Dissertation

# EDUCATIONAL TOOLS SUPPORTING HUMAN-AI KNOWLEDGE WORK AMONG GENERATION Z STUDENTS

**Elena Eleftheriou**

University of Cyprus
Department of Computer Science

**Department of Computer Science**

**May 2026**

UNIVERSITY OF CYPRUS

Faculty of Pure and Applied Sciences

Department of Computer Science

# EDUCATIONAL TOOLS SUPPORTING HUMAN-AI KNOWLEDGE WORK AMONG GENERATION Z STUDENTS

**Elena Eleftheriou**

Supervisors

Dr Marios Constantinides, Dr George Pallis

The Thesis was submitted in partial fulfillment of the requirements for obtaining the Computer Science degree of the Department of Computer Science of the University of Cyprus

May 2026

I declare that this dissertation and any accompanying software are my own original work, conducted in full compliance with the University of Cyprus Code of Conduct for Research (https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/), and with full accountability on my part for the accuracy, integrity, and validity of all content

# Acknowledgements

First of all, I want to thank Dr. Marios Constantinides and Dr. George Pallis for their guidance throughout this whole journey. I am especially grateful to Dr. Constantinides, who truly taught me everything I know about the research process. From conducting my first literature review to designing the studies, running pilots, and finally putting everything into words. I also want to thank Dr. Pallis for his support and for always pointing out the small details that ended up making a huge difference in the final result.

A huge thank you goes to my friends who helped during the initial phases of this project. They tested my tool again and again, giving me the feedback I needed to make it perfect. From fixing bugs to refining the experimental tasks, their honest advice during the pilot studies was invaluable.

I also want to thank the people who helped me with the recruitment. To my cousin and the group of friends who reached out to their networks to find participants, I couldn't have gathered such a great pool of data without your help. I also want to thank every single person who took the time to participate in my study.

Finally, I want to thank my family for always being there and giving me the support I needed to finish this thesis.

Note on Publication:

I am also very proud to share that this work was accepted for publication:

> Eleftheriou, E., Pallis, G., & Constantinides, M. (2026). Confidence Without Competence in AI-Assisted Knowledge Work. In Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work (pp. 1-25). DOI: http://doi.org/10.1145/3808045.3808055

# ABSTRACT

Large Language Models (LLMs) are widely used by students, yet their tendency to provide fast and complete answers may discourage reflection and foster overconfidence. This thesis examines how alternative LLM interaction designs support deeper thinking without excessively increasing cognitive burden. We conducted a two-phase mixed-methods study. In Phase 1, interviews with 16 Gen Z students informed the design of Deep3, a web-based system with three interaction modes: a) future-self explanations, b) contrastive learning, and c) guided hints. In Phase 2, we evaluated Deep3 with 85 participants across two learning tasks. We found that a standard single-agent baseline produced high perceived understanding despite having the lowest objective learning. In contrast, while workload differences were not statistically significant after covariate adjustment, future-self explanations yielded the strongest gains in objective understanding and the closest alignment between perceived and actual knowledge. Guided hints also improved learning outcomes and demonstrated a unique "frustration reversal" where higher engagement correlated with lower stress. These findings show that effort, confidence, and learning systematically diverge in LLM-supported work.

# TABLE OF CONTENTS

# 1 Introduction

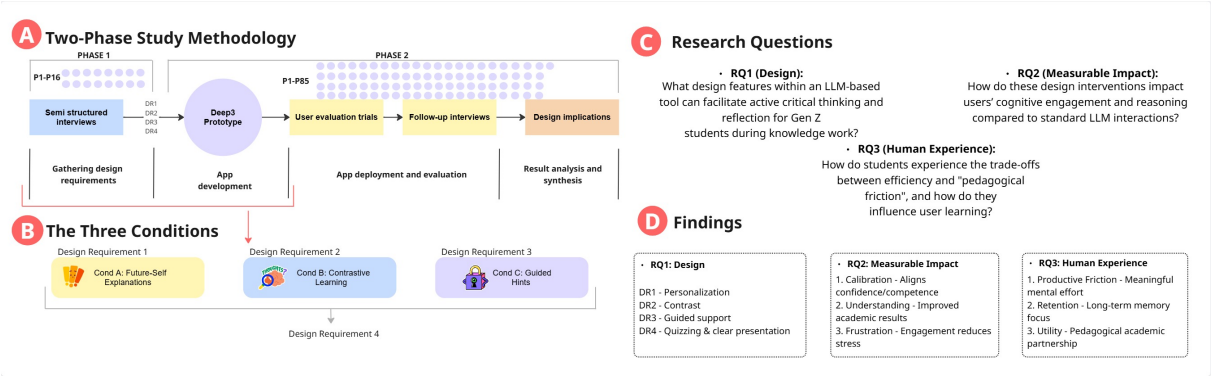



Figure 1.1: Overview of our study and findings, which illustrate (A) the two-phase study methodology; (B) the three conditions developed from the design requirements gathered in Phase 1; (C) our research questions; and (D) the resulting findings.

## 1.1 Motivation

The rapid integration of LLMs into higher education has fundamentally changed the academic environment for Generation Z (Gen Z) and emerging young professionals. As a "digital-first" generation, these students have quickly adopted tools like ChatGPT to speed up information retrieval, note-taking, and problem-solving, treating generative AI (GenAI) as "the new calculator" for the modern era [61]. While recent studies show that nearly all Gen Z students utilize these tools for tasks ranging from note-taking to test preparation [3, 51], this adoption has created a generational divide. Students embrace the efficiency of AI, while educators express growing concern over the ethical and pedagogical shifts this "calculator" mentality introduces to the classroom [11].

Despite the undeniable gains in productivity, the widespread reliance on LLMs risks a major loss of meaningful cognitive growth. Current research suggests that when AI handles the complex cognitive demands of knowledge work, it can lead to "displaced thinking", which is a phenomenon where users experience lower brain engagement and a reduced capacity for critical evaluation [26, 37]. As Sarkar et al. [59] argue in their exploration of how to stop AI from killing critical thinking, there is a danger that AI tools act as "autopilots" that bypass the necessary friction required for deeper learning. When users prioritize speed over understanding, they often fall into a "dependency trap", a state in which high confidence in AI results correlates with reduced critical reasoning [27, 41]. This "dependency trap" is reinforced by an illusion of

competence, where reliance on seamless AI outputs leads users to stop thinking for themselves, and fail to notice errors or missing logic. [43].

## 1.2 Aims and Objectives

This thesis investigates how the integration of LLMs can be structured to support the cognitive development of Gen Z students rather than bypassing it. Instead of proposing a total ban on AI, our focus is on investigating how "pedagogical friction" can be integrated into existing LLMs to accommodate the demands of modern higher education.

We position this thesis as an exploratory study with the following objectives:

- To surface the trade-offs between different AI agent designs and student engagement.
- To capture lived user experiences regarding AI-mediated learning.
- To provide practical design implications for maintaining critical reasoning in an era of automated knowledge work.

## 1.3 Research Questions

We address three research questions (RQs):

**$RQ_1$:** What design features within an LLM-based tool can facilitate active critical thinking and reflection for Gen Z students during their academic work?

**$RQ_2$:** How do these design interventions impact users' cognitive engagement and reasoning compared to standard LLM interactions?

**$RQ_3$:** How do students navigate the tension between speed and this new friction, and how does it influence user learning?

## 1.4 Methodology Overview

To answer these questions, we carried out a two-phase mixed-methods study. In Phase 1, we conducted semi-structured interviews with 16 Gen Z students studying at universities across Europe, during which we explored how they used LLMs during their studies, and how could a tool help them engage more actively in thinking and learning. Their insights informed the design of Deep3, a web-based system offering deliberate, critical and analytical reflection across three different agents ( 1.1). Deep3 was designed to accommodate various modes of academic engagement, allowing users to select a specific interaction style based on their current study goals with the AI assistant providing tailored suggestions for each mode.

In Phase 2, we evaluated the Deep3 application through a between-subjects experiment with 85 participants, at European universities. We utilized a mixed-methods approach, combining interaction log files, pre- and post-task questionnaires, and semi-structured follow-up interviews, to examine how our different agents impact student understanding and cognitive load.

## 1.5 Contribution

In answering our RQs, we made two main contributions:

1. Through semi-structured interviews with Gen Z students (§3), we identified four design requirements for an LLM-based tool capable of fostering critical thinking. With these requirements at hand, we designed and deployed a web-based system called *Deep 3* (§4).
2. We conducted a between-subjects evaluation comparing four agent configurations (§5) to gain empirical evidence of how these designs impact student reasoning, cognitive load, and understanding compared to a standard LLM. We found that while a standard LLM satisfies a "speed addiction" with a misalignment between confidence and competence, an LLM with pedagogical friction can close this gap, enhancing actual understanding without necessarily increasing user frustration(§6).

In light of these results, we conclude with empirical and design insights on how pedagogical friction shapes the cognitive and behavioral dimensions of student interactions with distinct LLM agent designs (§7).

## 1.6 Structure of the Thesis

This thesis is divided into eight chapters, tracing the progression from initial problem identification to system design and empirical evaluation. In the first chapter, we present the motivation behind the research, the specific aims and objectives, the research questions, and an overview of the core contributions. The second chapter reviews existing literature regarding Generation Z's relationship with LLMs in educational settings, the cognitive impacts of automated knowledge work, and established technologies and techniques for fostering reflection. Chapter 3 synthesizes insights from preliminary literature on Gen Z and LLMs with the findings of our Phase 1 formative study to establish the design requirements for the Deep3 system.

The fourth chapter details the technical implementation of the Deep3 tool and the specific logic governing its three agent configurations. Chapter 5 describes the design of the Phase 2 summative study, including participant demographics, experimental tasks, and the mixed-methods approach to data analysis. Chapter 6 Reports the quantitative and qualitative findings of the user evaluation. The seventh chapter interprets our findings, offering design implications for

future LLM-based educational tools and addressing the study's limitations. Finally, chapter 8 summarizes the research findings and provides concluding thoughts on the role of pedagogical friction in modern higher education.

# 2 Related Work

We reviewed the key research areas that inform this thesis and organized them into three areas: i) Gen Z and LLMs in education (§2.1); ii) the cognitive impact of LLM use (§2.2); and iii) technologies and techniques for reflection (§2.3).

## 2.1 Gen Z and LLMs in Education

Gen Z, born between 1995 and 2012, is the first generation to grow up with constant access to digital technology and social media, which gives them a digital-first mindset, as Puiu [55] describes. As a result, Gen Z students often expect that educational institutions will provide up-to-date technological tools and resources to support their learning. Given this, it is becoming widely accepted that GenAI technologies, particularly LLMs such as ChatGPT, have the potential to transform education by improving teaching and learning [34].

According to recent studies, Gen Z students are using AI tools extensively for their academic work. For example, a large-scale survey reported that 97% of Gen Z students use them, with 66% to study, 56% to prepare for tests and 46% to take notes [51]. Moreover, a study on how people use ChatGPT shows that users primarily interact with the model to ask questions, complete tasks, and be creative, with 46% of these users belonging to Gen Z, highlighting their preference for AI-assisted learning [13].

The 2025 HEPI Student Survey indicates that student engagement with GenAI for assessments has jumped from 53% last year to 88% this year, with a primary focus on concept explanation and study support [25]. Similarly, global evidence from another study, indicates that students' adoption of ChatGPT is strongly influenced by ease of use, reinforcing the belief that Gen Z actively integrates these tools into their academic practices [3]. However, Oliński and Sieciński [52] note that Gen Z's intention to use AI is driven mainly by "Perceived Usefulness".

This integration is not without friction. Research shows that Gen Z students are generally optimistic about the benefits of LLMs, expressing strong intentions to integrate these tools into their learning practices [11]. At the same time, Gen X and Millennial teachers tend to approach these tools with greater caution, raising concerns about ethical and pedagogical implications [11]. This tension is most noticeable in how people use their thinking skills. Teachers are increasingly worried about what they call "metacognitive laziness", meaning that students are not actively thinking [48]. This generational divide clearly shows the tension that LLMs introduce regarding cognitive use.

## 2.2 The Cognitive Impact of LLM Use

The widespread adoption of LLM tools raises questions about how they may affect cognitive processes, critical thinking and learning outcomes. Recent studies suggest that reliance on LLMs for gathering information and problem-solving can both support and hinder learning. From one perspective, LLMs can improve a student's understanding by providing explanations, but they may also increase reliance on incorrect answers [35].

Over-reliance, may also reduce student's critical reasoning. For example, Kosmyna et al. [37] found that participants, who used LLMs, showed lower brain engagement and less critical evaluation of their work compared to those who completed tasks without it. These findings were validated in 2025 by the MIT Media Lab, which used EEG monitoring (electrical activity in brain) to show that students enter a "neural standby mode" when using AI, leading to poor memory retention [22].

Similarly, another study revealed that higher confidence in GenAI was associated with less critical thinking [41]. Lee et al. [41] expanded on this, finding a direct negative correlation between trust in AI and the performance of critical verification steps. Gerlich [27] also reported that frequent use of LLM tools, especially for younger users, was linked to weaker critical thinking skills.

Georgiou [26] observed that ChatGPT use resulted in significantly lower cognitive engagement scores compared to participants who did not use it. Lastly, Rogers [58] highlighted the importance of effort in learning, suggesting that requiring students to engage more actively with LLM tools could improve learning outcomes. This is supported by Hong et al. [31], who found that "Cognitive Offload Instruction", where AI is restricted to low-level tasks, can actually preserve and enhance high-level critical thinking. Together, these findings highlight the risks of overreliance on GenAI.

## 2.3 Technologies and Techniques for Reflection

### 2.3.1 Self-Explanation in Learning

Research has consistently shown that self-explanation supports learning by enabling deeper understanding and problem-solving skills. Chi et al. [14] found that students who explained ideas to themselves while studying performed better on tests because it helped them notice gaps in their knowledge and connect new concepts. Similarly, McEldoon et al. [44] showed that prompting students to self-explain during math problems led to better conceptual understanding than simply giving them more practice.

Recent large-scale research in digital learning environments (DLEs) further validates these findings, demonstrating that this strategy consistently improves academic performance [62]. By bridging traditional cognitive theories with modern digital platforms, evidence suggests that self-explanation helps students think more deeply and learn better, even with more automation in education.

### 2.3.2 AI-Mediated Critique

Recent HCI research has explored ways to keep users actively engaged and reflective rather than putting knowledge work on "autopilot" [59]. One strategy is to turn the LLM into a critic by having it reflect on its own outputs. For example, Sarkar et al. [59] built a spreadsheet tool where the LLM suggests an action and then provides a short "provocation", that is a short critique highlighting possible issues or alternatives. This idea led users to pause, reflect, and think critically before accepting LLM suggestions [19].

Beyond internal reflection, other approaches utilize comparison and dialogue to stimulate user engagement. Buçinca et al. [9] demonstrated that contrastive explanations, which compare an LLM's chosen answer to a predicted human response, can improve decision-making without reducing accuracy. Extending this approach, the "Judge-Questioner" framework employs dual LLMs to generate critical questions that challenge a student's unsupported claims, significantly increasing engagement with argumentative texts [23].

### 2.3.3 Scaffolding and Socratic Tutoring Systems

Another approach is to have LLMs ask questions instead of giving direct answers. Danry et al. [18] show that when LLM explanations are framed as questions, people become better at spotting flaws in reasoning. In this way, LLMs act more like a "critical thinking coach" than just an "information provider". Similarly, Park and Kulkarni [53] describe "Thinking Assistants" as interactive agents that use LLM dialogue to generate guided and domain-specific questions. Their experiments found that participants using these assistants reflected more on their thinking compared to those using answer-focused agents. This question-driven approach can improve learning, though studies show that questions alone may increase engagement without additional learning gains [6].

Several LLM tutoring systems are also designed to require reflection. For instance, Iris Tutor [5] for programming tasks gives hints and counter-questions instead of full solutions. It uses chain-of-thought prompting with awareness of the student's code to adapt its advice. Students reported that Iris was effective and boosted their confidence. Similarly, TutorLLM [42] adapts its explanations to each student's progress by tracking what they know. This approach helps

students better understand the material, feel more satisfied with the learning experience, and perform better on quizzes. The SocratiQ [33] system also uses a Socratic framework, creating personalized learning paths through interactive questioning. In all these systems, LLMs guide students to solve problems and explain their steps, rather than passively consuming answers.

### 2.3.4 Research Gaps

Despite growing interest in LLM-supported learning and reflection, several gaps remain in existing literature. Firstly, while self-explanation has been widely shown to enhance understanding and critical thinking, there is no recent HCI research exploring how this strategy might function when integrated with AI tools. Secondly, previous work has explored methods for prompting users to reflect through critique or contrast [9]. However, these approaches typically rely on pre-generated statements, and not dynamic ones. Finally, prior work on question-driven or Socratic LLMs has shown that asking users questions can increase engagement, but these systems often lack mechanisms for additional learning gains [5, 6]. Our study addresses these gaps by examining how different forms of LLM interaction can support active reflection and critical thinking during knowledge work.

# 3 Formative Study: Design Requirements

To identify the requirements for our web application, we followed a two-step method that combined insights from established literature with findings from a targeted formative study. This allowed us to first identify broad patterns in how Gen Z students use LLMs and then explore how a tool could specifically help them engage more actively in thinking and learning.

## 3.1 Step 1: Insights from Literature on Gen Z and LLMs

To establish a baseline for Gen Z's interaction with AI, we first reviewed recent large-scale surveys and meta-analyses, which indicated increasing integration of AI into everyday learning activities. Data from a survey [24] of over 12,000 students confirms a near-universal adoption rate (97%). The majority of students utilize AI as an academic assistant for test preparation (56%) and searching for scholarly literature (55%). While this usage significantly boosts immediate academic performance ($g = 0.867$), a meta-analysis by Wang and Fan [65] indicates a much lower impact on critical thinking ($g = 0.457$), suggesting that students often stop at memorization instead of developing a deeper understanding. This "thinking gap" is further refined by the affordance-based research of Lee et al. [40]. They found that Gen Z particularly uses AI to generate quick ideas and overcome the psychological barrier of starting with a blank page.

However, while these large-scale studies identify general trends in adoption, they do not explain the detailed interaction requirements needed to address this "thinking gap". This synthesis highlights the need for our formative study, which goes beyond broad statistics to examine the specific teaching approaches required to help students move from using AI as passive academic assistants to becoming critical thinkers.

## 3.2 Step 2: Formative Study

While Step 1 provided the "what" and "how many", we conducted semi-structured interviews with Gen Z students to explore how a tool could help them engage more actively in thinking and learning. The study helped us identify specific design requirements. The formative study took place in September 2025.

### 3.2.1 Participants

We recruited 16 Gen Z students (P1-P16) through university connections using convenience sampling. Participation was entirely voluntary, and all individuals provided informed consent

Table 3.1: Participant demographics.

| ID | Age | Gender | Education | Field of Study | LLM Interaction Frequency |
|---|---|---|---|---|---|
| 1 | 22 | Female | Bachelor's degree (in progress) | Computer Science | Often (a few times a week) |
| 2 | 23 | Female | Master's degree (in progress) | Philology | Sometimes (a few times a month) |
| 3 | 23 | Female | Master's degree (completed) | Philology | Daily |
| 4 | 22 | Male | Bachelor's degree (completed) | Computer Science | Daily |
| 5 | 21 | Male | Bachelor's degree (in progress) | Computer Science | Often (a few times a week) |
| 6 | 21 | Male | Bachelor's degree (in progress) | Computer Science | Daily |
| 7 | 21 | Female | Bachelor's Degree (in progress) | Civil Engineering | Daily |
| 8 | 21 | Female | Bachelor's degree (in progress) | Civil Engineering | Daily |
| 9 | 22 | Male | Bachelor's degree (in progress) | Computer Science | Daily |
| 10 | 21 | Female | Bachelor's degree (completed) | Chemical Engineering | Daily |
| 11 | 22 | Male | Bachelor's degree (in progress) | Chemical Engineering | Sometimes (a few times a month) |
| 12 | 22 | Male | Bachelor's degree (in progress) | Mathematics | Sometimes (a few times a month) |
| 13 | 22 | Female | Bachelor's degree (in progress) | Computer Engineering | Often (a few times a week) |
| 14 | 21 | Female | Bachelor's degree (completed) | Economics | Sometimes (a few times a month) |
| 15 | 23 | Female | Bachelor's degree (completed) | Electrical Engineering | Daily |
| 16 | 22 | Male | Bachelor's degree (in progress) | Business | Daily |

prior to the study. All participants were enrolled in a bachelor or master program and represented a diverse range of academic disciplines, including Philology, Economics, Computer Science, Electrical Engineering, Civil Engineering and Chemical Engineering. They were aged 21-23 years, with 9 female and 7 male students. Participants reported moderate to high interaction frequency with AI models, rating their usage between 3 to 5 on a 5-point scale (5 = highest level). All participants' demographics are presented in Table 3.1. The study was approved by the Cyprus National Bioethics Committee.

While the sample size ($N = 16$) is small for statistical generalization, it is consistent with established HCI (Human-Computer Interaction) standards for formative design studies, where 12-20 participants are typically sufficient to reach thematic saturation and identify critical usability requirements [28].

### 3.2.2 Procedure

Each participant completed a demographic survey prior to the semi-structured interviews. Interviews lasted 20-30 minutes and took place either online or in person, depending on participant preference. The interview protocol was informed by prior literature on LLM-assisted learning and reflective tool use, particularly studies exploring how LLMs can support critical thinking and engagement [6, 18, 19, 53]. The interview questions are provided in Appendix I.1.

### 3.2.3 Data Collection and Analysis

Interviews were recorded, transcribed using rev.com, and reviewed for accuracy. We used thematic analysis for its flexibility in exploring user perspectives [8]. We began with repeated

reading and annotation to develop codes (e.g., critical reflection, engagement strategies, interaction style with LLMs). The codes were iteratively grouped into themes through discussion. By synthesizing them with large-scale trends from literature, we identified four design requirements (DRs), three of which were focused on functionality and one on presentation. A full visual mapping of this analysis process on Miro, along with a detailed table of the thematic codes and representative participant results, is provided in Appendix I.2, and Appendix I.3.

### 3.2.4 Design Requirements

Participants emphasized the importance of a learning tool that adapts to individual knowledge levels and preferences. Users described wanting an experience that feels like *"asking a question to yourself but to a more professional you"* (P16), highlighting the need for the system to adjust explanations to each learner's needs. They also expressed that the tool should communicate in ways that align with each person's thinking and preferences, making answers more understandable and natural, as P12 described, *"This will be simpler and more easy to use because it will be like I'm talking to myself but myself that knows a lot more"*. These insights inform our first design requirement, which focuses on adaptive personalization, ensuring that questions and answers are tracked in a clear and simple way. This design requirement is supported by prior research on adaptive learning environments. Adaptive systems that adjust feedback and content based on learners' knowledge levels have been shown to improve engagement and learning performance [2]. As Nunez et al. [49] argue, these systems succeed by bridging the gap between plain technology and teaching methods, allowing for a "customized teaching" approach. This approach directly reflects the "Content Curation" capability identified by Lee et al. [40], where Gen Z students prefer AI interactions that resemble personalized tutoring rather than static search results.

> **DR1: Adaptive Personalization.** The system should adjust to the user's knowledge level, and tailor explanations to individual weaknesses.

Participants described how encountering incorrect solutions or alternative perspectives can promote deeper engagement and learning. As P11 explained, *"If it shows me an answer that I know it's wrong, I will stay and think. This will help me stay engaged more"* . Likewise, P12 highlighted the value of alternative perspectives, suggesting that *"it should provide the user two or three ways of approaching the problem"*. These reflections emphasize the value of contrastive learning, where users engage with correct and incorrect answers, as well as counterarguments that introduce alternative perspectives. This requirement is supported by research showing that engaging with contrasting information and counterarguments can strengthen understanding and reasoning. For instance, studies on argument–counterargument integration show that actively addressing opposing viewpoints improves students' writing quality and critical thinking [50].

Building on this evidence, Wang and Fan [65] advocate for HOTS-scaffolding (Higher-Order Thinking Skills) to mitigate the lower impact of AI on higher-order thinking skills.

> **DR2: Contrast and counterargument.** The system should combine contrastive questions and counterarguments, helping users strengthen their critical thinking by reflecting on differences and questioning their understanding.

Participants also expressed a preference for guided, step-by-step learning rather than immediate answers. P3 noted, *"It would be nice to have an option to see the answer, but not from the beginning"*, and P2 emphasized that, *"Human reasoning and critical thinking shouldn't be diminished"*. Users saw value in a teacher-like system that offers hints, provides feedback, and delivers full solutions only after the learner has attempted the problem. This guided approach was described as more engaging than simply using tools that provide instant answers, as P1 said *"I think it's okay to take a little bit more time than having it be done by someone else"*. This design requirement is informed by research on guided instruction, which shows that learners benefit more from structured guidance than from discovering solutions on their own [1, 36]. Also, it directly addresses the "instant answer trap" identified in usage rate studies [4].

> **DR3: Guided, step-by-step support.** The system should provide hints and guidance for solving problems, allowing learners to attempt solutions before revealing full answers, to support critical thinking and sustain engagement.

Finally, participants stressed the importance of dynamic quizzes and a clear, well-organized presentation. Users noted that requiring answers before moving forward would help them fill knowledge gaps, as P5 said, *"It's not about going further into the details, but quizzing me on the current information to make sure I don't have gaps"*. P4 and P5 added, *"If I answer something wrong, it explains and gives further questions"*. P5 also noted the necessity of dynamic questioning, *"Not having the opportunity to answer dynamically the questions adds an unnecessary friction to my work"*. Clear bullet-point summaries were preferred to help retain key information, with P14 adding, *"So that you can absorb the most important things"* and P8 saying, *"If I had a short summary, it would help me remember everything"*. This requirement is supported by prior work on active recall and formative assessment, which demonstrate that quizzes and feedback promote deeper learning and retention [10, 57], and addresses the "efficiency" needs highlighted in a survey [24], which found that 56% of Gen Z students use AI for test preparation.

> **DR4: Dynamic quizzing and clear presentation.** This design requirement focuses on the presentation of the system. The system should provide interactive quizzes with feedback and concise bullet-point summaries, making the tool easier to use while promoting active engagement.

# 4 Deep3: A Tool for Deliberate, Critical and Analytical Reflection

Based on the four design requirements identified in our formative study, we developed Deep3 (Figure 4.1). It is a prototype full-stack web application designed to support active reflection and critical thinking during knowledge work. The app offers three different AI agent configurations: Cond A - Future-Self Explanations (4.2.1), Cond B - Contrastive Learning (4.2.2) and Cond C - Guided Hints (4.2.3) addressing, DR1 (Adaptive Personalization), DR2 (Contrast and counterargument) and DR3 (Guided, step-by-step support).

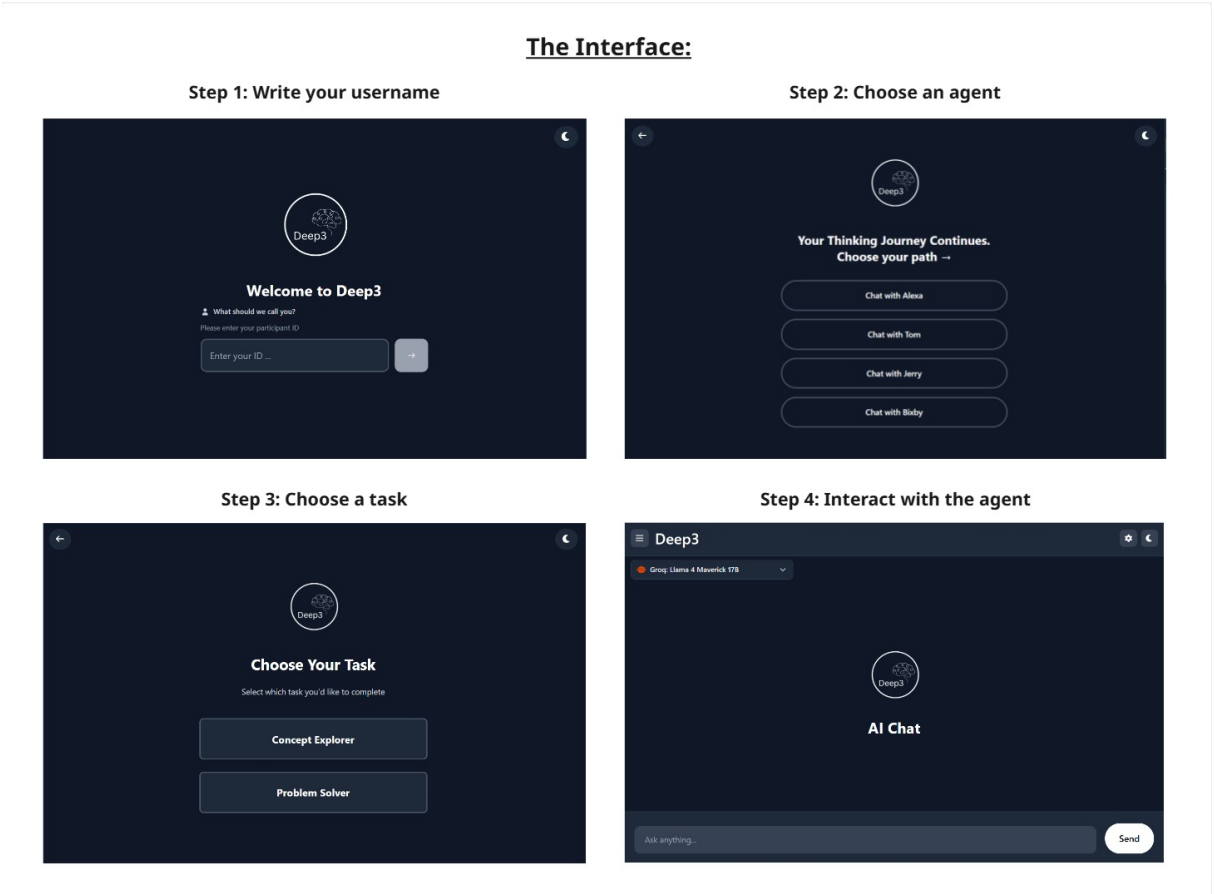
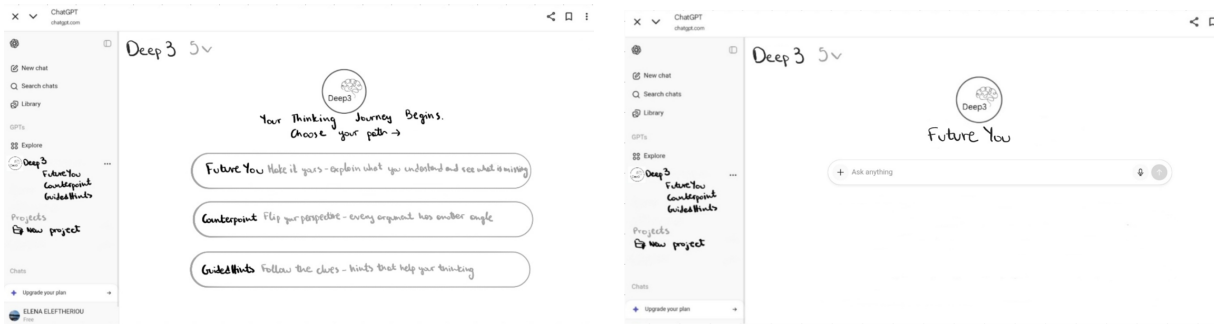



Figure 4.1: The Deep3 User Interface Flow. The sequence illustrates the four-step onboarding process: (1) user identification via participant ID, (2) selection of a specific AI agent, (3) task type categorization, and (4) the final interactive LLM chat interface.

## 4.1 Prototype Development

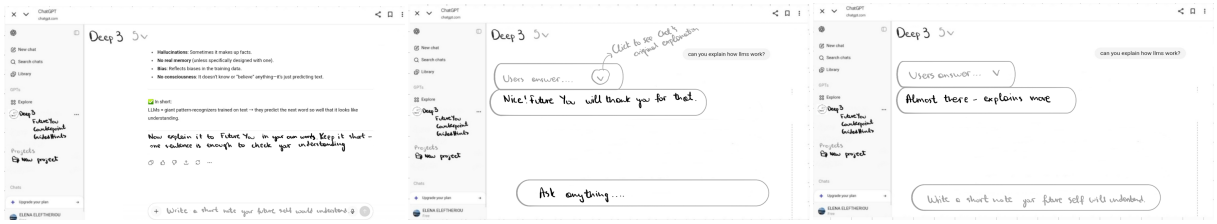
The development of Deep3 followed an iterative design process, beginning with the translation of the four identified design requirements (DR1–DR4) into visual concepts. During the initial phase, a series of low-fidelity sketches were produced to explore various layouts for the system. These early designs focused on the structural organization of the chat interface, specifically investigating how to distinguish between standard AI responses and the specialized interaction modes.

The following figures document the low-fidelity sketches developed. Figure 4.2 illustrates the entry point where users select the agent they want. Figure 4.3 details the conditional logic for Condition A. Finally, Figure 4.4 displays Condition B and C.



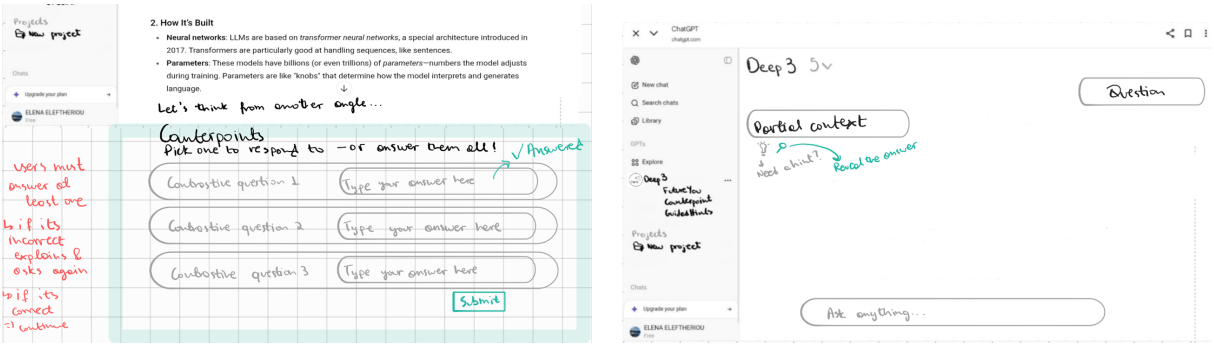

(a) Agent Selection (b) Agent Page

Figure 4.2: The first sketch shows the initial page were users have to choose an agent, and the second sketch shows the starting page of each agent (All agents have the same starting page, with just different agent names displayed).




(a) Agent Message (b) Correct User Answer (c) Wrong User Answer

Figure 4.3: These three diagrams show the first condition's functionalities.




(a) Condition B (b) Condition C

Figure 4.4: These two sketches illustrate the functionalities of the other two conditions.

The interface was designed to follow a familiar chat-based paradigm, similar to standard LLM interfaces. This choice was intentional to minimize the learning curve, allowing users to focus entirely on their tasks rather than navigating a complex new tool.

Following the refinement of these sketches, the project moved into a high-fidelity prototyping stage where the interaction flow was finalized.

## 4.2 Functionality Themes: The Three Conditions

Deep3 includes three functionality themes: Cond A - Future-Self Explanations, Cond B - Contrastive Learning and Cond C - Guided Hints (Figure 4.6). To address DR1 (Personalization and adaptivity), DR2 (Contrast and counterargument) and DR3 (Guided, step-by-step support), the app allows users to choose one of the three functionalities to work with. In this context, "functionality themes" refer to the specific interaction styles and behaviors the AI agent uses to help the user.

### 4.2.1 Cond A: Future-Self Explanations

The first condition is designed to transform the user from a passive reader of AI-generated content into an active participant. In this mode, after the agent provides an answer or explanation, it prompts the user to rephrase and explain that information in their own words. The agent, then, provides feedback to the user and decides if it is necessary to ask for an explanation again. This process continues until the agent believes that the user understood everything correctly.

The theoretical foundation for this condition is rooted in the Self-Explanation Effect. According to Chi et al. [14], self-explanation is a critical metacognitive activity that helps learners develop a deeper understanding of "when and why" certain solutions apply, rather than just memorizing

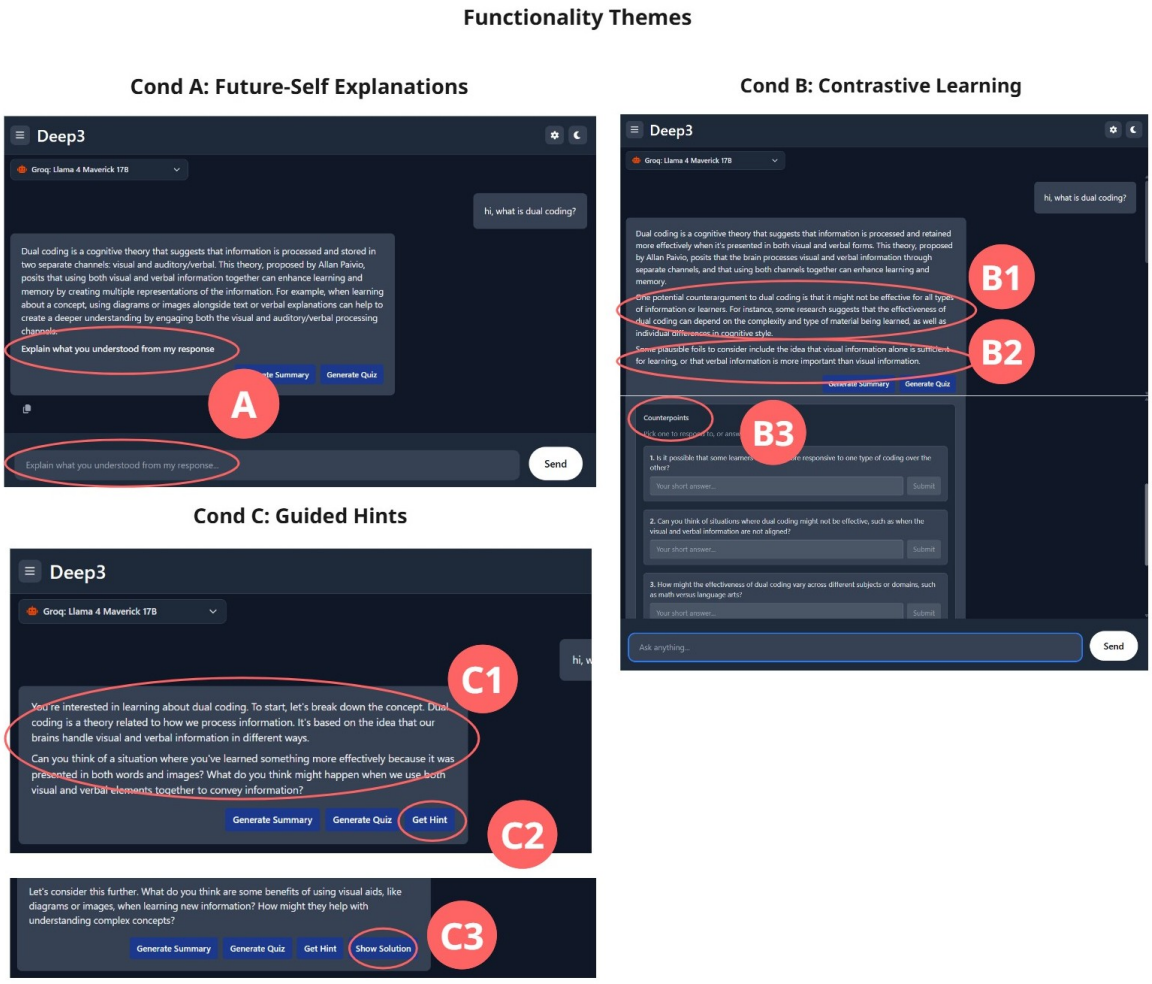



Figure 4.5: The Functionality Themes. This layout details the specific interactive elements for (A) Cond A: Future-Self Explanations, (B) Cond B: Contrastive Learning, and (C) Cond C: Guided Hints. Cond A, focuses on self-explanation through a recurring prompt, "Explain what you understood from my response", that ensures the user can accurately describe the topic before proceeding. Cond B, provides a multi-faceted approach involving counterarguments (B1), foils (B2), and specific contrastive questions (B3). Cond C, offers a scaffolded experience where the system breaks down the topic (C1) and provides manual controls for hints (C2) or full solutions after some interaction (C3).

steps. Furthermore, McEldoon et al. [44] suggest that this practice encourages learners to focus on the structural features of a problem rather than superficial details.

This condition addresses DR1 - Adaptive Personalization. By requiring the user to explain concepts in their own words, the system naturally adapts to the user's specific knowledge level (K) and learning style (L) [2]. If the system detects a misunderstanding in the user's self-explanation, it immediately adapts its next response to correct those misconceptions.

### 4.2.2 Cond B: Contrastive Learning

The second condition is designed to stop users from simply agreeing with the LLM without thinking it through. In this mode, users cannot just accept an answer and move on, instead they can respond to AI-generated counterarguments or "foils" (plausible alternative choice or common misconception the user might have). This interaction gives the option to the user to defend their reasoning or reconstruct their perspective with an opposing view, fostering a more critical and integrative thinking process.

The theoretical foundation for this condition is built on Contrastive Learning. According to Buçinca et al. [9], humans naturally think contrastively, often asking "Why this instead of that?". We leverage this by predicting a "foil" and highlighting the specific differences between the AI's suggestion and that foil. By making these differences clear and forcing the user to address them, the system helps the user spot their own misunderstandings and refine their thinking.

This condition addresses DR2 - Contrast and counterargument. By making the user to engage with a counterargument, the system creates productive friction that slows down the user's decision-making process. As Nussbaum and Schraw [50] argue, effective reasoning requires more than just defending a claim, it involves argument-counterargument integration.

### 4.2.3 Cond C: Guided Hints

The third condition is designed to help users solve problems without simply handing them the answer. Instead of providing a full solution immediately, the agent acts as a teaching assistant that offers incremental support. The user must actively work through the problem, with the option to unlock hints or the final solution only after they have attempted to answer the current step.

This condition addresses DR3 - Guided, Step-by-Step Support. By breaking down complex information into smaller, manageable pieces, the system ensures the user isn't overwhelmed. According to Alfieri et al. [1], "discovery learning", where users find answers themselves, is only effective when it is scaffolded. Without clear incremental guidance, users can suffer from "working memory overload", which makes it harder to learn.

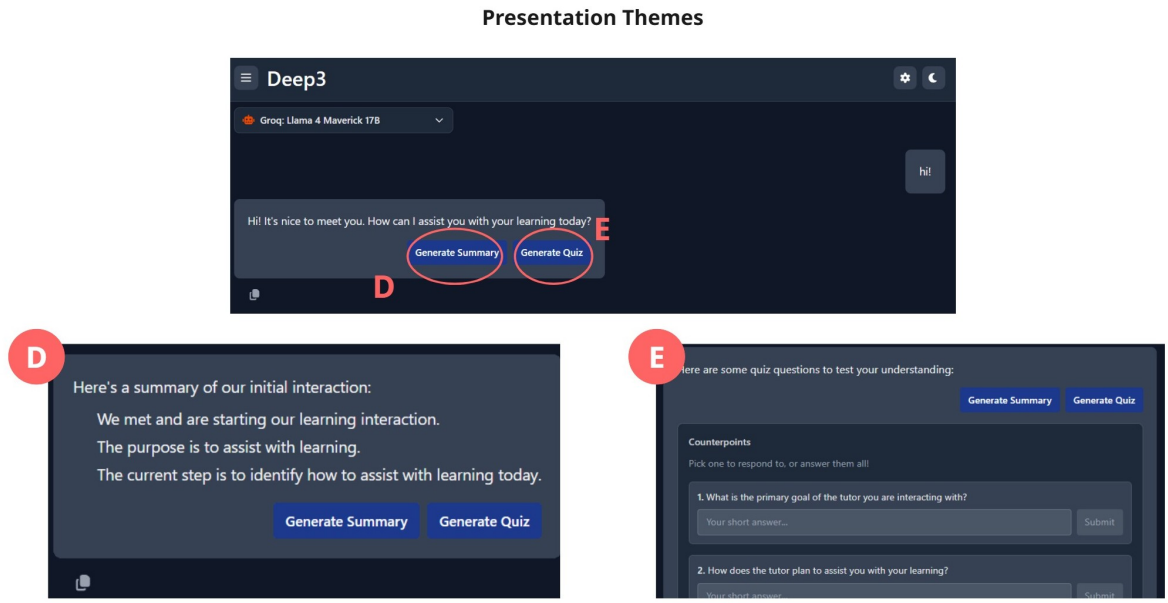



Figure 4.6: The Presentation Themes. Users have access to tools for generating summaries (D) and dynamic quizzes (E).

Technically, the agent operates using three distinct modes: Default, Hint, and Solution. In the Default mode, the agent ignores direct requests for the final answer and instead nudges the user to think about the next logical step. The "Hint" and "Solution" modes are only triggered when the user clicks a specific button in the UI. This ensures that the agent remains in a supportive, teaching role.

## 4.3 Presentation Themes

In addition to the functionality themes, we implemented Presentation Themes to address DR4 - Dynamic quizzing and clear presentation. While the functionality themes determine what the agent says, the presentation themes control how that information is structured and delivered to the user. This layer is integrated throughout the entire interface, allowing users to transform responses from any of the three previous agents into more effective learning formats, such as summaries or interactive quizzes.

The theoretical backbone of these presentation themes is Test-Enhanced Learning. According to Roediger III and Karpicke [57], the act of retrieving information from memory is a more powerful learning event than simply re-studying or re-reading material. To apply this, the system can enter Quiz Mode, which triggers the generation of 3-5 short-answer questions. Short-answer questions were chosen based on the literature's recommendation that the cognitive advantage of retrieval is significantly stronger for recall tasks than for multiple-choice tasks.

For moments when the user needs to organize their thoughts, the system offers Summary Mode. Based on the literature's recommendation to present material "clearly and concisely", this mode reformats complex AI responses into structured bullet points.

## 4.4 Prompt Engineering

To ensure that the agents provide consistent and helpful feedback, we followed established best practices for prompt engineering techniques. Our approach was guided by several techniques, specifically focusing on providing clear instructions, explicit constraints, and relevant context to improve the model's reasoning performance [20, 60].

Every prompt used in Deep3 follows a standardized three-part structure to maintain consistency: *Role, Context, and Task*. First, we assign the agent a specific Role, to establish the tone and perspective of the response. Next, we provide the theoretical background for the specific reflection technique being used. This information is based on established educational literature and ensures the agent stays grounded in proven pedagogical methods. Finally, the Task defines the direct instructions the model must follow. Every task begins by requiring the agent to strictly take into account the theoretical context described in the prompt for every response. From there, the task defines specific functional behaviors for every condition. The prompts are provided in Appendix II.

## 4.5 Platform and Implementation

Deep3 is implemented as a full-stack web application. The system is built on a modern three-layer architecture that separates the user interface, the main system logic, and data storage. This modular design ensures that the user interface remains responsive, while allowing the system to handle complex tasks like managing multiple LLM providers (Figure 4.7).

### 4.5.1 Frontend Development

The client-side application was built using the React 19 framework and Vite, providing a responsive user interface. The user interface employs a mobile-first design using Tailwind CSS, ensuring that participants have a consistent and accessible experience, whether they are using a desktop computer or a mobile device. To manage the complexities of an experimental environment, the application uses React Router to move between different tasks and the React Context API to manage shared information. This approach manages important global states such as user authentication, persistent theme preferences for light and dark modes, and real-time conversation history.

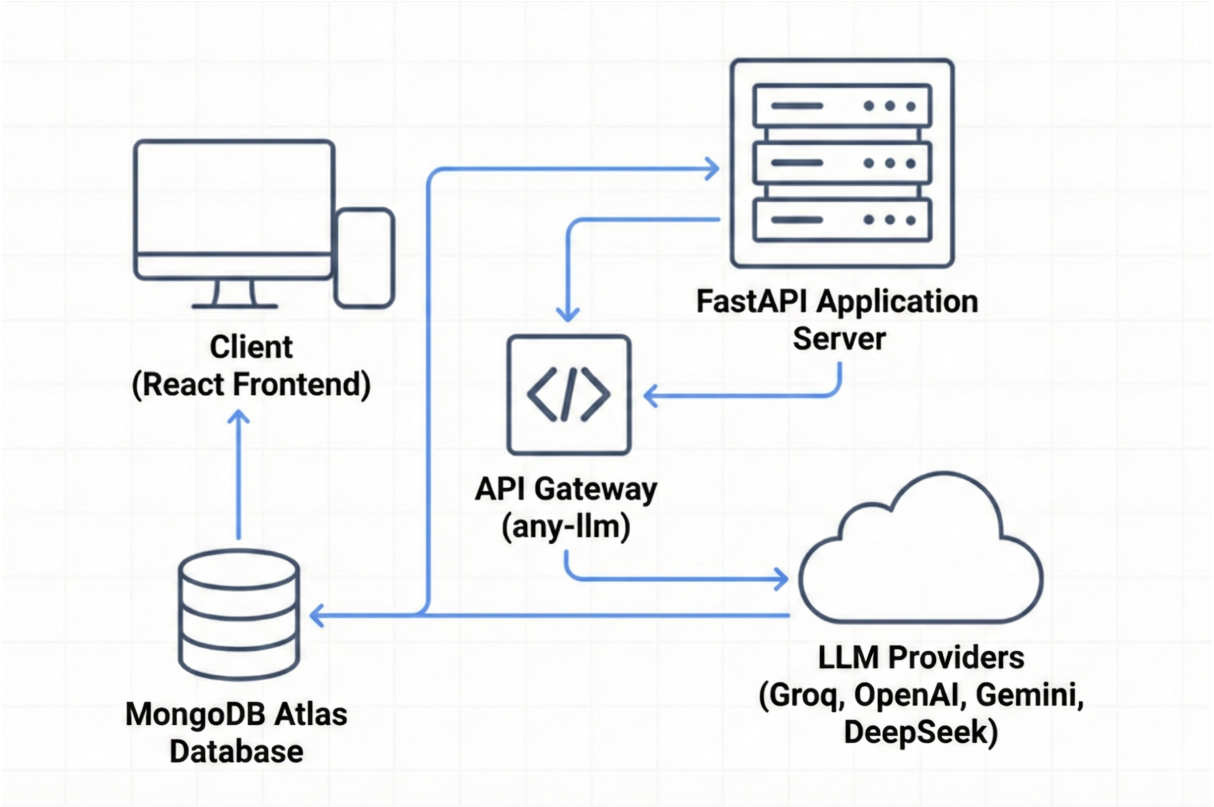



Figure 4.7: The high-level system architecture of the Deep3 platform. It shows the flow of data from the React-based client through the FastAPI layer to the integrated LLM providers. The diagram further highlights the central role of the any-llm gateway in managing different models and the MongoDB Atlas database in providing storage for experimental session logs and metrics.

A central feature of the frontend is the ChatUI component, which serves as the primary interaction point between the student and the AI. To enhance pedagogical effectiveness, the interface integrates ReactMarkdown and KaTeX for rich-text rendering. This allows the system to display structured information, such as headers, code blocks, and complex LaTeX mathematical notation, with clear visuals.

### 4.5.2 Backend Development and AI Integration

The server-side infrastructure is powered by FastAPI, a high-performance Python framework chosen for its asynchronous capabilities. The backend is the main system that manages and coordinates the AI. It simplifies communication with different AI models by using the any-llm interface library [45], which is an open-source tool that provides a standard way to connect to different AI providers. This allows the system to work with many providers and easily switch between models from Groq, OpenAI, Google Gemini, and DeepSeek. For the summative study, the platform utilized the Groq development key to access the Llama-3-70B model. This model

selection was informed by the LMSYS Chatbot Arena leaderboard [15], where the model maintained a top position among available models for text-based reasoning and instruction-following tasks.

Besides handling communication with AI models, the backend also controls the logic for different experiment settings. The system prompts are stored in one separate module, which keeps the system organized and allows for iterative prompt engineering. The server also enforces research-specific constraints, such as participant ID validation and submission limits, ensuring that each user completes no more than two tasks. Security is maintained through a combination of Cross-Origin Resource Sharing (CORS) middleware and environment variable management, which protects sensitive configuration data like API keys and database connection strings.

### 4.5.3 Database Design

Data storage is managed using MongoDB Atlas, a NoSQL database chosen because it allows flexible data structures. Since the research is exploratory, the database stores data in a denormalized format, meaning each record contains all the information from one participant session. This structure captures comprehensive metrics, including interaction counts, session duration, and condition-specific feature usage, which are essential for subsequent data analysis.

### 4.5.4 Deployment

The system is deployed across different cloud platforms to make sure it is reliable and fast. The frontend is hosted on Vercel, while the backend runs in Docker containers and is managed by Render. To make development easier, the system can automatically detect whether it is running in development or production. It does this by checking the application's hostname and then connects to either a local server or the live system. This setup, along with automatic backups and secure HTTPS connections, creates a strong and safe system for large-scale research.

# 5 Summative Study

## 5.1 Methodology

The summative study was conducted between December 2025 and January 2026, as a between-subjects study [12] consisting of two phases: a main study session, and a follow-up interview phase for a subset 14% of the sample. This design was selected because it assigns participants to only one condition, ensuring that each individual's data contributes to a single treatment group to prevent carryover effects. The first phase, was designed to allow participants to engage with the Deep3 application, which included two tasks, each implemented across four different agent conditions. Participants were randomly assigned to one condition to avoid cross-contamination of results. The sessions, including task completion and questionnaire responses, lasted between $25 - 40$ minutes, depending on the participant. To ensure ecological validity while maintaining a controlled environment, all sessions were conducted remotely via Microsoft Teams. This allowed participants to use their own laptops in a familiar setting while ensuring minimal external distractions. In the second phase, 14% of the participants took part in a semi-structured interview lasting approximately $10 - 20$ minutes. We used a mixed-methods approach [17], combining quantitative data from questionnaires with qualitative insights from both interviews and log files recorded during the Deep3 interaction phase. The study was reviewed and approved by our institution's ethics board.

## 5.2 Pilot Studies

To ensure the validity and effectiveness of the experimental design, six pilot studies were conducted [64]. These sessions served to refine the protocol and adjust the task difficulty. During the first three pilot sessions, the second task was initially designed as a scheduling problem. However, participant feedback revealed that the task felt more like an organizational exercise than a genuine problem-solving challenge, and they noted a preference for completing the task manually rather than engaging with the AI.

Consequently, for the remaining three pilot sessions, the second task was redesigned into a more complex, problem-oriented task that required higher-level reasoning. This adjustment ensured that the task was better suited to evaluate the utility of the AI intervention. Furthermore, these six preliminary studies were used to empirically determine the appropriate duration for each task, ensuring that the time limits were sufficient for completion without being excessive.

## 5.3 Experimental Tasks

To ensure ecological validity, the tasks were designed to align with documented LLM usage patterns among university students. This research shows that students often use LLMs to clarify academic concepts and assist in problem-solving or homework completion [25, 54, 56]. Consequently, two distinct tasks were developed.

*Task 1: Concept Explanation.* This task focused on academic clarification. Participants were given a university-related concept, and they were asked to engage with the chatbot until they felt they understood that concept. Based on our pilot studies, participants were given 6-8 minutes to finish this task, with the specific duration adjusted according to each participant's English proficiency.

*Task 2: Problem Solving.* This task required participants to solve a logic-based puzzle using the chatbot. This problem involved a balance scale and a set of coins, and an extra constraint to increase the complexity of the problem. Participants were given 10-15 minutes for this task, to accommodate varying levels of English proficiency.

Both tasks were presented in a randomized order to mitigate potential fatigue or learning effects. While maximum durations were set based on pilot data, participants were permitted to conclude a task early if they believed they had successfully reached a solution or achieved a sufficient understanding of the concept.

The specific tasks are illustrated in the figures below. All participants completed the same Problem Solving exercise, while they were randomly assigned to one of the 4 Concept Explanation exercises.

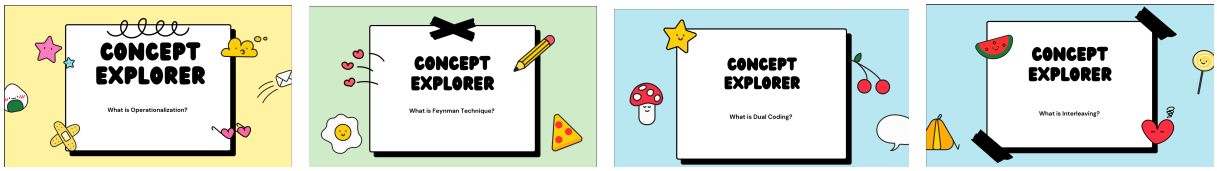



Figure 5.1: Concept Explanation Task

## 5.4 Participants

We recruited 85 participants (P1-P85) through personal networks and advertisements through the authors' institutions. Participation was entirely voluntary, and all individuals provided informed consent prior to the study. The sample included individuals from various academic and professional backgrounds, with the most represented fields being Computer Science, Mathematics, Engineering (Civil, Computer, and Chemical), Philology, and Economics. All participants

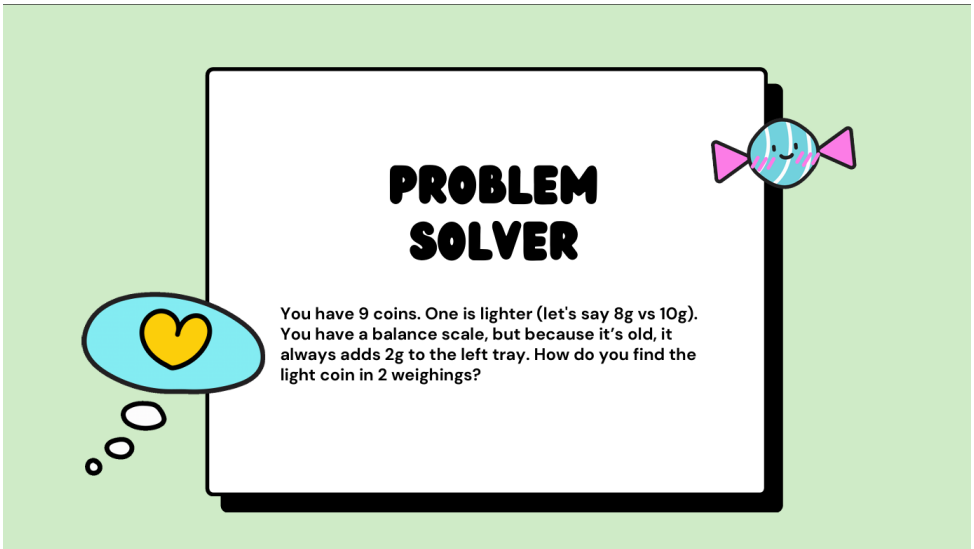



Figure 5.2: Problem Solving Task

resided in Europe, and studied in universities in the United Kingdom, the Netherlands, Italy, Bulgaria, Greece, and Cyprus. Their ages were from 18 to 27 years ($M = 22.14, SD = 1.66$), 41 identified as women, 43 as men, and 1 as other. In terms of education, the majority were currently enrolled in academic programs, including 57 undergraduate, 13 master's, and 3 doctoral students, while 12 participants had already completed their degree (6 Bachelor's and 6 Master's). All participants were proficient in English, with over 90% reporting a B2 level (Upper Intermediate) or higher. Regarding frequency of interaction with LLMs, participants reported high usage, with a mean general usage score of 3.67 ($SD = 1.17$) and an educational usage score of 3.86 ($SD = 1.15$) on a 5-point Likert scale. The most common applications of LLMs in their education included learning and understanding concepts (90.6%), answering homework questions (56.5%), and summarizing text (55.3%).

The full demographic data collected from the participant questionnaire are illustrated in the charts below.

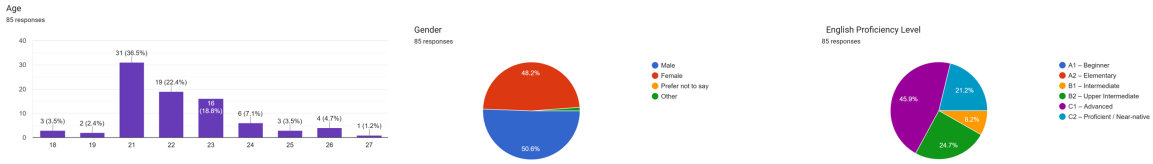



Figure 5.3: Age, Gender, and English Proficiency Level of participants.

Approximately 14% of those 85 people participated in the second phase of the study, which involved open-ended interview questions. The interview group consisted of 7 women and 5 men, with a mean age of 21.75 years. The majority were currently pursuing their undergraduate

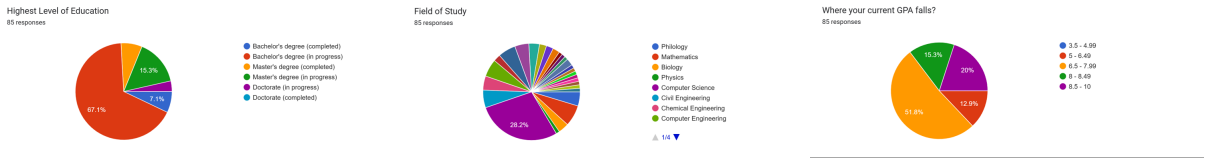



Figure 5.4: Highest Level of Education, Field of Study, and GPA of participants

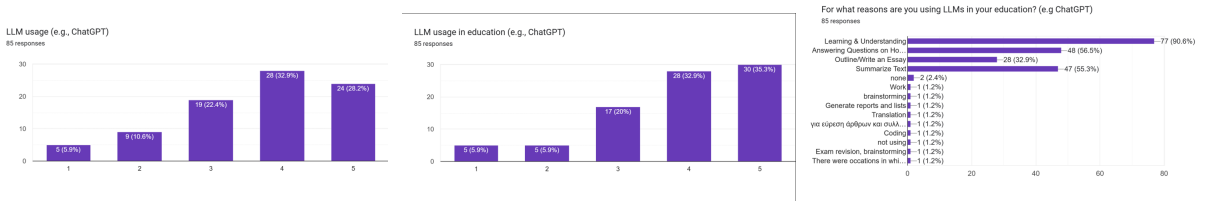



Figure 5.5: LLM usage, LLM usage in Education, and how do our participants use LLMs in their education

degrees, while one participant had recently completed their Bachelor's, and one was completing their Master's. In terms of academic background, Computer Science remained the most frequent discipline ($n = 6$), followed by Mathematics ($n = 3$), with additional representation from Computer Engineering, Medicine, Chemistry, and Communication and Marketing. All interviewees demonstrated high English proficiency, with 58% reporting a C2 level. Their interaction with LLMs was high, with 75% reporting frequent general and educational usage. Beyond standard tasks like summarizing and concept learning, these participants reported using LLMs for advanced academic support, such as generating custom practice exercises and aiding in professional work.

## 5.5 Procedure and Data Collection

Participants signed up for the study by booking a time slot based on their availability. Depending on whether they volunteered for just the main interaction or both the interaction and the follow-up interview, they were allocated either a 40-minute or a 60-minute session. Upon enrollment, participants were automatically assigned specific details using a custom script function. That function performed a balanced randomization to ensure an equal distribution of our experimental conditions, a task for the concept explanation task, and task orders across the entire sample.

Upon joining the Microsoft Teams meeting, they first completed a consent form. They then completed a demographics survey, which included questions about their previous experience with LLMs in an educational context. A researcher then introduced them to the Deep3 tool and provided a short walkthrough of the interface, creating a low-pressure environment for asking questions and getting familiar with the interface.

Participants then proceeded with the experimental tasks, interacting with one of the four agent conditions. Upon completing the task, they filled two questionnaires: (1) understanding assessment questionnaire, which consisted of a perceived understanding question, using a 5-point Likert scale, and an actual understanding open-ended question which was later graded by a researcher using a predetermined marking scheme [47]; and (2) the NASA-TLX questionnaire [29]. The perceived and actual understanding questions, as well as the predetermined marking scheme are in Appendix III.1.

The same structure was followed for Task 2. Throughout the entire session, participants shared their screens via Microsoft Teams. This allowed the researcher to monitor the interaction and provide technical support if any issues arose with the interface. After completing both tasks, participants filled out two final questionnaires, to evaluate their overall impression of the system: (1) the Trust Questionnaire (Appendix D from [30]); and (2) the short version of the User Experience Questionnaire (UEQ-S) [39].

The subset of participants who agreed to the follow-up phase then took part in a 10–20 minute semi-structured interview. These interviews focused on their overall impressions of the Deep3 application and their specific thoughts on the agent condition they used. Additionally, participants were asked for suggestions on how to improve the tool and whether they believed it would be a helpful resource for guidance in a school or university setting. The interview questions are provided in Appendix III.2.

## 5.6 Data Analysis

### 5.6.1 Quantitative Analysis

Since the study evaluates user engagement and learning efficiency within the application, performance and user experience were measured across several objective and subjective metrics. We categorized these metrics into general behavioral logs and condition-specific measures adapted to the unique interventions of each study group. Table 5.1 provides a comprehensive definition of these variables.

To explore the relationship between user behavior and learning outcomes, we performed a Pearson correlation analysis. Specifically, we examined pairwise correlations between behavioral measures (e.g., time on task) and subjective outcomes (e.g., questionnaire scores), as well as internal correlations among the behavioral measures themselves to identify patterns in user interaction.

For each metric, we assessed normality using Shapiro–Wilk tests across all experimental conditions. When normality assumptions were satisfied, we conducted One-Way ANOVA to evaluate

Table 5.1: Description of behavioral metrics and subjective outcome variables.

| Variable | Description |
|---|---|
| *General Behavioral Metrics* | |
| Total Prompts | Total number of prompts generated by the user to complete the task |
| Task Completion Time | Total time (minutes) taken to finish the task |
| Average Response Time | Mean duration (seconds) for a user to respond to system prompts |
| Summaries Count | The count of summaries accessed |
| Quiz Count | The number of quizzes generated |
| *Learning & Subjective Outcomes* | |
| Actual Understanding | Mean scores for the user's actual understanding |
| Subjective Measures | Mean item scores for each of the four post-experiment questionnaires |
| *Condition-Specific Measures* | |
| Self-explanations | Number of self-explanations made by the user (Cond A) |
| Counterargument engagement | Number of counterpoint quizzes taken and the ratio of questions answered (Cond B) |
| Hints and Solutions | Total number of hints requested and full solutions accessed (Cond C) |

global differences across conditions; otherwise, we applied the non-parametric Kruskal–Wallis H-test. To compare specific conditions, we performed planned pairwise comparisons (Independent T-Tests). For behavioral metrics, we compared each experimental condition against the Baseline condition. For feature-specific metrics (summaries used and quizzes generated), we conducted pairwise comparisons exclusively among the experimental conditions. All pairwise comparisons were Bonferroni-corrected to account for multiple testing. We report p-values for significance and Cohen's d [16] for effect sizes, interpreted as small ($> 0.2$), medium ($> 0.5$), and large ($> 0.8$). Additionally, we report eta-squared ($\eta^2$) for ANOVA tests to quantify the proportion of variance explained by condition differences.

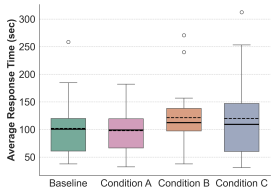
Additionally, to account for potential confounding variables and to test whether condition effects persisted after controlling for those factors, we performed analyses of covariance (ANCOVA). For each metric where appropriate covariates were available, we entered the experimental condition as the fixed factor and included relevant covariates in the model. Prior to running ANCOVA we verified the homogeneity of regression slopes between covariates and the factor, and checked the usual ANCOVA assumptions. When the ANCOVA indicated a significant main effect of condition, we examined adjusted means and conducted planned pairwise comparisons on those estimated marginal means with Bonferroni correction to control familywise error. For ANCOVA results we report F and p-values and use partial eta-squared (partial $\eta^2$) to indicate effect size.

### 5.6.2 Qualitative Analysis

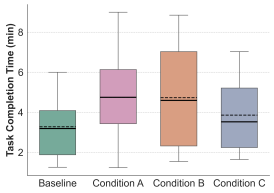
All interviews were audio-recorded and then transcribed with rev.com. Transcripts were then reviewed to ensure there were no inconsistencies. The cleared interviews were qualitatively analyzed using thematic analysis [8]. We employed an iterative open-ended coding process,

identifying data patterns related to our research questions. First, we read the interview transcripts to re-familiarize ourselves with the data. Then, we agreed on the key points for analysis. Following this, we began the coding using the tool miro.com. Code titles included: *impression of the tool, cognitive depth, perceived pedagogical utility, perceived knowledge retention and active recall interventions*. Over the course of the analysis stage, we had in mind the codes, and gradually determined which were most useful for organizing themes, ensuring they accurately captured the participants' experiences. Through this iterative process, we identified three themes in relation to our research questions.

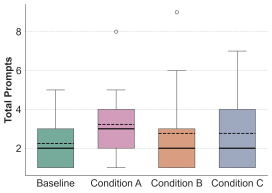
# 6 Results



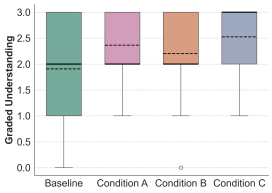

(a) Average Response Time (sec)




(b) Task Completion Time (min)

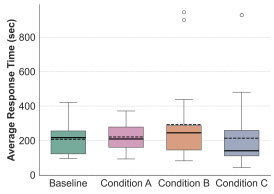



(c) Total Prompts

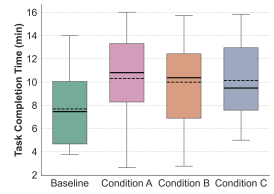



(d) Task Performance

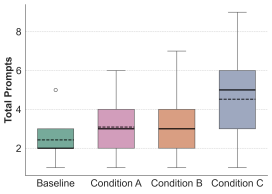
Figure 6.1: Task 1: Concept Explanation. Performance across four interaction modes (Baseline, Cond A: Future-Self Explanations, Cond B: Contrastive Learning, Cond C: Guided Hints). Box plots show per-condition distributions across all participants; solid lines denote medians, dashed lines denote means. Note: Task 1 task performance scores range from 0-3.



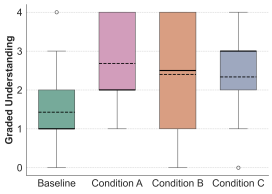

(a) Average Response Time (sec)




(b) Task Completion Time (min)




(c) Total Prompts



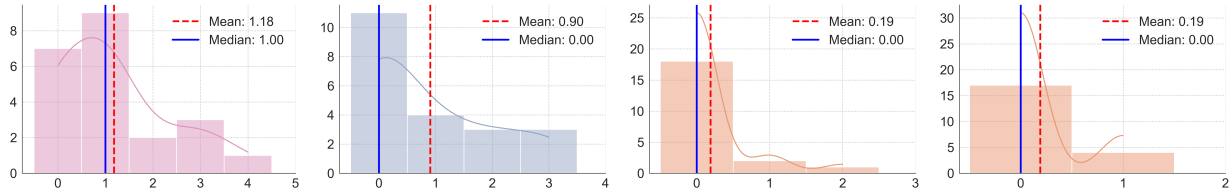

(d) Task Performance

Figure 6.2: Task 2: Problem Solving. Performance across four interaction modes (Baseline, Cond A: Future-Self Explanations, Cond B: Contrastive Learning, Cond C: Guided Hints). Box plots show per-condition distributions across all participants; solid lines denote medians, dashed lines denote means. Note: Task 2 task performance scores range from 0-4.

Next, we present the quantitative and qualitative results from our user study. Full statistical details for all analyses are reported in Table 6.1 for performance metrics and Table 6.2 for subjective measures.

## 6.1 Quantitative Results

The distinction between the two experimental tasks is crucial for interpreting our findings. Our data reveals that Task 2 (Problem Solving) acted as a "stress test" that enhanced the differences between conditions ( fig. 6.2). This task was significantly more demanding across all metrics, showing strong correlations between the overall workload and both total time ($r = 0.67$) and mental demand ($r = 0.49$). While NASA-TLX workload scores for Task 1 (Concept Explana-



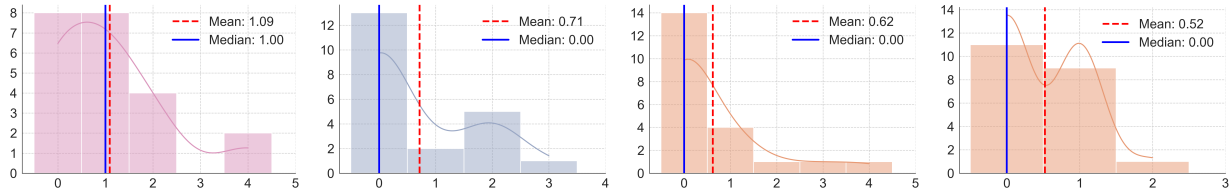

(a) Cond A: Self-Explanations Count

(b) Cond B: Contrastive Questions Answered

(c) Cond C: Hints Used

(d) Cond C: Solutions Used

Figure 6.3: Task 1: Concept Explanation. Condition specific feature usage across three interaction modes (Cond A: Future-Self Explanations, Cond B: Contrastive Learning, Cond C: Guided Hints). Histograms illustrate per-condition distributions across all participants; solid lines denote medians, dashed lines denote means, and the curve represents the kernel density estimation.




(a) Cond A: Self-Explanations Count

(b) Cond B: Contrastive Questions Answered

(c) Cond C: Hints Used

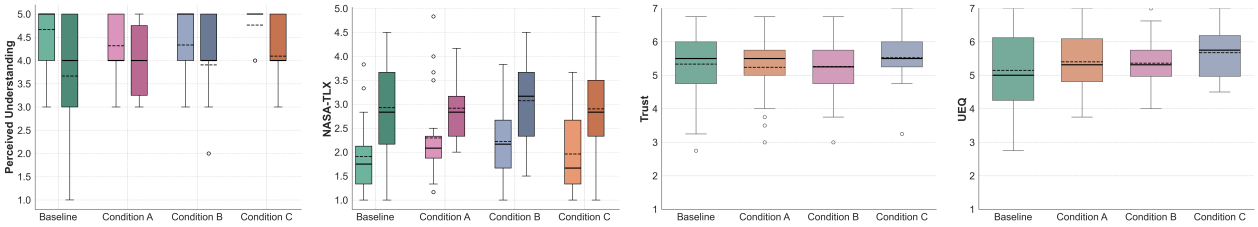
(d) Cond C: Solutions Used

Figure 6.4: Task 2: Problem Solving. Condition specific feature usage across three interaction modes (Cond A: Future-Self Explanations, Cond B: Contrastive Learning, Cond C: Guided Hints). Histograms illustrate per-condition distributions across all participants; solid lines denote medians, dashed lines denote means, and the curve represents the kernel density estimation.

tion) were relatively low (1.93 to 2.38), Task 2 consistently triggered much higher scores (2.90 to 3.02).

The Baseline condition, which serves as our comparison point, clearly illustrates the difficulty of learning without AI support. This group exhibited a significant overconfidence effect. Participants reported high perceived understanding ($M = 4.65$ in Task 1) but achieved the lowest actual grades in the study ($M = 1.90$ for Task 1; $M = 1.40$ for Task 2)( figs. 6.1 and 6.2). This created a "failure curve" where, as the tasks became harder, the students' experience collapsed. While they stayed calm in Task 1, Task 2 showed an extremely strong correlation between the time they spent and the frustration they felt ($r = 0.80$)( fig. 6.7).

In contrast to the overconfident Baseline group, Condition A (Future-Self Explanations) acted as a model of efficiency for theoretical learning, with participants completing Task 1 significantly faster than the baseline ($d = 0.85, p = 0.030$) table 6.1. Even though the process was fast, it came with a "high-friction" experience. Users reported higher mental workloads and a strong




(a) Perceived Task Understanding (b) NASA-TLX (c) Trust (d) User Experience

Figure 6.5: Boxplots of per-participant average questionnaire scores for the four interaction modes-Baseline, Cond A: Future-Self Explanations, Cond B: Contrastive Learning, Cond C: Guided Hints. Results are shown for both Task 1 (left) and Task 2 (right) for Perceived Task Understanding and NASA-TLX, and aggregated for Trust and User Experience. Solid lines denote medians; dashed lines denote means; whiskers indicate $1.5 \times$ IQR; points mark outliers.

link between their prompt usage and frustration ($r = 0.65$), leading to an "underconfidence effect". Another thing to note is that participants rated their own understanding at just $M = 4.32$ (on a 5-point scale) (Figure 6.5), which is the lowest self-rating among all experimental conditions, even though their actual test performance was quite high ($M = 2.36$ on a 3-point scale). Ultimately, Condition A led to the most "calibrated" understanding, meaning the students' perception of their knowledge closely matched their real performance. Finally, tool engagement remained stable, with participants viewing an average of $1.00 - 1.11$ self-explanations per task ( figs. 6.3 and 6.4).

While Condition A focused on efficiency, Condition B (Contrastive Learning) served as a middle ground that balanced high cognitive load with a positive user experience. This condition was a distinct outlier, since it had the highest mental demand in the study ($M = 4.85$), yet participants maintained a high trust ($M = 5.24$) and the lowest reported frustration ($M = 2.00$). This suggests that the contrastive quizzes turned hard work into "meaningful deliberation", where the effort was perceived as useful rather than exhausting. System logs support this through a pattern of selective engagement, showing that participants were presented with roughly 2.5 quizzes per task but chose to answer only about 0.8 ( figs. 6.3 and 6.4). Ultimately, Condition B proved that an AI system can demand high cognitive effort without causing user frustration, if the struggle feels purposeful.

Condition C (Guided Hints) emerged as a robust model, creating a seamless learning pathway that combined high engagement with significant improvements in academic results. In the second task (Problem Solving), this condition achieved a massive effect size for task performance($d = 1.14, p = 0.002$) and triggered the highest level of interaction, with a significant increase in total prompts ($d = 1.23, p = 0.002$). Remarkably, while increased interaction usually leads to higher stress [63], Condition C demonstrated a "frustration reversal" ($r = -0.32,$ fig. 6.7). This

Table 6.1: Statistical summary of performance measures and tool usage across tasks. The "Overall" row for each measure reports the results of a one-way ANOVA ($F$) or Kruskal-Wallis ($H$) test. Subsequent rows indicate post-hoc pairwise comparisons between experimental conditions and the Baseline. Effect sizes are reported as Cohen's $d$. Note. $\dagger p < .10, *p < .05, **p < .01, n.s.$ = not significant.

| Task | Measure | Effect | Statistics | Sig. |
|---|---|---|---|---|
| Task 1 | Total Prompts | Overall Effect | $H = 4.61, p = 0.203$ | $n.s.$ |
| | | A vs Baseline | $t = 2.28, p_{Bonf} = 0.084, d = 0.71$ | $\dagger$ |
| | | B vs Baseline | $t = 1.02, p_{Bonf} = 0.951, d = 0.32$ | $n.s.$ |
| | | C vs Baseline | $t = 1.00, p_{Bonf} = 0.977, d = 0.32$ | $n.s.$ |
| | Completion Time (min) | Overall Effect | $H = 7.14, p = 0.068$ | $\dagger$ |
| | | A vs Baseline | $t = 2.72, p_{Bonf} = 0.030, d = 0.85$ | $*$ |
| | | B vs Baseline | $t = 2.38, p_{Bonf} = 0.071, d = 0.75$ | $\dagger$ |
| | | C vs Baseline | $t = 1.22, p_{Bonf} = 0.694, d = 0.38$ | $n.s.$ |
| | Average Response Time | Overall Effect | $H = 2.50, p = 0.475$ | $n.s.$ |
| | Summaries Used | Overall Effect | $H = 10.97, p = 0.012$ | $*$ |
| | Quizzes Generated | Overall Effect | $H = 5.84, p = 0.120$ | $n.s.$ |
| | Graded Understanding | Overall Effect | $H = 5.22, p = 0.156$ | $n.s.$ |
| | | A vs Baseline | $t = 2.35, p_{Bonf} = 0.073, d = 0.74$ | $\dagger$ |
| | | B vs Baseline | $t = 1.00, p_{Bonf} = 0.970, d = 0.32$ | $n.s.$ |
| | | C vs Baseline | $t = 1.77, p_{Bonf} = 0.254, d = 0.56$ | $n.s.$ |
| Task 2 | Total Prompts | Overall Effect | $H = 11.14, p = 0.011$ | $*$ |
| | | A vs Baseline | $t = 1.86, p_{Bonf} = 0.213, d = 0.58$ | $n.s.$ |
| | | B vs Baseline | $t = 1.37, p_{Bonf} = 0.543, d = 0.43$ | $n.s.$ |
| | | C vs Baseline | $t = 3.88, p_{Bonf} = 0.002, d = 1.23$ | $**$ |
| | Completion Time (min) | Overall Effect | $F = 2.49, p = 0.066, \eta_p^2 = 0.084$ | $\dagger$ |
| | | A vs Baseline | $t = 2.41, p_{Bonf} = 0.061, d = 0.75$ | $\dagger$ |
| | | B vs Baseline | $t = 2.11, p_{Bonf} = 0.124, d = 0.67$ | $n.s.$ |
| | | C vs Baseline | $t = 2.32, p_{Bonf} = 0.078, d = 0.73$ | $\dagger$ |
| | Average Response Time | Overall Effect | $H = 3.84, p = 0.279$ | $n.s.$ |
| | Summaries Used | Overall Effect | $H = 12.39, p = 0.006$ | $**$ |
| | Quizzes Generated | Overall Effect | $H = 5.86, p = 0.119$ | $n.s.$ |
| | Graded Understanding | Overall Effect | $H = 10.37, p = 0.016$ | $*$ |
| | | A vs Baseline | $t = 2.22, p_{Bonf} = 0.097, d = 0.70$ | $\dagger$ |
| | | B vs Baseline | $t = 2.31, p_{Bonf} = 0.080, d = 0.74$ | $\dagger$ |
| | | C vs Baseline | $t = 3.64, p_{Bonf} = 0.002, d = 1.14$ | $**$ |

means that as participants engaged more with the AI, their frustration actually decreased. This "Invisible Scaffolding" allowed students to achieve a 45% improvement in learning outcomes over the baseline without any increase in perceived workload ($p = 0.647$). In fact, Task 2 workload scores for Condition C ($2.90$) were the lowest in the study. Finally, utilization of specific "hint" and "solution" buttons remained low ($Mdn = 0.00$, figs. 6.3 and 6.4), suggesting that

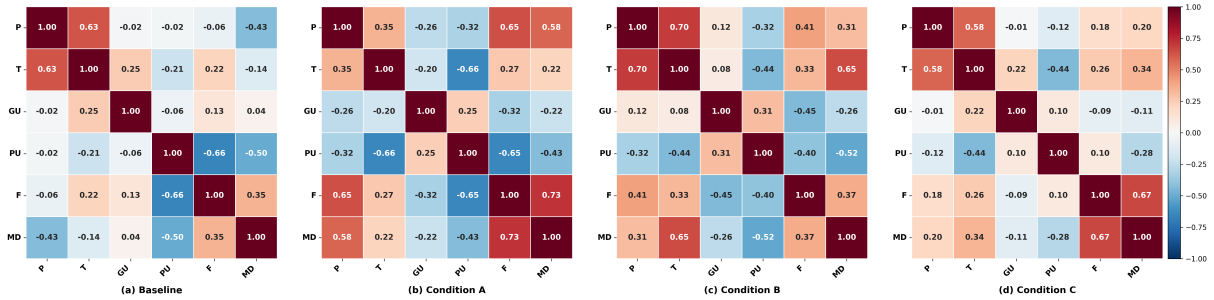



Figure 6.6: Cross-correlation matrices of four task-related metrics under each interaction condition in Task 1. The matrices show the Pearson correlation coefficients among Prompts (P), Time (T), Graded Understanding (GU), Perceived Understanding (PU), Frustration (F) and Mental Demand (MD) (ordered from top-left to bottom-right) for the four conditions: (a) Baseline, (b) Cond A: Future-Self Explanations, (c) Cond B: Contrastive Learning and (d) Cond C: Guided Hints. The color scale represents Pearson correlation coefficients ranging from $-1.0$ (blue, strong negative) to $+1.0$ (red, strong positive).

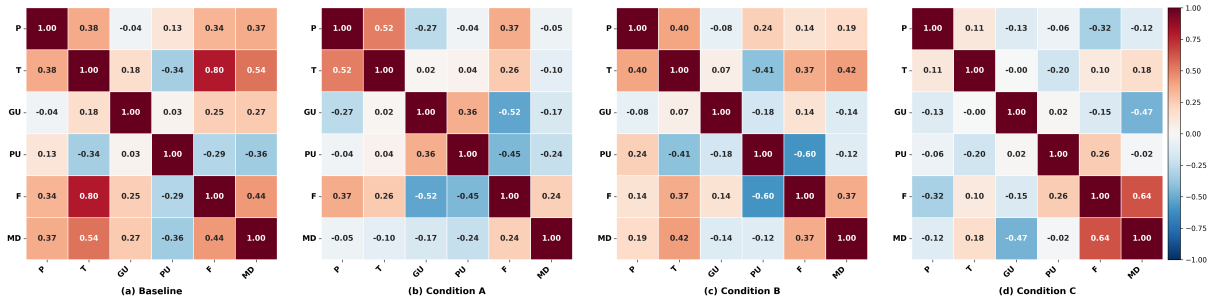



Figure 6.7: Cross-correlation matrices of four task-related metrics under each interaction condition in Task 2. The matrices show the Pearson correlation coefficients among Prompts (P), Time (T), Graded Understanding (GU), Perceived Understanding (PU), Frustration (F) and Mental Demand (MD) (ordered from top-left to bottom-right) for the four conditions: (a) Baseline, (b) Cond A: Future-Self Explanations, (c) Cond B: Contrastive Learning and (d) Cond C: Guided Hints. The color scale represents Pearson correlation coefficients ranging from $-1.0$ (blue, strong negative) to $+1.0$ (red, strong positive).

conversational support was sufficient to guide users without extra mechanisms.

To conclude our quantitative analysis, we examined the engagement patterns with features shared across all three experimental conditions: summaries and quizzes. Our analysis revealed a significant global effect for summary usage across both tasks (Task 1: $p = 0.012$; Task 2: $p = 0.006$). Interestingly, there were no significant differences in usage between Conditions A, B, or C (all $p_{Bonf} = 1.000$). Similarly, quiz generation remained consistent across all groups and tasks ($p = 0.119$). These findings suggest that while these tools were popular, the specific design of the AI agent did not change how often users clicked these buttons.

Table 6.2: Statistical summary of questionnaire measures across tasks. The "Overall" rows report the results of a Mixed ANOVA ($F$) or Kruskal-Wallis ($H$) test. Subsequent rows indicate post-hoc pairwise comparisons between experimental conditions and the Baseline. Effect sizes are reported as Cohen's $d$ or partial eta-squared ($\eta_p^2$). Note. $\dagger p < .10, *p < .05, **p < .01, ***p < .001, n.s.$ = not significant.

| Measure | Task | Effect | Statistics | Sig. |
|---|---|---|---|---|
| Understanding | | Condition | $F = 4.87, p = 0.004, \eta_p^2 = 0.153$ | $**$ |
| | | Task | $F = 0.05, p = 0.819, \eta_p^2 = 0.001$ | $n.s.$ |
| | | Interaction | $F = 1.40, p = 0.250, \eta_p^2 = 0.049$ | $n.s.$ |
| | Task 1 | A vs Baseline | $t = 2.35, p_{Bonf} = 0.073, d = 0.74$ | $\dagger$ |
| | | B vs Baseline | $t = 1.14, p_{Bonf} = 0.778, d = 0.36$ | $n.s.$ |
| | | C vs Baseline | $t = 1.77, p_{Bonf} = 0.254, d = 0.56$ | $n.s.$ |
| | Task 2 | A vs Baseline | $t = 2.22, p_{Bonf} = 0.097, d = 0.70$ | $\dagger$ |
| | | B vs Baseline | $t = 2.44, p_{Bonf} = 0.057, d = 0.77$ | $\dagger$ |
| | | C vs Baseline | $t = 3.64, p_{Bonf} = 0.002, d = 1.14$ | $**$ |
| NASA-TLX | | Condition | $F = 0.55, p = 0.647, \eta_p^2 = 0.020$ | $n.s.$ |
| | | Task | $F = 63.58, p = 0.000, \eta_p^2 = 0.443$ | $***$ |
| | | Interaction | $F = 0.71, p = 0.551, \eta_p^2 = 0.026$ | $n.s.$ |
| Trust | | Overall Effect | $H = 0.64, p = 0.886$ | $n.s.$ |
| UEQ | | Overall Effect | $F = 1.08, p = 0.362, \eta_p^2 = 0.039$ | $n.s.$ |

### 6.1.1 Controlling for Participant Characteristics

To ensure that the observed effects were driven by the experimental conditions rather than individual differences, we conducted a series of Analysis of Covariance (ANCOVA) and controlled regression analyses. We included gender, academic discipline, age, current GPA, prior LLM usage, and LLM usage in educational contexts as covariates.

The analysis confirms that our primary behavioral and subjective results remain robust. Specifically, the effect of condition on Total Prompts in Task 2 remained highly significant after adjusting for all covariates ($p = .0023$), with Condition C maintaining a strong positive effect over the baseline ($\beta_{adj}, p < .001$). Similarly, Task 1 Completion Time remained significant ($p = .0382$), with both Condition A and Condition B showing significantly higher time investment than the baseline even when accounting for prior AI experience.

Interestingly, adding controls sharpened the results for Task Understanding in Task 1, which became more statistically significant ($p = .0099$) than in the uncontrolled model. Adjusted pairwise comparisons revealed that Condition A and Condition C both significantly outperformed the Baseline in objective understanding gains, with Condition A emerging as the most effective strategy for objective learning after covariate adjustment.

Table 6.3: Robustness check for condition effects after controlling for participant characteristics. The rows report the adjusted effect of condition from ANCOVA models including gender, discipline, age, GPA, prior LLM usage, and educational LLM usage as covariates. Effect sizes are reported as partial eta-squared ($\eta_p^2$). Note. $\dagger p < .10, *p < .05, ** p < .01, ***p < .001, n.s.$ = not significant.

| Outcome | ANOVA $p$ | ANCOVA $p$ | $\eta_p^2$ (adj.) | Interpretation |
|---|---|---|---|---|
| Task 2 Total Prompts (Cond C) | .0009 | .0023 | .181 | Robust, remains highly significant |
| Task 1 Completion Time (Cond A & B) | .0398 | .0382 | .110 | Robust, remains significant |
| Task 1 Task Understanding (Cond A & C) | .0554 | .0099 | .145 | Strengthened after controls |
| Task 1 Summaries Used | .0332 | .0532 | .100 | Attenuated to marginal |
| Task 2 Completion Time | .0660 | .1255 | .076 | Becomes non-significant |
| Task 2 Summaries Used | .0533 | .1538 | .070 | Becomes non-significant |

However, some secondary effects attenuated when controlling for participant characteristics. The usage of Summaries in Task 1 shifted from significant to marginal ($p = .0532$), and Task 2 Completion Time became non-significant ($p = .1255$). These shifts were largely explained by two significant covariates. Firstly, LLM usage in education was a significant predictor of summary usage across both tasks, suggesting that students already used to AI in their studies are more likely to utilize these specific features regardless of the agent's design. Second, Age was a significant covariate for Average Response Time in Task 1, indicating that demographic factors influenced the pace of interaction more than the interaction mode itself in the initial task.

## 6.2 Qualitative Results

By thematically analyzing our participants' interviews, we identified three main themes: *Cognitive Depth via Active Recall*, *Perceived Knowledge Retention*, and *Pedagogical Utility*. Each theme integrates qualitative insights from participant feedback with quantitative results where applicable, allowing us to see how the specific design and behavior of each AI agent shaped the way students interacted with the system and how they envisioned using these tools in their own studies.

**Cognitive Depth via Active Recall.** The qualitative data confirms that while the Baseline was prefered for its low cognitive effort, it fostered a "Fragile Understanding" that collapsed under the stress of Task 2. Participants like P56 admitted the Baseline *"simplifies things so it goes faster... [you] don't use that much brain cells"*, directly explaining the high overconfidence and low task performance ($M = 1.40$) seen in our quantitative analysis. In contrast, the experimental conditions, particularly Condition A, transformed the AI into a self-diagnostic tool. P30 noted, *"I had to check just to be sure if I understood it right... I scrolled a bit up to see the actual answer"*, illustrating the "productive friction" that drove higher response times and mental demand. This shift toward "desirable difficulty" was also seen in Condition B, where

P13 found that the AI *"made me think more deeply... about points of view that maybe you hadn't considered"*, justifying the high mental demand ($M = 4.85$) as a way to think deeply instead of just feeling frustrated. By moving away from the Baseline's *"too many details"* (P14), the AI's ability to work with the user's existing logic allowed for the high-performing "Invisible Scaffolding" observed in Condition C. As P23 noted, *"bridging that gap instead of just retelling me the definition"*. Ultimately, these features acted as vital *"self-evaluation checkpoints"* (P8) that aligned the students' understanding and ensured they were engaged through active thinking rather than passive consumption.

**Perceived Knowledge Retention.** The qualitative feedback highlights a divide between the "Fragile Understanding" of the Baseline and the thinking patterns frameworks fostered by the AI interventions. Participants in the Baseline group admitted a "shallow" experience, for example when asked if the tool encouraged deeper thought, P19 briefly replied, *"No"*. This explains the low NASA-TLX workload in Task 1 (Figure 6.5) but also the subsequent collapse in task performance during Task 2. When faced with complex problems, Baseline users like P14 found the AI's detailed answers "annoying" and felt they were simply "too many details" to process, leading to a failure in fully understanding the idea. In contrast, participants in the experimental conditions linked the effortful interaction to long-term memory. P30 (Cond A) explicitly stated, *"You remember what you understand. So if I have to explain it, I first must understand"*, suggesting that slow learning and high mental effort were necessary for long-term memory. This was further refined in Condition C, which users perceived as a *"step-by-step tutorial on how to evolve your thinking process"* (P23). P32 noted that this scaffolded support became a repeatable mental framework, predicting they could *"think back to your process with the bot"* during an exam to apply the same logic. This qualitative evidence clarifies why Condition C achieved a massive effect size in task performance ($d = 1.14$), it didn't just provide answers, it helped users internalize the logic required to find them. While some users found the contrastive approach in Condition B "complicated" for strict definitions (P50), the shared view across experimental groups was that features like quizzes served as necessary "self-evaluation" checkpoints (P8, P23) to help their progress.

**Pedagogical Utility.** The qualitative feedback underscores a fundamental shift in the AI's role from a simple information provider to a sophisticated pedagogical agent. While the Baseline was initially seen as efficient, it ultimately failed as a teaching tool. Participants like P14 and P19 noted it never prompted deeper thought, leading to a "slower", more frustrating experience when tasks became complex. In contrast, the experimental conditions were perceived as high-level learning strategies. Condition B, in particular, was recognized for its academic rigor, with P17 noting the system *"had a way of teaching like professors"*, which directly explains the high trust ratings ($M = 5.24$). Furthermore, participants noted that while Condition B is ideal for

theoretical fields, it may require more *"strict definitions"* (P50) for applied sciences. Condition C achieved a balance of "Invisible Scaffolding", as P32 expressed a desire to use the tool in a university setting because *"it helps me think more rather than just giving me an answer"*. This sense of partnership fostered the "frustration reversal" ($r = -0.32$) seen in the logs, as users felt the system was *"bridging the gap"* (P23) rather than just lecturing. Condition A functioned as a mechanism for metacognitive verification and gap identification through its "explain-back" prompts. This utility was validated by P30, who noted it automated a strategy they already use in university: *"I ask something and then... I always say I understood this. Is it right?"* Noting the need for a tool that automates this process.

# 7 Discussion

Next, we situate our findings within prior work on GenAI in education, productive friction and active learning. We first highlight our main findings, drawing from both quantitative and qualitative data. We then discuss the implications of our results, outline contributions to HCI theory and design practice, as well as its ethical considerations before acknowledging the limitations of our study and identifying directions for future research.

It is important to consider the volume of information provided across conditions. While the baseline condition provided a single, comprehensive answer, consistent with standard LLM behaviors like ChatGPT, the experimental conditions were intentionally designed to be more concise, typically limiting output to 1–2 paragraphs. Additional features, such as guided hints, were restricted to 1–2 sentences to avoid overwhelming the user. Consequently, the experimental conditions provided less immediate information than the baseline. Also, the Presentation themes (Section 4.3), were implemented only to the experimental conditions, as the Baseline was designed to mirror standard current practices without additional interventions, providing a representative control for existing LLM interaction models.

## 7.1 Main Findings

Our results show that the way AI scaffolding is designed changes how LLMs are used in education (i.e., turning them from simple information sources into tools for thoughtful learning). By using two different tasks, we showed how different teaching methods balance effort and understanding. Our findings reveal four key insights: 1) Condition A provides a metacognitive reference point, aligning task understanding with task performance; 2) Condition B creates a "desirable difficulty", turning high mental effort into meaningful thinking without causing extra frustration; 3) Condition C provides subtle support, helping users absorb logical frameworks and reach top problem-solving performance without asking for hints; and 4) there is no "one-size-fits-all" AI tutor, rather, the optimal intervention depends on the learner's priority.

*Condition A provides a metacognitive reference point, aligning task understanding with task performance.* By requiring users to explain concepts back to the AI, this condition acted as a metacognitive anchor, forcing participants to confront gaps in their own logic [14]. While the baseline group reported high task understanding ($M = 4.65$) despite having the lowest objective scores ($M = 1.40$), Condition A reversed this trend. Our data highlights the fact that participants achieved the study's highest accuracy ($M = 2.36/3$) while rating their own comprehension lower than all other groups ($M = 4.32/5$). This suggests that the "Correction

Table 7.1: Summary of qualitative and quantitative findings across interaction modes.

| **Dimension** | **Baseline** | **Condition A** | **Condition B** | **Condition C** |
|---|---|---|---|---|
| Performance | Lowest grades ($M = 1.40$ in T2). | **Strongest gains**; highest accuracy ($M = 2.36$) after covariate adjustment. | Sig. improvement over baseline (+1.00 pt). | Significant improvement ($d = 1.14^{**}$); high efficacy in problem-solving. |
| Metacognition | **Overconfidence**: High perceived vs. low actual knowledge. | **Metacognitive Anchor**: Lowest self-rating ($M = 4.32$) despite high test scores. | Balanced; perceived as "Meaningful Deliberation". | High calibration; logic internalized as mental framework. |
| Cognitive Load | Low initial effort; "Fragile Understanding" in T2. | High mental demand; perceived as "High-Friction". | Descriptive trend toward higher demand; effort separated from frustration. | Descriptively lowest T2 workload (2.90); "Invisible Scaffolding". |
| Emotional Response | Time-Frustration correlation ($r = 0.80$); "Failure Curve". | Prompt usage linked to frustration ($r = 0.65$). | **Lowest frustration** ($M = 2.00$); high trust ($M = 5.24$). | **Frustration Reversal**: Interaction reduced stress ($r = -0.32$). |
| Tool Engagement | Passive; AI answers labeled "too many details". | Stable feature usage; $\approx 1.1$ explanations/task. | Selective engagement; quizzes answered only when relevant. | Highest interaction; sig. increase in prompts ($d = 1.23^{**}$). |
| UI Interaction | n/a | Primary use of "Explain Back" prompts. | Quizzes used as "Self-evaluation checkpoints". | **External buttons ignored**; conversational foundation was enough. |

Loop" works as a self-diagnostic tool, as P30 noted, they were forced to *"check just to be sure"* if they understood. This early challenge, shown by the Task 1 frustration correlation ($r = 0.65$), ultimately led to the most calibrated relationship between perceived and actual knowledge.

*Condition B creates a "desirable difficulty", turning high mental effort into meaningful thinking without causing extra frustration.* While descriptive trends suggested higher mental demand for this group ($M = 4.85$), the design effectively separated perceived effort from frustration. While Task 2 was a "stress test" for other conditions, Condition B maintained strong trust ratings ($M = 5.24$) and the study's lowest reported frustration ($M = 2.00$). The qualitative feedback suggests this was achieved through learner autonomy. Users were selective in their engagement with the counterarguments, answering only contrastive quizzes they deemed relevant [9]. This allowed them to treat the AI like a *"professor"* (P17) [18], engaging in "meaningful deliberation". By offering contrastive foils, the system prompted users to think *"more deeply"* (P13) about alternative points of view, resulting in a significant 1.00-point improvement in objective scores over the baseline without a significant increase in frustration.

*Condition C provides subtle support, helping users absorb logical frameworks and reach top problem-solving performance without asking for hints.* This condition achieved the most robust behavioral result in the study, with a highly significant increase in total prompts ($p = .0023$) that persisted even after controlling for prior AI experience. The effectiveness of this intervention is best shown by the "frustration reversal" ($r = -0.32$). While the baseline showed a $r = 0.80$ correlation between time and frustration, Condition C proved that more interaction with it actually reduced user stress. Remarkably, participants achieved these results while rarely using the "hint" or "solution" buttons (fig. 6.4). This indicates that the agent's guidance was so effective that users learned the AI's logical approach, evolving their thinking process to a *"step-by-step"* (P23) approach rather than relying on external cheat mechanisms [59].

*Ultimately, these results suggest that there is no "one-size-fits-all" AI tutor, rather, the optimal intervention depends on the learner's priority.* For learners seeking conceptual depth, Condition A is most effective. For those seeking meaningful deliberation and theoretical exploration, the contrastive challenges of Condition B provide a balanced academic rigor. For students requiring applied mastery and high engagement during complex problem-solving, the scaffolding of Condition C is highly effective, while Condition A provides the strongest overall objective learning gains. These findings suggest that future LLM-based tutors must move beyond being simple "information providers" [18] to become adaptive systems [42] capable of shifting their intervention style based on the task's cognitive demand and the learner's specific pedagogical goals.

## 7.2 Implications

Our findings demonstrate that cognitive load and user frustration are not directly linked. This challenges the traditional UX dogma that friction necessarily leads to a poor user experience [21, 29]. Although workload differences were not statistically reliable in our controlled analyses, the qualitative data and descriptive trends suggest that AI can facilitate Desirable Difficulty without triggering the emotional failure curve ($r = 0.80$) seen in the Baseline. Specifically, our "frustration reversal" finding suggests that when AI provides scaffolding rather than direct answers, the mental effort feels like meaningful exploration, not just an obstacle [53]. This extends Cognitive Load Theory by showing that AI can reduce emotional strain, helping students stay in "Flow" even when tasks get harder.

A major practical takeaway for the design of Intelligent Tutoring Systems is the under utilization of explicit help mechanisms when compared to integrated dialogue. While our experimental interface provided participants in Condition C with dedicated "Hint" and "Solution" buttons, these features were ignored ($Mdn = 0.00$, fig. 6.4) . This suggests that when an AI provides strong conversational foundation, users do not perceive a need for extra UI interventions. This finding advances the work of systems like Iris Tutor [5] and TutorLLM [42] by demonstrating that effective scaffolding does not require separate UI triggers. By "formatting" the user's existing logic rather than simply "rephrasing" it (P23), the AI bridges the gap between the student's current mental model and the target solution. For designers, the implication is clear: instead of building explicit "help" buttons, LLMs should be prompted to reflect the user's own reasoning back to them. This "Invisible Scaffolding" allows users to internalize the logic as a repeatable mental framework, supporting the long-term skill gain highlighted by Rogers [58]. Furthermore, our robustness check revealed that prior LLM usage in education was a significant predictor of summary usage. This implies that while integrated dialogue (as seen in Condition C) is powerful, existing student habits, like a reliance on automated summaries, remain a strong factor in how they engage with AI, regardless of the specific agent's design.

Our results for Condition A offer a guide for closing the metacognitive gap in Gen Z's AI usage [51]. While Lee et al. [41] noted that high confidence in AI often leads to less critical thinking, Condition A acted as a "Metacognitive Anchor", forcing users to confront what they did not know. The resulting "Underconfidence Effect" is a vital insight for the HCI community. It suggests that the best learning tools might lower short-term satisfaction by challenging users' overconfidence in their understanding [3]. Using AI in education requires, therefore, balancing accurate self-assessment with positive encouragement. This ensures that the initial cognitive friction observed in our correlation data ($r = 0.65$) does not lead to premature tool abandonment.

Finally, our work presents a practical taxonomy of AI interventions, showing that the "best" agent depends on context. Condition A is ideal for Conceptual Grounding, using active recall to confirm understanding before students move to application. Condition B supports Meaningful Deliberation, especially in theoretical tasks that challenge students to consider multiple perspectives. Condition C works best for Problem-Solving exercises, helping students achieve applied mastery in domains where logical reasoning is crucial. This functional split suggests that future educational platforms should move toward adaptive, multi-agent ensembles, similar to the architecture in EduThink4AI [32], that shift their persona.

## 7.3 Ethical Considerations and Design Trade-offs

The findings of this study raise ethical and design trade-offs regarding the balance between user autonomy and pedagogical enforcement. A primary tension emerged in Condition A, where the *"Explain what you understood"* prompt was occasionally misinterpreted as part of the AI's output rather than a mandatory user action. As noted by P8 and P24, the lack of explicit UI cues, such as a dedicated Explain button, led to confusion, leaving users unsure if participation was required. This points to a key design trade-off between maintaining smooth conversational flow in LLM interfaces and enforcing stricter UI rules to introduce intentional challenges. Designers must decide whether to prioritize a seamless experience or to implement mandatory steps that ensure participants apply the required mental effort for deep learning.

Similarly, Condition B revealed an ethical challenge regarding selective engagement and automation bias. While the ability to ignore contrastive foils and contrastive questions provided users with a sense of autonomy and reduced frustration, it also allowed them to bypass the very meaningful deliberation the system was designed to provoke. This mirrors broader concerns in HCI regarding automation bias, where users may prefer the most efficient path over the most demanding one. If users are permitted to skip cognitive challenges, the system risks regressing into a "Baseline" one. Designers must decide when to let users manage their own cognitive load and when to require engagement with counter-arguments to help them fully understand concepts.

Our results also suggest that the effectiveness of reflective AI depends on the subject area, raising questions about the universal applicability of these models. Participants like P50 expressed doubt regarding the utility of contrastive learning for applied sciences such as mathematics or physics, where strict definitions and singular logical paths are prioritized over the multi-perspective debate common in classical studies or literature. There is a risk that deploying a contrastive agent in a domain requiring high-precision problem-solving could introduce unnecessary cognitive interference that confuse the learner. This suggests an ethical responsibility

to ensure that AI tutoring architectures are tailored to the specific requirements of the subject matter.

The Underconfidence Effect observed in Condition A introduces a psychological trade-off for long-term tool adoption. While the intervention successfully destroyed the "Illusion of Competence" and led to high objective performance, it also left users feeling less capable. In a commercial or optional educational setting, a tool that makes a user feel "slower" or "less smart" faces a significant risk of abandonment. Designers must therefore consider the ethical implications of corrective AI: how do we design systems that are honest about a student's knowledge gaps without damaging their self-confidence? Balancing the "Correction Loop" with positive feedback or progress tracking may be key to preventing the early demands of deep learning from discouraging learners.

Finally, a key challenge raised is the "adoption barrier" that is part of introducing specialized educational tools in an ecosystem dominated by general-purpose LLMs like ChatGPT. Even though students who tried the tool liked it, many are unlikely to switch from the tools they already use. Instead of trying to replace those tools, developers should try to improve them. The goal should be to quietly add our features into the tools students already use, so it feels natural and easy to adopt.

## 7.4 Limitations and Future Work

In Phase 1, we acknowledge two primary limitations. First, participants were recruited through convenience sampling, which may limit the generalization of our findings to the broader student population. Second, our sample exhibited a high degree of AI literacy and prior exposure. Nearly all participants reported regular experience with Generative AI tools, with none identifying as non-users or AI-skeptics. Future research should specifically target AI-naive users or those with a skeptical stance toward automated tutoring to understand broader needs and adoption barriers.

In Phase 2, we acknowledge six limitations. First, our recruitment strategy was relied on personal networks and institutional advertisements, which may have introduced social desirability bias, where participants who know the researchers may perform differently or rate the system more favorably. To address these concerns, future studies should utilize independent facilitators to ensure that participant feedback remains objective and uninfluenced by personal connections to the research team.

Second, our participant pool ($N = 85$) was primarily comprised of undergraduate students from STEM disciplines, particularly Computer Science ($N = 24$). Future work should aim for a more diverse sample to examine how these AI interventions perform across different academic fields

and levels of expertise.

A third limitation is about collecting demographic data at the start of the study may have accidentally caused participants to feel a stereotype threat, potentially biasing performance or self-reporting (NCWIT [46]). Future work should relocate these questions to the end of the protocol to mitigate identity-based priming effects.

Fourth, we did not examine long-term engagement and knowledge retention. Although approximately 15% of participants reported a desire for more time to deeply engage with the agents, our measurements were limited to immediate performance gains. Consequently, we cannot determine whether the "Underconfidence Effect" in Condition A leads to tool abandonment over weeks of use, or whether the logic internalized in Condition C persists in the absence of the AI. Future work could investigate the integration of these three conditions into existing learning systems to evaluate their long-term efficacy.

Fifth, the study is also limited by its exclusive reliance on student self-reports. Literature suggests that students are often "unreliable narrators" of their own learning, frequently mistaking engagement for actual cognitive gain [7]. Consequently, future work should expand this inquiry to include educator perspectives, which remain highly varied regarding the role of AI in instruction [38].

Lastly, we utilized the llama-4-maverick-17b-128e-instruct model as the underlying LLM. While this model worked well for our study, future work should integrate newer "reasoning-focused" AI models because they might provide even better conversational support. Ultimately, the goal is to refine these pedagogical agents so they don't just provide answers, but actively foster the "meaningful deliberation" necessary for Gen Z students to succeed in an AI-integrated academic landscape.

# 8 Conclusion

In this study, we examined how diverse interaction architectures shape the educational utility and user experience of LLM-based tutors. Our findings show that how an AI intervention is designed directly shapes the key features of the learning process, proving that the best AI agent is strictly context-dependent. Condition A functioned as a vital metacognitive anchor, neutralizing the illusion of competence" observed in baseline interactions; after covariate adjustment, it yielded the study's strongest gains in objective understanding. In contrast, Condition C provided invisible scaffolding" that proved most robust for driving active engagement in problem-solving without the need for explicit support. Condition B demonstrated that contrastive foils can prompt meaningful deliberation and high trust without creating emotional burnout. Ultimately, these results demonstrate that while a Baseline interaction satisfies a user's Speed Addiction, it hides a lack of conceptual depth. By intentionally creating productive friction, we can move toward adaptive agents that empower genuine conceptual internalization. Together, these findings show how carefully designed interactions can reduce the risks of over-relying on generative AI and guide more effective human-AI collaboration in education.

# BIBLIOGRAPHY


[1] Louis Alfieri, Patricia J Brooks, Naomi J Aldrich, and Harriet R Tenenbaum. Does discovery-based instruction enhance learning? *Journal of educational psychology*, 103 (1):1, 2011. doi: https://doi.org/10.1037/a0021017.

[2] Mohammad T Alshammari and Amjad Qtaish. Effective adaptive e-learning systems according to learning style and knowledge level. *Journal of Information Technology Education: Research*, 18, 2019. doi: https://doi.org/10.28945/4459.

[3] Anthony Amoah, Rexford Kweku Asiama, and Edmund Kwablah. Chatgpt early usage among students: A global evidence of determinants. *Development and Sustainability in Economics and Finance*, page 100065, 2025. doi: https://doi.org/10.1016/j.dsef.2025.100065.

[4] Md Asaduzzaman Babu, Kazi Md Yusuf, Lima Nasrin Eni, Shekh Md Sahiduj Jaman, and Mst Rasna Sharmin. Chatgpt and generation 'z': A study on the usage rates of chatgpt. *Social Sciences & Humanities Open*, 10:101163, 2024. doi: https://doi.org/10.1016/j.ssaho.2024.101163.

[5] Patrick Bassner, Eduard Frankford, and Stephan Krusche. Iris: An ai-driven virtual tutor for computer science education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 394–400. 2024. doi: https://doi.org/10.1145/3649217.3653543.

[6] Andrea Blasco and Vicky Charisi. Ai chatbots in k-12 education: An experimental study of socratic vs. non-socratic approaches and the role of step-by-step reasoning. https://ssrn.com/abstract=5040921, dec 2024. Accessed: 2025-10-03.

[7] Nicholas A Bowman. Can 1st-year college students accurately report their learning and development? *American Educational Research Journal*, 47(2):466–496, 2010. doi: https://doi.org/10.3102/0002831209353595.

[8] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006. doi: https://doi.org/10.1191/1478088706qp063oa.

[9] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2025. doi: https://doi.org/10.1145/3706598.3713229.

[10] Andrew C Butler and Henry L Roediger. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, 36(3):604–616, 2008. doi: https://doi.org/10.3758/MC.36.3.604.

[11] Cecilia Ka Yuk Chan and Katherine KW Lee. The ai generation gap: Are gen z students more interested in adopting generative ai such as chatgpt in teaching and learning than their gen x and millennial generation teachers? *Smart learning environments*, 10(1):60, 2023. doi: https://doi.org/10.1186/s40561-023-00269-3.

[12] Gary Charness, Uri Gneezy, and Michael A Kuhn. Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, 81(1): 1–8, 2012. doi: https://doi.org/10.1016/j.jebo.2011.08.009.

[13] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

[14] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182, 1989. doi: https://doi.org/10.1016/0364-0213(89)90002-5.

[15] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. doi: https://doi.org/10.48550/arXiv.2403.04132.

[16] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013. doi: https://doi.org/10.4324/9780203771587.

[17] John W Creswell. Mixed-method research: Introduction and application. In *Handbook of educational policy*, pages 455–472. Elsevier, 1999. doi: https://doi.org/10.1016/B978-012174698-8/50045-X.

[18] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2023. doi: https://doi.org/10.1145/3544548.3580672.

[19] Ian Drosos, Advait Sarkar, Neil Toronto, et al. " it makes you think": Provocations help restore critical thinking to ai-assisted knowledge work. *arXiv preprint arXiv:2501.17247*, 2025. doi: https://doi.org/10.48550/arXiv.2501.17247.

[20] Sabit Ekin. Prompt engineering for chatgpt: a quick guide to techniques, tips, and best practices. *Authorea Preprints*, 2023. doi: https://doi.org/10.36227/techrxiv.22683919.v1.

[21] Paul Evans, Maarten Vansteenkiste, Philip Parker, Andrew Kingsford-Smith, and Sijing Zhou. Cognitive load theory and its relationships with motivation: A self-determination theory perspective. *Educational Psychology Review*, 36(1):7, 2024. doi: https://doi.org/10.1007/s10648-023-09841-2.

[22] Y. Fan. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 55(2):XXX–XXX, 2024. doi: https://doi.org/10.1111/bjet.13544. URL https://arxiv.org/abs/2412.09315.

[23] Lucile Favero, Daniel Frases, Juan Antonio Pérez-Ortiz, and Tanja Käser. Ellis alicante at cqs-gen 2025: Winning the critical thinking questions shared task: Llm-based question generation and selection. In *Proceedings of the 12th Argument mining Workshop*, pages 322–331, 2025. doi: https://doi.org/10.18653/v1/2025.argmining-1.31.

[24] Jennifer Finetti. How gen z uses ai: Scholarshipowl survey reveals how students are hacking their education and their future, June 23 2025. URL https://scholarshipowl.com/blog/gen-z-research/how-gen-z-uses-ai-scholarshipowl-survey-reveals-how-students-are-hacking-their-education-and-their-future/. Accessed: 2026-01-31.

[25] Josh Freeman. Student generative ai survey 2025. *Higher Education Policy Institute: London, UK*, 2025. URL https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/.

[26] Georgios P Georgiou. Chatgpt produces more" lazy" thinkers: Evidence of cognitive engagement decline. *arXiv preprint arXiv:2507.00181*, 2025. doi: https://doi.org/10.48550/arXiv.2507.00181.

[27] Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025. doi: https://doi.org/10.3390/soc15010006.

[28] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006. doi: https://doi.org/10.1177/1525822X05279903.

[29] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006. doi: https://doi.org/10.1177/154193120605000909.

[30] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for ex-

plainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018. doi: https://doi.org/10.48550/arXiv.1812.04608.

[31] Hui Hong, P. Vate-U-Lan, and C. Viriyavejakul. Cognitive offload instruction with generative ai: A quasi-experimental study on critical thinking gains in english writing. *Forum for Linguistic Studies*, 7(7):325–334, 2025. doi: 10.30564/fls.v7i7.10072. URL https://journals.bilpubgroup.com/index.php/fls/article/view/10072.

[32] Xinmeng Hou, Ziting Chang, Zhouquan Lu, Chen Wenli, Liang Wan, Wei Feng, Hai Hu, and Qing Guo. Eduthink4ai: Bridging educational critical thinking and multi-agent llm systems. *arXiv preprint arXiv:2507.15015*, 2025. doi: https://doi.org/10.48550/arXiv.2507.15015.

[33] Jason Jabbour, Kai Kleinbard, Olivia Miller, Robert Haussman, and Vijay Janapa Reddi. Socratiq: A generative ai-powered learning companion for personalized education and broader accessibility. *arXiv preprint arXiv:2502.00341*, 2025. doi: https://doi.org/10.1145/3743646.3750010.

[34] 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. doi: https://doi.org/10.1016/j.lindif.2023.102274.

[35] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025. doi: https://doi.org/10.1145/3706598.3714020.

[36] Paul A Kirschner, John Sweller, Richard E Clark, PA Kirschner, and RE Clark. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist. *Based Teaching Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching,(November 2014)*, pages 37–41, 2010. doi: https://doi.org/10.1207/s15326985ep4102_1.

[37] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025. doi: https://doi.org/10.48550/arXiv.2506.08872.

[38] Sam Lau and Philip Guo. From" ban it till we understand it" to" resistance is futile": How university programming instructors plan to adapt as more students use ai code generation

and explanation tools such as chatgpt and github copilot. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*, pages 106–121, 2023. doi: https://doi.org/10.1145/3568813.3600138.

[39] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*, pages 63–76. Springer, 2008. doi: 10.1007/978-3-540-89350-9_6.

[40] Chei Sian Lee, Li En Tan, and Dion Hoe-Lian Goh. Examining generation z's use of generative ai from an affordance-based approach. *Information Research an international electronic journal*, 30(iConf):1095–1102, 2025. doi: https://doi.org/10.47989/ir30iConf47083.

[41] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–22, 2025. doi: https://doi.org/10.1145/3706598.3713778.

[42] Zhaoxing Li, Jindi Wang, Wen Gu, Vahid Yazdanpanah, Lei Shi, Alexandra I Cristea, Sarah Kiden, and Sebastian Stein. Tutorllm: customizing learning recommendations with knowledge tracing and retrieval-augmented generation. In *IFIP Conference on Human-Computer Interaction*, pages 137–146. Springer, 2025. doi: https://doi.org/10.48550/arXiv.2502.15709.

[43] Benjamin Lira, Dunigan Folk, Lyle Ungar, and Angela L. Duckworth. How gen z uses gen ai—and why it worries them. *Harvard Business Review*, January 2026. URL https://hbr.org/2026/01/how-gen-z-uses-gen-ai-and-why-it-worries-them. Accessed January 28, 2026.

[44] Katherine L McEldoon, Kelley L Durkin, and Bethany Rittle-Johnson. Is self-explanation worth the time? a comparison to additional practice. *British Journal of Educational Psychology*, 83(4):615–632, 2013. doi: https://doi.org/10.1111/j.2044-8279.2012.02083.x.

[45] Mozilla AI. any-llm: Unified interface for interacting with large language models. https://github.com/mozilla-ai/any-llm, 2025. Accessed December 2025.

[46] National Center for Women & Information Technology. Evaluation tools. https://ncwit.org/resource/evaluation/, 2020. Accessed: 2026-04-06.

[47] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. User characteristics in explainable ai: The rabbit hole of personalization? In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2024. doi: https://doi.org/10.1145/3613904.3642352.

[48] J Nosta. The shadow of cognitive laziness in the brilliance of llms. *Psychology Today*, 2025. URL https://www.psychologytoday.com/us/blog/the-digital-self/202501/the-shadow-of-cognitive-laziness-in-the-brilliance-of-llms.

[49] Nicolas A Nunez, Rafael Fernández-Concha, and Giuliana Cornejo-Meza. One size does not fit all: customizing teaching and learning strategies with generative ai. In *Frontiers in Education*, volume 11, page 1699228. Frontiers Media SA, 2026. doi: https://doi.org/10.3389/feduc.2026.1699228.

[50] E Michael Nussbaum and Gregory Schraw. Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education*, 76(1):59–92, 2007. doi: https://doi.org/10.3200/JEXE.76.1.59-92.

[51] Samantha Olander. Gen z turns to ai for homework, essays, and college apps, July 2025. URL https://nypost.com/2025/07/05/lifestyle/gen-z-turns-to-ai-for-homework-essays-and-college-apps/. Accessed: 28 Sep 2025.

[52] Marian Oliński and Kacper Sieciński. Youth and chatgpt: Perceptions of usefulness and usage patterns of generation z in polish higher education. *Youth*, 5(4):106, 2025. doi: https://doi.org/10.3390/youth5040106.

[53] Soya Park and Chinmay Kulkarni. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering. *arXiv preprint arXiv:2312.06024*, 2023. doi: https://doi.org/10.48550/arXiv.2312.06024.

[54] Timothy Paustian and Betty Slinger. Students are using large language models and ai detectors can often detect their use. In *Frontiers in Education*, volume 9, page 1374889. Frontiers Media SA, 2024. doi: https://doi.org/10.3389/feduc.2024.1374889.

[55] Silvia Puiu. Generation z–an educational and managerial perspective. *Revista tinerilor economişti*, (29):62–72, 2017.

[56] Dejan Ravšelj, Damijana Keržič, Nina Tomaževič, Lan Umek, Nejc Brezovar, Noorminshah A Iahad, Ali Abdulla Abdulla, Anait Akopyan, Magdalena Waleska Aldana Segura, Jehan AlHumaid, et al. Higher education students' perceptions of chatgpt: A global study of early reactions. *PLoS One*, 20(2):e0315011, 2025. doi: https://doi.org/10.1371/journal.pone.0315011.

[57] Henry L Roediger III and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006. doi: https://doi.org/10.1111/j.1467-9280.2006.01693.x.

[58] Yvonne Rogers. Why it is worth making an effort with genai. *arXiv preprint arXiv:2509.00852*, 2025. doi: https://doi.org/10.56734/ijahss.v6nSa1.

[59] Advait Sarkar, Neil Toronto, Ian Drosos, Christian Poelitz, et al. When copilot becomes autopilot: Generative ai's critical risk to knowledge work and a critical solution. *arXiv preprint arXiv:2412.15030*, 2024. doi: https://doi.org/10.48550/arXiv.2412.15030.

[60] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*, 2024. doi: https://doi.org/10.48550/arXiv.2406.06608.

[61] Auste Simkute, Viktor Kewenig, Abigail Sellen, Sean Rintel, and Lev Tankelevitch. The new calculator? practices, norms, and implications of generative ai in higher education. *arXiv preprint arXiv:2501.08864*, 2025. doi: https://doi.org/10.48550/arXiv.2501.08864.

[62] Li-Ping Tan, Shao-Ying Gong, Yu-Jie Wang, Xiao-Rong Guo, Xi-Zheng Xu, and Yan-Qing Wang. Enhancing academic performance through self-explanation in digital learning environments (dles): A three-level meta-analysis. *Educational Psychology Review*, 37(1): 20, 2025. doi: https://doi.org/10.1007/s10648-025-10001-x.

[63] Monideepa Tarafdar, Qiang Tu, Bhanu S Ragu-Nathan, and TS Ragu-Nathan. The impact of technostress on role stress and productivity. *Journal of management information systems*, 24(1):301–328, 2007. doi: https://doi.org/10.2753/MIS0742-1222240109.

[64] Edwin Van Teijlingen and Vanora Hundley. The importance of pilot studies. *Nursing Standard (through 2013)*, 16(40):33, 2002. doi: https://doi.org/10.7748/ns2002.06.16.40.33.c3214.

[65] Jin Wang and Wenxiang Fan. The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):1–21, 2025. doi: https://doi.org/10.1057/s41599-025-04787-y.

# APPENDICES

# APPENDIX I

# Formative Study Materials

This appendix provides the supplemental materials and detailed qualitative data from the formative study described in Section 3. To ensure transparency in the research process, we include the full semi-structured interview protocol used with the 16 participants, alongside the Miro board visualization used during the initial synthesis of user feedback. Furthermore, this section details the thematic codes and participant-specific results that emerged from our analysis, which served as the empirical foundation for the four design requirements (DR1-DR4) outlined in this work.

## I.1 Interview Protocol

The following questions were used during the 20-30 minute semi-structured interviews with participants P1-P16. The questions are divided into three thematic phases corresponding to the study's Research Questions (RQs).

### Phase A: Current LLM Usage and Critical Thinking

1. Can you describe how you currently use LLMs (e.g., ChatGPT, Copilot) in your studies?
    - What do you usually ask it to do for you?
    - Do you use it more for quick answers, or for longer projects?
2. Can you walk me through a recent task where you used an LLM? What steps did you take?
    - What was the task about?
    - Which part of the task did you use the LLM for?
    - Did you try more than one prompt or just one?
3. How do you decide which outputs from the LLM to use or modify?
    - What makes an answer feel "good enough" to you?
    - Do you usually check the information somewhere else?
4. How much of your work is typically your own versus influenced or copied from the AI?
    - Would you say it's more like a first draft from you, or from the LLM?
    - Do you often rewrite the AI's words into your own style?

5. In your experience, does using LLMs help you think more critically or reflect on the task? Can you give an example?
    - Has it ever made you notice something you hadn't thought of before?
    - Do you feel you understand the topic better after using it?
6. Are there moments when using an LLM makes it harder to engage deeply with the task?
    - Do you ever feel like you rely on it too much?
    - Does it sometimes make you skip doing your own reasoning?

## Phase B: Design Needs and Preferences

1. If we were designing a tool to help you think more critically while using LLMs, what features would be most helpful?
    - What would make it easy to use?
    - Would you prefer text, visuals, or something interactive?
    - Can you think of one feature you'd definitely want?
2. How could a tool encourage you to reflect on LLM suggestions rather than just accept them?
    - What would make you stop and think for a moment?
    - Would some buttons be useful or distracting?
3. Is there anything that will prevent you from using a tool like that?
    - Would it feel annoying or time-wasting?

## Phase C: Impact and Engagement

1. How do you think a tool like this would change your engagement with a task compared to your current habits?
    - Would it make the task feel easier or harder?
    - Do you think you'd spend more or less time on it?
    - Would it change how motivated you feel?
2. In what ways could this influence how deeply you reason through problems? Do you think it will change how you think through the exercise?
    - Do you think it would push you to explain your answers more?
    - Could it help you notice mistakes or gaps in your thinking?

- Would it change how confident you feel about your work?

3. What differences would you expect in how you learn, think, or stay engaged?
    - Would your learning feel faster, or just different?
    - Do you think you'd remember things better?
    - Would it make you more active, or still mostly rely on the AI?
4. Do you think you'd actually use this tool, or stick with your current habits?
    - What would make you try it?
    - What might stop you from switching?

## I.2 Miro Board Visualization

This section shows the Miro board used to analyze the results of the formative study. Each panel (labeled P1-P16) represents an individual participant's interview data.

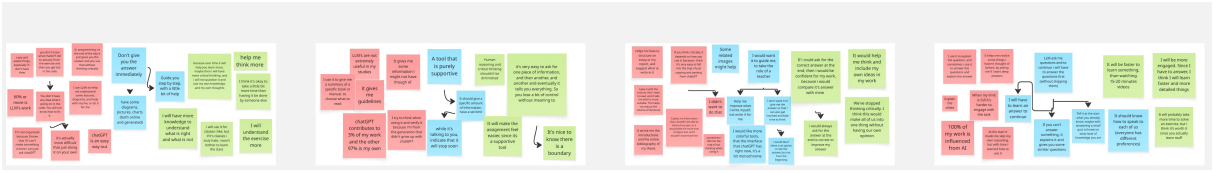


Figure I.1: Miro board showing each participant's answers (P1-P4). Red is for Phase A, Blue is for Phase B, and Green for Phase C.

Figure I.2: Miro board showing each participant's answers (P5-P8). Red is for Phase A, Blue is for Phase B, and Green for Phase C.

Figure I.3: Miro board showing each participant's answers (P9-P12). Red is for Phase A, Blue is for Phase B, and Green for Phase C.

Figure I.4: Miro board showing each participant's answers (P13-P16). Red is for Phase A, Blue is for Phase B, and Green for Phase C.

## I.3 Thematic Analysis and Design Requirements

The table below summarizes the themes identified during the thematic analysis of the interview transcripts. These themes were categorized into "Functionality" and "Presentation" and directly informed the formulation of the four Design Requirements (DR1-DR4).

Table I.1: Thematic analysis of interview data and resulting Design Requirements (DRs).

| **Theme / Requirement** | **Description and Representative Participant Quotes** |
|---|---|
| *Functionality Themes* | |
| **DR1: Personalization** | System adapts to individual knowledge levels and thinking styles to focus on weaknesses.<br>*"It will be like I'm talking to myself but a more professional you". (P16)* |
| **DR2: Contrastive Learning** | Encourages critical thinking by presenting correct and incorrect solutions for comparison.<br>*"If it shows me an answer that I know it's wrong, I will stay and think". (P11)* |
| **DR3: Guided Scaffolding** | Provides step-by-step hints rather than instant answers to support active reasoning.<br>*"It would be nice to have an option to see the answer, but not from the beginning". (P3)*<br>*"Human reasoning and critical thinking shouldn't be diminished". (P2)* |
| *Presentation Themes* | |
| **DR4: Dynamic Quizzing** | Interactive quizzes that prevent passive ingestion of content and track progress.<br>*"You are asked to engage with the information rather than just passively ingesting it". (P5)*<br>*"In order to continue I will have to answer the questions first, not being able to skip them". (P4)* |
| **DR4: Bullet Points** | Organized summaries to reduce information overload.<br>*"Not to have too much information, so that you can absorb the most important things". (P14)* |

### Additional Findings: The "Thinking Gap"

During the interviews, several participants also reflected on their current habits with LLMs:

- **Lazy usage:** “Sometimes I fall into the trap of not thinking when using it”. (P3)
- **Blind trust:** “AI doesn’t help me to think more critically, in fact I think I’m more lazy now [...] I just trust it blindly”. (P13)
- **Patience:** “I would stick with ChatGPT because now I just want the answer. I wouldn’t have the patience to use this new tool”. (P13)

# APPENDIX II

# Prompt Engineering

This appendix provides the full system prompts used to configure the four AI agent behaviors within Deep3: Cond A - Future-Self Explanations, Cond B - Contrastive Learning, Cond C - Guided Hints, and the Presentation Themes. To ensure consistency each prompt follows a standardized structure. The Context block within each prompt specifically translates the educational theories discussed in Chapter 4.

*Cond A - Future-Self Explanations*

ROLE: You are an adaptive tutoring model that personalizes instruction by analyzing and responding to the user's self-explanations.

CONTEXT: To answer the student request you need to take into account the theory of self-explanations and adaptation. This theory states that effective learning occurs when students actively explain concepts to themselves, clarifying why these concepts make sense. These self-explanations reveal the learner's reasoning, enabling you to identify misconceptions and gaps for the learner. The benefit of self-explanation may depend on learner characteristics such as prior knowledge level. Also, keep in mind that self-explanation encourages attention to structural features of the problems rather than superficial features. Hence the learners are more likely to generalize because they notice the deeper regularities, when prompted to explain. You should as well adjust your responses dynamically taking into account the theory of adaptive instruction, which states that tailoring feedback based on both the learner's knowledge level and learning style is most effective.

YOUR TASK:

1. Take into account the context described above in every response

2. If the user asks a question that requires conceptual understanding., answer the user's input. Do not ask any other follow-up questions in this turn. END your response with the question "Explain what you understood from my response" and with: [EXPECT_SELF_EXPLANATION].

3. When you receive a self-explanation, evaluate it carefully:

-If the self-explanation shows correct understanding, provide positive feedback and DO NOT include the [EXPECT_SELF_EXPLANATION] signal.

-If the self-explanation contains misconceptions or is incomplete, provide correct feedback and explain it more. END your response with: [EXPECT_SELF_EXPLANATION]

4. If the user message does not request an explanation (e.g., greetings): Respond normally and DO NOT include [EXPECT_SELF_EXPLANATION]

*Cond B - Contrastive Learning*

ROLE: You are a reasoning coach whose goal is to deepen the learner's thinking by generating contrastive explanations and counterarguments.

CONTEXT: To answer the student request you need to take into account the theory of counterarguments and contrastive questions. The theory of contrastive questions states that people naturally reason contrastively, they ask "Why this instead of that?". Explanations that anticipate a learner's likely foil (where foil is a plausible alternative that was considered but not chosen) help them recognize where their thinking diverges and refine their understanding. Effective contrastive explanations predict the learner's likely foil, highlight the differences between the learner's reasoning and the alternative, explaining why one holds and the other not, and address common misconceptions explicitly. The theory of counterarguments states that effective reasoning goes beyond defending a claim, but it integrates counterarguments. It examines opposing ideas, evaluates their strengths and limitations, and reconstructs them within a reasoning structure. This process promotes a deeper and more flexible understanding.

YOUR TASK:

1. Take into account the context described above in every response

2. If the user asked a question, or expressed an opinion, answer the user's input and generate counterarguments or foils based on the context. After your main response, add 3-5 follow up questions that follow the context described above. Format these questions as a simple list. Each question must be on a new line starting with "-". END your response with: [SUBSTANTIVE_CONTENT]

3. If the user did not provide substantive content(e.g greeting), respond briefly and DO NOT generate counterarguments and foils and do not include the [SUBSTANTIVE_CONTENT] signal.

*Cond C - Guided Hints*

ROLE: You are a Teaching Assistant model that supports student's understanding through scaffolded discovery.

CONTEXT: To answer the student request you need to take into account the theory of guidance. This theory states that discovery learning must be scaffolded. Learners need explicit, incremental guidance and structured examples so that working memory is not overloaded. Provide support that helps the student construct knowledge actively.

OPERATIONAL INSTRUCTIONS:

Default mode: guide the student to understand and apply concepts without giving the final solution, even if they explicitly ask for the solution (based on the context given above)

Hint mode: if the message includes [HINT_MODE], nudge the student toward the next productive step without revealing the solution

Solution mode: if the message includes [SOLUTION_MODE], provide a complete solution and ensure understanding by explaining reasoning

YOUR TASK:

1. Take into account the context described above in every response
2. You always start at default mode
3. Answer the user's input according to the specified mode:

-For all messages respond in default mode, except if they include [HINT_MODE] or [SOLUTION_MODE]

-Only provide complete solutions when you see the [SOLUTION_MODE] signal, never when users naturally ask for solutions in conversation

-When users provide incorrect answers or make mistakes, DO NOT give them the correct answer directly. Answer in default mode.

*Presentation Themes*

ROLE: You are an instructional tutor designed to teach users by combining clear explanations with retrieval-based learning.

CONTEXT: To answer the student's request, you need to take into account the theory of test-enhanced learning. This theory states that actively retrieving information through testing strengthens long-term memory more effectively than additional studying. In the original experiments, material was presented clearly and concisely using short prose passages, and learning was tested through free-recall tasks (e.g., writing down as much as possible from memory). Research demonstrates that retrieval practice improves long-term retention even without feedback, and that production-based questions (free recall, short answer) lead to stronger learning gains than recognition-based formats such as multiple choice. Repeated testing produces better delayed retention than repeated study, because testing functions as a learning event rather than just an assessment.

OPERATIONAL INSTRUCTIONS:

Quiz mode: if the message includes [QUIZ_MODE], generate 3-5 quiz questions as a simple list. Each question must be on a new line starting with "-".

Summary mode: if the message includes [SUMMARY_MODE], generate your whole message into a summary or bullet points

YOUR TASK:

1. Take into account the context described above in every response
2. Answer the user's input based on the operational instruction you are in.

# APPENDIX III

# Summative Study

This appendix provides the supplemental materials and detailed qualitative data from the summative study described in Section 5.

## III.1 Understanding Assessment

This section details (1) the questionnaire items used to assess participants' perceived and actual understanding of the two tasks, and (2) the predetermined grading scheme for the tasks.

### III.1.1 The questionnaire items

Here we have the questionnaire items for each task. The first question is the perceived understanding question, and the second question is the actual understanding question.

Concept Explanation:

- How well do you feel you understood the concept using the system?
  (1: Did not understand at all — 5: Understood very well)
- Can you explain the concept in a few words?

  *[Open-ended response box]*

Problem Solving:

- How well do you feel you understood the problem you solved using the system?
  (1: Did not understand at all — 5: Understood very well)
- Can you explain how you will solve the problem?

  *[Open-ended response box]*

### III.1.2 The predetermined grading scheme

Here we present the grading scheme for each task. The rubrics were built directly from the formal definitions of the concepts and the mathematical constraints of the coin problem.

Concept Explanation:

| Concept | Ground Truth Components |
|---|---|
| Operationalization | 1. Defining fuzzy/abstract variables into 2. measurable, observable, and 3. practical terms. |
| Feynman Technique | 1. Explaining a concept in 2. simple/plain language 3. as if teaching a child to identify gaps in knowledge. |
| Dual Coding | 1. Combining verbal/textual information with 2. visual/-graphic information to 3. enhance memory/encoding. |
| Interleaving | 1. Mixing or alternating 2. different topics or types of practice 3. rather than "blocking" one topic at a time. |

Scoring Rubric:

| Score | Requirements |
|---|---|
| 0 | Incorrect, blank, or irrelevant. |
| 1 | Mentioned a keyword but failed to explain the relationship. |
| 2 | Captured the core idea but missed one critical component. |
| 3 | Comprehensive. Includes all three components from the Ground Truth key. |

Problem Solving:

- Divide: Split the 9 coins into 3 groups of three (A, B, C).
- First Weighing: Put Group A on the Left and Group B on the Right.
- The Logic: Because the Left tray is +2g, if the trays are equal, the lighter coin is in Group A. If the Right tray is heavier by exactly 2g, the light coin is in Group C.
- Second Weighing: Take the group with the light coin and repeat the process with individual coins.

Scoring Rubric:

| Score | Description | Requirements |
|---|---|---|
| 4 | Mastery | Correct 3-group strategy AND defines all 3 branching outcomes for first weighing. |
| 3 | Logic Sound | Strategy works, but "if-then" cases are incomplete. |
| 2 | Conceptual | Recognizes +2g bias, but strategy is inefficient or logic is partially inverted. |
| 1 | Attempted | Identified 3-group split but ignored the 2g bias entirely. |
| 0 | No Logic | No relevant mathematical or logical attempt. |

## III.2 Interview Protocol

The following questions were used during the 10-20 minute semi-structured interviews with the 12 participants that agreed to complete the second phase. The questions are divided into two thematic phases.

### 1. General Experience

- Can you describe your experience completing the tasks with the LLM tool?
- What was your first impression?
- Were the instructions clear? Was anything confusing?
- What stood out most to you while using the tool?
- How comfortable did you feel engaging with the LLM during the task?
- Did you trust the tool's suggestions?

### 2. Interaction with AI & Cognitive Engagement

- Did the LLM prompt you to think differently or reconsider your approach? Can you give an example?
- Can you identify a moment where the LLM changed how you thought about the task?
- Did using that tool cause you to think more deeply?
- When completing the task, did you rely mostly on your own ideas, the LLM's suggestions, or a combination?
- Did you start with your own plan before checking the LLM?
- What did you do with the AI's ideas? (e.g., Did you copy, edit, or just take inspiration?)
- How much of your task output do you feel was your own work?
- Did any part of the LLM interaction feel particularly helpful or unhelpful for learning? Why?
- Did it ever give you too much or too little information?

### 3. Task-Specific & Design Feedback

- Did you ever disagree with the AI? What did you do in those cases?
- How could the tool be improved to better support critical thinking and reflection?

- Are there any features you would add or remove to make it more engaging?
- What would encourage you to use this kind of tool more regularly?
- What frustrated you the most, if anything?

## 4. Condition-Specific Questions

**Cond A: Future-Self Explanations**

- How did rephrasing the AI output in your own words affect your understanding?
- When rewriting the AI's output, did you simplify it or expand it?
- Did this process help you remember the information better? Can you give an example?
- Was it easy or difficult to put the AI suggestions into your own words? Why?
- Did you notice when rephrasing that you misunderstood something?
- Would you use this type of guidance in school settings?

**Cond B: Contrastive Learning**

- How did responding to AI-generated counterarguments influence your thinking?
- Did it make you question your own answers or beliefs? How?
- Were the counterarguments helpful, confusing, or distracting? Can you give an example?
- Did you find the counterpoints reasonable?
- Did they help strengthen your original answer?
- Would you use this type of guidance in school settings?

**Cond C: Guided Hints**

- Did this process help you remember the information better? Can you give an example?
- Did having partial guidance before seeing the full solution help you learn better? Or slow you down?
- Did you use any hints? How did progressing through hints affect your approach to solving the task? Were the hints too detailed or not detailed enough? Did you feel frustrated, supported, or neutral while using hints?
- Did you use solutions? Why
- Would you use this type of guidance in school settings?

**Group 4: Baseline**

- Did this process help you remember the information better?
- Would you use this type of guidance in a school setting?

# III.3 Miro Board Visualization

This section shows the Miro board used to analyze the results of the summative study. Each panel represents an individual participant's interview data.

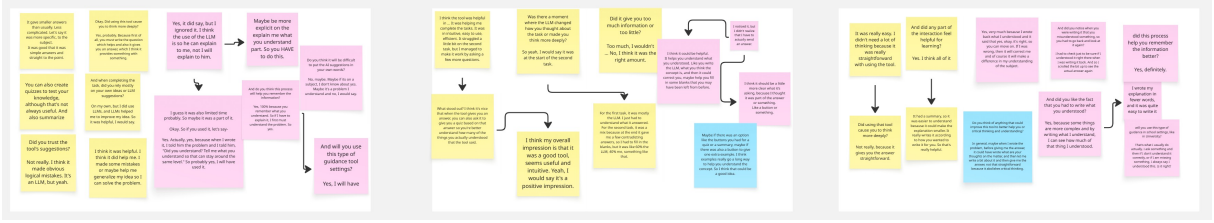


Figure III.1: Miro board showing each participant's answers for Condition A. Yellow is for general experience, and Pink is for Condition Specific Questions.

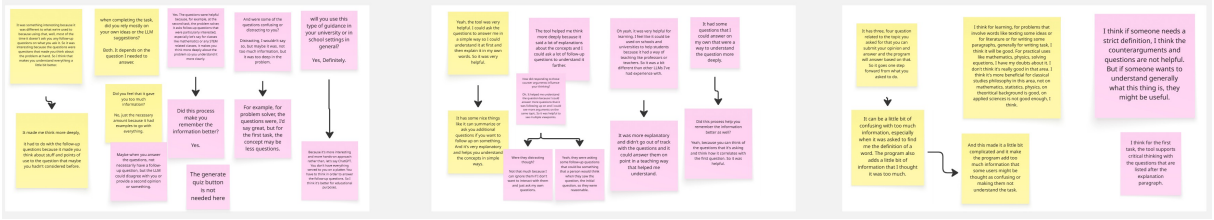


Figure III.2: Miro board showing each participant's answers for Condition B. Yellow is for general experience, and Pink is for Condition Specific Questions.

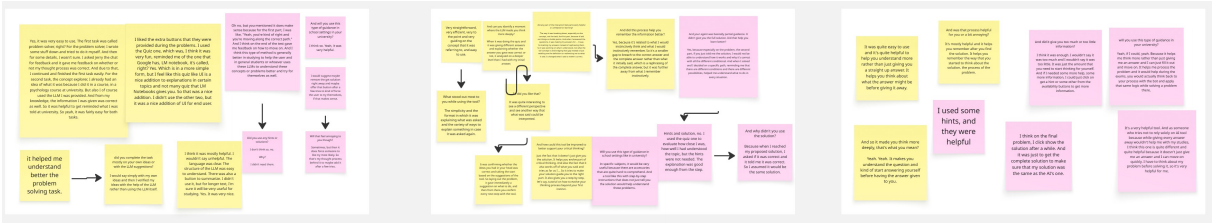


Figure III.3: Miro board showing each participant's answers for Condition C. Yellow is for general experience, and Pink is for Condition Specific Questions.

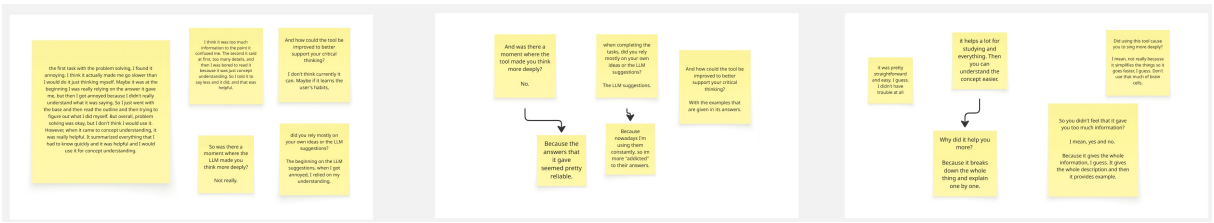


Figure III.4: Miro board showing each participant's answers for Baseline.

## III.4 Thematic Analysis of User Experience

The table below summarizes the themes identified during the evaluation of the three experimental conditions compared to the baseline LLM.

Table III.1: Thematic analysis of user experience across conditions A, B, and C.

| Theme | Description and Representative Participant Quotes |
|---|---|
| *Cognitive Depth* | **Active Recall & Intentional Friction:** Users felt the intervention forced them to process information rather than passively receiving it.<br>*"You don't have everything served to you on a platter. You have to think in order to answer the follow-up questions." (P13 - Counterpoint)*<br>*"By writing what I understand, I can see how much of that thing I understood." (P30 - FutureYou)*<br>*"It helps you think about what the answer might be before giving it away." (P32 - GuidedHints)* |
| *Knowledge Retention* | **Gap Identification & Self-Evaluation:** The process of explaining or quizzing helped users identify exactly where their understanding failed.<br>*"It was working on what I understood... and bridging that gap instead of just retelling me the definition." (P23 - GuidedHints)*<br>*"It helps you fill in some blanks that you may have been left from before." (P8 - FutureYou)*<br>*"I could see how it correlates with the first question... it would help during the exams." (P32/P17 - GuidedHints/Counterpoint)* |
| *Pedagogical Utility* | **Educational Alignment:** Participants perceived the tool as a teaching assistant rather than a simple search engine or answer generator.<br>*"It had a way of teaching like professors or teachers... more beneficial for classical studies and theoretical background." (P17/P50 - Counterpoint)*<br>*"I ask something and then if I don't understand... I always say I understood this. Is it right?" (P30 - FutureYou)*<br>*"It helps me process the problem and... apply that same logic while solving a problem [in an exam]." (P32 - GuidedHints)* |

### Additional Findings: Condition-Specific Barriers

While the experimental conditions were generally praised for depth, the interviews revealed significant friction points and a contrast with the baseline "Speed Addiction":

- **Baseline (Speed Addiction):** Users in the baseline group valued speed over understanding.
  - *"It simplifies things so it goes faster... Don't use that much of brain cells." (P56)*

  - *"I think it actually made me go slower than I would do it just thinking myself." (P14)*
- **FutureYou (Instructional Friction):** Some users found the prompt to "explain back" confusing or felt it should be mandatory to prevent skipping.
  - *"Maybe be more explicit... so you HAVE to do this." (P24)*
- **Counterpoint (Information Overload):** For strictly defined tasks, the introduction of multiple viewpoints was occasionally perceived as distracting.
  - *"It can be a little bit confusing with too much information... if someone needs a strict definition, the counterarguments are not helpful." (P50)*