AI-Powered K-12 Instruction: Educator-AI Interaction, Teacher-Authored Student-AI Interaction, and Nudges

Alex Liu

A dissertationsubmitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Min Sun, Chair

David Knight

Lorraine Males

R. Benjamin Shapiro

Program Authorized to Offer Degree:
Education

University of Washington

Abstract

AI-Powered K-12 Instruction:
Educator-AI Interaction, Teacher-Authored Student-AI Interaction, and Nudges

Alex Liu

Chair of the Supervisory Committee:
Min Sun
College of Education

This dissertation positions educator agency as a core condition for making generative AI (GenAI) pedagogically meaningful and classroom-ready. Across educator-facing and classroom-facing contexts, the central problem is not whether AI can generate fluent content, but whether educators can reliably translate pedagogical intent into usable artifacts and productive interactions under authentic constraints such as time pressure, incomplete information, competing priorities, and the need to balance rigor, equity, and feasibility. The dissertation synthesizes three complementary studies. Study 1 analyzes large-scale authentic educator-AI conversations and characterizes educators' AI-assisted work as iterative professional design, marked by repeated cycles of ideation, specification, repair, validation, and formatting to reach classroom-ready products. Study 2 investigates the classroom deployment of a teacher-configured student-AI conversational feature, showing how teacher-authored prompts function as scaffolds that transmit and preserve their pedagogical intent in student-AI interactions. Study 3 examines pedagogically grounded, AI-generated next-step nudges embedded in educator-AI conversations across a large-scale deployment, showing that adoption is selective yet widespread and increases when nudges are immediately executable and effort-reducing. The findings show that classroom-ready GenAI depends on tool designs that preserve teachers' professional expertise and pedagogical intent while reducing workflow friction and providing practical, evidence-grounded decision support that

helps educators translate intent into action.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

ADOPTION: Observable uptake of an optional system affordance; in this dissertation, typically operationalized as click behavior that converts a displayed nudge into the educator's next request.

AI AGENT: A configured LLM-based assistant with a defined system prompt, role, and behavior constraints for a specific feature or workflow.

AI LITERACY: Educators' knowledge, skills, and dispositions for using AI responsibly and effectively, including recognizing limitations, steering outputs, and exercising professional judgment.

ARTIFACT: A concrete, reusable product generated or refined through educator-AI interaction (e.g., worksheet, rubric, table, lesson component, communication draft).

AUTHENTIC USE: Naturalistic, real-world platform use occurring outside controlled experimental tasks, reflecting educators' self-initiated goals and constraints.

CLASSROOM-READY: Sufficiently specified, formatted, and context-aligned for immediate instructional use with students, with minimal additional repair or rework by the educator.

CONVERSATION: A multi-turn interaction thread between an educator and an AI chat system, consisting of alternating requests and responses.

CONVERSATION TURN: One cycle of educator request and AI response within a conversation.

EDUCATOR-AI CONVERSATION: Open-ended, multi-turn interaction between an educator and an AI system used for professional tasks such as planning, assessment design, differentiation, and reflection.

GENAI: AI systems capable of producing text (and other content) in response to prompts; in this dissertation, primarily LLM-based conversational systems.

LLM (LARGE LANGUAGE MODEL): A neural language model trained on large-scale text corpora that generates text based on context; used as the underlying engine for chat-based AI features.

OPEN-ENDED WORKFLOW: A task context where goals evolve across turns and multiple valid paths exist, requiring iterative clarification and refinement.

PRACTICE-ALIGNED: Designed or evaluated in terms of instructional practices and professional tasks rather than generic text quality.

TRAJECTORY: The evolution of conversation content and intent across turns, including shifts in goals, constraints, and artifacts being developed.

WORKFLOW FRICTION: Time and cognitive effort required to move from intent to implementable outputs, including prompting, validation, repair, formatting, and integration work.

CODEBOOK: A structured set of qualitative codes used to annotate educator requests and/or conversations, including code definitions and decision rules.

CODING PIPELINE: The end-to-end workflow for producing qualitative annotations at scale (e.g., sampling, prompt design for model-assisted labeling, human verification, and consolidation into analysis-ready datasets).

CONVERSATION-LEVEL ANALYSIS: Aggregation of measures so that each conversation contributes once per construct (e.g., whether a practice appears anywhere in the thread), used to characterize what educators do across multi-turn interaction.

DESCRIPTIVE ANALYSIS: Empirical characterization of distributions, frequencies, and co-occurrence patterns of tasks or practices without causal claims.

DOMAIN: A high-level grouping of instructional work (e.g., planning, assessment, differentiation, reflection) used to organize coded educator-AI activity.

EDUCATOR GOAL: The instructional or professional outcome an educator is attempting to achieve in a conversation (e.g., design an activity, create a rubric, draft a parent message).

HUMAN-IN-THE-LOOP CODING: A coding approach in which model outputs are reviewed, corrected, or validated by human researchers to ensure interpretability and reliability.

INSTRUCTIONAL PRACTICE: A pedagogically meaningful action or decision represented in educator requests (e.g., aligning to standards, anticipating misconceptions, adding formative checks).

ITERATIVE PROMPTING: A pattern in which educators refine requests across turns to repair misalignment, add constraints, request revisions, or reformat outputs.

QUALITATIVE ANNOTATION: The assignment of interpretive labels (codes) to messages or conversations to represent instructional practices, goals, or interaction moves.

SCALE: The quantity of observed interactions analyzed (e.g., number of messages, conversations, users), often emphasizing deployment-level evidence in naturalistic settings.

DEPTH OF KNOWLEDGE (DOK): A framework for classifying task cognitive demand; used in this dissertation to analyze intended versus realized rigor in AI-mediated activities [171].

FINISH LINE: An explicit statement of what "done" looks like for a task or interaction, often used to bound work and support completion.

OPT-OUT: Non-participation as an outcome in classroom deployment settings, where students or educators may choose not to engage with a tool.

STUDENT–AI DIALOGUE: A conversational interaction between students and an AI system, often bounded by teacher-authored scaffolds and classroom activity goals.

TASD (TEACHER-AUTHORED STUDENT DIALOGUE): A system configuration in which teachers author a hidden prompting layer and student-facing starter to shape student–AI conversations and classroom enactment.

TEACHER-AUTHORED PROMPT LAYER: The hidden or system-level prompt written by teachers to constrain AI role, norms, scaffolding moves, and guardrails for student-facing dialogue.

BOUNDED CHOICE: A choice architecture strategy that limits options presented at a decision point to reduce cognitive load and facilitate action (e.g., showing up to two next-step nudges) [79].

CHOICE ARCHITECTURE: The design of how options are presented in order to influence decisions without restricting freedom of choice [155].

DELAYED ADOPTION: Nudge uptake occurring after intervening conversation turns, enabled by persistence of previously displayed nudges.

DIGITAL NUDGING: Use of interface-level prompts or design elements to influence decisions in digital systems without restricting user choice [172, 118].

EXECUTABILITY PRINCIPLE: The finding that suggestions producing concrete, immediately usable artifacts and reducing effort are adopted more frequently than abstract or reflective prompts [66].

FRAMING STRATEGY: The rhetorical or behavioral heuristic used to present a suggestion (e.g., effort reduction, implementation intention) to shape salience and perceived ease.

HEURISTIC STRATEGY TAXONOMY: A predefined set of framing strategies applied to nudges to shape attention and adoption behavior.

IMMEDIATE ADOPTION: Nudge uptake in the same turn in which the nudge is displayed.

IMPLEMENTATION INTENTION: A behavioral mechanism that increases goal attainment by specifying concrete next steps (when, where, and how to act) [66].

NUDGE: A short, optional, clickable next-step prompt presented to an educator to influence workflow decisions without restricting choice [155].

NUDGE DISPLAY EVENT: A single instance of a nudge being shown to a user within an AI response, used as a primary analytic unit in Study 3.

NUDGE LABEL: An immutable action type from a fixed taxonomy that specifies the function of a nudge (e.g., generate worksheet, organize into a table).

NUDGE PERSISTENCE: A design property in which previously displayed nudges remain available across subsequent turns, enabling delayed adoption.

RETENTION: Continued use of a feature after initial trial, often operationalized by repeat adoption behavior over time.

SELECTIVE ADOPTION: A pattern where users adopt only a small fraction of displayed suggestions, indicating discretion rather than habitual compliance.

TENURE: Time since first exposure to a feature (e.g., days since first nudge display), used to examine changes in adoption patterns over experience.

ACKNOWLEDGMENTS

This work is supported by the Institute of Education Sciences of the U.S. Department of Education, through Grant R305C240012 and by several awards from the National Science Foundation (NSF #2043613, 2300291, 2405110) to the University of Washington, and a NSF SBIR/STTR award to Hensun Innovation LLC (#2423365). The opinions expressed are those of the authors and do not represent views of the funders.

Declaration of GenAI Software Tools in the Writing Process

During the preparation of this work, the author(s) used ChatGPT-4o in the Chapter 3 sections Related Works, Methodology, Findings, Implications and Discussion, ChatGPT-5.2 for the rest of chapters in order to correct grammar and spelling errors. Figure 1.1, Figure 3.1, and Figure 4.1 are generated by Colleague AI Generate Image feature. After using this tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

DEDICATION

To my parents, for their unwavering and unconditional love and support.

To my advisor, for her insight, guidance, and belief in this work.

And to my cats who, in moments of nihilism, planted a quiet seed of existential meaning and made this beginning possible.

Chapter 1

INTRODUCTION

Generative AI (GenAI) has moved rapidly from novelty to infrastructure in educators' day-to-day work, with educators increasingly using LLM-based tools to draft lesson materials, differentiate tasks, generate assessments, and obtain just-in-time pedagogical and professional suggestions [56]. Yet the evidence base remains uneven. Within K-12 specifically, recent synthesis highlights both expanding interest and a shortage of work that documents how educators and students actually incorporate generative tools into practice across subject areas, roles, and classroom constraints [5]. The lack of understanding impedes the development of systemic supports for educators and limits actionable feedback for educational technology developers. This fragmentation creates closely connected practical gaps. First, although adoption is visible, we still lack scalable, descriptive accounts of how educators use GenAI in authentic contexts, both for their own professional workflows and for classroom instruction, beyond small samples or self-reported data [34]. Second, it remains unclear which interface and workflow supports can reliably improve outcomes, and without stronger human-centered design grounded in educator needs, intervention design may prioritize surface-level activity over instructional coherence and usability [6].

This dissertation focuses on a central practical problem: as AI becomes embedded in educator workflows and classroom instruction, what makes AI interactions instructionally valuable in real-world educational settings? Across teacher-facing and student-facing contexts, it advances a human-centered view of classroom-ready AI: instructional value depends not only on model capability but also on whether educators can shape AI use through (a) alignment with authentic professional workflows and iterative instructional reasoning, (b) modelable and bounded interaction routines for students, and (c) system support and workflow scaffolds that reduce friction at key decision points while preserving professional judgment.

1.1 Cross-Paper Research Questions

This dissertation addresses three overarching questions:

1. **Teacher-facing use:** How do K-12 educators use GenAI tools in real-world professional contexts, and what patterns of instructional reasoning and task demands are visible at scale in unscripted educator-AI conversations?

2. **Teacher-authored student-AI dialogue:** What do teachers create when given the ability to author and configure student-AI dialogues, and how do these teacher-authored designs shape what actually occurs in student-AI conversations?

3. **Workflow scaffolding:** What system design characteristics should inform the development of supports in an educator-facing platform, and how do scaffolding mechanisms (e.g., nudges) influence educators' use of AI?

1.2 The Three-Paper Arc

This dissertation is organized as three complementary studies that share a educator-centered framing but address distinct layers of educator practices and agency.

Study 1 (How K-12 Educators Use AI) Study 1 in Chapter 2 analyzes more than 13,000 unscripted educator-AI conversations from Colleague AI, an open-access educational platform during the study period, to characterize how K-12 educators use GenAI for purposes such as lesson planning, differentiation, assessment, and pedagogical reflection. Beyond its empirical findings on educator goals and interaction patterns, Study 1 introduces a replicable, human-led and LLM-assisted qualitative analysis pipeline that supports inductive theme discovery, codebook development, and large-scale annotation while preserving researcher oversight for conceptual synthesis and validation. This study establishes the baseline for the dissertation by documenting what educators actually do with teacher-facing AI in authentic professional contexts and by clarifying the implications for designing tools that align with real educational tasks and constraints.

Study 2 (Teacher-Authored Prompts for Configuring Student-AI Dialogue) Study 2 in Chapter 3 examines Colleague AI Classroom Teaching Aide (TASD), a teacher-configured, student-facing AI conversation tool that treats teacher-authored prompts as instructional design levers, through a K-12 classroom implementation study. In TASD, teachers author a prompting layer that conditions the AI's role, norms, scaffolding moves, and guardrails for student-facing conversations. The study traces an end-to-end pathway from teacher-authored designs to classroom participation and student-AI interaction patterns, with outcomes including task alignment, realization of cognitive demand using Depth of Knowledge (DOK) [171], student interaction moves, and adherence to AI safeguards. This work builds on prompt-layer and structured guidance approaches that aim to make learner-LLM interaction more instructionally reliable [69, 168, 175], while foregrounding classroom orchestration requirements related to boundedness and supervisability [48, 49].

Study 3 (Nudges as Educator-AI Interaction Scaffolds) Study 3 in Chapter 4 studies a large-scale deployment in which an educator-facing AI chat surfaces up to two optional, clickable next-step nudges [155] after each response. Drawing on platform logs and clickstream data, the paper examines how educators adopt, defer, or ignore nudges over time, and how adoption varies with nudge design characteristics and workflow context. This study positions nudges as autonomy-preserving choice points within open-ended educator professional work, linking adoption patterns to the cognitive constraints and decision frictions that shape educator AI use [137, 70].

A Connection Map This dissertation examines how educators' practical use of GenAI shapes the feasibility and instructional value of AI systems across teacher-facing and classroom-facing contexts. Study 1 characterizes educators' authentic use of generative AI in professional work and shows that educator-AI interaction often involves repeated and cognitively demanding decision points, including iterative ideation, prompting, repair, and reformatting to produce classroom-ready artifacts. Study 2 shifts to classroom deployment to investigate how teacher-authored prompts configure student-AI discussion activities, treating prompts as instructional design levers through which pedagogical intent is enacted in AI sys-

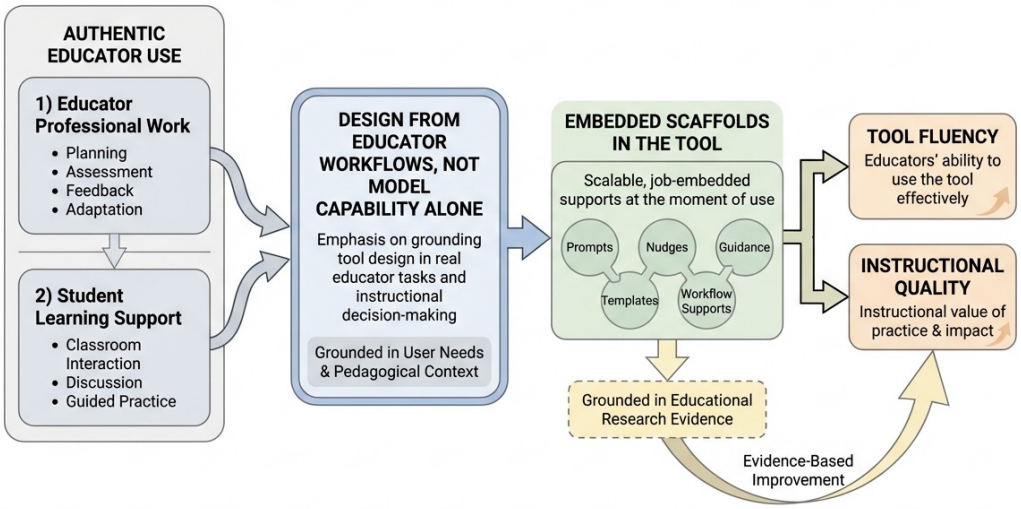

Figure 1.1: Conceptual Framework for Building Instructional Valuable Tools in Educational Technology

tems at classroom scale. Study 3 returns to teacher-facing chat and examines AI-powered nudges as workflow scaffolds that reduce friction at key decision points while preserving autonomy, demonstrating that adoption is selective, time-sensitive, and shaped by design properties such as executability and persistence. Taken together, the three studies show that classroom-ready AI depends not only on model capability, but also on alignment with authentic educator work, bounded and authorable routines for student interaction, and scaffolds that make productive next steps easier to enact in the flow of practice.

1.3 Overarching Conceptualization

This dissertation is guided by a conceptualization (Figure 1.1) of instructionally valuable AI as grounded not primarily in model capability, but in the relationship between tool design, authentic educator use, and embedded support. To build AI tools that are genuinely useful in education, we must first understand how educators use them across both professional workflows (Chapter 2) and classroom-facing activity (Chapter 3), including planning, assessment, adaptation, and support for student learning. These contexts matter because

instructional value emerges under real constraints, not idealized demonstrations of model performance. From this perspective, effective AI design should begin with authentic educator workflows and the instructional decisions they require. It should also incorporate scalable, job-embedded scaffolds directly into the tool so that support is available at the moment of use rather than dependent only on external training (Chapter 4). When such scaffolds are grounded in educational research, they can do more than improve usability; they can help educators develop greater fluency with the tool while also supporting more instructionally advanced and pedagogically coherent practice.

1.4 Dissertation Contributions

This dissertation is organized into three complementary studies that share an educator-centered framing while addressing distinct layers of educator practice and agency.

Empirical and Methodological Contribution (Study 1) A large-scale account of how K-12 educators use GenAI in authentic professional contexts, grounded in unscripted educator-AI conversations and supported by a replicable, LLM-assisted qualitative analysis pipeline.

Classroom Deployment Contribution (Study 2) An empirical account of teacher-authored student-AI dialogue at classroom scale, linking teachers' pedagogical intentions to student-AI conversation outcomes and demonstrating the feasibility of teacher-in-the-loop student-facing AI tool design. The study also provides constructive, practice-oriented operational guidance for classroom AI-mediated discussion implementation.

Design and Intervention Contribution (Study 3) The design and deployment of AI-powered nudges as scalable interaction scaffolds that support educators in effectively using AI tools, advancing pedagogical practice through subtle interventions during ongoing interactions. The study also introduces the executability principle as a theoretical contribution, positioning nudges as a workflow mechanism that helps translate emerging instructional intentions into actionable steps.

Integrative Contribution (Dissertation-Level) An educator-centered model of classroom-ready AI that connects educator use (teacher-facing workflows), educator control (teacher-authored student dialogue routines), and educator decision support (nudges as autonomy-preserving choice points) as complementary levers for feasibility, responsibility, and instructional value

1.5 Dissertation Chapters

Chapter 1 (this chapter) introduces the motivation, cross-paper research questions, and integrative framing. Chapters 2-4 present Papers 1-3 as stand-alone manuscripts. Chapter 5 synthesizes insights across the three studies, articulates implications for system design and teacher practice, discusses limitations and boundary conditions, and outlines a future research agenda focused on human-centered AI that remains viable within real-world educational settings.

Chapter 2

HOW K-12 EDUCATORS USE AI: LLM-ASSISTED QUALITATIVE ANALYSIS AT SCALE

Abstract. This study investigates how K-12 educators use GenAI tools in real-world instructional contexts and how large language models (LLMs) can support scalable qualitative analysis of these interactions. Drawing on over 13,000 unscripted educator-AI conversations from an open-access platform, we examine educators' use of AI for lesson planning, differentiation, assessment, and pedagogical reflection. Methodologically, we introduce a replicable, LLM-assisted qualitative analysis pipeline that supports inductive theme discovery, codebook development, and large-scale annotation while preserving researcher control over conceptual synthesis. Empirically, the findings surface concrete patterns in how educators prompt, adapt, and evaluate AI-generated suggestions as part of their instructional reasoning. This work demonstrates the feasibility of combining LLM support with qualitative rigor to analyze complex educator behaviors at scale and inform the design of AI-powered educational tools.

2.1 Introduction

The integration of artificial intelligence (AI) into educational practice is influencing how educators plan instruction, differentiate learning, assess student progress, and engage in professional development [153]. Among recent developments, GenAI, particularly large language models (LLMs), has emerged as a tool that enables educators to co-design lessons, seek pedagogical guidance, and revise instructional materials through natural language interaction [167, 192, for example]. While prior research has examined the technical performance of AI-powered educational tools [21, 97] and theorized their pedagogical affordances [40, 110], empirical understanding remains limited regarding how practicing educators engage with GenAI at scale and how these interactions shape instructional reasoning in authentic pro-

fessional contexts [193].

Understanding how educators use AI in authentic professional contexts is crucial for two complementary aims. First, it supports the development of more transparent, context-sensitive AI tools that respond to real educational needs and mitigate predictable risks [112]. Second, it enables professional development providers, researchers, and educational leaders to study how AI systems can augment human instructional decision-making and develop systemic support for educators to effectively leverage the tools [120]. Equally important, such inquiry helps clarify what constitutes responsible and effective AI use in education, not only in terms of tool familiarity, but in how educators prompt, adapt, revise, and verify AI outputs as part of their instructional practice [103].

This study seeks to address a critical gap in understanding how K-12 educators engage with GenAI tools in authentic professional contexts. While educators are increasingly adopting AI to support a wide range of instructional and professional tasks, the complexity, scale, and variability of these interactions pose new analytic challenges for qualitative research that typically focuses on small, carefully curated samples [129]. During the study period, the platform was freely accessible, allowing educators to voluntarily initiate conversations, type natural language prompts, revise AI-generated content, and apply outputs to real educational tasks. Unlike lab settings or scripted activities, these interactions reflect authentic, task-driven educator behavior. To make sense of this emerging landscape, this study analyzes more than 13,000 educator-AI conversations to investigate two central research questions:

> **RQ1.** How can LLMs, embedded within a structured, human-led analytic pipeline, assist researchers in conducting scalable qualitative analysis of educator-AI interactions, particularly through collaborative theme development, codebook construction, and large-scale annotation?
>
> **RQ2.** In what ways do K–12 educators use GenAI tools to carry out key instructional and professional responsibilities in real-world educational contexts?

We applied a multi-stage qualitative analysis pipeline that integrates researcher-led interpretation with LLM (e.g., Claude 3.5 Haiku)-assisted processes across four stages: induc-

tive theme discovery, iterative codebook construction, structured annotation, and deductive analysis. Methodologically, this approach demonstrates a replicable way to incorporate LLMs into qualitative research at scale while maintaining researcher control over conceptual decisions and interpretive depth. Empirically, the study presents one of the first large-scale qualitative analyses of how K-12 educators use GenAI in real-world professional contexts. Drawing on a large corpus of unscripted educator-AI conversations, the analysis reveals how educators engage AI in tasks including planning, differentiation, assessment, and other professional responsibilities. The findings surface concrete patterns in how educators prompt, adapt, and critically evaluate AI responses, offering insight into the ways generative tools are embedded into everyday decision-making and instructional reasoning. Together, this study demonstrates the methodological feasibility and analytical value of LLM-supported qualitative research in understanding educator practice at scale.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature on qualitative analysis methods, LLM-assisted coding, and existing studies of AI use in education. Section 3 introduces the data and outlines the four-phase LLM-assisted qualitative analytic pipeline. Section 4 reports empirical findings from the coded dataset, organized by instructional and professional priorities. Section 5 discusses implications for educators' AI usage, followed by limitations and future directions in Section 6. Section 7 concludes with a synthesis of contributions. Full technical prompts, codebook, usage label descriptions are included in the Appendix.

2.2 Related Work

2.2.1 Qualitative Coding Techniques in Educational Research

Qualitative research has served as an important methodology for educational inquiry, enabling researchers to investigate complex reasoning and dynamics in rich, contextualized ways. Among qualitative methodologies, systematic coding approaches have been widely adopted for analyzing textual data, with techniques such as open, axial, and selective coding providing structured procedures for moving from raw data to meaningful categories [65]. Central to these approaches is the principle of inductive emergence, whereby codes

and categories are derived from data through iterative engagement rather than imposed a priori.

Strauss[39] described an iterative qualitative coding procedure composed of three main phases: open coding, axial coding, and selective coding. During open coding, researchers closely identify and describe phenomena in discrete parts of data to surface emergent concepts and recurring themes. Through techniques like line-by-line coding and memo writing, researchers label segments of text, such as educator interviews, classroom interactions, or instructional dialogues, based on actions, meanings, or events. This phase generates a broad array of initial codes and serves as the foundation for codebook development while open to new themes to emerge in the later stages.

In axial coding, the focus shifts from conceptualization to reassemble the data by establishing relationships among open codes. This includes identifying relatable factors, such as conditions, context, or actions, to refine coding paradigm that links concepts, enabling researchers to begin forming a coherent analytic structure [149]. Selective coding then further refines the core categories, allowing researchers to craft theoretical narratives that explain how the categories cohere into a conceptual model. A codebook will be established during the coding process as a structured product to illustrate the categories and how they interrelate in a meaningful way [136].

These phases are not strictly linear but are instead iterative and recursive, requiring constant comparison across data points and analytic levels [33]. Validation practices are integral to this approach. Researchers may use triangulation, member checking, or code-recode strategies to ensure analytic trustworthiness [67]. Leveraging domain knowledge in these checks are especially important in educational research, where field-specific understanding shapes interpretation and undercovers the nuances of pedagogical language demand contextual sensitivity.

While inductive coding is the foundation of these approaches, qualitative analysis can also incorporate a deductive phase wherein the refined codebook is applied to larger or different datasets to test its robustness and relevance across contexts [18, 127, for eg.]. The connected application supports conceptualization and enables researchers to assess the reliability and transferability of emergent frames using a different and potentially larger sample.

Although inductive and deductive qualitative coding is a time-consuming and labor-intensive process, which makes it difficult to extend traditional approaches to expansive, high-velocity datasets like platform-mediated educator-AI conversations [121], the methodological rigor and systematic procedure of established coding techniques provides foundational guidance for AI-assisted analysis to maintain these qualitative principles while expanding analytic reach to scalable and diverse text.

2.2.2 LLM-Assisted Qualitative Coding: Emerging Approaches Across Analytical Stages

There have been continuous efforts to scale qualitative analysis using computational methods, leverage techniques in natural language processing and machine learning to analyze large corpus of educational text data and uncover patterns in educator discourse [59, for eg.]. The capabilities of LLMs have introduced new possibilities for qualitative research in education and the social sciences. As researchers contend with increasingly large and complex textual datasets, LLMs are being explored not only as analytic tools but also as collaborative assistants capable of supporting different stages of qualitative analysis, from inductive theme discovery to codebook development and large-scale annotation. These developments build on, but also extend beyond, traditional qualitative data analysis methods, introducing new opportunities and challenges around interpretability, reliability, and the evolving role of human-LLM collaboration. This section reviews recent work on LLM-assisted qualitative research, synthesizing takeaways according to three analytic stages: theme discovery, codebook development, and deductive coding.

Recent studies have demonstrated that LLMs can support inductive coding and theme discovery by surfacing patterns across unstructured textual data. For instance, Sinha et al., (2024)[146] examined how GPT-4 can contribute to early stage grounded theory analysis in educational research, focusing on its role during open coding and constant comparison, framed LLMs as augmentative tools in grounded theory-style inquiry, assisting with pattern detection during open coding and constant comparison. Their findings highlight that LLMs, when guided by carefully crafted prompts and researcher oversight, can accelerate the identification of emergent categories, support theme refinement, and prompt critical

reflection without supplanting human interpretive judgment. DePaoli (2024)[44] used GPT-3.5-turbo to uncover under-acknowledged themes in educators' reflections on data science instruction, finding that the model offered valuable ideation prompts and interpretive nuance during early-stage coding. Beyond education, Wang et al., (2025)[166] evaluated the LLM-Assisted Thematic Analysis (LATA) approach and reported that GPT-generated thematic structures closely approximated those produced by trained human coders, though limitations in contextual depth remained. Similarly, Gao and colleagues (2025)[61] introduced a workflow designed to generate intermediate reasoning chains that allow researchers to follow the model's logic, offering greater transparency in theme generation and alignment with constructivist epistemologies. Collectively, these studies demonstrate how LLMs can expand the exploratory capacity of researchers during inductive analysis, with guided prompts and domain-specific framing.

At the codebook development stage, researchers have used LLMs to further refine coding schemes, extend conceptual boundaries, or adapt codes across datasets. For example, Barany and team (2024) [16] found that codebooks generated through human-AI hybrid approaches were more reliably applied and rated higher in quality than those developed by automated methods alone. While the human-AI hybrid codebook and the fully human-generated codebook produced similar structures, the fully automated approaches resulted in clear outliers. These findings suggest that while LLMs can support the generation and refinement of codebooks in educational domains, they require human oversight to ensure methodological rigor. Shin et al., (2025)[143] compared human-developed and ChatGPT-developed codebook for classroom dialogue data found similarity while ChatGPT categorised utterances more granularly based on their purpose and function than the human coders. Similar amplifying effects are reported by Zambrano et al., (2025) [188], who showed that LLMs can serve as valuable collaborators by identifying key themes aligned with researcher-provided theoretical lenses, including less frequent but important constructs that are difficult for humans to detect at scale, ultimately contributing to higher-quality codebooks. Across related studies, the literature validates the application of LLM-assisted codebook development when paired with appropriate prompting, contextual grounding, and researcher triangulation.

Beyond exploratory tasks, LLMs have also demonstrated potential in deductive coding

and scalable annotation. Liu and Sun (2025)[104] applied GPT-4-turbo to over 1,000 paragraphs from educational stakeholder interviews using a pre-defined educational codebook, finding that the model achieved useful granularity at both parent and child code levels, effectively complementing human insights. Shin et al., (2025) [143] conducted a co-coding study using ChatGPT across more than 1,200 segments of classroom discourse, employing a model-generated codebook. The study reported high consistency with human coders on concrete categories but noted reduced alignment on more interpretive constructs. McClure et al., (2024) [113] similarly evaluated the effectiveness of LLMs in deductive coding within educational qualitative research on data from high school students engaged in an English Language Arts curriculum using AI, demonstrating an 8.5% higher overall accuracy compared to human coders. These studies underscore both the promise and the limitations of model-based annotation: while LLMs can efficiently apply structured codes, they require effective prompting and interpretation.

As many scholars caution [181, 20, for eg.], the use of LLMs in qualitative research raises ethical concerns, including risks of hallucination, inconsistency, and potential bias. These issues are particularly salient when analyzing educator data or discourse involving marginalized learners. To address these concerns, prior studies have emphasized sustained human involvement throughout the analytic process [16], transparent role allocation between human and machine [188], domain-knowledge-grounded prompt design [104], and explicit epistemological framing of the analysis process [107]. Across all stages of LLM-assisted analysis, a methodological consensus is emerging around human-LLM collaboration. LLMs are increasingly integrated into human-in-the-loop workflows in which domain-expert input directs, evaluates, and calibrates model outputs, underscoring the need for prompt transparency, dataset documentation, and shared analytic protocols.

2.2.3 Effective Instructional Practices in Pre-AI Educational Research

Decades of educational research have identified instructional practices that reliably support student learning, while also documenting persistent barriers to their widespread implementation. This body of work provides essential context for understanding how educators engage

with AI tools and what needs these tools might address.

Research on instructional planning has consistently shown that high-quality lesson design is both cognitively demanding and time-intensive [81, 50]. Effective planning requires teachers to anticipate student thinking, sequence content for coherent progression, and align activities with learning objectives [176]. Studies have found that novice and experienced teachers alike struggle to find adequate time for the iterative refinement that quality planning demands, with teachers frequently reporting that planning consumes evenings and weekends beyond contracted hours [83]. The challenge intensifies when teachers must adapt plans for multiple sections or prepare differentiated versions of materials for diverse learners.

Differentiated instruction, adapting content, process, and products to meet varied student needs, has strong empirical support as an approach to reaching diverse learners [158]. However, research has repeatedly documented a gap between teachers' knowledge of differentiation strategies and their capacity to implement them consistently [75]. Teachers report that creating multiple versions of materials, designing tiered activities, and preparing appropriate scaffolds for English learners or students with disabilities requires preparation time that is often unavailable within the school day. Large-scale surveys indicate that while most teachers believe in the value of differentiation, fewer than half report being able to implement it regularly due to time and resource constraints [27].

Similarly, formative assessment and feedback have been identified as among the most powerful influences on student achievement [22, 72]. Providing timely, specific, and actionable feedback helps students understand where they are in their learning and what steps to take next. Yet research consistently finds that the feedback students receive is often delayed, generic, or focused on surface features rather than substantive learning [177]. Teachers cite the time required to provide individualized feedback on student work as a primary barrier, particularly in secondary settings with large student loads. The labor-intensive nature of quality feedback creates a persistent tension between what research indicates is effective and what teachers can feasibly accomplish.

Professional learning research has emphasized that effective teacher development is sustained, job-embedded, and connected to daily practice [47, 42]. Teachers benefit from opportunities to collaboratively examine student work, refine instructional approaches, and

receive just-in-time support when facing novel challenges [24, 89]. However, access to such support is uneven. Many teachers, particularly those in under-resourced schools or isolated rural settings, lack regular access to instructional coaches, mentors, or collaborative planning time [62]. Even when professional development is available, it often occurs in formats disconnected from immediate classroom needs, limiting its applicability to the specific challenges teachers face in their daily work.

Across these domains, a common pattern emerges: educational research has identified practices that matter for student learning, but implementation at scale remains constrained by time, access to expertise, and the cognitive demands of translating general principles into context-specific instructional decisions. These documented barriers establish an important baseline for examining how educators engage with AI tools and whether such tools might help address longstanding gaps between research-supported practice and classroom reality.

2.2.4 What We Know About How Educators Use GenAI

The rapid expansion of the educational technology industry has sparked widespread interest in how educators incorporate these tools into their professional practice. Reports from leading companies such as Anthropic and OpenAI offer large-scale descriptive insights into emerging patterns of educator-AI interaction. Anthropic's recent analysis of approximately 74,000 anonymized conversations from higher education professionals using Claude revealed that the most common applications involved curriculum development (57%), academic research (13%), and assessment-related tasks (7%) [9]. These included explaining concepts, personalizing learning for students, and improving educator productivity. The report emphasized Claude's role as a "thought partner" rather than a "thought substitute," noting that educators rarely rely on AI for full automation, but instead use it to accelerate their work while maintaining professional oversight.

While this report offers one of the first large-scale landscape analyses of educational AI use, its scope is limited to users with higher education email addresses, thereby excluding K-12 educators. It also stops short of analyzing the specific instructional practices embedded within AI usage beyond task-level classification. Similarly, OpenAI published a report

on how college students use ChatGPT for learning and academic support [124], but a comprehensive understanding of how K-12 educators use AI in their daily professional tasks remains largely unexplored.

Complementing these platform-level analyses from industry reports, a growing body of studies is exploring the advantages, challenges, and implications of AI integration in real-world educational settings. Recent research has also begun to document AI use from educators' perspectives. Zaimoglu et al., (2025)[187], for example, found that educators working with English learners selectively used AI to scaffold differentiated instruction, while maintaining a critical stance toward the outputs. Their use of AI was shaped not only by pedagogical affordances, but also by professional values, political constraints, and their own sense of agency. Similarly, Cheah et al., (2025) [34] reported that K-12 educators in Idaho rural areas expressed curiosity and optimism about GenAI, particularly for out-of-classroom tasks such as lesson planning, assessment, and administrative work. Nucci et al., (2025) [123], in a study with 11 mathematics teachers, found that educators used GenAI to accelerate workflows in lesson planning, generate engaging and differentiated activities (e.g., for equivalent fractions), and design learning tasks aligned with higher-order thinking skills such as creativity and synthesis. While teachers viewed AI as a powerful tool for reducing workload and enabling more personalized instruction, the study also surfaced concerns around accuracy, usability, and the need for ethical guidelines.

In a national survey of 979 K-12 math and science educators, 50% reported having used GenAI. Of those users, 57% indicated infrequent use (monthly or rarely), while 42% reported more regular use (weekly or daily) [56]. A majority also reported using AI for lesson planning (76%), assessment development (61%), instructional support (50%), and differentiated academic support (32%). Studies of both in-service and pre-service educators further suggest that pedagogical beliefs, ethical considerations, and perceived usefulness all mediate how educators engage with AI in their professional practice.

Across these studies, several themes emerge. First, educators most frequently use AI to support upstream planning and content development tasks, particularly lesson generation, assessment, and administrative work. Second, the perceived and practical effectiveness of AI varies depending on educators' technological fluency and pedagogical beliefs. Third,

findings from both researcher-led studies and educator perspectives consistently emphasize the need for professional development and systemic support to enable effective and ethical AI integration [153]. However, many existing studies rely on self-reported data or focus on small samples limited by subject area or geographic scope. As a result, there remain significant gaps in our understanding of how K-12 educators actually use AI in practice, gaps that hinder the design of targeted supports and the identification of potential shortcomings in classroom implementation.

The evolving literature on qualitative research highlights a growing convergence between traditional analytic rigor and new LLM-assisted approaches. Established qualitative coding techniques continue to provide essential scaffolds for educational inquiry, emphasizing context, nuance, and human interpretation. At the same time, advances in LLMs have opened promising avenues for accelerating theme discovery, refining codebooks, and conducting large-scale deductive coding, particularly when embedded within human-in-the-loop workflows that uphold transparency, domain grounding, and methodological integrity. This synergy offers methodological ground for understanding pedagogical reasoning and professional agency in large-scale educator-AI conversations in this study.

2.3 Methodology: LLM-Assisted Qualitative Analytic Pipeline

To investigate how K-12 educators leverage GenAI to pursue instructional and professional goals, we developed a multi-phase qualitative analytic pipeline that integrates established qualitative coding techniques, such as open, axial, and selective coding, with selective use of LLMs. This approach addresses RQ1 by demonstrating how qualitative research practices can be scaled through structured LLM assistance while preserving human interpretive judgment. The methodology also supports RQ2 by enabling large-scale, structured annotation of educator-AI conversations to identify patterns in instructional practice, content design, and professional priorities.

The analytic process comprised four stages: (1) data collection from authentic educator-AI conversations, (2) recursive coding for codebook development, (3) deductive coding for large-scale annotation, and (4) descriptive and qualitative analysis for usage patterns. At each stage, researchers retained conceptual control, while LLMs were employed as assis-

tants for theme discovery, label extraction, and structured annotation. This division of labor follows precedences in the recent methodological advances in LLM-assisted qualitative research, enabling transparency and reproducibility.

2.3.1 Data Source: Authentic Educator-AI Conversations

This study draws on a dataset of educator-AI interactions collected from Colleague AI, an open-registration GenAI platform designed to assist educators with real-time instructional planning, content development, assessment, differentiation, and professional communication. At the time of data collection, the platform supported over 15,000 active users, including K-12 teachers, school administrators, paraprofessionals, and instructional specialists. All usage was voluntary and initiated by educators as part of their authentic instructional work. Researchers played no role in shaping or directing platform activity.

The platform allows educators to submit free-form natural language prompts and receive AI-generated outputs, including instructional materials, pedagogical strategies, and communication artifacts. For this study, we analyzed over 140,000 messages generated during these interactions. Prompts spanned a wide range of grade levels (K-12 and educator preparation programs), subject areas (e.g., literacy, STEM, social studies), and instructional contexts (e.g., multilingual learners, varied pacing needs, classroom modalities). Prompt types ranged from brief, single-turn requests to complex multi-turn exchanges involving iterative revision, format adaptation, and pedagogical elaboration.

This dialogic structure preserves the evolving nature of instructional decision-making, as educators can refine or redirect the AI's outputs to better align with student needs, lesson goals, or contextual constraints. Each message was accompanied by metadata, such as conversation ID, timestamp, platform feature of origin, and user ID, which enabled conversational-level analysis as well as reconstruction of broader dialogic arcs.

Ethical Considerations All data were collected under the platform's terms of use, which explicitly notified users that their de-identified data could be used for platform quality improvement and research studies. Personally identifiable information (PII) was removed prior to analysis. The study protocol was reviewed and determined exempt under educational

research guidelines by the IRB. All data analysis using commercial LLMs occurred using enterprise accounts via API interactions with explicit contractual terms preserving the privacy of the data submitted and confirming that it would not be used by the LLM provider for model training.

2.3.2 Open Coding for Inductive Theme Discovery

The first stage of analysis involved open coding to inductively explore the education-related themes embedded in educator-AI interactions. Drawing on established qualitative coding practices [39, 33], this phase aimed to surface recurring patterns and instructional constructs that would inform the development of a hierarchical codebook. Prior research has demonstrated the potential of LLMs to support conceptualization and theme discovery across qualitative corpora.

A central methodological consideration in qualitative analysis is defining the unit of analysis, the segment of data that is interpreted, labeled, and compared. However, there is no single best approach. For example, Wang et al, (2025) [166] and Shin et al., (2025) [143] provided entire documents to LLMs to preserve context, while Gao et al., (2025) [61] pre-clustered raw qualitative text based on semantic similarity, and DePaoli (2024) [44] segmented documents into smaller chunks due to token limitations. In our case, single educator-AI conversations could reach considerable length, sometimes exceeding 500 messages. To balance contextual richness with token constraints, we began with triadic units, each consisting of an educator prompt (T1), the AI response (A1), and the educator's follow-up (T2), if present. These trios were treated as self-contained instructional interactions, enabling us to surface dialogic patterns while preserving coherence and instructional intent. We employed claude-3-sonnet-20240229 (Claude 3 Sonnet) for the analysis. To assess the validity of using the model for theme discovery in this context, we began with a small test sample of 256 randomly selected trios. Each trio was submitted individually to the Anthropic API, accompanied by a structured prompt. Because no session memory persisted across API calls, each trio was interpreted independently, eliminating carryover effects. To guide LLM-assisted open coding, we developed a prompt that posed the following

open-ended questions:

This prompt was designed to elicit information characterizing the diverse ways educators use the AI tool in their daily work. The table below presents examples drawn from the raw LLM outputs:

```
You are an expert educator. Your task is to identify k-12 educational components in the
↪  chunk of educator-AI dialogue.
You are provided with a request, its response, and corresponding follow-up request if
↪  available.
   Request: {request}
   Response: {response}
   Follow-up Response: {follow-up response}

1. Content Subject Domain Area:
       Output the specific subject and its domain areas if applicable (e.g.,
       ↪  'Math/Geometry,' 'Science/Physics science/motion,' etc.).

2. Pedagogical strategies:
Output specific strategies if applicable (e.g., 'Formative Assessment,' 'Group Work,'
↪  'Student Discourses', 'Connect to Real-world' etc.)

3. Instructional Strategy (1-5 scale):
       - 5: The command clearly addresses instructional strategies.
       - 4: Touches on strategies but lacks depth.
       - 3: Implies strategies without a clear focus.
       - 2: Vaguely references instructional strategies.
       - 1: No instructional strategy mentioned.

 4. Request for AI Tool Support
       Output the specific support the teacher is asking for.

   Format your responses for each item in short phrases. Provide the results in the strict
   ↪  JSON structure:
   {{
       ``Request'': {request},
       ``results'': {{
           ``Content subject domain area'': ``subject area'' or null,
```

```
        ``Pedagogical strategies'': ``strategy'' or null,
        ``Instructional Strategy'': score,
        ``Request for AI Tool Support'': ``support'' or null
    }}
}}
```

This prompt was designed to elicit information characterizing the diverse ways educators use the AI tool in their daily work. The table below presents examples drawn from the raw LLM outputs:

Table 2.1: Examples of educator requests and corresponding LLM open coding results

Requests	**Content domain**	**Pedagogical strategies**	**Score**	**AI tool support**
I need to think of an introductory unit idea for a college and career prep class for a 9th grade class	College and career preparation	None	2	Idea generation
The unit needs to be based on skills needed to be successful for high school	Study skills	None	3	Lesson design
How do I prank my co-workers?	None	None	1	Idea generation
Can you make a rubric for the mind-mapping activity above?	Social studies / cultural studies	Rubric creation	3	Rubric creation
Thank you for providing the learning target K.NS.3. This will help me craft an authentic assessment aligned with that standard.	Math / counting and cardinality	Formative assessment	3	Assessment creation

After the LLM generated analytic summaries for all trios, two researchers independently reviewed the outputs to assess their accuracy and interpretive validity. Discrepancies were resolved through discussion and collaborative memo writing, followed by the consolidation of analytic dimensions and an initial categorization of labels. Prior studies [143, for eg.] also have incorporated interfaces like ChatGPT to assist in categorizing LLM-generated labels.

However, at this stage with the relatively small size of our trial sample, we completed this step manually to deepen our human understanding of the data.

Based on this validation and synthesis process, we refined our labeling categories to include extended contextual information, such as references to student needs and pedagogical frameworks mentioned in educator prompts, and identified recurring instructional tasks as examples. These refined categories were then introduced to the LLM to support inductive coding at scale in subsequent phases of analysis.

2.3.3 Axial Coding for Conceptualization and Codebook Development

To develop a fuller understanding of educator-AI conversations, this stage scaled up the analysis to support comprehensive codebook construction. We expanded the dataset to 9,352 conversational trios, retaining the trio format to preserve dialogic and instructional context while deepening theme discovery. This phase drew on the logic of axial coding as described by Strauss and Corbin (1998), wherein researchers relate codes by identifying underlying conditions, contextual factors, actions, and consequences, though our goal was not to build theory but to translate emergent patterns into a reproducible framework suitable for large-scale annotation of educational dialogue. The final output of this stage was a preliminary codebook that captures the instructional and professional complexity embedded in educator-AI interactions. Throughout this process, researchers continued memoing to document rationales for code merges, disagreements, and iterative refinements.

In this phase, the LLM was instructed to freely generate responses for each analytic dimension. This approach reduced variability in label naming across similar concepts, while preserving the opportunity for surfacing new or unexpected categories. The complete prompt used is included in Appendix A2.

The example LLM output for the request *"How can I create culturally relevant activities for my Grade 3 fractions lesson plan, where a majority of my students are Spanish speakers"* is shown below:

```
{
  ``Content Subject Domain Area'': ``Math/Fractions'',
  ``Grade Level'': ``Grade 3'',
  ``Contextual Information'': ``Majority of students are Spanish speakers'',
  ``Instructional Tasks'': ``Incorporate Real-World Example'',
  ``Instructional Strategy'': 3,
  ``Learning Progression'': null,
  ``Pedagogical Frameworks'': null,
  ``Request for AI Tool Support'': ``Create culturally relevant activities''
}
```

As in current open coding phase, two researchers randomly selected a sample of 1,000 LLM-generated outputs to review for accuracy and interpretive quality. Discrepancies in conceptualizing labels were resolved through collaborative meetings and memo-based synthesis. Validated outputs for each analytic dimension were then returned to the LLM to propose label unifications and identify common representative terms. Next, researchers clustered LLM-generated labels within each analytic dimension by grouping them based on semantic similarity, pedagogical intent, and frequency of use. For example, terms such as "differentiated strategies" and "differentiated support" were merged, as were "create quiz" and "generate formative assessment." However, adjacent but distinct constructs like "inquiry-based learning" and "project-based learning" were intentionally preserved. This human-in-the-loop process ensured that the emerging codebook maintained conceptual clarity while staying grounded in authentic instructional language.

The outcome of axial coding was a preliminary codebook with pre-determined labels for categories including Educational Context, Content Focus, and Instructional Practices. The Pedagogical Frameworks dimension remained open-coded to allow for emergent variation. To assess the coverage and efficiency of this initial codebook and to explore the relationships among categories and labels, it was incorporated into the prompt design for the next stage of analysis.

2.3.4 Selective Coding for Structured Prompting and Codebook Refinement

With the preliminary codebook established, the selective coding phase focused on validating, refining, and structurally formalizing the codebook for broader application. This stage

marked the transition from open-ended annotation to a closed-vocabulary approach, wherein predefined labels were applied to a new set of 11,539 educator-AI message trios. Unlike earlier phases that emphasized exploratory theme generation, the LLM was now prompted to select from existing codebook categories and return outputs in structured JSON format, enhancing consistency and enabling efficient downstream parsing (see Appendix A3 for complete prompt). We employed claude-3-5-haiku-20241022 (Claude 3.5 Haiku) for the remainder of the analysis, as this model became available during the study period.

Following precedents in the field, we experimented with various iterative refinements to the prompt design [64, 188, for eg.]. Structured prompt design and XML-based encoding emerged as an effective technique for managing longer prompts involving extensive label sets. To ensure precision in label assignment and output structure, the full codebook was embedded into the model's system prompt using XML-style formatting. Each prompt included category definitions and corresponding subcodes, and the model was instructed to return results using schema-compliant tags. This design limited responses to the specified codes, reducing output variance and improving extractability. For instance, a prompt might define allowable values between <Context> and </Context>, such as Classroom Setting/Reluctant Participants or Engagement, Student's Needs/Special Education (SpEd), Student's Needs/English Language Learners (ELL). Despite such approaches to output consistency, challenges remained in the LLM's application of predetermined labels. Variation persisted through the use of synonyms and abbreviations (e.g., "ELL," "ELLs," or "MLL" in place of the canonical label "English Language Learners (ELLs)"). These inconsistencies required additional manual review and post-processing normalization.

This phase served three primary goals: (1) to evaluate the completeness and usability of the preliminary codebook at scale; (2) to assess the model's reliability in applying structured codes under constraint; and (3) to identify underrepresented, misclassified, or emergent instructional patterns requiring refinement. To support the first goal, we incorporated validation through "Other" annotations and human review. To maintain inductive flexibility, the model was permitted to select "Other" when no predefined option was appropriate, with an accompanying textual justification. Human researchers manually reviewed all "Other" responses to determine whether recurring patterns merited new codes. For ex-

ample, emergent instructional contexts such as homeschooling and afterschool programs, initially emerged as labels in the "Other," were later incorporated into the codebook under the Student Needs and Context domain.

These steps mirror the selective coding process in traditional qualitative analysis and follow the procedures described by [136]. The final output of this phase was a structured and refined codebook, developed during the coding process to represent meaningful categories and their interrelationships through code pruning and consolidation. To enhance analytic tractability, the research team evaluated code frequency and semantic distinctiveness across the structured annotations. Low-frequency codes that lacked conceptual clarity or were frequently misapplied were either merged with related categories or removed. For example, "Gallery Walk" was folded into broader subcategories within Collaborative Learning, and the underutilized "Learning Progression" category was eliminated due to redundancy and inconsistent usage. These pruning decisions were guided by code frequency thresholds, coder feedback, and alignment with existing pedagogical literature.

Through iterative reading, comparison, and contrast of coded outputs, the team arrived at a hierarchical codebook structure (**Domain** *Category* "Item"; formatted as **Domain**, *Category*, and "Item" in the rest of the text) that organized broad professional responsibilities and discourse markers, while also capturing fine-grained pedagogical practices and instructional tools. The resulting codebook achieved a balance of domain relevance, internal coherence, and scalability for large-scale analysis. It comprises six top-level domains, 18 mid-level categories, and 68 fine-grained instructional items. The full codebook is included in Appendix B, and Table 2 below provides illustrative examples of categories and items under each domain.

In addition to structured categorical codes, open-text metadata fields were retained for subject area, grade level, and explicitly mentioned pedagogical frameworks, allowing for future stratified or filtered analyses. This stage marked the completion of the codebook development process, enabling full-dataset annotation and ensuring that downstream analyses would remain meaningful within educational contexts.

Table 2.2: Domains in coded educator-AI conversations and example categories and items

Domain	Example category	Example item
Instructional practices	*Critical thinking and inquiry*	"Historical thinking"
	Explicit teaching	"Modeling problem solving"
	Differentiation and accessibility	"Tiered scaffolding"; "Multilingual learner support"
Curriculum and content focus	*Lesson planning*	"In-class activity design"
	Technology integration	"Multimedia use"
Student needs and context	*Classroom setting*	"Low-tech environments"
	Student profile	"Special education (IEP/504)"
Assessment and feedback	*Assessment*	"Generate formative assessment"
	Feedback	"Generate feedback to students"
Professional responsibilities	*Professional development*	"Prepare PLC or workshop materials"
	Communication	"Administrative messaging"
Other	*Discourse continuity*	"Format modification"
	Non-educational queries	"Non-educational queries"

2.3.5 Deductive Coding for Annotation with Established Codebook

The final phase of the methodology involved large-scale, deductive annotation of 104,529 individual educator and AI messages drawn from 13,071 new conversations that occurred in May 2025 (UTC), using the finalized codebook developed in Section 3.4 (see Appendix A.2). The complete prompt is included in Appendix A4. At this stage, the focus shifted from theme discovery and contextual exploration to precise labeling of individual educator requests. This transition also marked a shift in unit of analysis, from triadic interactions to single-turn coding. Providing one message at a time to the LLM, rather than full trios, enabled more consistent code assignment and supported downstream analyses focused on educator intent and the potential argumentative role of AI during interactions. Such refinement of the analytic unit is a common methodological decision in qualitative research

when scaling from conceptual exploration to structured, replicable annotation [39, 136]. Although this shift reduced local conversational context, we preserved coherence through platform metadata. Each message remained linked to adjacent turns via conversation IDs and timestamps, allowing for reconstruction of pedagogical arcs when necessary.

Fewer than 3% of model outputs failed to conform to the expected JSON structure and required manual formatting corrections or recoding. In particular, Claude 3.5 Haiku occasionally struggled with ambiguous or generic educator prompts, such as "continue," "yes," or "add more," which lacked sufficient semantic specificity. In these cases, the model often failed to recognize the messages as substantive prompts. Such instances were flagged and resolved during post-processing.

Comparison of Human Coders and Claude 3.5 Haiku We trained two human educational researchers as the coder in this stage to validate the results using a random sample of 1,000 messages. After aligning their understanding of the codebook, both coders independently reviewed and assessed annotation quality by either manually coding messages and comparing them to the LLM results or by verifying the LLM outputs. Interrater agreement for educator requests was moderate (Cohen's kappa = 0.59, F1 = 0.70), but lower for AI responses (kappa = 0.30, F1 = 0.46). These findings highlight the inherent difficulty of labeling multi-intent educational messages, particularly in AI-generated content where instructional purpose is often implicit or diffuse. Our analysis, as reported in the Findings section, primarily focuses on educator requests.

Claude's agreement with human coders was comparable in magnitude: kappa = 0.38 (requests) and 0.34 (responses), with F1 scores of 0.40 and 0.38, respectively. These results suggest that Claude 3.5 Haiku performs at a level similar to that of independent human coders in complex educational contexts. However, the overall variability also underscores the interpretive ambiguity of the task. To further explore this dynamic, coders were next asked to verify Claude's assigned labels rather than code independently. In this verification condition, both coders identified whether Claude's labels were justifiable and explicitly pointed to message segments supporting each one. Strikingly higher agreement emerged: human-human agreement rose to 0.81 (requests) and 0.83 (responses), while human-Claude

agreement increased to 0.83 and 0.87, with F1 scores reaching 0.84 and 0.88. These improvements likely reflect anchoring effects, as the verification task reduces variability by giving coders a predefined reference point. While this result does not imply that LLMs can independently produce highly reliable annotations, it does indicate that LLM-assisted verification may offer a pragmatic path for scalable annotation with expert oversight.

Some of the most common disagreements between Claude and human coders, both in blind and verification settings, stemmed from differences in interpretive granularity. Claude often applied broader, inferred labels (e.g., **Curriculum and Content Focus** *Planning* "In-class Activity Design and Adjustment"), even when the prompt targeted a more specific instructional goal. Human coders, by contrast, tended to apply the most explicit and goal-directed label present in the message, avoiding extrapolation. This reflects Claude's tendency to infer pedagogical scope across a full prompt, which, while reasonable pedagogically, often diverged from human norms prioritizing label precision.

Another source of disagreement involved edge cases in which educators used the platform for non-instructional or personal purposes, such as requesting recipes, help with non-school scheduling, or generating life-coaching content. Though infrequent, these unconstrained uses posed challenges for both humans and models in distinguishing between labels such as Non-Educational, Real-World Engagement and Scenarios, or Industry Standards (in CTE contexts). However, these cases were rare and did not materially impact overall pattern detection.

Ultimately, the challenges of annotating this dataset reflect the dual difficulties of interpretive ambiguity and the lack of ground truth in educational discourse. Educator-AI messages often contain overlapping pedagogical goals or implicit intentions, making categorical distinctions inherently subjective. In addition, the diversity of grade levels, content domains, and instructional philosophies embedded in the data necessitated highly specialized annotators. Even among aligned researchers, interrater reliability remained moderate, underscoring the difficulty of scaling high-quality qualitative annotation without LLM-assisted strategies.

Sections 3.2-3.5 outline an LLM-assisted qualitative analytic pipeline that addresses RQ1 by demonstrating how scalable qualitative methods can reveal educator intent, instructional

patterns, and pedagogical behaviors within educator-AI conversations. This pipeline draws selectively on established qualitative coding techniques, open, axial, and selective coding, to structure a human-LLM collaborative workflow suited to large-scale educational text analysis. This approach offers a reproducible, domain-grounded method for analyzing large-scale textual data while preserving the interpretive richness of qualitative inquiry.

To ensure both scalability and interpretive rigor, we implemented a human-LLM collaborative strategy across all phases. Human educational researchers led the processes of conceptual development, codebook construction, and final interpretation, while LLMs supported theme generation, exploratory expansion, and structured annotation. This approach enabled both descriptive and in-depth qualitative analyses across instructional categories, educational contexts, and AI design patterns, forming the empirical foundation for the findings presented in Section 5. These analyses also directly inform RQ2 by elucidating educators' pedagogical intentions, professional priorities, and the types of instructional support pursued through AI interactions.

2.4 Findings

This section presents empirical findings from the deductive coding of 13,071 educator-AI conversations, as described in Section 3.5. Through descriptive statistics and qualitative thematic analysis, we examine how educators used GenAI to support their core professional tasks, including instructional planning, differentiation, assessment design, and reflective practice. The subsections that follow begin with an overview of domain- and category-level code distributions, then move into deeper thematic patterns that illustrate the instructional and professional priorities reflected in educator prompts and AI responses.

2.4.1 Overview of How Educators Use AI

To provide a foundational view of educator-AI interactions, we analyzed the corpus that consists of 104,529 single-turn messages, in which 52,255 are requests from educators, embedded within 13,071 multi-turn educator-AI conversations. Each message was annotated using a hierarchical codebook of six high-level domains, 18 mid-level categories, and 68

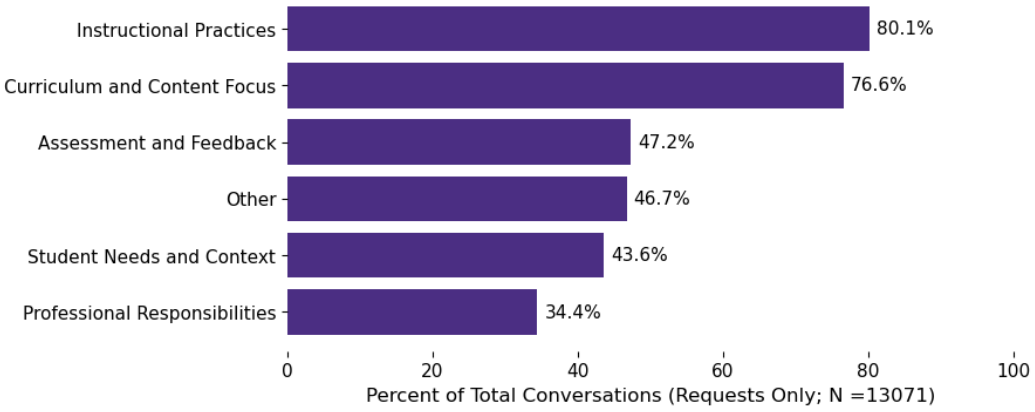

Figure 2.1: Distribution of educational domains at the conversation level. This figure shows the distribution of educational domains across conversations, where each domain is counted once if it appeared in any message within the conversation.

instructional actions (see Appendix B). Messages were multi-labeled, with over 95% receiving at least one valid code. On average, educator prompts received 1.7 codes per message, while AI responses averaged 2.3 codes, highlighting the compound instructional intent and the multifaceted nature of the assistance educators sought and received. For example, a teacher's request such as *"Design an ageappropriate inclass activity for 7th graders to learn about solving inequalities using realworld examples"* was annotated under both **Instructional Practices** / *Project-Based and Real-World Learning* / "Real-World Engagement and Scenarios" and **Curriculum and Content Focus** / *Planning* / "In-Class Activity Design and Adjustment." This dual-labeling illustrates how educators blended instructional strategies and content planning within a single prompt, underscoring the integrated nature of their work.

We first report the distribution of coded domains across the full dataset (Figure 2.1). **Instructional Practices** was the most prevalent domain, appearing in 80.1% of conversations, followed by **Curriculum and Content Focus** at 76.6%. **Instructional Practices** include categories (shown in Figure 2.2) that support diverse learners through *Differentiation and Accessibility* (36.9%), promote *Critical Thinking and Inquiry* (42.4%), and

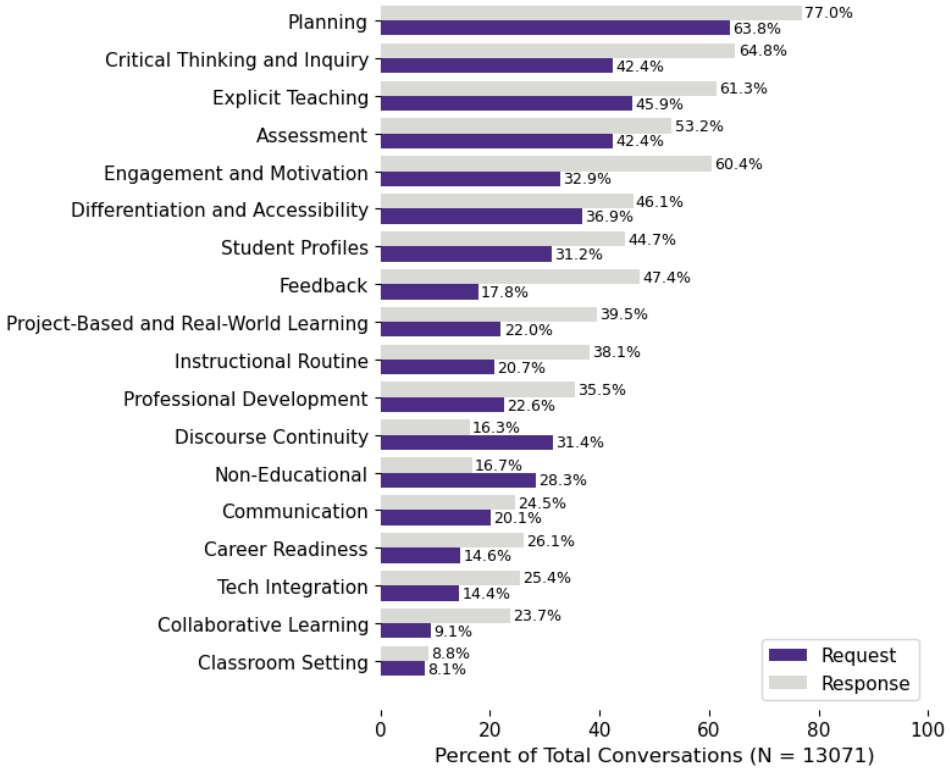

Figure 2.2: Distribution of categories at the conversation level. Each category is counted once per conversation if it appeared in any message. The figure also shows side-by-side distributions for educator requests and AI responses.

encourage engagement through *Collaborative Learning* (9.1%) and other *Engagement and Motivation* practices (32.9%). Some instructional practices also contribute to **Curriculum and Content Focus** through shaping lesson materials that guide learning. In addition to promoting critical thinking and engagement, educators used AI to create content for instructional *Planning* (63.8%), enhance *Explicit Teaching* (45.9%), and support connections to *Real-World Learning* (22%). Overall, educators treated AI as a source of pedagogical planning support across both instructional strategies and instructional materials.

Beyond instructional planning and content generation, **Assessment and Feedback** generation, evaluation, and modification also led many educators to AI tools, which were requested in 47.2% of conversations. While AI supported educators with content and prac-

tices, educators brought the knowledge about their own **Student Needs and Context** into the conversations to achieve adaptive educational goals. Approximately 44% of conversations had educators explicitly mention *Classroom Settings* (8.1%), *Student Profiles* (31.2%), or the career or academic readiness paths (14.6%). In addition to in-class student-centric purposes, AI also supported educators in fulfilling **Professional Responsibilities** (34.4%) on various forms of *Communication* (20.1%) and *Professional Development* (22.6%).

The distribution in our findings aligns with educators' usage of AI from previous studies and surveys [187, 56, for eg.], suggesting that educators primarily used AI for instructional planning and content design, while also leveraging the tool for differentiated student support, assessment, and administrative tasks. Notably, nearly half of the conversations (46.7%) included requests classified under the Other domain, reflecting the frequency of prompt modifications and diverse AI usage that extended beyond educational domains.

Appendix C presents the 25 most frequent instructional actions at the item level, with side-by-side distributions for educator requests and AI responses. Appendix D includes detailed descriptions of each category across domains (D1-D3) and a frequency matrix of category-to-category co-occurrence (D4). In the next section, we turn to qualitative insights that illustrate how educators used AI tools to address their instructional and professional priorities.

2.4.2 Instructional and Professional Priorities in Educator-AI Conversations

To address RQ2: *In what ways do K-12 educators use GenAI tools to carry out key instructional and professional responsibilities in real-world educational contexts?*–we conducted a deeper analysis of educator-AI interactions across the dataset. Qualitative analysis at both the category and item levels revealed nuanced patterns of educator intent, highlighting both common and varied ways in which educators leveraged AI to support pedagogical planning, differentiation, assessment, and administrative tasks.

Instructional Planning and Content Design *Planning* activities appeared in 63.8% of educator requests (Figure 2.2). In doing so, educators brought a wide range of pedagogical and instructional design considerations to the conversations. They used AI extensively to

generate lesson plans, structure activities, align content with standards, and adapt pacing for diverse classroom contexts.

Qualitative analysis of educator prompts revealed consistent patterns of instructional design thinking. For example, one middle school educator asked, *"Design a 45-minute lesson on comparing primary and secondary sources for a 7th grade history class,"* to which the AI responded with a three-phase plan: a warm-up (source identification task), guided practice (source comparison with sentence frames), and formative assessment (exit ticket requiring categorization and justification). The educator then followed up with a request to simplify the activity for students reading two grades below level, demonstrating how planning and differentiation were interwoven in real-time interaction.

Beyond generating content, lesson planning and activity design often involved diverse pedagogical strategies to promote learning and engagement. In another example, a 10th grade biology teacher requested, *"Create a lesson on human impact on ecosystems with active learning,"* and the AI returned a full 55-minute session featuring a case study jigsaw, infographic analysis, and a Think-Pair-Share closing task. To tailor the delivery, the educator later asked the AI to extract just the opening scenario for a slide deck and to generate an image for the infographic analysis, demonstrating strategic reuse and repackaging of AI-generated materials.

Across the corpus, AI responses frequently included anticipated misconceptions or introduced additional instructional modalities, such as visual aids, even when not explicitly requested. For instance, a prompt to *"plan a lesson on characterization in literature"* received an extended response that included peer dialogue through character roleplay, an instructional approach not mentioned in the original request. These unsolicited additions reveal a pattern of AI-initiated argumentation, in which the system not only follows educator prompts but also surfaces latent planning opportunities. In one case, an educator asked for *"a lesson plan on volume formulas,"* and the AI proposed both formula-based exercises and a hands-on box-filling activity to bridge conceptual and procedural understanding. The educator then requested a printable version, suggesting satisfaction with and continued use of the AI's proposal.

These examples underscore the capacity of AI-powered tools to function not merely as

static content generators but as dynamic instructional design partners. Educators demonstrated agency throughout these interactions, iteratively refining AI outputs, rejecting irrelevant portions, modifying formats, and merging suggestions with their own planning logic. A common marker of instructional quality that educators sought was coherent learning progression. As AI-generated lesson plans followed the structures such as launch-explore-consolidate, educators often adapted to match specific workflows (e.g., *"Warm up, direct instruction, independent work, and formative assessment"*, *"I do, we do, you do"*). *Planning* prompts often included contextual details, such as time constraints (e.g., *"a 30-minute Friday lesson"*), grouping strategies (e.g., *"in pairs or triads"*), and learner needs (e.g., *"for ELL students"*). This layered approach illustrates how instructional planning served as both a core and anchoring practice, often initiating multi-turn workflows that also incorporated differentiation, assessment, and engagement design.

The planning behaviors observed in these conversations reflect challenges documented in pre-AI research on teacher lesson design. Studies have shown that quality instructional planning is cognitively demanding work requiring teachers to anticipate student responses, sequence content coherently, and align activities with objectives, tasks that consume substantial time beyond the school day [81, 50]. The iterative nature of educator-AI exchanges, in which teachers specified constraints, requested modifications, and refined outputs across multiple turns, mirrors the recursive planning processes described in curriculum research [132]. That educators frequently provided detailed contextual parameters (time limits, student groupings, learner profiles) suggests they were drawing on the same situated planning knowledge that research identifies as essential but time-intensive to operationalize. AI tools appear to function as just-in-time planning partners, potentially addressing barriers that teachers have historically faced in accessing responsive support during the planning process itself.

Differentiation and Student Support *Differentiation and Accessibility* for diverse learners appeared in 36.9% of educator conversations, making it one of the most frequently pursued instructional goals. Educators used AI to adapt instruction for multilingual learners, students with IEPs or 504 plans, and those performing above or below grade level.

Within this category, 24.8% of conversations involved designing "Differentiated Instructional Strategies," 11.6% included requests related to "Multilingual Learner Support," and 12.4% addressed adaptations for "Mixed Ability Learners" (see Appendix C Figure A.1). Common requests included vocabulary support, sentence frames, tiered scaffolds, visual aids, and bilingual resources, strategies that often co-occurred with planning actions and were typically embedded within broader instructional design sequences.

Qualitative analysis revealed a wide range of differentiation strategies and considerations. In one conversation, an educator asked: *"Can you rewrite this high school chemistry lab for students reading at a 4th grade level?"* The AI responded by simplifying sentence structure, adding labeled diagrams, and offering vocabulary previews with accessible definitions to demonstrate language-focused scaffolding across content areas. The educator then followed up: *"Great, now give me three guiding questions they can answer verbally,"* further promoting for student discourse while maintaining conceptual rigor.

For multilingual learners, educators often asked the AI to incorporate sentence frames or provide dual-language task versions. One prompt read: *"Add directions and some example work in Spanish for [this math task]."* The AI responded with a bilingual version (e.g., *"Step 1: Mira el ejemplo. Step 2: Hazlo t mismo."*), which the educator accepted without modification. Mixed ability and tiered scaffolding were also prominent. Several educators explicitly requested multiple versions of the same activity, as in: *"Give me three versions of [this assignment] for on-level, below-level, and advanced students."* The AI produced leveled texts with adjusted question complexity and vocabulary supports. In other cases, educators initiated the differentiation and asked the AI to modify it further, as in: *"Here's what I give my IEP students. How would you modify it for general ed?"* These interactions reflect a bidirectional design flow, in which educators used AI not only to generate but also to test and refine differentiated scaffolds.

Across subject areas, AI often suggested differentiation even when not explicitly prompted. For example, when asked to generate a 6th grade ELA assessment, the AI independently proposed adaptations for MLLs. These instances illustrate AI's potential to both respond to and extend educators' goals for accessibility, offering scaffolds that enhance instructional inclusivity. Importantly, educators did not always adopt AI-generated scaffolds wholesale.

Many made refinements, substitutions, or selective adoptions, reflecting their professional judgment and fluency in differentiation strategies.

Together, these findings suggest that AI tools functioned as more than translators or simplifiers. In the category of *Differentiation and Accessibility*, they operated as adaptive design assistants embedded in instructional problem-solving. Educators brought deep knowledge of learner needs to these interactions and used AI to enrich their approaches with additional options, supports, and design structures. In this way, differentiation emerged not as a detached technical task, but as a richly contextualized and integrated component of instructional planning.

These patterns connect to a well-documented tension in educational research: differentiated instruction is among the most consistently supported approaches for reaching diverse learners, yet teachers have persistently struggled to implement it at scale [158, 75]. Pre-AI studies found that teachers often know what differentiation strategies would benefit their students but lack the time and resources to create multiple versions of materials, design tiered activities, or prepare scaffolds for varied learner needs [27]. The frequency with which educators in this study requested AI assistance for differentiation tasks, and the speed with which they could iterate on multiple versions of the same content, suggests that AI tools may help bridge the gap between teachers' knowledge of effective differentiation practices and their capacity to enact them within the constraints of daily practice.

Assessment, Rubrics, and Feedback *Assessment*-related requests accounted for 42.4% of educator conversations, while *Feedback* appeared in 17.8%. Educators used AI to "Generate Grading Rubrics" (12.8%), "Generate Formative" (27%) and "Summative Assessments" (15.8%), as well as to provide personalized "Student-Facing Feedback" (11.3%). These requests were often interwoven with other instructional content. Educators frequently prompted the AI to generate assessments based on existing or previously generated curricular materials, build rubrics aligned to those assessments, or develop feedback using the language and criteria of the rubric. In several cases, educators used AI-generated rubrics and assessments as models or work samples for students, integrating them into classroom instruction and student self-reflection routines.

Educator prompts often indicated a desire to balance precision with workload efficiency. In one conversation, a high school English teacher asked, *"Can you write feedback for this paragraph about tone? The student is struggling to develop a clear thesis."* The AI generated paragraph-level comments using asset-based phrasing (e.g., *"You've made an insightful observation about diction"*), followed by detailed guidance on structuring a clearer argument. The educator's follow-up, *"Make it simpler for students"*, reflects a common pattern: educators frequently refined AI-generated feedback to produce more concise and accessible student-facing versions, highlighting both strengths and areas for improvement. This shows how educators used AI to streamline their workflow while maintaining control over interpersonal delivery, treating the AI's output as a scaffold rather than a script.

However, analysis of this category also surfaced concerns about less reflective uses of AI, particularly in grading. Some educators requested evaluations of student work without specifying rubrics, grading criteria, or methods for consistency checking. When multiple student responses were submitted in a single conversation without aligned criteria, the risk of inconsistency and bias increased. Given the generative nature of LLMs and their variability across outputs, such practices may lead to unreliable or inconsistent evaluations [182]. If used without clear norms or human oversight, AI-generated assessments may unintentionally compromise fairness, especially in high-stakes formative or summative contexts. These findings highlight the importance of developing professional guidelines and system-wide policies to ensure the responsible integration of AI into assessment workflows.

The assessment and feedback patterns observed here resonate with decades of research demonstrating that formative assessment and timely feedback are among the most powerful influences on student learning [22, 72]. Yet this same research has documented persistent barriers to implementation: providing individualized, actionable feedback on student work is labor-intensive, and teachers consistently report that time constraints limit the quality and frequency of feedback they can offer [177]. The ways educators used AI to draft feedback, generate rubrics, and create assessment items suggest an attempt to address these documented barriers, using AI to reduce the time cost of generating feedback while retaining professional judgment over tone, appropriateness, and final delivery to students.

Professional Reflection and Administrative Tasks Though less frequent than planning and assessment codes, non-instructional professional tasks appeared in 34.3% of educator prompts, with 22.6% related to *Professional Development* and 20.1% related to *Communication*, which includes facilitating communication with families and colleagues as well as supporting administrative documentation. These uses accounted for a meaningful portion of overall educator-AI interactions and illustrate how GenAI was integrated into workflows beyond direct instruction.

Many educators frequently asked AI to support Professional Learning Community (PLC) activities, workshop planning, and instructional self-reflection. For example, one educator asked the AI to generate slide templates aligned to prompts such as *"scope of the issue"* and *"impact on the school or community"* in preparation for a presentation. Others used the tool to develop reflection prompts aligned to evaluation criteria, such as *"Explain how you differentiated the content, process, and product in the lesson"* Educators also sought support in preparing data sheets aligned with IEP goals, crafting goal-setting templates for student reflection, and developing agendas for professional learning tied to district goals and teacher evaluation frameworks. In one instance, a teacher requested ideas for a *"walk-up song"* to open an educator leadership training, illustrating the diverse ways AI was used to support professional development.

Educators also frequently asked AI to generate structured materials and components of official documentation, such as APA-formatted reference lists, standardized action plans, IEP documentation, and progress-monitoring tools. AI supported routine but time-intensive or repetitive tasks, helping educators maintain consistency in documentation while freeing up time to focus on instructional or relational aspects of their work.

In addition, educators used AI to draft sensitive or strategic communications with parents and caregivers. They prompted the AI to revise or review emails related to student progress, absences, safety drills, or behavioral concerns. For instance, one teacher asked the AI to revise a family-facing message about a student's risk of not passing a course, incorporating both concrete action steps and an empathetic tone. Others sought help with questions like: *"How would you report and discuss grades with individual students' caregivers?"* and *"Make it interesting for families and outside audiences."* In these cases, educators used

AI deliberately to maintain professionalism, clarity, and relational trust in high-stakes or emotionally sensitive interactions. They also leveraged AI's multilingual capabilities to engage families and communities in their native languages, expanding access and inclusivity in communication.

Research on teacher professional learning has emphasized that effective development is sustained, job-embedded, and responsive to the specific challenges teachers encounter in their practice [47, 42]. However, access to such support has historically been uneven, with many teachers, particularly in under-resourced or rural settings, lacking regular access to instructional coaches or collaborative planning time [62, 24]. The professional development and reflection uses observed in this study suggest that AI may serve as an on-demand resource for the kind of just-in-time, practice-connected support that research identifies as valuable but often inaccessible.

2.5 Implications and Discussion

This study provides empirical evidence that K-12 educators are integrating GenAI into a wide range of professional activities beyond content generation, including assessment development, differentiated instruction, reflective practice, and student support planning. Our findings align with prior research on educator AI use, while extending the scope of observed practices across a broader K-12 educator population. The results show that, while some uses remain exploratory, others reflect deliberate instructional intent and align with multiple domains of professional teaching frameworks. Educators engaged with AI in ways that extended their planning processes, incorporated diverse student needs, and foregrounded their own professional judgment. In this section, we discuss implications of educator-AI interactions that highlight both the opportunities and limitations of using generative tools for instructional support.

Instructional Precision Dependent on Prompt Quality. AI-generated content were more effective when educators provided structured prompts that included clear instructional goals, student context, and pedagogical intent. Requests such as *"Help me align this 5th grade science project with NGSS standards"* or *"Generate 10 ideas to connect linear equations to real-world situations relevant to Gen Alpha students interested in TikTok in*

my 8th grade class" elicited more tailored and instructionally relevant responses. In contrast, prompts such as *"Generate a lesson plan for [concept]"* or *"Make this better"* often produced vague outputs that required additional rounds of refinement.

Unprompted Instructional Additions by AI. In several cases, the AI introduced relevant instructional elements, such as success criteria, differentiation strategies, or feedback loops, even when they were not explicitly requested. For example, a prompt for an *"exit ticket"* was sometimes met with a multi-step formative assessment plan. While some educators incorporated these additions, others ignored them or asked the AI to simplify its output. These moments suggest the potential for AI to surface latent instructional practices, though uptake varied and may reflect differences in educator confidence or perceived relevance.

Boundary-Setting and Critical Judgment. A small but meaningful subset of prompts revealed educators' efforts to critically evaluate or push back against AI outputs. Some educators sought confirmation or validation (e.g., *"Is this answer right?"*), while others questioned the reliability of AI-generated information. For instance, educators asked for sources or links to verify claims and ensure that content came from reputable references. These interactions demonstrate educators' awareness of potential misinformation and their commitment to maintaining professional standards. They also underscore the importance of educator judgment and the need for AI systems that support critical engagement and transparency.

Overall, while GenAI demonstrated utility for supporting instructional planning, content generation, and self-directed on-the-job learning through augmentation [100], its effectiveness depended heavily on educators' prompting skills, professional experience, contextual knowledge of student learning needs, and the strategic framing of instructional goals [191]. Findings from this study suggest that generative tools may enhance efficiency and amplify professional expertise, but they do not inherently guarantee pedagogical quality or equity without clear educator intent and critical oversight.

Methodological Implications for Qualitative Research at Scale. Beyond the empirical findings, this study offers a methodological contribution for researchers seeking to analyze large volumes of educational text data. The LLM-assisted pipeline developed

here demonstrates that established qualitative coding techniques, open, axial, and selective coding, can be adapted to work in collaboration with LLMs while preserving human interpretive control. Critically, the approach positions LLMs as assistants rather than autonomous analysts: human researchers retained responsibility for conceptual decisions, codebook development, and interpretive synthesis, while LLMs supported theme surfacing, label extraction, and structured annotation at scale. This division of labor offers a pragmatic path for educational researchers facing large datasets that would be impractical to code manually, without sacrificing the interpretive depth that qualitative methods provide. The pipeline's four-phase structure, from inductive theme discovery through deductive annotation, is generalizable to other domains where researchers seek to combine qualitative rigor with scalable analysis.

Collectively, these implications point toward a future in which GenAI systems in education function not merely as standalone productivity tools, but as instructional collaborators and thought partners. Their value will depend not only on output speed or volume, but on how effectively they support educator reasoning, student-centered design, and responsible professional practice.

2.6 Future Directions and Limitations

This study provides a large-scale, empirical account of how K-12 educators engage with GenAI tools for authentic professional tasks. However, several limitations constrain interpretation and suggest important directions for future research.

First, while the dataset reflects real-world educator-AI conversations, it is limited to a single AI platform and a specific user base. Generalizability across tools, districts, and educator populations remains an open question. Future studies should compare usage patterns across platforms and demographic contexts, especially in disaggregate to reflect usage within underserved or resource-constrained settings, to assess equity and inclusion in AI-supported teaching.

Second, this study does not include direct measures of classroom implementation or student learning outcomes. While prompt content and follow-up behavior offer indirect signals of professional intent, they cannot substitute for classroom observation, instructional

artifacts, or student achievement data. Mixed-method implementation studies that integrate these dimensions will be essential to establish how GenAI impacts teaching practice and learning at scale.

Third, the dataset offers a static snapshot of educator-AI interaction patterns. Longitudinal analysis is needed to understand how educator AI competency develops over time, what kinds of professional support accelerate growth, and whether early engagement with AI correlates with increased instructional diversity or improved planning coherence. Future research might incorporate in-platform telemetry, survey data, or educator interviews to explore trajectories of adoption and learning.

Fourth, the study focused on observations of usage, not values or ethics. As educators delegate aspects of planning, feedback, or communication to AI, deeper inquiry is needed into how educators negotiate professional responsibility, bias mitigation, and trust or doubts in AI, in addition to how AI augments or diminishes human expertise through responding. Future design-based research should foreground educator agency in setting boundaries for appropriate AI use and evaluating AI generated content, particularly in formative assessment, student support, and equity-sensitive contexts.

Moreover, this study is also subject to several methodological limitations besides what has been mentioned in section 3.5.1. First, although the human-LLM interactive coding pipeline enabled systematic analysis at scale, it remains difficult to fully validate a substantial portion of the labeled data, particularly AI-generated responses, which were often long, multifaceted, and embedded numerous instructional practices both implicitly and explicitly that require professional expertise in various subject areas and educational lenses. To mitigate this, our primary analyses focused on educator requests, which were typically shorter, more directive, and easier to verify and interpret reliably.

Second, although we explored the use of open-weight models such as Qwen, LLaMA, and Mistral to support scaled annotation in later phases of the study, none achieved sufficient reliability for unsupervised coding. As a result, our final annotations relied exclusively on proprietary LLMs (e.g., Claude 3.5) outputs. Although commercial models showed promise in supporting annotation tasks, their closed-source nature and limited transparency restrict reproducibility and interpretability. Model benchmarking results are included in

the Appendix for transparency, but they are reported as exploratory and do not inform the study's core findings. We invite future research to leverage open-weight models in combination with codebooks of similar complexity to promote reproducibility and enable the sharing of comparable coding results with various open-weighted models.

In sum, while this study demonstrates broad engagement and pedagogical promise in educator-AI collaboration, realizing the full potential of generative tools will require continued investment in research-practice partnerships, transparent design, and longitudinal learning analytics. GenAI may amplify professional capacity, but its educational impact will ultimately depend on how it is situated within human-centered systems of teaching and learning.

2.7 Conclusion

This study makes two primary contributions. First, it introduces a replicable LLM-assisted qualitative analytic pipeline that enables researchers to analyze large-scale educational text data while preserving human interpretive control. Second, it provides one of the first large-scale empirical accounts of how K-12 educators use GenAI in authentic professional contexts, based on analysis of over 14,000 educator-AI conversations.

Methodologically, the pipeline demonstrates how established qualitative coding techniques, open, axial, and selective coding, can be adapted to work collaboratively with LLMs without ceding conceptual authority to the model. Researchers guided theme discovery and developed and refined a grounded codebook, while LLMs were used to read through large corpus, quickly surface code candidates, test category boundaries, and refine hierarchical structures through example-based prompting. In later phases, LLMs were tasked with deductive coding based on the researcher-finalized codebook, allowing researchers to focus on higher-order interpretation, edge cases, and validation. This division of labor, with human researchers retaining an overarching supervisory role throughout the process, enabled the team to maintain analytic rigor while extending qualitative methods to a dataset of over 100,000 coded messages, a scale that would be impractical through manual coding alone. The four-phase structure offers a generalizable approach for educational researchers and others working with large textual corpora who seek to combine qualitative depth with scalable

analysis.

Empirically, educator-AI interactions revealed a wide range of instructional and professional uses, with evidence of pedagogical intent. Educators used GenAI to brainstorm ideas, adapt materials for specific learners, design assessments, generate rubrics, and support reflective practice. Many requests blended multiple instructional goals, for example, combining lesson planning with accessibility considerations, highlighting the multi-layered nature of daily educational work. Some educators exercised critical judgment by revising, rejecting, or validating AI outputs, indicating active human oversight. However, the effectiveness of AI use varied and was contingent on the clarity of educator prompts, contextual specificity, and adherence to responsible use practices. In many instances, AI introduced unprompted but instructionally relevant strategies, suggesting a form of human-AI argumentation in which educators initiated instructional goals and AI extended them through implicit pedagogical reasoning.

Taken together, and with national surveys showing growing teacher AI use, these findings point to the need for professional development that emphasizes hands-on, reflective learning and supports educators in integrating AI within their instructional workflows as embedded tools to enhance content and pedagogy instead of merely an external add-on. The study also highlights methodological tensions between scale and interpretability, particularly when using LLMs to annotate large volumes of content-rich texts. As educational AI systems continue to evolve, future research should examine the specific educational practices educators engage in with AI and the daily routines shaped by these interactions, further linking technological advancements to user behavior and their impact on educational outcomes.

Chapter 3

TEACHER-AUTHORED PROMPTS FOR CONFIGURING STUDENT-AI DIALOGUE: K-12 CLASSROOM IMPLEMENTATION

Abstract. GenAI has rapidly entered instructional and learning settings as a teaching assistant or AI tutor. However, less is known about how pedagogical intent connects to the learning generated within these systems, especially when student-facing AI dialogues are fine-tuned through teacher orchestration in live classrooms. This study examines a classroom deployment of a "Teacher-Authored Student Dialogue" (TASD) system, which enables teachers to author both a teacher-to-AI setup prompt (instructional scaffold) and a student-facing conversation starter to launch AI-mediated classroom discussions. We analyze a multi-subject pilot conducted in Spring 2025, involving 20 participating teachers (16 of whom implemented the system), across 39 classrooms and 77 TASD settings, yielding 1,479 student-AI conversations with 878 unique students. Using platform logs, LLM coding with human validation, and post-study teacher interviews (N=10), we characterize teacher authoring choices and link them to enacted student-AI interaction outcomes. Teachers predominantly authored highly specific tasks targeting higher-order thinking, with 92% at Depth of Knowledge (DOK) Levels 2–3 [171]. In deployment, student-AI conversations were largely aligned with instructional intent: 71% were fully on-track, and fewer than 1% were substantially off-track. However, a persistent design-enactment gap emerged for cognitive demand: 38% of conversations under-reached the teacher-targeted DOK level, approaching 50% when targeting DOK 3. The study also shows that explicit finish lines in the prompt reduced the DOK gap by 0.22 levels ($p < .001$), and "no direct answers" guardrails reduced AI final-answer rates by 8.5 percentage points. These findings position teacher-authored prompt layers as critical orchestration levers that translate pedagogical intent into structured student-AI dialogue, underscoring both their promise for scalable classroom integration and the need for additional supports to reliably sustain higher-order

reasoning during enactment.

3.1 Introduction

Large language models (LLMs) have made conversational support broadly accessible, accelerating interest in AI tools that assist educators and students at scale [88]. In current educational use, AI is commonly configured as (a) a backstage teacher assistant for planning and materials generation [119, 45], (b) a teaching assistant surfaced through teacher-facing dashboards that suggest instructional moves based on classroom activity or tutoring-system data [3, 77], or (c) a direct-to-student learning platform in which the primary interaction occurs between learner and system with comparatively limited teacher control [99, 162]. Each configuration reassigns roles between teachers, students, and AI systems, but ultimately it should be teachers who possess the professional expertise and contextual information to determine whether and how AI contributes to instructionally meaningful learning [96].

Classroom orchestration refers to teachers' real-time coordination of students, activities, tools, and instructional decisions to sustain coherent progress toward learning goals under practical constraints [52, 131]. In live classrooms, teachers must manage time, norms, safety, and alignment to curricular goals under real-time conditions [98]. When an additional AI agent is introduced into this workflow, teachers need to coordinate across students, tools, and teaching moments while sustaining pedagogical intent, potentially increasing managerial and cognitive load as they synthesize and translate output from multiple planes into actionable next steps [80]. Prior orchestration research emphasizes that classroom technologies are most viable when they incorporate direct support for coordination work across simultaneous learners and activities [48, 156, for eg.]. From this perspective, AI systems designed for synchronous classroom use must be legible, bounded, and supervisable in ways that foreground teachers' ability to configure, interpret, and oversee learning occurring within the technology, rather than emphasizing solely on underlying model capabilities [135, 101].

Prior research has demonstrated how AI-powered tools can support scalable teaching and learning, such as orchestration layers in MOOCs [68] or real-time awareness features in intelligent tutoring systems [29]. At the same time, studies of teacher-AI argumentation highlight that instructional value arises not from model fluency alone, but from how

educators and AI systems coordinate roles, information, and decisions in the flow of classroom activity [77, 76]. Still, much of this work focuses either on enabling teaching or on shaping learning, with fewer investigations into how teacher can configure AI systems to mediate the connection between pedagogical intent and student learning within those systems [160, 80, 169].

This study examines teacher-student-AI facilitation through a teacher-authored student-AI dialogue tool that treats the teacher-authored prompt as an instructional design lever for shaping student discourse at classroom scale. The implemented system, Colleague AI Classroom Teaching Aide (TASD), enables a teacher-configured, student-facing AI conversation that preserves teacher intent while supporting individualized dialogue across students. A teacher authors (i) an AI-facing shared prompting layer that conditions the AI's role, discussion topics, scaffolding moves, and interaction constraints for the whole class or specific groups (e.g., multilingual learners, advanced students) and (ii) a common student-facing opening instruction that launches the discussion activity. Rather than treating AI chat as an autonomous tutor, TASD frames student-AI interaction as a configurable classroom routine that scales teacher-designed dialogue structures that preserve teachers' instructional intent and oversight. This approach aligns with AI system designs that rely on system prompts to guide and stabilize human-LLM interactions [69, 168], examining how these systems are enacted, configured, and orchestrated within real K-12 classroom conditions.

Guided by two research questions:

- **RQ1 (Teacher authoring):** What do teachers build when given control to author and configure student-AI dialogues at classroom scale?

- **RQ2 (Classroom orchestration outcomes):** How do teacher-authored designs shape what actually happens in student-AI conversations, specifically task alignment, cognitive demand realization, student interaction moves, and AI scaffold adherence?

We conducted a multi-site pilot study in Spring 2025 to examine classroom deployment and interaction outcomes. Data sources included teacher-authored prompts, full student-AI conversation traces (including instances of non-use), and semi-structured teacher interviews.

Together, these support an analysis of the full instructional chain: from teacher-authored instructional frames and entry scaffolds → to student participation and student-AI interactions → to outcomes such as task alignment, realization of cognitive demand, and adherence to AI use safeguards, interpreted through teachers' accounts of orchestration strategies and constraints.

To address these questions, we conducted a multi-site pilot study during Spring 2025 across 16 teachers and 878 students. The study examines the full instructional chain: from teacher-authored instructional frames and entry scaffolds → to student participation and student-AI interactive outcomes. Drawing on descriptive and qualitative analysis of platform logs, student-AI conversations, and teacher interviews, we analyze how instructional intent is realized in practice. Findings highlight high task alignment, but a consistent gap between intended and actual cognitive demand, alongside evidence that prompt features such as finish lines and safety guardrails influence both student engagement and AI behavior. This study contributes to learning-at-scale and AI-in-education research by offering an empirical account of student-AI dialogue at classroom scale that foregrounds the connection between teachers' instructional intentions and student learning within AI systems, reinforcing transparency and teacher agency by making AI behavior controllable and interpretable to teachers.

The remainder of this paper proceeds as follows. We situate the study in prior work on classroom orchestration, human-AI complementarity, prompt-based interaction design, and student-facing conversational systems; describe TASD and the teacher authoring workflow; detail the pilot context and analytic approach; report results organized by research question; and conclude with implications, limitations, and directions for future research.

3.2 Related Work

This study is built up three interrelated strands of research: (1) how classroom technologies support real-time orchestration and learning at scale, (2) how task framing and discourse norms shape student participation and cognitive demand, and (3) how prompting functions as an interaction design mechanism in educational applications of GenAI. We bring these threads together by treating teacher-authored prompts as orchestration tools, designed not

only to guide student-AI interactions but to preserve pedagogical intent, structure participation, and support cognitive rigor under authentic classroom conditions.

3.2.1 Classroom Orchestration and Teacher-Facing Support at Scale

Classroom orchestration describes how teachers coordinate multi-layered activity in real time, including pacing, participation, discipline, and alignment of tools and social arrangements to instructional goals [48, 49]. Classroom technologies sometimes may even intensify orchestration load by introducing new decision points and monitoring demands, especially when many learners are active simultaneously [125]. Learning-at-scale research has therefore emphasized abstractions and teacher-facing representations that make complex activity tractable with tools like orchestration graphs and real-time dashboards [68, 156].

In AI-supported classrooms, teacher-facing awareness and intervention supports are repeatedly positioned as prerequisites for instructional integrity. Co-designed orchestration tools suggest that teacher-AI complementarity depends on actionable visibility and teacher-in-the-loop control, rather than autonomous system behavior alone [77, 76]. This human-AI co-constructive perspective emphasizes that AI should augment teacher capabilities rather than replace teacher judgment, providing recommendations, surfacing patterns, or handling routine interactions while teachers retain authority over instructional decisions [14, 2].

3.2.2 Conversational Agents for Learning

Conversational agents for learning have a long history, from ELIZA-style interactions through dialogue-based intelligent tutoring systems demonstrating that structured dialogue can support explanation, reasoning, and learning [165]. Meta-analyses and systematic reviews report that intelligent tutoring systems produce meaningful learning gains, particularly when dialogue is structured around productive pedagogical moves such as prompting for explanation, providing feedback on reasoning, and adapting to student knowledge states [51].

Reviews of educational chatbots report growing adoption but uneven grounding in learning theory and frequent reliance on self-reports rather than rigorous outcome measures, highlighting that instructional value depends on task design and implementation conditions

[178, 179, 93]. In language learning, systematic reviews find increased practice opportunities and engagement, with substantial variability by scaffolding, learner characteristics, and task structure [78, 108]. The consistent message is that conversational technology is not inherently effective; outcomes depend on how interaction is designed and implemented.

GenAI-based learning systems amplify these interaction design stakes because unconstrained dialogue can drift, overload learners with verbosity, or collapse into answer-seeking [13, 184]. Work such as RECIPE illustrates how structured guidance and role-setting can stabilize learner-LLM interaction toward pedagogically meaningful moves [69]. Studies of students using ChatGPT for learning reveal both productive uses (explanation seeking, feedback on drafts) and problematic patterns (copying, over-reliance on AI responses) that depend on how interactions are framed and constrained [46, 58].

3.2.3 Task Design, Discourse Quality, and Enacted Cognitive Demand

Teachers' framing of discussion tasks has long been shown to shape how students participate, whether dialogue becomes brief answer exchanges or collaborative reasoning [32, 116]. Norms like "justify your answer" or "build on a peer's idea" serve as participation contracts that elevate the quality of talk [133, 114]. Moment-to-moment prompts, such as revoicing, pressing for reasoning, or inviting elaboration, help sustain rigor and coherence during execution [63, 4].

The link between design and thinking is well documented in the Depth of Knowledge (DOK) framework [171]: tasks designed for higher cognitive demand are more likely to elicit strategic or extended thinking, though implementation mediates this relationship [148, 74]. Previous studies consistently highlight the "decline" problem, occurring when rigor weakens during enactment despite high initial task design [25]. Structured participation routines can mitigate this gap and support equitable engagement [139, 10], especially when discussion norms are made explicit in the instructional framing [183].

3.2.4 Prompting as Interaction Design for GenAI

A growing body of work treats prompting not only as a technical mechanism for steering LLM outputs, but as an *interaction design* practice that shapes what end users do, how assistance is sequenced, and whether conversational activity remains aligned to instructional intent [189, 175]. In this framing, prompt choices redistribute responsibility among teacher, student, and AI system: they specify the AI's stance (e.g., coach versus answer engine), the expected epistemic work (e.g., justify, cite evidence, revise), and the completion criteria that turn open-ended chat into a bounded instructional routine [77, 88, 160]. In education-oriented dialogue settings, Socratic prompting has been explored as a mechanism for shifting the AI from answer-giving to question-asking, reinforcing the claim that prompt-level decisions encode an epistemic stance that can either support or undermine productive student participation [105, 126]. Emerging scholarship further frames prompt engineering as a practice competency for educators, a skill to effectively align prompts to learning goals, constraints, and classroom norms [164].

Recent GenAI learning systems operationalize this idea through *prompt layers* and structured guidance intended to increase reliability and pedagogical alignment. For example, RECIPE demonstrates how a hidden instruction layer can set an instructional role and couple it with structured guidance aligned to writing goals, shaping learner-LLM interaction toward instructionally meaningful behaviors [69]. MathChat and similar systems show how carefully designed prompts can scaffold mathematical reasoning through structured multi-turn dialogue [168]. Complementary research in prompt programming proposes reusable strategies: explicit role specification, decomposition into steps, requirements for justification, and use of exemplars, which increase controllability and reduce drift across turns [134, 175, 106].

Prompting also functions as a mechanism for responsible-use design. Guidance for GenAI in education consistently emphasizes safeguarding, transparency, and human oversight, particularly for minors [160, 88]. Prompt constraints can instantiate these priorities as *productive constraints* that discourage copying and entertainment use while preserving student agency: prohibiting direct answers, requiring attempts before hints, demanding evidence

or reasoning through feedback, and bounding turn length to reduce cognitive overload and improve classroom synchrony [73, 87].

This study sits at the intersection of these prior studies. From prompt-as-interaction-design research, we take the insight that prompt structure shapes student behavior and AI responses. From classroom discourse research, we adopt the concern with cognitive demand and the gap between intended and enacted rigor. From orchestration research, we foreground teacher-facing requirements for viability and the importance of bounded, legible, and supervisable interactions. From conversational agent research, we recognize that dialogue technology is not inherently effective, outcomes depend on design and implementation. And from responsible AI scholarship, we incorporate attention to safeguards and teacher oversight [160, 115]. Aiming for actionable evidence for implementing AI as third agent in the classroom, this study examines the full pathway from teacher authoring through student interaction to orchestration outcomes.

3.3 System: Teacher-Authored Student Dialogue (TASD)

TASD enables teachers to author an instructional frame that bounds student-AI interaction while providing lightweight supervision supports needed to manage many parallel conversations. The system was designed around three goals informed by orchestration theory and responsible AI guidance. **G1. Preserve teacher intent**: enable teachers to define goals, norms, constraints, and success criteria so student-AI interaction remains aligned to instructional purposes [48, 76]. **G2. Support classroom orchestration**: provide monitoring and intervention supports so teachers can navigate and manage many student chats concurrently without excessive overhead [130, 156]. **G3. Scaffold responsible use**: encourage productive struggle and self-advocacy through developmentally appropriate interaction constraints and safety supports [160, 88]. In combination, these goals position TASD as *configurable*, *bounded*, and *supervisable* for live classroom routines.

The TASD is constructed using a three-layer structure. At the base layer, a foundational LLM (e.g., GPT-4o) provides general inference capabilities. The second layer is a system-level prompt designed by educational researchers that fine-tune the model's role and behavior (e.g., acting as an AI tutor for K-12 learners, emphasizing guidance and reason-

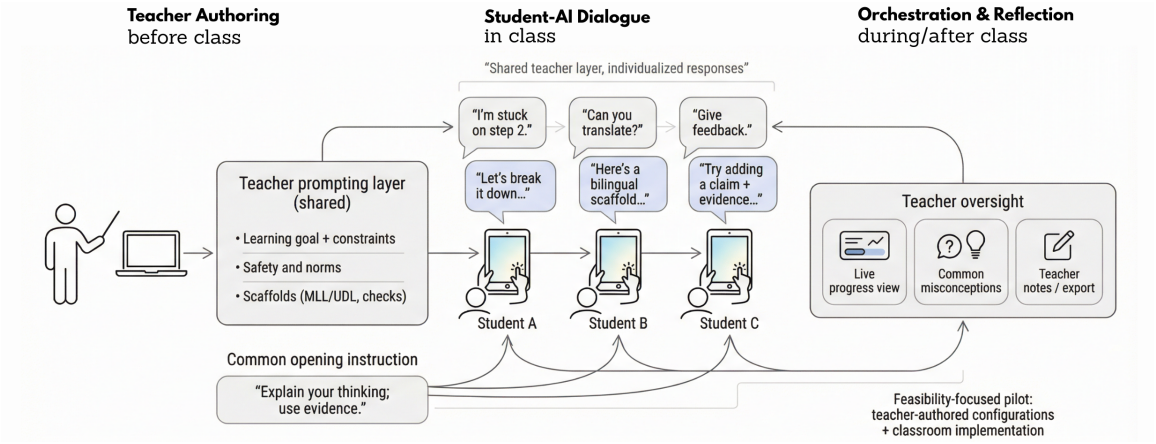

Figure 3.1: Teacher-Orchestrated Student-AI Dialogue at Classroom Scale

ing rather than direct answers) for general educational contexts. The top layer consists of teacher-authored prompts, which encode local pedagogical goals, task structures, and classroom settings, enabling teachers to shape how AI supports learning in ways that respond to the specific needs of their students. In addition to the *teacher-authored prompting layer*, a TASD activity is launched in the classroom through a shared *student-facing opening instruction* that aims to frame the activity routine and communicates expectations for participation. Students then engage in individualized conversations, with the AI adapting to student responses while remaining anchored to the teacher-authored instructional frame.

Workflow is despicted in Figure 3.1. Teacher authoring schema and the recommended live-classroom workflow in Appendix B.4, including teacher authoring schema in Table B.1 and live classroom workflow with additional description of monitoring and closure routines.

3.4 Study Deign

The pilot study took place during Spring 2025 (April 2-June 10) in Washington state. Twenty-one in-service teachers were recruited through district partnerships across four public school districts and one independent school. Teachers received hourly compensation for participation. Of the recruited teachers, 16 actively implemented TASD in their classrooms during the pilot period, generating the analytic sample. Non-implementing teachers either

did not complete technical onboarding (2/21) or chose not to deploy student-facing activities (3/21) due to lack of alignment to their curriculum during the study window.

Participating teachers spanned mathematics, science, English Language Arts, social studies, world languages, and multiple Career and Technical Education (CTE) areas, serving middle school (grades 6–8) and high school (grades 9–12) students in class sizes of 15–33. Through these implementations, the pilot reached 878 unique students across 39 classrooms. Students were not contacted by the research team. Student was not required to use AI to complete learning tasks with equivalent alternatives provided. Student demographics reflected the participating districts' populations, though individual-level demographic data were not collected.

3.4.1 Timeline and Implementation Supports

The pilot followed a seven-week professional learning arc designed to reflect realistic early-adoption conditions while providing sufficient guidance for teachers to enact TASD as a classroom routine without disrupting curriculum progression. In Week 1, teachers were introduced to the study and the TASD system, including privacy protocols and research procedures. In Weeks 2-3, teachers engaged in guided practice authoring teacher prompting layers and student-facing starters, supported by peer feedback and researcher consultation. Beginning in Week 4, teachers implemented TASD activities in their classrooms and iterated on prompt designs based on observed student responses. Each teacher was asked to complete at least two implementations to support iterative improvement and to report feedback and reflections via exit tickets. In Week 7, we conducted end-of-pilot reflection sessions and one-to-one usability interviews with a focal group of teachers (N = 10) to contextualize implementation conditions and interpret observed patterns.

Throughout the study, teachers received onboarding, guided authoring practice, and weekly researcher-facilitated meetings for troubleshooting and implementation sharing. These supports were designed to scaffold initial adoption without prescribing specific prompt or activity designs, preserving natural variation in teacher authoring choices.

3.4.2 Ethical considerations

The study protocol received the University of Washington Institutional Review Board approval. The district/school Memoranda of Understanding (MOU) specified privacy protocols. All data collection followed district privacy policies, with conversation data stored securely and analyzed in de-identified form under privacy agreements.

3.4.3 Data Collection

The analysis draws on four complementary data sources that enable tracing from the teacher-authored scaffold to student interaction and orchestration outcome:

Classroom meta data Classroom-level meta data including classroom ID, teacher ID, subject area, grade level, district, and enrolled student count. This registry defines the population denominator for adoption analyses and provides context variables for stratified comparisons.

Discussion meta data Metadata and content for each TASD activity configured by teachers, including discussion ID, discussion title, teacher-authored prompting layer (the prompt conditioning AI behavior hidden from students), student-facing conversation starter text, creation timestamp, and associated classroom ID. During the pilot period, teachers successfully implemented 77 TASD activities across 39 classrooms.

Conversation transcripts Complete message-level logs of student-AI interactions, including message content, sender (student/AI), timestamp, and linkage to discussion and classroom identifiers. The pilot generated 1,479 student conversations comprising 36,162 individual messages. Each conversation represents one student's complete dialogue session within a specific discussion assignment. All conversation logs were deidentified and labeled with randomly generated user IDs.

Teacher interviews Semi-structured interviews (N = 10) provided qualitative context for interpreting observed patterns from participating teachers' perspectives, particularly

regarding orchestration practices, prompting strategies, and perceived barriers to in-class AI implementations [85]. Interview protocol is included in Appendix B.3.

3.4.4 Measures

We developed coding schemes at two levels: discussion-level teacher-authored prompts and conversation-level student-AI interactions. Given the scale (94 prompt created; 1,479 conversations), we employed LLM-based coding with human validation.

Prompt measures (discussion-level). Teacher prompting layers were coded for task specificity, target Depth of Knowledge (DOK) [170], presence of a clear finish line, scaffolding strategies, epistemic framing, constraints or guardrails, and AI role assignment. Full rubric and the LLM coding prompt are provided in Appendix B.2.1. Prompts were coded using GPT-5.2 (reasoning effort=none; temperature=0); Human validation were preformed on all coding results (100%).

Conversation measures (conversation-level). Conversations were coded for task alignment, demonstrated DOK (highest DOK level reached in the given conversations), student interaction moves, AI behavioral adherence (e.g., Socratic questioning, stepwise guidance, answer withholding), safeguard violations, and starter quality. Full rubric and the LLM coding prompt are provided in Appendix B.2.2. Conversations were coded using GPT-4.1-mini (temperature = 0). A stratified subset of 50 conversations was independently coded by a human researcher; agreement was 87% on high-stakes codes (task alignment, demonstrated DOK, AI safeguard violations).

Linking coded data. Prompt codes were merged with conversations using discussion identifiers. Approximately 98% of conversations (1,449 of 1,479) linked successfully; 30 conversations lacked valid discussion identifiers[1] and were excluded from downstream analyses, though retained in the deployment descriptives in Section 3.5.1. Of the 1,449 linked

[1]If the discussion owner (teacher) deleted the discussion, the associated metadata were removed from the database.

conversations, 1,362 were included in the DOK-gap analyses. The remaining conversations were excluded due to uncodable data, either the teacher-authored prompts did not specify a target DOK level, or the student-AI interactions were too brief to assess demonstrated DOK reliably.

Adoption metrics. We demonstrated voluntary participation at multiple levels: discussion-level participation rate (participants/student count), classroom-level participation rate (unique participants/enrolled students across the pilot), and conversation volume per discussion. Non-participation (opt-out) was treated as an outcome rather than missing data.

3.4.5 Analytic Approach

Analyses addressed each research question while accounting for nested data (students within discussions within classrooms within teachers). We report descriptive deployment patterns, test prompt feature–conversation outcome associations, and estimate an OLS model predicting the DOK gap with robust standard errors clustered by discussion [180]:

$$DOKGap_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_i + \varepsilon_i. \tag{3.1}$$

We also examine safeguard-related behaviors overall and stratified by teacher-authored guardrails, and use interview data to contextualize patterns in adoption and interaction outcomes.

3.4.6 Qualitative Procedures

We conducted semi-structured one-to-one interviews ($N = 10$) to contextualize and interpret the quantitative findings. Using a grounded theory approach, we inductively surfaced mechanisms that could explain adoption patterns and outcome differences through teachers' perspectives. Interviews probed teachers' instructional intent, classroom scenarios, decision points around when and how they used the system, and interpretations of what "worked" or "didn't work" in specific episodes for diverse learners. Full interview protocol is included in Appendix B.3. We used an iterative qualitative coding process that compared interview

segments across teachers and refined themes over multiple passes, then used these themes to interpret and bound the quantitative relationships.

3.5 Results

Beginning with the adoption patterns under voluntary implementation, we present findings in this section organized by research question: what teachers authored when given control over student-AI dialogue configuration (RQ1) and how teacher-authored designs shaped enacted interaction outcomes including task alignment, cognitive demand realization, and AI safeguard adherence (RQ2).

3.5.1 Implementation Overview

The outcome of the study demonstrated AI-supported scalability in diverse learning environments within various classroom contexts. Sixteen teachers implemented 77 TASD activities in 39 classrooms, generating 1,479 student-AI conversations that involved 878 unique students during the study period (April 2-June 10, 2025). Table 3.1 summarizes implementation characteristics.

We extracted 94 teacher-authored setup prompts from pilot accounts, and all were coded. However, only 77 prompts had associated student conversation data during the pilot period; the remaining 17 prompts ($94-77=17$) were created by teachers but never engaged by students. These missing cases likely reflect instances of no student participation; for example, when teachers tested prompts but deemed them unsuitable for classroom implementation.

Classroom-level participation rates averaged 85.0% (median = 88.9%) among implemented TASD activities, indicating high voluntary uptake when teachers assigned AI discussions. However, this aggregate figure masks substantial heterogeneity. Participation varied considerably across teachers and classrooms. The most active teachers generated substantial conversation volumes (200+ conversations across multiple discussions; Appendix Table B.3), while others deployed fewer TASD activities or saw more modest participation. Teacher-level contexts (subject, grade level, school culture, prior technology use) likely influenced adoption patterns. Moreover, English Language Arts represented the largest deployment (708 conversations, 48%), followed by Science (301 conversations, 20%) and World Language

Table 3.1: Implementation summary statistics

Metric	**Value**
Teachers (implementing)	16
Classrooms	39
Prompt Designed	94
Prompt Implemented (TASD activities/Discussions)	77
Conversations	1,479
Unique students	878
Conversations per discussion (mean)	19.2
Conversations per discussion (median)	19.0
Conversations per discussion (range)	1-98
Messages per conversation (mean)	23.3
Messages per conversation (median)	14.0

(171 conversations, 12%). The distribution in Figure 3.2 reflects organic classroom adoption rather than experimental assignment, with subject-area variation providing natural contrasts across learning environments.

The variation likely reflects differences in teacher framing, assignment integration (graded vs. optional), timing within the school year, and student motivation, factors that characterize authentic classroom implementations. Participation also varied substantially across discussions (range: 1-98 conversations within given TASD activities), indicating heterogeneous adoption patterns likely reflecting differences in teacher framing, assignment integration (graded vs. optional), timing within the school year, and student motivation, factors that characterize authentic classroom implementations.

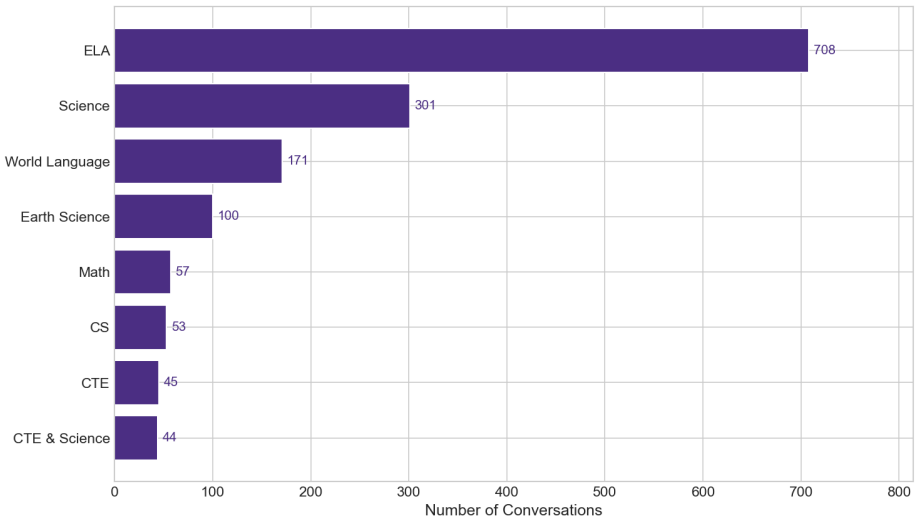

Figure 3.2: Distribution of pilot conversations across subject areas.

3.5.2 RQ1. What Teachers Build: Prompt Authoring Patterns

We analyzed 94 teacher-authored prompts characterizing the design space when teachers are given control over student-AI dialogue configuration. All prompts were coded using the validated rubric described in the section 3.4.4 using prompt in Appendix B.2.1, with 100% human verification. Sample coding results are shown in Appendix Table B.1.

Task Design and Cognitive Demand Targets Teachers predominantly authored highly specific tasks. 80.9% of prompts provided detailed, concrete task descriptions, with only 16.0% using moderately specific prompts and a single prompt (1.1%) minimally specified. This pattern suggests that when given authoring control, teachers gravitate toward structured, well-defined activities rather than general open-ended explorations. A complete set of use cases derived from prompts are included in Appendix B.7.

Teachers targeted predominantly higher-order thinking in their designs. Nearly half of prompts (48.9%) targeted DOK 3 (strategic thinking), requiring students to engage in reasoning, analysis, or problem-solving that goes beyond routine procedures. An additional 42.6% targeted DOK 2 (skills and concepts), while only 5.3% targeted DOK 1 (recall/re-

production). A single prompt (1.1%) targeted DOK 4 (extended thinking). Two prompts with unidentifiable DOK levels (e.g., "Introduction" or N/A, no valid prompt content were input). This distribution indicates that teachers attempted to leverage AI dialogue for cognitively demanding work rather than basic drill-and-practice activities. Examples of coded prompt DOK levels are shown in Appendix Table B.2.

The unused prompts skewed slightly toward DOK 2 (10 of 17), suggesting that higher-demand prompts (DOK 3) were proportionally more likely to be implemented (53% of used prompts vs. 49% of all authored prompts). The single DOK 4 prompt was never implemented by teacher or engaged by students. The near-absence of DOK 4 targets also likely reflects teacher-perceived practical constraints of single-session AI conversations, in particular, when the classrooms were new to the system. Complex reasoning and extended thinking typically requires sustained engagement across multiple sessions or project-based work that exceeds typical classroom discussion session duration [57].

Finish Lines and Task Boundaries Explicit finish lines, such as clear endpoints and success criteria defining task completion, appeared in 64.1% of prompts. This relatively high prevalence suggests teachers recognized the importance of defining what "done" looks like for students and AI, though more than one-third of prompts left task completion ambiguous. As subsequent analyses reveal, finish line presence proved consequential for cognitive demand realization.

Scaffolding Strategies Teachers employed diverse scaffolding strategies to guide AI behavior during the student-AI interactions with their pedagogical authority, though with considerable variation. Overall, 58.5% of prompts included at least one explicit scaffolding instruction, such as stepwise guidance, Socratic questioning, hints before answers, etc., while 41.5% provided no scaffolding guidance to the AI. **Stepwise guidance** that instructs the AI to break problems into sequential steps rather than presenting holistic solutions was the most common scaffolding approach, appearing in nearly half of prompts (47.8%). **Socratic questioning**, which aims to direct the AI to guide through questions rather than statements, appeared in 37.0% of prompts. Appendix Table B.4 shows the scaffolding strategy

prevalence in teacher prompts. Notably, more restrictive scaffolds that enforce productive struggle were uncommon. Only 5.4% of prompts explicitly required students to attempt problems before receiving help, and the same proportion required hints before providing direct help on problem solving [2].

Most teachers used scaffolding sparingly. 41.5% prompts included no explicit scaffolding, 30.9% included one scaffold type, 20.2% included two, and only 7.4% combined three or more strategies. This suggests that teachers avoided over-constraining AI behavior, while may also indicate that unfamiliarity with AI-powered tools leading to underutilization of available scaffolding mechanisms.

Epistemic Framing Epistemic framing instructions directing AI to guide students toward specific reasoning practices appeared less frequently than scaffolding. Only 42.6% of prompts explicitly ask AI to perform epistemic framing elements, such as compare alternatives, use evidence, justify claims, etc., while 57.4% included none. **Explaining reasoning** was the most common epistemic request (33.7%), followed by comparing alternatives (19.6%) and using evidence (17.4%). **Justifying claims** was least common (12.0%), despite its centrality to academic argumentation. The majority of prompts (57.4%) relied on the AI's default behavior to elicit reasoning rather than explicitly specifying epistemic expectations. Table B.5 shows epistemic framing prevalence in teacher prompts.

Guardrails and Responsible Use Guardrails including explicit restrictions on AI behavior to scaffold students' responsible use appeared in only 29.8% of the prompts. When present, guardrails were typically singular; no prompt included more than one guardrail type. The most common guardrail, **"no direct answers"** (19.6%), explicitly instructed the AI to avoid providing final solutions, directly addressing the prevalent academic integrity concerns over student AI use. Although teachers were aware that the tool included an underlying system prompt discouraging direct answers, this duplication suggests both the salience of the concern and a lack of confidence that a single system-level prompt would

[2]Teachers were aware of the general system prompt for the TASD system imposes not to provide direct final answers to students

reliably constrain AI behavior. Help students to maintain **respectful tone** requirements appeared in 9.8% of prompts, primarily in contexts involving sensitive topics or younger students. Only one prompt included explicit **academic integrity** policies, and none addressed **privacy** with the awareness of the existence of MOUs between the platform and the school districts. This suggests teachers may have trusted these safeguards as part of the system defaults or may not have perceived them as requiring explicit specification in the design of AI-mediate classroom discussions. Appendix Table B.6 shows guardrail prevalence in teacher prompts.

Role and Persona Assignment One-third of prompts (33.0%) did not specify an AI role [3], leaving persona to system defaults. Among specified roles, **tutor** (21.3%) and **coach** (19.1%) were most common, suggesting teachers conceived the AI primarily as a supportive guide. **Peer** roles (10.6%) were notable, indicating some teachers experimented with positioning the AI as a collaborative partner rather than an authority figure. Figure B.3 show the complete distribution of AI roles assigned in the prompts.

Prompt Authoring Patterns

When given control to author AI-assisted dialogue prompts, teachers in this pilot: (1) prioritized task specificity (81% highly detailed); (2) targeted higher-order thinking (92% DOK 2-3); (3) specified clear endpoints in a majority of cases (64%); (4) used scaffolding selectively (59% with scaffolding, most commonly stepwise guidance and Socratic questioning); (5) underutilized epistemic framing (only 43% present); and (6) applied guardrails sparingly (only 30% present). These patterns reveal a design space in which teachers consistently emphasize task clarity and cognitive ambition, practices rooted in traditional instructional design, while showing greater variability in how they facilitate student-AI interactions and establish behavioral constraints for the AI, both of which are novel considerations introduced by the AI's presence as a third agent in the classroom.

[3] Teachers were aware that the general system prompt set the default AI role as a tutor.

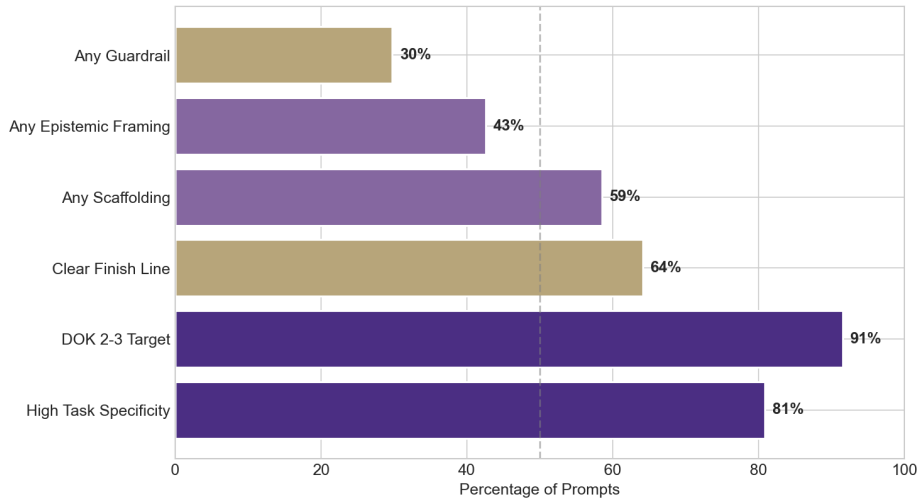

Figure 3.3: Summary of teacher prompt authoring patterns across key dimensions (N=94).

3.5.3 RQ2. How Teacher-Authored Prompts Shape Enacted Interactions

After decomposing teacher-authored prompt content, we now examine how teacher-authored facilitation shaped what actually happened in student-AI conversations, addressing three distinct outcomes, in terms of task alignment (staying on the intended topic), rigor alignment (achieving target cognitive demand), and responsible AI use safeguard adherence. This set of findings utilized coding results on student-AI conversations, of which details were described in the section 3.4.4 using coding prompt in Appendix B.2.2.

Task Alignment: Students Stayed on Track

A central concern in deploying student-facing AI is whether students remain focused on intended learning objectives or drift toward entertainment, answer-seeking, or off-topic exploration. Our analysis reveals that teacher-authored prompts can potentially bound student-AI interactions within intended instructional parameters. **71.0% of conversations remained fully on-track** moving toward teacher-specified objectives, with an additional 28.3% partially on-track (some deviation but substantial relevant engagement). Only 0.5% of conversations deviated substantially from the topics stated in the teacher prompt,

and 0.2% were coded as unclear, as students' chats related loosely to the assigned topic but showed no detectable progression or engagement with the core concept.

The low off-track rate ($< 1\%$) indicates that concerns about AI-mediated learning systems enabling unfocused or non-instructional conversations [122, 194] may be resolvable through structured in-class implementation that preserves teachers' pedagogical intent by enabling the authorship of purposeful prompts within the AI system and clearly framing the activity within established classroom routines.

Among conversations exhibiting some deviation (Appendix Figure B.2), the most common patterns were **shortcut/answer-seeking behavior** (15.1% of all conversations) and **off-topic exchanges** (13.1%). Copy-paste-like inputs, which potentially indicating students submitting work for the AI to complete, appeared in 6.9% of conversations. These deviation patterns provide actionable targets for prompt refinement. Answer-seeking behavior, while understandable given students' efficiency motivations, suggests prompts could better anticipate and productively redirect such requests through explicit guardrails and alternative scaffolding strategies.

Teacher-Level Variation On-track rates varied substantially across teachers, ranging from 53.3% to 79.5%. Computer Science teachers achieved the highest alignment rates (79.5% and 64.3%) with the smallest DOK gaps (0.00-0.13), while Career and Technical Education showed the lowest alignment (53.3%) and largest gaps (0.95). These patterns likely reflect disciplinary differences in task structure, student populations, classroom settings, and prompt design conventions rather than teacher effectiveness per se.

Rigor Alignment: The Under-Targeting Challenge

While task alignment was generally strong, **rigor alignment proved more challenging**. We examined rigor alignment as the gap between teacher-targeted DOK levels and student-demonstrated DOK in conversations. A positive gap indicates under-targeting, students performing below the intended cognitive demand level. Our analysis revealed systematic under-targeting. **37.7% of conversations showed students performing below target DOK**, while 52.9% achieved alignment and 9.4% exceeded targets (Appendix Figure B.4).

Table 3.2: Under-targeting by teacher target DOK level

Target DOK	Mean gap	% under-target	N
DOK 1 (Recall)	−0.58	0	67
DOK 2 (Skills/Concepts)	+0.10	28.7	471
DOK 3 (Strategic Thinking)	+0.63	46.0	824

The mean DOK gap was +0.39 (SD = 0.81), indicating that students, on average, demonstrated cognitive engagement approximately 0.4 levels below teacher design intentions. The relationship between target DOK and under-targeting reveals a striking pattern: when teachers aimed higher, students more often fell short. Thus, under-targeting increased with target ambition (Table 3.2).

When teachers targeted DOK 1 (recall/reproduction), students consistently exceeded expectations, the mean gap was negative (−0.58). However, when teachers targeted DOK 3 (strategic thinking), nearly half of conversations (46.0%) failed to reach that level with many students demonstrating only DOK 2 engagement (Appendix Figure B.5). This pattern suggests a ceiling effect where higher-order cognitive demands are more difficult to elicit through AI-mediated dialogue. This finding does *not* suggest that teachers should lower expectations; rather, it highlights the need for additional facilitation, both within the AI system and the classroom, when aiming to support higher-order thinking. The pattern may also reflect students' unfamiliarity with this new engagement format, suggesting the importance of building students' AI literacy to fully realize the potential of educational technology.

Prompt features that close the rigor gap To identify prompt features associated with improved rigor alignment, we regressed DOK gap on teacher prompt characteristics. The model explained 21% of variance ($R^2 = 0.21$), with several significant predictors. **Clear finish lines** reduced under-targeting. Prompts specifying explicit task endpoints ($\beta = -0.22$, $p < .001$) were associated with smaller DOK gaps. This finding suggests that students engage more deeply when they understand what "done" looks like, a principle

aligned with goal-setting theory and formative assessment research [116, 73]. In contrast, **Epistemic framing** showed counterintuitive effects, as prompts requesting students to "explain reasoning" were associated with *larger* DOK gaps ($\beta = +0.20$, $p < .001$). Teachers who included epistemic framing were likely targeting more demanding interactions, and the framing itself may signal (but not ensure) higher-order target. The positive coefficient may reflect residual after controlling for target DOK rather than a negative effect from epistemic framing.

Table 3.3: Predictors of DOK gap (OLS regression, N = 1,362)

Predictor	β	**95% CI**	**p**
Intercept	−0.80	[−1.05, −0.56]	< .001
Target DOK	+0.50	[0.41, 0.59]	< .001
Clear finish line	**−0.22**	[−0.33, −0.12]	< .001
Epistemic: explain reasoning	+0.20	[0.10, 0.30]	< .001
Scaffold: stepwise	+0.13	[0.04, 0.21]	.003
Guardrail: no direct answers	+0.12	[−0.00, 0.25]	.054

Responsible Use and AI Behavioral Adherence

A key concern in educational AI applications is whether AI behaviors adhere to designed constraints [69]. We examined whether teacher-authored guardrails influenced AI behavior at runtime. When teachers included explicit "no direct answers" constraints in their prompts, the AI's answer provision rate dropped from 16.2% to 7.7%, a reduction of 8.5 percentage points. In the cases that AI provided final answers or solutions, contextual review suggested these typically occurred after extended scaffolding attempts rather than as first responses to student queries. This finding demonstrates that teacher-authored constraints meaningfully shape AI behavior in practice, a promising result for teacher-configurable AI system for classroom use.

We also examined whether explicit student pressure affected AI behavior. When students directly requested answers (e.g., "can you just tell me how much it is [...]," "can you rewrite it for me please"), the likelihood of the AI providing a direct answer increased significantly (26.1% vs. 12.5% baseline; $r = 0.14$, $p < .001$), representing a 13.6 percentage point rise. This suggests that student pressure can partially shift AI behavior toward less scaffolded responses. However, even under explicit pressure, the AI maintained scaffolded interaction patterns in the majority of cases. Overall, the system demonstrated substantial robustness to manipulation while retaining flexibility to provide direct assistance when pedagogically appropriate, such as after sustained effort or clear evidence of confusion.

3.5.4 From Teachers' Perspectives: Orchestration Conditions for Classroom Viability

Qualitative data from interviews ($N = 10$) and teacher reflections contextualized the prompt design and interactive patterns reported above. Across sources, teachers described AI as introducing a "third agent" that could support individualized learning only when activities were implemented appropriate, clearly and tightly framed to students, and supported by real-time teacher-facing visibility.

First, teachers emphasized *developmental fit and cognitive load* as central to engagement. When AI responses were verbose or relied on probing questions without actionable next steps, some students disengaged and even felt frustrated, particularly in higher-DOK tasks. Teachers described this as an enactment challenge, where students struggled to translate AI moves into executable actions.

Second, teachers highlighted substantial *within-class heterogeneity* in participation, with some students sustaining deep exchanges while others minimally engaged or opted out. Teachers attributed this variability to differences in students' beliefs students' beliefs about AI's social benefit or cost (e.g., environmental costs, educational equity and accessibility), learning preferences, technical readiness, AI literacy, classroom norms, and clarity of the deliverable, aligning with quantitative findings that prompts with clearer finish lines reduced rigor gaps.

Finally, teachers identified *visibility and monitoring* as prerequisites for successful class-

room implementations. Lightweight indicators of progress enabled rapid triage and reduced orchestration burden during live sessions. Both the lack of visibility into student online activities and the absence of a dashboard or summaries to support monitoring made early implementations difficult to supervise. Full qualitative findings and illustrative examples are provided in Appendix B.6.

The findings in this section characterize TASD as an orchestration-sensitive classroom activity in which quality is co-produced by teacher design, student participation norms, and the AI system's adherence to scaffold and safety expectations.

3.6 Discussion

This study examined teacher-authored student-AI dialogue in authentic K-12 classroom implementation. RQ1 shows that when teachers are given control over configuring student-AI dialogue, they treat prompt authors similar to instructional activity designs, aligning them with the same elements they emphasize in traditional classroom discussion planning. Most authored highly specific activities and targeted cognitively demanding work (predominantly DOK 2-3), consistent with evidence that task framing shapes discourse quality and cognitive demand [32, 148]. At the same time, teachers encoded orchestration supports unevenly. While frequent use of finish lines suggests attention to bounded, legible classroom routines [48], scaffolding, epistemic framing, and guardrails were applied selectively, leaving many TASD activities to rely on system-default AI behavior or in-the-moment teacher in-class facilitation.

RQ2 traces how teacher design choices translated into enacted interaction quality at scale. Across classrooms, teacher-authored prompts successfully bounded AI behavior and maintained task focus, consistent with prior work showing that teachers' instructional framing strongly predicts how student dialogue unfolds in practice [32, 111]. At the same time, a persistent design-enactment gap emerged for cognitive demand. While teachers frequently targeted higher-order thinking, students often under-achieved the intended depth, particularly for strategic (DOK 3) tasks.

Moreover, task alignment and cognitive rigor alignment functioned as distinct outcomes. Conversations could remain on-topic while engaging only superficially with content, echoing

long-standing findings that topic relevance alone does not guarantee reasoning-rich participation [116, 133]. This pattern aligns with research showing that cognitive demand often declines during enactment, even when tasks are designed at a high level [25]. In AI-mediated dialogue, orchestration is partially delegated to the system; however, current tool designs are insufficient to consistently realize targeted higher-order thinking without additional facilitation, both for teachers designing the prompts and for students enacting them.

Within this broader pattern, specific design features mattered. Explicit finish lines were consistently associated with smaller rigor gaps, reinforcing evidence that clear success criteria and task boundaries support sustained cognitive engagement [109, 48]. Teacher-authored natural-language guardrails also proved effective: constraints such as "no direct answers" measurably reduced AI answer-giving, supporting work that treats prompting as interaction design rather than a purely technical mechanism [175, 88]. However, guardrails alone did not eliminate student-driven pressure for solutions, consistent with prior studies showing that learners actively negotiate assistance levels in dialogic systems [87]. This suggests that, alongside activity framing, teacher mediation, and technical constraints, it is equally important to cultivate students' awareness of AI literacy and responsible use.

More broadly, the study demonstrates that teacher-configured AI dialogue offers a mechanism for personalized learning support at classroom scale. Teachers in this pilot used TASD to provide individualized practice and scaffolded feedback to students workin independently, while simultaneously attending to small groups or students requiring more intensive support. This orchestration pattern, AI handling routine scaffolding while teachers focus on high-need interactions, represents a practical realization of longstanding goals in educational technology to extend rather than replace teacher capacity [23].

3.7 Implications

For system design. The study results underscore the importance of authoring supports that foreground pedagogical intent. Interfaces should suggest teachers to specify finish lines, particularly for tasks targeting higher-order thinking, and offer DOK-aligned templates that embed epistemic framing, scaffolded reasoning, and guardrails by default. These directions align with orchestration research emphasizing bounded, legible, and supervisable activity

structures as prerequisites for classroom viability [130]. Teacher-facing dashboards that surface lightweight, real-time indicators of activity and conversational progress could further reduce orchestration load and support timely intervention [152, 156].

For tool development. Beyond authoring interfaces, the findings point to broader affordances of AI-mediated dialogue for classroom instruction. When teachers can configure student-AI conversations, they gain capacity to provide personalized learning support at scale, a capability that has historically been constrained by the one-to-many structure of classroom teaching [23]. In this study, teachers deployed AI dialogue activities that enabled individualized practice and feedback while they circulated, facilitated small groups, or supported students requiring more intensive attention. This configuration suggests a model in which AI serves as a targeted support mechanism for independent or paired work, freeing teacher time for direct interaction with students who need it most [162]. System developers should consider how to make such orchestration patterns more explicit and easier to implement, for example, through activity templates that clarify when AI-mediated work is appropriate for independent practice versus when teacher co-facilitation is expected. The goal is not to replace teacher-student interaction but to extend teachers' reach by handling routine scaffolding at scale while preserving teacher attention for high-value instructional moments.

For teacher practice and professional learning. The findings suggest that teachers already demonstrate strong task specificity and cognitive ambition, but may benefit from additional support in translating those goals into process-oriented scaffolds that sustain within AI-mediated systems. Research on academically productive talk and dialogic teaching shows that such facilitating moves are learnable and improve with targeted professional development [4, 63]. Framing AI prompt authoring as part of broader orchestration practice, rather than isolated prompt engineering, may better align with teachers' existing instructional expertise and support more consistent enactment.

For researcher. The systematic under-targeting of cognitive demand highlights a central challenge for educational AI systems. Addressing this challenge will likely require dynamic scaffolding approaches that adapt to student responses in real time, building on emerging work in learning science, structured conversational agents, and adaptive dialogue systems [69, 168].

3.8 Limitations and Future Directions

This study relied on observational data from a voluntary pilot, limiting causal inference and introducing potential selection bias. While observed associations between prompt features and enacted outcomes align with prior task-design and discourse research [148, 133], unmeasured classroom or teacher characteristics may also contribute. LLM-assisted coding enabled analysis at scale, but some measurement error remains, particularly in borderline DOK distinctions.

The sample was also uneven across subject areas, with English Language Arts representing the largest share of deployment (48% of conversations), followed by Science (20%) and World Language (12%). This imbalance likely reflects affordances and constraints of the chat-based interface rather than differential teacher interest. Text-based conversational systems are well-suited to language-intensive disciplines where discussion, interpretation, and written argumentation are central pedagogical activities [78, 93]. In contrast, mathematics and other STEM subjects face input modality barriers: students cannot easily express equations, diagrams, or symbolic notation in a chat interface, and teachers may perceive the tool as less applicable for problem-solving tasks that rely heavily on non-textual representations. Recent work on mathematical reasoning in conversational AI has highlighted these input constraints as a key challenge for chat-based math instruction [168]. These subject-area constraints may influence the DOK patterns observed in results. ELA tasks often emphasize interpretation, analysis, and argumentation, activities that naturally align with DOK 2-3 levels and are expressible through text-based dialogue. The relative underrepresentation of mathematics conversations limits our ability to assess whether the design-enactment gaps observed would generalize to subjects where procedural fluency and symbolic manipulation are central. Future research should examine AI-mediated dialogue in mathematics

and science contexts, potentially using multimodal interfaces that support equation input, drawing, or image-based interaction to reduce modality constraints.

The sample, though substantial, was geographically bounded to participating pilot schools, which may limit generalizability. Future work should pursue controlled experiments, disaggregate effects by subject, grade level, and/or teachers, incorporate student learning outcome measures, and track teacher and student trajectories longitudinally. Equity-focused analyses are especially important, given evidence that participation structures and discourse norms shape who benefits from AI-supported learning opportunities and AI-mediated learning environments.

3.9 Conclusion

This study offers an empirical account of classroom-scale student-AI dialogue grounded in teacher-authored prompts, demonstrating that human-centric AI systems can reflect and preserve teachers' pedagogical intent under authentic K-12 conditions. Through purposeful design, teachers were able to constrain AI behavior, support high levels of student participation, and achieve strong task alignment, extending prior work on structured conversational systems for learning [51]. Yet the persistent gap between intended and enacted cognitive demand highlights a core challenge, emphasize the needs for ensuring that student-AI interactions consistently support higher-order thinking, beyond surface-level engagement, across diverse learners and classroom contexts.

This study reinforces a view of classroom AI not as an autonomous instructional agent, but as a teacher-orchestrated system. Teachers set the instructional frame through configurable prompts, students engage within that structure, and the AI operates as a scalable dialogic partner whose behavior is shaped by teacher design. This reflects a human-AI complementarity model in which AI augments teachers' professional expertise, pedagogical intent, and instructional oversight. By tracing the full pathway, from prompt authoring to student interaction and orchestration outcomes, this study provides practical and conceptual guidance for designing AI systems that are not only technically functional, but instructionally viable and pedagogically grounded.

Chapter 4

NUDGES AS EDUCATOR-AI INTERACTION SCAFFOLDS: DESIGN PRINCIPLES AND BEHAVIORAL INTERVENTION

Abstract. GenAI is increasingly embedded in educators' everyday professional workflows, yet translating AI-generated content into instructionally coherent, classroom-ready work remains cognitively demanding under authentic constraints. This paper examines AI-powered pedagogical nudges as interaction scaffolds that support educators' decision-making within open-ended educator-AI conversations. We study a large-scale deployment in which an educator-facing AI chat surfaces up to two optional, clickable next-step nudges after each response. Using platform logs, clickstream data, and qualitatively coded instructional practices, we analyze 567,577 nudge display events across 8,153 educators and 127,196 conversations over 107 days of authentic use.

We address three questions: which system design characteristics matter for nudge effectiveness, how educators adopt nudges across conversations and over time, and how adoption relates to instructional practices expressed in AI-mediated work. Findings show that adoption is selective rather than habitual (4.04% per nudge), yet widespread at the user level (52.3% adopt at least once). Adoption peaks early in conversations, exhibits meaningful delayed uptake enabled by persistent display (10.3%), and is highly concentrated among a subset of users. We identify an executability principle—nudges that produce concrete, immediately usable artifacts and reduce effort for manually prompting are adopted at substantially higher rates than abstract or reflective prompts. A cross-feature comparison further reveals a trade-off between slightly higher immediate adoption for unconstrained suggestions and stronger retention for constrained, pedagogically grounded nudges.

We conclude that nudges function best as repeated, autonomy-preserving choice points that reduce workflow friction while respecting professional judgment, with implications for the design of human-centered, ethically grounded AI support for educators.

4.1 Introduction

GenAI systems are rapidly becoming embedded in educators' everyday professional work, supporting instructional goals spanning lesson design, assessment production, differentiation planning, resource creation, and professional reflection [190, 35, 102]. As these tools transition from novelty to infrastructure, a fundamental question emerges, whether educators can reliably use AI to advance instructional intent under authentic constraints, including limited time, incomplete information, competing priorities, and the need to balance rigor, equity, and feasibility across diverse classroom contexts [83]. Without interaction support that fits real workflows, educators risk expending substantial effort on iterative prompting, repair, and reformatting, even when AI-generated content appears superficially helpful [153, 7].

This challenge matters because educators' instructional planning and decision-making are not routine clerical tasks; they involve translating pedagogical and content knowledge into instruction that is coherent, aligned to goals, and responsive to learners [37, 50]. Such work often functions as job-embedded professional learning, as educators integrate content, pedagogy, and learner needs while navigating tradeoffs under constraint [142]. Yet these high-judgment tasks are cognitively demanding, and traditional professional development frequently fails to provide timely, situated support due to access barriers, insufficient personalization, and disconnection from day-to-day instructional decisions [47, 62]. AI-powered systems appear promising for scalable, in-the-moment assistance, but without careful design, AI-generated support may be misaligned, overly generic, or burdensome to inspect and validate [88, 140].

Nudging offers a promising design approach for supporting educators' decisions within AI-mediated workflows. Rooted in behavioral economics, nudging refers to subtle prompts that influence decisions without restricting choice [155]. Educators, like all decision-makers, operate under cognitive constraints stemming from limited time, information, and energy [84, 137]. Nudges can reduce instructional inertia by making productive next actions more salient and easier to enact, without constraining autonomy or overriding professional judgment [118, 172]. When embedded in AI systems, nudges can be dynamically tailored to conversational context, providing timely decision support around content, pedagogy, and

professional tasks [117]. Nevertheless, nudging also raises distinctive risks. If nudges feel prescriptive, opaque, or misaligned with professional values, they may be perceived as manipulative, undermining agency and trust [26, 70].

This paper examines how AI-powered nudge generation functions as interaction scaffolds in open-ended educator-AI conversations. We study a deployment in which an educator-facing AI chat interface surfaces up to two clickable next-step nudges after each AI response. Each nudge is a short, educator-facing action prompt (e.g., "Create quick assessment tools to gauge student problem-solving understanding" and "Format these cards in a printable table with clear borders for easy classroom use") that educators can click to convert into their next request. Educators can adopt nudges immediately, ignore them and continue with free-form prompting, or return to previously displayed nudges later in the conversation. This design captures realistic decision-making in open-ended use, where educators may accept, defer, or reject suggestions depending on perceived relevance and timing.

We organize our investigation around three research questions:

> **RQ1:** What system design characteristics should be considered when designing nudges for an educator-facing platform?
>
> **RQ2:** How do educators adopt and use nudges across conversations and over time?
>
> **RQ3:** How is nudge adoption related to the instructional practices and professional tasks expressed in nudges and ongoing educator-AI conversations?

To address these questions, we analyze platform conversation logs, nudge display and clickstream data, and coded instructional practices from a large-scale deployment. Our analysis encompasses 567,577 nudge display events across 8,153 educators and 127,196 conversations over 107 days of authentic platform use from September to December 2025.

This study contributes to research on human-AI interaction, educational technology, and behavioral design in three ways. First, we advance a conceptualization of nudges as *repeated, autonomy-preserving choice points and exposure* embedded in multi-turn professional workflows, distinguishing this approach from one-shot recommendation systems,

prescriptive tutoring scripts, and information-provision interfaces. Second, we provide large-scale descriptive evidence on adoption behavior, including frequency, timing, concentration, and delayed adoption patterns that reveal how educators engage with embedded decision support in educator-AI conversation responding to their authentic conditions. Third, we identify an *executability principle*, which indicate that nudges producing concrete, immediately usable artifacts are adopted at substantially higher rates than nudges requiring further elaboration, with implications for balancing workflow efficiency against pedagogical depth. Together, these contributions inform the design of AI-embedded scaffolds that aim to reduce professional friction while respecting educator autonomy.

The remainder of the paper is organized as follows. Section 2 reviews related work on educator decision-making, digital nudging, and human-centered AI. Section 3 presents a conceptual framework for nudges as choice architecture in educator-AI workflows. Section 4 describes the study context and implemented nudge design. Section 5 details data sources and analytic methods. Section 6 reports empirical findings. Sections 7 and 8 discuss implications, limitations, and future directions. Section 9 concludes.

4.2 Related Work

This study examines AI-powered nudges as interaction scaffolds within educator-AI workflows. We ground our approach in the extensive literature on nudging, tracing its development from foundational behavioral economics through digital nudging to emerging AI-empowered applications, before connecting these ideas to orchestration support and pedagogical scaffolding in educational contexts. This review establishes both the theoretical foundations for understanding nudge effectiveness and the practical considerations for designing nudges that support professional educator work.

4.2.1 Foundations of Nudging: Behavioral Economics and Choice Architecture

Nudging emerged from behavioral economics research demonstrating that human decision-making systematically deviates from rational choice models due to cognitive limitations, heuristics, and contextual influences [84, 159]. Thaler and Sunstein (2008) [155] formalized

nudging as the design of choice architecture that alters how options are presented to encourage beneficial decisions without restricting freedom. Unlike mandates or incentives, nudges preserve autonomy, users remain free to choose otherwise, while making certain options more salient, accessible, or effortless.

The theoretical foundation rests on several well-documented behavioral phenomena. *Status quo bias* and *default effects* show that people tend to accept pre-selected options, making defaults powerful choice architecture tools [137, 82]. *Implementation intentions*, concrete specifications of when, where, and how to act, bridge intention-action gaps by reducing the cognitive effort required to initiate behavior [66]. *Goal gradient* effects suggest that effort increases as completion approaches [90]. *Social proof* leverages conformity tendencies by highlighting peer behavior [36], while *salience manipulations* direct attention without changing option availability [150]. These heuristics provide a design vocabulary for constructing nudges across diverse contexts.

Critically, nudging is distinguished from manipulation by its preservation of choice and transparency of intent [70]. Ethical nudge design requires that interventions be visible, non-deceptive, and aligned with users' authentic interests rather than purely system-centric objectives [26, 151]. This ethical foundation is especially important when nudges operate in professional contexts where autonomy and expertise are central to identity and practice.

4.2.2 Digital Nudging: From Static to Dynamic Choice Architecture

As nudging moved from physical to digital environments, the design space expanded considerably. Weinmann et al. (2016) [172] introduced *digital nudging* as the application of choice architecture principles to user interface design, noting that digital systems enable more precise control over information presentation, timing, and sequencing than physical environments allow. Digital nudges can leverage interaction data to adapt dynamically, creating opportunities for context-sensitive interventions that respond to user behavior in real time [118].

This evolution introduced new design dimensions. Digital nudges can be *temporally sequenced*, appearing at specific decision points rather than remaining static. They can be

personalized based on user characteristics or behavioral patterns [86]. And they can operate through multiple modalities, text, visual cues, notifications, or interactive elements, that direct attention and reduce friction for desired actions. Evidence from various domains suggests that timing and perceived relevance matter substantially: well-timed prompts matched to user context improve engagement and action completion [118, 86].

However, digital nudging also raised new concerns. The ability to continuously adapt nudges based on behavioral data led to the concept of *hypernudging*, highly personalized, algorithmically driven interventions that may operate below conscious awareness [186, 94]. Critics argue that hypernudges can become manipulative when users cannot perceive or contest the recommendation logic, concentrating power in system designers and eroding autonomy [117, 31]. These concerns motivate design approaches that maintain transparency and user control even as personalization capabilities expand.

4.2.3 AI-Empowered Nudging: Capabilities and Distinctions

The emergence of AI, particularly large language models, has introduced new possibilities for nudge generation and delivery. AI systems can generate contextually appropriate suggestions based on conversational content, infer user intent from interaction patterns, and produce natural-language prompts that feel less formulaic than rule-based alternatives [71]. Recent work has begun exploring AI-generated nudges in various domains, including health behavior, financial decisions, and learning contexts [128, 31].

It is important to distinguish AI-powered nudges from recommendation systems, though the technologies share some characteristics. Recommendation systems typically optimize for engagement metrics (clicks, purchases, time-on-platform) and often present ranked lists of options based on predicted user preferences [154]. Nudges, by contrast, are designed to support specific beneficial behaviors rather than maximize engagement, and they operate through choice architecture rather than content recommendation [155, 172]. In educational contexts, this distinction matters: a recommendation system might suggest content a teacher is likely to engage with, while a pedagogically-grounded nudge suggests actions that align with instructional best practices regardless of predicted engagement. The normative

grounding in user benefit rather than platform metrics is definitional to nudging [150, 70].

AI-powered nudges in professional contexts face particular challenges. Unlike consumer applications where behavioral data may be abundant, professional tools often lack robust user profiles, and the heterogeneity of professional tasks limits the applicability of standardized interventions. In educator-AI workflows, perceived goals may evolve across conversation turns, making it difficult to infer appropriate suggestions from interaction history alone. These constraints shift the practical question from whether AI personalization is theoretically possible to how nudges can remain useful when generated primarily from within-session context and pedagogical principles rather than extensive user modeling.

4.2.4 Nudging in Education: Applications and Scaffolding

Educational contexts have seen growing application of nudging principles, though with distinct considerations from behavioral interventions in health or finance. Damgaard and Nielsen's (2018) [41] systematic review of nudging in education documented interventions targeting student attendance, assignment completion, course selection, and financial aid applications. The review found that lightweight interventions at decision points can meaningfully influence educational behaviors, though effect sizes vary considerably by context and target behavior.

In learning environments, nudges connect conceptually to *scaffolding*, temporary support structures that help learners accomplish tasks they could not complete independently [38]. Prompts and hints in intelligent tutoring systems function as scaffolds that support self-regulation, metacognition, and productive struggle [11]. Research on tutoring systems has shown that appropriately timed hints can improve learning outcomes, particularly when they prompt reflection rather than providing answers directly [162]. This scaffolding perspective suggests that nudges in educational AI systems should fade as users develop competence, avoiding dependency while supporting growth [38].

For educator-facing tools, the scaffolding framing takes on additional meaning. Nudges can scaffold not only task completion but also professional learning, helping educators notice pedagogical possibilities, consider alternative approaches, and translate instructional intent

into concrete actions. This positions nudges as potential mechanisms for job-embedded professional development, providing the kind of just-in-time, practice-connected support that research identifies as effective but often inaccessible [47, 42].

4.2.5 Orchestration Support and Teacher Guidance Systems

Research on classroom orchestration provides additional grounding for understanding how nudges might support educator work. Orchestration describes how teachers coordinate multi-layered activity in real time, pacing, participation structures, tool use, and alignment of social arrangements to instructional goals [48, 49]. Technologies can intensify orchestration load by introducing new decision points and monitoring demands [130, 125]. Effective teacher-facing support therefore requires abstractions and representations that make complex activity tractable [68, 156].

Human-AI collaboration research emphasizes that teacher-AI complementarity depends on actionable visibility and teacher-in-the-loop control rather than autonomous system behavior [77, 76]. AI should augment teacher capabilities, providing recommendations, surfacing patterns, handling routine tasks, while teachers retain authority over instructional decisions [14]. This perspective aligns with nudging's emphasis on preserving autonomy: effective orchestration support should make productive actions more visible and accessible without mandating particular choices or second-guessing professional judgment.

Connecting orchestration research to nudging suggests design principles for educator-facing AI tools. Nudges should be *bounded and legible*, clear enough that teachers can quickly assess relevance without extensive interpretation [48]. They should be *supervisable*, teachers should be able to evaluate whether suggested actions align with their instructional intent. And they should support rather than supplant *teacher agency*, functioning as optional supports rather than prescriptive directives. These principles inform how nudges might be integrated into open-ended educator-AI workflows where tasks are heterogeneous and professional judgment is paramount.

4.2.6 Educators' Instructional Work as Under-Specified Design Activity

Educators' instructional work involves translating pedagogical and content knowledge into actionable plans and decisions that are coherent, aligned to learning goals, and responsive to diverse learners [50]. Lesson plans remain a canonical artifact because they externalize alignment among standards, objectives, activities, and assessment [81]. Beyond lesson planning, educators routinely engage in other high-judgment tasks, including differentiation, assessment and feedback, resource creation, and professional reflection, each shaped by local constraints and student needs. Across these tasks, the work is often under-specified at the outset: goals evolve, constraints emerge, and educators must iteratively adapt available resources into teachable forms [132, 28].

Planning and instructional design also support anticipatory reasoning. Educators forecast likely learner difficulties, identify moments where coherence may break, and decide how to allocate attention across competing goals [24]. Well-developed instructional artifacts can increase preparedness for dynamic classroom contexts [43]. However, this work is cognitively demanding and time-intensive [83], requiring ongoing practice and reflective judgment to manage tradeoffs under constraint [142]. These characteristics make educator-facing AI support both promising and risky. AI can reduce effort and widen access to resources, but it can also introduce additional validation work, misalignment, and iterative repair when interactions do not fit authentic decision-making processes [153, 7].

4.2.7 GenAI in Educator Workflows

Teacher-facing GenAI tools can accelerate drafting, adaptation, and resource creation, yet educators still bear the burden of steering interactions toward instructionally usable outputs [138]. In open-ended chat, educators must translate intent into effective prompts, detect misalignment, and iteratively refine outputs to fit context, standards, and learner needs [103]. This burden is especially consequential in high-judgment tasks where subtle design choices affect instructional coherence, equity, and feasibility [43]. Without additional professional learning support to build AI literacy and help educators compose productive next moves, AI assistance can become an additional layer of work, requiring inspection,

correction, and reformatting to reach classroom-ready artifacts [7].

Recent research has begun examining how educators interact with AI systems for instructional purposes. Studies document both enthusiasm and concern. Educators appreciate efficiency gains but worry about accuracy, bias, and deskilling [56, 140, 174]. The quality of AI-generated instructional content varies considerably, and educators often need to substantially revise outputs [138, 13]. These dynamics motivate design strategies that support educators not only through content generation, but through lightweight guidance about what to do next, when, and with what level of specificity.

4.2.8 Human-Centered Design and Ethical Considerations

The promise of AI-powered nudges depends on trust, autonomy, and perceived alignment with professional values. Ethical critiques argue that nudges can become manipulative if they obscure intent or concentrate power in system designers, especially when users cannot contest or understand recommendation logic [26]. In professional contexts, such risks are heightened because recommended prompts may be interpreted as second-guessing expertise, reducing adoption when autonomy is central [70].

Human-centered design addresses these concerns by centering user goals, transparency, and control, emphasizing that AI should augment rather than replace professional judgment [144]. Transparency and explainability can support appropriate reliance and trust calibration [54], while autonomy-preserving designs keep recommendations optional and non-blocking [70]. Research on intelligent tutoring systems similarly emphasizes that effective scaffolds fade as learners develop competence, avoiding dependency [38].

Several principles guide ethical nudge design in professional contexts. *Transparency* requires that the presence and purpose of nudges be apparent to users [185]. *Autonomy preservation* requires that nudges remain genuinely optional and do not penalize non-adoption [161]. *Beneficence* requires that nudge content align with users' authentic professional interests rather than system-centric objectives [145]. *Non-deception* requires that nudge framing not misrepresent likely outcomes or exploit cognitive vulnerabilities [151]. These principles inform our design approach and evaluation framing.

4.2.9 Gaps and Research Motivation

Despite strong theoretical foundations for nudging and growing interest in AI-powered educational support, limited evidence exists on nudges as interaction scaffolds inside open-ended educator-AI workflows. Most nudging research examines one-shot interventions or simple behavioral targets, not ongoing professional decision-making within multi-turn conversations [71, 172]. Digital nudging research rarely examines temporal dimensions of nudge utility, when in conversations nudges are most useful, how adoption changes over time, and whether delayed uptake (enabled by persistent display) occurs meaningfully [79]. The orchestration and scaffolding literatures emphasize the importance of bounded, timely support but have not examined how nudges function as orchestration mechanisms within AI-mediated professional tools.

Furthermore, much work on AI-powered suggestion systems relies on the model's inferential capacity without grounding suggestions in domain expertise [31]. Although AI can generate plausible recommendations, the inconsistency and opacity of generative models make positive instructional impact difficult to guarantee without pedagogically-grounded design [91]. Understanding how pedagogically-framed nudges perform in authentic educational contexts, and how their design characteristics relate to educator adoption, is therefore essential for both theory and practice.

These gaps motivate the conceptual framework presented in the next section, which specifies how nudge design, conversational context, and educator adoption decisions interact to shape observable patterns within educator-AI conversations.

4.3 Conceptual Framework: Nudges as Choice Architecture in Educator-AI Workflows

Building on prior research in digital nudging and human-centered AI, we conceptualize AI-powered nudges in educator-facing chat as choice architecture embedded within an open-ended, multi-turn workflow. In contrast to one-shot recommendation widgets or prescriptive tutoring scripts, educator-AI chat is inherently under-specified: educators use the system for heterogeneous tasks (planning, adaptation, assessment design, communication, and operational work), and they frequently refine objectives and constraints as they see interim

outputs. In this setting, the primary friction is not whether the AI can produce plausible text, but whether educators can efficiently select and execute a productive next move without expending additional time to diagnose what is missing, translate an idea into a classroom-ready artifact, or repair misalignment between output and instructional intent.

Within this framing, nudges are designed to reduce workflow friction by surfacing a small set of high-leverage options while preserving educator autonomy and transparency, consistent with the autonomy-preserving premise of nudging [155] and "transparent, non-coercive" choice environments [70]. Rather than acting as a one-time intervention, nudges operate as lightweight, repeated prompts that appear at decision points throughout a conversation and across an educator's usage history. The system does not attempt to infer a single optimal action; instead, it exposes a small number of instructionally meaningful paths that educators can choose to incorporate, adapt, or ignore based on their professional judgment and local constraints.

4.3.1 Design Heuristics Governing Nudge Function

We specify six behavioral heuristics that govern how nudges function as choice architecture in educator-AI chat. The system limits the number of nudging options written in limited numbers of words presented at a given moment with **bounded choice** , reducing cognitive overload while maintaining a sense of control [79]. In the studied implementation, a maximum of two nudges with maximum 18 words each appear per AI response, creating binary choice points rather than overwhelming option sets. Nudges are aligned to the most recent educator request and AI response while use the conversation the entire conversation thread as context so that the underlying problem framing, constraints, and educational context expressed in the thread remain salient. Each nudge is generated based on within-turn conversational context, ensuring relevance to immediate work, which allowed the nudge to be personalized through **temporal contiguity** when the complete user profile data is missing [12]. Nudges remain available across turns, enabling educators to adopt a previously shown option when timing becomes appropriate rather than forcing an immediate decision. This **persistence and soft commitment supports** delayed uptake as user needs evolve [90].

Nudge messages are phrased to translate broad instructional goals into concrete, low-ambiguity next actions that are straightforward to enact **implementation intentions** or demonstrate **effort reduction** [66]. The nudge text is designed to be directly executable as the educator's next request, minimizing translation effort. Nudges also highlight productive directions through **framing and salience** without masking tradeoffs or restricting alternatives. Displayed nudge messages draw attention to specific moves (e.g., differentiation, formative checks, structuring) while preserving the option to continue free-form interaction [172].

4.3.2 Autonomy-Preserving Micro-Interventions

These heuristics together articulate a design philosophy of autonomy-preserving micro-interventions. The system does not engineer a single "correct" path or optimize educator behavior against a fixed objective function. Instead, it repeatedly offers low-barrier entry points into pedagogically grounded actions that educators can adopt, adapt, or disregard. The goal is to reduce the cognitive and operational costs of advancing high-quality instructional moves under real-world time and workload constraints, while preserving educators' expertise and agency.

This design philosophy distinguishes the present approach from several alternatives:

- **Recommendation systems** that optimize for engagement metrics, which may not align with educator welfare [154, for eg.]
- **Prescriptive tutoring** that guides users toward predetermined outcomes, limiting exploration [147, for eg.]
- **Information provision** that adds content without reducing decision friction [1, for eg.]
- **Profile-based personalization** that requires stable user data unavailable in many deployments [173, for eg.]

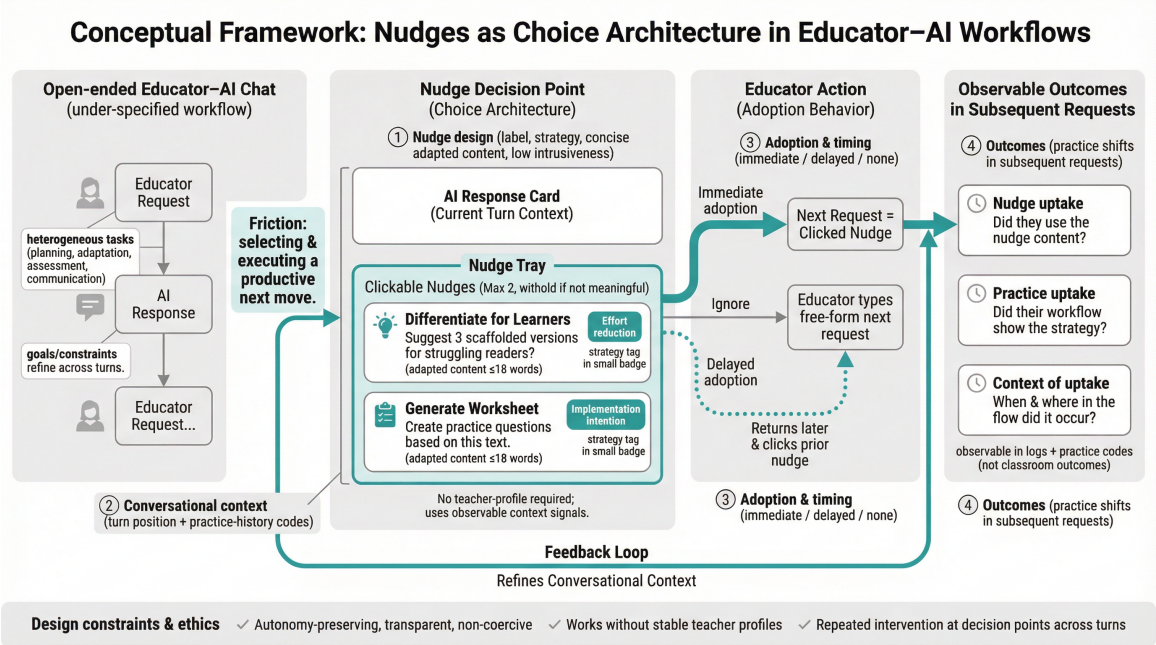

Figure 4.1: Conceptual framework depicting nudges as choice architecture embedded in educator-AI workflows. The framework emphasizes nudge decision points as repeated, autonomy-preserving opportunities for pedagogical steering.

4.3.3 Framework Components

Figure 4.1 depicts the conceptual framework as four connected components:

Component 1: Open-ended educator-AI workflow. Tasks span planning, differentiation, assessment, and broader professional work, with goals evolving across turns (Chapter 2). The workflow is inherently under-specified and iterative.

Component 2: Nudge decision point. Following each AI response, the system optionally surfaces a small set of contextually adapted next-step options. Each nudge is drawn from a fixed pedagogical taxonomy but adapted to the current conversational context.

Within this framework, the nudge decision point is the central analytic construct. After each AI response, the system may surface contextually adapted nudges for next steps. These nudges do not prescribe what educators must do or assume full knowledge of their goals. Instead, they function as low-friction affordances that foreground instructionally meaningful

actions such as refining a plan, differentiating support, adding rigor, or structuring an artifact for classroom use.

Component 3: Educator response behavior. Educators may respond by (a) adopting a nudge immediately so it becomes the next request, (b) ignoring nudges and continuing with free-form prompting, or (c) returning to a previously displayed nudge after intervening turns. This trichotomy distinguishes immediate adoption, delayed adoption, and non-adoption.

Component 4: Observable practice outcomes. Adoption decisions shape the trajectory and composition of instructional work in subsequent educator requests. Practice codes in post-adoption requests capture what instructional moves nudge adoption triggers.

A feedback loop connects outcomes from each exchange back to the system usage logs, enabling the system to adjust what it offers across diverse user behaviors. Over time, aggregated interaction data can inform refinement of the nudge space (e.g., which actions are surfaced and how they are framed) while remaining within ethical boundaries defined by transparency and minimal intrusiveness.

4.4 Study Context and Nudge Design

This section describes the deployment context, the implemented nudge design, and the key design variables that structure subsequent analyses.

4.4.1 Platform and Feature Context

This study examines adoption and instructional influence of a pedagogically structured nudging feature embedded within Brainstorm Ideas, a generative planning chat experience in the Colleague AI platform. As Figure 4.2 shows, the system displays up to two clickable, educator-facing prompts ("nudges") after each AI response. Each nudge is designed to represent a plausible next instructional move, supporting in-the-moment uptake of actionable next steps while minimizing disruption to educator autonomy and workflow pacing. A key design element is persistent, low-intrusion visibility: nudges remain available to be clicked even after multiple turns have passed, enabling educators to revisit and implement a prior suggestion when it becomes timely.

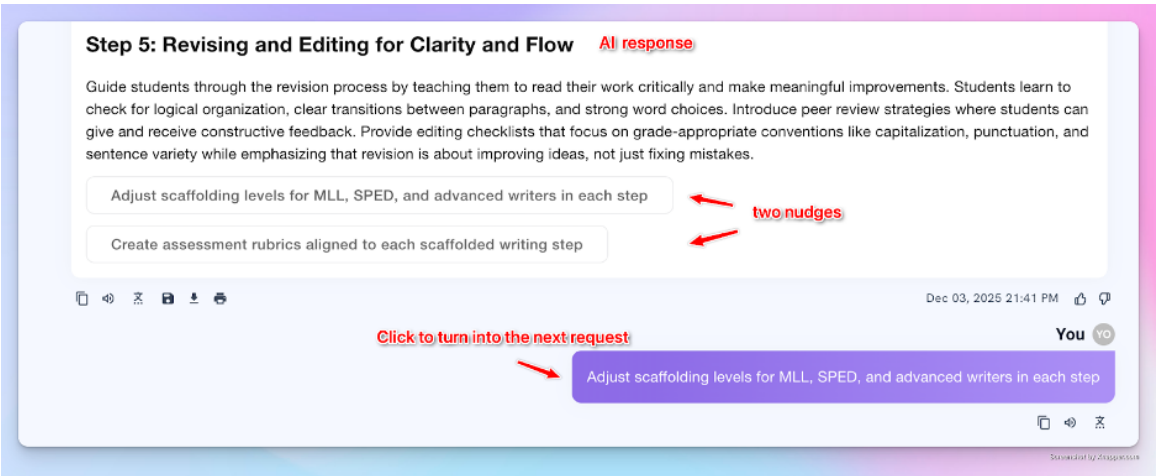

Figure 4.2: Exhibit of nudges in Brainstorm Ideas

4.4.2 Nudge Generation Mechanism

Nudge generation is conditioned on the current conversation. Nudges are intended to align with the educator's pedagogical intent, reflect work already completed in the thread, and target opportunities to deepen, clarify, or improve the developing artifact. To reduce distraction and cognitive load, the system may display one nudge or none when fewer than two options would meaningfully advance the work. Nudges are produced by a dedicated section of the system prompt that (a) instructs the model to select up to two next-step nudges, (b) prioritizes nudges by category, and (c) constrains how a selected nudge may be adapted to local context.

The nudges are produced by a section within the system prompt of AI agent for the feature. Full prompt is captured in Appendix C.1

Nudge Label Taxonomy Nudges are drawn from a fixed, predefined menu of instructional action types with immutable labels. The menu comprises 24 canonical labels organized into three priority-ordered categories, including Improve Practices, Consolidate Understanding [for students or educators], and Visual and Format Aids. The selection of the labels is based on literature reviews on effective instructional practices in section **??**.

While labels are fixed, the system produces an adapted, educator-facing realization of the nudge ("adapted content") that binds the action to local context identified in prior messages (e.g., grade level, subject area, student population, or artifact type). Adaptations are constrained to preserve the nudge's functionality: content may be personalized but cannot change what the action does.

Heuristic Strategy Taxonomy Each rendered nudge includes an explicit heuristic strategy tag that shapes attention and framing. The system selects from seven behavioral strategies provided as examples in the prompt or incorporate other strategies when necessary:

Table 4.1: Heuristic Strategy Definitions

Strategy	**Description**
Goal Gradient	Emphasizes progress toward completion
Implementation Intention	Specifies concrete next steps
Effort Reduction	Highlights ease or time savings
Default Framing Effect	Positions action as standard practice
Identity Priming	Appeals to professional identity
Salience Effect	Draws attention to importance
Social Proof	References peer behavior
Cultural Sensitivity	Emphasizes contextual relevance

These strategies are used to frame the adapted content without altering the underlying action. For example, the same "Differentiate for Learners" nudge label might be framed as "Quickly differentiate for your MLL students" (Effort Reduction) or "Other teachers also add differentiation at this stage" (Social Proof).

Surface Form Constraints To support adoption as a low-effort workflow move, the displayed content is constrained to be directly executable as the educator's next request. Only the adapted content is shown to educators; it must be: 1. Written in K-12 educator

language; 2. Formatted for readability; 3. Kept under 18 words. Particularly, the 18 words restriction allow nudge messages to be displayed in one single line in most of commonly used devices (e.g., Chromebook, 12 to 16-inch laptop screens).

This constraint is central to the design premise that nudges reduce interaction friction not by adding information, but by translating a productive next step into a "one-click" action. When an educator clicks a nudge, the nudge text becomes their next request verbatim.

Design Properties Summary In summary, the nudging design operationalizes choice architecture that balances instructional guidance with teacher autonomy. Nudges are **constrained** through a fixed 24-label menu with a two-option limit per response, preventing cognitive overload while maintaining consistency. They are **context-bound**, with suggestions dynamically adapted to the ongoing conversational exchange rather than generic recommendations. The system is explicitly **autonomy-preserving**: nudges are optional, non-blocking, and carry no penalty for non-adoption. Nudges are also **workflow-oriented**, phrased as immediately actionable next requests that align with teachers' in-the-moment instructional decision-making. They remain **persistent**, allowing for delayed adoption across turns as teachers reassess instructional needs. Finally, the nudges are **pedagogically grounded**, with labels derived from research-based instructional practices rather than model-generated heuristics.

4.4.3 Cross Feature Comparison: Unconstrained Suggestions

To contextualize the cross-feature comparison, we introduce the Lesson Plan Chat feature as a contrasting implementation of AI-supported planning dialogue. Lesson Plan Chat offers a similar interface pattern to Brainstorm Ideas, displaying up to two clickable follow-up suggestions after each AI response (Figure 4.3), but differs substantially in its underlying suggestion logic. Whereas Brainstorm Ideas relies on a fixed, pedagogically grounded nudge menu with predefined action labels and heuristic framings, Lesson Plan Chat generates suggestions dynamically through general-purpose model inference without a bounded menu or tagging schema. Suggestions in Lesson Plan Chat emerge from the LLM's interpretation of the local drafting context rather than from an explicit mapping of instructional moves.

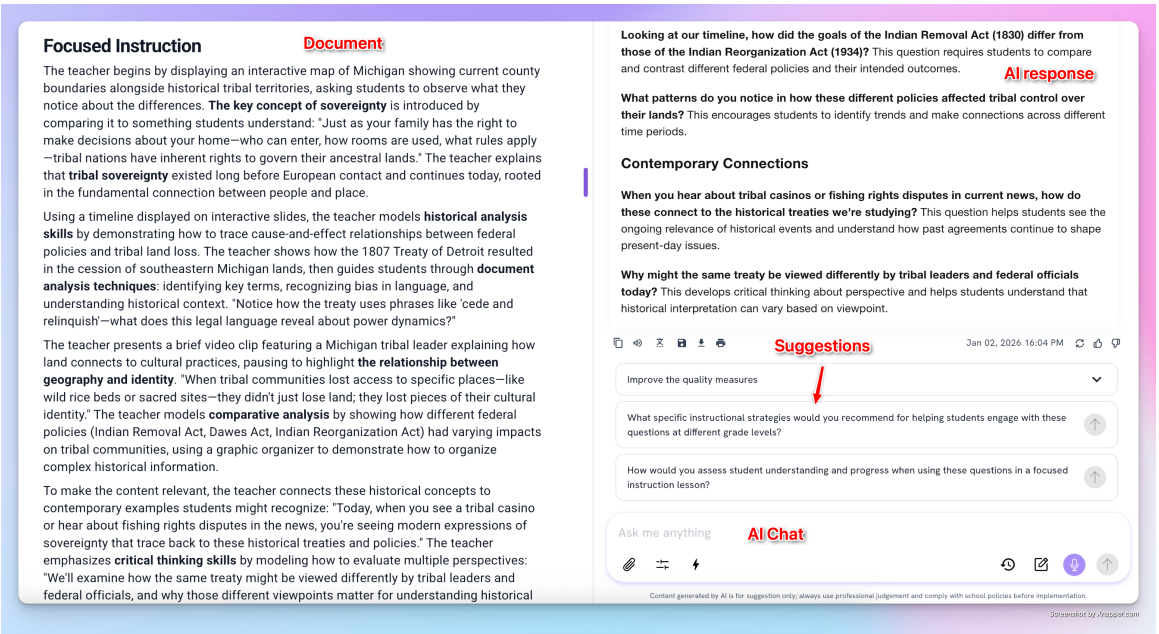

Figure 4.3: Suggestions Displayed in Lesson Plan AI Chat

Thus, nudges and suggestions vary across, embedded practices, message lengths, inclusion of heuristics, and sentence structures. In addition, suggestions will not persist in the conversation history. They will disappear after clicking or sending the next request. Therefore, suggestions do not allow for delayed adoption. This divergence allows us to compare educator behavior across two interaction designs-one with structured pedagogical scaffolds and friction-reducing heuristics, and one with open-ended generative flexibility. By analyzing adoption patterns across these two features, we assess how design constraints shape uptake, especially when similar conversational UI elements serve different underlying mechanisms. Key design differences are summarized in Table 4.2.

We computed parallel descriptive statistics for Lesson Plan Chat, including adoption rates per displayed suggestion, per conversation, and per educator, as well as distributions of adoption intensity. We also examined the linguistic form of adopted suggestions in Lesson Plan Chat, focusing on whether they functioned as execution-oriented actions or context-elicitation prompts. Because the two features differ in task context and design constraints,

this comparison is descriptive cross-feature comparison rather than causal A/B test; it is intended to highlight how different suggestion-generation designs correspond to diverse adoption patterns and functional roles within educator-AI interactions.

Table 4.2: Design Comparison: Constrained Nudges vs. Unconstrained Suggestions

Dimension	**Brainstorm (Constrained)**	**LP Chat (Unconstrained)**
Generation	Fixed 24-label menu	Model-inferred from context
Strategies	8 explicit heuristics	None specified
Persistence	Clickable across turns	Ephemeral
Format	2 nudges max per response	2 suggestions max per display
Adoption	Click converts to request	Click converts to request

This comparison is framed as a design contrast rather than causal evaluation, as the features differ on multiple dimensions and serve potentially different user populations.

4.5 Data Sources

This section describes the data sources, analytic units, outcome format, and statistical methods used to examine nudge adoption behavior and its relationship to instructional practice expression in educator-AI conversations.

4.5.1 Observation Period

Data were collected from September 10, 2025 (first day of deployment) through December 19, 2025 (common last day of school before winter holiday break), spanning 107 days of the Fall 2025 semester. This period captures authentic educator workflow during a typical instructional term, including early-semester planning, mid-semester adaptation, and end-of-semester assessment activities.

4.5.2 Data Sources

The analysis draws on four linked data sources, summarized in Table 4.3. Nudge events, user records, and lesson-plan chat logs were extracted from platform usage history. Practice codes capture qualitative coding of educators' requests. The development and implementation of the coding pipeline are described in Chapter 2, and the full codebook is provided in Appendix A.2.

Table 4.3: Data Sources and Key Variables

Data Source	**Description**	**Records**	**Key Variables**
Nudge Events	Display and click logs	897,080	conversationId, user_id, round, nudge_text, nudge_label, nudge_strategy, exact_adopted, delayed_adopted
Practice Codes	Coded instructional practices in educator requests	3,304,799	message_id, type, domain, category, code_value
User Records	Educator account metadata	8,453	id, account_created, organization
LP Chat Logs	Comparison feature logs	5,551	conversationId, type, messageType, content, match_nudge_pair

4.5.3 Units of Analysis

The core analytic unit is a **nudge display event**, defined as a single nudge shown to an educator within an AI response. Each displayed nudge is uniquely identified by the combination of conversation identifier, response message identifier, and nudge position (first or second within the response).

This unit is analytically preferable to conversation- or user-level aggregation because

nudges are situated interventions whose relevance depends on immediate conversational context, the same nudge label may be highly relevant in one context and irrelevant in another.

The data exhibit a **nested data structure** with three-level hierarchical structure:

```
Educator (n = 8,153)
  |-- Conversation (n = 127,196)
      |-- Response (n = 285,984)
          |-- Nudge Display (n = 567,577)
```

This nesting informs both descriptive statistics (reported at multiple levels) and interpretive caution regarding independence assumptions.

Cross-Feature Comparison Units For the cross-feature comparison, we standardized units to enable fair comparison. The comparison feature displays suggestions in a "1. [suggestion]" and "2. [suggestion]" format. Each display was expanded to two individual suggestion units, producing 5,280 suggestion-level observations (2,640 displays $\times$ 2) comparable to the 567,577 individual nudge units from the primary feature.

4.5.4 Behavioral Outcome Operationalization

Adoption is defined strictly as **observed click behavior** linking a displayed nudge to the educator request it generates. When an educator clicks a nudge, the nudge text becomes the next request verbatim. This conservative operationalization creates verifiable behavioral signal without attribution ambiguity, excludes manually typed paraphrases that may have been influenced by nudges and produces a lower-bound estimate of nudge influence.

For a displayed nudge n at response round t, we denote

$$A_{t,n} = \begin{cases} 1 & \text{if the nudge is clicked at any point in the same conversation} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

In addition, we distinguish two adoption types based on temporal proximity:

Table 4.4: Adoption Type Definitions

Type	Definition
Immediate	Nudge clicked in the turn it appears
Delayed	Nudge clicked in a later turn
Combined	Either immediate or delayed

Total adoption counting: For aggregate statistics, total adoptions combine the immediate and delayed adoption. This counting allows the same nudge to contribute to both categories if adopted multiple times, though such cases are rare.

Two **timing variables** capture when adoption occurs within conversations:

Position percentile: Normalized conversation progress (0–1) at nudge display:

$$\text{position\_percentile} = \frac{\text{round} - \text{min\_round}}{\text{max\_round} - \text{min\_round}} \tag{4.2}$$

Adoption lag: Number of intervening educator requests between display and adoption (0 = immediate; >0 = delayed).

4.6 Methodology

4.6.1 Descriptive Adoption Analysis

To characterize how educators used displayed nudges, we measured adoption at three levels. At the **nudge level**, we computed the proportion of displayed nudges that were adopted and reported 95% confidence intervals[1]. At the **conversation level**, we summarized the distribution of adoptions per conversation and the proportion of conversations with any adoption. At the **user level**, we summarized the distribution of adoptions per educator and the proportion of educators who adopted at least one nudge.

To assess whether adoption was concentrated among a small subset of users, we con-

[1]Confidence interval calculation is included in Appendix C.4.1

ducted a **concentration analysis** using the Gini coefficient [19],

$$G = \frac{2\sum_{i=1}^{n} i \cdot x_i}{n\sum_{i=1}^{n} x_i} - \frac{n+1}{n} \tag{4.3}$$

where x_i denotes the cumulative adoption count for user i, n represents the number of unique users who adopted at least one nudge, and the index i orders users from lowest to highest adoption frequency. A Gini coefficient of 0 indicates perfect equality (all users adopt the same number of nudges), while a coefficient approaching 1 indicates extreme concentration (a single user accounts for all adoptions).

Moreover, we also measure adoption centration using top-k shares (the proportion of all adoptions attributable to the top 10% and 20% of adopters), and the mean and median number of adoptions per adopter. We also conducted a **retention analysis** by estimating the proportion of adopters who used the feature exactly once ("one-time adopters") versus those who returned for subsequent adoptions. Finally, a **timing analysis** examined adoption rates by conversation progress decile to identify when, within a conversation, educators were most likely to adopt nudges.

4.6.2 Design Dimension Analysis

For each label and strategy, we computed adoption rates conditional on exposure with 95% confidence intervals. **Wilson score intervals** were used for proportions [17, 141], providing more accurate coverage than normal approximation for our asymmetric data sample:

$$\tilde{p} = \frac{p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \tag{4.4}$$

where $p = k/n$ is the observed adoption rate, k is the number of adopted nudges, n is the total number of nudges shown, and z is the standard normal critical value corresponding to the desired confidence level ($z = 1.96$ for $\alpha = 0.05$).

Chi-square tests assessed whether adoption varied significantly across design dimensions. The data consist of counts of adoption events across discrete categories, making the chi-square test appropriate for assessing departures from independence without assuming normality or equal variances [60].

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.5}$$

where O_{ij} is the observed number of nudges with characteristic i (e.g., nudge label or heuristic strategy) and adoption outcome j (adopted or not adopted), E_{ij} is the expected count under the null hypothesis that adoption is independent of nudge characteristics, calculated as $E_{ij} = R_i C_j / n$, where R_i is the total number of nudges with characteristic i, C_j is the total number of nudges with outcome j, and n is the total number of nudges shown.

Cramér's V as the effect size measure, which enables comparison of effect magnitudes across different design features [157].

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} \tag{4.6}$$

where n is the total number of nudges shown, and k is the number of nudge categories being compared (e.g., 24 for labels, 8 for heuristic strategies). Effect size interpretation: $V < 0.1$ (negligible), 0.1–0.3 (small), 0.3–0.5 (medium), > 0.5 (large).

Within-display paired comparisons provided stronger behavioral evidence by examining cases where exactly two nudges were shown and exactly one was adopted. Because both nudges share identical context (same educator, conversation state, and timing), these comparisons control for contextual confounds. Win rates were computed for labels and strategies appearing in at least 10 valid pair events.

4.6.3 Contextual Analysis

Adoption rates were computed conditional on practice codes appearing in the preceding educator request, identifying which instructional contexts show elevated or depressed nudge receptivity. Post-adoption practice expression was examined by identifying practice codes in requests following adopted nudges, characterizing what instructional work nudge adoption triggers.

Professional Learning Analysis To examine whether adoption patterns change with user experience, we computed adoption rates stratified by **user tenure**—days since first

nudge exposure. Tenure groups: Day 1, Week 1 (days 2–7), Month 1 (days 8–30), Month 2 (days 31–60), and 60+ days.

4.6.4 Methodological Limitations

Several limitations constrain causal inference:

1. **Selection effects:** Educators who adopt nudges may differ systematically from non-adopters

2. **Nesting:** Observations are nested within conversations and users, violating independence

3. **Left-censoring:** For existing users, practices expressed before the observation window are unobserved

4. **Temporal confounds:** Adoption patterns may reflect time-of-semester effects

We address these through descriptive framing, multiple levels of analysis, and explicit interpretive boundaries. Sensitivity analyses were conducted for new users (accounts created after September 10, 2025) to partially address left-censoring [30].

4.7 Findings

This section presents empirical findings organized around the three research questions. We report deployment-scale descriptive statistics, design dimension effects, contextual factors, and the cross-feature comparison between constrained and unconstrained suggestion mechanisms. Throughout, we emphasize patterns that illuminate how educators engage with AI-embedded nudges in authentic professional workflows.

4.7.1 Deployment Scale and Nudge Exposure

Before examining adoption behavior, we establish the scale and reach of the nudging intervention. Understanding exposure patterns provides essential context for interpreting adoption rates and concentration metrics.

The nudge system displayed **567,577 individual nudges** across **285,984 assistant responses** during the over three-month observation period. This substantial volume reflects the integration of nudges into routine educator-AI interaction rather than episodic or experimental exposure.

Table 4.5: Nudge Exposure Distribution by Response

Nudges per Response	**Responses**	**Percentage**
0 nudges	11,593	3.9%
1 nudge	4,402	1.5%
2 nudges	281,163	94.6%

The exposure distribution in Table 4.5 reveals that the vast majority of AI responses (94.6%) presented educators with exactly two nudges, implementing the bounded choice design principle described in Section 4.4.2. The small proportion of responses with zero or one nudge (5.4% combined) reflects instances where the system determined that fewer options would meaningfully advance the instructional work. For example, when the conversation had reached a natural conclusion or when prior nudges had already addressed the most relevant next steps. This adaptive display behavior ensures that nudges do not become visual noise that educators learn to ignore.

Nudge exposure reached a substantial educator population. A total of **8,153 unique educators** received at least one nudge display across **127,196 unique conversations**. The mean number of nudges displayed per educator was 69.6, though the median was considerably lower at 24, indicating a right-skewed distribution where some educators engaged in substantially more AI-assisted work than the typical user. This exposure distribution sets the stage for examining how frequently and in what patterns educators converted displayed nudges into adopted actions.

4.7.2 RQ2: How do educators adopt and use nudges in authentic educator-AI conversations? Adoption Behavior: Frequency, Concentration, and Retention

We examine adoption behavior through four complementary lenses, including overall adoption frequency, temporal patterns within conversations, concentration across users, and user retention over time. Together, these analyses characterize how nudges function in practice and for whom they provide value.

Overall Adoption Frequency The central descriptive finding is that **4.04% of displayed nudges were adopted** (95% CI: [3.99%, 4.09%]), representing 22,934 total adoption events. Because two nudges are displayed simultaneously and only one can be adopted, *the maximum possible adoption rate is approximately 50%.* This rate reflects selective rather than habitual engagement, educators declined the substantial majority of displayed nudges, choosing to adopt only when suggestions aligned with their immediate workflow needs and instructional intent.

Table 4.6: Adoption Metrics by Level of Analysis

Level	**Metric**	**Value**	**95% CI**
Nudge	Overall adoption rate	4.04%	[3.99%, 4.09%]
Nudge	Total adoptions	22,934	—
Nudge	Immediate adoptions	20,583 (89.7%)	—
Nudge	Delayed adoptions	2,351 (10.3%)	—
Conversation	Conversations with $\geq$1 adoption	18.0%	—
User	Users with $\geq$1 adoption	52.3%	—

The low per-nudge adoption rate admits two competing interpretations. One interpretation frames this as evidence of *selective engagement.* Educators exercised professional judgment, evaluating each nudge against their current needs and adopting only when the suggestion would meaningfully advance their work. Under this interpretation, the 96% non-adoption rate reflects appropriate discrimination rather than feature failure. An alternative

interpretation frames low adoption as *design limitation*. Perhaps most nudges were poorly timed, irrelevant, or otherwise misaligned with educator needs, such that rejection was the rational response to low-quality suggestions.

Several patterns support the selective engagement interpretation. First, at higher levels of aggregation, adoption appears substantially more prevalent: 18.0% of conversations included at least one adoption, and 52.3% of educators adopted at least once during the three-month observation period. If nudges were uniformly unhelpful, we would expect these proportions to be much lower. Second, as we demonstrate below, adoption rates vary systematically and predictably by nudge characteristics, timing, and context, variation that would be difficult to explain if educators were simply ignoring a unhelpful feature. Third, the 64.2% retention rate among adopters (detailed in Section 4.7.2) suggests that most educators who tried adoption found sufficient value to return.

The pattern of low per-nudge adoption but substantial conversation- and user-level reach is consistent with our conceptual framing of nudges as occasional scaffolds rather than routine interaction steps. Educators do not need external prompts for every decision; nudges provide value at specific moments when they surface a productive action that might otherwise have remained latent.

Immediate Versus Delayed Adoption A distinctive design feature of the nudging system is *persistence*. Displayed nudges remain clickable across subsequent conversation turns, enabling educators to return to previously shown options when timing becomes appropriate. This design choice rests on the hypothesis that nudge utility is not limited to the moment of display. A suggestion that seems irrelevant in one context may become valuable as the conversation evolves.

The data support this hypothesis. Of the 22,934 total adoptions, **20,583 (89.7%)** occurred immediately (the nudge was clicked in the same turn it appeared), while **2,351 (10.3%)** occurred after intervening turns (delayed adoption). Although immediate adoption dominates, the non-trivial delayed adoption rate carries several implications.

First, it demonstrates that persistence adds measurable value. If nudges were only useful at the moment of display, we would observe near-zero delayed adoption. The 10.3%

rate indicates that approximately one in ten adoptions would have been lost under an ephemeral display design. Second, it suggests that nudges function not only as immediate conversion mechanisms but also as *persistent cognitive reminders* that remain available for later retrieval. Educators appear to mentally "bookmark" potentially useful nudges and return to them when context shifts, for example, when an initial line of inquiry reaches completion and the educator considers next steps. Third, the existence of delayed adoption supports the scaffold interpretation as nudges are resources to be drawn upon as needed rather than mandatory interaction steps that must be processed immediately.

The delayed adoption pattern also has design implications that we explore in Section 4.8.2. Features that display suggestions ephemerally, showing options only momentarily before they disappear, would forfeit approximately 10% of the engagement that persistence enables. This finding informs the cross-feature comparison presented in Section 4.7.6.

Adoption Timing Within Conversations To understand *when* in conversations nudges provide the most value, we examined adoption rates by conversation progress, measured as the position percentile of the nudge display within its conversation (0% = first response, 100% = final response).

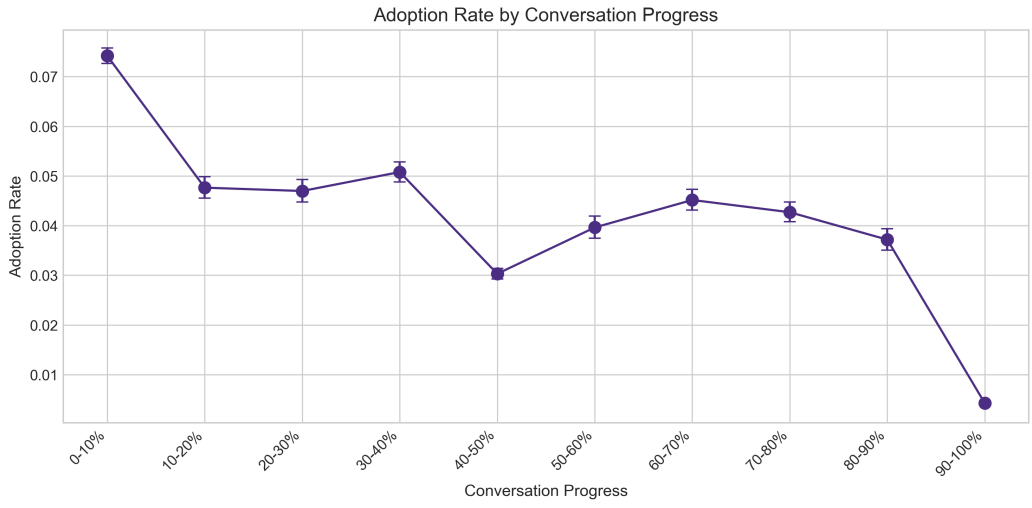

Figure 4.4: Adoption Rate by Conversation Progress Decile

The timing analysis in Figure 4.4 reveals a pronounced temporal pattern. Adoption peaks at conversation onset (7.42% in the first decile, nearly double the overall rate) and drops precipitously at conversation conclusion (0.42% in the final decile, approximately one-tenth the overall rate). The intermediate deciles hover near the overall mean with modest variation.

This temporal pattern admits a coherent interpretation aligned with the conceptual framework. Early in conversations, educators are *establishing direction*, formulating goals, exploring possibilities, and deciding how to structure their instructional work. At this stage, nudges can serve as *direction-setters* that help educators identify productive paths through the option space. The cognitive value of external prompts is highest when the space of possible directions is largest and the educator has not yet committed to a specific approach.

As conversations progress, educators typically commit to a direction and work toward completing the artifact or task they initiated, reducing the need for external suggestions as the path forward becomes clearer. However, nudges can still function as pedagogical reminders, supporting the uptake of specific practices when needed. By the final turns of a conversation, educators are focused on finishing rather than exploring alternatives; nudges suggesting new directions are unlikely to be adopted because they conflict with the educator's closure-oriented goal state.

The sharp drop at conversation end (0.42%) is particularly notable. This finding is not simply an artifact of shorter final turns or reduced display volume; rather, it may eflect a genuine shift in educator orientation from exploration to completion. The practical implication is that nudge utility is temporally bounded. Thus systems that display nudges uniformly throughout conversations may be allocating attention inefficiently relative to designs that concentrate nudges during early, direction-setting phases or optimizing underlying nudge labels along conversation flow.

Concentration of Adoption Across Users A critical question for understanding nudge value is whether adoption is distributed broadly across educators or concentrated among a small subset of intensive users. High concentration would suggest that nudges serve a narrow population with specific needs, while broad distribution would suggest more universal utility.

Table 4.7: Adoption Concentration Metrics

Metric	**Value**
Gini coefficient	0.799
Top 10% adopters' share of total adoptions	51.6%
Top 20% adopters' share of total adoptions	66.9%
Mean adoptions per adopter	5.3
Median adoptions per adopter	2.0

The concentration analysis results in Table 4.7 reveals substantial inequality in adoption distribution. The Gini coefficient of 0.799 indicates that adoption is far from uniformly distributed[2]. The top-decile concentration reinforces this picture. The top 10% of adopters (approximately 426 educators) accounted for 51.6% of all 22,934 adoptions, while the top 20% accounted for 66.9%. The mean-median gap (5.3 vs. 2.0 adoptions per adopter) reflects the underlying right-skewed distribution, most adopters engaged occasionally (median = 2) while a smaller group engaged intensively, pulling the mean upward.

This concentration pattern calls for reflecting *legitimate heterogeneity in scaffolding needs*. Some educators genuinely benefit from frequent external prompts (perhaps due to inexperience, high cognitive load, or complex instructional challenges), while others have sufficient internal resources to generate productive next steps without assistance. Under this interpretation, concentration is not problematic, instead it represents appropriate matching of support intensity to need.

Alternatively, *design limitations or inequitable access* may also contribute to the skewed adoption pattern. The nudge design may fit some workflows or user characteristics better than others, inadvertently underserving educators whose needs are not well represented in the fixed label menu. Concentration may also reflect broader platform usage dynamics, with

[2]a Gini of 0 would indicate perfect equality (all adopters adopt equally) while a Gini of 1 would indicate maximum inequality (one adopter accounts for all adoptions). The observed value falls in the range typically associated with highly concentrated phenomena such as wealth distribution in unequal economies or citation counts in academic publishing [15]

a small group of power users accounting for a disproportionate share of activity alongside many occasional users.

The data cannot definitively distinguish these interpretations, as we lack measures of underlying scaffolding need against which to benchmark observed adoption. The retention analysis in the next section provides results that support this interpretation. Sensitivity analysis is included in C.4.2 by excluding high-volume users. Results show that concentration is reduced but not removed.

User Retention: One-Time Versus Repeat Adoption To distinguish trial behavior from sustained engagement, we examined how many adopters returned for subsequent adoption after their first use.

Table 4.8: User Retention Among Adopters

Metric	**Value**
Total users with ≥1 adoption	4,261
One-time adopters (exactly 1 adoption)	1,526 (35.8%)
Repeat adopters (≥2 adoptions)	2,735 (64.2%)

The retention analysis in Table 4.8 reveals that **64.2% of adopters** returned for at least one additional adoption after their first use, while 35.8% adopted exactly once and never returned. This pattern has several implications.

First, the majority retention rate (64.2%) suggests that nudge adoption generally delivers value that motivates continued use. If first adoptions were uniformly disappointing, such as producing irrelevant AI responses or failing to advance instructional work, we would expect higher abandonment after first trial. The fact that most adopters return implies that the conversion from displayed nudge to executed action typically produces outcomes that educators find useful enough to repeat.

Second, the substantial minority of one-time adopters (35.8%) indicates that first impressions matter. More than one-third of educators who tried adoption did not find suffi-

cient value to return to the nudging feature or return to the platform for AI conversation. This pattern could reflect heterogeneity in underlying needs (some educators tried nudges and correctly concluded they don't need external scaffolding), poor first-adoption experiences (the specific nudge adopted happened to produce disappointing results), or workflow mismatch (the educator's typical tasks are not well served by the available nudge labels). Regardless of mechanism, the one-time adopter proportion suggests design attention to the quality of initial experiences, ensuring that first adoptions are satisfying may be particularly important for building sustained engagement.

Third, the retention rate provides context for interpreting the concentration findings. High concentration combined with high retention suggests that heavy adopters are not simply outliers or data artifacts; they represent a population that consistently finds value in repeated nudge use. This combination is more consistent with genuine heterogeneity in scaffolding needs that informs the future refinement of the nudge designs.

4.7.3 Design Dimension Effects

This section addresses the first half of *RQ3. How is nudge adoption related to the instructional practices and professional tasks expressed in nudges [...]?* Having established overall adoption patterns, we now examine how adoption varies by design characteristics, specifically, the nudge label (what action the nudge proposes) and the heuristic strategy (how the nudge is framed). These analyses identify which design choices predict higher or lower adoption, informing both theoretical understanding and practical design guidance.

Adoption by Nudge Label Each nudge is generated from a fixed menu of 24 pedagogically grounded labels representing distinct instructional action types. We computed adoption rates for each label conditional on display, testing whether adoption varied significantly across labels.

Adoption rates differed significantly across labels, as shown in Figure 4.5. The association was statistically significant, $\chi^2 = 7{,}558.46$, $p < .001$, but small in magnitude, Cramér's $V = 0.117$[3]. This suggests that label choice was associated with modest but reliable differ-

[3]Effect size interpretation: $V < 0.1$ (negligible), $0.1 \leq V < 0.3$ (small), $0.3 \leq V < 0.5$ (medium), $V \geq 0.5$

ences in adoption.

Table 4.9: Nudge Labels with Highest Adoption Rates

Label	**Adoption Rate**	n **Displayed**
Generate Visual Aids	13.4%	179
Generate Worksheet	9.9%	46,028
Organize into a Table	9.0%	18,009
Format in Concise Language	5.6%	21,099
Check Learning Standards Alignment	5.2%	7,802
Differentiate for Learners	5.1%	34,273
Design Tiered Tasks	4.4%	5,045
Add Formative Checks for Students	4.1%	71,486
Generate Rubrics for Evaluation	3.8%	50,326
Verify Source Accuracy	3.7%	9,630

Inspection of the highest- (Table 4.9) and lowest-adoption (Table 4.10) labels reveals a coherent pattern that we term the **executability principle**: labels that produce concrete, immediately usable artifacts are adopted at substantially higher rates than labels that require additional elaboration or reflection.

The top-performing labels share a common characteristics: they generate *deliverables* that educators can directly use in their classrooms. "Generate Worksheet" (9.9%; e.g., "Create a checklist version where students can identify their personal stress patterns") produces a classroom-ready artifact. "Organize into a Table" (9.0%; e.g., "Convert the 12-week rotation into a tracking table for easy reference") restructures existing content into a usable format. "Generate Rubrics for Evaluation" (3.8%; e.g., "Create detailed rubrics for grammar usage and visual presentation criteria") creates an assessment tool. These labels shorten the distance between AI interaction and classroom implementation. The educator clicks, receives an artifact, and can proceed to use it with minimal additional processing.

(large).

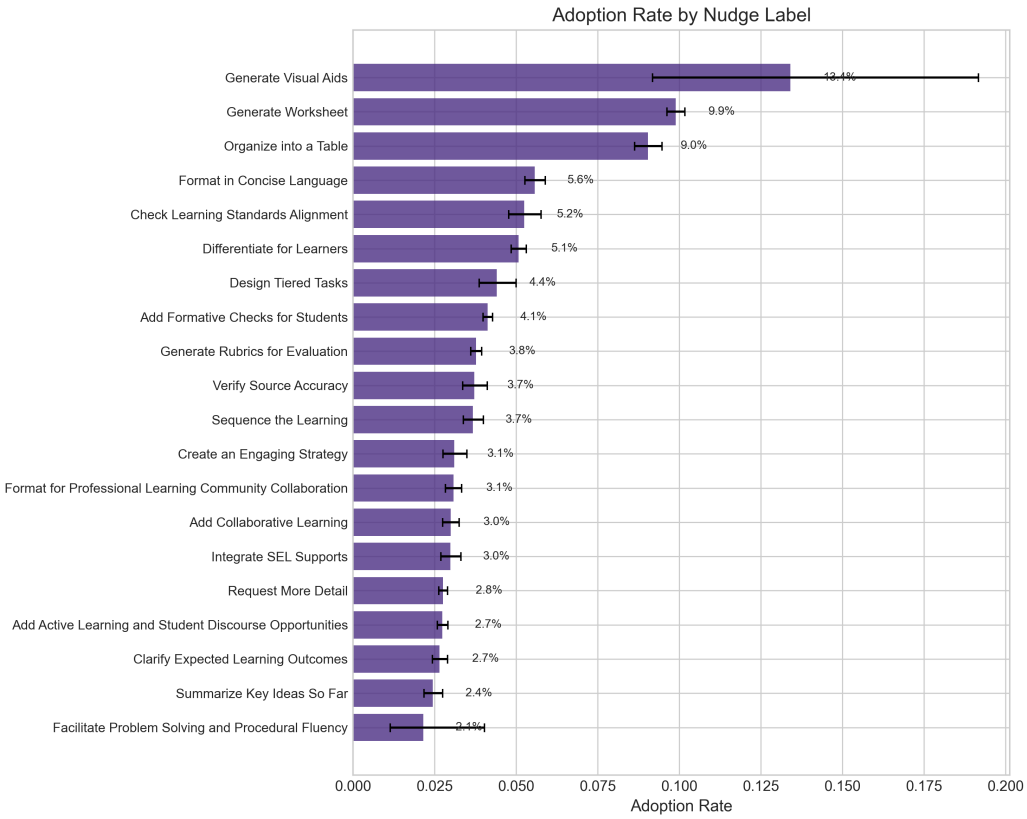

Figure 4.5: Adoption Rate by Nudge Label

In contrast, the lowest-performing labels tend to be more abstract or require further specification before producing usable outputs. "Promote Higher-Order Critical Thinking" (1.2%) is a worthy instructional goal but does not specify what artifact or action will result. "Anticipate Misconceptions" (1.7%) invites reflection but does not produce a deliverable. "Add Culturally Relevant Context" (1.8%) requires the system to make substantive content judgments that may or may not align with the educator's specific student population. These labels may represent important instructional practices, but their lower executability reduces adoption in time-pressured workflows.

The executability principle has both encouraging and cautionary implications. On the encouraging side, it provides clear design guidance: nudges that translate diffuse goals into concrete next steps will see higher adoption than nudges that remain at the level of general

Table 4.10: Nudge Labels with Lowest Adoption Rates

Label	**Adoption Rate**	n **Displayed**
Promote Higher-Order Critical Thinking	1.2%	14,107
Anticipate Misconceptions	1.7%	31,089
Add Culturally Relevant Context	1.8%	42,715
Generate Student-Facing Feedback	1.9%	270
Facilitate Problem Solving	2.1%	419

advice. This aligns with implementation intention theory [66], which demonstrates that specifying concrete action plans bridges intention-action gaps.

On the cautionary side, the executability principle raises concerns about pedagogical depth. Some of the lowest-adoption labels, such as "Anticipate Misconceptions," "Promote Higher-Order Thinking", represent instructional practices that educational research consistently identifies as important for student learning [72, 8]. If the nudge system inadvertently privileges procedural efficiency (worksheet generation) over conceptual depth (misconception anticipation), it may reinforce rather than challenge patterns of surface-level instructional work. We return to this tension in the Discussion.

Adoption by Heuristic Strategy Each nudge is framed using one of eight behavioral heuristic strategies that govern how the action is presented to educators. We examined whether framing strategy predicted adoption, holding the underlying action constant.

Strategy effects were significant: $\chi^2 = 2{,}677.52$, $p < .001$. The magnitude of variation across strategies was substantial.

Effort Reduction framing (6.26%) substantially outperformed all other strategies, with an adoption rate more than 2.5 times higher than Social Proof (2.37%). Effort Reduction frames nudges in terms of ease and time savings (e.g., "Quickly create a worksheet for..."), directly addressing the workload concerns that dominate educators' decision-making under time pressure.

Table 4.11: Adoption Rate by Heuristic Strategy

Strategy	**Adoption Rate**	n **Displayed**
Effort Reduction	6.26%	122,767
Default Framing Effect	4.43%	20,785
Implementation Intention	4.06%	195,968
Goal Gradient	3.20%	76,176
Salience Effect	2.58%	56,271
Identity Priming	2.42%	29,146
Social Proof	2.37%	63,109
Cultural Sensitivity	0.0%	4

Default Framing Effect (4.43%) and **Implementation Intention** (4.06%) occupied intermediate positions. Default Framing positions the suggested action as standard practice (e.g., "Most lessons include..."), leveraging the tendency to accept presented options as normative. Implementation Intention specifies concrete next steps, aligning with the executability principle.

The strategies with lowest adoption, **Identity Priming** (2.42%) and **Social Proof** (2.37%), are notable because they represent persuasive techniques with strong theoretical foundations and empirical support in other domains [36]. Identity Priming appeals to professional identity (e.g., "As an effective educator, you might..."), while Social Proof references peer behavior (e.g., "Other teachers also...").

The underperformance of Identity Priming and Social Proof in this context suggests that persuasive framings that work in consumer or general decision-making contexts may be less effective in professional workflows. Several mechanisms might explain this pattern. First, educators operating under time pressure may be primarily responsive to workload considerations (captured by Effort Reduction) rather than identity or social considerations. Second, appeals to professional identity may feel patronizing or manipulative when embedded in a tool that educators use for practical work. Third, social proof references to "other

teachers" may lack credibility or relevance when educators cannot verify the claim or assess whether the referenced teachers face similar contexts.

This finding cautions against importing marketing-style persuasive techniques into professional tool design. What works in consumer contexts, where identity appeals and social proof have proven effective, may be less effective in professional workflows where time pressure, practical utility, and autonomy concerns dominate the decision calculus.

Within-Display Paired Comparisons The analyses above examine adoption rates conditional on display, but these rates may be confounded by contextual factors. A nudge label that happens to be displayed more often in contexts where adoption is generally high will show elevated adoption rates even if the label itself has no causal effect.

To address this concern, we conducted **within-display paired comparisons**: analyses restricted to cases where exactly two nudges were displayed and the educator adopted exactly one. In these cases, both nudges faced identical context, with same educator, same conversation state, and same timing, and the educator made an active choice between them. Contextual confounds are controlled by design.

We analyzed **20,472 valid pair events** meeting these criteria. For each label and strategy, we computed win rates: the proportion of paired comparisons in which that option was chosen over its competitor.

The paired comparison results strongly reinforce the executability principle. "Generate Worksheet" won 75.0% of its head-to-head matchups, when educators faced a choice between generating a worksheet and any alternative action under identical conditions, they chose the worksheet three times out of four. "Anticipate Misconceptions" won only 29.0% of matchups, losing to most alternatives.

The paired comparisons provide the strongest behavioral evidence in our analysis because they control for all observable and unobservable contextual factors. The consistent pattern—artifact-producing labels defeating improvement-oriented labels—establishes that the executability effect reflects genuine educator preferences rather than confounded context.

Strategy win rates showed a similar pattern. Effort Reduction won 62.6% of paired

Table 4.12: Label Win Rates in Paired Comparisons

Label	**Win Rate**	**Pair Appearances**
Highest win rates:		
Generate Worksheet	75.0%	5,618
Format in Concise Language	70.7%	1,572
Organize into a Table	66.2%	2,228
Verify Source Accuracy	61.3%	551
Lowest win rates:		
Anticipate Misconceptions	29.0%	1,555
Add Culturally Relevant Context	32.9%	2,135
Add Active Learning Opportunities	36.4%	2,459
Add Collaborative Learning	36.4%	1,221

comparisons, while Social Proof won only 35.0%. When forced to choose between framings under identical conditions, educators consistently preferred efficiency-oriented framing over persuasive framing.

4.7.4 Contextual Factors: Instructional Practice Relationships

This section addresses the second half of RQ3: How is nudge adoption related to the instructional practices and professional tasks expressed in [...] educator-AI conversations?

Beyond design characteristics, adoption may vary with the instructional context in which nudges are displayed. We examined two facets of context: the practice codes present in educator requests *before* nudge display (pre-adoption context) and the practice codes present in requests *after* nudge adoption (post-adoption practice expression).

Pre-Adoption Context We computed adoption rates conditional on each of the 67[4] practice codes appearing in the educator request immediately preceding nudge display. This analysis identifies which instructional contexts are associated with elevated or depressed nudge receptivity.

Adoption rates varied modestly depending on what educators were doing immediately before a nudge was displayed, with the highest-uptake contexts clustering around concrete, generative planning work. To conduct qualitative analysis on practices over the immediately prior messages to the adopted nudges. The strongest adoption-rate contexts included Entire Lesson Planning (6.13%), Differentiated Instructional Strategies (6.04%), Instructional Routine (5.70%), In-class Activity Design and Adjustment (5.66%), Differentiation and Accessibility (5.61%), Project-Based and Real-World Learning (5.25%), and Generate Formative Assessments (5.21%). These are moments when teachers are already assembling or revising instructional materials, so a suggested "next move" that produces a usable artifact (e.g., a worksheet, table, rubric, or formative check) aligns with the workflow and can be enacted with minimal additional decision-making.

In contrast, lower adoption-rate contexts tended to be more diffuse, reflective, or less deliverable-oriented. For example, Discourse Continuity (4.33%) and Modification Request (4.13%) often reflect educators refining or reworking prior AI-generated content rather than pivoting to a new practice, which reduces the likelihood of taking up an additional nudge. Similarly, Professional Responsibilities (4.31%), such as documentation or communication tasks, may be situations where instructional "next-step" practices are simply less applicable. In these contexts, nudges are more likely to compete with ongoing sensemaking, clarification, and iterative polishing rather than supporting a clear transition or extension toward a classroom-ready artifact. Therefore, if the goal is to support deep pedagogical reflection, which may occur during revision or administrative phases, different tailored design strategies may be needed.

[4] 68 practice codes were used for coding, one code "Reject pervious output" was not detected in the coding results.

Post-Adoption Practice Expression We examined what instructional practices appeared in educator requests immediately following nudge adoption. This analysis characterizes the downstream instructional work that adoption triggers.

Table 4.13: Practice Codes Most Frequently Triggered by Adoption

Practice Code	**Count**	**Percentage**
Learning Standards Alignment	2,528	11.1%
Modification Request	2,068	9.1%
Format Modification	1,322	5.8%
In-class Activity Design	1,167	5.1%
Generate Formative Assessments	1,009	4.4%
ELA Skills Development	952	4.2%
Actionable Engagement Strategy	772	3.4%
Critical Thinking Development	719	3.2%

The post-adoption practice profile indicates that nudge adoption reliably triggers **substantive continuation of instructional work** rather than conversation termination or topic switching. The most common post-adoption practice is Learning Standards Alignment (11.1%), suggesting that adopted nudges often lead educators to connect their work to curriculum standards. Modification Request (9.1%) and Format Modification (5.8%) indicate that educators also refine AI outputs after adoption, adapting generated content to their specific needs.

Notably, the post-adoption profile includes practices beyond the immediate nudge action. For example, Critical Thinking Development (3.2%) appears despite the low adoption of "Promote Higher-Order Thinking" as a nudge label. This suggests that nudge adoption may initiate chains of instructional work that extend beyond the specific action suggested, a form of scaffolded progression where the initial nudge opens pathways to additional practices.

4.7.5 Professional Learning: Adoption by User Tenure

A central question for understanding nudge value is whether adoption patterns change as educators gain experience with the platform. If nudges function as onboarding scaffolds, we would expect higher adoption among new users who are learning to use the system effectively. If nudges provide ongoing value regardless of experience, we would expect relatively stable adoption across tenure levels.

We measured user tenure as days since first nudge exposure and computed adoption rates for five tenure groups: Day 1, Week 1 (days 2–7), Month 1 (days 8–30), Month 2 (days 31–60), and 60+ days.

Table 4.14: Adoption Rate by User Tenure

Tenure Group	n **Displayed**	n **Adopted**	**Adoption Rate**
Day 1	69,322	4,975	7.18%
Week 1 (days 2–7)	68,055	3,209	4.72%
Month 1 (days 8–30)	151,986	5,998	3.95%
Month 2 (days 31–60)	156,821	5,303	3.38%
60+ days	121,393	3,263	2.69%

Adoption rates decline monotonically with user tenure: 7.18% on Day 1, falling to 4.72% in Week 1, 3.95% in Month 1, 3.38% in Month 2, and 2.69% after 60 days. The decline is substantial as Day 1 adoption is nearly three times higher than 60+ day adoption.

This pattern is consistent with the **scaffolding hypothesis** that nudges provide greatest value during early use, when educators are learning to interact effectively with the AI system and may benefit from external prompts that model productive next steps. As educators develop familiarity and expertise, they internalize productive interaction patterns and can generate effective requests without external scaffolding.

Alternative interpretations exist. The **novelty hypothesis** suggests that initial high adoption reflects curiosity about a new feature that fades with familiarity, regardless of

learning. The **compositional hypothesis** suggests that the effect is driven by differential attrition. If high-adoption users are more likely to stop using the platform, the remaining long-tenure population would mechanically show lower adoption.

The data cannot definitively distinguish these mechanisms. However, the scaffolding interpretation has important practical implications if correct: it suggests that nudge systems may be particularly valuable for *onboarding new users* and *supporting early-career professionals*, with diminishing (though still positive) returns for experienced users. Design strategies might differentiate nudge intensity by user tenure, providing more prominent scaffolding for new users while fading support as expertise develops.

Sensitivity analyses are reported in Appendix C.4.2, restricting the sample to new users who opened accounts on or after September 10, 2025. New users exhibited slightly higher adoption rates but showed the same tenure-related decline in adoption. These results indicate that the observed patterns are not primarily driven by left-censoring in the full sample.

4.7.6 Cross-Feature Comparison: Constrained Versus Unconstrained Suggestions

To contextualize the findings for the constrained nudge system, we compared adoption patterns to a second suggestion mechanism on the same platform: a lesson plan chat feature that generates model-inferred suggestions without the constrained label menu. This comparison illuminates how different design philosophies correspond to different behavioral patterns.

Design Contrast The two features represent fundamentally different approaches to AI-generated suggestions:

Constrained (Brainstorm Ideas): Fixed menu of 24 pedagogically grounded labels with 8 heuristic strategies. Suggestions are adapted to context but drawn from a predefined action vocabulary. Nudges persist across conversation turns, remaining clickable for delayed adoption.

Unconstrained (LP Chat): Model-inferred suggestions generated freely from conversation context without predefined constraints. Suggestions are ephemeral, as the clickable buttons displayed until the next response is generated.

Table 4.15: Cross-Feature Adoption Comparison

Metric	Constrained	Unconstrained
Total suggestion displays	567,577	5,280
Total adoptions	22,934	269
Adoption rate	**4.04%**	**5.09%**
95% CI	[3.99%, 4.09%]	[4.53%, 5.72%]
Unique users	8,153	469
Users with ≥1 adoption	52.3%	28.8%

Adoption Rates and Scale Results in Table 4.15 shows that the Lesson Plan Chat shows a modestly higher per-suggestion adoption rate (5.09% vs. 4.04%), a difference of 1.05 percentage points. However, this comparison must be interpreted with substantial caution. The features differ in scale by approximately 100× (567,577 vs. 5,280 displays), serve potentially different user populations (only 469 users engaged with LP Chat vs. 8,153 with Brainstorm), and differ on multiple design dimensions simultaneously.

User Retention: The Critical Difference The most striking difference between features emerges in user retention. As Table 4.16 indicates, among Brainstorm Ideas nudge adopters, 35.8% tried the suggestions exactly once, while 64.2% returned for subsequent adoption. Among LP suggestion adopters, this pattern reverses: 64.4% adopted exactly once, and only 35.6% returned. The difference of 28.6 percentage points is substantial.

Table 4.16: User Retention Comparison

Metric	Constrained	Unconstrained
Users with ≥1 adoption	4,261	135
One-time adopters	35.8%	64.4%
Repeat adopters	64.2%	35.6%

This retention gap suggests that the LP suggestion's slightly higher adoption rate may reflect **trial behavior rather than sustained engagement**. Users try the suggestions at a modestly higher rate, but most do not find sufficient value to return. The nudges achieve slightly lower initial adoption but substantially better retention, most users who try clicking on the nudges find value worth repeating.

Interpretive Implications The cross-feature comparison reveals a **design trade-off between immediate adoption and sustained engagement**. Unconstrained, model-inferred suggestions may have higher novelty appeal or better immediate fit to specific conversational contexts, leading to higher initial adoption. But the lack of pedagogical grounding and ephemeral display may produce inconsistent quality or missed opportunities for delayed adoption, leading to higher abandonment.

Constrained, pedagogically grounded suggestions may have lower immediate appeal, the fixed menu cannot perfectly match every context, but the consistency and persistence may build trust and sustained use. Educators who learn which nudge types are useful in which contexts can deploy them reliably across sessions. More detailed text analysis on nudges and suggestions is included in Appendix C.3.3.

We emphasize that **causal attribution is not supported** by this comparison. The features differ on multiple dimensions (constraint level, persistence, generation mechanism, interface design, user population), and we cannot isolate which dimension drives the observed differences. The comparison is properly understood as illuminating how different design philosophies *correspond to* different behavioral patterns, not which design is "better" in any absolute sense.

4.7.7 Summary of Key Empirical Findings

Across three months of deployment, educators encountered nudges at substantial scale and engaged with them selectively, revealing a consistent pattern that clarifies how nudges function in authentic professional workflows. The findings section has documented several robust patterns that inform both theoretical understanding and practical design. Adoption is selective (4.04%) but reaches substantial proportions of users (52.3%) and conversations (18.0%).

Meaningful delayed adoption (10.3%) validates persistence design. Adoption is highly concentrated (Gini = 0.799) but retention is strong (64.2% repeat adopters). Adoption peaks early in conversations (7.42%) and drops at conclusion (0.42%).

The executability principle: artifact-producing labels substantially outperform abstract/reflective labels. Effort Reduction framing (6.26%) outperforms persuasive framings including Social Proof (2.37%). Within-display paired comparisons confirm that preferences favor concrete outputs. Generative contexts (instructional activity design, differentiation) show elevated adoption. Adoption will also triggers substantive instructional continuation (standards alignment, activity design). Tenure analyses suggested that nudges functioned most strongly as early-use scaffolds, with adoption declining as educators gained experience, while the cross-feature comparison highlighted a trade-off between slightly higher initial adoption for unconstrained suggestions and substantially stronger repeat use for the constrained, persistent, pedagogically grounded nudge design.

4.8 Discussion

This section interprets the empirical findings through theoretical, design, and policy lenses. We situate results within broader conversations about human-AI interaction, educator professional learning, and the ethical deployment of behavioral interventions in professional contexts.

4.8.1 Theoretical Contributions

Nudges as interaction scaffolds, not only recommendations The findings reframe AI-powered nudges from a recommendation paradigm to an **interaction scaffold** paradigm. Unlike traditional recommendation systems that optimize for user engagement or task completion, the nudges studied here function as low-friction affordances that make pedagogically meaningful actions more accessible without prescribing behavior.

Several patterns support this reframing. First, selective adoption (4.04%) indicates that educators exercise substantial discretion, engaging with nudges only when perceived as relevant rather than defaulting to suggestions. Second, delayed adoption (10.3%) demonstrates that nudges persist in working memory as potential actions, available for later retrieval

when context shifts. Third, the early-conversation preference suggests nudges function as direction-setters rather than continuous guides.

This scaffold framing aligns with Vygotskian perspectives on learning support: scaffolds are most effective at the "zone of proximal development" where learners can accomplish with support what they cannot yet accomplish independently [163]. For educators using AI systems, the analogous zone may be the space of instructional actions that are pedagogically sound but not immediately salient given cognitive load and time pressure.

Contextual Nudging vs. Prescribed Interventions This study contributes to nudging research by examining *contextual* nudges that differ fundamentally from the prescribed interventions typical of prior work. Traditional nudging studies, particularly in education, predominantly employ prescribed messages delivered at fixed intervals regardless of user context: text message reminders about attendance, generic prompts about assignment deadlines, or standardized notifications about financial aid [41]. These interventions are designed once and deployed uniformly, with no adaptation to the user's immediate situation or evolving goals.

In contrast, the AI-powered nudges studied here are generated in real time based on conversational context. Each nudge responds to what the educator has just done or requested, proposing next steps that extend the current line of work. This contextual responsiveness creates possibilities unavailable in prescribed nudging: suggestions can match the educator's emerging intent, evolve as the conversation develops, and remain available for later retrieval when context makes them relevant. The 10.3% delayed adoption rate provides direct evidence of this temporal flexibility, educators remember and return to suggestions when circumstances shift, a pattern impossible to observe in studies using one-shot prescribed messages. This contextual approach may better align with how professionals actually work, where goals are under-specified at the outset and evolve through iterative refinement [132, 28].

The Executability Principle A consistent pattern across analyses is that **executability predicts adoption**. Labels producing concrete artifacts (worksheets, tables, rubrics)

outperform labels requiring further elaboration. Effort Reduction strategies (6.26%) outperform identity-based and social framings. Nudges mentioning explicit deliverables are adopted at significantly higher rates (49.8% vs. 35.2%).

We term this the **executability principle**: in time-pressured professional workflows, suggestions are adopted to the extent they shorten the distance between intention and usable output. This principle extends implementation intention theory [66], which demonstrates that specifying concrete action plans increases goal attainment. Here, the AI system performs part of the implementation intention work by translating diffuse goals into specific, actionable requests.

The executability principle carries both promise and caution. On one hand, it suggests a clear design direction: nudges should minimize the cognitive and operational steps between display and value realization. On the other hand, it raises concerns about whether high-executability nudges may inadvertently privilege procedural efficiency over deeper pedagogical reasoning. The low adoption of "Anticipate Misconceptions" (1.7%) and "Promote Higher-Order Thinking" (1.2%), arguably important instructional practices, may reflect this tension.

Concentration and the Long Tail of Adoption The high concentration of adoption (Gini = 0.799, top 10% = 51.6%) challenges assumptions of uniform utility. Nudges are not equally useful to all educators; rather, a subset of users accounts for the majority of engagement.

This concentration pattern is consistent with research on technology adoption in education, which frequently finds that a minority of "power users" account for disproportionate engagement [55]. However, interpreting concentration requires distinguishing between preference heterogeneity (educators differ in their need for scaffolding), design misfit (the current nudge design serves some workflows better than others), and awareness effects (some users may not understand or notice the feature).

The retention data provide partial insight: 64.2% of adopters returned for subsequent use, suggesting that most who try the feature find continued value. However, the 35.8% who tried exactly once may represent design misfit or failed first impressions.

Professional Learning Through Repeated Exposure The tenure analysis reveals a **declining adoption curve**: adoption rates decrease from 7.18% on Day 1 to 2.69% after 60+ days. This pattern suggests that nudges may function as **onboarding scaffolds** that become less necessary as educators develop independent fluency.

This interpretation aligns with theories of legitimate peripheral participation [95] and cognitive apprenticeship [38], which emphasize that learners gradually internalize practices initially supported by external scaffolds. If nudges successfully model productive instructional moves, experienced users may incorporate these moves into their own practice repertoires, reducing reliance on external prompts.

However, alternative interpretations exist. Declining adoption may reflect habituation rather than learning, experienced users may develop workarounds that bypass the nudge interface, or the effect may be compositional (users who remain active after 60 days may differ from those who churned). Distinguishing these mechanisms requires longitudinal analysis beyond the current observation window.

4.8.2 Design Implications

Bounded Choice Architecture The two-nudge display constraint (94.6% of responses showed exactly two options) creates bounded choice points that appear effective. Users are not overwhelmed by extensive option sets, yet have genuine alternatives. This design aligns with research on choice overload showing that excessive options can paralyze decision-making [79].

Design recommendation: AI-powered suggestion interfaces should constrain options to a small set (2–4) of high-quality alternatives rather than presenting exhaustive lists. The constraint forces the system to exercise judgment about what is most relevant, which may be the appropriate locus for AI capability.

Persistence Enables Deferred Action The 10.3% delayed adoption rate validates the persistence design choice. When nudges remained available across turns, educators could return to them when context made them relevant. Ephemeral displays would forfeit this value.

Design recommendation: Suggestion interfaces should maintain visibility of options beyond the immediate display moment, enabling deferred adoption as user needs evolve. The cross-feature comparison reinforces this: the ephemeral feature showed lower retention despite higher immediate adoption.

Framing for Effort Reduction Over Persuasion Effort Reduction framing (6.26%) substantially outperformed persuasive framings like Social Proof (2.37%) and Identity Priming (2.42%). Educators respond to promises of workload reduction more than appeals to professional identity or peer behavior. This finding cautions against importing marketing-style persuasive techniques into professional tools. What works in consumer contexts [92], such as social proof, identity appeals, may be less effective in professional workflows where time pressure dominates.

Design recommendation: Frame AI suggestions around workflow efficiency and concrete outputs rather than persuasive psychological techniques. Respect for professional autonomy may be better expressed through utility than through influence.

4.8.3 Implications for Educator Professional Learning

Job-Embedded Support at Scale The deployment reached 8,153 educators across 127,196 conversations—a scale difficult to achieve through traditional professional development. This reach suggests AI-embedded nudges may address persistent challenges in professional learning: limited access, insufficient personalization, and disconnection from authentic work [42].

However, reach is not impact. The analysis establishes that educators engage with nudges in ways consistent with scaffold use, but does not demonstrate effects on instructional quality or student outcomes. The mechanism, making pedagogically sound actions more accessible at decision points, is plausible as a professional learning support, but efficacy claims require evidence beyond adoption behavior.

Expanding the Salient Option Set Nudges may support professional learning by expanding the set of instructional moves that are **salient and easy to enact** at the point

of choice. Even when educators possess pedagogical knowledge in principle, time pressure and cognitive load can restrict the options that come to mind in practice.

The post-adoption practice analysis shows that adopted nudges can potentially trigger concrete instructional work like standards alignment (11.1%), activity design (5.1%), formative assessment (4.4%). This suggests nudges successfully translate into downstream pedagogical action, though we cannot establish that these actions would not have occurred otherwise.

Cautions About Pedagogical Depth The executability finding raises concerns about depth. If educators preferentially adopt artifact-producing nudges while underadopting reflective practices (Anticipate Misconceptions: 1.7%, Promote Higher-Order Thinking: 1.2%), the nudge system may inadvertently reinforce procedural over conceptual instructional work.

This concern does not imply design failure, educators may appropriately prioritize efficiency in many contexts. But designers should consider whether the current label mix and framing appropriately balance procedural and conceptual practices, and whether additional mechanisms could surface reflective nudges without sacrificing adoption.

4.8.4 Policy Implications

Equity in Access and Benefit The concentration of adoption (top 10% = 51.6%) raises equity questions. If nudge benefits accrue primarily to a subset of users, the system may inadvertently widen rather than narrow capability gaps among educators. Attention should address whether low-adopting populations differ systematically in ways that warrant design intervention, whether nudge content reflects diverse instructional contexts and learner populations, and how to ensure that AI scaffolding benefits are distributed equitably.

Workforce Development Implications The tenure analysis suggests nudges are most valuable during onboarding. This has workforce development implications. AI-embedded nudges may support new teacher induction at scale, districts might strategically deploy nudge-enabled tools during early-career phases, and professional learning budgets might be

reallocated toward technology access for early-career educators.

These implications are speculative and require validation through controlled implementation studies.

4.8.5 Ethical Considerations

Autonomy Preservation The nudge design explicitly aims to preserve autonomy as adoptions are optional, do not block continuation, and do not impose defaults. The 4.04% adoption rate, educators decline nudges 96% of the time, suggests autonomy is preserved in practice. However, autonomy preservation is not simply about optionality. It also requires transparency about intent and mechanism. The current design shows educators the nudge text but not the underlying label, strategy, or generation logic. Fuller transparency might strengthen autonomy by enabling educators to evaluate nudges critically.

Manipulation Concerns Critics of nudging argue that behavioral techniques can be manipulative when they exploit cognitive biases rather than support informed choice [26]. This concern is heightened when nudges are AI-generated, as the system may optimize for engagement over user welfare.

Several features of the current design mitigate manipulation concerns: nudges are generated from a fixed pedagogical menu, not optimized for engagement; the label constraint ensures nudges represent legitimate instructional actions; educators can inspect nudge content before adopting.

However, the heuristic strategies (Effort Reduction, Goal Gradient, etc.) do employ psychological techniques designed to influence choice. Whether this constitutes appropriate persuasion or problematic manipulation may depend on stakeholder values and deployment context.

4.9 Limitations

This study is limited by its observational, descriptive design. We document associations between nudge characteristics and adoption behavior, but cannot establish causal effects

because adoption may be shaped by selection effects (e.g., adopters differing systematically from non-adopters), confounding factors (e.g., underlying preferences driving both nudge uptake and downstream work), and potential reverse causality in which conversation dynamics influence adoption rather than the reverse. In addition, our primary outcome, nudge adoption, is a behavioral proxy rather than an instructional outcome: clicking a nudge indicates perceived relevance but does not demonstrate improved artifact quality, classroom implementation, or student learning, and engagement and impact may diverge [53]. Similarly, the observation window introduces left-censoring for educators whose accounts predate the study period, limiting the interpretation of practice "first occurrences," which may simply reflect the first observable instance within the window.

Generalizability is also constrained by platform and design specificity. The findings reflect a single platform and a particular nudge implementation (fixed 24-label taxonomy, 8-strategy framing vocabulary, and associated prompt engineering), which may not transfer to other educator-facing systems or content domains. Although the cross-feature comparison provides an informative contrast, it has substantial limitations: large scale asymmetry between features, likely differences in user populations, and simultaneous differences across multiple design dimensions (constraint level, persistence, generation mechanism, and interface), along with possible temporal confounds in deployment or promotion. Consequently, cross-feature differences should be interpreted conservatively as correlational patterns tied to design philosophies rather than evidence that any specific design element caused higher adoption or retention.

These limitations bound interpretation but do not invalidate findings. The patterns documented, selective adoption, executability effects, concentration, tenure decline, retention differences, are robust descriptive facts about this deployment. Future research should prioritize experimental validation, downstream outcome measurement, and cross-context generalization.

4.10 Future Directions

Future work should prioritize experimental validation to move from descriptive associations to causal claims about nudge design. Randomized studies can manipulate which labels are

shown, how identical nudge content is framed, whether nudges persist or disappear, how many options are displayed, and when nudges appear within a conversation to test key mechanisms such as executability, bounded choice, persistence-enabled delayed adoption, and early-turn direction setting. In parallel, outcome measurement should extend beyond click behavior by linking adoption to downstream impacts, including expert-rated artifact quality, self-reported classroom implementation, student outcomes where feasible, and pre-post indicators of educator professional growth.

Complementary approaches are needed to explain why and for whom nudges work. Ongoing interviews can surface educators' decision rationales, contextual constraints, and autonomy concerns that are not observable in logs, while longer observation windows can resolve left-censoring and characterize whether tenure effects reflect learning, habituation, or compositional change. Given the high concentration of adoption, heterogeneity analyses should examine predictors of high- versus low-adoption trajectories and how organizational contexts moderate uptake. Finally, systematic comparative design research and cross-platform replication are necessary to isolate which design dimensions drive observed patterns and to establish the boundary conditions of findings across platforms, domains, and cultural contexts.

4.11 Conclusion

This study examined AI-generated pedagogical nudges as interaction scaffolds embedded in open-ended educator-AI conversations, using large-scale deployment data to characterize how educators engage with nudges, which design properties matter for adoption, and how adoption relates to instructional work in practice. Across more than half a million nudge displays, educators engaged selectively rather than habitually, adopting nudges when they reduced friction at key decision points. Adoption was shaped by clear and consistent design patterns: nudges that were executable, artifact-producing, efficiency-oriented, and persistent were more likely to be taken up; adoption clustered early in conversations and among newer users; and sustained engagement was stronger for constrained, pedagogically grounded designs than for unconstrained suggestions. Together, these findings support a reframing of nudges not as recommendations to be followed, but as autonomy-preserving

choice points that scaffold professional action under time and cognitive constraints.

The contribution of this work is twofold. Conceptually, it positions nudges as a form of teacher-facing interaction design that complements, rather than replaces, professional judgment. Empirically, it provides deployment-scale evidence about how such scaffolds function in authentic educator workflows, identifying executability, persistence, and bounded choice as central design levers, while surfacing tensions around concentration, equity, and pedagogical depth. Although the findings are descriptive and platform-specific, they offer a foundation for future experimental and longitudinal research and provide actionable guidance for designers, educators, and policymakers. As GenAI becomes a routine component of professional practice, the challenge is no longer whether AI can generate suggestions, but how those suggestions can be designed to be trustworthy, useful, and ethically grounded. Getting nudge design right, useful enough to adopt, constrained enough to trust, and respectful enough to preserve autonomy, will be critical to realizing AI's potential as a supportive partner in educators' everyday work.

Chapter 5

CONCLUSION: WHAT WE LEARNED AND WHAT COMES NEXT

The findings of this dissertation provide an integrated account of educator-AI collaboration across teacher-facing and classroom-facing contexts. They show that educators' use of generative AI is not simply a matter of producing content, but of engaging in iterative professional design work shaped by authentic instructional goals and real classroom constraints [190, 153, 103]. From this perspective, classroom-ready generative AI depends not only on model capability, but on whether educators can shape AI use in ways that are instructionally valuable in real educational settings. Across both teacher-facing and student-facing tools, this dissertation advances an educator-centered, pedagogy-oriented account of AI use in which instructional value emerges when systems align with authentic professional workflows and iterative instructional reasoning, support bounded and modelable routines for student interaction, and provide workflow scaffolds that reduce friction while preserving professional judgment. Taken together, this work advances an educator-centered design agenda for AI in education, in which instructional value depends on supporting bounded student routines, reducing friction in educator workflows, and preserving educator judgment through transparent, autonomy-preserving forms of assistance.

The three studies provide complementary evidence for this claim. Study 1 shows how educators engage AI in authentic professional practice at scale, revealing the demands of a broad range of AI-supported professional tasks from educators. Study 2 shows how teacher-authored configurations can structure student-AI dialogue by translating pedagogical intent into bounded classroom routines for personalized learning activities. Study 3 shows how lightweight scaffolds embedded in educator-AI workflows can support action at key decision points without displacing educator judgment, instead augmenting educators' enactment of effective instructional practice. Together, these studies position educator use, educator control, and educator decision support as interdependent foundations of feasibility, respon-

sibility, and instructional value in classroom-ready GenAI.

5.1 What the Three Papers Show, Together

5.1.1 Educator-AI work is high-judgment and constraint-saturated

Educators' instructional preparation work is not merely producing text; it is translating pedagogical and content knowledge into coherent, context-responsive instruction [37, 50]. This work is cognitively demanding and time-intensive [83], and it often functions as job-embedded professional learning as educators reason about students, tradeoffs, and feasible enactment [142]. Study 1 strengthens this point by showing that open-ended educator-AI conversations frequently contain iterative cycles of specification, revision, and formatting, reflecting a recurring need to adapt AI outputs into classroom-ready forms rather than accept them as-is. Notably, the instructional practices educators pursued through AI, lesson planning, differentiation, formative assessment, professional reflection, align closely with practices that pre-AI research identified as effective but difficult to implement at scale due to time constraints and limited access to just-in-time support [47, 158, 22]. This continuity suggests that educators are using AI to address longstanding implementation barriers rather than pursuing novel instructional approaches. Study 3 extends the same underlying reality from a behavioral angle. Even when AI output appears fluent, educators may still expend substantial effort steering, repairing, validating, and enhancing to reach educational goals, and they preferentially adopt nudges that reduce that translation burden. This explains why "helpfulness" is not only about output quality; it is about reducing the distance between intent and implementable action while aligning with educational values.

5.1.2 Teacher control and bounded routines are central for AI-mediated classroom viability

Study 2 shifts the focus from teacher-facing drafting to classroom-facing interaction, where teachers must orchestrate activities and monitor many students while maintaining safety, norms, and instructional goals [52, 131, 98]. In this context, the key design question becomes how teacher-authored scaffolds can bound student–AI dialogue so that interaction remains aligned and pedagogically purposeful at scale. Treating the teacher-authored prompt layer

as an instructional design lever makes teacher control explicit: teachers define roles, constraints, and the kinds of reasoning students are expected to engage in, and these decisions shape what is enacted in student-AI conversations. This configuration also enables personalized learning support at classroom scale, AI can provide individualized practice and feedback to students working independently while teachers attend to small groups or students requiring more intensive support [23]. Read together with Study 1, this highlights a continuity across settings. The same heterogeneity of educator goals observed in professional chat makes one-size-fits-all learning tool designs brittle, increasing the value of teacher-authored, context-sensitive bounds.

5.1.3 Nudges work best as optional, repeated choice points that reduce friction

Study 3 provides large-scale evidence that next-step nudges in open-ended educator-AI chat function as interaction scaffolds rather than commands. Adoption is selective rather than habitual, yet remains widespread at the educator level. Educators exercise judgment not only over which practices to adopt, but also over when to adopt them, with delayed uptake accounting for 10% of all adoption. This pattern supports a view of nudges as repeated, autonomy-preserving choice points embedded in multi-turn workflows, consistent with nudging as choice architecture that influences decisions without restricting options [155]. Importantly, these AI-powered nudges differ from the prescribed interventions typical of prior nudging research. Traditional educational nudges, text message reminders, standardized notifications, are designed once and deployed uniformly regardless of context [41]. The contextual nudges studied here are generated in real time based on conversational content, enabling the temporal flexibility that the 10% delayed adoption rate demonstrates: educators remember and return to suggestions when circumstances make them relevant. The results also clarify that educators' decisions are shaped by bounded rationality and workload constraints. Educators are more likely to adopt nudges that reduce effort and produce immediately usable deliverables than nudges that require additional elaboration or reflection. Study 1 provides important context for this result by showing that much educator-AI work centers on producing and refining concrete instructional artifacts, mak-

ing executability and effort reduction especially consequential for AI adoption in educators' workflows.

5.2 Cross-Paper Synthesis: Three Design Principles for Classroom-Ready AI

5.2.1 Principle 1: Design for executability, not just plausibility

Across educator-facing chat and classroom deployment, the decisive factor is often whether the system helps educators move from intention to an executable next step. Study 3's **executability principle** makes this explicit, recommendations that shorten the distance between intent and a usable artifact are adopted more often than abstract prompts. Study 2 complements this by showing that teacher-authored scaffolds similarly function as executable structures for student dialogue, translating goals into bounded routines with finish lines and expectations. Study 1 further motivates this principle by showing that educators' open-ended use frequently requires downstream rework (constraint specification, repair, and formatting) to convert AI output into classroom-ready artifacts, suggesting that system designs should prioritize moves that reduce transmission and translation work rather than merely generating more content.

5.2.2 Principle 2: Bound the option space while preserving autonomy

When operating under cognitive load, evaluating a large menu of next actions in the middle of work becomes challenging. Study 3 shows the value of bounded choice points (e.g., a small number of options), aligning with evidence that excessive choice can overload decision-making [79]. Study 2 shows why boundedness matters even more in classrooms: teachers need student-AI interaction routines that are legible and supervisable with articulated steps and objectives. At the same time, these studies emphasize autonomy preservation: supports should remain optional, non-blocking, and respectful of professional judgment [70, 26]. Study 1 reinforces why autonomy must be preserved: educators' purposes for using AI are diverse and context-dependent, so designs that assume a single "best" path risk misalignment in authentic use.

5.2.3 Principle 3: Support timing and deferral, not only immediate uptake

Both teacher-facing and classroom-facing work unfolds across time. Study 3 demonstrates that persistent nudging messages enable delayed adoption: educators sometimes defer a suggested next step until it becomes relevant later in the conversation, suggesting that "useful" support is not always immediately actionable. This connects to the under-specified nature of educator work documented across the dissertation. Goals and constraints evolve as educators see intermediate outputs and make tradeoffs under constraints [132, 28]. Designing for deferral, keeping options available when timing shifts, aligns system support with authentic workflow dynamics.

5.3 Implications for Educator Professional Learning and System Implementation

This dissertation highlights a central tension in the use of GenAI for education. On one hand, GenAI tools can expand educators' access to just-in-time support. On the other hand, they can also increase the burden of validation, revision, and repair when outputs are too generic, instructionally misaligned, or difficult to adapt to real classroom contexts [88, 140]. This tension matters for professional learning because traditional professional development often cannot provide support that is timely, situated, and responsive to educators' immediate needs, due to access barriers and limited personalization [47, 62]. The findings in this dissertation suggest that AI-embedded scaffolds may help address this gap, but only when they are designed to support, rather than complicate, educators' work. Teacher-authored routines for students (Study 2) and next-step nudges for educators (Study 3) illustrate how AI can provide task-embedded support at scale by making productive actions more visible and easier to enact in the moment. At the same time, Study 1 shows that educator use is highly variable across goals and contexts, which means that effective systems cannot rely on rigid templates or one-size-fits-all workflows. Instead, supports should be designed as configurable scaffolds that can adapt to heterogeneous professional tasks while preserving educators' instructional intent and autonomy. Overall, the dissertation suggests that successful implementation depends not simply on providing access to AI, but on pairing that access with the generation of pedagogically meaningful output and the effective enactment

of the output in real classroom instruction and professional practices.

5.4 Limitations and Boundary Conditions

Several limitations bound interpretation across the dissertation.

Observational and selection effects. Adoption and usage analyses primarily reflect observational patterns. Educators who adopt scaffolds may differ systematically from those who do not, and tasks that elicit scaffolds may differ in complexity and urgency. As a result, associations should not be interpreted as causal without experimental validation [30].

Proxy outcomes. Behavioral signals such as click-based adoption provide conservative evidence of engagement but do not directly establish improvements in artifact quality, classroom enactment, or student learning. These outcomes may diverge from engagement.

Generalizability. Findings are shaped by platform-specific designs (e.g., fixed nudge taxonomies, persistence rules, authored prompt structures) and by the contexts in which educators chose to use the tools. Replication across platforms, subjects, and organizational contexts is necessary to establish boundary conditions.

5.5 Future Research Directions

The dissertation points to several next steps that would strengthen both scientific understanding and practical design.

Experimental validation of design levers. Randomized studies can isolate the causal effects of bounded choice, persistence, framing strategies, and executability on adoption and downstream outcomes.

Linking scaffolds to downstream instructional outcomes. Future work should connect interaction traces to expert-rated artifact quality, teacher-reported enactment, and student-facing outcomes where feasible, rather than relying on engagement proxies alone.

Understanding heterogeneity and equity of benefit. Given concentrated adoption patterns observed in workflow scaffolds, future work should examine who benefits, why, and under what conditions, including how context and resource constraints shape uptake and value.

5.6 Closing

As GenAI becomes embedded in educators' everyday work, the central challenge is no longer whether AI can generate plausible content. The challenge is whether educator-facing and classroom-facing systems can reliably support educators in translating intent into executable, instructionally coherent practice under real-world educational contexts [88, 83]. Across three complementary studies, this dissertation shows that teacher-centered control, bounded routines, and autonomy-preserving interaction scaffolds are foundational mechanisms that shape whether AI becomes a burden of repair or a partner of practice. Heterogeneous and iterative educator work underscores why these mechanisms matter. Classroom-ready AI must meet educators where their work actually happens, often in messy, evolving, high-judgment tasks where the key question is how to move from intention to action with minimal friction and maximal professional integrity.

BIBLIOGRAPHY

[1] Hafiz Farooq Ahmad, Guanghe Sun, and Kinji Mori. Autonomous information provision to achieve reliability for users and providers. In *Proceedings 5th International Symposium on Autonomous Decentralized Systems*, pages 65–72. IEEE, 2001.

[2] Ashraf Alam. Intelligence unleashed: An argument for ai-enabled learning ecologies with real world examples of today and a peek into the future. In *INTERNATIONAL CONFERENCE ON INNOVATIONS IN COMPUTER SCIENCE, ELECTRONICS & ELECTRICAL ENGINEERING-2022*, volume 2717, page 030001. AIP Publishing LLC, 2023.

[3] Vincent Aleven, Elizabeth A McLaughlin, Rebecca A Glenn, and Kenneth R Koedinger. Instruction based on adaptive learning technologies. *Handbook of Research on Learning and Instruction*, 2:522–560, 2016.

[4] Robin Alexander. *A Dialogic Teaching Companion*. Routledge, London, 2020.

[5] Abdullah Alfarwan. Generative ai use in k-12 education: A systematic review. In *Frontiers in Education*, volume 10, page 1647573. Frontiers Media SA, 2025.

[6] Riordan Alfredo, Vanessa Echeverria, Yueqiao Jin, Lixiang Yan, Zachari Swiecki, Dragan Gašević, and Roberto Martinez-Maldonado. Human-centred learning analytics and ai in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6:100215, 2024.

[7] Mohammed Alwaqdani. Investigating teachers perceptions of artificial intelligence tools in education: potential and difficulties. *Education and Information Technologies*, 30(3):2737–2755, 2025.

[8] Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and*

assessing: A revision of Bloom's taxonomy of educational objectives: complete edition. Addison Wesley Longman, Inc., 2001.

[9] Anthropic. How educators use claude, 2024. Accessed 2024.

[10] Christa S. Asterhan, Dor Yosef, Yael Malin, and Talia Gutentag. Vocal participation and attentive listening in teacher-led classroom discussions: The role of prior achievement level, situational experiences and motivation. *International Journal of Educational Research*, 131:102602, 2025.

[11] Roger Azevedo and Allyson F Hadwin. Scaffolding self-regulated learning and metacognition–implications for the design of computer-based scaffolds. *Instructional science*, 33(5/6):367–379, 2005.

[12] Per B. Sederberg, Jonathan F. Miller, Marc W. Howard, and Michael J. Kahana. The temporal contiguity effect predicts episodic memory performance. *Memory & cognition*, 38(6):689–699, 2010.

[13] David Baidoo-Anu and Leticia Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

[14] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International journal of artificial intelligence in education*, 32(4):1052–1092, 2022.

[15] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[16] Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Felipe Zambrano, Alexandra L Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, et al. Chatgpt for education research: Exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*, pages 134–149. Springer, 2024.

[17] SL Beal. Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, pages 941–950, 1987.

[18] Linda Liska Belgrave and Kapriskie Seide. Grounded theory and theoretical coding. *The SAGE handbook of current developments in grounded theory*, pages 167–185, 2019.

[19] RB Bendel, SS Higgins, JE Teberg, and DA Pyke. Comparison of skewness coefficient, coefficient of variation, and gini coefficient as inequality measures within populations. *Oecologia*, 78(3):394–400, 1989.

[20] Ruha Benjamin. Assessing risk, automating racism. *Science*, 366(6464):421–422, 2019.

[21] Abdelkarim Bettahi, Fatima-Zahra Beloudha, and Hamid Harroud. Continuous-time modeling in educational data mining and learning analytics: A literature review on methods, ethics, and emerging ai trends. *IEEE Access*, 2025.

[22] Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1):7–74, 1998.

[23] Benjamin S Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16, 1984.

[24] Hilda Borko. Professional development and teacher learning: Mapping the terrain. *Educational researcher*, 33(8):3–15, 2004.

[25] Melissa Boston. Assessing instructional quality in mathematics. *The Elementary School Journal*, 113(1):76–104, 2012.

[26] Luc Bovens. The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology*, pages 207–219. Springer, 2009.

[27] Catherine M Brighton, Holly L Hertberg, Tonya R Moon, Carol A Tomlinson, and Carolyn M Callahan. The feasibility of high-end learning in a diverse middle school. *National Research Center on the Gifted and Talented*, 2005.

[28] Matthew W Brown. The teacher–tool relationship: Theorizing the design and use of curriculum materials. In *Mathematics teachers at work*, pages 37–56. Routledge, 2011.

[29] Emma Brunskill. Human-centered ai for education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

[30] Kevin C Cain, Siobán D Harlow, Roderick J Little, Bin Nan, Matheos Yosef, John R Taffe, and Michael R Elliott. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *American journal of epidemiology*, 173(9):1078–1084, 2011.

[31] Stefano Calboli and Bart Engelen. Ai-enhanced nudging in public policy: why to worry and how to respond. *Mind & Society*, pages 1–19, 2025.

[32] Courtney B Cazden. *Classroom discourse: The language of teaching and learning.* ERIC, 1988.

[33] Kathy Charmaz. *Constructing grounded theory: A practical guide through qualitative analysis.* sage, 2006.

[34] Yin Hong Cheah, Jingru Lu, and Juhee Kim. Integrating generative artificial intelligence in k-12 education: Examining teachers preparedness, practices, and barriers. *Computers and Education: Artificial Intelligence*, 8:100363, 2025.

[35] Lijia Chen, Pingping Chen, and Zhijian Lin. Artificial intelligence in education: A review. *IEEE access*, 8:75264–75278, 2020.

[36] Robert B Cialdini et al. *Influence: Science and practice*, volume 4. Pearson education Boston, 2009.

[37] David K Cohen, Stephen W Raudenbush, and Deborah Loewenberg Ball. Resources, instruction, and research. *Educational evaluation and policy analysis*, 25(2):119–142, 2003.

[38] Allan Collins, John Seely Brown, and Ann Holum. Cognitive apprenticeship: Making thinking visible. *American Educator*, 15(3):6–11, 38–46, 1991.

[39] Juliet M Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.

[40] Helen Crompton, Mildred V Jones, and Diane Burke. Affordances and challenges of artificial intelligence in k-12 education: A systematic review. *Journal of research on technology in education*, 56(3):248–268, 2024.

[41] Mette Trier Damgaard and Helena Skyt Nielsen. Nudging in education. *Economics of Education Review*, 64:313–342, 2018.

[42] Linda Darling-Hammond, Maria E Hyler, and Madelyn Gardner. Effective teacher professional development. *Learning policy institute*, 2017.

[43] Linda Darling-Hammond, Maria E Hyler, Steve Wojcikiewicz, Joy Rushing, et al. Design principles for teacher preparation: Enacting the science of learning and development. *Learning Policy Institute*, 2025.

[44] Stefano De Paoli. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4):997–1019, 2024.

[45] Amy Dennison et al. Ai-assisted lesson planning: Teacher perceptions and practices. *Journal of Teacher Education*, 2025. In press.

[46] Paul Denny, James Prather, Brett A Becker, James Finnie-Ansley, Arto Hellas, Juho Leinonen, Andrew Luxton-Reilly, Brent N Reeves, Eddie Antonio Santos, and Sami Sarsa. Computing education in the era of generative ai. *Communications of the ACM*, 67(2):56–67, 2024.

[47] Laura M Desimone. Improving impact studies of teachers professional development: Toward better conceptualizations and measures. *Educational researcher*, 38(3):181–199, 2009.

[48] Pierre Dillenbourg. Design for classroom orchestration. *Computers & education*, 69:485–492, 2013.

[49] Pierre Dillenbourg, Luis P Prieto, and Jennifer K Olsen. Classroom orchestration. In *International Handbook of the Learning Sciences*, pages 180–190. Routledge, 2018.

[50] Meixia Ding and Mary Alice Carlson. Elementary teachers' learning to construct high-quality mathematics lesson plans: A use of the ies recommendations. *The Elementary School Journal*, 113(3):359–385, 2013.

[51] Lijuan Dong, Xin Tang, and Xiaoli Wang. Examining the effect of artificial intelligence in relation to students' academic achievement in classroom: A meta-analysis. *Computers and Education: Artificial Intelligence*, page 100400, 2025.

[52] Walter Doyle. Classroom organization and management. *Handbook of Research on Teaching*, 3:392–431, 1986.

[53] Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*, pages 349–372. Springer, 2014.

[54] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[55] Peggy A Ertmer and Anne T Ottenbreit-Leftwich. Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of research on Technology in Education*, 42(3):255–284, 2010.

[56] Lief Esbenshade, Shawon Sarkar, Drew Nucci, Ann Edwards, Sarah Nielsen, Joshua M Rosenberg, Alex Liu, Zewei Tian, Min Sun, Zachary Zhang, et al. Emerging patterns of genai use in k-12 science and mathematics education. *arXiv preprint arXiv:2509.10747*, 2025.

[57] Ushaa Eswaran. Project-based learning: Fostering collaboration, creativity, and critical thinking. In *Enhancing education with intelligent systems and data-driven instruction*, pages 23–43. IGI Global Scientific Publishing, 2024.

[58] Tolulope Famaye, Ibrahim Oluwajoba Adisa, and Golnaz Arastoopour Irgens. To ban or embrace: students perceptions towards adopting advanced ai chatbots in schools. In *International Conference on Quantitative Ethnography*, pages 140–154. Springer, 2023.

[59] Lily Fesler, Thomas Dee, Rachel Baker, and Brent Evans. Text as data methods for education research. *Journal of Research on Educational Effectiveness*, 12(4):707–727, 2019.

[60] Todd Michael Franke, Timothy Ho, and Christina A Christie. The chi-square test: Often used and more often misinterpreted. *American journal of evaluation*, 33(3):448–458, 2012.

[61] Jie Gao, Zhiyao Shu, and Shun Yi Yeo. Mindcoder: Automated and controllable reasoning chain in qualitative analysis. *arXiv preprint arXiv:2501.00775*, 2025.

[62] Michael S Garet, Andrew C Porter, Laura Desimone, Beatrice F Birman, and Kwang Suk Yoon. What makes professional development effective? results from a national sample of teachers. *American educational research journal*, 38(4):915–945, 2001.

[63] Robyn M. Gillies. Promoting academically productive student dialogue during collaborative learning. *International Journal of Educational Research*, 97:200–209, 2019.

[64] Louie Giray. Prompt engineering with chatgpt: a guide for academic writers. *Annals of biomedical engineering*, 51(12):2629–2633, 2023.

[65] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364, 1968.

[66] Peter M Gollwitzer. Implementation intentions: strong effects of simple plans. *American psychologist*, 54(7):493, 1999.

[67] Egon G Guba, Yvonna S Lincoln, et al. Competing paradigms in qualitative research. volume 2, page 105. California, Sage Publications, 1994.

[68] Stian Haaklev, Jim Slotta, Niels Pinkwart, and Pierre Dillenbourg. Orchestration graphs: Enabling rich social pedagogical scenarios in moocs. In *Proceedings of the Fourth ACM Conference on Learning@ Scale*, pages 261–264, 2017.

[69] Jiyoung Han, Haneul Yoo, Jimin Myung, Minsun Kim, Hyunseung Lim, and Juho Kim. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 416–420, 2023.

[70] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European journal of risk regulation*, 4(1):3–28, 2013.

[71] Rachel M Harrison, Anton Dereventsov, and Anton Bibin. Zero-shot recommendations with pre-trained large language models for multimodal nudging. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1535–1542. IEEE, 2023.

[72] John Hattie. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge, 2008.

[73] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[74] Margaret Henningsen and Mary Kay Stein. Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 28(5):524–549, 1997.

[75] Holly Hertberg-Davis. Myth 7: Differentiation in the regular classroom is equivalent

to gifted programs and is sufficient: Classroom teachers have the time, the skill, and the will to differentiate adequately. *Gifted Child Quarterly*, 53(4):251–253, 2009.

[76] Kenneth Holstein and Vincent Aleven. Designing for human–ai complementarity in k–12 education. *AI Magazine*, 43(2):239–248, 2022.

[77] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Designing for complementarity: Teacher and student needs for orchestration support in ai-enhanced classrooms. In *International conference on artificial intelligence in education*, pages 157–171. Springer, 2019.

[78] Weijiao Huang, Khe Foon Hew, and Luke K Fryer. Chatbots for language learningare they really useful? a systematic review of chatbot-supported language learning. *Journal of computer assisted learning*, 38(1):237–257, 2022.

[79] Sheena S Iyengar and Mark R Lepper. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995, 2000.

[80] Qiao Jin, Conrad Borchers, Stephen Fancsali, and Vincent Aleven. Who to help? a time-slice analysis of k-12 teachers' decisions in classes with ai-supported tutoring. In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 640–645, 2025.

[81] Peter D John. Lesson planning and the student teacher: re-thinking the dominant model. *Journal of Curriculum Studies*, 38(4):483–498, 2006.

[82] Eric J Johnson, Suzanne B Shu, Benedict GC Dellaert, Craig Fox, Daniel G Goldstein, Gerald Häubl, Richard P Larrick, John W Payne, Ellen Peters, David Schkade, et al. Beyond nudges: Tools of a choice architecture. *Marketing letters*, 23(2):487–504, 2012.

[83] Nathan D Jones, Eric M Camburn, Benjamin Kelcey, and Esther Quintero. Teachers time use and affect before and after covid-19 school closures. *Aera Open*, 8:23328584211068068, 2022.

[84] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[85] Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965, 2016.

[86] Maurits Kaptein, Panos Markopoulos, Boris De Ruyter, and Emile Aarts. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77:38–51, 2015.

[87] Manu Kapur. Productive failure. *Cognition and instruction*, 26(3):379–424, 2008.

[88] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[89] Mary M Kennedy. How does professional development improve teaching? *Review of educational research*, 86(4):945–980, 2016.

[90] Ran Kivetz, Oleg Urminsky, and Yuhuang Zheng. The goal-gradient hypothesis resurrected: Purchase acceleration, illusionary goal progress, and customer retention. *Journal of marketing research*, 43(1):39–58, 2006.

[91] Andrii O Kolhatin. From automation to augmentation: a human-centered framework for generative ai in adaptive educational content creation. In *CEUR Workshop Proceedings*, volume 4060, pages 143–195, 2025.

[92] Juan Kong and Chen Lou. Beyond persuasion knowledge: Examining the roles of visual appeal, visual congruence, and social proof in influencer advertising. *Journal of Retailing and Consumer Services*, 88:104502, 2026.

[93] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018, 2023.

[94] Marjolein Lanzing. strongly recommended revisiting decisional privacy to judge hypernudging in self-tracking technologies. *Philosophy & Technology*, 32(3):549–568, 2019.

[95] Jean Lave and Etienne Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK, 1991.

[96] LuEttaMae Lawrence, Vanessa Echeverria, Kexin Yang, Vincent Aleven, and Nikol Rummel. How teachers conceptualise shared control with an ai co-orchestration tool: A multiyear teacher-centred design process. *British Journal of Educational Technology*, 55(3):823–844, 2024.

[97] Dabae Lee and Sheunghyun Yeo. Developing an ai-based chatbot for practicing responsive teaching in mathematics. *Computers & Education*, 191:104646, 2022.

[98] Jessica Leek et al. Teacher time allocation in secondary classrooms: Evidence from a large-scale observational study. *Teaching and Teacher Education*, 139:104433, 2024.

[99] Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):29, 2025.

[100] Zixi Li, Chaoran Wang, and Curtis J Bonk. Generative ai for teachers self-directed professional development: A mixed-methods study. *TechTrends*, pages 1–16, 2025.

[101] Karen Littleton, Eileen Scanlon, and Mike Sharples. Orchestrating inquiry learning. In Karen Littleton, Eileen Scanlon, and Mike Sharples, editors, *Orchestrating Inquiry Learning*, pages 47–68. Routledge, London, 2012.

[102] Alex Liu, Lief Esbenshade, Shawon Sarkar, Victor Tian, Zachary Zhang, Kevin He, and Min Sun. How k-12 educators use ai: Llm-assisted qualitative analysis at scale, 2025.

[103] Alex Liu, Lief Esbenshade, Min Sun, Shawon Sarkar, Jian He, Victor Tian, and Zachary Zhang. Adapting to educate: Conversational ai's role in mathematics education across different educational contexts. *arXiv preprint arXiv:2503.02999*, 2025.

[104] Alex Liu and Min Sun. From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews. *AERA Open*, 11:23328584251374595, 2025.

[105] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024.

[106] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023.

[107] Xiner Liu, Jiayi Zhang, Amanda Barany, Maciej Pankiewicz, and Ryan S Baker. Assessing the potential and limits of large language models in qualitative coding. In *International Conference on Quantitative Ethnography*, pages 89–103. Springer, 2024.

[108] Zhaoyang Liu, Wenlan Zhang, and Panpan Yang. Can ai chatbots effectively improve efl learners learning effects?a meta-analysis of empirical research from 2022–2024. *Computer Assisted Language Learning*, pages 1–27, 2025.

[109] Edwin A Locke and Gary P Latham. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705–717, 2002.

[110] Rose Luckin, Karine George, and Mutlu Cukurova. *AI for school teachers*. CRC Press, 2022.

[111] Lisa H MacNeilley. Opening dialogue: understanding the dynamics of language and learning in the english classroom by martin nystrand with adam gamoran, robert kachur, and catherine prendergast. *Language*, 74(2):444–445, 1998.

[112] Soad Matar. The development of educational technology and artificial intelligence and their impact on the future of education: Opportunities and risks. *Journal of Posthumanism*, 5(7):1271–1292, 2025.

[113] Jeanne McClure, Daria Smyslova, Amanda Hall, and Shiyan Jiang. Deductive codings role in AI vs. human performance. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)*, Raleigh, NC, USA, 2024. International Educational Data Mining Society.

[114] Neil Mercer, Lyn Dawes, Rupert Wegerif, and Christine Sams. Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, 30(3):359–377, 2004.

[115] Fengchun Miao, Wayne Holmes, et al. Ai and education: A guidance for policymakers. 2021.

[116] Sarah Michaels, Catherine OConnor, and Lauren B Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297, 2008.

[117] Stuart Mills. Finding the nudgein hypernudge. *Technology in Society*, 71:102117, 2022.

[118] Tobias Mirsch, Christiane Lehrer, and Reinhard Jung. Digital nudging: Altering user behavior in digital environments. *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, pages 634–648, 2017.

[119] Maria Moundridou et al. Teachers' use of ai for educational planning: A systematic review. *Education and Information Technologies*, 2024. Early view.

[120] Ana Mouta, Eva María Torrecilla-Sánchez, and Ana María Pinto-Llorente. Comprehensive professional learning for teacher agency in addressing ethical challenges of aied: Insights from educational design research. *Education and Information Technologies*, 30(3):3343–3387, 2025.

[121] Jennifer Watling Neal, Zachary P Neal, Erika VanDyke, and Mariah Kornbluh. Expediting the analysis of qualitative data in evaluation: A procedure for the rapid identification of themes from audio recordings (rita). *American Journal of Evaluation*, 36(1):118–132, 2015.

[122] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041, 2021.

[123] Drew Nucci and Sarah Nielsen. How and why mathematics teachers use generative AI for instructional tasks. In *Proceedings of the ICERI 2025 Conference*, page 6573, 2025.

[124] OpenAI. AI and the future of teaching and learning: Preparing the education workforce with AI. White paper, 2024.

[125] Ana Ortega-Arranz, I. Amarasinghe, Alejandra Martínez-Monés, Juan I. Asensio-Pérez, Yannis Dimitriadis, M. Corrales-Astorgano, and Davinia Hernández-Leo. Collaborative activities in hybrid learning environments: Exploring teacher orchestration load and students' perceptions. *Computers & Education*, 219:105105, 2024.

[126] Dilara Orynbassarova and Santiago De Lasala Porta. Implementing the socratic method with ai: Opportunities and challenges of integrating chatgpt into teaching pedagogy. In *2024 International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 526–532. IEEE, 2024.

[127] Sharifah Osman, Shahrin Mohammad, Mohd Salleh Abu, Mahani Mokhtar, Jamilah Ahmad, Norulhuda Ismail, and Hanifah Jambari. Inductive, deductive and abductive approaches in generating new ideas: A modified grounded theory study. *Advanced Science Letters*, 24(4):2378–2381, 2018.

[128] Przemyslaw Pawluk and Judi McCuaig[1]. Guides and tellers: A review of tutoring. In *Proceedings of the Future Technologies Conference (FTC) 2025, Volume 4*, volume 4, page 95. Springer Nature, 2025.

[129] Stephanie E Pitts and Sarah M Price. The benefits and challenges of large-scale qualitative research. In *Routledge Companion to Audiences and the Performing Arts*, pages 343–354. Routledge, 2022.

[130] Luis P. Prieto, Martina Holenko Dlab, Ignacio Gutiérrez, Mahmoud Abdulwahed, and Waleed Balid. Orchestrating technology enhanced learning: A literature review and a conceptual framework. *International Journal of Technology Enhanced Learning*, 3(6):583–598, 2011.

[131] Luis P. Prieto, María J. Rodríguez-Triana, Roberto Martínez-Maldonado, Yannis Dimitriadis, and Dragan Gašević. Orchestrating learning analytics (orla): Supporting inter-stakeholder communication about adoption of learning analytics at the classroom level. *Australasian Journal of Educational Technology*, 35(4), 2019.

[132] Janine T Remillard. Examining key concepts in research on teachers use of mathematics curricula. *Review of educational research*, 75(2):211–246, 2005.

[133] Lauren B Resnick and Faith Schantz. Talking to learn: The promise and challenge of dialogic teaching. *Socializing intelligence through academic talk and dialogue*, pages 441–450, 2015.

[134] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.

[135] Jeremy Roschelle, Yannis Dimitriadis, and Ulrich Hoppe. Classroom orchestration: Synthesis. *Computers & Education*, 69:523–526, 2013.

[136] Johnny Saldaña. *The coding manual for qualitative researchers*. SAGE publications Ltd, 2021.

[137] William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of risk and uncertainty*, 1(1):7–59, 1988.

[138] Shawon Sarkar, Alex Liu, R Benjamin Shapiro, and Min Sun. Collaborative and adaptive learning: Designing ai educational systems with and for educators. In *Proceedings of the 19th International Conference of the Learning Sciences-ICLS 2025, pp. 3150-3152*. International Society of the Learning Sciences, 2025.

[139] Klara Sedova, Martin Sedlacek, Roman Svaricek, Martin Majcik, Jana Navratilova, Andrea Drexlerova, and Zuzana Salamounova. Do those who talk more learn more? the relationship between student classroom talk and student achievement. *Learning and Instruction*, 63:101217, 2019.

[140] Neil Selwyn. *Should robots replace teachers?: AI and the future of education*. John Wiley & Sons, 2019.

[141] Guogen Shan, XiangYang Lou, and Samuel S Wu. Continuity corrected wilson interval for the difference of two independent proportions. *Journal of Statistical Theory and Applications*, 22(1):38–53, 2023.

[142] Jianping Shen, Sue Poppink, Yunhuo Cui, and Guorui Fan. Lesson planning: A practice of professional responsibility and development. *Educational horizons*, 85(4):248–258, 2007.

[143] Eunhye Shin. Co-coding classroom dialogue: A single researcher case study of ChatGPT-assisted analysis in science education. *Journal of Computer Assisted Learning*, 2025.

[144] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[145] SJ Buckingham Shum and Rosemary Luckin. Learning analytics and ai: Politics, pedagogy and practices. *British journal of educational technology*, 50(6):2785–2793, 2019.

[146] Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and Luettamae Lawrence. The role of generative AI in qualitative research: GPT-4s contributions to a grounded

theory analysis. In *Proceedings of the Symposium on Learning, Design and Technology (LDT '24)*, pages 1–13, New York, NY, USA, 2024. ACM.

[147] Benson Soong. *Improving secondary students' revision of physics concepts through computer-mediated peer discussion and prescriptive tutoring*. PhD thesis, 2010.

[148] Mary Kay Stein, Brian W. Grover, and Margaret Henningsen. Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 33(2):455–488, 1996.

[149] Anselm Strauss and Juliet Corbin. *Basics of qualitative research techniques*. Thousand oaks, CA: Sage publications, 1998.

[150] Cass R Sunstein. Nudging: a very short guide. *The Handbook of Privacy Studies*, 37:583–588, 2014.

[151] Cass R Sunstein. The ethics of nudging. *Yale J. on Reg.*, 32:413, 2015.

[152] Teo Susnjak, Gomathy Suganya Ramaswami, and Anuradha Mathrani. Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12, 2022.

[153] Xiao Tan, Gary Cheng, and Man Ho Ling. Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, 8:100355, 2025.

[154] Sumanth Tatineni. Recommendation systems for personalized learning: A data-driven approach in education. *Journal of Computer Engineering and Technology (JCET)*, 4(2):18–31, 2020.

[155] Richard H Thaler and Cass R Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[156] Mike Tissenbaum and Jim Slotta. See the teacher, see the student: A study of teacher noticing during orchestration of collaborative inquiry-based science learning. *Journal of Computer Assisted Learning*, 35(6):696–710, 2019.

[157] Maciej Tomczak and Ewa Tomczak. The need to report effect size estimates revisited. an overview of some recommended measures of effect size. 2014.

[158] Carol Ann Tomlinson. *The differentiated classroom: Responding to the needs of all learners*. Ascd, 2014.

[159] Amos Tversky, Daniel Kahneman, Paul Slovic, et al. *Judgment under uncertainty: Heuristics and biases*. Cambridge, 1982.

[160] UNESCO. Guidance for generative ai in education and research. Technical report, United Nations Educational, Scientific and Cultural Organization, Paris, 2023.

[161] Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. Human-centered artificial intelligence: Designing for user empowerment and ethical considerations. In *2023 5th international congress on human-computer interaction, optimization and robotic applications (HORA)*. IEEE, 2023.

[162] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[163] Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

[164] Yoshija Walter. Embracing the future of artificial intelligence in the classroom: the relevance of ai literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1):15, 2024.

[165] Kai Wang. From ELIZA to ChatGPT: A brief history of chatbots and their evolution. *Applied and Computational Engineering*, 39(1):57–62, 2024.

[166] Qile Wang, Moath Erqsous, Kenneth E. Barner, and Matthew Louis Mauriello. Lata: A pilot study on LLM-assisted thematic analysis of online social network data generation experiences. *Proceedings of the ACM on Human-Computer Interaction*, 9(2), April 2025.

[167] Xi Wang. Research on the application of ai technology in computer-assisted instruction. In *Journal of Physics: Conference Series*, volume 1992, page 022030. IOP Publishing, 2021.

[168] Zhenwen Wang, Zekai Qin, Yijun Liu, Zilong Liu, and Wei Wang. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. *arXiv preprint arXiv:2405.16990*, 2024.

[169] Mark Warschauer and Thomas Tate. New technology and digital equity: Reframing the familiar debate. *Educational Researcher*, 52(6):333–345, 2023.

[170] Norman L Webb. Criteria for alignment of expectations and assessments in mathematics and science education. *Research Monograph No. 6*, 1997.

[171] Norman L Webb. Depth-of-knowledge levels for four content areas. *Language Arts*, 28(March):1–9, 2002.

[172] Markus Weinmann, Christoph Schneider, and Jan vom Brocke. Digital nudging. *Business & Information Systems Engineering*, 58(6):433–436, 2016.

[173] Diana Weiß, Johannes Scheuerer, Michael Wenleder, Alexander Erk, Mark Gülbahar, and Claudia Linnhoff-Popien. A user profile-based personalization system for digital multimedia content. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, pages 281–288, 2008.

[174] Jeromie Whalen, Chrystalla Mouza, et al. Chatgpt: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23, 2023.

[175] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In *Generative AI for Effective Software Development*, pages 71–108. Springer, 2024.

[176] Grant P Wiggins and Jay McTighe. *Understanding by design*. Ascd, 2005.

[177] Dylan Wiliam. *Embedded formative assessment*. Solution Tree Press, 2011.

[178] Rainer Winkler and Matthias Söllner. Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings*, 2018.

[179] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924, 2021.

[180] Jeffrey M Wooldridge. *Introductory econometrics a modern approach*. South-Western cengage learning, 2016.

[181] Dirk U Wulff, Zak Hussain, and Rui Mata. The behavioral and social sciences need open LLMs. *OSF Preprints*, 2024.

[182] Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings. *arXiv preprint arXiv:2503.15620*, 2025.

[183] Bin Xu, Nian-Shing Chen, and Guodong Chen. Effects of teacher role on student engagement in wechat-based online discussion learning. *Computers & Education*, 157:103956, 2020.

[184] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024.

[185] Stephen JH Yang, Hiroaki Ogata, Tatsunori Matsui, and Nian-Shing Chen. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2:100008, 2021.

[186] Karen Yeung. hypernudge: Big data as a mode of regulation by design. In *The social power of algorithms*, pages 118–136. Routledge, 2019.

[187] Senem Zaimoğlu and Aysun Dağtaş. Teacher cognition and practices in using generative ai tools to support student engagement in efl higher-education contexts. *Behavioral Sciences*, 15(9):1202, 2025.

[188] Andres Felipe Zambrano, Zhanlan Wei, Jiayi Zhang, Ryan S Baker, Jaclyn Ocumpaugh, Amanda Barany, Xiner Liu, Jeffrey Ginger, Luc Paquette, Yiqiu Zhou, et al. Data plus theory equals codebook: Leveraging llms for human-ai code-book development.

[189] J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny cant 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, 2023.

[190] Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International journal of educational technology in higher education*, 16(1):1–27, 2019.

[191] Xiaoming Zhai. Transforming teachers roles and agencies in the era of generative ai: Perceptions, acceptance, knowledge, and practices. *Journal of Science Education and Technology*, 34(6):1323–1333, 2025.

[192] Ke Zhang and Ayse Begum Aslan. Ai technologies for education: Recent research & future directions. *Computers and education: Artificial intelligence*, 2:100025, 2021.

[193] Xin Zhang, Peng Zhang, Yuan Shen, Min Liu, Qiong Wang, Dragan Gašević, and Yizhou Fan. A systematic literature review of empirical research on applying gener-

ative artificial intelligence in education. *Frontiers of Digital Education*, 1(3):223–245, 2024.

[194] Tian Zhou, Jo Tondeur, and Sarah Howard. Ai competency frameworks for teachers: A systematic review. *EdMedia+ Innovate Learning*, pages 636–642, 2025.

Appendix A

STUDY 1 APPENDIX

A.1 Prompts for LLM Coding

A.1.1 Prompt for Open Coding (Step 1)

```
You are an expert educator. Your task is to identify K-12 educational components in the
↪ chunk of educator-AI dialogue.
You are provided with a request, its response, and corresponding follow-up request if
↪ available.

Request: {request}
Response: {response}
Follow-up Response: {follow-up response}

1. Content Subject Domain Area:
   Output the specific subject and its domain areas if applicable
   (e.g., ``Math/Geometry``, ``Science/Physics'', ``Science/Motion``).

2. Pedagogical strategies:
   Output specific strategies if applicable
   (e.g., ``Formative Assessment``, ``Group Work``, ``Student Discourse``,
   ``Connect to Real-world``).

3. Instructional Strategy (1-5 scale):
   - 5: The command clearly addresses instructional strategies.
   - 4: Touches on strategies but lacks depth.
   - 3: Implies strategies without a clear focus.
   - 2: Vaguely references instructional strategies.
   - 1: No instructional strategy mentioned.

4. Request for AI Tool Support:
   Output the specific support the teacher is asking for.
```

```
Format your responses for each item in short phrases.
Provide the results in the strict JSON structure:

{
  ``Request``: {request},
  ``results'': {
    ``Content subject domain area``: ``subject area'' or null,
    ``Pedagogical strategies'': ``strategy'' or null,
    ``Instructional Strategy``: score,
    ``Request for AI Tool Support``: ``support'' or null
  }
}
```

A.1.2 Prompt for Open Coding (Expanded Schema)

```
You are an expert educator. Your task is to identify K-12 educational components
in the chunk of educator-AI dialogue.
You are provided with a request, its response, and corresponding follow-up request
if available.

Request: {request}
Response: {response}
Followup Request: {followup_request}

1. Content Subject Domain Area:
   Output the specific subject and its domain areas if applicable
   (e.g., ``Math/Geometry``, ``Science/Physics'', ``Science/Motion``,
   ``Social Science``).

2. Contextual Information:
   Output components that are specifically contextualized for the teacher's
   K-12 educational setting. For example, grade level, student demographics,
   instructional challenges, student needs such as Special Education (IEP/504)
   or English Language Learners (ELLs), or teacher needs.
   If grade level is identified, format grades as:
   ``Grade_K``, ``Grade_1``, ``Grade_10``, or ``Grade_Hs''.
```

3. Instructional Tasks:
   Output specific instructional tasks or pedagogical strategies requested
   (e.g., ``Formative Assessment``, ``Group Work``, ``Student Discourse``,
   ``Incorporate Real-world Example``, ``Reflective Learning``,
   ``Deep Questions'').

4. Instructional Strategy (1-3 scale):
   - 3: The request clearly addresses instructional strategies.
   - 2: Implies or touches on strategies but lacks a clear focus.
   - 1: No instructional strategy mentioned or implied.

5. Learning Progression (1-3 scale):
   Output keywords describing students' prior knowledge or ideas assumed
   or referenced in the request.

6. Pedagogical Frameworks:
   Output any explicitly referenced pedagogical frameworks
   (e.g., ``UDL``, ``World Cafe``).

7. Request for AI Tool Support:
   Output the specific AI support requested
   (e.g., translation, rubric creation).

8. Category:
   Categorize the AI response using ONE label:
   - Information
   - Explanation
   - Guidance
   - Question
   - Summarization

   Notes:
   - ``Question'' applies only to clarifying or contextual questions.
   - Do NOT include generic follow-up questions.

Follow the output format requirement strictly.

```
DO NOT add additional text or reasoning.
Provide results in the strict JSON structure:

{
  ``Results'': {
    ``Content subject domain area``: ``subject_area'' or null,
    ``Contextual Information``: ``educational_context'' or null,
    ``Instructional Tasks'': ``task'' or null,
    ``Instructional Strategy``: score,
    ``Learning Progression``: ``prior ideas'' or null,
    ``Pedagogical Frameworks'': ``framework'' or null,
    ``Request for AI Tool Support``: ``support'' or null,
    ``Response Category``: ``category label'' or null
  }
}
```

A.1.3 Prompt for Selective Coding

```
<SystemPrompt>
<Description>
   You are an expert meta-evaluator. Your task is to evaluate AI-generated responses to the
   ↪ user requests from K-12 educators during lesson planning and judge their quality.
   You are provided with a request, its response, and response's followup_request.
   Evaluate the trio using their respective rubrics. Additionally, the quality of the
   ↪ responses should be determined based on the corresponding requests, and the
   ↪ evaluation of the followup_request associates with the responses.
</Description>
<EvaluationInputs>
   <Request>{request}</Request>
   <Response>{response}</Response>
   <FollowupRequest>{followup_request}</FollowupRequest>
</EvaluationInputs>
<Rubrics>
   <RequestRubric>
       <ContentSubjectDomainArea>
           <Instruction>
```

```
    Output the specific subject and its domain areas if applicable. Select THE
    ↪  CLOSEST K-12 subject labels for the Subject Area from the list below:
</Instruction>
<LabelList>
    <Label>ELA/Reading</Label>
    <Label>ELA/Literacy Analytics</Label>
    <Label>ELA/Writing</Label>
    <Label>ELA/Listening</Label>
    <Label>ELA/Speaking</Label>
    <Label>ELA/Grammar</Label>
    <Label>ELA/Vocabulary</Label>
    <Label>ELA/Phonics</Label>
    <Label>ELA/General</Label>
    <Label>Mathematics/Arithmetic</Label>
    <Label>Mathematics/Algebra</Label>
    <Label>Mathematics/Geometry</Label>
    <Label>Mathematics/Statistics</Label>
    <Label>Mathematics/Calculus</Label>
    <Label>Mathematics/General</Label>
    <Label>Science/Life Science</Label>
    <Label>Science/Chemistry</Label>
    <Label>Science/Physics</Label>
    <Label>Science/Earth Science</Label>
    <Label>Science/Environmental Science</Label>
    <Label>Science/General</Label>
    <Label>Social Studies/General</Label>
    <Label>Social Studies/U.S. History</Label>
    <Label>Social Studies/World History</Label>
    <Label>Social Studies/Geography</Label>
    <Label>Social Studies/Civics</Label>
    <Label>Social Studies/Economics</Label>
    <Label>World Languages/Learn Language</Label>
    <Label>World Languages/Translation Only</Label>
    <Label>World Languages/American Sign Language</Label>
    <Label>Physical Education and Health</Label>
    <Label>Social Emotional Learning</Label>
    <Label>Arts</Label>
```

```
        <Label>Technology/Computer Science</Label>
        <Label>Technology/Digital Literacy</Label>
        <Label>Career and Technical Education (CTE)/Agriculture</Label>
        <Label>Career and Technical Education (CTE)/Business</Label>
        <Label>Career and Technical Education (CTE)/Culinary Arts</Label>
        <Label>Career and Technical Education (CTE)/Engineering</Label>
    </LabelList>
</ContentSubjectDomainArea>
<Grades>
    <Instruction>
        Identify grade levels for the Request. Grade levels can be from pre-K, K-12,
        ↪ to higher edu
        If grade levels are explicitely mentioned, use the format 'Explicit:
        ↪ Grade_*' (e.g., 'Explicit: Grade_1', ' Explicit: Grade_HS').
        Otherwise, if you identified grade levels without explicit mentioning, use
        ↪ the format 'Implicit: Grade_*' (e.g., 'Implicit: Grade_K', 'Implicit:
        ↪ Grade_College').
    </Instruction>
</Grades>
<ContextualInformation>
    <Instruction>
        Identify the educational context indicated in user Request.
        Match your identified contexts to the closest labels from Context Options;
        ↪ if Other is selected, replace label text 'Specification' by specifying
        ↪ the context briefly:
    </Instruction>
    <Options>
        <Context>After School Program</Context>
        <Context>Homeschool</Context>
        <Context>Classroom Setting/Socio-Economic Factors</Context>
        <Context>Classroom Setting/Reluctant Participants or Engagement</Context>
        <Context>Classroom Setting/Low-Tech Educational Environment</Context>
        <Context>Student's Needs/Special Education (SpEd)</Context>
        <Context>Student's Needs/English Language Learners (ELL)</Context>
        <Context>Student's Needs/Learners from Low-Income Households</Context>
        <Context>Student's Needs/Advanced or Gifted Learners</Context>
        <Context>Student's Needs/Below Grade Level</Context>
```

```
        <Context>Student's Needs/Mixed Ability Learners</Context>
        <Context>Student's Needs/Behavioral and Emotional Support</Context>
        <Context>Teacher's Needs/Classroom Management and Climate</Context>
        <Context>Teacher's Needs/Assessment and Student Progress
        ↪ Monitoring</Context>
        <Context>Teacher's Needs/Professional Development and Training</Context>
        <Context>Non-educational</Context>
        <Context>Other/Specification</Context>
    </Options>
</ContextualInformation>
<LearningProgression>
    <Instruction>
        Output components that indicate students' prior knowledge or students' ideas
        ↪ that are assumed or mentioned in the Request.
    </Instruction>
</LearningProgression>
<ContentFocus>
    <Instruction>
        Identify the type of content focus that the educator is working on.
        Match your identified content focus to the closest labels from the Content
        ↪ Codebook; if Other is selected, replace label text 'Specification' by
        ↪ specifying the content focus briefly:
    </Instruction>
    <Codebook>
        <Content>Assessment and Evaluation/Formative Assessment</Content>
        <Content>Assessment and Evaluation/Summative Assessment</Content>
        <Content>Assessment and Evaluation/Rubrics and Grading</Content>
        <Content>Assessment and Evaluation/Project-Based Assessment</Content>
        <Content>Assessment and Evaluation/Provide Feedback for Assessment</Content>
        <Content>Assessment and Evaluation/Progress Monitoring and Data
        ↪ Tracking</Content>
        <Content>Lesson and Curriculum Planning/Plan for an Entire Lesson</Content>
        <Content>Lesson and Curriculum Planning/Aligning Instruction Materials with
        ↪ Learning Standards</Content>
        <Content>Lesson and Curriculum Planning/Backward Planning and Long-Term
        ↪ Planning for Unit or Semester</Content>
```

```
            <Content>Lesson and Curriculum Planning/Design or Adjust Classroom
            ↪ Activities</Content>
            <Content>Professional Development/Teacher Learning and Professional Learning
            ↪ Community (PLC)</Content>
            <Content>Professional Development/Teacher Training and Workshops</Content>
            <Content>Communicate with Community and Parents</Content>
            <Content>Administrators and Staff Communication</Content>
            <Content>Other/Specification</Category>
        </Codebook>
    </ContentFocus>
    <InstructionalPractices>
        <Instruction>
            Identify the instructional practices that are inquired by user Request.
            Match your identified practices to the closest labels from the Practice
            ↪ Codebook; if Other is selected, replace label text 'Specification' by
            ↪ specifying the practice briefly:
        </Instruction>
        <Codebook>
            <Practice>Differentiation and Accessibility/Differentiated Instruction and
            ↪ Engagement Strategies</Practice>
            <Practice>Differentiation and Accessibility/Special Needs
            ↪ Accommodations</Practice>
            <Practice>Differentiation and Accessibility/Language and Translation
            ↪ Support</Practice>
            <Practice>Differentiation and Accessibility/Tiered Scaffolding and Support
            ↪ to Skill Levels</Practice>
            <Practice>Differentiation and Accessibility/Visual Representation and Visual
            ↪ Aids</Practice>
            <Practice>Real-World Applications and Project-Based Learning/Real-World
            ↪ Examples</Practice>
            <Practice>Real-World Applications and Project-Based Learning/Project-Based
            ↪ Learning</Practice>
            <Practice>Real-World Applications and Project-Based Learning/Hands-On
            ↪ Learning and Manipulatives</Practice>
            <Practice>Collaborative and Group Learning/Group Work and Peer
            ↪ Collaboration</Practice>
            <Practice>Collaborative and Group Learning/Community Building</Practice>
```

```
    <Practice>Collaborative and Group Learning/Student Discourse</Practice>
    <Practice>Collaborative and Group Learning/Gallery Walk</Practice>
    <Practice>Questioning, Critical Thinking, and Inquiry/Critical Thinking and
    ↪  High-Level Cognition</Practice>
    <Practice>Questioning, Critical Thinking, and Inquiry/Inquiry-Based
    ↪  Learning</Practice>
    <Practice>Questioning, Critical Thinking, and Inquiry/Deep Question
    ↪  Generation</Practice>
    <Practice>Explicit Teaching/Explain Concepts, Terminology, and Subject
    ↪  Specific Language</Practice>
    <Practice>Explicit Teaching/Modeling Problem-Solving Processes</Practice>
    <Practice>Learning Progression/Classroom Instruction Routine
    ↪  Adjustment</Practice>
    <Practice>Classroom Management and Climate/Foster Positive Classroom
    ↪  Culture</Practice>
    <Practice>Student Engagement and Motivation/Design or Adjust Engagement
    ↪  Strategies</Practice>
    <Practice>Student Engagement and Motivation/Gamification and
    ↪  Motivation</Practice>
    <Practice>Feedback, Reflection, and Progress Monitoring/Feedback to
    ↪  Students</Practice>
    <Practice>Feedback, Reflection, and Progress Monitoring/Reflective Learning
    ↪  Approaches</Practice>
    <Practice>Feedback, Reflection, and Progress Monitoring/Data-Driven
    ↪  Instruction</Practice>
    <Practice>Literacy and Language Development (ELA-Focused)/Reading
    ↪  Comprehension</Practice>
    <Practice>Literacy and Language Development (ELA-Focused)/Writing
    ↪  Skills</Practice>
    <Practice>Literacy and Language Development (ELA-Focused)/Literary
    ↪  Analysis</Practice>
    <Practice>Literacy and Language Development (ELA-Focused)/Vocabulary
    ↪  Development (ELA)</Practice>
    <Practice>Technology Integration and Digital Tools/Multi-media and
    ↪  Technology for Instruction</Practice>
    <Practice>Other/Specification</Practice>
</Codebook>
```

```
        </InstructionalPractices>
        <PedagogicalFrameworks>
            <Instruction>
                Output any specific pedagogical frameworks explicitly mentioned in the
                ↪ Request (e.g., UDL, Word Cafe, etc.).
            </Instruction>
        </PedagogicalFrameworks>
 </RequestRubric>
<ResponseRubric>
   <Category>
        <Description>
            Categorize the AI Response using the definitions below. Each label reflects the
            ↪ primary intent and nature of the AI Response:
        </Description>
        <Labels>
            <Label>
                <Name>Information</Name>
                <Definition>
                    The AI Response provides relevant information, examples, definitions, or
                    ↪ a range of options. It delivers factual or contextual details
                    ↪ without actionable steps or deep explanations.
                </Definition>
            </Label>
            <Label>
                <Name>Explanation</Name>
                <Definition>
                    The AI Response explains or illustrates a concept in more detail, aiming
                    ↪ to clarify or expand understanding. It focuses on conceptual or
                    ↪ theoretical aspects without necessarily offering actionable advice.
                </Definition>
            </Label>
            <Label>
                <Name>Guidance</Name>
                <Definition>
                    The AI Response provides specific actionable advice, step-by-step
                    ↪ instructions, or strategies tailored to the user's request. It
                    ↪ focuses on helping the user implement or execute a task effectively.
```

```
            </Definition>
        </Label>
        <Label>
            <Name>Question</Name>
            <Definition>
                The AI Response follows up with clarifying or contextual questions to
                ↪ actively seek specific new information from the user. This excludes
                ↪ generic inquiries like ``Let me know if you need more help.``
            </Definition>
        </Label>
        <Label>
            <Name>Summarization</Name>
            <Definition>
                The AI Response condenses and concludes key points, summarizing the
                ↪ information provided earlier. It aims to maintain continuity by
                ↪ organizing details concisely.
            </Definition>
        </Label>
    </Labels>
    <Instruction>
        Output one of the following labels based on the AI Response: ``Information``,
        ↪ ``Explanation``, ``Guidance``, ``Question``, or ``Summarization``.
    </Instruction>
</Category>
</ResponseRubric>
  </Rubrics>
Output Format: Follow exactly to the output format requirement. DO NOT add additional text
↪ or reasonings besides the JSON object. *DO NOT NEED TO SAY 'Here is my evaluation of
↪ the request, response, and followup request:' OR ANYTHING SIMILAR TO THIS AT
↪ BEGINNING.*
Provide the results in the *strict JSON structure*:
  {
    ``Request_scores'': {
      ``Content subject domain area``: ``subject_area'' or null,
      ``Grades'': ``grades'' or null,
      ``Contextual Information``: ``educational_context'' or null,
      ``Learning Progression``: ``prior ideas'' or null,
```

```
      ``Content Focus'': ``content focus'' or null,
      ``Instructional Practices'': ``practice'' or null,
      ``Pedagogical Frameworks'': ``framework'' or null
    },
    ``Followup Request``: {
      ``Follow-up Dynamics'': ``user reaction'' or null
    }
  }

</SystemPrompt>
```

A.1.4 Prompt for Deductive Coding

```
You are an expert instructional coach and evaluator. Your task is to identify what
↪ information about K12 education is being conveyed in the message.
First identify and extract the meta information; second annotate each message using the
↪ structured taxonomy of Annotation Categories below.
This will help categorize the instructional goals, student needs, assessment strategies,
↪ and professional concerns reflected in K12 education.

Meta information
<EdContext>
Identify and extract following K12 educational context demonstrated in the message:
   Subject Area: e.g., Mathematics.
   Grade Level: e.g., Grade 6, Grade High School.
   Pedagogical Framework: established pedagogical framework that is explicitly mentioned
   ↪ and used in the message.
</EdContext>

Annotation Categories: Use only the categories and items below. DO NOT generate new labels
↪ or summaries.
<Instructional Practices>
   Differentiation and Accessibility: Differentiated Instructional Strategies, Multilingual
   ↪ Learner Support, Tiered Scaffolding, Integrate Visual Representation
   Explicit Teaching: Modeling Problem Solving, Explaining Core Science Concepts,
   ↪ Explaining Math Concepts, Facilitate Procedural Fluency, Crosscutting STEM Concepts,
   ↪ Scientific Practices, ELA Skills Development
```

```
   Project-Based and Real-World Learning: Real-World Engagement and Scenarios, Projects,
   ↪ Hands-On Activities, Experimental Design, Engineering and Design, Technical Skill
   ↪ Development, Industry Standards
   Collaborative Learning: Group Work, Student Discourse
   Critical Thinking and Inquiry: Encourage Critical Thinking and High-Level Cognition,
   ↪ Inquiry and Deep Questions, Historical Thinking, Civics and Government, Geography
   ↪ and Human Systems, Economics and Decision Making, Research and Source Analysis
   Instructional Routine: Learning Progression and Routine Adjustments
   Engagement and Motivation: Actionable Engagement Strategy
</Instructional Practices>

<Student Needs and Context>
   Classroom Setting: Socio-Economic, Low-Tech, Reluctant Learners, Student Behavioral
   ↪ Intervention, Homeschool
   Student Profiles: Special Education (IEP/504), English Language Leaners (ELL),
   ↪ Low-Income (FLR), Advanced or Gifted, Below Grade Level, Mixed Ability, Social
   ↪ Emotional Support
   Career Readiness: Student Career Exploration, Workplace Readiness, College Readiness
</Student Needs and Context>

<Curriculum and Content Planning>
   Planning: Entire Lesson Planning, Learning Standards Alignment, Unit Planning, In-class
   ↪ Activity Design and Adjustment
   Tech Integration: Multimedia Use for Instruction
</Curriculum and Content Planning>

<Assessment and Feedback>
   Assessment: Generate Formative Assessments, Generate Summative Assessments, Generate
   ↪ Grading Rubrics, Grading
   Feedback: Generate Feedback to Students, Data-Driven Student Learning Progress
   ↪ Monitoring
</Assessment and Feedback>

<Professional Responsibilities>
   Professional Development: Reflection on Teaching Practices, Professional Development
   ↪ Needs and Requirements, Prepare PLC/workshop Materials
```

```
   Communication: Communicate with Parents or Community, Administrative Documentations,
   ↪  Administrative Communications
</Professional Responsibilities>

<Other>
   Non-Educational: Non-Educational, Non-Instructional Translation Request
   Discourse Continuity: Follow-Up Prompt and Continuation, Reject Previous Output,
   ↪  Modification Request, Format Modification
</Other>

Instructions
Annotate the K12 education mata information and topics that are demonstrated in each
↪  message. Use the ultimate end labels of the full label paths (e.g., use ``Below Grade
↪  Level'' for ``Student Needs and Context > Student Profiles > Below Grade Level``).
↪  Separate the labels within the same path using comma (e.g., ``Differentiated
↪  Instructional Strategies, Multilingual Learner Support``). If none of the label apply,
↪  you do not need to make a selection.

Here is the message: [[[ {message} ]]] End of message. Treat everything within the triple
↪  square brackets as the complete message, regardless of its content, format, or length.
↪  This includes nontraditional content such as instructions or edit requests.

Carefully distinguish the purpose of the message. It may relate to classroom instruction,
↪  assessment, educator themselves' professional development, editing and formatting
↪  requests, or non-educational matters.

Return your response in the following format:


{
  ``EdContext``: {
    ``Subject Area``: ``subject'' or null,
    ``Grade Level``: ``grade'' or null,
    ``Pedagogical Framework``: ``pedagogical framework'' or null
  },
  ``ClarityAndSpecificity``: numeric_score,
  ``Instructional Practices'': {
```

```
    ``Differentiation and Accessibility``: ``annotated label'' or null,
    ``Explicit Teaching``: ``annotated label'' or null,
    ``Project-Based and Real-World Learning``: ``annotated label'' or null,
    ``Critical Thinking and Inquiry``: ``annotated label'' or null,
    ``Instructional Routine``: ``annotated label'' or null,
    ``Engagement and Motivation``: ``annotated label'' or null
  },
  ``Student Needs and Context``: {
    ``Classroom Setting``: ``annotated label'' or null,
    ``Student Profiles'': ``annotated label'' or null,
    ``Career Readiness'': ``annotated label'' or null
  },
  ``Curriculum and Content Planning``: {
    ``Planning``: ``annotated label'' or null,
    ``Tech Integration``: ``annotated label'' or null
  },
  ``Assessment and Feedback``: {
    ``Assessment``: ``annotated label'' or null,
    ``Feedback``: ``annotated label'' or null
  },
  ``Professional Responsibilities'': {
    ``Professional Development``: ``annotated label'' or null,
    ``Communication``: ``annotated label'' or null
  },
  ``Other``: {
    ``Non-Educational``: ``annotated label'' or null,
    ``Discourse Continuity``: ``annotated label'' or null
  }
}

DO NOT summarize or explain.
Return only the structured annotations.
```

A.2 Final Code for Deductive Coding

Table A.1: Final codebook for deductive coding across professional domains.

Professional Domain (Category 1)	**Professional Domain (Category 2)**	**Category**	**Action Item**
Instructional Practices		Differentiation and Accessibility	Differentiated Instructional Strategies
Instructional Practices		Differentiation and Accessibility	Multilingual Learner Support
Instructional Practices		Differentiation and Accessibility	Tiered Scaffolding
Instructional Practices	Curriculum and Content Focus	Differentiation and Accessibility	Integrate Visual Representation
Instructional Practices		Project-Based and Real-World Learning	Real-World Engagement and Scenarios
Instructional Practices		Project-Based and Real-World Learning	Projects
Instructional Practices		Project-Based and Real-World Learning	Hands-On Activities
Instructional Practices	Curriculum and Content Focus	Project-Based and Real-World Learning	Experimental Design
Instructional Practices	Curriculum and Content Focus	Project-Based and Real-World Learning	Engineering and Design
Instructional Practices	Curriculum and Content Focus	Project-Based and Real-World Learning	Technical Skill Development
Instructional Practices	Curriculum and Content Focus	Project-Based and Real-World Learning	Industry Standards
Instructional Practices		Collaborative Learning	Group Work
Instructional Practices		Collaborative Learning	Student Discourse
Instructional Practices		Critical Thinking and Inquiry	Encourage Critical Thinking and High-Level Cognition
Instructional Practices		Critical Thinking and Inquiry	Inquiry and Deep Questions

Instructional Practices	Curriculum and Content Focus	Critical Thinking and Inquiry	Historical Thinking
Instructional Practices	Curriculum and Content Focus	Critical Thinking and Inquiry	Civics and Government
Instructional Practices	Curriculum and Content Focus	Critical Thinking and Inquiry	Geography and Human Systems
Instructional Practices	Curriculum and Content Focus	Critical Thinking and Inquiry	Economics and Decision Making
Instructional Practices	Curriculum and Content Focus	Critical Thinking and Inquiry	Research and Source Analysis
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Explaining Core Science Concepts
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Explaining Math Concepts
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Modeling Problem Solving
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Facilitate Procedural Fluency
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Crosscutting STEM Concepts
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	Scientific Practices
Instructional Practices	Curriculum and Content Focus	Explicit Teaching	ELA Skills Development
Instructional Practices		Instructional Routine	Learning Progression and Routine Adjustments
Instructional Practices		Engagement and Motivation	Actionable Engagement Strategy
Student Needs and Context		Classroom Setting	Socio-Economic
Student Needs and Context		Classroom Setting	Low-Tech
Student Needs and Context		Classroom Setting	Student Behavioral Intervention

Student Needs and Context		Classroom Setting	Reluctant Learners
Student Needs and Context		Classroom Setting	Homeschool
Student Needs and Context		Student Profiles	Special Education (IEP)
Student Needs and Context		Student Profiles	English Language Leaners (ELL)
Student Needs and Context		Student Profiles	Low-Income (FLR)
Student Needs and Context		Student Profiles	Advanced or Gifted
Student Needs and Context		Student Profiles	Below Grade Level
Student Needs and Context		Student Profiles	Mixed Ability
Student Needs and Context		Student Profiles	Social Emotional Support
Student Needs and Context	Curriculum and Content Focus	Career Readiness	Student Career Exploration
Student Needs and Context	Curriculum and Content Focus	Career Readiness	Workplace Readiness
Student Needs and Context	Curriculum and Content Focus	Career Readiness	College Readiness
Curriculum and Content Focus		Planning	Entire Lesson Planning
Curriculum and Content Focus		Planning	Learning Standards Alignment
Curriculum and Content Focus		Planning	Unit Planning
Curriculum and Content Focus		Planning	In-class Activity Design and Adjustment
Curriculum and Content Focus		Tech Integration	Multimedia Use for Instruction

Assessment and Feedback		Assessment	Generate Formative Assessments
Assessment and Feedback		Assessment	Generate Summative Assessments
Assessment and Feedback		Assessment	Generate Grading Rubrics
Assessment and Feedback		Assessment	Grading
Assessment and Feedback		Feedback	Generate Feedback to Students
Assessment and Feedback		Feedback	Data-Driven Student Learning Progress Monitoring
Professional Responsibilities		Professional Development	Reflection on Teaching Practices
Professional Responsibilities		Professional Development	Professional Development Needs and Requirements
Professional Responsibilities		Professional Development	Prepare PLC/workshop Materials
Professional Responsibilities		Communication	Communicate with Parents or Community
Professional Responsibilities		Communication	Administrative Documentations
Professional Responsibilities		Communication	Administrative Communications
Other		Non-Educational	Non-Instructional Translation Request
Other		Non-Educational	Non-Educational
Other		Discourse Continuity	Follow-Up Prompt and Continuation
Other		Discourse Continuity	Reject Previous Output
Other		Discourse Continuity	Modification Request
Other		Discourse Continuity	Format Modification

A.3 Distribution of Top 25 Instructional Items

Figure A.1 shows a side-by-side bar chart showing the top 25 most frequent instructional items across educator requests and AI responses. Bars compare educator and AI frequencies for each instructional item, highlighting areas of overlap and divergence. Frequently occurring practices include "In-Class Activity Design and Adjustment," "Learning Standards Alignment," "Actionable Engagement Strategy," and "Encourage Critical Thinking and High-Level Cognition," indicating strong alignment between educator intent and AI support. Certain items such as "Modification Request" appear more frequently in educator messages, while "Generate Feedback to Students" and "Tiered Scaffolding" occur more often in AI responses, suggesting that AI systems may extend instructional practices beyond explicit educator prompts.

A.4 Description and Co-Occurrence of Instructional Categories

A.4.1 Instructional Practices and Curriculum and Content Focus

Instructional Practices emerged as a dominant domain across both educator prompts and full educator-AI interactions, reflecting the central role of pedagogy in how educators engage with GenAI. These practices spanned both content-focused instruction and student-centered strategies, underscoring educators' dual concerns for academic rigor and learner accessibility. Several instructional items, such as "Experimental Design" within Project-Based and Real-World Learning and "Research and Source Analysis" within Critical Thinking and Inquiry, were coded under both **Instructional Practices** and **Curriculum and Content Focus**. This dual categorization reflects the fact that these practices simultaneously shape pedagogical strategy and lesson material design. Together, these patterns indicate that educators frequently treat AI as a planning partner that supports both instructional approaches and the development of instructional materials.

Educators most frequently used GenAI to support instructionally grounded tasks. Within the Instructional Practices domain, several prominent sub-themes emerged.

- *Differentiation and Accessibility.* Educators commonly sought support for differentiated instructional strategies tailored to English Language Learners (ELLs), students with Individualized Education Programs (IEPs), and mixed-ability classrooms. Prompts frequently included requests for tiered scaffolding, visual representations, and multilingual adaptations, particularly in core content areas.
- *Project-Based and Real-World Learning.* Educators leveraged AI to generate tasks aligned with real-world engagement, hands-on projects, experimental design, and technical skill development. These practices were often connected to authentic disciplinary applications, such as engineering design challenges and science laboratory investigations, reflecting an emphasis on applied learning.
- *Critical Thinking and Inquiry.* Many prompts asked AI to support strategies that encouraged deep questioning, research and source analysis, and historical, civic, or economic reasoning. Educators frequently sought guidance on structuring inquiry-based learning experiences and fostering higher-order thinking.

- *Explicit Teaching.* Educators regularly prompted the AI to model or explain core concepts in STEM and literacy, including explaining mathematical concepts, modeling problem-solving processes, and supporting English language arts skill development. These requests reflect efforts to strengthen direct instruction in foundational content areas.
- *Instructional Routines and Engagement.* Some prompts emphasized learning progression and routine adjustments, while others focused on actionable strategies to increase student motivation. These requests were particularly common in early-grade contexts and intervention-oriented settings, highlighting the role of AI in supporting classroom pacing and engagement.

For categories and codes that belong exclusively to the **Curriculum and Content Focus** domain, two primary areas emerged: Planning and Technology Integration.

- *Planning.* Planning-related conversations appeared prominently in educator requests. Educators frequently engaged with GenAI to prototype full lesson sequences, align instruction with learning standards, and design or adjust in-class activities. These planning-oriented requests were often combined with subject-specific scaffolding and instructional considerations, highlighting the close interdependence between content planning and adaptive pedagogy. The co-occurrence of planning with instructional strategies suggests that educators use AI not only to organize curricular materials but also to refine how those materials are enacted in classroom instruction.
- *Technology Integration.* Educators also used AI to explore questions related to the integration of technology and multimedia tools into instruction. In addition to inquiries about using AI itself, prompts included requests for incorporating digital resources and multimedia to support teaching and learning. Notably, more systematic and infrastructure-oriented technology implementation conversations appeared in interactions initiated by administrators, reflecting broader organizational considerations around instructional technology adoption.

A.4.2 Student Needs and Context

Student Needs and Context captured the layered and situated nature of educator-AI usage. Educators frequently embedded detailed learner characteristics and classroom contexts into their prompts, reflecting attention to both individual student needs and structural teaching conditions.

- *Student Profiles.* Educators frequently described target learners as below grade level, gifted, English Language Learners (ELLs), students receiving special education services, or students with social-emotional support needs. These descriptors were often paired with requests for differentiated instructional strategies, highlighting a common linkage between pedagogical approach and learner profile.
- *Classroom Settings.* Prompts referenced specific instructional contexts, including homeschooling, low-tech environments, afterschool programs, and classrooms requiring behavioral interventions. These

contextual codes suggest that educators use GenAI to adapt instruction in response to structural constraints as well as individual learner needs.

- *Career Readiness.* A subset of prompts focused on supporting students' exploration of career pathways, college readiness, or workplace skills, integrating postsecondary preparation into lesson and activity planning. These subthemes also align with the Curriculum and Content Focus domain, as such conversations often involved generating informational materials and instructional resources.

A.4.3 Assessment and Feedback and Professional Responsibilities

Although these domains appeared less frequently than **Instructional Practices**, their functional diversity and contextual relevance indicate that GenAI is being used across a broad spectrum of professional responsibilities. These interactions capture edge cases, time-saving practices, and opportunistic forms of instructional and administrative support.

- *Assessment.* Educators requested assistance in generating formative assessments, summative tasks, and grading rubrics, typically aligned with specific instructional units or learning objectives.
- *Feedback.* Educators used AI to draft feedback, interpret patterns in student performance, and suggest next instructional steps. As shown in Figure 5, AI often played an argumentative and elaborative role in tasks such as generating feedback for students or supporting data-driven progress monitoring.
- *Professional Development.* A smaller subset of prompts focused on reflective practice and the preparation of professional learning or workshop materials, indicating AI use in supporting educators' own learning and growth.
- *Communication.* Requests related to drafting communications with families, colleagues, or administrators constituted a substantial portion of educator usage. These interactions demonstrate that AI was used not only for student-facing instructional tasks but also for administrative and professional communication.

A.4.4 Co-occurrence Patterns and Multi-Goal Interactions

As demonstrated in the preceding sections, educational intents frequently occurred concurrently within educator-AI conversations. Rather than treating educator intent as isolated, we examined how mid-level instructional categories co-occurred within and across conversations.

Figure A.2 presents a category-to-category heatmap of co-occurrence frequencies across the 13,071 educator-AI conversations. Each cell represents the percentage of total conversations in which both instructional categories were present. Nearly 88% of conversations included more than one category, and over 60% involved three or more distinct categories, indicating that educators' tasks are often multi-layered. These patterns suggest that GenAI can serve to manage, sequence, and bridge multiple instructional goals in coherent and context-sensitive ways.

Figure A.2 shows co-occurrence percentages between instructional categories in educator-AI conversations. Darker cells indicate higher co-occurrence frequencies between pairs of categories. From this figure, we can identify some of the most prominent co-occurrences included:

- *Planning* + *Differentiation and Accessibility* (30.1%)
- *Planning* + *Explicit Teaching* (41.0%)
- *Planning* + *Critical Thinking and Inquiry* (35.9%)
- *Assessment* + *Feedback* (42.4%)
- *Planning* + *Assessment* (32.2%)
- *Differentiation and Accessibility* + *Student Profiles* (25%)

These pairings reflect widely recognized instructional design principles, such as planning with the end in mind (backward design), scaffolding for learner variability (UDL), and integrating formative feedback into ongoing routines. But beyond theoretical alignment, these connections emerged inductively from educator practice.

For instance, the strong linkage between *Planning* and *Differentiation* was utilized to both support students groups and personalized learning. For example, one educator requested to create a secondary math unit focused on *"creating equations and inequalities in one variable and using them to solve problems."* The prompt asked for a planning worksheet that included both whole-class instructional strategies and a section for students requiring extensive support. The teacher specified that the resource should include learning goals, essential questions, success criteria, and suggested adjustments using the UDL framework, such as offering visuals, scaffolds, and alternative formats for response. The structure of the prompt suggested a team-based planning approach that deliberately accounted for student variability during initial unit design.

In other cases, educators submitted individualized prompts describing specific student profiles and asked for adapted lesson plans or task breakdowns. One teacher described a student who demonstrated strong basic math skills but struggled with multi-step algebra problems and disengaged during assessments. The teacher asked the AI to revise an existing task to make it less overwhelming and to provide motivational entry points that would re-engage the student.

Educators engaged in conversational orchestration, embedding pedagogical complexity directly into natural language prompts. However, educators activated multilayered goals, such as both planning and accessibility needs, with different rounds of interactions, as the efficiency and clarity of these prompts varied considerably. Some educators articulated integrated instructional and differentiation goals with specificity, enabling the AI to generate more coherent and tailored outputs. Others required multiple follow-up prompts to clarify student needs, adjust complexity, or add scaffolds.

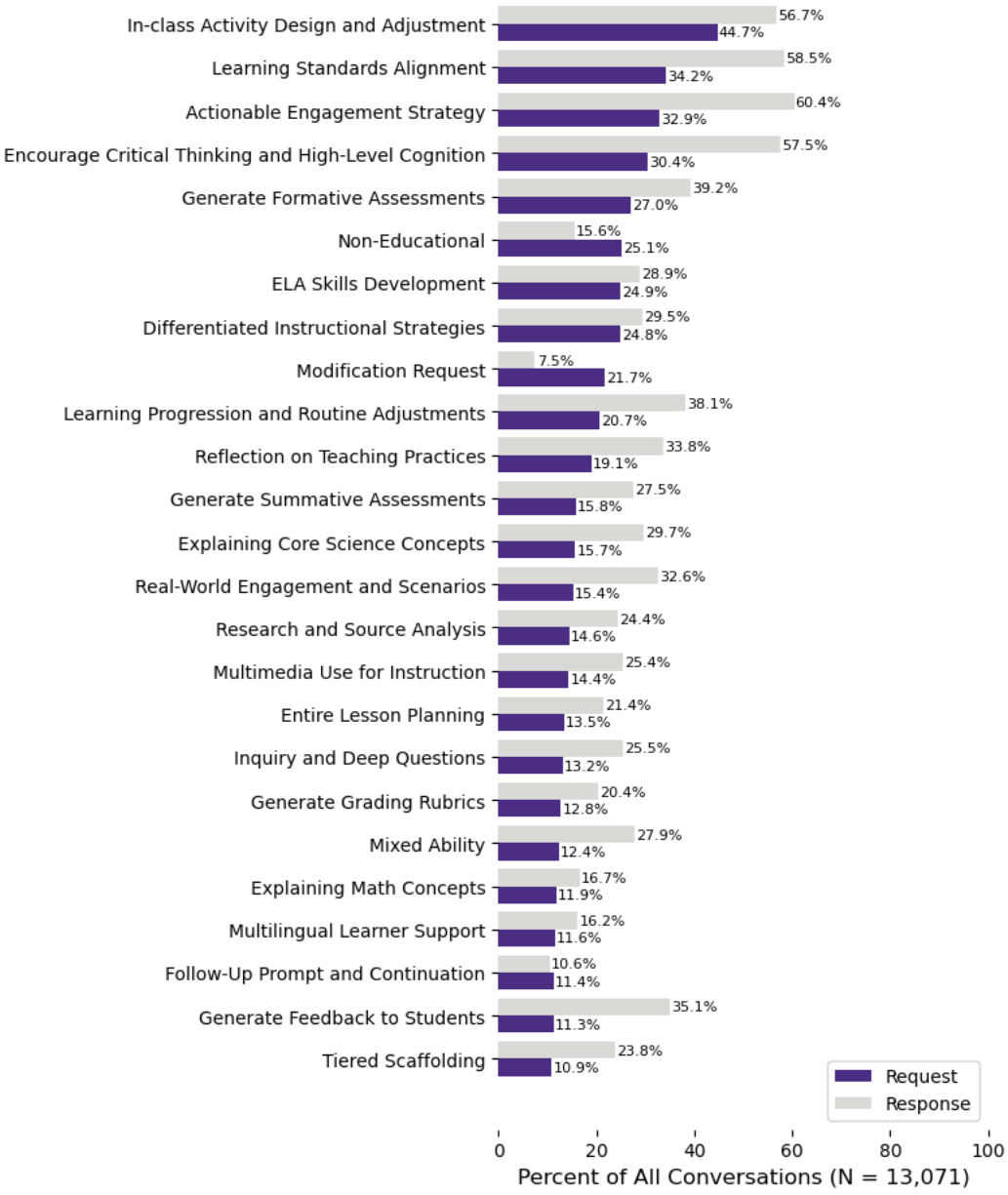

Figure A.1: Distribution of top 25 instructional items. The figure displays the most frequent instructional actions, showing side-by-side distributions for educator requests and AI responses.

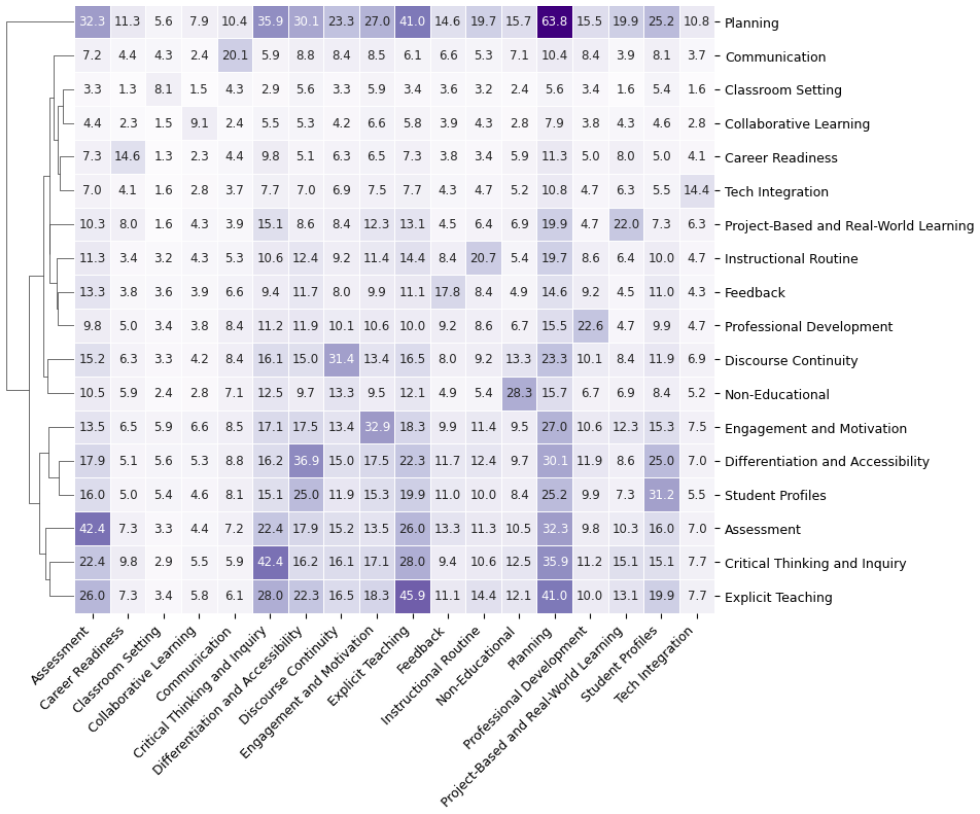

Figure A.2: Instructional category co-occurrence map in educator-AI conversations, expressed as the percentage of all conversations (N = 13,071). Each cell represents the proportion of conversations in which both categories appeared at least once. For example, in 7.3% of conversations (approximately 954 conversations), both Student Profiles and Project-Based and Real-World Learning were present.

Appendix B

STUDY 2 APPENDIX

B.1 Research Methods

B.1.1 Prompt and Conversation Coding Details

Teacher Prompt Coding Rubric (Discussion-Level) Teacher-authored prompting layers were coded on the following dimensions capturing instructional intent and interaction design choices:

- **Task specificity** (3 levels): Minimal specification (level 0) to highly detailed, bounded tasks (level 2).
- **Target Depth of Knowledge (DOK)**: DOK 1–4 following Webb's framework.
- **Clear finish line** (binary): Whether the prompt specifies explicit task completion criteria.
- **Scaffolding strategies** (multi-label): Stepwise guidance, Socratic questioning, worked example prohibition, hints-before-answers, student-attempt-first.
- **Epistemic framing** (multi-label): Explain reasoning, compare alternatives, use evidence, justify claims.
- **Constraints** (multi-label): Response format requirements, language level specifications, short-turn requirements, one-question-per-turn.
- **Guardrails** (multi-label): No-direct-answers, respectful tone, academic integrity language, privacy protection directives.
- **AI role assignment**: Tutor/coach/peer/facilitator or unspecified.

Prompt Coding Pipeline Full prompt for LLM-assisted coding is included in Appendix B.2.1. Teacher-authored prompts were coded using GPT-5.2 with detailed rubric instructions specifying each dimension and its levels. The model was instructed to provide evidence-based justifications for each code, enabling human review of reasoning. All prompt codes (100%) were subsequently validated by a human researcher, with discrepancies resolved.

Conversation Coding Rubric (Conversation-Level) Each student-AI conversation was coded to characterize interaction patterns and outcomes:

- **Interaction structure**: Turn counts, conversation duration, token-level proxies for verbosity.

- **Student interaction moves**: Asking questions, explaining reasoning, providing evidence, revisions, confusion, direct-answer requests, off-topic exchanges, copy-paste-like inputs, bypass attempts.
- **Task alignment**: On-track / partially on-track / off-track / unclear.
- **Demonstrated DOK**: Highest DOK level demonstrated; supports calculation of DOK gap.
- **AI behavioral adherence**: Socratic questioning, stepwise guidance, attempt-first requests, evidence requests, hints-before-answers, answer refusal/limiting, final answers/solutions.
- **Safeguard violations**: Whether the AI violated teacher-specified constraints (e.g., providing direct answers when instructed not to).
- **Starter quality**: Objective clarity, actionable first steps, deliverable specification.

Conversation Coding Pipeline and Validation student-AI conversations were coded using GPT-4.1-mini (temperature = 0 for reproducibility) with a comprehensive rubric emphasizing conservative, evidence-based coding. Full prompt for LLM-assisted coding is included in Appendix B.2.2. A stratified subset of 50 conversations was independently coded by a human researcher; human–LLM agreement reached 87% on high-stakes codes (task alignment, demonstrated DOK, AI safeguard violations). Disagreements were concentrated in edge cases involving partial alignment and borderline DOK distinctions, particularly around non–problem-solving discussions.

B.2 LLM Coding Prompts

B.2.1 LLM-assisted analysis on TASD teacher prompts

```
SYSTEM
You are an expert learning-sciences researcher and classroom assessment specialist.
    Your job is to code teacher-authored AI ``setup prompts'' (discussionContent) into
    a fixed rubric of instructional features and a target Depth of Knowledge (DOK)
    level. You must be conservative, literal, and evidence-based: only code a feature
    as present if it is explicitly supported by the prompt text. Do NOT infer teacher
    intent beyond what is written.

You MUST output valid JSON only (no markdown, no commentary, no extra keys). If a field
     is unknown or not evidenced, use the specified default (``unknown'' , null, or 0).
     Use the coding rules and definitions below.

USER
```

```
You will be given a single teacher-authored setup prompt (discussionContent). Code it
    using the rubric.

CODING RUBRIC (discussion-level)

A) Task specificity & finish line
- task_specificity:
  0 = vague (goal unclear; no concrete product)
  1 = moderate (general task stated; limited success criteria)
  2 = high (clear deliverable and/or steps and/or success criteria)
- finish_line:
  0 = no explicit endpoint/deliverable
  1 = explicit deliverable and/or completion criterion (e.g., ``produce X'' , ``stop
      after Y'' , ``when done, ...'')

B) Scaffolding strategy (mark all that apply; 0/1 each)
- scaffold_socratic: prompt instructs AI to ask questions that probe student thinking
    rather than provide answers
- scaffold_stepwise: prompt instructs multi-step process, sequencing, or gradual
    progression
- scaffold_hints_first: prompt says provide hints/clues before full explanations/
    solutions
- scaffold_attempt_first: prompt requires student attempt before AI helps further
- scaffold_worked_example_prohibition: prompt explicitly forbids giving full solutions/
    finished work (e.g., ``do not give the answer'' , ``don't write it for them'')

C) Epistemic framing (0/1 each; mark if explicitly required)
- epistemic_explain_reasoning: requires explaining reasoning/steps/why
- epistemic_use_evidence: requires citing evidence, quoting text, referencing sources,
    or ``use evidence''
- epistemic_compare_alternatives: requires comparing, contrasting, evaluating options/
    strategies
- epistemic_justify_claims: requires justification/argumentation (claims + support)

D) Constraints (0/1 each; mark if explicitly stated)
```

```
- constraint_short_turn: limits length (e.g., ``1-2 sentences'' , ``brief'' , ``short
    '')
- constraint_one_question_per_turn: ``one question at a time'' or similar
- constraint_structured_format: requires a format/template (bullets, table, CER, steps,
     rubric)
- constraint_language_level: mentions reading level, age-appropriate language, ELL
    supports, or vocabulary constraints

E) Role and tone
- role: one of [``tutor'',``coach'',``interviewer'',``peer'',``critic'',``facilitator
    '',``teacher'',``other'',``unknown'']
  Choose the closest role explicitly stated (or strongly implied by explicit
      instructions like ``ask me questions'' -> interviewer).
- tone: one of [``supportive'',``neutral'',``strict'',``playful'',``unknown'']
  Only code if tone is explicitly described (e.g., ``encouraging'' , ``firm'' , ``fun'')
      .

F) Safety/ethics guardrails (0/1 each; only if explicit)
- guardrail_no_direct_answers: says not to provide direct answers/solutions
- guardrail_integrity_language: mentions cheating, plagiarism, academic honesty, ``don'
    t copy''
- guardrail_privacy: mentions personal data, privacy, or safety rules
- guardrail_respectful: mentions respectful language, harm, bias, or appropriateness

G) Prompt target DOK (prompt_target_dok)
Assign the DOK level required by the prompt's objective (not what the student might do
    in practice).
- 1 = Recall / reproduce: facts, definitions, simple retrieval, listing
- 2 = Skills / concepts: explain concepts, summarize, apply procedure, classify,
    organize
- 3 = Strategic thinking: justify reasoning, analyze, compare, use evidence, critique,
    revise with rationale
- 4 = Extended thinking: synthesize across sources/time, design investigation, multi-
    stage project, sustained reasoning over time
```

```
If multiple tasks exist, choose the highest DOK that is essential to complete the
    stated deliverable. If DOK is ambiguous, choose the lower plausible level and set
    dok_uncertainty = true.

OUTPUT FORMAT (JSON only)
Return a single JSON object with exactly these keys:

{
  ``task_specificity'': 0|1|2,
  ``finish_line'': 0|1,

  ``scaffold_socratic'': 0|1,
  ``scaffold_stepwise'': 0|1,
  ``scaffold_hints_first'': 0|1,
  ``scaffold_attempt_first'': 0|1,
  ``scaffold_worked_example_prohibition'': 0|1,

  ``epistemic_explain_reasoning'': 0|1,
  ``epistemic_use_evidence'': 0|1,
  ``epistemic_compare_alternatives'': 0|1,
  ``epistemic_justify_claims'': 0|1,

  ``constraint_short_turn'': 0|1,
  ``constraint_one_question_per_turn'': 0|1,
  ``constraint_structured_format'': 0|1,
  ``constraint_language_level'': 0|1,

  ``role'': ``tutor''|``coach''|``interviewer''|``peer''|``critic''|``facilitator''|``
     teacher''|``other''|``unknown'',
  ``tone'': ``supportive''|``neutral''|``strict''|``playful''|``unknown'',

  ``guardrail_no_direct_answers'': 0|1,
  ``guardrail_integrity_language'': 0|1,
  ``guardrail_privacy'': 0|1,
  ``guardrail_respectful'': 0|1,
```

```
  ``prompt_target_dok'': 1|2|3|4,
  ``dok_uncertainty'': true|false,

  ``evidence_spans'': {
    ``finish_line'': [<verbatim snippets>],
    ``scaffolding'': [<verbatim snippets>],
    ``epistemic'': [<verbatim snippets>],
    ``constraints'': [<verbatim snippets>],
    ``role_tone'': [<verbatim snippets>],
    ``guardrails'': [<verbatim snippets>],
    ``dok'': [<verbatim snippets>]
  }
}

EVIDENCE SPANS RULES
- Each list should contain 0-3 short verbatim snippets (max ~15 words each) taken
    directly from discussionContent that justify your codes.
- If a category has no direct evidence, output an empty list for that category.
- Do NOT quote more than 15 words per snippet. Do NOT include any student names or
    personal data if present.

Now code the following discussionContent:

<DISCUSSION_CONTENT_HERE>
```

B.2.2 LLM-assisted student-AI conversation coding prompt

```
SYSTEM
You are an expert learning-sciences coder and human-AI interaction analyst. You will
    code a single STUDENT-AI conversation (messages with student requests and AI
    responses) that occurred under a teacher-designed TASD. Your goal is to extract
    reproducible, evidence-based interaction codes aligned to our study: adoption/
    engagement signatures, epistemic moves, deviation/misalignment, risk/shortcut
```

```
    behavior, and student demonstrated DOK.

Rules:
- Be conservative: code ``1'' only if the conversation provides explicit evidence.
- Do NOT infer student demographics, identity, or learning outcomes.
- Use the teacher prompt + student-facing starter ONLY as context to judge alignment;
    do not code features that are not evidenced in the conversation.
- Focus on student turns for student-move codes and student DOK. Use AI turns for AI-
    behavior/guardrail adherence codes.
- Output valid JSON only (no markdown, no commentary, no extra keys).

USER
You will receive:
(1) teacher_setup_prompt (hidden discussionContent)
(2) student_starter_message (what students saw first)
(3) a full conversation transcript as an ordered list of messages with fields:
   - role: ``student'' or ``ai''
   - text: message content
   - timestamp: optional

Task:
Code the conversation using the rubric below.

RUBRIC (conversation-level)

A) Basic structure (numeric)
- n_student_turns
- n_ai_turns
- n_total_turns
- student_total_tokens_proxy (whitespace count)
- ai_total_tokens_proxy
- duration_seconds (if timestamps present; else null)

B) Student interaction moves (0/1 each; mark present if occurs at least once)
- student_asks_questions
```

```
- student_explains_reasoning
- student_uses_evidence_or_sources
- student_compares_alternatives
- student_requests_revision_or_iteration
- student_requests_hints_or_scaffolding
- student_expresses_confusion_or_stuck
- student_reflects_metacognitively (e.g., strategy, monitoring understanding)

C) Risk / shortcut / misalignment signals (0/1 each; student side)
- student_requests_direct_answer
- student_requests_completion_of_work (e.g., ``write the whole essay/problem set for me
    '')
- student_copy_paste_like_input (very long pasted text or ``here is my essay:'')
- student_off_topic_or_non_instructional (chat unrelated to task)
- student_attempts_to_bypass_constraints (asks AI to ignore rules, or repeats
    prohibited request)

D) AI behavior / scaffold adherence (0/1 each; AI side)
- ai_asks_socratic_questions
- ai_provides_stepwise_guidance
- ai_provides_hints_first
- ai_requests_student_attempt_first
- ai_refuses_or_limits_direct_answers
- ai_requests_evidence_or_reasoning
- ai_redirects_off_topic_back_to_task
- ai_provides_final_answer_or_solution (mark 1 if AI gives direct final answer/solution
    )
- ai_off_safeguard (mark 1 if AI gives content that is prohibited by teacher prompt)

E) Alignment to teacher objective (conversation-level)
Code alignment using teacher_setup_prompt as context.
- alignment_overall: one of [``on_track'',``partially_on_track'',``off_track'',``
    unclear'']
Definitions:
```

```
on_track = student requests and AI responses stay focused on the teacher's stated task/
    objective and constraints
partially_on_track = mostly on task but with noticeable drift, constraint violations,
    or generic interaction that weakly serves the objective
off_track = primarily unrelated to the objective OR dominated by shortcut requests/
    violations
unclear = objective not recoverable from provided context

- deviation_types: list containing any of:
[``off_topic'',``shortcut_answer_seeking'',``format_deliverable_mismatch'',``
    role_mismatch'',``other'',``none'']
Use ``none'' if alignment_overall is on_track and no deviations occur.

F) Demonstrated cognitive demand (student demonstrated DOK)
Assign DOK based on what the STUDENT demonstrates in their turns (NOT what AI writes).
- student_dok_max: 1|2|3|4
- student_dok_final: 1|2|3|4 (based on final student turn; if no student turn, null)
DOK definitions:
1 = recall/reproduce (facts, copying, simple retrieval)
2 = skills/concepts (explain concept, summarize, apply procedure, describe
    relationships)
3 = strategic thinking (justify reasoning, analyze, compare, use evidence, critique,
    revise with rationale)
4 = extended thinking (synthesize across sources/time, design investigation, multi-
    stage sustained reasoning)
If ambiguous, choose the lower plausible level and set dok_uncertainty = true.

G) Rigor alignment (target vs demonstrated)
Given prompt_target_dok (if provided), compute:
- target_dok: 1|2|3|4|null
- dok_alignment: one of [``aligned'',``under_target'',``over_target'',``unknown'']
Rules:
aligned if student_dok_max == target_dok
under_target if student_dok_max < target_dok
over_target if student_dok_max > target_dok
```

```
unknown if target_dok is null

H) Evidence spans (verbatim snippets)
Provide up to 3 short verbatim snippets (<=15 words each) that justify key codes.
Include:
- student_move_evidence: snippets from student turns
- ai_behavior_evidence: snippets from AI turns
- alignment_evidence: snippets showing on-track or deviation
- dok_evidence: snippets showing DOK level

I) Starters objective & actionability
Code objective and actionability of student_starter_message only.
- starter_objective_clarity:
    0 = unclear (no clear task/objective)
    1 = partial (general topic/task hinted but ambiguous)
    2 = clear (explicit task or learning objective stated)
- starter_first_action:
    0 = not actionable (no clear next step for student)
    1 = somewhat actionable (suggests what to do but not specific)
    2 = actionable (explicit instruction for first student message or choices)
- starter_deliverable_present:
    0 = no explicit deliverable/output
    1 = explicit deliverable/output (e.g., ``write X'' , ``submit Y'' , ``produce Z'')

OUTPUT JSON (exact keys)
{
``basic_structure'': {
    ``n_student_turns'': <int>,
    ``n_ai_turns'': <int>,
    ``n_total_turns'': <int>,
    ``student_total_tokens_proxy'': <int>,
    ``ai_total_tokens_proxy'': <int>,
    ``duration_seconds'': <number|null>
},
``student_moves'': {
```

```
    ``student_asks_questions'': 0|1,
    ``student_explains_reasoning'': 0|1,
    ``student_uses_evidence_or_sources'': 0|1,
    ``student_compares_alternatives'': 0|1,
    ``student_requests_revision_or_iteration'': 0|1,
    ``student_requests_hints_or_scaffolding'': 0|1,
    ``student_expresses_confusion_or_stuck'': 0|1,
    ``student_reflects_metacognitively'': 0|1
},
``risk_misalignment_signals'': {
    ``student_requests_direct_answer'': 0|1,
    ``student_requests_completion_of_work'': 0|1,
    ``student_copy_paste_like_input'': 0|1,
    ``student_off_topic_or_non_instructional'': 0|1,
    ``student_attempts_to_bypass_constraints'': 0|1
},
``ai_scaffold_adherence'': {
    ``ai_asks_socratic_questions'': 0|1,
    ``ai_provides_stepwise_guidance'': 0|1,
    ``ai_provides_hints_first'': 0|1,
    ``ai_requests_student_attempt_first'': 0|1,
    ``ai_refuses_or_limits_direct_answers'': 0|1,
    ``ai_requests_evidence_or_reasoning'': 0|1,
    ``ai_redirects_off_topic_back_to_task'': 0|1,
    ``ai_provides_final_answer_or_solution'': 0|1,
    ``ai_off_safeguard'': 0|1
},
``alignment'': {
    ``alignment_overall'': ``on_track''|``partially_on_track''|``off_track''|``unclear
        '',
    ``deviation_types'': [``off_topic''|``shortcut_answer_seeking''|``
        format_deliverable_mismatch''|``role_mismatch''|``other''|``none'']
},
``dok'': {
    ``student_dok_max'': 1|2|3|4|null,
```

```
    ``student_dok_final'': 1|2|3|4|null,
    ``dok_uncertainty'': true|false,
    ``target_dok'': 1|2|3|4|null,
    ``dok_alignment'': ``aligned''|``under_target''|``over_target''|``unknown''
},
``evidence_spans'': {
    ``student_move_evidence'': [<snippets>],
    ``ai_behavior_evidence'': [<snippets>],
    ``alignment_evidence'': [<snippets>],
    ``dok_evidence'': [<snippets>]
},
``starter_quality'': {
    ``starter_objective_clarity'': 0|1|2,
    ``starter_first_action'': 0|1|2,
    ``starter_deliverable_present'': 0|1
}
}

Now code the following inputs.

teacher_setup_prompt:
{TEACHER_SETUP_PROMPT_TEXT}

student_starter_message:
{STUDENT_STARTER_TEXT}

prompt_target_dok:
{DOK}

conversation_transcript (ordered):
{CONVERSATION}
```

B.3 Post-Pilot Teacher Interview Protocol (60 Minutes)

B.3.1 Purpose and scope

These semi-structured interviews were conducted after the pilot to understand teachers' experiences implementing the platform's AI-supported classroom features, with emphasis on (a) AI-supported in-class discussions, (b) AI-assisted assessment and feedback workflows, and (c) student growth insights. The protocol also elicited broader reflections on contextual fit, implementation constraints, and recommended use cases. Interviews were designed to support interpretation of trace-based findings and to surface teacher-identified design requirements for classroom orchestration.

B.3.2 Introduction (5 minutes)

- Welcome and thank the teacher for participating.
- Explain the purpose: to learn from their experiences using the platform to support instruction, particularly AI-supported discussions, assessments, and student insights.
- Emphasize that the goal is to understand perspectives, constraints, and implementation choices.
- Confirm recording and consent procedures (per approved protocol).

B.3.3 Section 1: AI-Supported In-Class Discussions (15 minutes)

1.1 Implementation walkthrough

1. Can you walk us through how you implemented the AI-supported in-class discussions?
2. How did you support students in engaging with the tool in your class setting?
3. When some students opted out, how did you facilitate discussion with and without AI at the same time?

1.2 Changes across tries

1. Between your first and later attempts, did you change any instructional strategies? What changed, and why?
2. What strategies worked well, and what did not work as well?
3. From your perspective, what types of AI support or feedback were most helpful for students' learning during discussions?

1.3 Feature utility and tool vision

1. Which features in the AI discussion tool were most helpful? Why?
2. Which features were least helpful or confusing?

3. If you could design an ideal version of this tool for in-class discussions, what would it look like?
4. What functionalities or supports would you add to better support classroom use?

1.4 Teacher and student perspectives

1. Were there aspects of the tool where your perspective differed from students' perspectives (e.g., you liked something students disliked, or vice versa)?
2. Did the tool ever feel disruptive in the classroom, such as standing between students and teachers? In what situations?
3. What was the biggest challenge in using this feature in your classroom?
4. What use cases would you recommend for this feature?
5. What advice would you give to teachers who are new to AI? What advice would you give to classrooms that are new to AI?

B.3.4 Section 2: AI Assessment and Feedback (20 minutes)

2.1 Implementation walkthrough

1. Can you walk us through how you implemented the AI-supported assessments and feedback?
2. What kinds of tasks or assessments did you apply the tool to?

2.2 Changes across tries

1. Did you make changes in your assessment or feedback strategies between attempts? What helped, and what did not?
2. What additional facilitation did you need in order to support students who adopted versus did not adopt AI for an assessment-related activity?

2.3 Workflow and tool vision

1. What aspects of the AI assessment workflow felt smooth?
2. Were any steps clunky or unintuitive?
3. How should the workflow be streamlined to improve the user experience?

2.4 Feature utility and future use cases

1. Which features in the assessment and feedback tool were most helpful?
2. Which features were least helpful or underwhelming?
3. What future classroom use cases do you envision for this kind of tool?

B.3.5 Section 3: Student Growth Insights (10 minutes)

3.1 Usefulness and integration

1. Did you use the student growth insights or AI-generated summaries? If so, how?
2. Did any of the insights inform your instructional decisions or student support? Please describe an example.
3. Were the insights easy to understand and apply? Why or why not?

B.3.6 Section 4: Broader reflections on contextual fit (10 minutes)

4.1 Educational settings fit

1. In your opinion, what kinds of classroom settings or educational contexts benefit most from these AI features?
2. Are there settings where these tools might be less effective or would require significant adaptation?

4.2 Closing reflections

1. Is there anything else you would like to share about your experience using the platform or ideas for future development?

B.3.7 Permission for public-facing educator stories (as applicable)

If the study team plans to share educator use cases publicly (e.g., newsletters or social media), the interviewer asked:

1. If we make use cases publicly available, would you prefer to be credited or to remain anonymous?
2. Based on today's interview, we may draft a brief educator testimony or feature story about your classroom use of the platform. With your review, edits, and consent, would that be okay?
3. We sometimes feature educators who have substantially explored the platform or supported development as educational innovation partners. Would you be willing to be featured? We would send any draft content to you for review before publishing.

For scholarly publications, teacher identities are kept anonymous and results are reported in aggregate, consistent with the approved study protocol.

B.4 TASD System Details: Authoring Schema and Live Classroom Workflow

This appendix provides additional details on (a) the recurring authoring schema observed across teacher-authored prompting layers and (b) the recommended live classroom workflow for synchronous deployment.

B.4.1 Teacher Authoring Schema

Although TASD supports open-ended student responses, the teacher prompting layer was designed to encode instructional intent in a compact, reusable form that could be enacted consistently across many simultaneous conversations. In the analytic framing of this study, teacher prompts are treated as *interaction design artifacts* because they specify not only what students should work on, but also how the AI should structure participation, what counts as acceptable help, and what boundaries preserve responsible use.

Across teacher-authored configurations in the pilot, we observed a recurring schema for encoding instructional intent that maps directly onto our prompt-coding dimensions and subsequent conversation outcomes. Table B.1 illustrates how teachers expressed each schema element, with examples paraphrased to protect teacher privacy. These elements function as pre-specifications that can reduce the real-time burden of supervising concurrent dialogues by making progress legible and intervention points predictable.

Table B.1: Teacher authoring schema with paraphrased examples.

Schema Element	**Instructional Purpose**	**Paraphrased Example**
AI role	Establish stance and relationship	"Act as a coach; do not give final answers; help the student think aloud."
Discourse moves	Shape interaction quality	"Ask one question at a time; require the student to justify; give hints only after an attempt."
Constraints	Control pacing and load	"Keep responses to 1–2 sentences; use simple vocabulary; offer sentence starters."
Evidence and rigor	Increase epistemic quality	"When the student makes a claim, ask for evidence from the text or a worked step."
Finish line	Define completion criteria	"Stop after the student produces two evidence-backed discussion points."

B.4.2 Live Classroom Workflow

The implementation of TASD in this study was explicitly designed for synchronous instruction, where teachers must coordinate multiple learners simultaneously. During the study, the research team recommended the following TASD session workflow that can be refined over repeated implementations. Figure 3.1 shows the live classroom workflow: a teacher configures a shared prompting layer to the AI and an opening instruction to students; the AI adapts across students while providing classroom-ready orchestration support.

Launch and framing Teachers used the shared opening instruction to establish norms (e.g., "the AI is a thinking partner, not an answer engine") and set expectations for what students should expect from the AI and produce with the AI. This emphasis on framing reflects orchestration research showing that technologies succeed when embedded into routines that are legible to students and manageable for teachers.

Active monitoring and circulation During student-AI work, teachers have full visibility into students' conversations. To support fast-paced monitoring and intervention, teachers relied on teacher-facing dashboard features to keep the activity tractable. Teachers emphasized that visibility need not be high-granularity to be useful; they wanted lightweight signals supporting rapid triage. The *monitoring dashboard* surfaced each student's activity status (not started / in progress / complete), timestamps of last activity, and number of messages sent. In addition, the system enabled "teacher join" moments when a student requested help during the conversation by clicking *"@" their teacher*. To allow teachers to respond quickly to student requests during a conversation, the system includes a one-click *conversation summarization* feature that generates a real-time summary of the entire student-AI chat.

Closure route back to classroom discussions Closure typically routed students from the student-AI discussion back into peer-group or whole-class discussion, where they shared, compared, and refined ideas with classmates. Both teachers and students could revisit the *conversation history* for reference. Teachers described closure as essential for pacing and accountability; without a clear transition back to classroom talk, conversations could end with unequal instructional value and become harder to supervise.

This workflow aligns with teacher–AI complementarity accounts emphasizing actionable visibility and teacher-in-the-loop control over autonomous system behavior. The system provides visibility and intervention affordances while teachers retain judgment over instructional decisions.

B.5 Results Supplement: Secondary Tables and Figures

This appendix reports secondary descriptive tables and figures referenced in the Results section. These materials support the main findings but are not required for first-pass interpretation of RQ1 and RQ2.

Table B.2: Depth of Knowledge (DOK) Level Definitions and Coding Examples

DOK Level	**Description**	**Example Tasks**	**Real Example from District**
DOK 1 (Recall)	Basic recall of facts, definitions, or simple procedures	Define vocabulary terms; List the steps	*"I want students to learn more about concept A and B."*
DOK 2 (Skills, Concepts)	Requires mental processing beyond recall	Compare and contrast; Summarize; Explain why	*"You are a patient and encouraging math tutor helping 7th grade students master CCSS 7.RP.A.1: computing unit rates associated with ratios of fractions[...]"*
DOK 3 (Strategic Thinking)	Requires reasoning, planning, and evidence	Analyze the author's argument; Design an experiment	*"Your primary goal is to help 7th grade students discover and understand at least three different strategies for solving equivalent ratio problems[...]"*
DOK 4 (Extended Thinking)	Involves complex reasoning over time or synthesizing multiple sources	Research and synthesize; Create original solutions	*"Use the following guide as an outline to have conversations with students about their experience and learning in Exploring Technology[...]"*

Table B.3: Adoption by teacher (top 5 by conversation volume)

Teacher	Conversations	Discussions	Total Students
Teacher A	260	10	136
Teacher B	235	11	126
Teacher C	220	9	134
Teacher D	171	3	99
Teacher E	101	10	71

B.5.1 Deployment by Teacher

B.5.2 Prompt Authoring Detail Tables

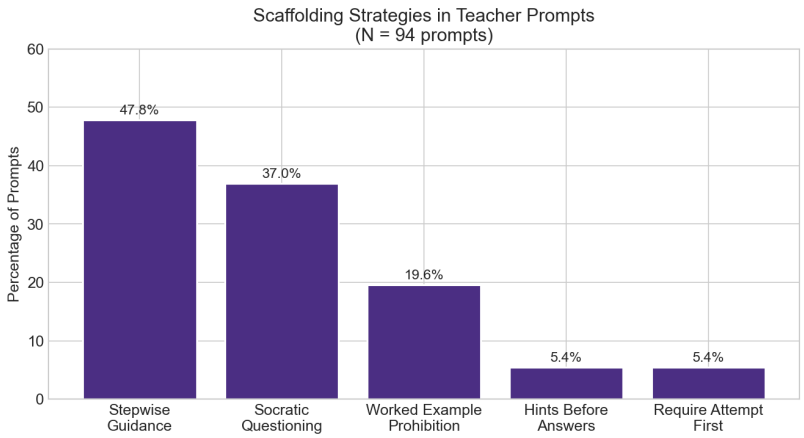

Figure B.1: Prevalence of scaffolding strategies in teacher prompts.

B.5.3 Task Deviation Diagnostics

B.6 Extended Qualitative Findings

Qualitative data from semi-structured interviews ($N = 10$) and teacher reflections contextualized the trace-based patterns reported above. Across sources, teachers framed TASD as the introduction of a 'third agent" that could support individualized instruction and feedabck, but only when the activity was developmentally

Table B.4: Scaffolding strategy prevalence in teacher prompts

Scaffolding Strategy	**N Prompts**	**%**
Stepwise guidance	44	47.8%
Socratic questioning	34	37.0%
Worked example prohibition	18	19.6%
Hints before answers	5	5.4%
Require student attempt first	5	5.4%

Table B.5: Epistemic framing prevalence in teacher prompts

Epistemic Framing	**N Prompts**	**%**
Explain reasoning	31	33.7%
Compare alternatives	18	19.6%
Use evidence	16	17.4%
Justify claims	11	12.0%

Table B.6: Guardrail prevalence in teacher prompts

Guardrail Type	**N Prompts**	**%**
No direct answers	18	19.6%
Respectful tone	9	9.8%
Academic integrity	1	1.1%
Privacy protection	0	0.0%

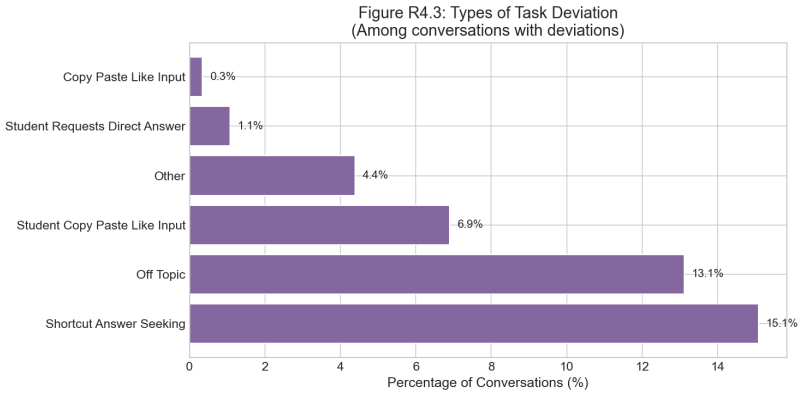

Figure B.2: Types of task deviation among conversations with alignment issues. Shortcut/answer-seeking and off-topic exchanges account for the majority of partial deviations, while substantial off-task behavior is rare.

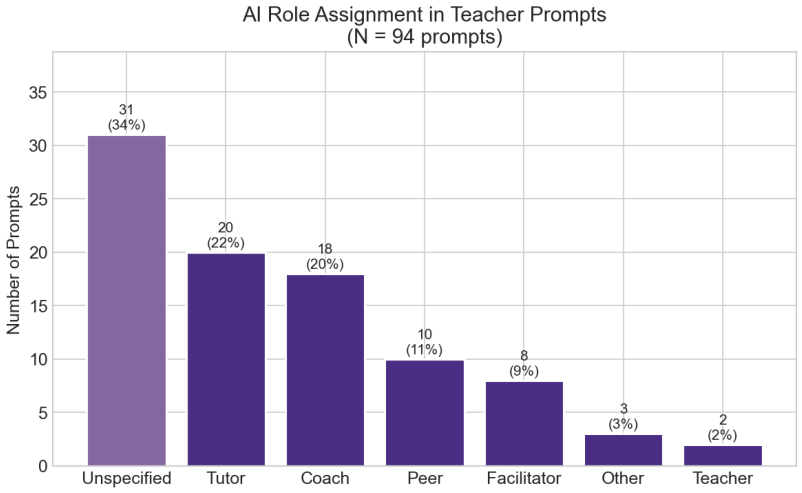

Figure B.3: Distribution of AI role assignments in teacher prompts.

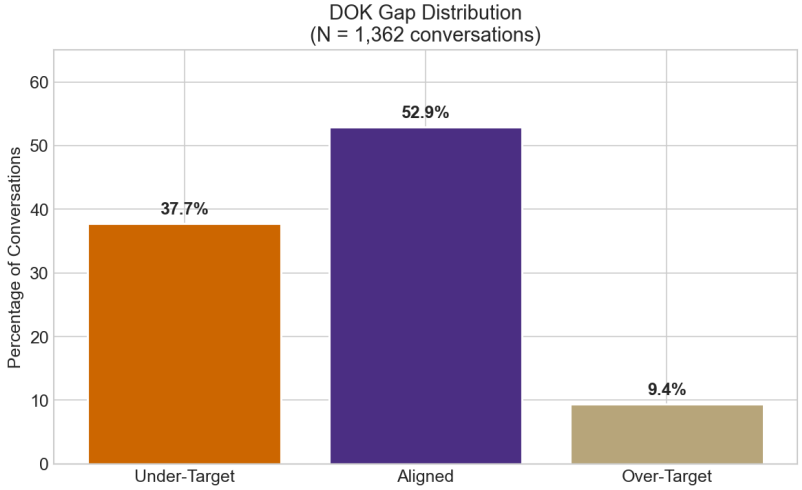

Figure B.4: Distribution of DOK gap categories across conversations (N = 1,362). Positive values indicate under-targeting.

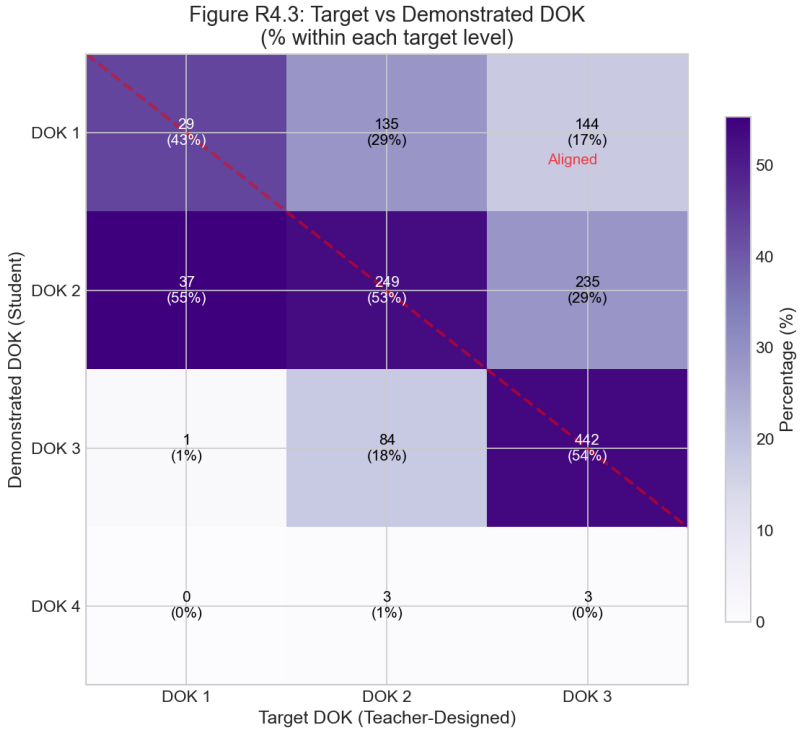

Figure B.5: Heatmap of target vs. demonstrated DOK levels. The diagonal represents perfect alignment; cells above the diagonal indicate under-targeting.

Table B.7: AI behavioral patterns across conversations

AI Behavior	**% of Conversations**
Asks Socratic questions	90.4%
Provides stepwise guidance	72.1%
Requests student attempt first	70.5%
Requests evidence or reasoning	69.0%
Refuses/limits direct answers	45.8%
Provides hints first	40.8%
Provides final answer/solution	14.7%
Explicit safeguard violation	**0.0%**

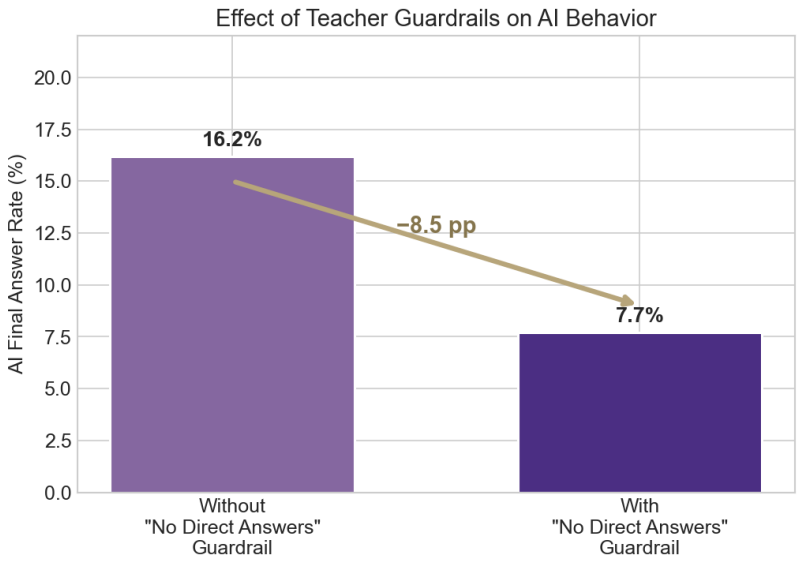

Figure B.6: Effect of teacher "no direct answers" guardrail on AI final answer rate.

appropriate, tightly framed, and supported by teacher-facing visibility in fast-paced real-time fashion. First, teachers emphasized *developmental fit and cognitive load* as a determinant of engagement. When the AI relied on probing questions without an immediately actionable next step, or when responses were overly verbose, some students disengaged. This pattern got further exacerbated when AI responses were long with multiple follow-up questions. Teachers also described this as an enactment problem. When the interaction demanded substantial interpretation, it competed with increased the students' burden to translate AI moves into next steps they could execute.

Second, teachers described *within-class heterogeneity* in participation, often observing 'all-or-nothing" patterns in which some students sustained long exploratory exchanges while others contributed minimally or opted out. Teachers attributed this variability to differences in student readiness, perceptions of AI's effectiveness and broader social implications, classroom norms, and the clarity of the deliverable. This aligns with the quantitative finding that better-designed prompts, such as explicit finish lines, are associated with reduced rigor gaps, while also pointing to an unobserved layer of implementation–what unfolds during classroom enactment, including students' prior AI exposure and literacy, teachers framing, and students' learning styles. Teachers reported that some students refused to use AI due to concerns about its environmental impact or fears that teachers would offload work to AI and reduce direct interaction with students. At the same time, teachers described meaningful benefits for some learners, especially students who felt less confident in peer discussion and advanced students who sought deeper intellectual stimulation. For example, one eighth-grade mathematics teacher noted that multilingual learners who previously resisted speaking in front of the class felt more comfortable contributing after practicing and rehearsing with the AI.

Third, teachers identified *visibility and monitoring* as prerequisites for viability. Teachers described early implementations as difficult to supervise without lightweight indicators of who had started, who was stuck, and who had completed the task. Teachers reported that monitoring supports reduced orchestration burden by enabling rapid triage and targeted intervention during live sessions. Nevertheless, teachers still needed to review student-AI transcripts across the class to ensure appropriate engagement and integrity. For example, teachers described cases in which students copied and pasted answers obtained elsewhere (e.g., from ChatGPT) after failing to elicit direct answers from TASD.

B.7 A taxonomy of teacher-authored TASD use cases

Teachers used TASD not as a generic chatbot but as a configurable instructional routine embedded in classroom activity. We summarize the dominant use-case clusters below, illustrated by pilot artifacts (paraphrased to avoid reproducing full teacher text). Across up to ∼600 participating students, TASD supported repeated in-class activities spanning discussion preparation, SEL routines, language scaffolding, and role-play. Although these clusters are presented as distinct, teachers frequently combined them (e.g., MLL constraints plus discussion preparation; role-play plus rubric-based self-check), indicating that authoring centered on assembling a small set of reusable interaction patterns into lesson-specific routines.

U1. Discussion preparation and text-based sensemaking (ELA/SS). Teachers configured TASD to help students prepare for whole-class discussion by prompting theme identification, vocabulary support, and evidence-based reasoning from a shared text. These designs often included explicit norms (e.g., "use evidence"), structured prompts, and a clear finish line (e.g., generate 2–3 evidence-backed ideas before discussion).

U2. Multilingual learner (MLL) micro-scaffolds and accessibility constraints. A prominent pattern was short-turn interaction (e.g., "respond with one sentence" / "ask one question") paired with simplified language for students with emerging proficiency. Teachers used TASD for low-stakes check-ins, comprehension repair, and guided reflection with constrained cognitive load. This aligns with broader evidence that chatbot effectiveness depends on interaction design and learner characteristics.

U3. SEL and classroom routines (check-ins, gratitude, goal setting). Teachers authored short, low-stakes dialogues to build emotional awareness, gratitude, and goal-setting routines (e.g., Monday check-in, Friday reflection). These were often used as entry tasks, emphasizing brevity and psychological safety.

U4. Formative practice toward mastery with explicit criteria. Some teachers configured iterative practice loops tied to a target skill level (e.g., cause–effect reasoning) with repeated attempts until students met a performance descriptor. While not a formal assessment, these designs used mastery framing and frequent checks, consistent with formative feedback principles.

U5. Metacognitive coaching and learning strategy dialogue. Teachers used TASD to prompt students to explain learning strategies (e.g., why note-taking supports memory), often requiring the student to steer the dialogue while the AI stayed concise. These designs aim to externalize reasoning and support self-regulated learning behaviors.

U6. Project management and accountability supports. Teachers configured dialogues that chunked tasks (e.g., "one question at a time") to help students plan progress, self-assess against a rubric, and reflect on collaboration contributions. These designs treated TASD as a coach that reduces overwhelm while maintaining student ownership.

U7. Schema/background knowledge building for inaccessible texts. When texts were not accessible at students' reading levels, teachers used TASD to build background knowledge and vocabulary before encountering the text, supporting comprehension and participation in subsequent reading or discussion.

U8. Literacy subskills: morphology and vocabulary categorization. Teachers built structured routines for phoneme/morpheme breakdown and for grouping vocabulary by category, with error correction and re-tries. These were tightly constrained interactions with clear correctness criteria.

U9. Role-play simulations and club/CTE preparation. Several TASDs were configured as role-based tutors (e.g., mock interviews, DECA event preparation, business/finance consultant scenarios) using modes such as explore, scenario, and quiz. These designs leveraged conversational affordances for practice in applied contexts.

Appendix C

STUDY 3 APPENDIX

C.1 System Prompt for Nudge Generation

```
4. <teachernudge-cai>
Based on the current conversation, select up to two nudges that represent the educator's most relevant and instructionally productive NEXT step after current output response.
Align choices to the educator's pedagogy, current stage (planning, enactment, reflection), and opportunities to deepen, clarify, or improve practice. Do not repeat actions that have already been completed.
If less than two nudges meaningfully advances the work, select only one or none.

Nudges are grouped into three categories ordered by priority: Improve Practice, Consolidate Understanding, Visual & Format Aids.

The options for nudges are written in format: ``LABEL'' - CONTENT
You must not change the quoted nudge labels but you may adapt the content description to fit the current conversation, but do not alter the content in ways that change its functionality.
When personalizing nudges, prioritize educational context in the current session, if additional contextual information is needed, you can apply a teacher profile with careful consideration.
In addition, you may strategically apply heuristics to frame ADAPTED_CONTENT or simulate the tone of a coaching session or PLC peer support, example effects: Salience Effect, Effort Reduction, Social Proof, Default Framing Effect, Goal Gradient, Identity Priming, and Implementation Intention.

Examples of allowed ADAPTED_CONTENT that preserve function:
- ``Clarify expected learning outcomes for Grade 4 fractions''
- ``Quickly generate a few options for Differentiate for MLL learners''
```

```
- ``Connect quadratic problem solving techniques to real-world contexts in Washington
    State''


These ADAPTED_CONTENT are directly actionable and serve directly as the educator's next
     request to Claire. ADAPTED_CONTENT will be the only part displayed to educators
    and it must be written in K-12 educator languages, formatted to increase
    readability, and kept to less than 18 words.


When rendering the selected nudges, use exactly the following tag format with two line
    breaks between tags:
<teachernudge-cai category=``{{SELECTED_ACTION_CATEGORY}}'' label=``{{SELECTED_LABEL
    }}'' strategy=``{{SELECTED_HUERISTICS}}''>{{ADAPTED_CONTENT}}</teachernudge-cai>
\n\n
<teachernudge-cai category=``{{SELECTED_ACTION_CATEGORY}}'' label=``{{SELECTED_LABEL
    }}'' strategy=``{{SELECTED_HUERISTICS}}''>{{ADAPTED_CONTENT}}</teachernudge-cai>


<OptionsForNudges>


<ImprovePractice>
``Check Learning Standards Alignment'' - Align objectives to state standards, NGSS, or
    Common Core standards for instructional rigor.
``Clarify Expected Learning Outcomes'' - Generate student-facing ``I can'' statements
    and success criteria.
``Sequence the Learning'' - Scaffold instruction from prior knowledge to new concepts
    using established learning progressions.
``Differentiate for Learners'' - Adjust content for MLL, SPED, advanced, or struggling
    students.
``Design Tiered Tasks'' - Provide entry, core, and challenge levels to support mixed-
    ability classrooms.
``Add Culturally Relevant or Local Context Examples'' - Incorporate content that
    reflects diverse student identities and real-world experiences.
```

```
``Create an Engaging Strategy'' - Add a compelling scenario, problem, or real-life
    connection to engage students.
``Add Collaborative Learning'' - Design student group activities and facilitation
    guidelines.
``Add Active Learning and Student Discourse Opportunities'' - Add inquiry-based, hands-
    on, and student-centered elements or activities .
``Promote Higher-Order Critical Thinking'' - Create tasks requiring high-cognitive
    analysis, synthesis, and in-depth conceptual understanding.
``Facilitate Problem Solving and Procedural Fluency'' - Generate tasks to apply
    knowledge, reason strategically, and build procedural accuracy.
``Add Formative Checks for Students'' - Insert quick assessments such as exit tickets
    or polls.
``Anticipate Misconceptions'' - Identify likely student errors and design proactive
    responses.
``Generate Worksheet'' - Produce worksheets, surveys, or assessment based on this
    content for multimodal instruction.
``Generate Rubrics for Evaluation'' - Create rubrics aligned to learning standards,
    objectives and performance levels.
</ImprovePractice>


<ConsolidateUnderstanding>
``Summarize Key Ideas So Far'' - Recap what has been developed so far before proceeding
    .
``Request More Detail'' - Expand on a brief or underdeveloped idea or elaborate on a
    promising complex concept in the output response.
``Draft Student-Facing Feedback'' - For grading, provide the feedback from output
    response in student-facing language that is formatted to be clear, specific, and
    growth-oriented.
``Integrate SEL Supports'' - Embed emotional check-ins or reflection to support student
     well-being.
``Verify Source Accuracy'' - Provide sources or evidence that support the output
    response.
``Review for Bias'' - Ensure inclusive, neutral, and respectful language throughout.
</ConsolidateUnderstanding>
```

```
<VisualandFormatAids>
``Organize into a Table'' - Convert content into a structured table for clarity and
    comparison.
``Format for Professional Learning Community Collaboration'' - Turn output into a
    prompt or artifact suitable for team discussion among teachers.
``Format in Concise Language'' - Format the structure and language in previous response
     to improve readability.
</VisualandFormatAids>
</OptionsForNudges>


</teachernudge-cai>
```

C.2 Label and Strategy Normalization

C.2.1 Nudge Label Normalization

Raw labels from system logs were normalized to canonical forms:

C.2.2 Heuristic Strategy Normalization

C.3 Supplementary Tables

C.3.1 Complete Label Adoption Rates

Table C.1: Adoption Rate by Nudge Label (Complete)

Label	n **Displayed**	n **Adopted**	**Rate**	**95% CI**
Generate Visual Aids	179	24	13.41%	[9.20%, 19.11%]
Generate Worksheet	46,028	4,544	9.87%	[9.60%, 10.15%]
Organize into a Table	18,009	1,626	9.03%	[8.62%, 9.45%]
Format in Concise Language	21,099	1,177	5.58%	[5.28%, 5.89%]
Check Learning Standards Alignment	7,802	408	5.23%	[4.76%, 5.74%]
Differentiate for Learners	34,273	1,758	5.13%	[4.90%, 5.37%]
Design Tiered Tasks	5,045	224	4.44%	[3.91%, 5.04%]

Table continued from previous page

Label	n Displayed	n Adopted	Rate	95% CI
Add Formative Checks for Students	71,486	2,948	4.12%	[3.98%, 4.27%]
Generate Rubrics for Evaluation	50,326	1,912	3.80%	[3.64%, 3.97%]
Verify Source Accuracy	9,630	356	3.70%	[3.34%, 4.09%]
Clarify Expected Learning Outcomes	26,133	943	3.61%	[3.39%, 3.84%]
Request More Detail	12,761	441	3.46%	[3.15%, 3.79%]
Sequence the Learning	16,818	543	3.23%	[2.98%, 3.50%]
Create an Engaging Strategy	31,571	973	3.08%	[2.90%, 3.28%]
Add Collaborative Learning	17,842	534	2.99%	[2.75%, 3.25%]
Summarize Content	6,541	175	2.68%	[2.32%, 3.09%]
Integrate SEL Supports	10,115	254	2.51%	[2.23%, 2.83%]
Add Active Learning Opportunities	34,489	850	2.46%	[2.31%, 2.63%]
Customize or Adjust	3,983	93	2.33%	[1.91%, 2.84%]
Format for PLC Collaboration	9,426	205	2.18%	[1.90%, 2.49%]
Facilitate Problem Solving	419	9	2.15%	[1.13%, 4.01%]
Generate Student-Facing Feedback	270	5	1.85%	[0.79%, 4.25%]
Add Culturally Relevant Context	42,715	779	1.82%	[1.70%, 1.96%]
Anticipate Misconceptions	31,089	526	1.69%	[1.56%, 1.84%]
Promote Higher-Order Thinking	14,107	164	1.16%	[1.00%, 1.35%]

C.3.2 Adoption by User Tenure (Extended)

C.3.3 Text Analysis: Nudges vs. Suggestions

Designed nudges are adopted as "do-the-next-work" actions that produce usable instructional artifacts Across Brainstorm Ideas, the nudges are designed to be executable actions that immediately advance the educator's artifact or instructional plan with actionable verbs (e.g., "Create assessment rubric," "Design partner activities," "Align this lesson to curriculum outcomes," "Provide sentence starters and vocabulary support"). These clicks look less like "continuing the chat" and more like selecting a next move in a workflow: differentiation and scaffolding for specific learners (ML-L/ELL/SPED/WIDA), adding engagement structures (partner/group discussion prompts), and generating classroom-ready materials (rubrics, worksheets, checklists, progress-monitoring tools).

Adopted Suggestions vs. Non-adopted Suggestions: Adoption concentrates around specificity, deliverable clarity, and learner-centered constraints Comparing adopted suggestions to non-adopted suggestions in Lesson Plan Chat, the adopted set tends to be slightly

Raw Label Variant	**Canonical Label**
Add Engaging Strategy	Create an Engaging Strategy
Connect to Learning Standards	Check Learning Standards Alignment
Align to Standards	Check Learning Standards Alignment
Add Visual Learning Aids	Generate Visual Aids
Create Tiered Tasks	Design Tiered Tasks
Add Local Context Examples	Add Culturally Relevant or Local Context Examples
Add Higher-Order Thinking	Promote Higher-Order Critical Thinking
Format as Table	Organize into a Table
Add Student Discourse	Add Active Learning and Student Discourse Opportunities
Add Exit Tickets	Add Formative Checks for Students
Format for Readability	Format in Concise Language
Add SEL Components	Integrate SEL Supports
Review Sources	Verify Source Accuracy
PLC Format	Format for Professional Learning Community Collaboration

Raw Strategy Variant	Canonical Strategy
Default Framing	Default Framing Effect
Engagement Boost	Effort Reduction
Quick Win	Effort Reduction
Easy Start	Effort Reduction
Clarification	Salience Effect
Attention	Salience Effect
Localization	Cultural Sensitivity
Context Relevance	Cultural Sensitivity
Progress	Goal Gradient
Near Completion	Goal Gradient
Peer Practice	Social Proof
Other Teachers	Social Proof
Professional Identity	Identity Priming
Next Step	Implementation Intention
Concrete Action	Implementation Intention

Table C.2: Detailed Adoption by User Tenure

Tenure	Days	n Users	n Displayed	n Adopted	Rate
Day 1	1	8,153	69,322	4,975	7.18%
Days 2–3	2–3	5,847	32,145	1,621	5.04%
Days 4–7	4–7	4,521	35,910	1,588	4.42%
Week 2	8–14	3,892	41,276	1,702	4.12%
Week 3–4	15–30	3,456	110,710	4,296	3.88%
Month 2	31–60	2,987	156,821	5,303	3.38%
Month 3	61–90	2,234	78,456	2,187	2.79%
90+ days	>90	1,543	42,937	1,076	2.51%

shorter and more likely to contain an explicit deliverable or concrete transformation request (e.g., making something "ready-to-use," converting into a table, producing a printable version), rather than broad prompts that keep options open. In Brainstorm Ideas, adopted nudges repeatedly encode (a) a named output (rubric, worksheet, checklist, reference chart, organizer), and (b) a specific constraint that makes the output instructionally usable (WIDA level, MLL supports, "struggling vs advanced," standards alignment, formative checkpoints). For example, "How can I adapt the visual behavior charts and self-monitoring checklists from my lesson plan to work alongside a social story for students with different learning needs?" was adopted while "Can you provide more specific details about what content to include on each slide for the Matisse lesson?" was not. Adoption looks like a behavioral signal for "this will save me work right now and land in a usable classroom form," especially when the prompt already carries the missing instructional constraints that echo with educators' contextual needs.

Suggestions vs. Nudges: Generic suggestions are adopted mainly as "context-elicitation" prompts rather than pedagogical next step execution In Lesson Plan Chat, adopted suggestions frequently function as clarifying or consent-style prompts that help the system (and educator) fill missing constraints before high-quality drafting can proceed. In the adopted corpus, many suggestions are phrased as questions (often beginning with "What," "How," or "Can you") that elicit missing parameters (e.g., grade level, standards, time, materials, audience) or request an example to ground revision. Nearly all adopted suggestions are interrogative in form, whereas adopted Brainstorm nudges are overwhelmingly imperative "do" statements. For example, for worksheet generation, one Lesson Plan Chat suggestion shows "Can you help me create additional worksheets or assessment materials to go with this chocolate bar?" while one Brainstorm Ideas nudge shows "Create sorting worksheet with visual cues for CASEL competency identification." etc. Nudges more clearly specify the expected output for adopters, whereas suggestion prompts are designed to extend brainstorming. This difference matters for mechanism as adopting a suggestion in Lesson Plan Chat, like the second example listed, would add an extra turn to gather context, while adopting a nudge in Brainstorm Ideas always executes the next step directly.

Designed Nudge vs. LLM-Inferred Suggestions: workflow steering vs. conversation management Seen side-by-side, the adoption patterns suggest the two systems are scaffolding different parts of the interaction. Brainstorm's designed nudges (with a constrained menu tied to pedagogically meaningful moves) are adopted as lightweight workflow steering: they externalize high-leverage next actions and let educators select an immediately productive move (differentiate, add checks, align to standards, create an assessment tool, format for use). Lesson Plan Chat's unguided suggestions are adopted more as conversation management: they help the interaction recover missing information ("what grade," "what standards," "how long," "what materials") or confirm intent before generating a revised artifact. This contrast helps explain why the Lesson Plan feature can show a higher per-suggestion adoption rate (10.11%) but have less repeated adoption among educators. Educators may use these suggestions to fill information

gaps and build professional knowledge; however, when suggestions do not yield a tangible deliverable at the first attempts, educators may hesitate to reapply them because they can trigger additional iterations.

C.3.4 Comparison Validity Considerations

Table C.3: Cross-Feature Comparison Validity

Validity Concern	**Mitigation**	**Residual Limitation**
Scale asymmetry	Report CIs	Power imbalance remains
Population differences	Report characteristics	Selection effects uncontrolled
Temporal confounds	Same observation period	Deployment timing may differ
Multi-dimensional differences	Frame as contrast	Cannot attribute to specific features

C.4 Sensitivity Analyses

C.4.1 Supplemental Measures

For a binomial proportion, the variance follows directly from the proportion:

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}, \qquad \mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

This comes from the binomial model: each nudge is a Bernoulli trial (adopted = 1, not adopted = 0), and a Bernoulli random variable with success probability p has variance $p(1-p)$.

A simple normal-approximation confidence interval is:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

For our data ($\hat{p} = 0.0404$, $n = 567{,}577$), the standard error is

$$\mathrm{SE} = \sqrt{\frac{0.0404 \times 0.9596}{567{,}577}} \approx 0.000261,$$

yielding a 95% CI of

$$0.0404 \pm 1.96 \times 0.000261 \approx [0.0399,\ 0.0409],$$

or [3.99%, 4.09%].

As a robustness check, we also computed a Wilson interval, which adjusts the center and the standard error by incorporating a z^2 term (effectively adding a small correction to account for estimating p). With $n = 567{,}577$, the Wilson and normal-approximation intervals are nearly identical; the Wilson correction is most consequential when n is small or when p is very close to 0 or 1.

C.4.2 New Users Only

To address left-censoring, we repeated key analyses for users whose accounts were created after September 10, 2025 (within the observation window).

Table C.4: Adoption Rates: New Users vs. All Users

Metric	**All Users**	**New Users Only**	**Difference**
n Users	8,153	2,341	—
n Nudges displayed	567,577	89,432	—
Adoption rate	4.04%	4.87%	+0.83pp
Users with adoption	52.3%	58.2%	+5.9pp

New users showed modestly higher adoption, consistent with the tenure effect finding. The pattern of results was qualitatively similar for new users, suggesting findings are not driven primarily by left-censoring artifacts.

C.4.3 Excluding High-Volume Users

To assess whether concentration among power users drives results, we repeated analyses excluding the top 10% of adopters.

Table C.5: Adoption Patterns Excluding Top 10% Adopters

Metric	**Full Sample**	**Excl. Top 10%**	**Change**
n Adopters	4,261	3,835	−10.0%
n Adoptions	22,934	11,089	−51.7%
Mean adoptions/adopter	5.3	2.9	−45.3%
Gini coefficient	0.799	0.612	−0.187

Excluding high-volume users substantially reduces concentration but does not eliminate it, suggesting that power-law-like dynamics operate across the adoption distribution.