one for each of the two reading interface conditions—SCIM and a baseline consisting of a standard document reader, specifically the built-in PDF viewer in the Chrome web browser. We designed the study to be completed in under one hour to limit participant fatigue. The studies were conducted remotely via Zoom. To minimize biases, we counterbalanced the order of the reading interfaces and papers used in each task. Below, we describe the three tasks.

- *Task 1:* Participants skimmed a paper and identified a passage in the paper that described a key feature (e.g., dataset creation or evaluation) of the paper. This task was intended to familiarize participants with the two interfaces, so we did not include any measures from this task in our analysis.
- *Task 2:* Participants skimmed a paper and answered two multiple-choice questions based on information found in the paper. Answers to these questions *could* be found in text highlighted by SCIM. We hypothesized the main points highlighted by SCIM should be easier to locate, and this task was designed to test that hypothesis.
- *Task 3:* Participants skimmed a paper and answered two multiple-choice questions based on information found in the paper. Answers to these questions *could not* be found in text highlighted by SCIM.

In contrast to Task 2, we hypothesized that SCIM might prove a hindrance when finding information outside of the highlights, and this task was designed to check this concern.

Participants skimmed a different paper for each of the sub-tasks. The six papers for these tasks [Abrams and Scheutz, 2022; Li et al., 2022; Stasaski and Hearst, 2022; Sulem et al., 2022; Ibraheem et al., 2022; Xie et al., 2022] were selected from the proceedings of NAACL 2022, and corresponded to the following types: (1) technical papers introducing new datasets or metrics, (2) exploratory papers investigating the effectiveness of current tools and proposing new design guidelines, and (3) technical papers proposing novel language models for specific applications. Questions in Tasks 2 and 3 focused on aspects of a paper a reader might be interested in while skimming, such as evaluation metrics or the motivation behind a proposed method. Regardless of interface, participants were given multiple attempts and asked to skim until they answered correctly.

For each question, we used the following quantitative metrics:

- *Time*. The number of seconds taken by the participant to answer the question, from when the paper was first opened to when the correct answer choice was selected.
- *Accuracy*. A binary variable indicating whether the participant’s first response to the question was correct.
- *Difficulty*. A five-point Likert scale variable indicating the participant’s self-assessment of the following prompt: “I found the task difficult.”

After completing all sub-tasks, participants were also asked to self-assess on a five-point Likert scale whether they found the tasks overall easier to complete with SCIM and whether SCIM’s highlights were distracting during skimming.

Analysis

We compared readers’ time, accuracy, and perceived task difficulty using linear mixed-effects models [Lindstrom and Bates, 1990] with reading interface as a fixed effect, task and question number as nested fixed effects, and participant as a random effect. We first conducted F -tests for any differences across the inter-

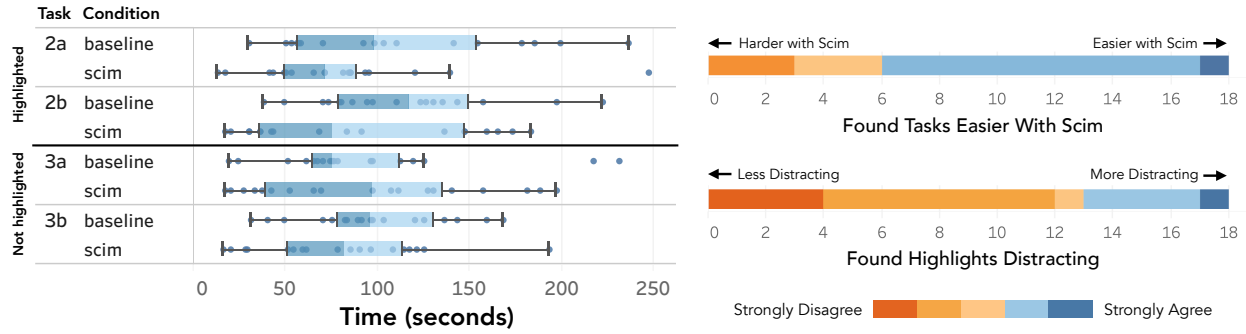


Figure 4.5: **(Left)** Time taken by participants to complete each information seeking question in Tasks 2 and 3 of the in-lab usability study. Overall, participants answered questions more quickly with SCIM than with a standard (baseline) document reader. **(Right)** Participants’ subjective responses regarding the ease of completing information seeking tasks with SCIM compared to a standard document reader, and whether they found SCIM’s highlights distracting.

face conditions, and then we conducted post-hoc *t*-tests when appropriate for differences in the estimated fixed-effects between conditions.

4.6.2 Results

Participants answered questions more quickly with SCIM ($\mu = 94.3s$, $\sigma = 74.9s$) than with a standard document reader ($\mu = 117.7s$, $\sigma = 76.4s$), a significant difference ($F(1, 126) = 4.17$, $p < .05$). The difference was more pronounced in Task 2, where the correct answer was located within one of the highlights ($F(1, 54) = 4.84$, $p < .05$): readers took an average of 93.8s with SCIM ($\sigma = 81.6s$) versus 127.3s with the standard reader ($\sigma = 77.8s$). In Task 3, where the correct answer was not located within one of the highlights, there was no significant difference in time ($F(1, 54) = 0.58$, $p = .45$), with participants taking 94.8s with SCIM ($\sigma = 68.7s$) versus 108.0s with the standard reader ($\sigma = 74.7s$).

There was no significant difference ($F(1, 126) = 0.22$, $p = .64$) in participants’ accuracy with SCIM ($\mu = 0.80$, $\sigma = 0.40$) compared to a standard document reader ($\mu = 0.76$, $\sigma = 0.43$). There was also no significant difference ($F(1, 119) = 0.01$, $p = .92$) in readers’ perceived difficulty in answering questions with SCIM ($\mu = 2.32$, $\sigma = 0.89$, with 5.00 indicating strong difficulty) compared to a standard document reader ($\mu = 2.31$, $\sigma = 1.02$). Altogether, the results show that SCIM reduced the time it took for readers to seek information in papers, with no difference in accuracy or effort.

4.7 Study 2: Longitudinal Diary Study

Participants in our usability study noted that it would take some time to acclimate to a novel reading interface like SCIM before they felt comfortable using it. To better understand realistic long-term use, we therefore also conducted a two-week long diary study. This study let readers use SCIM for papers of their choice from a list relevant to their discipline, leading to alignment of their motivation with typical motivations for skimming. Participants could choose when they read, and for how long, as long as they skimmed using SCIM at least once a day.

4.7.1 Study Design

We designed the diary study to provide insight into the following research questions:

RQ1. *What value can intelligent highlight-based skimming aids provide to researchers?*

RQ2. *How do researchers make use of skimming aids as they read?*

RQ3. *In what scenarios do researchers find skimming aids useful?*

RQ4. *What are the limitations of highlight-based skimming aids?*

RQ5. *What features should future intelligent skimming tools have?*

Participants

We recruited participants through university-affiliated mailing lists, Slack channels, and public posts from the authors' Twitter accounts. Participants were required to have prior experience reading or writing research papers. Preference was given to those with experience reading papers in the field of natural language processing (NLP), because the collection of papers we preprocessed for this study came from a recent NLP conference. A total of 12 participants were recruited for the study (6 male, 6 female). Two were master's students, and ten were PhD students. PhD students spanned a range of experience, with 1 first-year student, 3 second-year students, 2 third-year students, 3 fourth-year students, and 1 fifth-year student. No participants had participated in any of the prior lab studies. Participants were compensated \$100 USD at the end of the study.

Reading Materials

Though SCIM’s pipeline was able to process arbitrary papers within a few seconds, we wished to reduce the time it took for participants to load papers during the diary study. As a result, we preprocessed a set of papers we felt would be exciting for participants to read, specifically the proceedings of NAACL 2022, one of the most recent and widely-read NLP conferences. We selected these papers since SCIM had been fine-tuned primarily on NLP papers, and we expected it would perform appropriately for this collection. We also anticipated the NLP community would provide a broad audience from which we could recruit participants for a diary study. We also allowed participants to request other papers outside of this collection to read with SCIM throughout the study; in total, an additional 10 NLP papers were preprocessed.

Procedure

The diary study consisted of three stages: a welcome session, a two-week long observational period, and an exit interview. During the welcome session, participants completed a tutorial of how to use SCIM. They were given a few minutes to try out the interface, and to ask questions. Each participant was also shown the online diary (hosted in a Google Doc), and briefed on the protocol for recording their skimming experiences.

Then, during the observational period, participants were asked to spend 5–10 minutes each day, for 10 days, skimming at least one paper and completing a structured reflection in the diary.² On the first day of the observational period, participants skimmed papers using a standard (non-SCIM) document reader for the first day to provide a point of comparison with SCIM. During the subsequent nine days, they skimmed papers with SCIM. Following each skimming session, they completed a diary entry, consisting of the following questions:

1. Which papers did you skim today, and how long did you spend skimming each one?
2. What highlights (if any) drew your attention to something you might have otherwise missed?
3. Did highlights help you skim this paper? Explain.
4. List one or more ways the system could have helped you better skim this paper.

²Nearly all participants succeeded in completing 10 days of diary entries. Only 1 of 12 failed to complete all required entries; they completed only 7 of 10. When a participant fell behind in their diary entries, we sent them light email reminders.

Table 4.1: A summary of usage of SCIM’s features during the diary study. Notably, most readers used most features at least once. Use of the highlight controls varied widely, with some readers using them heavily (P6, P10, P12), and others less often (P1–3, P6, P7–9). All readers used the highlight browser on multiple occasions. Faceted highlights are omitted because we could not collect log data as to when readers looked at highlights.

Feature	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
Highlight Browser	5	10	8	10	9	4	12	3	5	20	6	19
Global Highlight Controls	1	0	3	0	2	16	3	3	0	22	9	4
Local Highlight Controls	3	2	1	8	0	12	0	0	0	16	3	34
Context Linking	0	1	0	3	30	2	0	0	0	6	0	8

After the observational period, we conducted exit interviews with participants. They were asked to reflect on their experience using SCIM in detail, including how it supported their skimming and opportunities for improvement.

Analysis

We conducted a thematic analysis on the qualitative data — diary entries and transcripts from exit interviews — following the approach described by Creswell and Poth [2016]. One author identified significant excerpts from the diary entries and transcripts, and iteratively developed and refined a set of themes represented in the data. A second author validated the analysis by reviewing the themes, checking their alignment with the excerpts, and proposing revisions. A total of 177 responses to diary prompts were analyzed (participants left responses to some questions blank). We also instrumented and analyzed behavioral logs detailing interactions with SCIM. In reporting results, we refer to participants with the pseudonyms P1–P12. The utterances presented below were edited to elide identifying information while preserving their meaning.

4.7.2 Results

Next, we present the findings of our diary study as they pertain to each of our research questions.

The value of SCIM as a skimming aid (RQ1)

For many readers, SCIM helped with skimming by allowing them to focus their attention and attain a high-level understanding of the paper (P5, P6, P9, P10). Furthermore, SCIM helped readers identify

key concepts and review the main ideas of papers. P5 described SCIM as guiding her to the important contributions of the papers she skimmed, and the highlights as offering a “gist of the paper beyond what was in the abstract.”

And though highlights helped readers review the paper as a whole, they could also help them orient to specific aspects of a paper they wanted to understand. For instance, P1 and P2 both noted that the highlights helped them to understand the results of the paper more quickly, which are often quite dense and text-heavy. SCIM’s highlights also helped readers attend to interesting details in sections of papers they might have otherwise skipped over (P1, P4, P11). This was described as “slowing down” and skimming with greater care:

This was a paper that is very light on methods and most content is about results, which I tend to skim over. So the highlights helped me slow down and slightly more carefully read a few places.

(P4)

For some readers, skimming without SCIM required two passes, first skimming a paper to identify relevant passages, and then re-reading passages of interest in greater detail (P5, P8). SCIM could alleviate the need for multiple passes:

With highlights, I usually spend more time reading and understanding the highlighted content and skimming the other content. Without the highlight[s], I need to scan the entire content first, identify the critical points and then understand it. The highlights save me time in skimming the whole paper. (P8)

How researchers made use of SCIM (RQ2)

Usage of SCIM entailed usage of its constituent features of highlights, the highlight browser, controls, and context linking. All readers made use of most features at least once (Table 4.1). We surmise that the most frequently used feature was the highlights, for several reasons. First, the feature was always turned on. Second, highlights figured prominently in our conversations with readers, as evidenced by the rest of this section. And third, most readers reported that the highlights helped them find useful information during their daily readings (see Section 4.7.2).

While the predominant method of interaction with SCIM was likely to view highlights within the paper, a second commonly used feature was the highlight browser: all readers opened SCIM's side bar more than once, with the average reader opening it 9.3 times. Readers described the highlight browser as supporting navigation and providing a rapid understanding of paper contents (P7, P9, P10). It was also described as an "extractive summary" (P2). One reader thought the highlight browser provided a "better way to skim" in comparison to highlights, which at the time of their diary entry, they believed made the paper "difficult to read" (P7).

Nearly all readers used both global and local controls to configure the number of highlights. Global controls were typically used a small handful of times to achieve an acceptable density of highlights (which was then persisted into subsequent skimming sessions). Only a few readers adjusted the highlights via the global controls across multiple papers skimmed. When asked, readers typically reported that the default density of highlights was appropriate (P2, P6, P7). That said, most did adjust the number of highlights with paper-level controls at least once. Exit interviews confirmed that readers tended to tune the level of highlights to the preferred level on the first day of use. One participant asked for highlight controls with coarser options, for instance enabling them to toggle between one mode showing only the most important highlights, and another with many highlights for a deeper skim.

Readers seemed to use SCIM to augment, rather than replace, their existing skimming strategies. Readers reported directing their attention both to the highlights and to conventional paper landmarks like section headers and visual content. For example, P9 described their process as navigating through the main sections of a paper as they might in a typical skim, and then using the highlights to identify important information within those sections. P4 similarly described skimming using the combination of section headers and highlights.

For some readers, it took some time to become accustomed to using SCIM (P8, P10). One issue seemed to be developing trust in what was highlighted (P8). In their exit interview, P10 described their how their trust and interactions with SCIM evolved over the course of the study:

I feel like I just got more used to the highlights. ... When I would see an objective highlight, I would trust it. I found the results highlights to be very helpful, so I would immediately focus on those. I would open the side panel right away instead of waiting during the end of the paper.

I just got used to the tool, and I learned how to use it fast, depending on the paper and what I wanted to get from the paper. (P10)

Circumstances in which SCIM was useful (RQ3)

Overall, the intelligent highlights appeared to be useful during a majority of skimming sessions. In response to the diary question, “Did highlights help you skim this paper?” 74 of 105 (70.4%) responses answered in the affirmative. There were a handful of circumstances in which readers reported SCIM as particularly useful.

One circumstance where SCIM was useful was in reading dense passages of text. The highlights made long passages that were absent of “visual support” such as figures more approachable (P3, P5). SCIM helped one reader skim a detailed experimental section and identify several important details which, due to the density of text, they “might have skipped if not for the highlights” (P5). Readers reported SCIM as helpful not just for dense passages, but also for papers that were text-heavy as a whole, such as survey papers (P5, P11).

Intelligent highlights were seen as useful to readers who sought information from papers on a topic they did not typically read about (P8, P10), assisting them in identifying and focusing on important paper content:

For me it was also generally useful for reading papers that were a little out of my comfort zone. ... In that case the highlighting helped me focus on, read, and conceptualize better certain parts of the methodology in order to better understand the conclusions. (P10)

The highlights also provided a summary of the paper in their own right. One reader described a situation where they were “not particularly interested in this paper.” For them, the highlights served as “a summary” that they could read in lieu of looking closely at the paper (P4). This suggests an interesting possibility for intelligent highlights to help not just highly-motivated skimmers, but also those skimming papers in lower-motivation contexts.

Limitations of SCIM’s model of intelligent highlighting (RQ4)

Readers identified several ways that intelligent highlights might be extended to be made more useful.

One concern was that highlighted passages sometimes lacked sufficient context to be understood alone (P1–3, P7–8, P11).

When reading the highlights, the context is often missing. Sometimes it is just in the lines before and after, but sometimes we need to find it which then makes reading difficult as there is now more back and forth instead of a linear reading. (P7)

SCIM was designed with the hope that readers would look for such “context” in the surrounding text by simply moving their focus from highlighted to unhighlighted text. In practice, it could be disruptive for readers to seek out this context. Necessary context could appear just before or after the highlight in the paragraph, and in some cases even in other sections. For one reader, skimming highlights that lacked context therefore became a process that resembled “more back and forth instead of a linear reading” (P7).

Some readers desired tighter integration between SCIM’s highlights and existing visual cues within a paper. While SCIM did not occlude or hide text that the author had emphasized (e.g., bolded text, section headers, or bulleted lists), this emphasized text was often not highlighted. As a result, readers discovered inconsistencies between the visual cues introduced by authors and the highlights suggested by SCIM (P2, P4, P5), such as bolded result statements or contributions in a list, which were not consistently highlighted.

Sometimes, text was highlighted in other unexpected and undesired ways. For instance, SCIM sometimes highlighted only one contribution from a list of bulleted contributions, when readers believed it should have highlighted all of them (P1, P7). SCIM was also unpredictable when highlighting passages that contained dense math notation (P1, P6, P11), and readers wished for highlights to apply to visual content like tables and figures (P2, P5, P7, P12).

Envisioning future intelligent skimming tools (RQ5)

SCIM represents just one way in which intelligent assistance could support the paper skimming process, and readers described alternative ways that future tools could help them skim. For some readers, SCIM’s highlights provided too much detail, particularly if they desired only a high-level understanding of the material (P6, P8). Readers suggested that an abstractive summarization of paper content (e.g., “with a bit of info pulled from tables/graphs/figures/examples” (P12)), could lessen the effort required to understand dense sections of papers (P1–2, P7–P8, P12). Recent large language models have achieved impressive advances in

summarizing scientific texts, and future tools could leverage these models to augment the reading experience with abstractive summaries.

Readers also believed they could be aided with better tools for navigation. One reader desired the ability to use a paper’s abstract or introduction as an index into related highlights in the rest of the paper (P2). Another reader wished to see the paper summarized in question-and-answer format, realizing they often sought answers to questions while they skimmed, such as “What are the research questions? What are the novelties/contributions of this study? What data/model/evaluation methods do they use? What are the main results? What are the limitations?” (P8). While SCIM addresses these information needs through faceted highlights, future tools could support more conversational interactions between readers and the papers they skim.

4.8 Study 3: Deployment of SCIM at Production Scale

To evaluate the impact of automatically-generated highlights at a larger scale, we adapted features from SCIM into Semantic Reader, a publicly-available augmented reading interface for scientific papers.³ Through their usage of the online reading interface, we envisioned scholars could interact with the features naturally throughout their skimming process. As part of the integration, we iteratively refined SCIM’s features by incorporating feedback from participants in our prior two studies, and explored opportunities to make the design more accessible for individuals with visual disabilities. Finally, in scaling SCIM’s automatically generated highlights to several hundred thousand papers, we added several architectural improvements to improve the robustness and efficiency of the paper processing pipeline.

4.8.1 Refining Customizability and Improving Accessibility

Findings from the two prior studies showed that colored highlights are an effective visual cue for communicating distinct facets of information in the text. However, one limitation of such a design is that it remains inaccessible to individuals who use assistive technology such as screen readers, those with difficulties seeing low-contrast visuals, or users with certain types of color blindness. Seen through the lens of inclusive

³As of the time of writing, SCIM is available on 521,000 papers, accessible at <https://www.semanticscholar.org/product/semantic-reader>.

design, we believed this one-size-fits-all approach of interface augmentations created an inequitable skimming experience. To accommodate such users who may engage with the text using a screen reader, we ensured snippets in the side bar are fully screen reader accessible. Thus, these users may read with the same attention guidance augmentations that users without assistive technology would encounter. While we find that this solution is a step toward a more accessible scientific paper reading experience, we acknowledge that reading side bar snippets is not the same experience as viewing highlighted text as visual cues within the paper itself. Future work could enable screen readers to verbally announce the beginnings and ends of highlighting in-line, which could help preserve the context around each highlight.

Furthermore, we iterated on SCIM's design to include three options that allow users to customize the presentation of skimming highlights based on their individual needs.

Margin flags We added optional labels adjacent to each highlight in the paper's margin which could indicate a highlight's facet at a glance without relying on color. Users who may have difficulty remembering the color associations or who cannot distinguish between the colors could benefit from these margin flags, which convey facet information via text rather than color.

Customizable highlight contrast and opacity We provided controls that allow users to customize the opacity and contrast of the highlights. Some users may find the bright highlights distracting, while others may be unable to distinguish low contrast highlights against a paper's white background. These contrast controls are designed to accommodate both needs.

Coarse-grained highlight density controls We simplified SCIM's original paper- and paragraph-level controls to reduce the effort required for users to determine the appropriate amount of highlights in a paper. Our diary study suggested that the predetermined number of highlights in SCIM was a reasonable amount to support skimming, and that few users cared to adjust the highlights at a finer-grained level while reading. As a result, we provided two global adjustments for highlight density, allowing users to either select less or more highlights than the default.

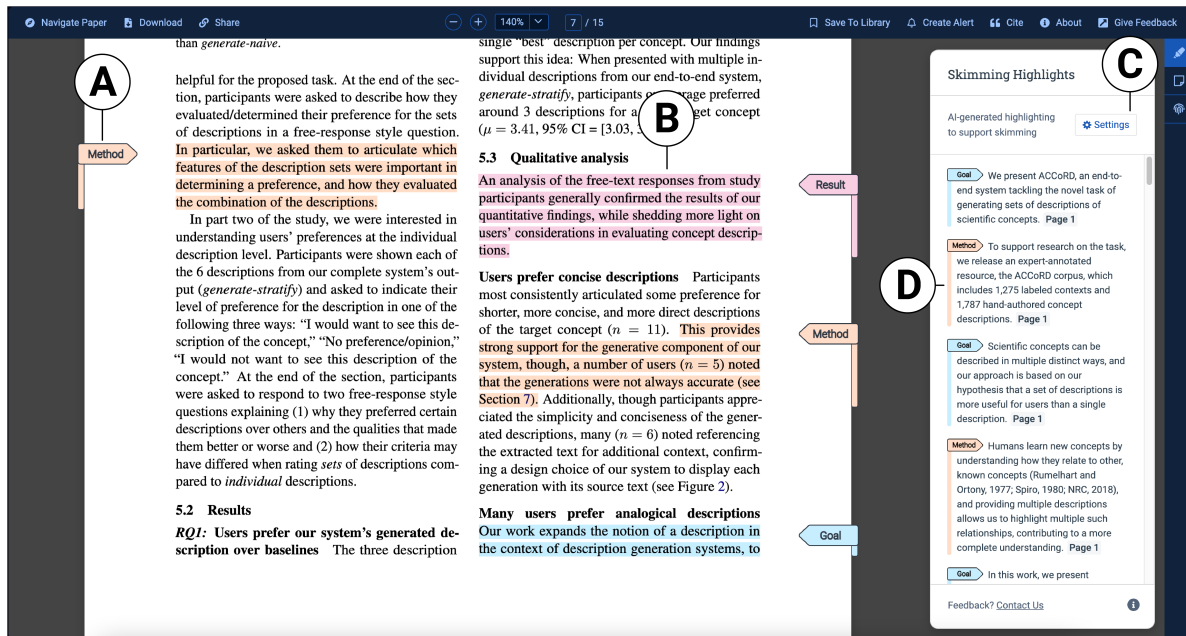


Figure 4.6: We integrated SCIM’s interface augmentations into a publicly-available reading interface for scientific papers, Semantic Reader, refined into the following set of features: margin flags which can be turned on or off (A); multi-faceted highlights (Goal, Method, and Results), the density and visual characteristics of which can be customized (B); user-controlled settings for further customizability of the interface (C); screen-reader accessible snippets containing the extractive text from each highlight (D).

4.8.2 Improving the Robustness of the Paper Processing Pipeline

In addition to highlights which were irrelevant, lacked sufficient context, or were classified with an incorrect facet, participants in the diary study also sometimes noted inaccuracies in the highlights that appeared to stem from errors in the document processing component. We found participants were particularly sensitive to these errors, which were easily distinguishable at a glance, and the prevalence of these errors became more noticeable as we scaled the number of papers for which highlights were available. For instance, some highlights were incorrectly split across multiple columns or pages in a paper, or on undesirable paper content such as footnotes, page numbers, and references. The presence of these errors may distract scholars as they skim, and even worse, potentially undermine their confidence in quality of the highlights overall.

Analysis of errors in highlights from document processing

To assess the frequency and types of potential errors that may appear in the displayed highlights, we conducted a two-phase evaluation. First, three members of the research team annotated and coded errors within a random, traffic-biased set of 30 papers (uniformly sampled from papers available on the Semantic Reader with more than two views over last 90 days). Annotators were asked to identify and categorize errors as they encountered them.

Annotators found a diversity of document processing issues that diminished the quality of the extracted highlights. These errors included:

1. **Non-body text errors** — The model responsible for extracting the main body text from paper PDFs occasionally included other text on the page as part of the main body, such as the paper’s preamble (e.g., title, author, affiliations), page numbers, headers, footnotes, and references. It also sometimes identified text from within figures and tables; we did not intend to highlight these visual or tabular components within papers in SCIM (though future aids may consider the utility of highlighting such content). Including non-main body text within an extracted sentence could degrade the accuracy of the subsequent highlight classification model, and also draw attention to irrelevant content within a page, visually distracting readers.
2. **Sentence splitting errors** — Sentences were sometimes incorrectly split on non-sentence boundary

punctuation (e.g., splitting on the period in “In Section 4.1, we first introduce...”) or on column and page boundaries. These resulted in highlights that seemed to begin or end abruptly, rather than capturing a complete sentence.

3. **Layout errors** — If the token bounding boxes were not accurately extracted, the highlights could appear shifted on the page. This resulted in some highlights that appeared over empty regions of a page, across sentence boundaries, or overlap figures.

In our error analysis, we noticed that papers with an arXiv source had fewer errors and generally higher quality data, since these papers were often had cleaner PDFs compiled from LaTeX source documents. We also found that the paper processing pipeline tended to perform better on computer science papers, possibly due to biases in the training data. As a result, we conducted a second annotation on 97 computer science papers with an arXiv source. For this set, we observed 60.8% of papers had at least one non-body text error, 54.4% of papers had at least one sentence splitting error, and 5.2% of papers had at least one layout error. No errors were detected in 35.1% of papers.

Overall, this error analysis revealed the existence of a substantial quantity and diversity in document processing errors that could propagate into the reading interface, introducing undesirable distractions for users while skimming. It further suggests how building effective and usable intelligent reading interfaces requires not only furthering the capabilities of models for language understanding, but also addressing the seemingly small and nuanced errors that may emerge when processing long and complex PDF papers, impacting the augmented reading experience.

Additional structural and linguistic predictors

To address these issues, we enhanced the original paper processing pipeline (Section 4.5) with additional heuristic and learned predictors.

1. **Margin predictor** — This uses a clustering algorithm to detect margins of a page containing the body text of a paper. Identifying these regions helps to remove extraneous symbols within the margins such as page numbers or watermark text (e.g., identifiers on arXiv preprints).
2. **Word predictor** — This fixes words that get split between rows (e.g., with hyphenation).

3. **Block predictor** — This introduces heuristics to improve VILA block predictions (originally described in Section 4.5.1) for non-body text components of papers. Accurately identifying these blocks helps in ensuring certain types of content are disregarded by the highlight selection model, such as text within references, figures, tables, equations, section headers, preamble (e.g., title, authors), headers, footers, and footnotes.
4. **Sentence predictor** — This merges sentences that span across columns and pages, and filters out overly long sentences (>160 tokens).
5. **Sentence perplexity predictor** — This estimates the perplexity of sentences using a unigram language model. Specifically, sentences with an average log word probability of less than -20 are removed.⁴ This is to ensure sentences that are mostly numbers (e.g., sentences parsed from tables) get filtered out.
6. **Sentence header predictor** — This matches each sentence with its appropriate (possibly nested) section headers. It improves upon the original header predictor in SCIM (which was originally used in a heuristic ensuring consistency between facet labels and sections; see Section 4.5.3), and is used to render the section header for each snippet in the side bar.

In an updated pipeline, these predictors are combined with the original faceted highlight detection and classification language models (described in Section 4.5.2). The specific language models used for classification could eventually be substituted with more performant models, e.g., generative large language models, which even with minimal prompt engineering and simple few-shot prompting strategies have exhibited impressive capabilities on many language understanding tasks [Brown et al., 2020b; Bubeck et al., 2023].

4.8.3 Deployment Study

Between August 20, 2023 and October 6, 2023, we ran a randomized controlled trial on 350,000 recent computer science, arXiv papers on Semantic Scholar’s Semantic Reader platform [Lo et al., 2023]. Fifty percent of users experienced a baseline version of the Semantic Reader and the other 50% were given the

⁴Log word probability is calculated using word frequencies from the Web 1T Ngram corpus [Brants and Franz, 2006]. See [Sol-daini and Lo, 2023] for similar details.

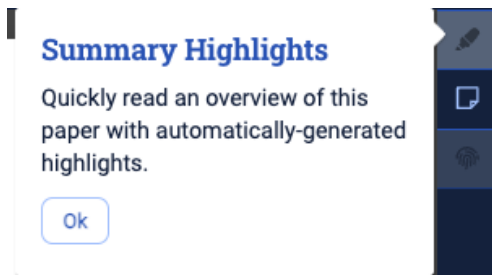


Figure 4.7: When users in the deployment study first encountered a paper with skimming features available, the capabilities were introduced with a popup window in the upper right, directing the user to controls for opening the sidebar with highlights and customizable settings.

reader with production skimming features as shown in Figure 4.6. On their first visit to a skimming-enabled paper, the feature was explained with a one-time, dismissable popup window (Figure 4.7), directing the user to controls to open the highlight panel and adjust controls.

Over the trial period, 20,200 unique users read 42,700 documents with skimming available. 19% of these users actively used the skimming capabilities, and an average of 12.7 times per user — with a mixture of clicking on paper highlights or flags, hovering or clicking on side bar snippets, and adjusting settings. 197 users (5% of users engaged with skimming) adjusted settings, such as the number, density or type of highlights displayed. 49 users (1.3%) turned the feature off. No statistically significant difference was seen between the control and treatment groups with respect to key metrics: papers viewed and scroll distance per user, nor in Semantic Scholar site-wide metrics such as the number of returning sessions per user or the number of discovery clicks per user. As a result, the test was concluded and skimming was enabled on 521,000 papers (as of October 22, 2023) for all users of Semantic Reader.

Enabling SCIM at production-scale across a large corpus of papers served to highlight a number of challenges in designing effective intelligent user interfaces, particularly in the context of scholarly support tools. Qualitative feedback collected throughout the deployment study suggests that a large opportunity remains in improving the models used in identification of relevant content to highlight. The current models we developed remain imperfect and also appeared to perform worse on out-of-distribution papers, such as those outside of computer science. Beyond improving the underlying automated techniques, a key tension in the design of SCIM was in augmenting papers in a way that accelerated rather than distracted from readers' own skimming processes. As a result, significant effort was dedicated to iteratively tailoring the distribution, density, and visual characteristics of the highlights to facilitate effective skimming support. The nontriv-

ial number of data processing improvements and guardrails that were incorporated into the pipeline before deployment also indicates the importance of mechanisms for identifying and resolving errors throughout complex, multi-component systems, as errors that propagate from any component, even beyond the natural language processing models themselves, can negatively influence users' perceived quality of these intelligent tools. Our design process also revealed the importance of customizability within these interfaces, for instance in allowing users to tailor the information presentation according to their specific needs, research backgrounds, and familiarity with a particular paper. Future systems can also continue to ensure equitable access by involving inclusive design principles as a core consideration from the onset, ensuring that these tools are universally usable.

4.9 Discussion

In this chapter, we explored how intelligent user interfaces could support readers in skimming scientific papers, examining the utility of one form of cognitive scaffolding—namely through visual augmentations—that strategically direct attention and provide a structured mental model aligned with the scholarly reasoning processes. Specifically, we designed and evaluated SCIM, a tool that augments a paper reading interface with automatically-generated faceted highlights of salient paper content. A lab study showed SCIM reduced the amount of time to complete short information seeking tasks in scientific papers, with no significant difference in readers' self-reported task difficulty. In a subsequent diary study, we observed how researchers might use SCIM in everyday reading environments. Participants believed SCIM helped them develop a high-level understanding of papers and determine which passages to skim or skip. SCIM was seen as particularly useful for skimming long or dense texts and papers from unfamiliar domains. Next, we consider these results, amidst limitations of our approach, and suggest implications for the agenda of developing intelligent skimming aids.

4.9.1 Skimming versus Scanning

In the lab study, the information seeking tasks we used were intended to measure participants' speed and accuracy while skimming. However, we noticed that participants often exhibited behavior that more closely resembled *scanning*. Unlike skimming which involves a rapid high-level comprehension of text, scanning

is a subtly different reading process concerned with locating specific pieces of information within a body of text. In the lab study, some participants utilized conventional scanning strategies such as “Control+F”, using keywords in the question or answer choices as prompts. While this strategy was typically unsuccessful since the questions were designed with text which nullified this keyword-based scanning strategy, this behavior suggests participants did not necessarily attempt to skim the paper to gain an understanding of the paper to answer the questions, but instead scanned the text for keywords to locate the exact answer to the question. For readers like these, our results may be less indicative that SCIM helps with the skimming process but rather the scanning process.

4.9.2 Supporting Experts and Novices

SCIM was designed to help experienced skimmers get more out of skimming. It was not, however, designed to help inexperienced skimmers develop proficiency with skimming. Skimming assistants for teaching skimming may or may not have a lot in common with SCIM. As AI models become increasingly adept at identifying salient paper content, those models may be useful not just to identify important content, but to coach skimmers to find this content as well. The development of such AIs and accompanying interfaces poses the interesting design problem of ensuring that readers have a consistently productive experience learning the essentials of skimming, regardless of their prior background or the documents they skim.

4.9.3 Risks to Attention

One risk of introducing technologies like SCIM is that they may have unintended consequences for a reader’s attention. Skimming requires significant attention to understand the idiosyncrasies and nuances within papers. Tools that augment the paper with assistive affordances may inadvertently deplete a reader’s limited attention if they impose additional cognitive burden, as might be the case if it highlights content in a way readers do not expect. The tools could also lead a reader to pay less attention to the paper as they skim, once they are no longer required to drive the skimming process themselves. Furthermore, should such skimming assistants be readily available, readers may not be incentivized to deeply read papers, but rather enticed to skim by the presence of highlights.

Skimming aids like SCIM therefore need to be designed in a way that respects a reader’s attention and the

value of deeply reading. They should be accurate and reliable. They should be deployed alongside studies that understand their effect on readers' engagement with texts. Furthermore, they should be developed and deployed in tandem with tools that support and encourage researchers to deeply read, and in general encourage good reading practices within the research community. For instance, affordances for skimming might be made available only when searching quickly through multiple papers or while reading on the go, but then are limited in scenarios befitting a deeper read.

4.9.4 Limitations of Highlights

Without sophisticated controls and affordances enabling more goal-driven or personalized skimming, highlights only present a single pathway through a paper. Highlighting is a cueing mechanism that directs reader attention and assists in the foraging of information, but it does not address other aspects of the information gathering and sensemaking processes scholars undergo while skimming. As readers suggested in the diary study, there are numerous ways in which skimming aids could provide more holistic support. For instance, augmented intelligence could offer additional context for highlighted passages, integration with existing visual cues, highlighting of visual content, complementary usage of abstractive summarization, or enhanced navigation support.

4.9.5 Future Work

Improving Highlight Quality

The effectiveness of SCIM, like that of many other AI-infused user interfaces, is limited by the accuracy of its underlying AI models. Future research could improve the utility of highlights by building upon our models with other computational approaches. One promising direction could involve long-form summarization models sensitive to our highlight-relevant design guidelines, and another could leverage the discriminative capabilities of general-purpose large language models. Features including a paper's hierarchical structure, author-cued content, or visual content may also be leveraged to improve highlighting accuracy. Furthermore, our findings suggest improvements to PDF processing are necessary to improve the user experience for augmented reading interfaces like SCIM. Minor errors in SCIM's PDF processing resulted in content like footnotes, section headers, tables, or figures being concatenated with paper sentences, which led both to

poor classification of those passages and highlights that included disparate content. Resolving such issues in PDF processing could significantly improve classification accuracy and the cleanliness of the presentation of highlights within a reading interface.

Social Annotations

Our studies suggested that readers may be hesitant to adopt an augmented reading interface like SCIM due to distrust in the AI's ability to provide the relevant highlights. Some readers mentioned that they might trust highlights created by other people (e.g., fellow researchers) more than those generated by AIs. Could social annotations be used to produce better highlights? Social highlights have been extensively explored in other settings, including studies on the effect of social annotation on attention within public multimedia content [Carter et al., 2004], news reading [Kulkarni and Chi, 2013], and education [Glassman et al., 2015; Yoon et al., 2016; Zyto et al., 2012]. Modern online publishing platforms such as Medium also show “popular highlights,” suggesting the potential for social highlights in reading tools for scientific literature as well. Such affordances could port nicely into a system like SCIM, and the coordination of social highlights with AI-generated highlights can be an interesting area for future research. More broadly, the development of community-driven features, such as enabling readers to share and discover highlighted content curated by their peers within the same field, suggests how collective intelligence can be leveraged to personalize and enrich the skimming experience.

Personalization of Skimming Aids

As readers continue to interact with augmented reading interfaces, we envision an opportunity for AI-infused systems to learn from repeated reader interactions, providing personalized and proactive reading support to help mitigate undesirable cognitive overhead introduced by these systems. They could also be tailored to readers' individual reading behaviors by considering their experience reading papers within a particular field, their typical information needs, or their goals for reading a particular paper. Providing this interactivity will empower users to fine-tune the systems to their own preferences, enhancing the relevance of the content surfaced over time. Furthermore, while the current design of SCIM highlights content according to four common facets, future skimming aids could allow users to dynamically define and refine these facets beyond

this initial set. And unlike the current paper-agnostic facets, additional benefit may lie in augmenting the reading interface with more informative or fine-grained facets, dependent on the specific content in a paper users are actively reading.

4.10 Conclusion

In this chapter, we presented the design, development, and evaluation of SCIM, an intelligent augmented reading interface that supports scholars in skimming scientific papers with faceted, evenly-distributed, minimally intrusive, configurable highlights. A lab usability study found participants located information in papers more quickly with SCIM than with a standard document reader. In a two-week-long diary study, participants remarked ways in which SCIM supported a rapid, high-level skimming of papers. Based on feedback from these two studies, we iteratively refined SCIM's design and architecture, before conducting a production-scale deployment study by integrating SCIM's features into the Semantic Reader, a public augmented paper reading interface. SCIM was found to be particularly useful for dense passages of text and for papers from unfamiliar domains. Altogether, these studies suggest the potential for intelligent tools to support researchers in skimming scientific literature, and highlight opportunities for further model improvement and personalization.

Chapter 5

Challenges and Opportunities for Human-AI Scholarly Synthesis

In the previous chapters, we explored how particular AI-assisted representations can help mediate researchers and dense paper content while reviewing individual papers. QLARIFY, for instance, illustrated how fluid text expansions enable readers to progressively enrich static abstracts with clarifying information. This notion of expansion as a mechanism for sensemaking extends beyond single documents: for most scholarly work, researchers ultimately must synthesize insights across entire literature collections—a process that becomes increasingly challenging as the volume of literature to review continues to grow. Literature reviews and surveys are a primary mechanism for this synthesis, organizing vast literature into structured frameworks and narratives. However, these knowledge artifacts quickly become outdated, creating tension between their value and the effort required to maintain them. Just as QLARIFY reframed abstracts as expandable entry points that adapt to a reader’s evolving information needs, living reviews can be seen as expandable knowledge artifacts that adapt to the evolving state of a field. Both cases highlight expansion as a mechanism for maintaining fluidity between stable representations such as abstracts or surveys and the richer bodies of evidence they summarize. To understand how AI might collaborate with researchers to support the creation and maintenance of these syntheses as the underlying knowledge evolves over time, we examined how surveys in computing research are currently created and the key challenges and leverage points for computational assistance. Our findings from retrospective interviews with 11 authors reveal three

distinct types of updates required for living reviews—empirical, structural, and interpretative—each presenting unique opportunities for AI-human collaboration in scholarly synthesis. This study was previously published at ACM CHI 2025 [Fok et al., 2025b] and has been adapted for this document.

5.1 Introduction

Reviewing and building upon existing knowledge is fundamental to scholarly research. However, with the acceleration of research production and publication, keeping at the forefront of all available literature has become increasingly complex. Literature reviews offer tremendous value in synthesizing research, but are time-consuming and resource-intensive to create [Tricco et al., 2008; Michelson and Reuter, 2019a]. Moreover, they quickly become outdated as new research emerges [Shekelle et al., 2001]. One survival analysis suggests some reviews may be outdated by the time of their publication and over 25% require updating just two years after publication [Shojania et al., 2007]. Such information staleness presents a threat to validity, potentially under-informing researchers about research opportunities or critically misleading decision-makers.

In response, researchers have explored the concept of “living reviews”—documents that are continually updated as new evidence emerges [Elliott et al., 2014; Wijkstra et al., 2021; Higgins et al., 2019], typically centered around living *systematic* reviews. Maintaining living reviews can be challenging, however, with studies finding many are never updated after initial publication [Tricco et al., 2008; Heron et al., 2023]. Prior work on living reviews has primarily focused on accelerating literature discovery and appraisal [Thomas et al., 2017b; Vergara-Merino et al., 2020], but recent advances in artificial intelligence (AI) and large language models suggest an opportunity for supporting more complex cognitive aspects of updating reviews, such as evidence synthesis [Martin-Boyle et al., 2024; Wang et al., 2024b]. Such capabilities are particularly relevant for living *narrative* reviews, such as surveys in computing research, which rely more on expert interpretation and conceptual synthesis than the standardized protocols and meta-analyses of systematic reviews [Snyder, 2019].

To examine how narrative reviews are created and maintained in practice, we conducted semi-structured, retrospective interviews with 11 survey authors across diverse areas of computing research. Authors reflected on their processes and points of friction throughout their authoring and revision workflows, and

shared their perspectives on when, how, and with what content these reviews should be updated. In understanding these practices, we sought to identify opportunities to support the updating process, including the potential role of AI assistance.

Our findings reveal the varied methodologies used and challenges encountered when authoring and updating survey articles, especially in paper discovery, taxonomy development, and synthesis. We identify three key types of updates for maintaining narrative reviews: 1) empirical updates involving evidence and examples, 2) structural updates to taxonomies and paper organization, and 3) interpretive updates to syntheses and framing. Each type offers distinct opportunities for AI assistance, from routine tasks like recalculating numerical values to more complex support for identifying potential biases and emerging research gaps. While authors saw the potential of AI to assist with routine tasks, they were skeptical of its ability to handle more nuanced interpretive tasks like constructing a compelling narrative. Moreover, the subjective, expertise-driven nature of their workflows raises important considerations for future tools enabling living narrative reviews.

5.2 Background

There exist different types of literature reviews, varying across their goals, processes, and degrees of systematicity, with one typology identifying 14 distinct review types (e.g., critical, scoping, systematic, qualitative synthesis, and umbrella reviews) [Grant and Booth, 2009]. More broadly, reviews can be categorized into one of three major approaches: systematic, semi-systematic, or integrative [Snyder, 2019]. Systematic literature reviews, developed for and commonly utilized in clinical medicine, use “explicit, systematic methods to identify, appraise, and synthesize all the empirical evidence that meets pre-specified eligibility criteria to answer a specific research question.”¹ As the gold standard of evidence synthesis for clinical decision-making, systematic review methods have been studied extensively (e.g., [Chandler and Hopewell, 2013b; Pati and Lorusso, 2018; Lasserson et al., 2019]). On the other hand, semi-systematic or *narrative* reviews are preferable when reviewing every relevant article is not possible. These reviews may rely on quasi-systematic search and appraisal methods, combine both qualitative and quantitative evidence, and hold broader aims beyond a specific research question, such as to survey a research area or track its progression

¹<https://www.cochranelibrary.com/about/about-cochrane-reviews>

over time. Closely related to narrative reviews, integrative or critical reviews further critique the literature to develop new theoretical perspectives, often integrating heterogeneous sources beyond peer-reviewed research articles to provide a synthesis that goes beyond descriptive or historical.

Creating these different reviews to synthesize the literature is often time-consuming and costly [Tricco et al., 2008; Michelson and Reuter, 2019a]. Researchers have therefore explored the potential for AI to support literature review creation, among other research activities [Van Noorden and Perkel, 2023; Morris, 2023b; Messeri and Crockett, 2024]. However, the accuracy and utility of reviews can decay over time [Shekelle et al., 2001; Elliott et al., 2017], leading to interest in the potential of “living reviews” that are continually updated, incorporating relevant evidence as it becomes available [Elliott et al., 2014]. Organizations such as Cochrane² offer guidance toward living systematic reviews (LSRs) [Elliott et al., 2017], contributing a series of works that provide context and justification for LSRs [Elliott et al., 2017], reason about the role of technology in supporting the creation of LSRs [Thomas et al., 2017b], develop methods for updating the statistical meta-analyses found in LSRs [Simmonds et al., 2017], and outline the potential for living guidelines that offer dynamic recommendations based on LSRs [Akl et al., 2017].

Despite this guidance on the conduct, reporting, and publication of LSRs [Higgins et al., 2019; Elliott et al., 2017], studies suggest practical challenges in sustaining these workflows have limited the success of such living reviews [Tricco et al., 2008; Heron et al., 2023]. One recent analysis of LSRs on COVID-19 found that most LSRs were never updated after their initial publication, underscoring the overall difficulty of keeping such reviews alive if reliant on largely manual updating efforts [Heron et al., 2023]. Similarly, in climate science, US federal law mandates a process of updating a summary of relevant scientific knowledge every four years. This process is laborious and costly, with the most recent update involving over 500 authors.³ While AI tools are being considered as aids for this updating process, they have yet to be adopted [Khourdajie, 2024]. Overall, existing guidance on LSRs appears to insufficiently recognize the potential for technological support in what remains a predominantly manual review updating process, and the aging considerations ought to be reexamined in light of recent AI advancements.

Moreover, while existing work has focused on living *systematic* reviews, supporting the process of updating semi-systematic or narrative reviews has received little attention. Systematic reviews which adhere

²<https://www.cochrane.org/>

³<https://nca2023.globalchange.gov/>

Table 5.1: Participant biographies and self-reported research expertise.

ID	Biographical Information	Research Expertise
P1	40, male, assistant professor	Advanced cyber systems, digital equity, distributed systems
P2	26, male, PhD student	Critical HCI, queer HCI
P3	29, male, PhD student	Hardware security
P4	33, female, PhD student	Safety technologies, transformative justice
P5	27, male, PhD student	Persuasive interventions for absent-minded smartphone use
P6	28, male, PhD student	Low-power wide-area network, LoRa networking
P7	27, male, PhD student	Signal processing, video anomaly detection
P8	35, female, assistant professor	Human-computer interaction, realism, audio, games
P9	33, female, PhD student	Algorithmic systems in housing services
P10	36, male, assistant professor	Human aspects of software engineering
P11	33, male, research engineer	Knowledge distillation, NLP, CV, IR, symbolic regression

to explicit search protocols and often integrate findings through meta-analysis require distinct considerations for incorporating new evidence. In contrast, narrative reviews adopt less formal search and appraisal methodologies and serve broader aims of overviewing a research area or informing an agenda for further research. Our study examines the narrative review authoring process, exploring how and when these reviews created with semi-systematic methods should be updated.

5.3 Methods

We conducted in-depth, retrospective interviews with authors of narrative reviews, drawing on their expertise to inform the updating process. To identify authors, we searched dblp⁴ for peer-reviewed narrative reviews (equivalently, *survey articles* or *surveys*) in ACM Computing Surveys (CSUR), a premier journal for surveys in computing research. We identified additional surveys published in ACM CHI and CSCW using the search query “review OR survey.” We filtered surveys to those published 1-3 years prior, allowing time for new research to emerge while ensuring participants could recall and reflect on their processes. We then excluded any that were not actual surveys of computing literature, for instance papers on improving methodologies for online user surveys or analyzing app reviews. Due to IRB restrictions, we excluded papers whose corresponding authors were based outside the United States or Canada. This process yielded a total of 128 surveys, and we invited each corresponding author to participate via email.

⁴<https://www.dblp.org>

5.3.1 Participants

Eleven of the invited authors participated in our study and were included in the subsequent analysis. Six of their surveys were published in CSUR, four in CHI, and one in CSCW. The surveys were published between early 2022 and 2024, with the majority published in 2023. Due to the lengthy peer review and publication process, most participants noted the relevant research work occurred one to two years prior to official publication. Participants' ages ranged from 26 to 40 ($M = 32, SD = 4.5$). Three were female and eight were male. Three were research tenure-track assistant professors, seven were PhD students, and one was a research engineer. There was considerable diversity among participants, both demographically and in survey article topic (Table 5.1). We provide participants' self-identified areas of research expertise (which align closely with the topics of their survey articles they were invited to discuss) but refrain from disclosing the surveys to maintain confidentiality. We anticipated our study size would allow for sufficient theoretical saturation, and that the variety of participants' demographic and research backgrounds would allow us to elicit a rich diversity of survey authoring and updating practices.

5.3.2 Interview Protocol

Our interviews consisted of open-ended questions prompting participants to recall and reflect on their research practices. For instance, participants were asked to describe how they conceived of their survey topic, the composition and responsibilities of their research team, and their approaches for discovering and screening papers, extracting and organizing information, writing, and revising. Probing questions were used to elicit specific facets of participants' experiences and encourage more detailed discussion around particularly meaningful or evocative aspects of their process. Additional details are available in Appendix B.1. Where appropriate, participants were encouraged to share their screen to illustrate their recalled experiences and any artifacts they created. Interviews were conducted virtually via Google Meet, each lasting approximately one hour. Participants were compensated with a \$40 USD gift card. This study protocol was approved by the IRB at the research team's university.

5.3.3 Analysis

Data for analysis included transcripts automatically created from the interview audio recordings and manually cleaned. All personal identifiers were removed from the transcripts to ensure anonymity. We then followed a thematic analysis approach [Braun and Clarke, 2006] to analyze the data. Two team members reviewed all transcripts to familiarize themselves with the data, and then assigned specific codes to two transcripts using an open-coding approach [Strauss and Corbin, 1998, Chapter 8]. An initial codebook was formed by merging these codes and refined through discussion. To align divergent observations, coders identified overlapping codes, clarified code definitions, and excluded codes not aligned with the research objectives. Each coder then independently applied the established codes to half of the remaining data. Emergent higher-level themes were discussed and iteratively induced from the codes, informing the following findings.

5.4 Findings

We first summarize the processes and challenges participants described across four core survey authoring activities. We then examine their motivations, strategies, and obstacles for updating their surveys. Finally, we highlight their perspectives on the potential role of AI in aiding survey authoring and updating.

5.4.1 The Processes in Authoring Survey Articles: Work Practices and Challenges

Search Participants described a diversity of strategies for paper discovery, some more systematic than others. For example, most mentioned using a scholarly database to find papers, though several participants highlighted the challenge of forming an effective search query. P4 illustrated a lengthy process of iteratively searching for papers with an initial query based on his own expertise, scanning relevant papers returned by the initial query to identify other relevant keywords, and repeating the process until saturation. Another participant found it challenging to identify the right keywords to search for, especially in emergent research areas that lacked a consistent vocabulary (P8).

Participants also mentioned using more informal methods of paper discovery, including citation chaining [Webster and Watson, 2002] from foundational papers, receiving recommendations from their social

network, and monitoring prominent research groups in their field. Two participants had no dedicated means of paper discovery, relying instead on a collection of papers gathered throughout their research, and searching for and adding new papers only when necessary during the writing process (P1, P7).

Appraisal After collecting a set of papers, participants described carefully screening those papers for inclusion. Though most could be filtered based on their title and abstract alone, some participants noted how certain papers required more careful inspection of the introduction, implementation, or results to determine relevance. As with paper discovery, emergent research areas complicated paper screening, as P6 described how the lack of standardized terminology made it challenge to “comprehend the essence of what they were trying to say and implement.”

Participants also noted a challenge in identifying and removing duplicated papers (P1, P3, P5, P8, P10). Some papers could have multiple versions, e.g., a pre-print, a conference paper, and a journal article, yet offer the same contributions. De-duplication quickly became “very annoying” and “tedious” over a large corpus (P10). Other strategies appeared more arbitrary and driven by experience. For instance, several participants further filtered “pseudo-duplicate” papers; P1 described this aspect of his screening process as more art than science:

“I found a lot of semi-duplicates. There are some authors who do double dipping. They would write two papers, but they are based essentially on the same prototype. But this is more like art than an algorithm because sometimes I suspected authors use the same prototype, but I could not guarantee it. I tried to remove everything that I would suspect was either not really implemented or was double dipping. I tried to only keep the papers and projects that I knew that actually did something.”

However, these subjective processes rely on the authors’ expertise and are rarely made explicit in the survey article, making them inherently opaque, open to interpretation, and potentially challenging to replicate when updating the survey in the future.

Synthesis After screening papers for inclusion based on relevance and quality, participants often described finding the right organization as the most time-consuming and cognitively challenging phase of their workflow (P1, P3–P6, P10, P11). Organizing hundreds of papers into a multi-dimensional taxonomy was a

lengthy process of iterative refinement, sometimes with multiple authors collaborating to reach a consensus on the appropriate structure, dimensions, and evidence. Reflecting on his workflow, P1 described feeling humbled in his taxonomy development process:

“I had to discard a lot of ideas because initially what I thought it would be just didn’t work. It ended up being nonsense. Our initial dimensions, because they were correct but meaningless, they were not informative. It took a lot of time and humility, a lot of humbleness to accept that those initial things didn’t work.”

Many participants developed a codebook alongside their taxonomy, capturing key questions or dimensions to compare across papers. Predefined codes were based on their surveys’ research questions and own expertise, with new codes added as needed. Revisions were labor-intensive, with P10 noting how adding a code meant revisiting every paper and reestablishing a mental model of the paper to extract the relevant information.

Interpretation Finally, participants highlighted a key challenge in synthesizing and interpreting the literature. They described how beyond the organizational taxonomy, the real value of a survey lies in its identification of key challenges, future trends, and open research opportunities (P3, P6, P7, P9, P11). P6 emphasized this point, while acknowledging the difficulty in articulating those insights:

“Beyond the taxonomy, another important thing about this literature review is the challenges and the future trends you uncovered in the review...because the reason they want to read the paper is because they want to start their own research project. So I did spend a lot of time on these challenges and the trends, on how to make it clear and comprehensive and fancier to give back more insights to those readers.”

He elaborated that while the taxonomy he had developed worked well for categorization, it was too rigid for discussing future trends. Providing “coarse-grained” insights to inspire new research was “tricky” since they were not directly tied to the “logic” imposed by the taxonomy. Deriving meaningful insights can also be challenging due to limited perspective, with P6 adding that identifying emerging trends required insight he gained through discussion with other researchers rather than solely reviewing the literature.

“At that time, my vision was still quite limited. I can read lots of papers and try to summarize in the tables, even just write the sections to introduce each work, but it’s really hard for me to get the sense in five years, what LoRa networking research would be. It’s really hard for me to uncover this kind of future trends just by reading papers. I needed to talk to different researchers working on this research field to understand their vision for this topic. It’s just something I couldn’t do by myself. That’s kind of the most challenging part.”

Another challenge involved the selection and interpretation of key papers. In contrast to SLRs in clinical medicine, which use rigid inclusion and exclusion criteria to identify relevant studies for meta-analysis and avoid introducing bias, participants in our study described more informal search and appraisal processes, while aiming to be comprehensive in their overall search of the literature. Furthermore, not all relevant research could be synthesized given the length restrictions of a survey article, as one participant explained his nuanced process of selecting papers that balanced recency so “readers will find it interesting” and foundational work to avoid “forgetting the theory” (P3). Another participant, P7, included additional meta-commentary, drawing connections between contemporaneous papers with similar contributions and critically evaluating when a paper’s proposed approach was unsupported by its evidence.

5.4.2 The Dynamics of Updating Survey Articles: Motivations, Approaches, and Obstacles

Motivations Participants viewed keeping survey articles up-to-date as a valuable, albeit costly, endeavor. P10 felt keeping his survey current was important, effectively “showing a picture of how the research is at that point.” P11 similarly mentioned how updating his survey would be “helpful for the community,” as many researchers used it as a starting point, though he expressed concern about its relevance, acknowledging, “it’s already two years old.” Participants noted how this perception of value also carried an implicit expectation of currency with the latest research. Some recollected how peer reviewers asked for an update to the paper search, since a year had elapsed while under review and new, relevant research may have emerged (P8, P10). After a subsequent rejection, P8 also proactively updated their survey to include new literature, fearing future reviewers would be “annoyed” that the survey article was too “old and outdated.” For many participants, this expectation manifested as a pressure to publish surveys as quickly as possible and an anxiety surrounding the potential extra work in updating if the paper were not accepted.

Despite believing that surveys should ideally be kept up to date, participants held varying expectations on the desired frequency of updates. For instance, P8, working in a slower-moving field, noticed only a few relevant papers to their survey each year. In contrast, other participants reported encountering many relevant papers to include through passive monitoring of email alerts, conference and journal proceedings, and work citing their survey (P4, P5, P11).

Approaches Participants described three main types of updates they envisioned making to their surveys:

1. **Empirical**, involving the update of quantitative and qualitative evidence presented in text, tables, and figures.
2. **Structural**, involving the update of paper structure, organizational taxonomies, and frameworks.
3. **Interpretative**, involving the update of synthesis and interpretation of empirical evidence, such as the discussion of limitations, implications, emerging trends, unresolved and resolved challenges, and future directions.

First, participants identified *empirical* updates that aimed to incorporate the latest research findings. Empirical updates involve revising individual quantitative results to reflect the state-of-the-art, as well as any recalculating statistical meta-analyses and aggregated values, such as counts and proportions of papers within each dimension of the taxonomy. These updates may also involve refining the qualitative narrative by replacing less effective or outdated references with more compelling data, a process P10 paralleled to selecting the best quote from an interview study. One participant further characterized making empirical updates as a complex decision-making process, as each new piece of evidence requires consideration of whether “*you should replace this, you should add this, or you should just mention it in the table without any text*” (P6).

Second, participants believed empirical updates could eventually warrant *structural* updates, such as adding a new section to a survey to synthesize recently incorporated papers (P2, P5, P6). One participant described how changes in the momentum of a research area could induce a structural update for his survey:

“If some field is becoming more prominent and lots of work is going in a particular direction—maybe using SSVEP now more with VR—then we will highlight it in the discussion, make a

section for it, and explain the reason why we think it's gaining prominence now, highlighting some of the results.” (P5)

P6 further explained that the decision to create a new section in a survey—as opposed to integrating the evidence into an existing section—should depend on the sufficiency of new evidence, saying *“if it's just five papers for this section for this new topic, I don't think it's worth adding a new section. But if there are 20, 30, even 100, it will be really useful to make them into one independent section.”* Two participants described a similar type of structural update they made while revising their survey, splitting a dimension of their taxonomy into two as the number of included papers grew beyond a reasonable organization (P2, P11). Lastly, participants noted how structural updates could arise from significant changes to the social or technical status quo. For instance, many participants mentioned large language models as a notable paradigm shift they would incorporate into an updated survey (P3, P5, P6, P9, P11), with proposed revisions ranging from adding a section on their usage to restructuring around the pre- and post-LLM eras.

Third, participants described *interpretative* updates as a process of critically re-evaluating a survey article, aligning it with the current state of a research field. Participants saw this update as the key cognitive challenge, involving re-synthesizing the proposed challenges, trends, and opportunities given the updated empirical evidence. Interpretative updates could further be self-reflective, as P5 explained:

“It's not just generating new limitations and recommendations, but also considering these recommendations I already made, and seeing how the field is progressing and whether those limitations are being overcome.”

Some participants hoped new research would address prior limitations they had raised in an earlier iteration of their survey, resulting in a discussion of both resolved and emerging challenges in the updated survey (P1, P4, P5).

Participants were generally confident that the original organizational structures and taxonomies they had developed for their survey would remain relevant for several years after its initial publication. P1 highlighted the incremental nature of research he saw:

“If they happen to introduce a completely new method, then the methods need to be updated, a new workflow must be developed, which I doubt because I have not seen anything really

happening. People just squeeze from the previous one, they try to improve the performance of the existing methods.”

Participants therefore viewed empirical and interpretative updates as the bulk of maintaining a living survey, with structural updates necessary only after a critical mass of new work.

Given the perceived stability of their surveys' structure, in considering approaches to updating, participants emphasized the intention to reuse much of their original workflows. Maintaining the same processes would not only conserve effort, but also help ensure consistency in an updated survey. Participants mentioned reuse along logistical aspects—such as reforming the same research team—and mechanistic processes, like reapplying the original paper search criteria across the same scholarly databases, filtering papers with the same inclusion criteria, extracting data from new papers using the existing codebook, and rerunning automated scripts for quantitative analysis.

Obstacles Participants identified the lack of strong academic incentives as a key structural barrier to keeping surveys continuously updated. Despite being highly cited and valuable to the research community, participants saw the impact of surveys as tied to their initial publication. Without structural support in the form of academic recognition, participants found it difficult to invest the time and effort required to revise an existing publication. P4 explained how even when authoring their original survey, they needed to “draw a line” on the timeframe of surveyed literature, as adding new papers was “moving the goalposts” and delayed publication.

In addition to the lack of extrinsic incentives, continuous updating was seen as an “unmanageable” without additional support. One primary challenge was the need to continuously discover and screen new research papers. The search had to be comprehensive while avoiding duplication, and certain means of paper discovery, e.g., social recommendation, would be absent once the original survey team dissolved. In contrast, participants saw refining an organizational taxonomy as less challenging when updating a survey, as their previous efforts could largely be reused or adapted.

Participants also noted that updating a survey with new evidence could trigger a cascade of revisions (P1, P8, P10). For example, adding papers to the taxonomy would require recalculating statistics, revising text, updating charts and figures, and modifying conclusions. Managing these interconnections to ensure

consistency across a long survey article presented additional cognitive challenges. As a potential coping strategy, P1 suggested he would make updates in bulk, which he felt was more efficient than updating one paper at a time. In sum, while participants recognized the community value of living surveys, they struggled to justify the effort due to the lack of personal academic recognition.

5.4.3 Perspectives of AI Use for Survey Updating

AI Lacks Nuanced Understanding for Expert-Level Reviews

Participants expressed concerns about AI's ability to produce surveys that are as insightful and meaningful as those written by human experts (P1, P2, P5–P8). They emphasized that scholarly synthesis was more than just processing and summarizing data, demanding nuanced understanding, critical engagement, and the foresight to identify gaps and future trends that may advance the field. Several participants noted these aspects to be the core value of survey articles which drive meaningful discourse.

Participants also argued that without domain expertise, AI syntheses are likely to lack the depth and complexity needed for such discourse (P5–P8). P8 highlighted the importance of human oversight to avoid shallow analysis:

“I think you can use it. I just think you need to be way more careful about having a human-in-the-loop, and having a human being the final position of authority on what happens during the final part of the analysis. Because otherwise your analysis is going to be incredibly shallow.”

This need for experts as a critical voice was echoed by P5, who emphasized that relying too heavily on AI for updates risks making survey articles formulaic. P5 expressed his belief that if articles were written to be updated by AI, they would lack creativity and reader engagement:

“If we try to set up the paper in such a way that an AI can easily update sections of it to match data, it may not be an interesting paper to read... And I think papers are written to be fun.”

Overall, participants perceived current AI systems as lacking the ability to offer the depth, creativity, and engagement characteristic and required of expert-level survey articles. Without these elements, AI-generated syntheses risked becoming too shallow or formulaic.

Roles of AI in Supporting Updates

While participants were generally skeptical about AI's ability to replicate human-written survey articles, they recognized the potential for delegating repetition or tasks requiring less specialized expertise to the AI. Participants recommended using AI for tasks where mistakes would be less costly (P1, P7, P10). For instance, P10 described how he would only choose to use AI in places where “the results of the paper and the conclusion do not rely very strongly on it.” P7 suggested that AI should only be tasked with the “very easy work,” and for the creative work, such as discussing trends and challenges, AI should only be used to offer “suggestions.” The following summarizes three roles participants highlighted of AI assistance throughout authoring and updating a survey.

AI for routine automation First, participants identified routine numerical updates, such as tallying results or updating statistical information, as tasks where AI could be reliably employed (P4, P5). Automating these tasks could help ensure consistency and comprehensiveness when incorporating updates into a survey. For example, participants expressed the need to extract specific data from a corpus of papers to facilitate coding, and then subsequently ensure the quantitative aspects of the survey were updated to reflect the latest codes. One concern participants shared was the need to verify the work done by any AI system (P5, P10, P11). As P5 explained, “you may not be able to trust the extracted data,” highlighting the importance of proper guardrails to mitigate overreliance and providing affordances for user verification.

AI as a surrogate Participants described specific sub-tasks where AI could be valuable in managing a large corpus of data (P2, P5, P7–P11). For instance, they saw AI as a useful assistant in potentially making the search process more comprehensive by refining and augmenting search strings (P6, P10, P11). For large paper collections, AI was seen as helpful for doing an initial pass at screening and organizing of the papers (P2, P5, P7–P9). However, participants also expressed concerns about AI missing important information or adding irrelevant information. In identifying or updating a corpus of papers, P9 described how it is easier to spot irrelevance but more challenging to detect what may have been overlooked. Furthermore, some participants emphasized the importance of establishing proper ways to document AI usage for reproducibility and transparency in research (P8, P10).

Table 5.2: Opportunities for supporting various updates to narrative literature reviews.

Update Type	Description	Opportunities for AI Assistance in Updating
Empirical	Revision of quantitative and qualitative evidence presented in text, tables, and figures	<ul style="list-style-type: none"> • Identify and recalculate numerical values or meta-analyses • Update references to evidence in figures or tables • Change or revise representative examples in text
Structural	Revision of paper organization, taxonomies, and frameworks	<ul style="list-style-type: none"> • Reevaluate included evidence and definition of an existing taxonomy dimension • Group emerging research to align with or challenge existing taxonomy dimensions • Identify opportunities to improve paper organization (e.g., by splitting sections that exceed a threshold of research)
Interpretative	Revision of evidence synthesis, interpretation, and narrative framing	<ul style="list-style-type: none"> • Identify potential biases, assumptions, and narratives in the original survey that may be validated, challenged, or updated in light of new evidence or technologies • Analyze the extent to which prior gaps and limitations in the research have been addressed • Propose new unresolved challenges or under-explored areas

AI as a second opinion Participants discussed several places in the authoring and updates process where AI could serve as a co-author providing a second opinion (P1, P5, P6, P9–P11). For instance, when analyzing trends in a cluster of research papers, AI could do a “sanity check” in case the researcher misses anything (P9). Similarly, when working alone and screening papers, AI could “provide an opinion” alerting the researcher to papers that might require taking a closer look (P5). These examples highlight how integrating AI as a collaborative partner could be beneficial within authoring or updating, to enhance systematicity, reduce cognitive bias, and encourage a more comprehensive and insightful synthesis.

5.5 Discussion

5.5.1 Summary of Findings

Our study examined the work practices within survey article authoring and the challenges researchers foresee in maintaining their currency. Our findings reiterate four key processes of review—search, appraisal, synthesis, and interpretation—and highlight the many inherent subjective, expertise-driven decisions and idiosyncratic strategies. Participants noted the value in keeping their surveys updated for the research community and described three main types of updates they would make: empirical, structural, and interpretative. Participants characterized updating as a process of restoring institutional knowledge and attempting to adapt and reuse their original workflows. For instance, new research can be identified through established search methods and integrated into an existing taxonomy, while sufficient research could further motivate deeper structural and interpretative revisions.

Despite the perceived value, participants also expressed how continuous survey updating was infeasible and misaligned with existing academic incentives. Many therefore recognized the potential for AI support to lower the costs of updating, such as by helping to identify new, relevant papers or automating repetitive information extraction tasks, though skeptical of its ability to replace their own learned expertise for synthesizing evidence. Our findings, while focused on scholars in the computing field, align with perspectives from other empirical studies involving scholars across diverse disciplines [Messori and Crockett, 2024; Morris, 2023b; Chubb et al., 2022]. For instance, Messori and Crockett [2024] find that objective tasks are more conducive to helping scholars establish appropriate trust in AI assistance, and several other studies similarly suggest scholars are more receptive to AI’s role in narrow tasks that boost personal productivity but are hesitant to rely on AI in “emotional tasks” that require creativity and complex decision-making. For these nuanced tasks, participants instead see AI as suitable to “augment and assist human judgment” [Chubb et al., 2022].

5.5.2 Toward a Vision of Living Narrative Reviews

Considerations for AI-Assisted Updating of Narrative Reviews

Our findings offer several considerations for how future AI systems could aid updating workflows. First, systems could help determine *when* an update is needed by continuously monitoring new research and identifying those relevant to existing taxonomies and syntheses. They could further explain *why* new research is relevant, such as by highlighting how it aligns or challenges existing parts of a survey. Updating efficiency could be improved by localizing *where* updates are needed and directing authors' attention to specific components—text, figures, or tables—that should be revised. This could be especially helpful for cascading revisions, as participants noted, since integrating even a single paper may require updating many parts of a survey, such as where it is introduced, synthesized with related research, and visualized within tables or figures. Finally, systems could offer guidance on *how* to integrate new research into a survey, for instance by identifying the type of update required—empirical, structural, or interpretive—and providing tailored revision recommendations.

These three different types of updates in narrative reviews also present unique challenges that AI systems could help address. For empirical updates, systems could help recalculate numerical values and meta-analyses given new research, where providing transparency to allow researchers to verify the updated evidence is important. More challenging is supporting potential structural updates, such as determining when new research is sufficient to extend or revise a taxonomy and its dimensions, which participants felt required a deep and nuanced understanding of how new research may reshape the overarching conceptual framing of a research area. Future AI systems could also likely aid in interpretative updates, for instance by meaningfully synthesizing literature within a taxonomy dimension, reevaluating gaps in a research area, or informing an evolving agenda for future research. These technologies are promising yet nascent, with recent studies echoing participants' concerns regarding the shallow nature of AI-generated literature syntheses [Martin-Boyle et al., 2024]. Overall, narrative review updating remains for now a collaboration of AI and human effort, though given the high costs and low incentives of making frequent updates, realizing living narrative reviews may eventually require a concentration of human effort on verifying and steering AI-assisted updates.

Systematicity and Expertise in Narrative Reviews

Designing future systems to enable living narrative reviews requires balancing systematicity with the subjective, expertise-driven nature of survey authoring. While evidence-based disciplines like clinical medicine emphasize systematic methodologies—such as transparent, repeatable methods for identifying, appraising, and synthesizing literature—these standards may not fully align with semi-systematic methods such as in computing research surveys. Participants described workflows marked by subjective decisions informed by years of specific research expertise, across literature discovery, taxonomy creation, synthesis, and narrative construction. These choices, shaped by implicit knowledge and anticipation of reader needs, were rarely documented in surveys themselves, posing challenges for both subsequent human-driven and AI-supported updates. Without sufficient systematicity or documentation, restoring the institutional knowledge of a research team or replicating the original idiosyncratic processes to perform consistent updates becomes significantly harder.

Future AI systems may bridge this gap by supporting authors in externalizing the implicit strategies underlying their subjective decisions. Interactive tools could invite authors to articulate their search and synthesis methodologies more explicitly, clarify decision-making criteria, and document expertise-driven strategies, such as when participants described targeting specific journals or research groups. An exciting direction for future research lies in the design of mixed-initiative systems to then leverage these explicit methodologies and infer patterns in existing surveys to provide updating assistance aligned with the author's narrative and methodological intent, while making transparent any inferred subjective choices. Altogether, these approaches can serve to encourage systematicity in survey authoring and ensure subsequent updates to a survey are efficient, reliable, and consistent with prior survey iterations.

5.5.3 Limitations and Future Work

The retrospective nature of the interviews could introduce recall bias, as participants may have selectively reported the most salient challenges, overlooking others. Future work could use more direct observational methods, such as contextual inquiry, to capture authors' real-time processes as they update their literature reviews. To complement authors' perspectives on continuously updated reviews, additional studies could investigate *readers'* needs. For instance, what form should a living review take and how might new affor-

dances and interactions facilitate their use? Similar techniques could also aid peer review of review updates, which remains an open challenge for realizing living reviews.

5.6 Conclusion

In this chapter, we presented findings from a retrospective interview study with 11 authors of computing survey articles—a common manifestation of literature review synthesis—to examine their practices for creating, maintaining, and updating their surveys. We identified three main types of updates (empirical, structural, and interpretative) each with unique motivations, challenges, and opportunities for AI assistance. Although participants supported the idea of continuously updated reviews, they were skeptical about its feasibility due to current academic incentives, suggesting the important role AI tools may serve in facilitating or automating these updates. This study takes a first step toward a deeper understanding of when, why, and how these reviews should be kept up-to-date, and informing the development of future tools to enable living narrative reviews, an example of which we explore in the following chapter.

Chapter 6

DimInd: Scaffolding Synthesis with Intermediate Structured Representations

While previous chapters explored human-AI approaches for individual paper comprehension (QLARIFY in Chapter 3 and SCIM in Chapter 4), our empirical study of survey article updating revealed that a persistent challenge—and perhaps also greatest opportunity for AI assistance—lies in supporting the sensemaking processes over large paper collections through organization of knowledge into meaningful structural representations. To explore how structured mediation can address such literature synthesis at scale, we developed DIMIND, a system that implements the three core mechanisms: *discovery* through AI-suggested faceted columns that provide information scent across paper collections, *iteration* through interactive refinement of multi-level taxonomies and organizational schemas, and *verification* through clear provenance linking between generated summaries and specific text within source papers. DIMIND transforms unstructured paper collections into successive structured representations—faceted literature review tables, hierarchical concept taxonomies, and controllable narrative syntheses—supporting researchers across more repetitive tasks of information extraction, e.g., for the empirical and structural updates they identified as burdensome, while preserving their agency over the interpretive synthesis work they viewed as their core scholarly contribution. This research and corresponding preprint [Fok et al., 2025a] is under review and has not yet been published.

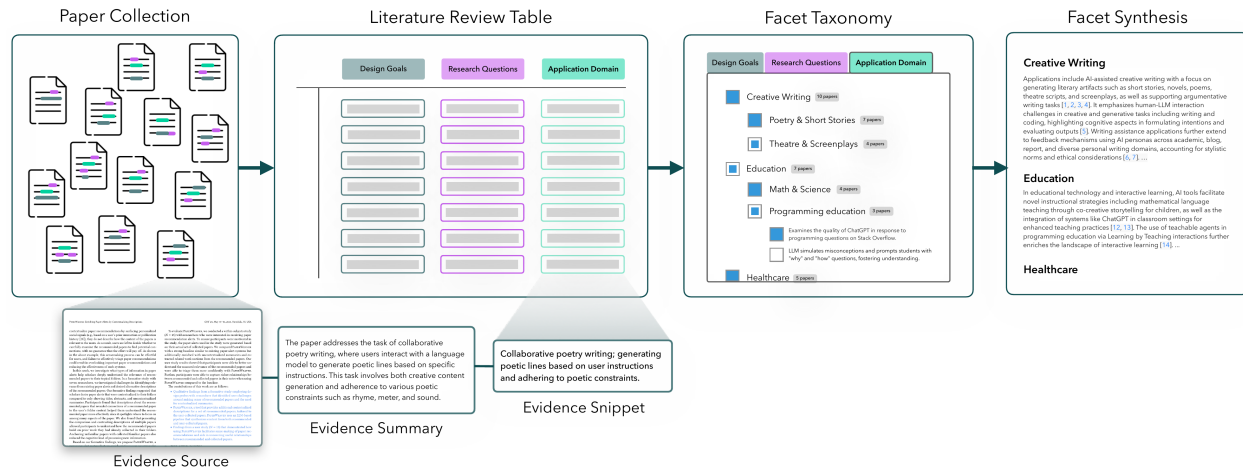


Figure 6.1: We present an LLM-assisted workflow aimed at scaffolding literature review over large paper collections, and instantiate it in a prototype system, DIMIND. Users can interactively construct and explore four successive *structured representations* of literature information: a *paper collection* listing papers and metadata, a *literature review table* with columns that render relevant evidence snippets from paper full texts along defined facets, a *facet taxonomy* that organizes faceted information into a higher-level conceptual overview, and a *facet synthesis* that provides a controllable faceted summary across the collection.

6.1 Introduction

Literature review offers a foundation for scientific progress. Researchers spend significant effort making sense of prior work throughout a research project, from finding inspirations for ideation, exploring techniques and datasets to develop methods, to writing related work to contextualize their own work. One way this effort can be shared and reused is in the form of survey papers that comprehensively review a large collection of papers. However, the process of reviewing and synthesizing a large set of papers is time-intensive [Michelson and Reuter, 2019b; Shojanian et al., 2007; Borah et al., 2017] and cognitively challenging [Knight et al., 2019; Granello, 2001]. As a result, this process is often conducted with a team of researchers over multiple weeks or months and rarely feasible for individual scholars. For example, Pang et al. [2025] had seven authors who collectively reviewed 186 papers, while Lee et al. [2024a] had an even bigger team of 36 researchers reviewing 274 papers. In the field of medicine, Cochrane coordinates a team of over 28,000 contributors to produce its timely review articles [Chandler and Hopewell, 2013a].

One main challenge is the sheer amount of information scattered across disparate papers that needs to be extracted, analyzed, compared, and synthesized [Snyder, 2019; Lee et al., 2024a]. As a result, researchers

often follow a step-by-step process to make the tasks more tractable, involving successively transforming information from one representation to another that is one degree more synthesized. For instance, the process begins with gathering a large collection of relevant papers, then reading and summarizing those papers individually, coding the summaries into common themes or aspects in large spreadsheets [Hashimoto et al., 2017; Newman et al., 2024], building out a taxonomy of research threads [Kang et al., 2022; Palani et al., 2023], and finally, synthesizing this structure into writing. Information management and coordination often involves large tabular spreadsheets to keep track of papers, summaries, qualitative codes, and themes.¹ Currently, much of this work is labor-intensive and low-level, involving significant back-and-forth—from manually inspecting the full text of papers and extracting information across many facets, to revisiting the full text to check for accuracy—incurring substantial cognitive costs [Pirolli and Card, 2005; Cooper, 2015; Khalil et al., 2022]. And rapid growth in scholarly publication means that the effort required to produce such reviews will only continue to increase [Bornmann and Mutz, 2015; Jinha, 2010].

Recent work has suggested large language models (LLMs) can be effective at facilitating this information extraction and compression process over large paper collections but these explorations have largely focused on “*data tables*” or task-specific structures, such as size and accuracy of AI models or intervention and patient demographic in clinical trials [Elicit, 2023; Wang et al., 2024a]. In contrast, their ability to extract nuanced qualitative insights that are useful for researchers to create more narrative literature reviews remains underexplored, and more importantly, existing approaches typically produce only flat, single-level representations (such as tables) rather than the multiple, interconnected levels of abstraction needed to support the progressive synthesis of comprehensive literature review.

In this chapter, we present DIMIND, an interactive system for LLM-powered literature review support that guides researchers through successive structured representations of paper information. Starting with a **collection of papers**, users iteratively construct a faceted **literature review table** that organizes relevant information extracted from papers’ full texts by defining custom facets in natural language or selecting from collection-aware facets suggested by the system. The system then transforms each faceted column into a hierarchical **taxonomy of concepts**, surfacing emergent themes across papers. These taxonomies can be then explored, refined, and transformed into **narrative syntheses**. Users are guided through these structured

¹e.g., <https://writing-assistant.github.io/#annotated-papers>, <https://github.com/Social-Futures-Lab/skin-deep/blob/main/Literature%20Review.csv>

representations via interaction and visual affordances. This progressive disclosure not only scaffolds sense-making, but also supports the steering and verification of information, which is critical when working with LLM-generated content.

To evaluate how DIMIND can better support analyzing and synthesizing large collections of papers, we conducted a within-subjects user evaluation with 23 computer science researchers where they reviewed two sets of 50 research papers in two 30-minute sessions. Comparing DIMIND against a baseline literature review approach assisted by a commercial LLM-based chat application (ChatGPT), we found that DIMIND’s structured representations effectively supported users in extracting, organizing, and verifying information across papers with less effort. Our qualitative findings further reveal how researchers used tables as information scent and taxonomies as navigational hubs, transforming information into manageable views that facilitate movement between high-level organization and detailed exploration, while balancing LLM assistance with their own scholarly agency. We conclude by discussing the tradeoffs of structured versus conversational assistance and highlight opportunities and tensions in LLM-assisted literature review workflows.

6.2 Design Goals

Our design goals are motivated by gaps in prior work in user-driven literature understanding tools. We also draw from cognitive psychology theories on how knowledge workers make sense of large amounts of information, specifically, sensemaking and information foraging theory.

Literature review requires identifying, understanding, and organizing multi-faceted relationships across many research papers [Snyder, 2019; Hashimoto et al., 2017]. We are inspired primarily by Pirolli and Card [2005]’s notional model of sensemaking, which describes how analysts iteratively collect, organize, and synthesize information to generate insights. This framework conceptualizes effective information analysis as “the process of creating a representation or *schema* to answer task-specific questions,” where schemas are structured, often externalized, knowledge artifacts created and refined during sensemaking [Russell et al., 1993]. Applied to literature review, such structured representations can scaffold the cognitive progression [Anderson et al., 2001] from low-level information retrieval to higher-order processes of analyzing, evaluating, and synthesizing in scholarly sensemaking.

Information foraging theory [Pirolli and Card, 1999] further highlights two forms of *environmental enrichment* that can optimize gathering relevant information from information patches (e.g., research papers)—1) reduce the cost of navigating between patches, and 2) make patches yield better returns of valuable information. A structured literature review table serves both these purposes: it functions as an externalized schema for sensemaking while organizing information across papers into a unified view (reducing navigation costs) and extracting key findings (improving information yield). These theories guide our following three design goals for LLM-assisted literature review workflows, addressing common challenges in literature review processes observed in prior work [Khalil et al., 2022; Daniel, 2022; Chen et al., 2016; Thomas et al., 2017a; Fok et al., 2025b].

6.2.1 Design Goals

DG1. Help researchers transform a large, unstructured literature collection into a structured repository of extracted relevant information.

Literature review requires researchers to process large volumes of unstructured content into meaningful, structured representations. A system should reduce the cognitive costs of extracting, organizing, and accessing relevant information across multiple papers. By creating a structured repository, i.e., a literature review table, that facilitates organization and easy navigation of key information across papers, researchers can shift their mental resources toward higher-level processes of evaluation and synthesis, rather than the tedious and repetitive process of information extraction.

DG2. Help researchers better make sense of vast information across many papers, transitioning between low-level details and high-level patterns.

Analyzing patterns and deriving insight across multiple papers is one of the key cognitive challenges in effective literature review, particularly as the number of papers in a review grows. A system should aim to support researchers in making sense of their information repository, and particularly in navigating between successive levels of analysis and structure, e.g., from individual paper review to facet-grounded comparisons to collection-wide synthesis. This could involve providing additional structures that transform the literature review table to scaffold and guide researcher judgment of extracted information, or surfacing potential pat-

terns across papers while enabling researchers to examine and evaluate these suggestions based on their expertise.

DG3. Help researchers quickly validate the veracity of AI-generated content.

Oversight and evaluation of information processed by an AI system and not researchers themselves is critical in maintaining the integrity of a literature review. As LLMs become increasingly integrate into these workflows—suggesting, extracting, and summarizing information—appropriate mechanisms are needed for researchers to efficiently verify the assistance these models provide.

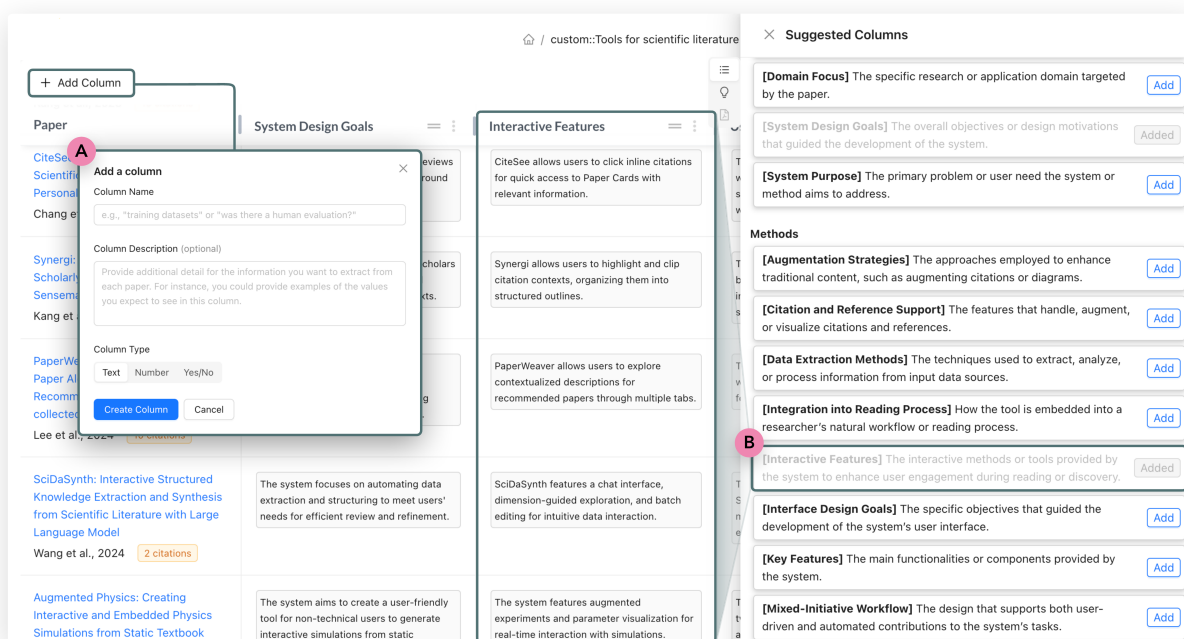


Figure 6.2: Columns can be added to the literature review table in two ways: A) *User-defined* columns precisely specify a faceted information need and allow additional context for steering LLM assistance; B) *System-suggested* columns offer collection-aware recommendations for columns that can be added with a single click.

6.3 System Design

Based on these design goals, we present DIMIND, an interactive system that supports researchers in exploring and making sense of large literature collections by using large language models (LLMs) to transform

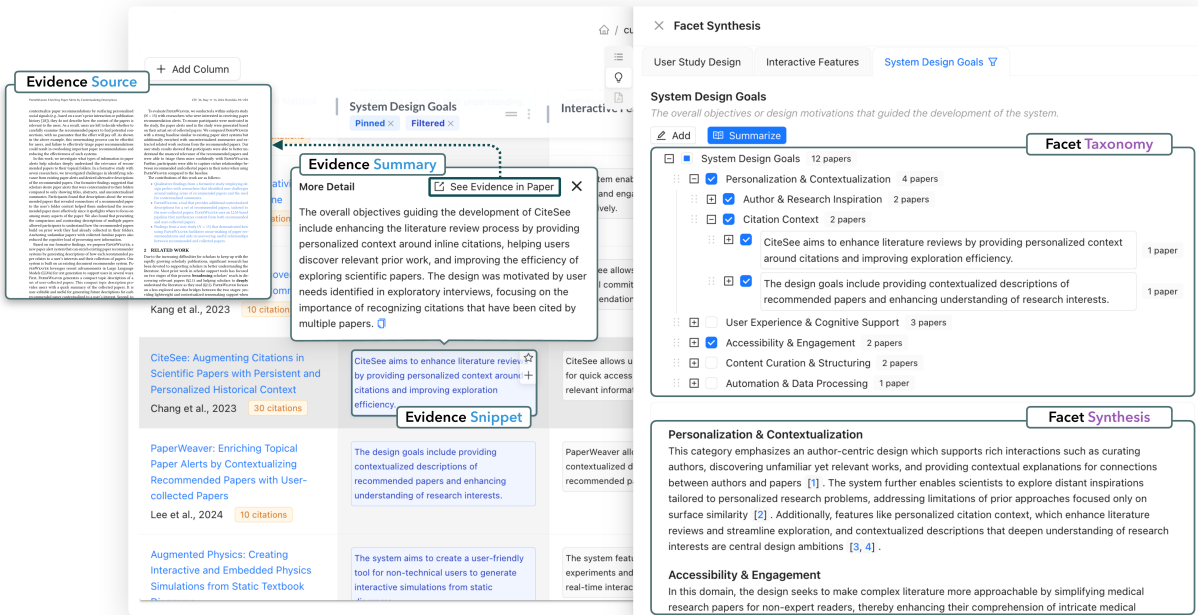


Figure 6.3: In DIMIND, users review large paper collections by navigating and analyzing information across various structured representations. Each cell in the literature review table is a snippet of faceted information from a paper (evidence snippet). Clicking on a snippet shows a popover with additional detail (evidence summary), with a button that can further open the paper PDF in an integrated paper reader with attributed paragraphs highlighted (evidence source). Faceted columns are transformed into distinct hierarchical taxonomies (facet taxonomy), which can be explored, refined, and used to controllably generate a narrative summary with citations (facet synthesis).

dense paper information into a series of linked, structured information representations. With DIMIND, users employ natural language to convey specific information needs, and in response, the system populates a literature review table by creating a column in the table generated using evidence retrieved from the full texts of papers in the collection. DIMIND also suggests several collection-aware columns to help users get started (**DG1**). DIMIND further organizes the information within each column into a hierarchical taxonomy overviewing the available themes across the collection, and allows controllable generation of a narrative synthesis that bridges the schematization and presentation phases of sensemaking (**DG2**). Connecting these information abstraction layers through interaction, DIMIND facilitates bi-directional exploration between the raw information within individual papers and broader themes that span across multiple papers (**DG3**).

6.3.1 Interface and Example User Scenario

To illustrate the design and features of DIMIND, we present a user scenario featuring Juno, a researcher in human-AI interaction who is interested in exploring the use of LLMs as evaluators for complex tasks. Before setting a specific research direction, she decides to consult the literature to better understand the existing work. She starts by compiling relevant papers she had saved and results from an academic search engine using the query *LLM-as-a-Judge*. We join Juno as she uploads her collection of 83 papers to DIMIND.

Defining faceted columns to extract relevant information across papers.

To start, DIMIND transforms her paper collection into a **literature review table** with a single *Paper* column, where each cell contains relevant metadata for a single paper (title, author, and citation count). Based on her prior knowledge and reading the papers' titles, Juno has several questions she's interested in exploring. For example, she had noticed two papers in her collection that use LLMs as evaluators for scientific ideation, and wonders, "*Where else has an LLM-as-a-Judge paradigm been used?*" She clicks to create a *user-defined column* in the table representing the information facet she wants to explore (Figure 6.2). In the column creation modal, she specifies the facet (*Application Area*) in the Column Name, offers specific examples ("*scientific ideation, creativity, etc*") in the Column Description to guide the information generated by the system, and leaves the Column Type as the default type of text.

Using Juno's specification, the system adds a new column to the table and populates each cell in the column with a short snippet of relevant information (**evidence snippet**) generated from the full text of the corresponding row's paper. Juno scrolls through the table, scanning keywords within the new column. She notices several commonalities at a glance—multiple papers have used LLM-as-a-Judge in applications across science, finance, and law.

Exploring system-suggested columns

Next, Juno checks out DIMIND's *system-suggested columns* by opening a side panel anchored to the right of the screen, containing a list of up to 20 suggested facets that are tailored to her specific paper collection (Figure 6.2). These suggestions could serve to guide novice researchers in exploring an unfamiliar collection, while also aiding experts who may find it easier to recognize facets of interest than to recall them from

memory. Each suggested facet has a short name in bold and an italicized sentence-long description. While browsing the list, Juno notices several relevant facets she hadn't considered. She clicks the **Add** button for one labeled *Challenges and Limitations*, prompting the system to add the faceted column to her literature review table and start extracting relevant information.

Information scent and progressively disclosed details

As she scans over the new column, one snippet catches her attention: “*Like human evaluators, LLMs evaluations are also found to have certain biases.*” DIMIND shows concise snippets in each table cell by default to reduce information overload, opting instead to progressively disclose relevant details on demand. Juno clicks on the snippet in the table, showing a popover with a paragraph of additional detail (**evidence summary**), which after reading, she finds has satisfied her previous information curiosity. If she wants to read more about these biases in the authors' own words, she can click the **See in Paper** button to open the paper's PDF in an integrated paper reader within the side panel. When opened this way, the reader also highlights a block of text relevant to the snippet, guiding her attention and offering a useful entry point into the full text for a deeper dive (**evidence source**). Instead, she decides to click on the title in the *Paper* column, revealing a popover with the paper's abstract. She skims over the abstract, making a mental note for her future self of how the paper offers a unique angle on biases in LLM-based evaluations. Together, these snippets in the table offer information scent, while the various levels of detail revealed through interaction allow Juno to control her own depth and direction of exploration.

Facet-focused sensemaking across documents

The facet columns allowed Juno to quickly extract relevant information across many papers and served as information scent to drill-down and read specific parts of each paper. However, it can still be difficult to go through the extracted values across all 83 papers to identify higher level themes and their distribution. For this, as Juno adds columns to the table, the system automatically organizes extracted information within each facet into a hierarchical **facet taxonomy** (one per column) (Figure 6.4). Switching to the synthesis panel, she selects a tab for *Application Area*, the first column she added, revealing a tree-like taxonomy grouping related snippets under high-level categories. Each category in the taxonomy displays the number

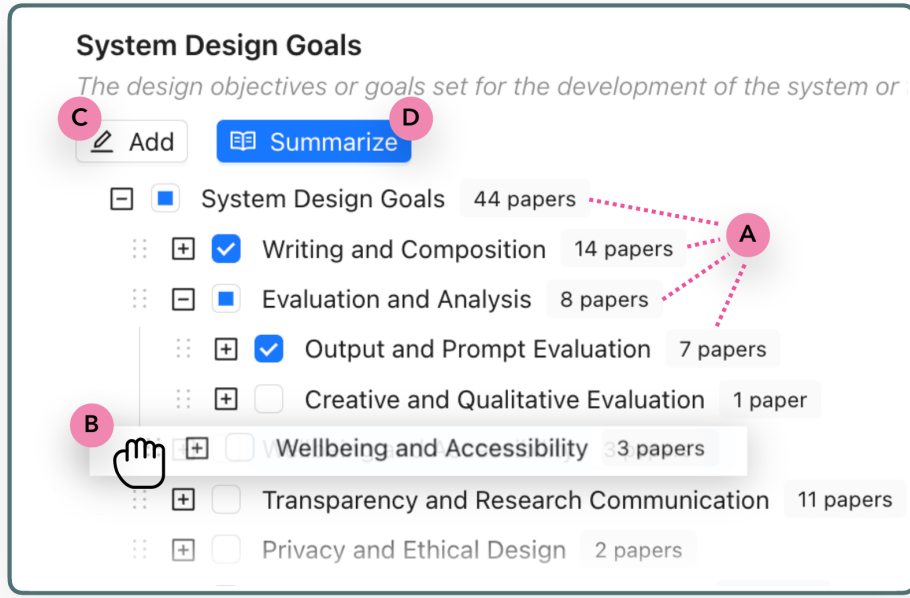


Figure 6.4: The facet taxonomy. Each category shows the number of included papers (A). Users can manually refine the taxonomy through drag-and-drop interactions (B) or add additional categories (C). If at least one category is selected, the taxonomy can be summarized into prose (D).

of associated papers, and the taxonomy is sorted to place common categories with more papers at the top, allowing her a quick overview of the information landscape.

Juno begins to explore the taxonomy, scanning the overarching categories before expanding specific ones to reveal their subcategories. Selecting a category highlights the corresponding snippets in the table (Figure 6.5), allowing her to quickly identify which papers are included in each category, ground the category in lower-level representations in the table (i.e., information extracted from each paper), and potentially reason over any gaps in coverage. While most categories appear appropriate, Juno notices two sub-categories in different levels that should be grouped together. She drags one category to the other, merging them under the same parent. As she continues refining the taxonomy, she shapes both the system’s organization and her own understanding along this facet.

Now satisfied with the structure, she selects the top three categories, comprising the majority of papers in her collection, and clicks the `Summarize` button, prompting the system to generate a **facet synthesis** aligned with her refined taxonomy (Figure 6.5). The synthesis presents a summary generated using only the papers from the selected taxonomy categories following the taxonomy structure, with inline citations which can be clicked to reveal citation cards containing the paper’s metadata and abstract, allowing a quick,

...highlights associated table cells

Paper	System Design Goals
ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing Arwip et al., 2023 55 citations	The design goals include supporting on-demand hypothesis testing and systematic evaluation of LLM outputs.
How Do Analysts Understand and Verify AI-Assisted Data Analyses? Gu et al., 2023 18 citations	The system design goals facilitate analysts' understanding of AI analyses and enhance verification granularity.
Behavioral: Simulating Conflict to Teach Conflict Resolution Shahh et al., 2023 32 citations	The design goals create an interactive tool for teaching conflict resolution strategies through simulated scenarios and feedback.
"It's a Fair Game", or is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents Zhang et al., 2023 42 citations	The system design goals create ethical LLM-based agents that enhance user privacy awareness and control.
EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria Kim et al., 2023 44 citations	The design goals automate evaluation of outputs based on user-defined criteria and facilitate inspection through explanations.

Selecting in the facet taxonomy...

System Design Goals
The design objectives or goals set for the development of the system or tool

- System Design Goals 44 papers
- Writing and Composition 14 papers
- Evaluation and Analysis 8 papers
- Output and Prompt Evaluation 7 papers
- Creative and Qualitative Evaluation 1 paper
- Transparency and Research Communication 6 papers
- Educational and Teaching Tools 5 papers
- Wellbeing and Accessibility 3 papers
- Privacy and Ethical Design 2 papers
- Programming and Prompt Engineering 2 papers

...controls the generated summary

Writing and Composition
The design objectives address enhancing user communication and writing quality by offering tools that improve writing style, emotional awareness, and cross-cultural communication [1]. Goals also target improved productivity and creative support, ensuring users maintain control over AI-assisted content and retain a sense of personal ownership in their tasks [2]. The frameworks cater for verifying and refining language outputs by providing semantic confidence measures and evidence-based annotations [3]. Specific initiatives include capturing the interactive writing process through diverse tasks, supporting iterative exchanges in environments ranging from social dialogues to structured creative scenarios [4, 5], and reducing discrepancies between user intentions and system operations [6].
Additional design strategies emphasize collaborative and interactive composition. Efforts involve integrating opinionated text suggestions to mimic social media interactions while also enabling on-demand AI feedback that reflects target reader personas for richer ideation and editing processes [7, 8]. The goals extend to supporting ideation in both individual and group settings through multiple suggestion modalities and exploring hierarchical content generation processes for narrative coherence [9, 10, 11, 12]. Furthermore, specific design considerations integrate provenance tracking, interactive co-authorship, and the flexibility for subjective interpretations, ensuring reusability and adaptability across various writing tasks [13, 14].

Evaluation and Analysis
The evaluation-oriented design goals focus on automating and systematizing the analysis of generated content and user prompts. This includes enabling user-defined criteria for output review and supporting hypothesis-driven evaluations of model performance with detailed explanations [15, 16]. A key objective is to ensure that such methods serve as valid proxies for human requirements while incorporating rich metadata that improves contextual understanding during model interactions [17, 18].
Furthermore, the design goals aim to support the rapid and systematic assessment of language model outputs without demanding extensive programming expertise, thereby facilitating on-demand hypothesis testing and iterative improvements in model behavior [19]. Efforts also concentrate on bridging data-oriented and procedure-oriented approaches to enhance the verification and communication of AI-generated analyses in real-world applications [20, 21].

Figure 6.5: Selecting specific categories in the facet taxonomy: 1) highlights cells for the included papers in the literature review table, allowing users to quickly delineate between and browse the selected (and not selected) papers; 2) controls the structure and papers included in the generated summary.

in-situ assessment of relevance (Figure 6.6). She saves her work by copying the generated summary with references in a single click, exporting it to her note-taking document for future use.

6.3.2 Technical Implementation

In this section, we describe DIMIND's computational pipeline, comprising four LLM-enabled components: **Facet Discovery** (§6.3.2) in which collection-aware comparative dimensions are generated, **Value Extraction** (§6.3.2) in which faceted information is retrieved from and attributed to papers' full texts, **Taxonomy Creation** (§6.3.2) in which information within a facet is clustered into emergent themes, and **Synthesis** (§6.3.2) in which information is organized into a coherent narrative for presentation.

Facet Discovery

DIMIND automatically identifies high-level comparative dimensions (facets) that serve as analytical lenses through which users can jumpstart or deepen their exploration of the literature. To induce these facets, DIMIND employs a three-stage process grounded in the paper collection. First, we randomly sample n subsets, each containing k papers from the overall collection. For each subset, we prompt an LLM to generate candidate facets using context formed from combining the titles and abstracts of papers in the subset. Finally, we use an LLM to consolidate these n sets of candidate facets into a cohesive final set,

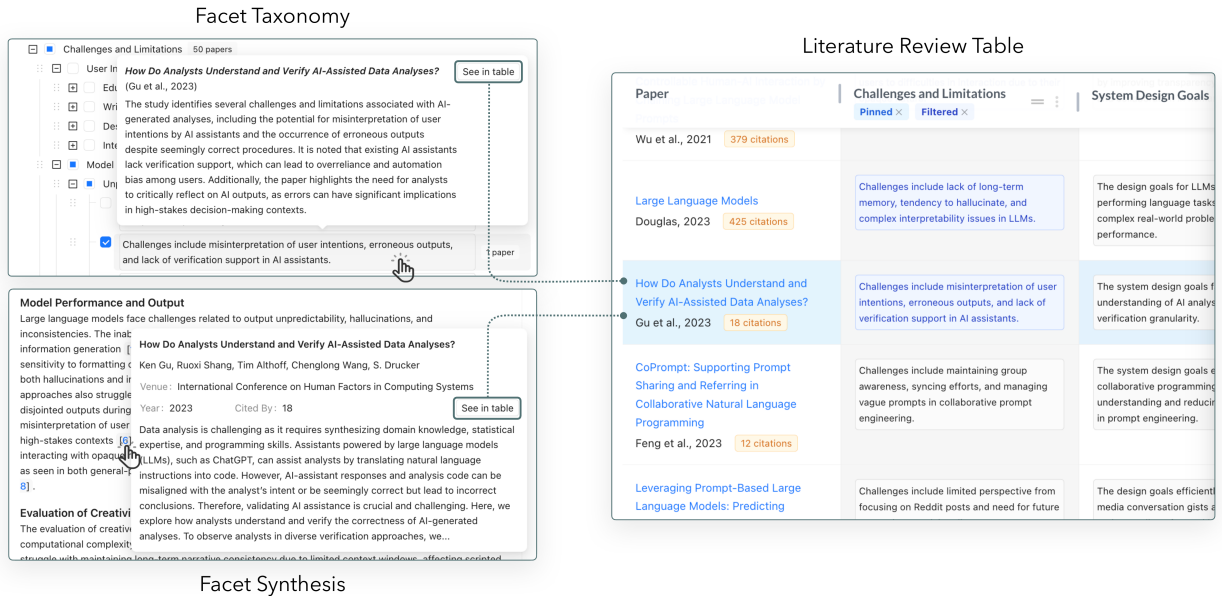


Figure 6.6: Users can view additional detail while exploring the synthesized representations: 1) Clicking an evidence snippet in the facet taxonomy shows the full evidence summary; 2) Clicking a citation in the facet synthesis shows an in-situ citation card. From either, users can click *See in Table* to scroll directly to the corresponding row in the literature review table.

prioritizing facets that appear across multiple subsets and merging semantically similar ones. We empirically chose $k = 4$ papers per subset—we found larger subsets led the LLM to produce many generic facets in an effort to form dense connections (e.g., “Main contribution” or “Findings”) that, while informative, lacked the desired specificity for analysis. Similarly, we found $n = 4$ subsets sufficient, as more subsets rarely yielded additional unique facets.

Value Extraction

When a user defines a facet of information they wish to explore (e.g., by creating a new column in the table), we use a retrieval-augmented LLM to generate a paragraph of relevant information from each paper’s full text. To avoid overwhelming users as the literature review table grows, we use a second LLM call to distill these dense paragraphs into single sentences. These LLM calls are heavily parallelized to minimize the time users must wait for values to be populated in the table after specifying a column. When users click to view evidence for a snippet attributed to the source PDF, the system encodes both the snippet and full text chunks using a text embedding model (all-MiniLM-L6-v2) and then highlights the chunk with maximal cosine

similarity.

Taxonomy Creation

To connect information across papers within a facet, we use an LLM to cluster the generated snippets into a hierarchical taxonomy. This taxonomy creates a nested tree structure with a dynamic depth, adapting to the natural organization of the information (up to a maximum of 5 levels). Each level represents increasingly specific categorizations, with leaf nodes containing the actual information snippets. The LLM is directed to create meaningful categories that avoid broad categories or excessive fragmentation, such that each category aims to have a reasonable number of conceptually similar snippets. Every paper is required to be included in the taxonomy, and information from a single paper may span multiple categories when appropriate.

Synthesis

The final component supports transformation of the facet-specific hierarchical taxonomy into a coherent narrative, making it more accessible for reuse and sharing. We use an LLM to generate a structured summary that follows the organization of selected branches in the taxonomy, allowing users to steer the narrative based on their specific research interests. The LLM is instructed to ensure their synthesis includes all papers associated with the selected nodes and that any generated claims are explicitly attributed, providing citations immediately after each claim. The result is a comprehensive and tightly attributed summary that aggregates information across papers while clearly tracing information back to its source.

Additional Details

DIMIND is implemented as a web application with Flask (Python) backend and React (Typescript) frontend. Paper collections can be created in the system from a list of titles or Semantic Scholar IDs, or created interactively from a search query, with relevant papers fetched using Semantic Scholar’s paper relevance search². Given a collection, the system downloads all available open-access PDFs and uses GROBID [Grobid, 2008–2025] to parse full text, section data, and token bounding boxes from the PDFs. Semantic Scholar

²https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_relevance_search

Survey Paper	Topic	Taxonomy Dimensions
Lee et al. [Lee et al., 2024a]	Intelligent and interactive writing assistants	Task , User, Interaction, Technology , Ecosystem
Pang et al. [Pang et al., 2025]	Use of LLMs in HCI research	Application Domains , LLM Roles, Limitations & Risks

Table 6.1: The survey topics and taxonomy dimensions used in the tasks, and their source papers.

APIs are used for fetching paper metadata, including authors, year, venue, and citation count. Most LLM-enabled components use OpenAI’s `o3-mini` (with “low” reasoning), except for value extraction which uses `GPT-4o-mini`. The specific models and parameters were selected to balance usable interaction latency and performance for each component. Additional details and LLM prompts are available in Appendix C.3.

6.4 Evaluation

We conducted a within-subjects user study with 23 researchers to evaluate the effectiveness and usability of DIMIND, in contrast to a more conventional workflow involving manual review with conversational LLM assistance. Our evaluation aimed to answer the following questions:

- RQ1.** How does DIMIND’s workflow of scaffolding literature review with generated structured representations compare to a manual approach assisted by an LLM chat-based baseline?
- RQ2.** How do researchers leverage different structured representations provided by DIMIND to make sense of information scattered across large collections of research papers?

6.4.1 Participants

We recruited 23 computer science researchers (13 female, 10 male; Age: $M = 27$, $SD = 4$) via university mailing lists, social media recruitment messages, and snowball sampling. Based on a screening survey, we filtered out those who self-reported no prior research experience or were not at all comfortable with reading abstracts and papers in HCI, since our study involved literature review tasks within this domain. Otherwise, participants were recruited on a first-come, first-served basis. Participants consisted primarily of PhD students (17/23), with most participants having at least 3 to 5 years of academic research experience

(17/23). Participants’ primary research areas were largely HCI (18/23), with focuses in AR/VR, accessibility, human-AI interaction, and AI ethics, among others. All participants were based in the United States. Most participants (19/23) used LLM-based applications at least weekly for general tasks, with 12 using them daily. Usage of LLMs for research varied, with 14 reporting extensive use of LLMs for research, 6 who used LLMs occasionally, and three who used them rarely or never. Additional participant details are available in Appendix C.1.1.

6.4.2 Task

We designed our study tasks to closely resemble the initial process of drafting an outline for organizing related literature, e.g., when writing a literature review or survey paper. Given our within-subjects study design, each participant used both system conditions, but with a different task, as repeating the same task would introduce learning effects. The two tasks were designed to be similar in difficulty and nature, allowing for valid comparisons of system performance across different tasks. For each task, we selected a survey paper in HCI [Lee et al., 2024a; Pang et al., 2025], extracted the top-level taxonomy explicitly defined in the paper, and sampled 50 papers from the surveyed literature. The collection size of 50 papers ensured consistency across tasks while balancing sufficient material for meaningful organization within the study timeframe. The selected taxonomy dimensions similarly balanced accessibility and depth across the two tasks: one dimension could be reasonably developed from abstracts (Task/Application Domains), while the other required deeper engagement with the full texts (Technology/Limitations & Risks). Participants were given the paper collection and survey topic, then asked to continue filling in the outline for the assigned sections (bolded). Their goal was to create meaningful subsections, save relevant information, and cite appropriate papers. Additional details are provided in Appendix C.1.2.

6.4.3 Baseline

In the baseline condition (BASELINE), participants used a system that displayed a searchable list of paper metadata—including title, authors, year, venue, abstract, and a PDF link. This design emulated the standard experience of browsing conference proceedings or results from an academic search engine. To reflect the growing integration of LLM-based tools in scholarly workflows [Liao et al., 2024], participants were en-

couraged to use ChatGPT,³ a commercial LLM chat application, to support their exploration.⁴ They could interact with the model via text prompts but were restricted from web search (since finding additional papers was beyond the scope of the task).

6.4.4 Procedure

The first author (study facilitator) conducted studies remotely with participants over Google Meet in March 2025. Each study lasted around 90 minutes. The study began with an introduction and overview of the two systems. Participants were first given a brief tutorial of DIMIND, and then allowed to use it to explore a sample paper collection involving “*interactive tools for scientific literature review*.” This familiarization period lasted up to ten minutes, during which the study facilitator answered any questions. Then, participants were asked to explore the same collection in the baseline interface. No tutorial was given for ChatGPT since all participants indicated in their screening survey some prior familiarity with LLM-based chat applications. Next, participants worked on the first task with the assigned system for 30 minutes, after which they completed a post-task survey. This process was repeated with the second task and assigned system. The pairing of task and system was fully counterbalanced across participants. After completing both tasks, any remaining time was used as a semi-structured interview in which participants elaborated on and compared their experience using both systems. The study facilitator also probed into any interesting usage behaviors participants exhibited during the study. Participants received a \$60 USD gift card upon completion of the study. This study was reviewed and exempted by our organization’s internal review board.

6.4.5 Measures and Analyses

For quantitative data, we collected participants’ post-task survey responses, including perceived system effectiveness, user control, information transparency, and overall satisfaction with the literature review process using each system (see Appendix C.1.3 for the specific questions). We also captured and analyzed interaction logs from the sessions to measure the quantity and types of actions participants took (e.g., create a user-defined column, add a system-suggested column, select or refine taxonomy categories, summarize).

³<https://chatgpt.com/>

⁴During the study, participants were provided with login credentials for a ChatGPT account created exclusively for this user study. This served two purposes: to standardize the version of ChatGPT each participant used and to allow us to export detailed logs of participants’ prompts and LLM responses.

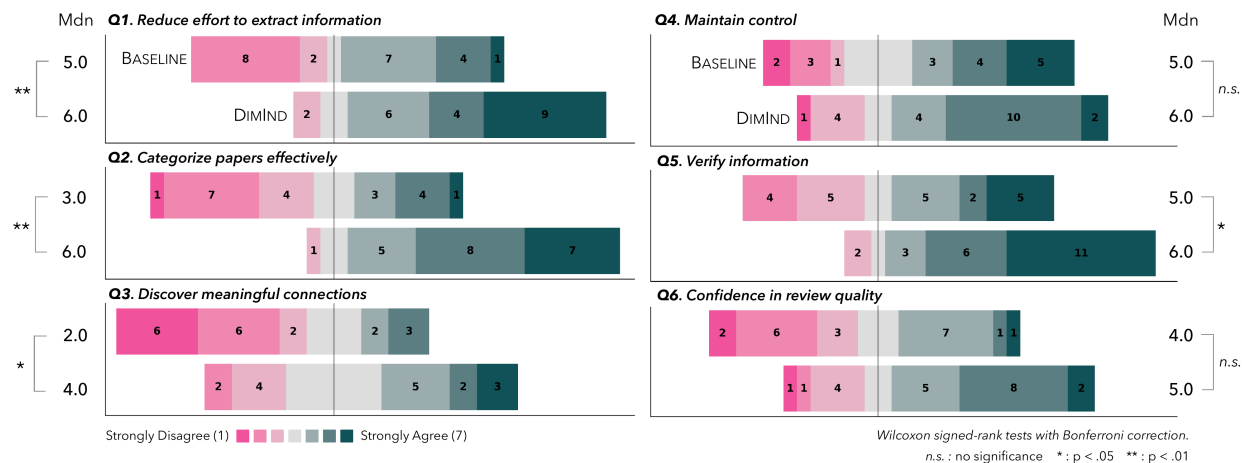


Figure 6.7: Participants' post-task ratings (7-point Likert scale) of literature review utility, control, information verifiability, and confidence across DIMIND and BASELINE.

We did not analyze the content of the outlines participants created during their tasks due to the significant diversity in form and content, complicating an unbiased expert evaluation. Instead, we treated the outline creation as a contextual activity that helped participants engage with the task while using each system, and relied on self-reporting and quantitative measures in the behavioral logs to examine the effectiveness of the two systems. Finally, participants' exit interviews were recorded and automatically transcribed. To analyze all paired Likert-scale data, we conducted Wilcoxon signed-rank tests with Bonferroni corrections for multiple comparisons. For qualitative data, we conducted a reflexive thematic analysis [Braun and Clarke, 2006], where the first author developed and iteratively refined codes through discussions with the research team to identify emerging themes.

6.4.6 Results

In this section, we present our quantitative and qualitative findings, organized by our two research questions. Participants are referred to with pseudonyms P1–P23. Quotes were lightly edited for brevity and clarity. We denote the median rating of BASELINE and DIMIND as M_B and M_D , respectively, and the Wilcoxon test statistic as W . Statistical significance was determined at $p < .05$.

How effectively does DIMIND support deeply exploring large collections of papers? (RQ1)

Tables with extracted information effectively alleviate foraging costs but risk information overload.

Based on the post-task ratings, we found participants appreciated the ability to quickly extract information from many papers at once through the creation of custom and system-suggested columns in DIMIND, reporting less perceived mental effort in extracting and organizing relevant information ($W = 21.5$, $p < .01^{**}$, $r = 0.74$) with DIMIND ($M_D = 6$, $IQR = 5 - 7$) than with BASELINE ($M_B = 5$, $IQR = 2 - 5$) (Fig. 6.7 – Q1). While many participants were able to eventually extract similar information with BASELINE with much more effort, they appreciated DIMIND’s support for easily and explicitly creating custom columns that matched the specific questions they had, and to have all the information organized together within the familiar table representation (P2, P13, P14, P18, P21). Contrastly, without careful prompting the baseline system often generated lots of information to mixed utility. More specifically, when using the baseline system, some participants felt they had to sift through information that “*tended to be a bit overly fluffy*” (P15), even to the point of irritation as P19 expressed: “*These lines are just kind of making me mad. . . ‘AI can act like the third teammate’, it doesn’t really tell me anything.*”

While participants appreciated being able to quickly extract relevant information across large collections of documents efficiently, we also observed a different flavor of information overload in DIMIND. Even when the information presented by the system was seen more useful, presenting it all at once without the users having to do much work can be overwhelming. P19 expressed ambivalent feelings, finding the table “*very useful to cut down the manual labor*” but also how “*having everything ready for me all at once can be overwhelming, since whenever I’m doing this it’s usually one row at a time.*” When reviewing large collections, even a relevant snippet for each paper felt unapproachable to use in drawing connections and organizing ideas across papers, as P4 described, “*it’s too detailed for me to organize it just by reading through these cells.*” Rather than directly interacting with the extracted information, P13 viewed the extracted information more as an auxiliary “*database to refer back to with all the columns and information*” when writing their literature review, and only then selectively investigating cells for papers they intended to discuss.

Instead of trying to consume all the information from the system at once, most participants using the LLM-generated information as *information scent* [Pirolli and Card, 1999], where keywords and categorizations served as navigational aids rather than as content directly integrated into a review. As P2 explained,

“these keywords are good enough for me to quickly skim through the PDF and find where it might lie,” suggesting that the generated information helped guide attention to relevant evidence within papers. For some, these cues served primarily as navigation—further interpretation would necessitate reading “*directly from the author’s mouth instead of a paraphrased version*” (P2).

Overall, these “*jumping-off points*” (P15) in the extracted evidence snippets facilitated efficient exploration, directing participants to relevant papers while enabling them to browse extracted information for the same facets across multiple papers simultaneously. For example, while P5 was examining a *Research Focus* column they had added to the table, they came across one snippet that mentioned “potential and risks of LLMs.” To see more context, they clicked to open the paper’s PDF, and the system scrolled to and highlighted the first paragraph of the discussion. They then briefly continued to browse snippets for other papers in the table, before directing their attention back to reader to skim the discussion section of the paper they had previously opened.

Taxonomies offer valuable cognitive scaffolding for top-down, thematic exploration. While participants leveraged the table structure as *information scent* for the underlying documents, most looked to the taxonomy structure in DIMIND for supporting higher-level synthesis. Overall, participants rated DIMIND ($M_D = 6$, $IQR = 5 - 7$) higher than BASELINE ($M_B = 5$, $IQR = 2 - 5$) for its ability to support paper categorization ($W = 0$, $p < .01^{**}$, $r = 0.88$) (Fig. 6.7 – Q2).⁵ Many participants attributed this advantage to DIMIND’s taxonomy representation clearly organizing emergent themes and relevant papers into an interactive hierarchical tree. P23 shared how the transformation from faceted column to taxonomy provided an efficient way to summarize themes across many papers for a facet:

“The tree structure was unexpectedly useful as a way to view themes. . . With [Baseline], you would have to compile relevant abstracts, titles, and metadata yourself before feeding it into the LLM. With the table [referring to DIMIND as a whole], it was a seamless way to direct information into the models and get thematic summaries.”

During exploration, the taxonomy structured representations served as external cognitive scaffolds, allowing participants to maintain multiple parallel organizing schemes at once. For example, P13, who had created a

⁵ $W = 0$ indicates all participants rated DIMIND’s support for paper categorization greater than or equal to that of BASELINE.

table facet about *ethical considerations*, and compared its cell values that included phrases such as *potential misuse of opinionated language models* and *perpetuation of societal biases* and the taxonomy tree that included higher-level themes such as *Bias*, *Transparency*, and *Societal Harms*, explained: “*I like how the tree structure gave a very intuitive way of grouping things together, and the way each tab[table facet] has a different tree structure. So it’s a different way of organizing your thoughts.*” This externalization of thought across different ‘tabs’ each with a unique taxonomy further serves to reduce the cognitive load of multi-faceted analysis.

However, while the taxonomy offered a rapid thematic overview, participants generally felt both BASELINE ($M_B = 2$, $IQR = 1.5 - 4$) and DIMIND ($M_D = 4$, $IQR = 3.5 - 5$) fell short of helping them discover very specific and meaningful connections (Fig. 6.7 – Q3), ones they would have been unable to discover without LLM assistance ($W = 15.5$, $p < .05^*$, $r = 0.78$). Combined with the length of the study, we found participants exhibited mixed levels of confidence in the quality and comprehensiveness of the outline they created (Fig. 6.7 – Q6); while participants reported slightly higher confidence in their outline using DIMIND ($M_D = 5$, $IQR = 3.5 - 6$) than BASELINE ($M_B = 4$, $IQR = 2 - 5$), this difference was not significant ($W = 34.5$, $p = .08$, $r = 0.66$). This is not surprising—while results are promising that DIMIND is able to significantly increase efficiency when deeply exploring a large collection of papers, the task of comprehensive literature reviews would still require significant time and effort beyond the scope of our lab study.

In sum, current models still fall short in generating deep intellectual connections between documents, highlighting the continued importance of a researcher-driven analytical process. While our findings are promising with regards to efficiency, designing for longer-term effects on confidence and the impact of prolonged usage remain directions for future work. At the same time, while most prior work focused on support literature understanding using a single representation, our results point to benefits of providing different structured representation, and the importance of allowing users to fluidly “zoom” between different levels of compression, allowing DIMIND to scale literature understanding support to large collections of papers.

Participants desired a blend of structured representations and conversation. While structured representations in DIMIND provided valuable scaffolding and more predictable interaction outcomes for the

literature review task, participants also valued the adaptability and flexibility offered by conversational interaction in BASELINE. We found no significant difference ($W = 65.5, p = 1, r = 0.78$) between DIMIND ($M_D = 6, IQR = 4 - 6$) and BASELINE ($M_B = 5, IQR = 3.5 - 6$) with respect to the ability to maintain control over the literature review process when aided by LLM assistance (Fig. 6.7 – Q4). Participants saw advantages and limitations of both interaction paradigms, and instead envisioned an ideal system that would blend structural organization with conversational control. For instance, participants appreciated the ability to trace information across structured representations, specifically drilling down into the specific evidence within a paper’s PDF and referring back to the paper’s abstract within the table and synthesis. As a result, participants reported finding it slightly easier ($W = 10.5, p < .05^*, r = 0.81$) to verify system-generated information when desired using DIMIND ($M_D = 6, IQR = 5.5 - 7$) than BASELINE ($M_B = 5, IQR = 3 - 6$) (Fig. 6.7 – Q5). At the same time, they also suggested several ways in which DIMIND could benefit from conversation, for example, to follow up with clarifying questions (P17), request explanations for certain decisions (e.g., why certain categories were created in the taxonomy and not others) (P15), explore relationships between specific papers (P23), or modify output formats and detail levels (P9).

How do researchers use and navigate across DIMIND’s structured representations? (RQ2)

Balanced use of user-defined and system-suggested facets for creating literature review tables. To transform the unstructured paper collection into a literature review table, most participants (14/23) took a balanced approach in adding both user-defined and system-suggested columns (6 added user-defined columns only; 3 added system-suggested columns only). Participants particularly valued the ability to customize the faceted columns based on their specific criteria. As P2 explained, “*I really like that I could customize what the criteria for generating an outline could be in that system. Because otherwise just asking an LLM to prompt—it defines its own criteria which might not be how researchers might approach writing [about] the literature.*” This flexible extraction from many papers at once helped reduce cognitive load, with P6 noting that after defining columns at a high level, their “*job is reduced down to only checking whether the information is correct.*” These results further showed the importance of supporting user-driven exploration, and that LLM generated structures might not always cover the nuanced and personal information needs of different users.

Facet taxonomies serve as central hubs for exploration. Facet taxonomies emerged as central navigational hubs during exploration. Participants saw them as DIMIND’s most useful information representation. The majority of participants primarily interacted with the taxonomy during their tasks, expanding and reviewing generated categories and used the taxonomy as *thematic filters* for papers in the table.

Most participants (19/23) used the taxonomy at least once, with participants expanding and collapsing categories an average of 18.0 times (SD=17.1). Similarly, most participants (20/23) selected categories in the taxonomy, using it to focus their attention on specific paper subsets, an average of 16.3 times (SD=18.1). As P4 noted, the tree structure helped overcome the challenge that specific paper evidence in the table representation could be “*too detailed for me to organize just by reading through,*” adding that “*after filtering through [the taxonomy] then it made more sense for me to look at what appeared.*” This behavior shows how participants used taxonomies to transform overwhelming tables into manageable views by first selecting relevant categories before examining filtered paper evidence to understand more deeply each generated category, and to confirm their validity.

Moreover, P23 described how DIMIND’s integrated workflow from specific faceted evidence to taxonomy addressed a challenge of traditional literature review fragmentation where researchers “*end up with 10 different documents of scattered information and a really clunky spreadsheet.*” She frequently navigated between taxonomy categories and specific papers by clicking connections between evidence in the taxonomy and table rows, summarizing:

“You’re able to scan through a lot of information, you’re able to expand things that you want to look at, and then most of the time it’s in this very compact form. I found myself using the tree the most, and I really like being able to switch back and forth between looking at the actual text, so the abstract or the whole paper, and then looking at the automatically generated information or summaries.”

Altogether, these findings suggest that the faceted taxonomies bring valuable high-level structure to literature review, effectively bridging overview and detailed exploration.

Transforming facet taxonomies to narrative summaries. About half of the participants transformed at least one taxonomy into a narrative summary using DIMIND (12/23). Five of these participants interacted

with the generated summaries, closely reading and inspecting more detail by clicking the inline citations (clicking these reveal a popover with metadata and the abstract for the cited paper) (P3, P6, P7, P9, P17). The other group of participants generated the summary to use as a reasonable artifact they could easily export into their task document.

Seven participants also tried to further improve the LLM-generated taxonomy, using those revisions to inform a more desirable summary. These participants iteratively refined the categories with the drag-and-drop interaction, often to group categories that the system had split but participants believed made more sense to be analyzed together. Some participants similarly expressed a desire to directly manipulate the taxonomy, but refrained from doing so given their time constraints. LLM-generated categories were not perfect, nor were they expected to be—when P20 noticed a misalignment between the LLM-generated taxonomy and her own understanding, e.g., “creative writing” nested under “creative” rather than “writing”, she described a simple repair:

“Sometimes it did classifications that didn’t make a ton of sense. With the writing one, where it separated creative writing from the rest of writing for some reason. But that felt a lot more natural. I can see that and be okay, I think of creative writing as writing. So in my own notes, let me just classify it like that.”

These findings suggest the utility of LLM-generated categories as an initial guide, with subsequent interactive, in-situ refinement or post-hoc refinement when exporting her understanding for presentation.

Despite engaging with the faceted table and taxonomies, participants were hesitant to delegate the final narrative synthesis to DIMIND (or other LLM tools) for their own work. This hesitation centered around a desire to preserve researcher agency, particularly in transitioning from the organizational schema of categories and relevant papers to a presentable artifact. In the notional model of sensemaking [Pirolli and Card, 2005], this suggests that most participants spent their time, and preferred LLM assistance, in the information extraction and schema refinement stages and less in the presentation stage. For instance, P20 explicitly preferred maintaining control over the final synthesis, saying *“I would probably not use the summary tool... That feels a little bit too close to writing my paper for me,”* while recognizing the utility of DIMIND in earlier stages.

Participants also shared concerns about automating what they saw as an essential scholarly process for

intellectual development: “*Literature review is to figure out what are all the keywords that you don’t know but convey the same meaning... having the automated process might accidentally limit some of the literature that you should know,*” and how over-automation with LLMs could “*kill the serendipity findings*” that emerge from personal exploration (P19). This suggests a nuanced relationship where researchers value LLM assistance for exploration and organization but still desire preservation of human authorship—a balance between LLM enhancement and maintaining scholarly agency and integrity.

6.5 Discussion

Literature review and synthesis at scale is a cognitively- and time-intensive activity, yet remains core to the research process. Recent advancements in AI—particularly LLM capabilities for processing and extracting information from long, complex documents—has led to growing research interest in leveraging these new technologies to support literature review.

While some prior work has sought to automate the entire process [Susnjak et al., 2025; Kasanishi et al., 2023; Altmami and Menai, 2022], we instead found participants often expressed a desire to retain a scholarly voice and steer the literature review process. Furthermore, user-driven processes can allow scholarly learning, as one participant noted how traditional approaches involving the manual construction of literature review tables can foster deeper engagement with the literature, helping to explore key ideas and possibly discover interesting connections. Similarly, some participants described appreciating the serendipity, curiosity, and intellectual development afforded by user-driven exploration in a literature review.

While participants found value in the generated table and taxonomy structures that provided information scent and broad overviews of many papers, they were more skeptical of the formulaic style of LLM-generated writing and highlighted ambiguity around intellectual ownership of the output, saying: “*Working with ChatGPT—yes, the dots are all connected for me, but it just feels very weird to say that this is mine.*” As a result, they were hesitant to save and share paragraphs directly generated by LLMs in both conditions, citing the potential for a loss of personal voice and unconscious plagiarism.

Compared to prior user-driven literature understanding systems that support exploring smaller sets of papers by only focusing on providing a single structured representation, such as paper pairs [Lee et al., 2024b; Murthy et al., 2022], tables [Elicit, 2023; Wang et al., 2024a; Newman et al., 2024] or taxonomies [Hsu

et al., 2024; Kang et al., 2022, 2023], our results instead suggest that providing multiple structured representations of paper information and allowing users to fluidly move between these different levels of information abstraction and compression can be effective in scaling a user-driven and LLM-assisted literature review process.

Most commonly, after adding a few columns of faceted information into the table, participants used a top-down review approach, structuring their exploration around the generated taxonomy. And while exploration of a facet often began in the taxonomy by reviewing the available concepts and their distribution in the collection, participants frequently navigated information at multiple levels of detail—for instance, interest in a particular taxonomy category could quickly transition into reading detailed evidence for the specific papers in that category within the table representation—to deepen and verify their understanding. At the same time, participants pointed to a gap between the table and taxonomy structures in DIMIND, where deeper connections between small subsets of papers were lacking. This suggests introducing an additional layer of structured representation that leverages prior work can potentially fill this gap. For example, leveraging PaperWeaver [Lee et al., 2024b] or ACCoRD [Murthy et al., 2022] to help users discover specific and nuanced connections between closely related pairs of papers in the table before looking at broader themes in the taxonomy structure.

We also found that leveraging LLMs to increase the efficiency of user-driven literature review highlighted a shift from interaction to cognitive costs, revealing several new design challenges. Specifically, given a particular facet of interest, the laborious manual process of reading and extracting relevant information from each individual paper can be greatly reduced with LLMs, yet the amount and rate at which information is presented to the user can become cognitively overwhelming. In DIMIND, we addressed this by presenting multiple levels of abstraction (tables, taxonomy, and narrative synthesis) as cognitive scaffolds, but the design space remains vast. What abstractions or interactive mechanisms best support sensemaking over large volumes of LLM-generated information without overwhelming or impeding serendipitous exploration. And how well do these designs transfer to other domains of information synthesis (e.g., medicine, clinical, or financial decision making) where AI is increasingly being studied and used [Duede et al., 2024; Liao et al., 2024]?

A final challenge involves designing to mitigate risks of LLM hallucination at scale. To this end, we

designed DIMIND with information transparency as a first-class goal, providing clear traces between transformations of paper information and interactive access to a paper’s abstract or full text to more easily verify generated information. Still, enabling verification at scale requires more accurate attribution techniques and more seamless interaction mechanisms. For example, DIMIND can build on recent advances in fine-grained attribution [Chuang et al., 2025; Zhang et al., 2024a; Slobodkin et al., 2024] to more effectively select and highlight relevant evidence when users drill down into the paper itself, or even proactively flag likely hallucinations for user review [Wadden et al., 2022a,b].

6.5.1 Limitations

We note several limitations of our evaluation that shape how our findings should be interpreted. Our within-subjects study was limited to 90-minute sessions to minimize participant fatigue, but literature review is inherently a complex and dynamic process that can take days or months to complete. Similarly, we set the number of papers to be explored in each task to be 50 papers sampled from real-world survey papers. While we believe this is a reasonable number balancing participant effort and scale, survey papers would typically reference more than 100 papers. At the same time, we believe this is a sufficient scale such that our findings may be generalizable to real-world literature review scenarios. We hope to continue to improve DIMIND based on participant feedback and conduct a deployment or longitudinal studies where researchers create and explore their own paper collections over extended periods. Such studies could reveal novel usage patterns and limitations of DIMIND, and provide deeper insights into the long-term effects of LLM-assisted cognitive scaffolding. Our evaluation also consisted primarily of CS researchers with an HCI focus; additional studies involving more diverse academic disciplines are necessary to establish broader generalizability and utility of LLM-assisted structures in literature review workflows.

Additionally, while all participants had prior experience with conversational LLM assistance (e.g., ChatGPT) for both everyday tasks and research purposes, DIMIND’s structured approach to LLM-enabled features presented a learning curve. Combined with familiarity and potential biases with conversational interfaces, these factors could have influenced participants’ expectations and perceptions of the reliability of ChatGPT and our system for the assigned tasks. Finally, our system’s interactive information traces with detailed paper information are only available when PDFs are openly accessible. Otherwise, DIMIND falls

back to paper abstracts, which lack sufficient detail for supporting the system’s overall goal of enabling the exploration and analysis of more nuanced paper facets.

6.5.2 Future Work

Based on our findings, we highlight several exciting directions for future work. We plan to release DIMIND as an open platform for researchers to further explore these structured representations and continue to improve the underlying LLM mechanisms for facet discovery, taxonomy creation, and synthesis.

Enhanced User Controls

Future work could introduce interaction mechanisms that improve steerability of LLM assistance. For instance, participants expressed interest in capabilities that would allow them to directly influence the system’s analysis, such as the ability to highlight specific papers or cells of evidence in the table to steer the generated taxonomy and synthesis (P18), hide papers from the analysis, directly edit or tag system-generated evidence (P11, P13, P18), and merge or split faceted columns (P9). These features could better shape the system’s output to align with researchers’ personalized needs.

Multi-Faceted Multi-Document Synthesis

Analysis support in DIMIND is currently limited to single facets using the taxonomy structure, but participants also expressed interest in analyzing relationships between multiple facets. For example, exploring interactions between table facets about *application domain* and *risks* could help discover risks that are understudied in a particular application domain. Future work could explore interactive 2D pivot visualizations—similar to [Suh et al., 2024] but where the two axes are themes—across two table facets to find correlations between facets, provide intelligent suggestions for which facets to analyze together, or introduce new interactions for simultaneously examining multiple facets in the table at once.

Fine-Grained Literature Discovery

The structured, faceted information in DIMIND’s literature review table present opportunities for more fine-grained approaches to literature discovery. For instance, users could explore papers that share similar values

to those found in a specific table cell, connect to broader conceptual clusters in the taxonomy, or introduce novel values that expand the current facet.

6.6 Conclusion

In this chapter, we presented DIMIND, an LLM-enabled system that scaffolds literature review by transforming unstructured paper content into navigable structured representations, with provenance of LLM-generated information to support paper-level verification. Our evaluation with 23 researchers demonstrated that DIMIND offers valuable cognitive support, particularly for extracting faceted information across papers and enabling top-down exploration via conceptual taxonomies. Our findings further suggest that these intermediate structured representations offer an effective way to leverage LLM assistance toward mediating the interaction between researchers and complex information within vast literature collections.

Chapter 7

Discussion and Conclusion

At the beginning of this dissertation, I set out to explore the following thesis: **We can improve how researchers gather and synthesize knowledge across scientific literature through the design of intermediate, interactive, and transparent representations that scaffold human-AI sensemaking.** I investigated this thesis through structured mediation—a design framework that creates interpretable representations operating through discovery, iteration, and verification mechanisms. Across three systems spanning individual paper comprehension (QLARIFY, SCIM) to collection-level synthesis (DIMIND), I demonstrated how AI can enhance scholarly reasoning while preserving researcher agency. Findings from lab and longitudinal evaluations examining researchers using these systems reveal both the potential of AI-augmented literature tools and critical questions about where and how AI should support scholarly knowledge work. Our analyses overall highlight tensions between automation efficiency and agency, many of which may extend beyond scholarly sensemaking to broader questions of how future human-AI systems should be designed to support and improve collaborative knowledge work in expertise-driven domains.

7.1 Evolving Models of Human-Information Interaction in the Age of AI

In information foraging theory, people navigate between static information patches using cues such as keywords, citations, and section headers to guide their exploration [Pirolli and Card, 1999]. Traditional models of sensemaking further characterize this as a process where people collect information, organize it into external representations, and iteratively build the mental models that inform their eventual decision-making [Rus-

sell et al., 1993; Pirolli and Card, 2005]. Often, these external representations manifest as supportive aids that facilitate but do not directly participate in the reasoning process, with human analysts (e.g., researchers) as the primary agents responsible for navigation, assessment, and synthesis.

Modern AI systems, e.g., those powered by LLMs, challenge these classical assumptions. The systems presented in this dissertation illustrate how AI reshapes both the structure of information and the nature of human engagement with it. For example, QLARIFY turns static abstracts into adaptive information patches that expand based on user queries grounded in a broader document context, SCIM surfaces AI-generated information scent that extends beyond surface-level textual cues to semantic understanding of content, and DIMIND generates entirely new information patches through representations like taxonomies and structured tables that would require substantial effort to produce manually. These systems shift the cost structure of information foraging through new discovery mechanisms and delegating routine organization and extraction tasks, allowing users to spend more effort on evaluation and synthesis.

This transformation also suggests a shift toward *distributed sensemaking*, where human and AI cognition operate in parallel across dynamic, externalized representations. Rather than serving passive roles as an automation tool, AI increasingly acts as a collaborative partner in the reasoning process—surfacing connections, restructuring information, and learning from and responding to evolving user goals. Future research can explore new theoretical frameworks for human–AI knowledge work that extend classical sensemaking models and draw on principles from human–AI interaction [Amershi et al., 2019; Lin et al., 2025] and mixed-initiative interfaces [Horvitz, 1999].

These shifts raise ethical design questions: how to preserve agency as AI assumes greater cognitive responsibility and how to prevent over-reliance or homogenization that narrow inquiry rather than expand it (e.g., mechanized convergence [Sarkar, 2024b] in AI-assisted work where homogenized outputs reduce idea diversity [Doshi and Hauser, 2024; Anderson et al., 2024]). Addressing epistemic agency, diversity of perspectives, and critical engagement is important to ensure AI assistance can expand rather than constrain scientific progress. The next two sections turn to design implications for ensuring AI systems preserve researcher agency and intellectual diversity while supporting efficiency.

7.2 Design Implications: Lessons from Structured Mediation

Guiding Attention Without Constraining Exploration Our findings suggest AI-driven discovery mechanisms should strive to guide attention without constraining exploration paths. For scholarly sensemaking, we found evidence that effective information scent often operates through suggestions rather than prescriptions. For example, QLARIFY’s expandable entities, SCIM’s faceted highlights, and DIMIND’s induced dimensions were often perceived as most effective when they provided “jumping-off points” for user-driven investigation rather than well-defined narratives.

A rich design space remains to explore how such discovery mechanisms may also become more personalized and context-aware, such as by learning from user interaction patterns. For instance, future iterations of QLARIFY and SCIM could personalize suggestions based on a user’s prior interactions such as previously explored entities or highlighted text, as well as patterns of collective usage across users reading the same paper. Another direction involves deepening the exploration of AI-driven discovery of latent dimensions that enable sensemaking and synthesis (beyond simple LLM prompting approaches explored in DIMIND) across vast information spaces, particularly for which manual surveying is infeasible.

More broadly, constructing these appropriate schemas is central to knowledge work; expert performance more often stems from well-developed, domain-specific structures for organizing information than from general ability [Pirolli and Card, 2005; Ericsson and Lehmann, 1996]. Future human-AI systems could aid knowledge work by automatically identifying and suggesting the most appropriate intermediate representations as cognitive scaffolds attuned to user goals and information context, narrowing the expertise gap between novices and experts.

Providing Mechanisms for Efficient Oversight Oversight of AI assistance, e.g., through verification in structured mediation, serves two important functions: calibrating reliance on AI and deepening understanding of the information content. However, as AI systems become more capable of synthesizing information across vast and disparate sources of knowledge (e.g., large paper collections), effective oversight may involve different verification and recovery strategies depending on task complexity and level of AI assistance. For instance, when AI generates abstract insights or identifies complex patterns, checking individual sources may not validate higher-level interpretations. For routine information tasks, manual verification could be

sufficient, e.g., QLARIFY and DIMIND’s deeply-linked attribution and SCIM’s in-context highlights allowed relatively easy validation of AI assistance. However, with increasingly agentic AI systems and the resulting reliance on AI for complex synthesis tasks, more sophisticated mechanisms for oversight may be needed to ensure alignment with user expectations, such as using multiple queries to cross-check results [Fok et al., 2024b; Zamfirescu-Pereira et al., 2023], providing alternative interpretations and cognitive functions that encourage critical verification [Bućinca et al., 2021], and decomposing multi-step reasoning processes [Wu et al., 2022a; Feng et al., 2025].

7.3 Designing for Metacognitive Support in AI-Assisted Knowledge Work

There is an inherent tension between reducing cognitive load through AI assistance and preserving the hands-on engagement needed for expertise development. Excessive automation risks deskilling, while insufficient support might leave researchers unnecessarily spending time on lower-level tasks rather than attending to activities where their expertise is more valuable. The ease of use and fluency of generative AI output may also promote overreliance, reducing the frequency of deeper cognitive engagement [Drosos et al., 2024; Prather et al., 2023]. Over time, this may contribute to the erosion of key analytical and synthesis skills [Sellen and Horvitz, 2024], highlighting a need to design systems that scaffold critical thinking rather than suppress it. For example, simple designs can prompt deeper reflection—in education research, metacognitive guiding questions (e.g., prompting students to reflect on their thinking) have been shown to improve critical engagement with texts [Salomon, 1988] and visual scaffolds for argumentation (e.g., juxtaposed, structured displays of competing evidence) can improve dialectical thinking [Tsai et al., 2015].

Our structured mediation framework offers one possible design perspective, through the use of intermediate representations that dynamically scaffold rather than replace scholarly reasoning processes. In QLARIFY, participants steered exploration through progressively disclosed, dynamic information artifacts, with many pursuing threaded, branching expansions. Similarly, in SCIM, participants integrated AI-suggested highlights into their existing reading strategies, using highlights as landmarks for potentially important information while still following their natural skimming patterns through papers. This suggests the potential for AI augmentation to guide attention without replacing researchers’ reading comprehension abilities.

In addition to intermediate representations, interaction via multi-turn drill-down conversations with AI

can also act as a form of metacognitive support. For example, in QLARIFY, participants often pursued threaded expansions that resembled conversational exploration, where successive interactions helped clarify evolving questions and surface overlooked connections. Such iterative dialogue not only improves foraging efficiency but also encourages users to test alternative frames and reflect on gaps in understanding, paralleling strategies like Socratic questioning [King, 1995].

However, DIMIND and the living syntheses study revealed more complex dynamics around expertise and automation. While participants appreciated AI assistance for information extraction and organization, they were skeptical of its ability for narrative synthesis, expressing concerns that automated writing felt too close to having their scholarly work done for them. Similarly, literature review authors welcomed AI support for routine updates like recalculating statistics or identifying new relevant papers, but were hesitant to delegate synthesis and narrative construction, viewing these as core scholarly activities that require human expertise and creativity.

Our findings point to two distinct challenges requiring different forms of metacognitive support [Tankelevitch et al., 2024]. For *experts*, deep domain knowledge can sometimes lead to cognitive entrenchment, where extensive experience creates fixed mental models that limit consideration of alternative approaches. Studies have suggested experts demonstrate specific strategies to avoid fixation—they may actively reflect on their own thought processes, seek out alternative perspectives through collaboration, or systematically analyze problems using structured methods to break free from established mental models [Crilly, 2015]. The structured representations in our systems—such as DIMIND’s taxonomies or QLARIFY’s expandable entities—demonstrate such metacognitive scaffolding by surfacing alternative organizational schemes or highlighting overlooked connections across literature. Experts particularly valued AI assistance serving as a “second opinion,” helping identify potential blind spots or biases in their review workflows.

For *novices*, the framework of deliberate practice suggests that expertise development requires sustained engagement with increasingly challenging tasks, coupled with immediate feedback and reflection opportunities [Ericsson, 2004]. This raises concerns about how automation might limit crucial learning processes. Participants using DIMIND noted that traditional literature review processes help researchers discover unfamiliar terminology and concepts—a learning process that could be compromised by excessive automation. Similarly, authors in the living syntheses study emphasized that the process of manually organizing and

synthesizing literature was itself valuable for intellectual development, noting that over-automation might “kill the serendipity findings” that emerge from personal exploration.

A future direction could involve moving beyond simple task automation toward systems that actively support metacognitive development. This suggests a shift in how we should conceptualize AI assistance: from reducing cognitive load to strategically engaging users in ways that promote expertise development while retaining the efficiency benefits of automation. The verification mechanisms across all three systems—QLARIFY’s attribution links, SCIM’s context preservation by design, and DIMIND’s multi-level provenance—offer one approach by facilitating active evaluation and tracing of AI-generated content back to attributed evidence in paper full texts.

Future human-AI systems could instead be designed to scaffold expertise development through guided practice rather than purely focus on reducing cognitive load. Fundamental models could be pedagogically steered towards using effective teaching strategies [Puech et al., 2024; Jurenka et al., 2024], and interactive systems could incorporate adaptive interfaces that gradually increase user autonomy as expertise develops over time, while occasionally prompting for reflection on decision-making processes, reflecting active learning practices in student education [Freeman et al., 2014; Merrill et al., 1992]. Integrating contextual explanations could also allow users to understand not just what information is relevant, but *why*, and enable structured comparison of user analyses with AI-generated alternatives to promote deeper understanding.

Some have further proposed reframing AI into roles that challenge user assumptions and prompt reevaluation of both AI- and user-generated content (e.g., provocateur [Sarkar, 2024a] or antagonist [Cai et al., 2024]). While such designs depart from the conventional goal of seamless, frictionless user experiences, they offer one possible strategy for encouraging deeper reflection in human–AI knowledge work. Importantly, these forms of “productive friction” should be used sparingly and only under the conviction that the AI is sufficiently performant (i.e., in most cases, significantly more accurate than the human); otherwise, they may risk frustrating users or eroding trust, leading to eventual disuse of the AI assistance altogether. Their utility should therefore be validated in context before integration into real-world scholarly support. That said, metacognition offers one possible design strategy that future systems could explore to improve the efficiency of existing knowledge work processes, enhance rather than diminish the development of scholarly expertise, and preserve researchers’ ownership over more desirable aspects of scholarly knowledge work,

such as creative synthesis, narrative construction, and serendipitous learning.

7.4 Concluding Remarks

In this dissertation, I showed how AI can augment scholarly sensemaking through a structured mediation framework. Across three systems and an empirical study of survey authors, we saw studies of how AI can effectively enhance comprehension, improve efficiency, and support synthesis while preserving information provenance and researcher control. A core motivation for this work is to help reframe how we view AI in scholarly work: not as a replacement for human reasoning, but as a proactive and collaborative tool with which cognitive responsibility is shared.

At the same time, the development and adoption of such tools raise broader ethical and social considerations. Just as ranking algorithms in scholarly search engines can inadvertently reinforce already well-cited work (i.e., the rich-get-richer effect), AI-mediated representations may similarly accelerate convergence around dominant framings or biased, AI-driven narratives. Tools that reduce effort also risk cognitive deskilling, shifting scholars toward an unhealthy reliance on AI rather than cultivating their own expertise through deeper engagement (e.g., conducting systematic literature reviews without AI assistance). These dynamics highlight how design choices across tools such as the ones explored in this dissertation that influence individual users could also collectively reshape the epistemic fabric of science. The future of AI-assisted scholarship holds tremendous potential, but realizing it requires not only technical innovation but also a commitment to aligning the AI systems we build with the enduring values of scholarship: fostering expertise, sustaining creativity, and ensuring meaningful human participation in the production of knowledge.

Bibliography

Science | AAAS. 2016. How to keep up with the scientific literature. Accessed: 2025-06-30.

Mitchell Abrams and Matthias Scheutz. 2022. Social norms guide reference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–11, Seattle, WA, USA. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Elie A Akl, Joerg J Meerpohl, Julian Elliott, Lara A Kahale, Holger J Schünemann, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, and 1 others. 2017. Living systematic reviews: 4. living guideline recommendations. *Journal of clinical epidemiology*, 91:47–53.

Noof Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1011–1028.

Saleema Amershi, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, Eric Horvitz, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, and Paul N. Bennett. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Glasgow, Scotland, UK. Association for Computing Machinery.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle

- Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, and 4 others. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans, LA, USA. Association for Computational Linguistics.
- Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&c '24, page 413–425, New York, NY, USA. Association for Computing Machinery.
- Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James D. Rath, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.
- Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact.*, 30(5).
- Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmquist. 2019. Elastic documents: Coupling text and tables through contextual visualizations for enhanced document reading. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):661–671.
- Michelle Q. Wang Baldonado and Terry Winograd. 1997. Sensemaker: an information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 11–18, Atlanta Georgia USA. Acm.
- Charles Bazerman. 1985. Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication - WRIT COMMUN*, 2:3–23.
- Jöran Beel and Bela Gipp. 2009. Google scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics*, volume 1, pages 230–241, Rio de Janeiro, Brazil. International Society for Scientometrics and Informetrics.

- Catarina G. Belem, Pouya Pezeskhpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2024. From single to multi: How llms hallucinate in multi-document summarization. *Preprint*, arXiv:2410.13961.
- Mark Bernstein. 2009. On hypertext narrative. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 5–14, Torino, Italy. Acm.
- Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool Publishers, San Rafael, CA, USA.
- Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint*. ArXiv:2212.08037 [cs].
- Francisco Bolanos, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2024. Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57(10):259.
- Rohit Borah, Andrew W Brown, Patrice L Capers, and Kathryn A Kaiser. 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. *BMJ open*, 7(2):e012545.
- 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):2215–2222.
- Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1.
- Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,

- Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.
- Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5(Cscw1).
- George Buchanan and Tom Owen. 2008. Improving skim reading for document triage. In *Proceedings of the Second International Symposium on Information Interaction in Context, IiiX '08*, pages 83–88, New York, NY, USA. Association for Computing Machinery.
- Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. Tldr: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.
- Alice Cai, Ian Arawjo, and Elena L. Glassman. 2024. Antagonistic ai. *Preprint*, arXiv:2402.07350.
- Patricia L. Carrell. 1985. Facilitating esl reading by teaching text structure. *TESOL Quarterly*, 19(4):727–752.
- Scott Carter, Elizabeth Churchill, Laurent Denoue, Jonathan Helfman, and Les Nelson. 2004. Digital graffiti: Public annotation of multimedia content. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems, CHI Ea '04*, pages 1207–1210, New York, NY, USA. Association for Computing Machinery.

- 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):1–21.
- Jackie Chandler and Sally Hopewell. 2013a. Cochrane methods - twenty years experience in developing systematic review methods. *Systematic reviews*, 2:1–6.
- Jackie Chandler and Sally Hopewell. 2013b. Cochrane methods-twenty years experience in developing systematic review methods. *Systematic reviews*, 2:1–6.
- Bay-Wei Chang, Jock D. Mackinlay, Polle T. Zellweger, and Takeo Igarashi. 1998. A negotiation architecture for fluid documents. In *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, UIST '98, pages 123–132, New York, NY, USA. Association for Computing Machinery.
- Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding comparison tables for online decision making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 391–405, Virtual Event USA. Acm.
- Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. Searchlens: composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 498–509, Marina del Ray California. Acm.
- Joseph Chee Chang, Amy X. Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S. Weld. 2023. Citesee: Augmenting citations in scientific papers with persistent and personalized historical context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Chi '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.
- ChatDOC. 2023. Chatdoc - chat with your documents.
- ChatPDF. 2023. Chatpdf - chat with any pdf.
- Der-Thanq “Victor” Chen, Yu-Mei Wang, and Wei Ching Lee and. 2016. Challenges confronting beginning researchers in conducting literature reviews. *Studies in Continuing Education*, 38(1):47–60.

- Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, Iui '17, pages 17–28, New York, NY, USA. Association for Computing Machinery.
- Li Chen, Feng Wang, Luole Qi, and Fengfeng Liang. 2014. Experiment on sentiment embedded comparison interface. *Knowledge-Based Systems*, 64:44–58.
- Ed H Chi, Michelle Gumbrecht, and Lichan Hong. 2007. Visual foraging of highlighted text: An eye-tracking study. In *Proceedings of the 12th International Conference on Human-Computer Interaction*, pages 589–598, Beijing, China. Springer.
- Ed H. Chi, Lichan Hong, Michelle Gumbrecht, and Stuart K. Card. 2005. Scenthighlights: highlighting conceptually-related sentences during reading. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, page 272, San Diego, CA, USA. Association for Computing Machinery.
- Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.
- Johan S. G. Chu and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118. Publisher: Proceedings of the National Academy of Sciences.
- Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen tau Yih. 2025. Selfcite: Self-supervised alignment for context attribution in large language models. *Preprint*, arXiv:2502.09604.
- Jennifer Chubb, Peter Cowling, and Darren Reed. 2022. Speeding up to keep up: exploring the use of ai in the research process. *AI & society*, 37(4):1439–1457.
- Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S. Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3691–3697, Online. Association for Computational Linguistics.

- Harris Cooper. 2015. *Research synthesis and meta-analysis: A step-by-step approach*, 5 edition, volume 2 of *Applied Social Research Methods Series*. Sage Publications, Thousand Oaks, CA.
- John W Creswell and Cheryl N Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications, Usa.
- Nathan Crilly. 2015. Fixation and creativity in concept development: The attitudes and practices of expert designers. *Design studies*, 38:54–91.
- Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Sigir '92, pages 318–329, New York, NY, USA. Association for Computing Machinery.
- Ben Daniel. 2022. Common challenges postgraduate students and early-career academics face when engaging with the scholarly literature. *Electronic Journal of Business Research Methods*, 20(3):142–152.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2024. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062.
- Mira Dontcheva, Steven M. Drucker, David Salesin, and Michael F. Cohen. 2007. Relations, cards, and search templates: User-guided web data integration and layout. In *Proceedings of the 20th Annual ACM*

- Symposium on User Interface Software and Technology*, UIST '07, page 61–70, New York, NY, USA. Association for Computing Machinery.
- Anil R Doshi and Oliver P Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290.
- Ian Drosos, Advait Sarkar, Xiaotong Xu, Carina Negreanu, Sean Rintel, and Lev Tankelevitch. 2024. "it's like a rubber duck that talks back": Understanding generative ai-assisted data analysis workflows through a participatory prompting study. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, Chiwork '24, New York, NY, USA. Association for Computing Machinery.
- Eamon Duede, William Dolan, André Bauer, Ian Foster, and Karim Lakhani. 2024. Oil & water? diffusion of ai within and across scientific fields. *Preprint*, arXiv:2405.15828.
- Geoffrey Duggan and Stephen Payne. 2009. Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of experimental psychology. Applied*, 15:228–242.
- Geoffrey B. Duggan and Stephen J. Payne. 2011. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, page 1141, Vancouver, BC, Canada. Association for Computing Machinery.
- Elicit. 2023. Elicit: The ai research assistant.
- Julian H Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, and 1 others. 2017. Living systematic review: 1. introduction—the why, what, when, and how. *Journal of clinical epidemiology*, 91:23–30.
- Julian H Elliott, Tari Turner, Ornella Clavisi, James Thomas, Julian PT Higgins, Chris Mavergames, and Russell L Gruen. 2014. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS medicine*, 11(2):e1001603.
- Martin J Eppler and Jeanne Mengis. 2004. The concept of information overload: A review of literature

- from organization science, accounting, marketing, mis, and related disciplines. *The information society*, 20(5):325–344.
- Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.
- K Anders Ericsson. 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic medicine*, 79(10):S70–s81.
- K Anders Ericsson and Andreas C Lehmann. 1996. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, 47(1):273–305.
- K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S. Weld, Amy X. Zhang, and Joseph Chee Chang. 2025. Cocoa: Co-planning and co-execution with ai agents. *Preprint*, arXiv:2412.10999.
- Jessica L. Feuston and Jed R. Brubaker. 2021. Putting tools in their place: The role of time and perspective in human-ai collaboration for qualitative analysis. *Proc. ACM Hum.-Comput. Interact.*, 5(Cscw2).
- Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024a. Qlarify: Recursively expandable abstracts for dynamic information retrieval over scientific papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.
- Raymond Fok, Joseph Chee Chang, Marissa Radensky, Pao Siangliulue, Jonathan Bragg, Amy X. Zhang, and Daniel S. Weld. 2025a. Facets, taxonomies, and syntheses: Navigating structured representations in llm-assisted literature review. *Preprint*, arXiv:2504.18496.

Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, Iui '23, pages 476–490, New York, NY, USA. Association for Computing Machinery.

Raymond Fok, Nedim Lipka, Tong Sun, and Alexa F Siu. 2024b. Marco: Supporting business document workflows via collection-centric information foraging with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. Association for Computing Machinery.

Raymond Fok, Alexa F. Siu, and Daniel S. Weld. 2025b. Toward living narrative reviews: An empirical study of the processes and challenges in updating survey articles in computing research. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Chi '25, New York, NY, USA. Association for Computing Machinery.

Raymond Fok, Luca Soldaini, Cassidy Trier, Erin Bransom, Kelsey MacMillan, Evie Cheng, Hita Kambhamettu, Jonathan Bragg, Kyle Lo, Marti A. Hearst, Andrew Head, and Daniel S. Weld. 2024c. Accelerating scientific paper skimming with augmented intelligence through customizable faceted highlights. *ACM Trans. Interact. Intell. Syst.*, 14(4).

Allen Institute for AI. 2022. pdf-component-library.

Allen Institute for AI. 2023. Multimodal document analysis (mmda).

Allen Institute for Artificial Intelligence. 2022. Mmda - multimodal document analysis.

National Science Foundation. 2021. Publications output: U.s. trends and international comparisons.

Robert L. Fowler and Anne S. Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, 59(3):358–364.

Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415.

- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.
- Charlie George and Andreas Stuhlmüller. 2023. Factored verification: Detecting and reducing hallucination in summaries of academic papers. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 107–116, Bali, Indonesia. Association for Computational Linguistics.
- Samira Ghodrathnama, Mehrdad Zakershahraei, and Fariborz Sobhanmanesh. 2021. Adaptive summaries: A personalized concept-based summarization approach by learning from users’ feedback. In *Service-Oriented Computing – ICSOC 2020 Workshops*, pages 281–293, Cham. Springer International Publishing.
- Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A spatially anchored census of student confusion for online lecture videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Chi ’15*, pages 1555–1564, New York, NY, USA. Association for Computing Machinery.
- Jamey Graham. 1999. The reader’s helper: A personalized document reading environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Chi ’99*, page 481–488, New York, NY, USA. Association for Computing Machinery.
- Darcy Haag Granello. 2001. Promoting cognitive complexity in graduate written work: Using bloom’s taxonomy as a pedagogical tool to improve literature reviews. *Counselor Education and Supervision*, 40(4):292–307.
- Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108.
- Grobid. 2008–2025. Grobid. <https://github.com/kermitt2/grobid>. *Preprint*, swb:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

- Tovi Grossman, Fanny Chevalier, and Rubaiat Habib Kazi. 2015. Your paper is dead!: Bringing life to research articles with animated figures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI Ea '15, pages 461–475, Seoul, Republic of Korea. Association for Computing Machinery.
- Nathan Hahn, Joseph Chee Chang, and Aniket Kittur. 2018. Bento browser: Complex mobile search without tabs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Chi '18, pages 1–12, New York, NY, USA. Association for Computing Machinery.
- Han L. Han, Junhang Yu, Raphael Bournet, Alexandre Ciorascu, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2022. Passages: Interacting with text across documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi '22, New York, NY, USA. Association for Computing Machinery.
- Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. 2024. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843.
- Hayato Hashimoto, Kazutoshi Shinoda, Hikaru Yokono, and Akiko Aizawa. 2017. Automatic generation of review matrices as multi-document summarization of scientific papers. In *Birndl @ Sigir*, pages 69–82, New York, NY, USA. Association for Computing Machinery.
- Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Chi '21, pages 1–18, New York, NY, USA. Association for Computing Machinery.
- Andrew Head, Amber Xie, and Marti A. Hearst. 2022. Math augmentation: How authors enhance the readability of formulas using novel visual design practices. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi '22, New York, NY, USA. Association for Computing Machinery.
- Marti A. Hearst and Duane Degler. 2013. Sewing the seams of sensemaking: A practical interface for tagging and organizing saved search results. In *Proceedings of the Symposium on Human-Computer Interac-*

- tion and Information Retrieval*, Hcir '13, pages 1–10, New York, NY, USA. Association for Computing Machinery.
- Marti A. Hearst and Emilia Stoica. 2009. Nlp support for faceted navigation in scholarly collections. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries - NLP4DL '09*, page 62, Suntec, Singapore. Association for Computational Linguistics.
- Leonie Heron, Diana Buitrago-Garcia, Aziz Mert Ipekci, Rico Baumann, Hira Imeri, Georgia Salanti, Michel Jacques Counotte, and Nicola Low. 2023. How to update a living systematic review and keep it alive during a pandemic: a practical guide. *Systematic Reviews*, 12(1):156.
- Julian P.T. Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch. 2019. *Cochrane handbook for systematic reviews of interventions*. wiley, Chichester (UK). Publisher Copyright: © 2019 The Cochrane Collaboration.
- William C. Hill, James D. Hollan, Dave Wroblewski, and Tim McCandless. 1992. Edit wear and read wear. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '92, pages 3–9, New York, NY, USA. Association for Computing Machinery.
- Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, pages 381–388, Hyderabad, India. Association for Computational Linguistics.
- Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '99, page 159–166, New York, NY, USA. Association for Computing Machinery.
- Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Wang, and Aakanksha Naik. 2024. CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 118–132, Bangkok, Thailand. Association for Computational Linguistics.

- Roderick Hunt. 1995. The subtlety of distinctiveness: What von restorff really did. *Psychonomic Bulletin & Review*, 2:105–112.
- Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the con in context: Identifying deceptive actors in the game of mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–168, Seattle, WA, USA. Association for Computational Linguistics.
- Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1011–1028.
- Edward W. Ishak and Steven K. Feiner. 2006. Content-aware scrolling. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, page 155, Montreux, Switzerland. Association for Computing Machinery.
- Emi Ishita, Yasuko Hagiwara, Yukiko Watanabe, and Yoichi Tomiura. 2018. Which parts of search results do researchers check when selecting academic documents? In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, Jcdl '18*, pages 345–346, New York, NY, USA. Association for Computing Machinery.
- Farnaz Jahanbakhsh, Elnaz Nouri, Robert Sim, Ryen W. White, and Adam Fourney. 2022. Understanding questions that arise when working with business documents. *Proceedings of the ACM on Human-Computer Interaction*, 6(Cscw2):341:1–341:24.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proc. ACM Hum.-Comput. Interact.*, 5(Cscw1).
- Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned publishing*, 23(3):258–263.

- Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.
- 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*, UIST '22, pages 1–15, New York, NY, USA. Association for Computing Machinery.
- Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–19, New York, NY, USA. Association for Computing Machinery.
- Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. SciReviewGen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6695–6715, Toronto, Canada. Association for Computational Linguistics.
- S. Keshav. 2007. How to read a paper. *Comput. Commun. Rev.*, 37:83–84.
- Hanan Khalil, Lotfi Tamara, Gabriel Rada, and Elie A. Akl. 2022. Challenges of evidence synthesis during the 2020 covid pandemic: a scoping review. *Journal of Clinical Epidemiology*, 142:10–18.
- Alaa Al Khourdajie. 2024. The role of artificial intelligence tools in climate change scientific assessments.
- Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018a. Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, UIST '18, page 423–434, New York, NY, USA. Association for Computing Machinery.
- Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018b. Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pages 423–434, Berlin, Germany. Association for Computing Machinery.

- Juho Kim, Amy X. Zhang, Jihee Kim, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Content-aware kinetic scrolling for supporting web page navigation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, pages 123–127, Honolulu, HI, USA. Association for Computing Machinery.
- Tae Soo Kim, Matt Latzke, Jonathan Bragg, Amy X. Zhang, and Joseph Chee Chang. 2023. Papeos: Augmenting research papers with talk videos. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.
- Alison King. 1995. Designing the instructional process to enhance critical thinking across the curriculum. *Teaching of Psychology*, 22(1):13–17.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint*.
- Aniket Kittur, Andrew M. Peters, Abdigani Diriye, Trupti Telang, and Michael R. Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2989–2998, Paris France. Acm.
- Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sense-making. In *Expertise out of context*, pages 118–160. Psychology Press.
- Ian A. Knight, Max L. Wilson, David F. Brailsford, and Natasa Milic-Frayling. 2019. Enslaved to the trapped data: A cognitive work analysis of medical systematic reviews. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, Chiir '19, page 203–212, New York, NY, USA. Association for Computing Machinery.
- Jeffrey W. Knopf. 2006. Doing a literature review. *PS: Political Science and Politics*, 39(1):127–132.
- Wei-Jen Ko, Te-yuan Chen, Yiyang Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6544–6555, Online. Association for Computational Linguistics.

- 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):154:1–154:35.
- Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting references between text and charts via crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 31–40, Toronto, Ontario, Canada. Association for Computing Machinery.
- Kayvan Kousha and Mike Thelwall. 2024. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*, 37(1):4–12.
- Chinmay Kulkarni and Ed Chi. 2013. All the news that’s fit to read: A study of social annotations for news reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi ’13, page 2407–2416, New York, NY, USA. Association for Computing Machinery.
- Vaibhav Kumar and Alan W Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.
- Natasha Lacroix. 1999. Macrostructure construction and organization in the processing of multiple text passages. *Instructional Science*, 27(3/4):221–233.
- Esther Landhuis. 2016. Scientific literature: Information overload. *Nature*, 535(7612):457–458.
- Toby J. Lasserson, Julian P.T. Higgins, and Jackie Chandler. 2019. Starting a review. In Julian P.T. Higgins, Jackie Chandler, Miranda Cumpston, Tianjing Li, Matthew J. Page, and Vivian A. Welch, editors, *Cochrane Handbook for Systematic Reviews of Interventions*, pages 1–12. Wiley, Uk.
- Byungjoo Lee, Olli Savisaari, and Antti Oulasvirta. 2016. Spotlights: Attention-optimized highlights for skim reading. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Chi ’16, pages 5203–5214, San Jose, CA, USA. Association for Computing Machinery.
- Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim,

- Simon Knight, Seyed Parsa Neshaei, and 17 others. 2024a. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. Association for Computing Machinery.
- Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024b. 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*, Chi '24, New York, NY, USA. Association for Computing Machinery.
- Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, WA, USA. Association for Computational Linguistics.
- Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.
- Maria Liakata, Simone Teufel, Advait Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2054–2061, Valletta, Malta. European Language Resources Association (ELRA).
- Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *Preprint*, arXiv:2411.05025.
- Li Lin, Yu Wang, Yayu Ping, Jian Gao, Zongbo Wang, and Shouyu Wang. 2025. Design guidelines for human-generative ai interaction. In *International Conference on Human-Computer Interaction*, pages 223–239. Springer.
- Mary J. Lindstrom and Douglas M. Bates. 1990. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687.

- Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A. Myers. 2019. Unakite: Scaffolding developers' decision-making using the web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 67–80, New Orleans LA USA. Acm.
- Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2021. To reuse or not to reuse?: A framework and system for evaluating summarized knowledge. *Proceedings of the ACM on Human-Computer Interaction*, 5(Cscw1):1–35.
- Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2022a. Crystalline: Lowering the cost for developers to collect and organize information for decision making. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, New Orleans LA USA. Acm.
- Michael Xieyang Liu, Andrew Kuznetsov, Yongsung Kim, Joseph Chee Chang, Aniket Kittur, and Brad A. Myers. 2022b. Wigglite: Low-cost information collection and triage. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA. Association for Computing Machinery.
- Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation*, 61(6):700–712.
- Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, and 36 others. 2023. The semantic reader project: Augmenting scholarly documents through ai-powered interactive reading interfaces. *Preprint*, arXiv:2303.14334.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Fernando Loizides and George Buchanan. 2009. An empirical study of user navigation during document

- trriage. In *Research and Advanced Technology for Digital Libraries*, volume 5714, pages 138–149. Springer, Berlin, Heidelberg.
- Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.
- Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *Preprint*, arXiv:2501.04306.
- Tonja Machulla, Mauro Avila, Pawel Wozniak, Dillon Montag, and Albrecht Schmidt. 2018. Skim-reading strategies in sighted and visually-impaired individuals: A comparative study. In *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference, Petra '18*, pages 170—177, New York, NY, USA. Association for Computing Machinery.
- Anna Martin-Boyle, Aahan Tyagi, Marti A. Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition. *Preprint*, arXiv:2402.12255.
- Damien Masson, Sylvain Malacria, Edward Lank, and Géry Casiez. 2020. Chameleon: Bringing interactivity to static digital documents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu, HI, USA. Association for Computing Machinery.
- Michael E. Masson. 1982. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):400–417.
- Michael E. J. Masson. 1983. Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition*, 11(3):262–274.
- Martha J. Maxwell. 1972. Skimming and scanning improvement: The needs, assumptions and knowledge base. *Journal of Reading Behavior*, 5(1):47–59.

- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Yan Meng, Liangming Pan, Yixin Cao, and Min-Yen Kan. 2023. Followupqg: Towards information-seeking follow-up question generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–271, Nusa Dua, Bali. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint*. ArXiv:2203.11147 [cs].
- Douglas C Merrill, Brian J Reiser, Michael Ranney, and J Gregory Trafton. 1992. Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the learning sciences*, 2(3):277–305.
- Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.
- Matthew Michelson and Katja Reuter. 2019a. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary clinical trials communications*, 16:100443.
- Matthew Michelson and Katja Reuter. 2019b. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. *Contemporary Clinical Trials Communications*, 16:100443.
- Microsoft. 2022. Scroll bar map mode and bar mode - visual studio (windows).
- Meredith Ringel Morris. 2023a. Scientists’ perspectives on the potential for generative ai in their fields. *Preprint*, arXiv:2304.01420.

Meredith Ringel Morris. 2023b. Scientists’ perspectives on the potential for generative ai in their fields. *Preprint*, arXiv:2304.01420.

Eliot Moss. 2021. Connected papers. Accessed: 2025-04-04.

Mozilla. 2022. Pdf.js.

Mui. 2025. Material ui.

Sonia Murthy, Kyle Lo, Daniel King, Chandra Bhagavatula, Bailey Kuehl, Sophie Johnson, Jonathan Borchardt, Daniel Weld, Tom Hope, and Doug Downey. 2022. ACCoRD: A multi-document approach to generating diverse descriptions of scientific concepts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 200–213, Abu Dhabi, UAE. Association for Computational Linguistics.

Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, and Narges Mahyar. 2023. How data scientists review the scholarly literature. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, Chiir ’23*, pages 137–152, New York, NY, USA. Association for Computing Machinery.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint*. ArXiv:2112.09332 [cs].

Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2022. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496.

Theodor H. Nelson. 1983. Literary machines: The report on, and of, project xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow’s intellectual revolution, and certain other topics including knowledge, education and freedom.

- Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 43–76. Springer US, Boston, MA.
- Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. Pawns: Pdf annotation with labels and structure. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 258–264, Online. Association for Computational Linguistics.
- Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.
- Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.
- Phong H. Nguyen, Kai Xu, Andy Bardill, Betul Salman, Kate Herd, and B.L. William Wong. 2016. Sensemap: Supporting browser-based online sensemaking through analytic provenance. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 91–100, Baltimore, MD, USA. Ieee.
- Donald A Norman. 1988. *The psychology of everyday things*. Basic books.
- Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Chi '23*, pages 1–20, Hamburg Germany. Acm.
- Pallets. 2023. Welcome to flask – flask documentation (2.3.x).

- Rock Yuren Pang, Hope Schroeder, Kynneddy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the llm-ification of chi: Unpacking the impact of llms at chi through a systematic literature review. *Preprint*, arXiv:2501.12557.
- Debajyoti Pati and Lesa N Lorusso. 2018. How to write a systematic review of the literature. *HERD: Health Environments Research & Design Journal*, 11(1):15–30.
- Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review*, 106(4):643.
- Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis*, 5:2–4.
- Pontus Plavén-Sigra, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. The readability of scientific texts is decreasing over time. *eLife*, 6:e27725.
- James Prather, Brent N. Reeves, Paul Denny, Brett A. Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. “it’s weird that it knows what i want”: Usability and interactions with copilot for novice programmers. *ACM Trans. Comput.-Hum. Interact.*, 31(1).
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. *arXiv preprint arXiv:2410.03781*.
- Riaz Qureshi, Daniel Shaughnessy, Kayden AR Gill, Karen A Robinson, Tianjing Li, and Eitan Agai. 2023. Are chatgpt and large language models “the answer” to bringing us closer to systematic review automation? *Systematic Reviews*, 12(1):72.
- Napol Rachatasumrit, Jonathan Bragg, Amy X. Zhang, and Daniel S Weld. 2022. Citeread: Integrating localized citation contexts into scientific paper reading. In *27th International Conference on Intelligent User Interfaces*, Iui ’22, page 707–719, New York, NY, USA. Association for Computing Machinery.

- Gonzalo Ramos, Napol Rachatasumrit, Jina Suh, Rachel Ng, and Christopher Meek. 2022. Forsense: Accelerating online research through sensemaking integration and machine research support. *ACM Trans. Interact. Intell. Syst.*, 12(4).
- David N. Rapp and Paul van den Broek. 2005. Dynamic text comprehension: An integrative view of reading. *Current Directions in Psychological Science*, 14(5):276–279.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, 11(3):269–282.
- Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, Nips’16*, page 3574–3582, Red Hook, NY, USA. Curran Associates Inc.
- Seema Rawat and Sanjay Meena. 2014. Publish or perish: Where are we heading? *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 19(2):87.
- Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. 2016. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1):4–34.
- William R. Reader and Stephen J. Payne. 2007. Allocating time across multiple texts: Sampling and satisficing. *Human–Computer Interaction*, 22(3):263–298.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- John P. Rickards. 1980. Notetaking, underlining, inserted questions, and organizers in text: Research conclusions and educational implications. *Educational Technology*, 20(6):5–11.

- Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):197:1–197:45.
- Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, Chi '93, pages 269–276, New York, NY, USA. Association for Computing Machinery.
- Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A. Rossi, and Franck Dernoncourt. 2023. Pdftriage: Question answering over long, structured documents. *arXiv preprint*. ArXiv:2309.08872 [cs].
- Gavriel Salomon. 1988. Ai in reverse: Computer tools that turn cognitive. *Journal of educational computing research*, 4(2):123–139.
- Advait Sarkar. 2024a. Ai should challenge, not obey. *Commun. ACM*, 67(10):18–21.
- Advait Sarkar. 2024b. Intention is all you need. *Preprint*, arXiv:2410.18851.
- Bill N. Schilit, Gene Golovchinsky, and Morgan N. Price. 1998. Beyond paper: Supporting active reading with free form digital ink annotations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '98, page 249–256, Usa. ACM Press/Addison-Wesley Publishing Co.
- m. c. schraefel, Yuxiang Zhu, David Modjeska, Daniel Wigdor, and Shengdong Zhao. 2002. Hunter gatherer: interaction support for the creation and management of within-web-page collections. In *Proceedings of the 11th international conference on World Wide Web*, Www '02, pages 172–181, New York, NY, USA. Association for Computing Machinery.
- m.c. schraefel, Max Wilson, Alistair Russell, and Daniel A. Smith. 2006. mspace: improving information access to multimedia domains with multimodal exploratory search. *Communications of the ACM*, 49(4):47–49.
- Athar Sefid and C. Lee Giles. 2022. Scibertsum: Extractive summarization for scientific documents. In *Document Analysis Systems: 15th IAPR International Workshop*, page 688–701, Berlin, Heidelberg. Springer-Verlag.

- Abigail Sellen and Eric Horvitz. 2024. The rise of the ai co-pilot: Lessons for design from aviation and beyond. *Commun. ACM*, 67(7):18–23.
- Leila Shahrzadi, Ali Mansouri, Mousa Alavi, and Ahmad Shabani. 2024. Causes, consequences, and strategies to deal with information overload: A scoping review. *International Journal of Information Management Data Insights*, 4(2):100261.
- Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Ido Dagan, and Yael Amsterdamer. 2022. Interactive query-assisted summarization via deep reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2551–2568, Seattle, WA, USA. Association for Computational Linguistics.
- Paul Shekelle, Eduardo Ortiz, Shannon Rhodes, Sally C. Morton, M. Eccles, Jeremy M. Grimshaw, and Steven H. Woolf. 2001. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? *Jama*, 286 12:1461–7.
- Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. 2022. Vila: Improving structured content extraction from scientific pdfs using visual layout groups. *Transactions of the Association for Computational Linguistics*, 10:376–392.
- Kaveh G Shojania, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher. 2007. How quickly do systematic reviews go out of date? a survival analysis. *Annals of internal medicine*, 147(4):224–233.
- Mark Simmonds, Georgia Salanti, Joanne McKenzie, Julian Elliott, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, Charlotte Pestridge, and 1 others. 2017. Living systematic reviews: 3. statistical methods for updating meta-analyses. *Journal of clinical epidemiology*, 91:38–46.
- Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

- Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339.
- Luca Soldaini and Kyle Lo. 2023. pes2o (pretraining efficiently on s2orc) dataset. Technical report, Allen Institute for AI. ODC-By, <https://github.com/allenai/pes2o>.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MpNet: Masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, Nips’20*, Red Hook, NY, USA. Curran Associates Inc.
- Ayah Soufan, Ian Ruthven, and Leif Azzopardi. 2022. Searching the literature: An analysis of an exploratory search task. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval, Chiir ’22*, page 146–157, New York, NY, USA. Association for Computing Machinery.
- Meta Open Source. 2023. React.
- Michael Spence, Christian Beilken, and Thomas Berlage. 1996. Focus: the interactive table for product comparison and selection. In *Proceedings of the 9th annual ACM symposium on User interface software and technology, UIST ’96*, pages 41–50, New York, NY, USA. Association for Computing Machinery.
- Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, WA, USA. Association for Computational Linguistics.
- A. Strauss and J.M. Corbin. 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Thousand Oaks, CA.
- 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 2024 CHI Conference on Human Factors in Computing Systems, Chi ’24*, New York, NY, USA. Association for Computing Machinery.

- Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.
- Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, no or idk: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1075–1085, Seattle, WA, USA. Association for Computational Linguistics.
- Nicole Sultanum, Christine Murad, and Daniel Wigdor. 2020. Understanding and supporting academic literature review workflows with litsense. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces*, Avi '20, New York, NY, USA. Association for Computing Machinery.
- Teo Susnjak, Peter Hwang, Napoleon Reyes, Andre L. C. Barczak, Timothy McIntosh, and Surangika Ranathunga. 2025. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Trans. Knowl. Discov. Data*, 19(3).
- Amanda Swearngin, Shamsi Iqbal, Victor Poznanski, Mark Encarnación, Paul N. Bennett, and Jaime Teevan. 2021. Scraps: Enabling mobile capture, contextualization, and use of document resources. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Yokohama Japan. Acm.
- Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Chi '24, New York, NY, USA. Association for Computing Machinery.
- Craig S. Tashman and W. Keith Edwards. 2011. Active reading and its discontents: The situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '11, pages 2927—2936, New York, NY, USA. Association for Computing Machinery.

- Carol Tenopir, Donald King, Sheri Edwards, and Lei Wu. 2009. Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib Proceedings*, 61(1):5–32.
- Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W. White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, Chiir '20, pages 43–52, New York, NY, USA. Association for Computing Machinery.
- Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.
- Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.
- James Thomas, Anna Noel-Storr, Iain Marshall, Byron Wallace, Steven McDonald, Chris Mavergames, Paul Glasziou, Ian Shemilt, Anneliese Synnot, Tari Turner, Julian Elliott, Thomas Agoritsas, John Hilton, Caroline Perron, Elie Akl, Rebecca Hodder, Charlotte Pestridge, Lauren Albrecht, Tanya Horsley, and 123 others. 2017a. Living systematic reviews: 2. combining human and machine effort. *Journal of Clinical Epidemiology*, 91:31–37.
- James Thomas, Anna Noel-Storr, Iain Marshall, Byron Wallace, Steven McDonald, Chris Mavergames, Paul Glasziou, Ian Shemilt, Anneliese Synnot, Tari Turner, and 1 others. 2017b. Living systematic reviews: 2. combining human and machine effort. *Journal of clinical epidemiology*, 91:31–37.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 41 others. 2022. Lamda: Language models for dialog applications. *arXiv preprint*. ArXiv:2201.08239 [cs].
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence

- scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Andrea C Tricco, Jamie Brehaut, Maggie H Chen, and David Moher. 2008. Following 411 cochrane protocols to completion: a retrospective cohort study. *PLoS One*, 3(11):e3684.
- Chun-Yen Tsai, Chih-Neng Lin, Wen-Ling Shih, and Pai-Lu Wu. 2015. The effect of online argumentation upon students’ pseudoscientific beliefs. *Computers & Education*, 80:187–197.
- Chris Tyler, KL Akerlof, Alessandro Allegra, Zachary Arnold, Henriette Canino, Marius A Doornenbal, Josh A Goldstein, David Budtz Pedersen, and William J Sutherland. 2023. Ai tools as science policy advisers? the potential and the pitfalls. *Nature*, 622(7981):27–30.
- Raymon van Dinter, Bedir Tekinerdogan, and Cagatay Catal. 2021. Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589.
- Richard Van Noorden and Jeffrey M Perkel. 2023. Ai and science: what 1,600 researchers think. *Nature*, 621(7980):672–675.
- Laura Vergara-Merino, Catalina Verdejo, Cynthia Carrasco, and Manuel Vargas-Peirano. 2020. Living systematic review: new inputs and challenges. *Medwave*, 20(11):e8092.
- Hedwig von Restorff. 1933. Über die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, 18(1):299–342.
- David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Find-*

- ings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, Nips’20*, Red Hook, NY, USA. Curran Associates Inc.
- Xingbo Wang, Samantha L. Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. 2024a. Scidasynth: Interactive structured knowledge extraction and synthesis from scientific literature with large language model. *Preprint*, arXiv:2404.13765.
- Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, pages 1–1, Vancouver, BC, Canada. NeurIPS. Poster.
- Jane Webster and Richard T. Watson. 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii.
- Alan J. Wecker, Joel Lanir, Osnat Mokryn, Einat Minkov, and Tsvi Kuffik. 2014. Semantize: visualizing the sentiment of individual document. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pages 385–386, Como, Italy. Association for Computing Machinery.
- Michel Wijkstra, Timo Lek, Tobias Kuhn, Kasper Welbers, and Mickey Steijaert. 2021. Living literature reviews. In *Proceedings of the 11th Knowledge Capture Conference, K-cap ’21*, page 241–248, New York, NY, USA. Association for Computing Machinery.

- Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint*. ArXiv:2109.10862 [cs].
- Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022a. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, Chi '22, New York, NY, USA. Association for Computing Machinery.
- Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022b. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, WA, USA. Association for Computational Linguistics.
- Yong Xie, Dakuo Wang, Pin-Yu Chen, Jinjun Xiong, Sijia Liu, and Oluwasanmi Koyejo. 2022. A word is worth a thousand dollars: Adversarial attack on tweets fools stock prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–599, Seattle, WA, USA. Association for Computational Linguistics.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria. OpenReview.net.
- Qian Yang, Gerard de Melo, Yong Cheng, and Sen Wang. 2017. Hitext: Text reading with dynamic salience marking. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 311–319, Perth, Australia. Association for Computing Machinery.
- Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.
- Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper

- summarization with citation networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33(01), pages 7386–7393, Washington, DC, USA. AAAI Press.
- Ji Soo Yi. 2014. Qndreview: read 100 chi papers in 7 hours. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pages 805–814, Toronto, Ontario, Canada. Association for Computing Machinery.
- Dongwook Yoon, Nicholas Chen, Bernie Randles, Amy Cheatle, Corinna E. Löckenhoff, Steven J. Jackson, Abigail Sellen, and François Guimbretière. 2016. Richreview++: Deployment of a collaborative multi-modal annotation system for instructor feedback and peer discussion. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, Cscw '16, page 195–205, New York, NY, USA. Association for Computing Machinery.
- J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21, Hamburg, Germany. Acm.
- Polle T. Zellweger, Bay-Wei Chang, and Jock D. Mackinlay. 1998. Fluid links for informed and incremental link transitions. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space*, Hypertext '98, page 50–57, New York, NY, USA. Association for Computing Machinery.
- Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work & Social Computing*, Cscw '17, page 2082–2096, New York, NY, USA. Association for Computing Machinery.
- Jiajie Zhang. 1997. The nature of external representations in problem solving. *Cognitive Science*, 21(2):179–217.
- Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024a. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *Preprint*, arXiv:2409.02897.

- Xiaolong Zhang, Yan Qu, C. Lee Giles, and Piyou Song. 2008. Citesense: supporting sensemaking of research literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '08, pages 677–680, New York, NY, USA. Association for Computing Machinery.
- 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*, pages 1–16, Hamburg, Germany. Acm.
- Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.
- Zekun Zhou, Xiaocheng Feng, Lei Huang, Xiachong Feng, Ziyun Song, Ruihan Chen, Liang Zhao, Weitao Ma, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, Ting Liu, and Bing Qin. 2025. From hypothesis to publication: A comprehensive survey of ai-driven research support systems. *Preprint*, arXiv:2503.01424.
- Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6790–6804, Singapore. Association for Computational Linguistics.
- Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful classroom deployment of a social document annotation system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Chi '12, page 1883–1892, New York, NY, USA. Association for Computing Machinery.

Appendix A

Additional Details for Qlarify

A.1 Evaluation Details

A.1.1 Interview Questions

The following questions were used to elicit qualitative insights from participants in the interview study. These questions were used to initially guide the discussion, and probing questions were used to further elaborate on responses.

- Summarize your overall impressions of this interactive abstract interaction. What did you like or dislike?
- What other features or improvements would you want in a future system with interactive abstracts?
- Did you prefer asking questions by highlighting text or by clicking one of the underlined entities, and why?
- How do you feel about the choice of three static questions provided in the question palette? Are there other questions you would have wanted?
- Did you use either the attributed evidence paragraph or the ability see the evidence in the context of the paper? If so, how did you use it and was it useful?
- How did you feel about the quality of the generated responses?

- How did you feel about the quality of the suggested question in the question palette?
- Do you think this interactive abstract would be useful in your own research workflows, and if so, how?

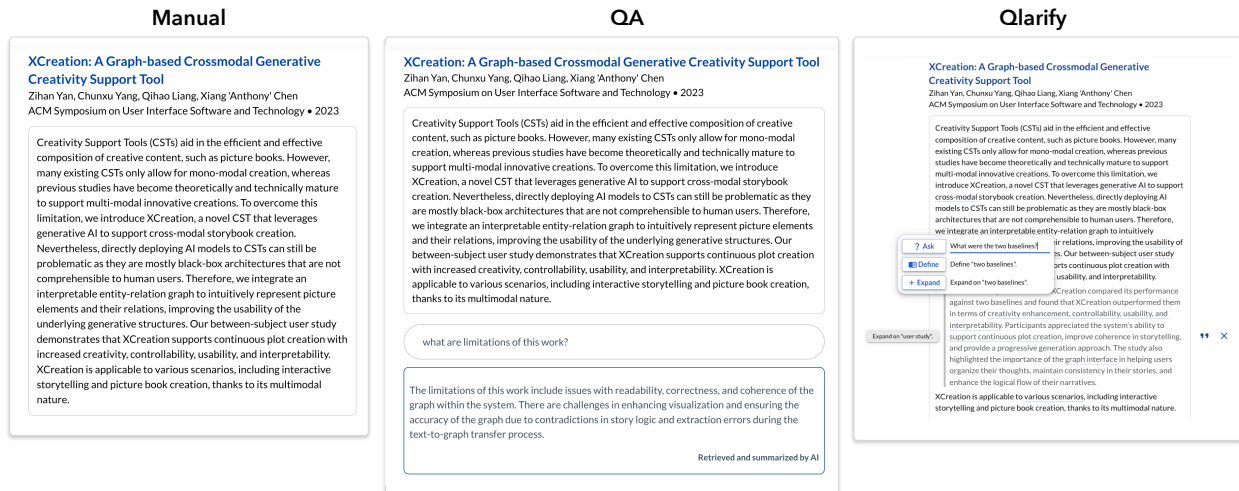


Figure A.1: Comparison of systems in the three conditions in Study 3: Comparative Evaluation.

A.1.2 Post-Task Survey Questions

For the post-task survey in the comparative evaluation, participants rated their agreement with the following statements on a seven-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

- *Satisfaction*: “I felt satisfied with my experience using the tool to explore the abstracts/papers.”
- *InfoGain*: “The tool helped me gain relevant information while reading the abstracts/papers.”
- *InfoAccurate*: “The information gained through my exploration was accurate.”
- *InfoReliable*: “The information gained through my exploration was reliable.”
- *Confidence*: “I felt confident while exploring the abstracts/papers.”
- *ExploreDeeper*: “I felt motivated to ask questions or explore deeper into abstracts/papers.”
- *ValueAdd*: “I believe the tool can add value to my research process.”
- *FutureUse*: “If publicly available in the future, I would use a similar tool for exploring abstracts/papers.”

A.1.3 System Updates for Comparative Evaluation

Figure A.2 illustrates the modifications made to Qlarify’s design based on feedback from the interview and field deployment studies.

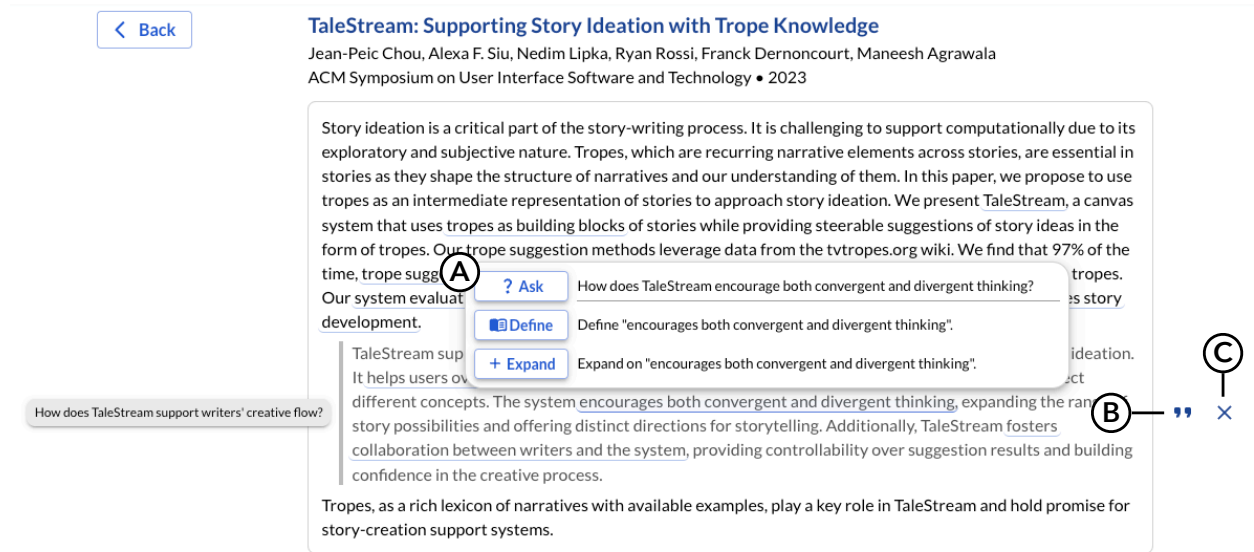


Figure A.2: Updated QLARIFY system interface for the comparative evaluation. Refinements included allowing users to edit the AI-suggested question (A), moving the button for viewing attributed evidence in the PDF to the right margin (B), and adding a button to remove a generated expansion (C).

A.1.4 User Action Maps

Figure A.3 illustrates how scholars used Qlarify’s features to explore paper abstracts, compared to baseline approaches of manual triage and standard question answering.

A.2 Example Qlarify Outputs

Table A.2 lists LLM-generated questions and expansions for each detected expandable entity, and Table A.1 lists example expansions for the static questions.

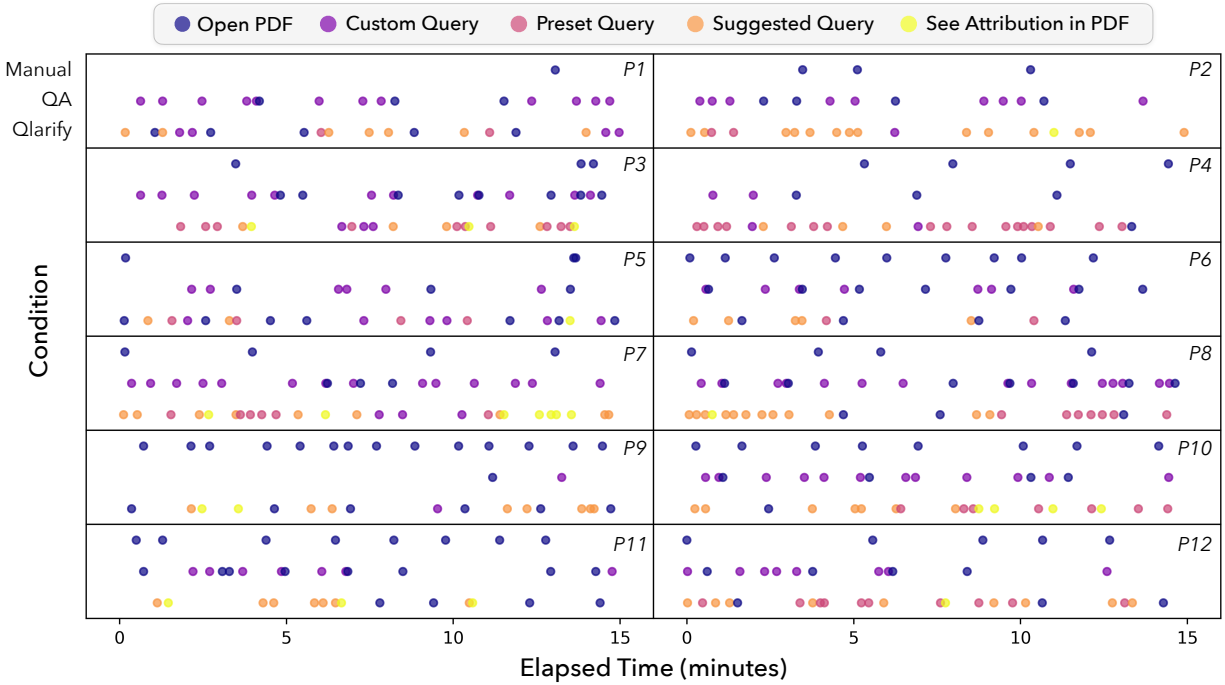


Figure A.3: Time-graph of user actions during task completion in the comparative evaluation.

A.3 LLM Prompts

Expandable Entity Extraction

```

You are a helpful research assistant that asks questions about abstracts of scientific papers.

List all questions that a curious reader might have after reading this abstract. These questions must not be answerable
given the abstract, but may be answerable given the full paper. These questions could help clarify vague terms, define
jargon, request for more detail, or ask for justification. Each question should be short and not contain multiple
sub-questions. Provide a phrase (three words or less) verbatim from the abstract that motivates each question.

Title: {Title}
Abstract: {Abstract}

{Examples}

Questions:

```

Question Generation

Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking

Hyeonsu B Kang, Sherry Wu, Joseph Chee Chang, A. Kittur • arXiv.org • 2023

Efficiently reviewing scholarly literature and synthesizing prior art are crucial for scientific progress. Yet, the growing scale of publications and the burden of knowledge make synthesis of research threads more challenging than ever. While significant research has been devoted to helping scholars interact with individual papers, building research threads scattered across multiple papers remains a challenge. Most top-down synthesis (and LLMs) make it difficult to personalize and iterate on the output, while bottom-up synthesis is costly in time and effort. Here, we explore a new design space of mixed-initiative workflows. In doing so we develop a novel computational pipeline, Synergi, that ties together user input of relevant seed threads with citation graphs and LLMs, to expand and structure them, respectively. Synergi allows scholars to start with an entire threads-and-subthreads structure generated from papers relevant to their interests, and to iterate and customize on it as they wish. In our evaluation, we find that Synergi helps scholars efficiently make sense of relevant threads, broaden their perspectives, and increases their curiosity. We discuss future design implications for thread-based, mixed-initiative scholarly synthesis support tools.

Figure A.4: QLARIFY’s user interface with expandable entities underlined and no expansions created (abstract shown is from [Kang et al., 2023]).

```
You are a helpful research assistant that predicts what question a reader might have.

A reader has highlighted a span of text in the abstract. What is the most likely question they could ask about the span?
The question must not be answerable given the abstract, but may be answerable given the full paper. The question may help
clarify vague terms, define jargon, request for more detail, or ask for justification. The question should be short and not
contain multiple sub-questions. Try framing the question as: How? Why? What? Such as?

Abstract: {Abstract}
Target span: "{Entity}", in the sentence "{Sentence}"
Question:
```

Question Answering

```
You are a helpful research assistant that answers questions about scientific papers.

Answer the question based on the following excerpts from the full text of the paper. Incorporate quotes verbatim from the
excerpts when relevant. If the question cannot be answered from the provided context, reply "No answer." Your answer should
be {Response Length}.
```

Table A.1: Examples of LLM-generated expansions for the three static questions in Qlarify (from [Kang et al., 2023]; abstract in Figure A.4).

Question	Entity	Expansion
<i>Define</i>	“mixed-initiative workflows”	Mixed-initiative workflows refer to a design approach that combines both human input and automated assistance in the process of conducting literature reviews and synthesizing knowledge from scientific papers. It involves the integration of user input, such as relevant seed threads, with computational algorithms and interaction features to expand and structure the threads. This approach allows scholars to start with a pre-generated structure and customize it according to their needs, facilitating efficient sense-making, broadening perspectives, and increasing curiosity.
<i>Expand</i>	“broaden their perspectives”	Participants in the study engaged with Synergi-generated threads of research to "broaden their perspectives." This means that the threads provided them with a wider range of information and insights, allowing them to consider different viewpoints and explore various aspects of the literature. By broadening their perspectives, participants were able to gain a more comprehensive understanding of the topic they were reviewing.
<i>Why</i>	“building research threads scattered across multiple papers remains a challenge”	Building research threads scattered across multiple papers remains a challenge because it requires synthesizing information from multiple sources and organizing them in a coherent manner. Prior approaches have supported largely manual bottom-up synthesis workflows, which can be time-consuming and inefficient. Synergi addresses this challenge by automatically retrieving important papers from a 2-hop neighborhood on the citation graph and summarizing them in a hierarchical manner, allowing users to quickly discover relevant threads and understand them through synthesis by other scholars.

```

{Examples}

Context: {Context}
Question: {Question}
Answer:

```

Table A.2: Examples of Qlarify’s LLM-generated expandable entities, suggested questions, and expansions. All examples are actual system output for the paper *Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking* [Kang et al., 2023] (abstract in Figure A.4).

Entity	Question	Expansion
“top-down synthesis”	What is meant by “top-down synthesis” and “bottom-up synthesis”?	Top-down synthesis refers to the approach of using large language models (LLMs) to generate summaries or syntheses of research papers. It is characterized by limited personalization and iteration on the output. On the other hand, bottom-up synthesis involves practices such as citation chasing and footnote chasing to discover important papers related to a research problem. It is more time-consuming but allows for a more comprehensive synthesis.
“mixed-initiative workflows”	What is the definition of a “mixed-initiative workflow”?	A mixed-initiative workflow is a type of workflow that combines both human input and automated processes. It involves user input of relevant seed threads, which are then expanded and structured using citation graphs and LLMs. Scholars can start with a pre-generated structure and customize it as they wish, allowing for personalization and iteration.
“relevant seed threads”	What is meant by “relevant seed threads”?	“Relevant seed threads” refers to the structured research threads that are generated based on the selected clips and their associated citation contexts and seed references. These threads are considered relevant because they align with the user’s interests and provide valuable information about specific research areas in the literature.
“citation graphs and LLMs”	How does Synergi use citation graphs and LLMs to expand and structure research threads?	Synergi uses citation graphs to retrieve important papers from a 2-hop neighborhood and summarize them in a hierarchical manner. It generates a 3-level hierarchy of research threads by capturing the core commonalities among the papers. It also utilizes LLMs like GPT4 to condense scholarly text and provide synthesis.
“iterate and customize”	What is the process for scholars to iterate and customize the threads-and-subthreads structure?	Scholars can review the system-generated threads, curate the ones that interest them, and build upon them iteratively. They can easily curate useful threads, references, and contexts into an editor using drag-and-drop. They can also prioritize references by using the group-by-reference view. They can expand and collapse the structured thread recommendations to access the relevant citation contexts and source materials.
“In our evaluation”	How was the evaluation of Synergi conducted?	The evaluation of Synergi involved 12 participants and domain experts. They compared Synergi to a baseline system, Threddy, and GPT4. The evaluation measured the quality of the outlines, efficiency in constructing them, and participants’ perception of Synergi-generated threads and experience. The evaluation used experts’ judgment, survey responses, qualitative data from interviews, and participants’ interaction logs.
“broaden their perspectives”	How does Synergi help scholars broaden their perspectives?	Synergi helps scholars broaden their perspectives by providing them with system-generated threads of research that they can quickly scan to understand how different sub-group structures map to the broader literature. This understanding allows scholars to orient their attention towards specific areas that align with their interests.
“future design implications”	What are the future design implications for thread-based, mixed-initiative scholarly synthesis support tools?	The future design implications for thread-based, mixed-initiative scholarly synthesis support tools include incorporating LLMs as components in computational pipelines, exploring the design space of interaction designs that benefit users in discovering, prioritizing, extracting, organizing, and synthesizing knowledge, and conducting additional ablation studies to understand the contributions of each component in the pipeline.

Appendix B

Interview Protocol for Living Syntheses Study

B.1 Interview Protocol

Our semi-structured interviews included the following guide questions. When appropriate, follow-up questions were used to encourage participants to elaborate on their responses, for instance to probe deeper into process details or recall motivations for a particular decision. In *Process.*, participants were asked to describe their workflows in as much detail as possible. Guide questions for specific scholarly activities were used only if unacknowledged by participants; otherwise, the discussion proceeded naturally, e.g., “*What did you do next?*”

- *Introduction.* Could you briefly describe your area of research and the survey paper?
- *Team Composition.* Could you describe the research team involved? How large was the team? Broadly, what were the responsibilities of each member?
- *Process.* Could you walk me through your entire process of writing this survey paper, starting from the beginning?
 - *Ideation.* How did you come up with the idea for the survey paper?

- *Paper Search*. Could you walk me through your initial paper search process? For instance, how did you find, filter, or organize papers?
 - *Coding*. How did you begin to make sense of the information within the papers?
 - *Writing*. How did you translate the information you collected and organized into writing the paper?
- *Perceived Challenges*. What did you feel were the most challenging aspects of the process?
 - *Tedious Aspects*. Were there any parts of the process that stood out as tedious, time-consuming, or cognitively challenging?
 - *Intermediate Artifacts*. What intermediate products (e.g., documents, spreadsheets) did you create, if any, during the process? How did you use those artifacts? What was helpful or not helpful about these artifacts?
 - *Tools*. What tools did you use throughout the process of writing this paper? What worked well, and what could have worked better?
 - *Process Adjustments*. If you were to write this survey paper again, what would you have done differently?
 - *Update Process*. Could you walk me through how you would update this paper? (Including similar sub-questions as in *Process*. above.)
 - *Update Elements*. Which parts of the paper do you envision would change from these updates? For instance, could you speak to potential revisions regarding the existing text, structure, figures and tables, or other elements?
 - *Update Challenges*. What do you believe are the main challenges in the process of updating this paper?
 - *Desired Support*. What kind of support would you want in the process of updating this paper?
 - *AI Perspectives*. Which parts of the survey paper authoring or updating process do you believe an AI would be capable or incapable of assisting you with?

Appendix C

Additional Details for DimInd

C.1 User Study Details

C.1.1 Demographic Details of User Study Participants

Table C.1 lists demographic details for all 23 participants included in our user study, including their research area, years of research experience, and use of LLMs for general and research tasks.

C.1.2 Additional Task Details

Task Scenario and Instructions

The following hypothetical survey paper writing scenario was presented to participants in each task, informing their exploration over the collection of 50 papers.

Imagine that you are part of a research team planning to write a survey paper reviewing recent research on the following topic: <topic>. You have conducted an initial search and found a set of possibly relevant papers. Today, your goal is to review this set of papers to get a broad understanding of this topic. The research team has sketched an initial outline and assigned you with two top-level sections to explore and add detail to. Using the provided papers, you could:

- Create one or more subsections that would be relevant to include in each section.
- Note down specific details, e.g., terms, concepts, or perspectives, relevant to each section.

ID	Age	Gender	Position	YoE	Research Area	LLM-Gen	LLM-Res
P1	25-34	Man	PhD student	3-5	GenAI safety and ethics	Extensively	Occasionally
P2	25-34	Woman	Incoming PhD student	3-5	HCI, AI, Accessible Computing	Frequently	Frequently
P3	18-24	Man	Undergraduate student	0-2	NLP, Reasoning	Extensively	Extensively
P4	25-34	Woman	PhD student	3-5	HCI, AI ethics	Occasionally	Never
P5	25-34	Woman	Master’s student	0-2	Data Visualization	Frequently	Frequently
P6	35-44	Man	Postdoctoral Researcher	3-5	Cybersecurity	Extensively	Frequently
P7	25-34	Man	PhD student	3-5	Security and Privacy	Extensively	Frequently
P8	18-24	Man	Undergraduate student	0-2	HCI, CSCW, AR/VR, AIMC	Extensively	Occasionally
P9	25-34	Man	PhD student	6-10	HCI, Human-AI	Extensively	Extensively
P10	25-34	Woman	PhD student	3-5	HCI, Health Tracking	Frequently	Occasionally
P11	18-24	Woman	PhD student	0-2	HCI, AR/VR applications	Extensively	Frequently
P12	25-34	Woman	Master’s student	0-2	HCI, Human-AI	Extensively	Occasionally
P13	25-34	Woman	PhD student	3-5	HCI	Rarely	Rarely
P14	25-34	Man	PhD student	10+	HCI, Metascience	Frequently	Frequently
P15	25-34	Woman	PhD student	6-10	HCI, LLM/Health	Frequently	Frequently
P16	25-34	Man	PhD student	3-5	HCI, AI	Occasionally	Occasionally
P17	18-24	Woman	Undergraduate student	0-2	HCI, ML	Extensively	Occasionally
P18	25-34	Man	PhD student	3-5	HCI, Social Computing	Extensively	Extensively
P19	25-34	Woman	PhD student	3-5	HCI, Usable Security	Occasionally	Rarely
P20	18-24	Woman	Incoming PhD student	3-5	HCI, AI, Accessibility	Frequently	Extensively
P21	25-34	Woman	PhD student	3-5	HCI	Extensively	Frequently
P22	25-34	Man	PhD student	3-5	HCI, ML, UbiComp	Extensively	Extensively
P23	25-34	Woman	PhD student	3-5	HCI, Emotion & Wellbeing	Frequently	Frequently

Table C.1: User study participants. **YoE** refers to years of experience conducting scholarly research. **LLM-General** refers to how frequently LLM applications are used for everyday tasks. **LLM-Research** refers to how frequently LLM applications are used for research activities.

- Cite specific papers relevant to each section, e.g., with title or author+year.

You will have up to 30 minutes to work on the outline. You are not expected to fully complete the outline or consider every paper—try to do as much as you can in the given time.

Creating the Sampled Paper Collections

For each survey paper, we retrieved the full list of references using the Semantic Scholar API. Papers without open-access PDFs were excluded. To standardize the collection size, each set was limited to the 50 most recent papers—a size for which cognitive challenges of large-scale review could be observed, while still being relatively manageable with current LLM tools for creating an initial review outline.

C.1.3 Post-Task Survey Questions

Table C.2 lists the post-task survey questions (7-point Likert scale) participants completed after system condition in our user evaluation.

Code	Statement
Reduce effort to extract information	The system helped reduce the mental effort required to extract and organize information.
Categorize papers effectively	I was able to identify and categorize relevant papers in a meaningful way.
Discover meaningful connections	The system helped me discover connections between papers that I might have missed otherwise.
Maintain control	I maintained appropriate control over the literature review process when using this system.
Verify information	I believe I could easily verify information provided by the system.
Confidence in review quality	I feel confident in the quality and comprehensiveness of the outline I created using this system.

Table C.2: Post-task survey questions and their corresponding labels (rated on 7-point Likert scale).

C.2 Creating Paper Collections

Paper collections in DIMIND can be interactively initialized, e.g., from a research question or search query (Figure C.1). Finding relevant papers is a critical yet challenging part of the literature review workflow and one that we leave for future work separate from our exploration of DIMIND.

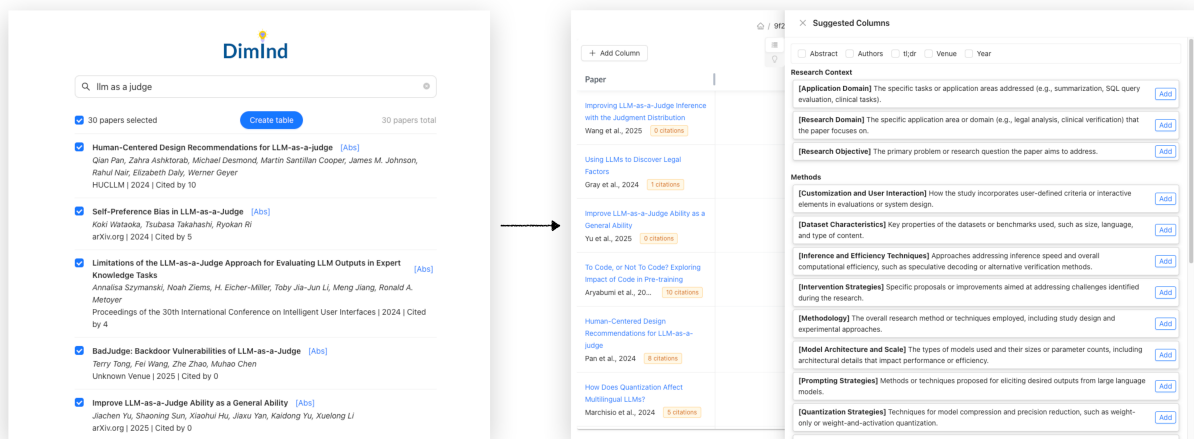


Figure C.1: Paper collections in DIMIND can be created interactively from a research question or search query.

C.3 DIMIND LLM Prompts

C.3.1 Inducing Collection-Aware Facets

Facet Induction

A user wants to write a literature review for a set of related research papers. The following is a list of contexts from the papers. Your task is to identify facets whose values would likely allow a user to meaningfully compare and contrast across the papers.

Context: {**context**}

Generate facets that can be used to compare and contrast different aspects of information across the set of papers. Each facet should be a short phrase that can be used to compare information across the papers. For each facet, generate a description in the form of a short question that the facet would help answer. Generate at most {**max_facets**} facets. The challenge is to find specific facets that are relevant to this set of papers, in addition to generally useful facets for comparing research papers, such as 'Study design' or 'Research questions'.

Keep each facet focused on a SINGLE concept. Do not combine multiple concepts into one facet. For example, instead of "Evaluation Metrics and Results", create separate facets for "Evaluation Metrics" and "Study Results". Instead of "Implications and Future Work", create separate facets for "Research Implications" and "Future Work Directions".

For example, specific facets such as 'User study methodology', 'Number of participants', or 'System design goals' could be relevant for helping a user explore a set of HCI papers. Or with a set of papers on machine learning for information retrieval, potentially informative facets may be 'Number of parameters' or 'Retrieval methods'.

Here are some other examples of facets: Intervention effects, Study design, Study objectives, Theoretical framework, Research questions, Dataset characteristics, Study count, Study duration, Statistical techniques, Algorithm type, Software tools, Participant demographics, Policy recommendations, Design goals, Research limitations, Ethical considerations, etc.

Return a single valid JSON list (without code block) containing objects with the name and description of each facet. Do not return any other text.

Example output: [{ "name": "<facet 1>", "description": "<description of facet 1>" }, { "name": "<facet 2>", "description": "<description of facet 2>" }, { "name": "<facet 3>", "description": "<description of facet 3>" }, ...]

Facet Merge

The following list contains facets used to analyze research papers. Your task is to:

- * Identify and consolidate truly duplicate facets (exact matches or synonyms)
- * Keep distinct concepts separate, even if they appear related
- * Select up to {**max_facets**} most important facets for understanding research papers

Guidelines:

- * Do not combine distinct concepts with "and" (e.g., keep "Data Collection" and "Data Analysis" as separate facets)

- * Only consolidate facets that mean exactly the same thing (e.g., "User Study" and "User Evaluation" could become just "User Study")
- * Each facet should represent a single clear concept
- * Preserve specificity where it adds value

The following is the list of facets to process:

```
{facets}
```

Return your output as a JSON array (without code block). Each facet should have:

- * 'name': A concise label for the facet
- * 'type': The data type (text, number, or boolean)
- * 'description': A clear, focused description of what the facet captures

```
Example output: [ { "name": "Research Objective", "type": "text", "description": "The primary goal or aim of the research" }, { "name": "Methodology", "type": "text", "description": "The research method used to conduct the study" } ]
```

C.3.2 Extracting Faceted Information

Value Extraction

Use the provided paper context to retrieve information relevant to the specified list of facets. Each facet has a number id, a "Name", and a "Description" (optional). When provided, the "Description" key provides context (e.g., additional instructions or example output formats) for the information expected to be extracted for that facet.

- * For each facet, generate a paragraph of detailed, accurate, and relevant information (typically between 3 to 5 sentences) by synthesizing relevant information from the provided paper context.
- * Use a passive or third-person voice when summarizing information for an facet. For instance, avoid using phrases such as "we", "our approach", etc.
- * If there is no relevant information for an facet in the provided context, return null for that facet value. Do not return any other text and do not make up an answer unsupported by the paper context.

Your output should contain a list of objects. Each object should have:

- * 'facet_id': The id of the facet (as provided in the input).
- * 'value': The information extracted for that facet, or null if no relevant information is found.

The output should contain exactly as many objects as the facets provided in the input.

Paper Context:

```
{context}
```

Facets:

```
{facets}
```

Output format:

Do not include any explanations, only provide a valid JSON response (without code block). For example, if you are provided with 3 facets, the output should be in the following format: [{ "facet_id": 1, "value": "<information for facet 1>" }, { "facet_id": 2, "value": "<information for facet 2>" }, { "facet_id": 3, "value": null }]

Value Distillation

Summarize information related to the following facet of research papers: **{facet}** (description (optional): **{facet_description}**) from research paper excerpts. For each excerpt, return exactly one sentence of clear and concise information about this specific facet. The goal is for your summary to allow a user to more quickly understand **{facet}**, e.g., during an initial exploratory phase of literature review, and return to the original, longer excerpt later if they desire additional detail.

Guidelines for your summaries:

- * Focus only on information directly related to **{facet}**
- * Keep each summary to a maximum of 20 words
- * Use present tense and consistent formatting across all summaries
- * If an excerpt contains no relevant information about **{facet}**, return an empty string
- * Include key statistics and metrics when present
- * Avoid subjective interpretations or evaluations

Your output must be a valid JSON object (without code block) where:

- * Keys are the original paperIds
- * Values are either null or a string containing a single sentence summary
- * Maintain the exact paperId format and order as provided in the input, making sure each paper in the input has a corresponding output value (null or string)

Example input for the facet "participant demographics":

```
{{"CorpusId:123": "The user study demographics consist of 32 trained undergraduates who have completed at least one course in computer science or statistics. A total of over 2100 responses were collected from these participants, ensuring a diverse pool of users with relevant academic backgrounds to assess the effectiveness of explanation methods.", "CorpusId:456": "The user study demographics include 72 participants, with a gender distribution of 19 women and 2 who declined to state their gender. Participants' ages range from 18 to over 50, with a majority falling between 30 and 39 years old. The study does not involve expert participants, but efforts were made to enhance their domain knowledge through training tasks.", "CorpusId:300": null }}
```

```
Example output: {"CorpusId:123": "32 trained undergraduates with computer science or statistics background provided over 2100 responses.", "CorpusId:456": "72 participants including 19 women, aged mostly 30-39, with no expertise but received training tasks.", "CorpusId:300": null }}
```

Now, summarize the following excerpts, returning your response as a valid JSON object (without code block):

{excerpts}

C.3.3 Synthesizing

Taxonomy Generation

A researcher is analyzing a collection of papers to write a literature review. They have extracted snippets of relevant information related to the following facet:

{facet_name_and_description}

Your task is to organize these snippets into a hierarchical structure that provides clear, specific, and informative categorization. The goal is to create a nested organization that logically groups similar evidence across papers together.

Hierarchy Requirements:

- * Construct a hierarchy with a maximum depth of 5 levels.
- * Let the natural structure of the snippets determine how many levels are appropriate.
- * The final level must contain only arrays of snippet indices (0-based indexing).
- * The top level must have no more than (**max_n_categories**) categories.
- * Aim to have categories that are not too small (e.g., 1 or 2 snippets) or too large (e.g., 20+ snippets). The goal is to have snippets within each category be meaningful when examined together.
- * Use subcategories when snippets cover diverse aspects that deserve separation.
- * Avoid vague or generic category names.
- * Ensure that ALL snippets are included somewhere in your hierarchy.
- * A snippet can belong to one or more categories (i.e., it may make sense for a snippet to be included in multiple relevant categories if it covers multiple topics).

Use "Miscellaneous" Categories Sparingly: * Do not create a "Miscellaneous" category unless absolutely necessary.

Check for patterns before grouping snippets under a catch-all category.

Split "Miscellaneous" categories into subcategories when possible.

Keep them small--only use them for genuinely diverse snippets.

Input paper snippets:

```
{snippets}
```

Output format:

The output must be a valid JSON object (without code block), where: * Keys represent categories and subcategories.

- * Values are either subcategories or arrays of snippet indices.
- * Do not include additional keys (e.g., "indices", "description", "items").
- * Do not include any explanation, preamble, or additional formatting.

Example:

If the input snippets were about "performance metrics", a good hierarchy could be:

```
{ "Model Performance": { "Text Processing": { "Classification": { "Accuracy Metrics": [0, 3] }, "Named Entity Recognition": { "Precision Metrics": [1] }, "Translation": { "Error Analysis": [2] } } }, "System Efficiency": { "Speed": { "Response Time": [4, 6], "Processing Throughput": [7] }, "Resource Utilization": { "Computational Resources": [5], "Memory Usage": [8], "GPU Performance": [9] } }, "User Experience": { "Satisfaction": [10, 12], "Reliability": [11, 13] } }
```

Summarization

Transform the following taxonomy of organized information snippets from different research papers into a clear and concise summary that captures the key points related to the given facet. Your synthesis should be formatted as a valid JSON object with a single key "summary_blocks" containing an array of objects, each with "header" and "content" keys.

Facet:

```
{facet_name_and_description}
```

Taxonomy Structure and Paper Excerpts:

Excerpts are extracted from various research papers and provides information relevant to a more detailed aspect of the facet above:

{**excerpts**}

Papers to Highlight:

When provided, the following papers should be highlighted in your summary. Incorporate them smoothly:

{**starred_papers**}

Additional Instructions:

- * Structure your response as a valid JSON dictionary with exactly one key "summary_blocks" containing an array of objects.
- * Each object in the array should have two keys: "header" (corresponding to a top-level category from the taxonomy structure) and "content" (containing your synthesized text for that category).
- * Headers should match the top-level structure provided in the taxonomy.
- * Ensure all papers from the taxonomy are included in your summary across the different blocks.
- * Your content should be primarily descriptive. Do not include introductory or concluding sentences. Do not generate statements that connect snippets in meaningless ways.
- * Synthesize the key details into cohesive and insightful content blocks, covering as much of the provided information as possible.
- * Use the provided hierarchical structure as a guide for organizing your summary blocks.
- * Break your content into paragraphs when appropriate to improve readability.
- * Prioritize the papers listed in the "Papers to highlight" section if any are listed, by ensuring they are incorporated into the summary. Make sure to incorporate them smoothly (e.g., by including more detail about those papers, but don't explicitly say phrases like "the highlighted papers examine..." or "the featured works on ...")
- * Your entire summary across all blocks should be {**length_constraint**}.
- * Remember that these snippets you are summarizing come from different papers. Avoid language like "This paper" or "This system", since the summary is synthesizing across multiple papers.

Citing Sources:

- * Your summary MUST cite all unique papers in the extracted snippets at least once, to the extent possible.
- * Your citations MUST use the exact paperId format provided in the extracted snippets. The paper id could be any string (e.g., "12345", "CorpusId:12345", "URL:12345", etc.), and you should cite using the exact and entire paper id.
- * Place citations [[paperId]] immediately after specific words, phrases, or concepts they support - not just at the end of sentences. For example: "The study found increased levels of protein X [[paperId1]] and decreased levels of enzyme Y [[paperId2]] in the treatment group."
- * When multiple sources support the same specific point, group those citations together like [[paperId1, paperId2, paperId3]].

Output format:

```
{ { "summary_blocks": [ { { "header": "First Top-Level Category", "content": "Your synthesized content with precise citations after specific terms [[paperId1]] or concepts [[paperId2]]. When listing multiple findings such as A [[paperId3]], B [[paperId4]], and C [[paperId5, paperId6]], citations should follow each item they support rather than appearing only at the end of sentences." }, { { "header": "Second Top-Level Category", "content": "More synthesized content with appropriate citations [[paperId7, paperId8]]." } } ] }
```