# More AI Assistance Reduces Cognitive Engagement: Examining the AI Assistance Dilemma in AI-Supported Note-Taking

XINYUE CHEN, University of Michigan, USA
KUNLIN RUAN, University of Michigan, USA
KEXIN PHYLLIS JU, University of Michigan, USA
NATHAN YAP, University of Michigan, USA
XU WANG, University of Michigan, USA

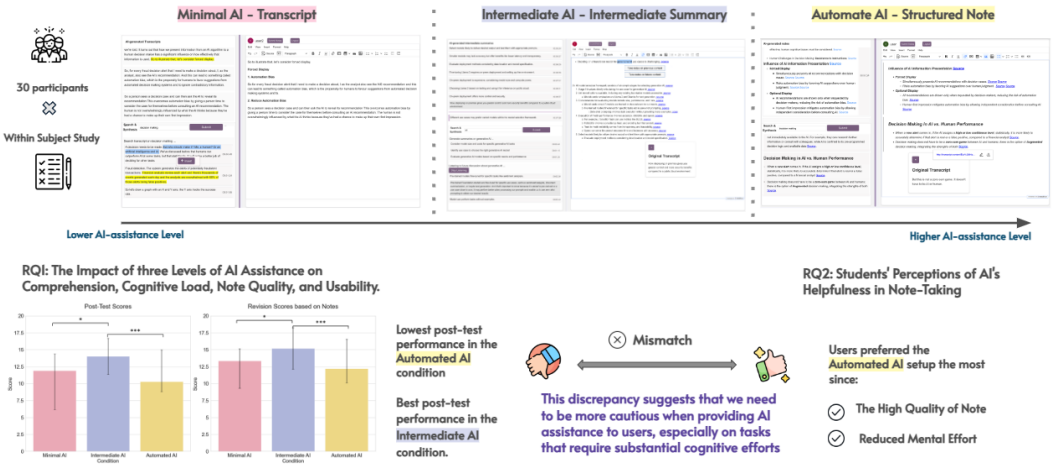


Fig. 1. In a within-subject experiment with 30 participants, we tested three conditions: `Automated AI` (high assistance), `Intermediate AI` (moderate assistance with summaries), and `Minimal AI` (transcript-only). Results show `Intermediate AI` yielded the highest learning outcomes, while `Automated AI` scored lowest despite being the most preferred for ease of use. These findings highlight the gap between preferred convenience and cognitive benefit, offering guidance for designing AI that enhances cognitive engagement in learning tasks.

As AI tools become increasingly integrated into cognitively demanding tasks, like note-taking, questions remain about whether they enhance or compromise cognitive engagement. This paper investigates the "AI Assistance Dilemma" in note-taking, examining how varying levels of AI support impact user engagement and comprehension. In a within-subject experiment, we asked participants (N=30) to take notes during lecture videos under three conditions: `Automated AI` (high assistance with structured notes), `Intermediate AI` (moderate assistance with real-time summary, and `Minimal AI` (low assistance with transcript). Results reveal that Intermediate AI yields the highest post-test scores and Automated AI the lowest. Participants,


*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

however, preferred the automated setup for its perceived ease of use and perceived lower cognitive effort, suggesting a discrepancy between preferred convenience and cognitive benefit. Our study provides insights on designing AI assistance that preserves cognitive engagement, offering implications for designing moderate AI support in cognitive tasks.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

**ACM Reference Format:**
Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2018. More AI Assistance Reduces Cognitive Engagement: Examining the AI Assistance Dilemma in AI-Supported Note-Taking. In . ACM, New York, NY, USA, 37 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

AI is increasingly used to facilitate cognitively challenging tasks, from providing next-sentence suggestions in writing [12] to generating meeting summaries [2, 7, 58]. The rise of AI productivity tools [18, 48, 50] raises the question: is AI assistance always desired and beneficial? When AI is leveraged to enhance productivity, could it compromise people's cognitive engagement in the task, leading to suboptimal outcomes? In this paper, we explore the AI assistance dilemma in the context of an important and common task: note-taking. We examine to what extent AI-powered automatic note-taking services might reduce valuable cognitive engagement of the note-taker and how we could design human-AI collaborative mechanisms to provide the necessary AI assistance to the note-taker while preserving the cognitive effort required for meaningful content engagement.

Note-taking is a common activity during real-time information consumption, such as in meetings [53, 55], lectures [19, 46], and presentations, where participants must quickly process information without the ability to revisit what was said [7, 33]. Note-taking has proven to be a helpful tool to assist real-time information consumption by externalizing information and reducing the load on participants' working memory [33]. Note-taking serves two key functions: it helps people internalize knowledge through active engagement (encoding function) and provides a lasting reference for future recall(storage function) [29, 30].

Note-taking offers a unique and valuable context to study the AI assistance dilemma due to its dual demands on cognitive processing and external documentation [17, 60]. Effective note-taking requires users to engage deeply with information, selecting and organizing key points in real time, which promotes active learning and understanding [14, 38]. However, when AI tools are introduced to automate or aid in this process, there is potential for reduced cognitive engagement, as users may rely on the AI rather than actively process information themselves [24].

It thus raises important questions about AI's role: Does AI assistance merely alleviate cognitive load, or does it undermine the mental engagement crucial for comprehension and retention? This concern reflects a longstanding issue known as the "Assistance Dilemma" in learning sciences, originally identified in the context of cognitive tutors deployed in classrooms [32]. This dilemma underscores that while AI-generated feedback can enhance learning, excessive assistance risks undermining learners' capacity for independent learning and critical thinking, whereas insufficient assistance may make tasks overly challenging [32, 44]. With the growing prevalence of AI assistance, we aim to investigate to what extent the AI assistance dilemma exists and how we might design human-AI collaborative mechanisms to address this challenge in a note-taking context.

To this end, we designed three levels of AI assistance to help people take notes during live lectures, namely 1) `Automated AI` , where participants received auto-generated AI notes around every

90 seconds, similar to common note-taking services; 2) `Intermediate AI` , where participants received real-time AI summary blocks; and 3) `Minimal AI` , where participants received real-time transcript blocks. In all three conditions, participants take notes in a text editor. Participants can drag and drop the AI-generated information (structured notes, summary blocks, and transcript blocks) into the text editor to compose their own notes. The `Automated AI` setup resembles popular note-taking services that provide a summary after a time interval [1]. With this setup, we aim to answer whether automated note-taking support would compromise people's cognitive engagement and negatively impact understanding of the lecture content. By introducing the `Intermediate AI` and the `Minimal AI` setups, we aim to address whether giving users more control and agency during the note-taking process would enhance their understanding of the content compared to the `Automated AI` setup. We used a post-test quiz after the lecture to measure people's understanding, which is a signal of valuable cognitive engagement during the lecture.

We performed a within-subject experiment with 30 participants, where they watched lecture videos and took notes in the three setups respectively. Our findings suggest that 1) the AI assistance dilemma does exist. People had the lowest post-test performance in the `Automated AI` condition and the best post-test performance in the `Intermediate AI` condition. The difference is statistically significant (Figure 7). `Minimal AI` has better performance than `Automated AI` , but the difference is marginally significant. This finding suggests that although AI may enhance productivity, it risks compromising people's cognitive engagement and may lead to lower comprehension and retention. 2) Providing intermediate summary blocks is a useful way to address this challenge, as users in the `Intermediate AI` condition had the best post-test performance. Users had more control and more valuable cognitive engagement during the note-taking process when they were provided with the building blocks and had the agency to compose notes by themselves. Using AI to generate summaries instead of presenting information verbatim reduces cognitive load and supports understanding. 3) However, users preferred the `Automated AI` setup the most. They found the auto-generated notes were of high quality and found them to significantly reduce their effort. This discrepancy suggests that we need to be more cautious when providing AI assistance to users, especially on tasks that require substantial cognitive efforts. As users naturally prefer to minimize effort, the risk of over-reliance on AI becomes even more pronounced.

## 2 RELATED WORK

### 2.1 Cognitive Costs and Benefits of Note-taking

Lecture note-taking has long been shown to be effective for learning. In-class note-taking helps students actively process information [38], improving memory retention [16], while the notes can also serve as an external representation for post-class review, which reinforces understanding and supports long-term knowledge retention [38, 42]. The underlying mechanism for these effects, known as the encoding-storage paradigm [29], suggests two key functions: ENCODING, where recording and taking notes itself promotes learning through increased attention and deeper processing, even without later review, and STORAGE, where notes serve as an external memory for later review, helping reinforce understanding by consolidating information, reconstructing unrecorded details, slowing forgetting, and relearning forgotten content.

Unfortunately, many studies have shown detrimental effects of taking notes while learning, as note-taking costs attentional resources that are needed for processing rapid and dense lecture presentations [5]. Research linking note-taking with cognitive load theory indicates that if appropriate note-taking strategies are not employed, comprehension might be impaired when students try to write down information verbatim due to an increase in extraneous cognitive load,

[1] http://otter.ai/

which adversely affects learning effectiveness [61]. Moreover, germane cognitive load—arising from actively organizing and integrating information—may be weakened when notes lack structure [17]. To optimize note-taking, researchers have explored structured techniques, finding that formats like pre-organized sheets [14] and outlines [30] reduce extraneous load and promote germane processes such as comprehension and synthesis [10, 25]. These structured approaches can also enhance metacognition, encouraging students to focus on understanding knowledge structures and how to prioritize, organize, and learn effectively [14].

In summary, past research underscores the importance of reducing unnecessary cognitive load while supporting note-taking's encoding and storage functions. Building on these theories, our study explores how varying levels of AI assistance impact encoding, storage, and cognitive load optimization in note-taking.

### 2.2 Tools to Enhance Note-taking

HCI researchers have developed numerous tools to support note-taking. One line of the research focuses on note-taking modalities and input methods, offering flexibility for capturing notes via voice [28], digital ink [22], sketching [65], and mobile inputs [43]. Studies comparing note-taking modalities reveal that while typing enables faster capture, handwriting better supports encoding due to its hands-on nature, though results vary across contexts [3, 17]. While these tools emphasize input affordances, they didn't address real-time, content-specific support within classroom settings. The researcher also seeks ways to scaffold note-taking, including, but not limited to, guiding students to take notes in structured formats or with predefined structures, such as concept maps [6, 51] or designing collaborative tools such as NotePals [11], LiveNotes [27], or NoteCoStruct [15] that reduce individual cognitive load by distributing note-taking responsibilities among group members. While collaborative notes are proved to be of higher quality [37], they may compromise individual agency since students lean towards relying on contributions made by their peers instead of personal involvement [9, 10, 16]. This dependence can weaken the personal encoding benefits of note-taking, thus reducing comprehension and retention [10].

Recently, AI has increasingly supported automated note generation, especially in complex contexts like clinical documentation [34, 54, 57] and online meetings [7]. Systems such as PhenoPad [57], MediNotes [34], and GazeNoter [54], as well as commercial tools like Zoom AI Companion [2], Teams Copilot [3], and Otter.ai [4], use AI to capture, summarize, and organize meeting content automatically. While AI automation is advantageous in clinical environments where efficiency is paramount, classroom note-taking differs in that it serves not only as information storage but as a critical encoding process [29]. Fully automated notes in classrooms may inhibit the encoding process, yet research on the ideal level of AI assistance in note-taking remains scarce. While a few recent studies started exploring using AI to generate an intermediate level of notes that users can further work on, providing insights into balancing user engagement and AI assistance [8], the impact of AI on real-time encoding and post-lecture storage or decoding processes in educational settings has yet to be systematically quantified and understood.

### 2.3 AI Assistance Dilemma in Cognitive Tasks

The assistance dilemma describes the challenge of balancing assistance with autonomy in learning, highlighting the need to provide optimal levels of support without fostering dependency [32]. In educational research, solutions to the assistance dilemma emphasize carefully calibrated scaffolding

[2]https://www.zoom.com/en/ai-assistant/
[3]https://support.microsoft.com/en-us/office/use-copilot-in-microsoft-teams-meetings
[4]https://otter.ai/

strategies to guide students' meta-cognitive skills [45] and adjust the level of help based on learners' skill and engagement levels [26]. In early problem-solving stages, worked examples are frequently paired with minimal guidance, transitioning to increased feedback or step-by-step hints as learners advance, encouraging their independent reasoning [39]. Other solutions use progressive prompts that adjust support based on performance, with some methods applying data-driven adaptive support to determine when learners need assistance based on their behavior [56].

### 2.4 Design to Balancing AI assistance in Sense-making Activities

In cognitive tutoring, studies indicate that varying levels of scaffolding are effective, yet these studies primarily pertain to predefined problem domains and structured worked examples, where they can effectively trace students' knowledge progression [39, 62]—techniques that do not directly apply to real-time note-taking. Recent research on using large language models (LLMs) for sense-making tasks in diverse scenarios provides insights into how to balance AI assistance and human engagement in note-taking [8, 12, 36, 40, 63]. A commonly useful approach involves positioning AI as an intermediary layer between the user and the cognitive task, providing a moderate level of assistance that aids without overstepping into autonomy [12, 40]. For example, in writing tasks, offering full paragraph suggestions may reduce user agency, while sentence-level assistance effectively reduces the difficulty of phrasing without limiting creative freedom [12]. Similarly, in LLM-assisted programming, providing next-step hints helps students learn coding skills, whereas full-block auto-completions often lead to more edits and increased dependency on the AI [40].

Another essential design aspect is that to preserve user autonomy, AI support often acts as middleware rather than being directly embedded into workflows. For instance, in the MeetMap system, researchers introduced a temporary AI-generated content holding area, allowing users to decide which suggestions to use or ignore [8]. This middleware design encourages metacognition rather than passive acceptance when interacting with AI [63]. Additionally, the level of AI abstraction and the potential uncertainty it brings is an important design consideration [18, 20]. For instance, a resilient interface can continuously highlight critical information, increasing AI's abstraction of original content as needed [20].

Our three levels of AI assistance draw insights from these designs, balancing the granularity of AI-provided content and the agency provided to users, incorporating a middleware interaction layer, and adjusting abstraction levels to support effective note-taking.

## 3 STUDY

With the growing prevalence of AI assistance, we aim to investigate to what extent the AI assistance dilemma exists and how we might design human-AI collaborative mechanisms to address this challenge in a note-taking context. First, to investigate whether an AI assistance dilemma exists, we examine how different AI assistance levels would impact people's cognitive engagement and understanding of the content in a note-taking task, where people attend live lectures and take notes to help them understand the content. On the one hand, we could hypothesize that AI-generated notes reduce the effort in note-taking, which will reduce the user's cognitive engagement and lead to lower comprehension. Conversely, we could also hypothesize that because AI generates structured notes at short intervals—every one to two minutes—it reduces the users' note-taking load and may enhance understanding by synthesizing information for them. Second, to mitigate the over-reliance issue, we examine how we can give users more control and engage them more in the encoding process of note-taking. We consider providing users with real-time summary blocks of the lecture, and real-time transcripts. Summary blocks are easier to read, whereas transcript blocks are more transparent since they are a true representation of the conversation. In this study, we compare the three setups of AI-assisted note-taking, namely providing slightly delayed AI-generated structured

notes, real-time summary blocks, and real-time transcript blocks. We introduce a post-test of the lecture to measure people's understanding, which is a signal of their cognitive engagement level during the lecture.

We performed a within-subject experiment with 30 participants, where they watched lecture videos and took notes in the three setups respectively.

### 3.1 Experiment design

*3.1.1 Apparatus.* We designed the note-taking tasks with three distinct levels of AI assistance, drawing from prior research on human-AI collaboration tools for sense-making activities, as discussed in section 2.4. In those studies, the assistance level was differentiated based on the degree of abstraction and scaffolding provided by AI, as well as the level of user agency retained—emphasizing how manual effort is distributed between humans and AI [13, 20]. In the context of lecture note-taking in this research, the assistance levels reflect how much of the encoding process AI automates and the degree of user autonomy in managing and engaging with content. Besides, the level of abstraction in such real-time tasks may also be related to the synchronicity users receive the AI assistance [36]. In this study, the assistance level apparatus is designed as follows:

- `Automated AI` : AI generates a structured note block every two minutes. Users can incorporate these structured blocks into their own notes - the notes are shown in a granularity that is similar to the business usual AI note-taking tools [5][6]
- `Intermediate AI` : AI produces turn-level summary blocks for each speaking turn. Users can select and incorporate these intermediate summary blocks into their notes.
- `Minimal AI` : AI provides a real-time transcript divided into transcript blocks after each completed speaking turn. Users can select and incorporate these transcript blocks into their notes.

The detailed system features and how the AI assistance differs in the three setups will be discussed in section 3.2

*3.1.2 Participants and Procedure.* We recruited 30 participants through mailing lists at a large public university in the US. We selected participants who indicated that they faced challenges in note-taking. The demographic information of participants are shown in the Table 1

| **Gender** | | **Age** | |
|---|---|---|---|
| Male | 16 | 0 - 20 years old | 9 |
| Female | 9 | 21 - 30 years old | 21 |
| Genderqueer | 1 | | |
| Prefer not to say | 4 | | |
| **Ethnicity** | | **Year in School** | |
| Caucasian | 7 | Freshman | 0 |
| Asian | 19 | Sophomore | 9 |
| Other/Unknown | 4 | Junior | 4 |
| | | Senior | 12 |
| | | Master | 1 |
| | | PhD | 4 |

Table 1. Demographic Information of Participants

[5]https://www.zoom.com/en/ai-assistant/

[6]https://support.microsoft.com/en-us/office/use-copilot-in-microsoft-teams-meetings

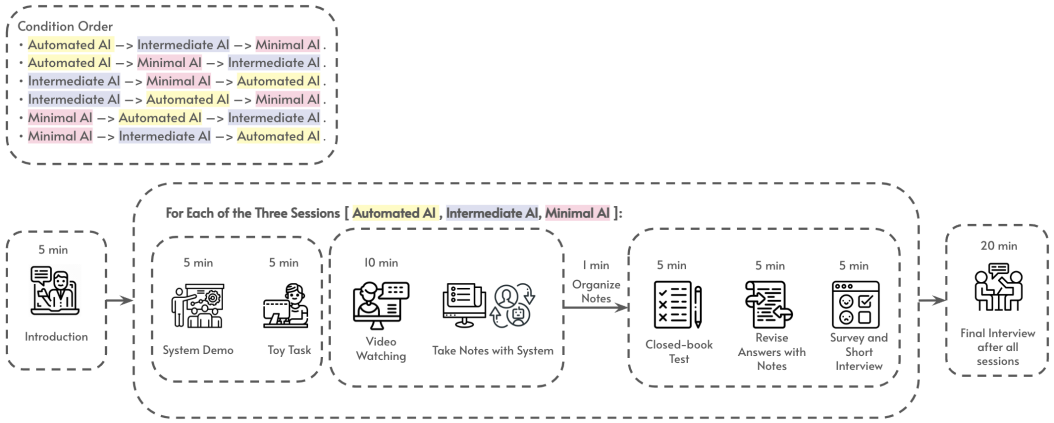


Fig. 2. **Study Design.** The study started with a 5-minute introduction outlining its objectives. Participants then experienced all three AI assistance conditions— `Automated AI` , `Intermediate AI` , `Minimal AI` , in a specified order. Each session began with a demo instruction and a toy task. Participants watched a 10-minute video while taking notes with the system, after which they had 1 minute to organize their notes. This was followed by a closed-book test lasting 5 minutes, after which participants revised their answers with their notes for an additional 5 minutes. Each session concluded with a survey and a brief interview lasting 5 minutes. A final interview was conducted after all sessions, lasting 20 minutes.

Each participant experienced all three AI assistance conditions in person, with the conditions— `Automated AI` , `Intermediate AI` , `Minimal AI` . The order of the conditions is counterbalanced across participants to mitigate order effects. Participants were randomly assigned to one of six possible sequences to ensure balanced exposure, as shown in Figure 2

The experiment began with an introduction to the study, providing participants with a thorough explanation of the tasks. A demo was conducted for each system variant, after which participants engaged in a toy task designed to familiarize them with the interface and functionalities of each system.

For each task, participants were seated in a study room with a large display simulating a classroom setting. The display played a 10-minute lecture video, mimicking a live classroom environment where the instructor delivered the lecture. During this time, participants used their own computers to take notes in real time through the customized AI-assisted system. To replicate the demands of in-situ note-taking, the video was designed to be non-replayable.

At the end of each video, participants were given one minute to finalize their notes before proceeding to a post-test, which included two multiple-choice questions (MCQs) and three open-ended questions (OEQs). Initially, participants completed the test without referring to their notes to assess the influence of the encoding process—the act and process of note-taking—on learning outcomes [29]. They then submitted their responses and were asked to revisit and revise their answers with access to their notes, aligning with prior research on the storage paradigm, where notes serve as external storage for review and reinforcement [29]. Following the post-test, participants completed a post-task survey assessing their cognitive load and user experience with the system. Additionally, each task concluded with a 10-minute interview to capture immediate feedback and insights on the specific condition they had just completed.

Upon completion of all three tasks, participants participated in a final 20-minute comprehensive interview to discuss their comparative perceptions of the three AI assistance levels. Questions were

tailored to explore participants' interactions with intermediate and structured AI-generated notes, their cognitive demands and agency during the process, and the trustworthiness of the AI.

In total, the study took approximately two hours per participant.

*3.1.3 Material.* We selected three lecture videos from YouTube, all presented by the same instructor and centered on an engaging, timely topic: large language models (LLMs). Each video spans approximately 8 minutes and 50 seconds to 9 minutes and 30 seconds. The topics covered are "Introduction to AI-Augmented Decision-Making," "Introduction to LLM Hallucination," and "Introduction to LLM model selection framework for your task," with the order fixed as listed. Importantly, the videos required no prior knowledge. Careful consideration was given to the selection and balance of video length and topic, minimizing potential differences in difficulty or familiarity and reducing these variables' impact on learning outcomes.

For each video, we designed a post-test comprising two multiple-choice questions (MCQs) and three open-ended questions (OEQs). An expert instructor in AI developed the tests, with their quality validated in a pilot study involving four students who watched the videos, answered the questions, and provided feedback on any points of confusion. This feedback ensured that the questions were clear and answerable based solely on the lecture content without needing outside resources. Full details on video sources and test questions are provided in the supplementary materials.

We opted not to include a pre-test in this study because the selected lecture videos are closely matched in length, complexity, and topic, minimizing video difficulty as a factor in learning outcomes. While a pre-test might enhance result validity, we excluded it to avoid lengthening the study since students had already experienced three conditions.

### 3.2 NoteCopilot: The Custom-Built AI-assisted Note-taking Tools

We developed NoteCopilot, an AI-assisted note-taking system offering adjustable levels of AI assistance - `Minimal AI` , `Intermediate AI` , `Automated AI` , each tailored to examine different degrees of cognitive assistance in note-taking.

*3.2.1 Common interaction across the three system variants.* The interface includes a Real-Time AI-generated Note Panel for proactive assistance, a Search and Synthesize Panel for expressing specific user intents and retrieving personalized results, and a rich text editor. In the text editor, users can fully interact with text through typing, selection, editing, deletion, and formatting, along with Markdown and other multimedia content enabled. The general interface and user interaction are shown in Figure 3

The Real-Time AI-generated Note Panel (Figure 3(a)) displays AI-generated notes in real-time. Depending on the condition, the content in this panel varies: from full transcripts (in `Minimal AI` ), to intermediate summaries (in `Intermediate AI` ), and structured notes (in `Automated AI` ). This feature enables users to view AI-generated content as the lecture progresses—serving as the system's core AI assistance. Here, AI proactively provides content in a temporary middleware layer, allowing users to opt to integrate it in their own notes or not. This design aligns with prior research that recommends a temporary holding space for AI content rather than direct insertion into the editor [8]

The text editor (Figure 3(c)) offers full note-taking functionality, supporting typing, selection, editing, and multimedia formatting. Users can drag and drop AI-generated content from panels (a) and (b) into the editor (c). When AI-generated content is dragged into the editor, a source button (c2) allows users to view the original transcripts associated with AI summaries, ensuring transparency and traceability of the information.

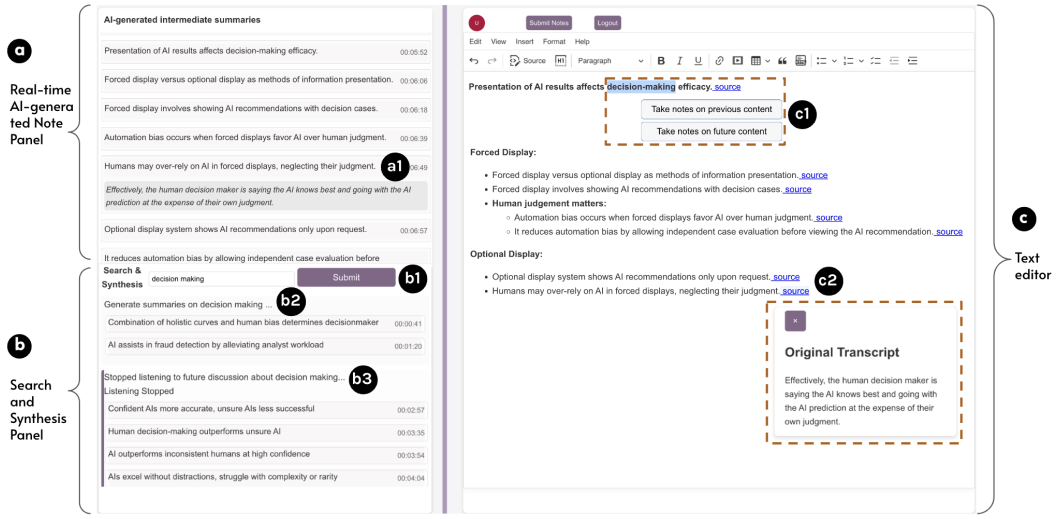


Fig. 3. **System Overview.** On the upper left is the Real-time AI-Generated Note panel (a), where users can see AI-generated notes and click on them to see related transcripts (a1). On the lower left is the Search and Synthesis Panel (b), where users can interact with AI by requesting notes centered around a topic (b1) based on previous course content and receive notes as (b2). On the right is the rich Text-Editor (c), where users can create their own notes or, with the help of AI-generated notes, drag and drop the AI content from (a) or (b) to the editor. They can use in-context selection (c1) to request topic-related notes based on previous course content as (b1) does and also ask AI to continuously generate topic-related notes by listening to future course content and receive results as (b3). In addition, users can check original related transcripts for each AI-generated note by clicking on the source button (c2). This is the interface for `Intermediate AI` .

The above two panels form the system's core mechanism, where AI generates notes in real-time with varying granularity and timing, allowing students the flexibility to selectively incorporate AI-generated content into their notes.

In addition to proactive AI assistance, we offer users the flexibility to search for specific information as needed. The Search and Synthesize Panel (Figure 3(b)) enables users to interact with the AI by requesting targeted information on lecture topics. Users can submit a query (b1), and the AI provides summaries or responses based on their request (b2). Additionally, users can initiate searches directly within the editor (c). By selecting typed content, users can access a drop-down menu with two AI options: "take notes on previous conversation" and "take notes on ongoing conversation" (c1). The former generates AI notes based on previous lecture discussions, while the latter activates a listening mode, capturing any mentions of the selected topic in future content as blocks in the search and summary panel (b2).

The interaction design remains consistent across the three assistance levels, enabling users to interact with the AI panels similarly regardless of the specific condition. This consistency allows us to focus on examining how varying levels of AI processing and automation impact user engagement and learning.

*3.2.2 Three AI-assistance Levels.* While users interact with AI in a similar way, each system variant—`Minimal AI` , `Intermediate AI` , and `Automated AI` —offers unique levels of AI assistance, allowing us to investigate the effects of different assistance levels while keeping user interactions controlled and similar across conditions. The three assistance levels reflect different

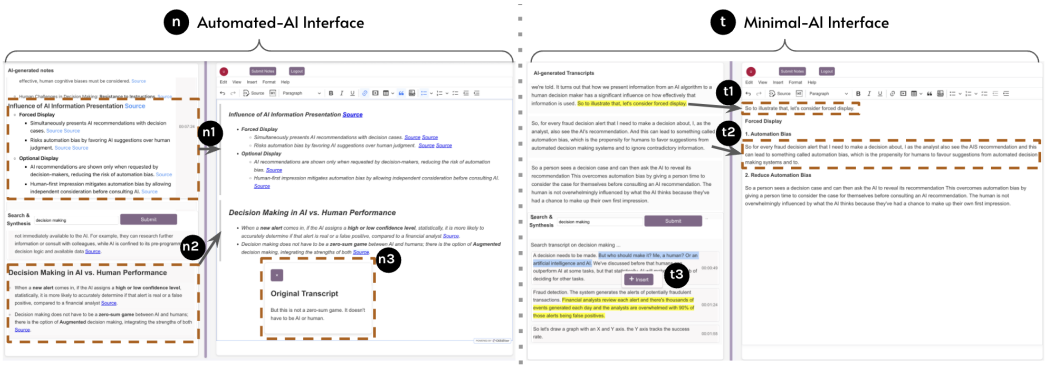


Fig. 4. **`Automated AI` and `Minimal AI` Interface.** On the left is the `Automated AI` Interface (n), where users can drag and drop AI-generated notes alongside related transcripts from the Real-time AI-Generated Note panel (n1) and Search and Synthesis Panel (n2) into the Text-Editor Panel. Users can also expand related transcripts by clicking the source button (n3). On the right is the `Minimal AI` Interface (t), allowing users to directly select text from the Real-time AI-Generated Note panel (t1) and Search and Synthesis Panel (t3) for highlighting and insertion into the Text-Editor Panel. Additionally, users can drag and drop entire transcript blocks into the Text-Editor Panel (t2).

extents to which AI automates the encoding process—the cognitive effort of interpreting, structuring, and recording information, and to which extent AI abstracts the information and adds its own interpretation to the content.

In the `Automated AI` condition, the system generates slightly delayed structured note blocks approximately every 1-2 minutes, which are displayed in the Real-Time AI-Generated Note Panel, as shown in Figure 4 (n1). This interval is chosen to provide users with timely content while allowing AI sufficient time to process and organize content into coherent structures. By offering structured notes, the AI automates much of the encoding process—organizing information, interpreting content, and presenting a synthesized version of the lecture. This design was assumed to reduce the users' note-taking load and may enhance understanding by synthesizing information for them [14, 38].

In the `Intermediate AI` condition, AI generates intermediate summary blocks right after each speaking turn, which appear in the Real-Time AI-Generated Note Panel, as shown in Figure 3(a). These summaries offer a brief interpretation of lecture segments without a fully structured format, providing users with essential information that they can selectively integrate into their notes. The interval between summary blocks is around 15 seconds, where a natural sentence is ended. Users can expand to check the original transcript of the summary block by clicking the block Figure 3(a1). This condition leverages the cognitive benefits of "building blocks" [23]. By using intermediate summaries, this level of support is designed to encourage students to engage with the content actively, reducing the cognitive load associated with verbatim note-taking while still fostering their encoding process through selective integration and organization.

In the `Minimal AI` condition, AI provides real-time transcripts of the lecture, divided into discrete "transcript blocks" based on turns, as shown in Figure 4(t). Users can incorporate these transcript bubbles into their notes through drag-and-drop. They can also select the specific content on one transcript block and click an "insert" button to put it into the editor, shown in Figure 4(t3). This condition represents low AI assistance, where the AI merely acts as a passive transcription tool without interpretation or structuring, thereby requiring the user to handle all aspects of encoding

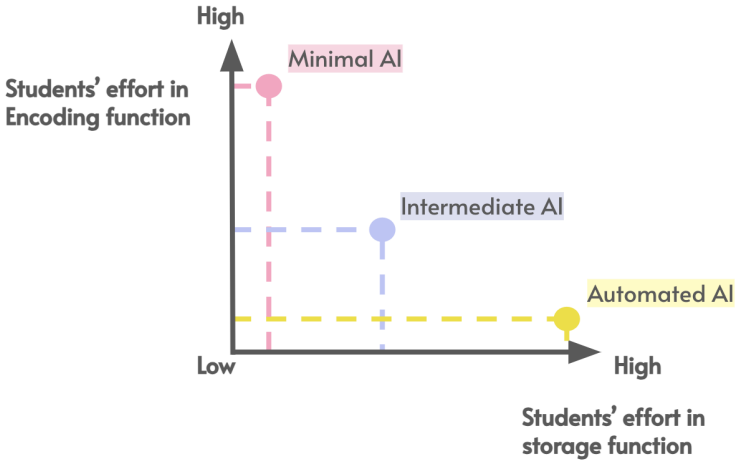


Fig. 5. **Presumed trade-off between students' effort in encoding and storage across AI assistance levels.** `Minimal AI` requires high encoding effort but requires less effort for later reviewing; `Intermediate AI` shares encoding between AI and user, balancing engagement; `Automated AI` minimizes encoding effort, potentially increasing review demands.

and organization. The choice of a `Minimal AI` condition instead of a non-AI condition allows us to control for system interaction across conditions while providing a lightweight AI aid similar to captions on online lectures in a business-usual setup. This design mitigates the potential cognitive overload associated with fully manual note-taking [42], while still offering a light-touch assistance that students can choose to engage with as needed.

We decided against a no-AI control condition because participants can fully control and opt to disregard AI-generated content in any of the three conditions; additionally, the system permits minimizing the AI panel to lessen visual distraction, thus providing the opportunity for a predominantly self-directed note-taking experience if preferred.

*3.2.3 Design Considerations.* We carefully design the three assistance levels in our system that reflect the varied impacts of AI on users' cognitive processes in note-taking, specifically regarding encoding and storage functions [29]. We assume that different levels of AI assistance impact cognitive effort during encoding and may influence demands in the storage phase (review and retrieval), as outlined below.

In the `Automated AI` condition, AI-generated notes aim to reduce encoding effort by automatically structuring information, allowing students more time to listen. Presumably, this reduced engagement in encoding may increase the effort needed for later review during storage. In the `Intermediate AI` condition, AI-generated summaries partially support encoding, reducing effort in capturing information while still requiring students to structure content. This approach aims to share the encoding task between AI and the user, promoting moderate engagement. In the `Minimal AI` condition, minimal AI support assumes higher encoding demands on the user, who captures, selects, and organizes content independently. This might benefit storage, as manually structured notes could better align with the user's mental model, aiding review. The presumed trade-offs of how the AI's assistance in taking away the encoding function may increase human storage effort is shown in Figure 5.

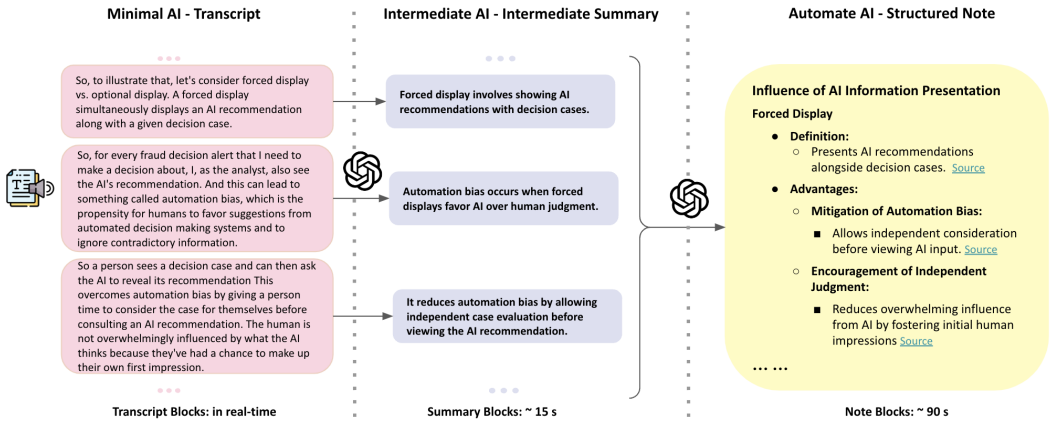


Fig. 6. **The granularity of AI notes in the three conditions.** From left to right, the AI-generated results of three systems—`Minimal AI` , `Intermediate AI` , and `Automated AI` —illustrate the topic of FORCED DISPLAY with increasing levels of AI assistance. The `Minimal AI` system presents users with raw AI-chunked transcripts, providing unfiltered information. In contrast, the `Intermediate AI` system offers concise summaries of each transcript chunk, enhancing clarity and facilitating user comprehension. Finally, the `Automated AI` system consolidates all short summaries into structured notes that include definitions and advantages of FORCED DISPLAY. This organized presentation provides detailed information for each point, allowing users to reference related source transcripts for further context.

To implement the three levels of assistance in the encoding function, we tailored the prompting and granularity of each variant to match the designated assistance level. Across all conditions, transcript and summary blocks are segmented based on natural turns—a complete sentence or a set of related short sentences—ensuring consistent granularity. The `Minimal AI` variant provides a real-time transcript. In the `Intermediate AI` variant, AI generates summaries immediately after each turn, averaging around 15 seconds per summary block. The difference between the two is that the summary blocks are easier to read, whereas transcript blocks are more transparent since they are true representations of the conversation. The `Automated AI` variant combines several turns into topic-based structured note blocks, with each note block between 1-2 minutes and traceable to the original transcript. Although the abstraction level varies, all AI-generated content offers a full representation of the lecture, ensuring comparability across the three system variants.

In our study setup, while our system is designed to offer real-time note support with functionalities built for live AI assistance, we pre-loaded each video's note data with timestamps in the real-time AI-generated note panel. This approach ensures that all participants receive the same content at consistent intervals, eliminating the potential variability of real-time AI performance. However, if participants use the search and synthesis features, the AI dynamically retrieves and synthesizes notes to cater to their specific queries. To produce a well-structured, high-quality summary and notes, researchers exerted significant effort in crafting the prompt to ensure the resulting notes appear reasonable and organized, as illustrated by the AI-generated transcript, summary, and notes shown in Figure 6.

## 4 DATA ANALYSIS AND MEASUREMENT

After collecting the data from N=30 participants, we proceeded to analyze the data. We performed a series of quantitative analyses using several statistical models, as discussed below, to answer the following two questions:

- How do different levels of AI assistance affect learning outcomes, cognitive load, note-taking behaviors, and usability?

We use our interview data to provide further explanations on the following:

- Why do these effects occur, how do students perceive the usefulness of AI, and how do they prefer to interact with different AI assistance levels?

### 4.1 Quantitative Outcome Measures

*4.1.1 Measurement of Understanding.* To assess users' understanding of the lecture, we graded students' post-test answers across two rounds of submissions. The first round, completed without reference to notes, measured students' real-time understanding of the lecture content—referred to as the Post-Test Score. The second round allowed students to revise their initial answers using their notes, called the Revision Score. This score represents the understanding achieved after reviewing and refining their responses with their notes.

We developed a rubric to evaluate open-ended question (OEQ) answers. One author initially created the rubric, with each question worth up to 4 points. Two authors then independently graded a sample (25%) of the students' answers using this rubric, blind to the experimental condition. The initial inter-coder reliability was 86.7%. After discussion and consensus on the grading standards, they divided the remaining answers for individual grading. Each test's total grade was then calculated. The rubric and example students' answers are provided in the B.3

*4.1.2 Measurement of Cognitive load.* We assessed cognitive load by employing questions tailored to evaluate overall mental effort [41], and we adapted scales from earlier studies [31, 35] to measure intrinsic (IL), extraneous (EL), and germane (GL) cognitive loads. We opted not to use a general task load measurement, such as NASA TLX and instead chose a scale specifically designed to measure the three types of cognitive load. This approach allows us to examine how each type of cognitive load is influenced by the three conditions, providing insights into how varying levels of AI assistance can enhance beneficial aspects of cognitive load, like germane load, while mitigating the more negative aspects, such as extraneous load.

*Overall Mental Effort:* Please rate your overall mental effort (from 1 - 9). Scale 1 = very, very low mental effort, and Scale 9 = very, very high mental effort [41]

*Intrinsic Load:* This type of cognitive load refers to the cognitive effort required to understand the inherent complexity of the learning content [31]. It depends on the difficulty of the material itself, independent of the instructional methods. Questions include:

- Q1: "The content explained in this lecture is very complex."
- Q2: "This lecture contains many terms and concepts that are unfamiliar to me."

The score ranged from 1 (strongly disagree) to 7 (strongly agree).

*Extraneous Load:* Extraneous load is the cognitive burden imposed by the instructional methods or tasks, which arise from unnecessary or confusing elements in the learning environment [31]. Questions related to extraneous load include:

- Q3: "The instructor failed to explain concepts clearly."
- Q4: "Manual note-taking required a lot of my effort."
- Q5: "Interacting with the system (e.g., using AI-generated notes) required significant effort."
- Q5: "Creating notes with AI assistance also required significant effort."
- Q6: "Reading AI-generated notes required significant effort."

The score ranged from 1 (strongly disagree) to 7 (strongly agree).

*Germane Load:* Germane load is the cognitive effort devoted to processes that directly support learning, understanding, and knowledge construction [31]. This is a type of beneficial load that helps deepen comprehension and consolidate knowledge. Questions include:

- Q7: "I was fully engaged in this lecture."
- Q8: "Note-taking with AI assistance enhanced my understanding of the lecture."
- Q9: "Reading AI-generated notes enhanced my understanding of the lecture."

The score ranged from 1 (strongly disagree) to 7 (strongly agree).

*4.1.3 Note-taking Behaviors and Outcomes.* Throughout the experiment, we systematically logged users' interactions with the interface. This included actions on the Real-time AI-generated Note Panel, Search and Synthesize Panel, and text-editor. We also recorded the content created in the Text Editor, distinguishing between notes generated by AI and those created by users. By analyzing this log data, we aim to gain a detailed understanding of how participants utilized the AI assistance in their note-taking process and how this influenced their sense-making process.

*Note-Taking Behaviors.* The study tracked several note-taking behaviors, which included:

- Clicking on the Text Editor: Frequency of checking the content in the text editor
- Typing in the Editor: Frequency of typing actions in the text editor
- Dropping AI-Generated Content: Inserting AI-generated content into the editor through drag and drop or 'insert' button
- Query Search: Searching via direct queries in the Search and Synthesize Panel
- In-Context Search for Previous Content: Searching for previously discussed content around the selected texts in the text editor
- In-Context Search for Upcoming Content: Searching for future content around the selected texts in the text editor

*Note Outcome Measures.* The outcome of the note-taking process was measured based on note quantity:

- Total Words in the Editor: The sum of all words, including both AI-generated content used by participants and manually typed content.
- Manually Typed Words: The count of words manually typed by participants in the editor.

This study only measured the quantity of notes and did not assess note quality. We acknowledge this as a limitation of our research and will address it in the discussion.

*4.1.4 Usability.* Usability was measured through a series of subjective usability questionnaire items, including scores for system ease of use, satisfaction, future usage intention, etc.

### 4.2 Interview Analysis

To understand students' perceptions of the usefulness and limitations of the three levels of AI assistance in note-taking, we conducted a thematic analysis of interview transcripts [4]. Initially, two researchers independently reviewed, commented on, and coded the transcripts. They developed 203 initial codes in this phase. They then discussed their findings to reach a consensus, after which one researcher refined the codes and identified key themes across the transcripts. All disagreements between coders were resolved through discussion. We uncovered 8 themes related to the benefits and limitations of each condition, 3 themes on how students used different AI assistance levels in note-taking, and 7 themes covering students' perceptions of agency, intention expression, needs, and trustworthiness. From these themes, we derived 5 high-level findings, which focused on AI's impact on students' learning, when and how students expressed their intentions toward the AI, and students' perceptions of agency and cognitive load.

# 5 FINDING

## 5.1 RQ1: How do different levels of AI assistance affect comprehension, cognitive load, note quality, and usability?

In RQ1, we evaluated how the different levels of AI assistance (`Automated AI` , `Intermediate AI` , `Minimal AI` ) influence students learning and understanding, cognitive load, and usability.

*5.1.1 Understanding Levels.* To evaluate the impact of different levels of AI assistance (`Automated AI` , `Intermediate AI` , `Minimal AI` ) on students' understanding of the lecture, we conducted mixed-effects linear regression analyses on both post-test and revision scores. In these models, we included Condition as a fixed factor, controlled for Video as a covariate, and incorporated interaction terms between Condition and Video. Random intercepts for Individual Participants were added to account for within-subject variability.

| **Coefficient (Std. Err)** | **Post-Test Score** | **Revision Score** |
|---|---|---|
| Intercept (`Automated AI` ) | 9.550 (0.980) | 12.033 (0.922) |
| Condition: `Intermediate AI` | 4.172 **(1.315)** ** | 2.848 **(1.190)** * |
| Condition: `Minimal AI` | 2.702 (1.339) * | 0.367 (1.216) |
| Video: Video 2 | 2.326 (1.338) | 0.565 (1.216) |
| Video: Video 3 | −0.298 (1.368) | −0.278 (1.190) |
| Condition: `Intermediate AI` × Video 2 | −1.146 (1.982) | −0.382 (1.806) |
| Condition: `Minimal AI` × Video 2 | −2.197 (1.974) | 2.528 (1.813) |
| Condition: `Intermediate AI` × Video 3 | −0.136 (1.933) | 1.339 (1.779) |
| Condition: `Minimal AI` × Video 3 | −0.993 (1.997) | −0.538 (1.805) |
| Group Var | 2.216 (0.603) | 4.096 (0.894) |

Table 2. Mixed-effects linear regression model results for Post-Test Scores and Revision Scores, examining the fixed effects of Condition and Video with random intercepts for User. The model controlled for the influence of Video and included interaction terms between Condition and Video. Results indicate that the `Intermediate AI` condition led to significantly higher Post-Test Scores and Revision Scores compared to `Automated AI` , with `Minimal AI` showing a higher Post-Test Score than `Automated AI` , and no significant interactions between Condition and Video. Significance codes: *** p<0.001, ** p<0.01, * p<0.05.

The analysis of Post-Test Scores revealed a significant main effect of `Intermediate AI` on learning outcomes. Specifically, students in the `Intermediate AI` condition (Mean = 13.88, SD = 2.34) scored significantly higher than those in the `Automated AI` condition (Mean = 10.28, SD = 3.75), with a coefficient of 4.17 ( p=0.002). The `Minimal AI` condition also showed a significant improvement over `Automated AI` , with a coefficient of 2.70 ( p=0.044). Post-hoc Tukey HSD analysis further confirmed these differences, showing that `Intermediate AI` outperformed `Automated AI` with a mean difference of 3.61 ( p<0.001), and significantly outperformed `Minimal AI` with a mean difference of 2.03 ( p=0.034).

The analysis of Revision Scores revealed a significant main effect of `Intermediate AI` on post-review learning outcomes. Specifically, students in the `Intermediate AI` condition (Mean = 15.29, SD = 2.99) scored significantly higher than those in the `Automated AI` condition (Mean = 12.12, SD = 2.76), with a coefficient of 2.85 (p=0.017). The `Minimal AI` condition showed no significant improvement over `Automated AI` (Coef = 0.37, p=0.763). Post-hoc Tukey HSD analysis supported these findings, showing that `Intermediate AI` outperformed `Automated AI` with a mean difference of 3.17 (p=0.0003), and significantly surpassed `Minimal AI` with a mean difference of 2.00 (p=0.031).

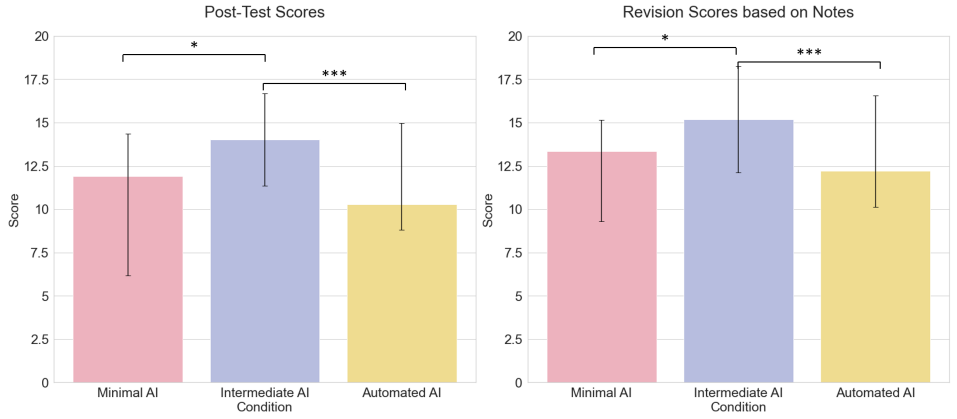


Fig. 7. **Post-hoc Tukey HSD analysis comparing post-test and revision scores across conditions.** The left bar graph shows the Post-test Scores, where the `Intermediate AI` condition significantly outperformed both the `Automated AI` and `Minimal AI` . The right bar graph displays the Revision Scores, with similar trends observed.

These findings reveal that `Intermediate AI` better supports real-time learning (Post-Test Scores) compared to `Automated AI` . However, contrary to our initial expectations, `Intermediate AI` also outperforms `Automated AI` in post-review tasks (Revision Scores). Interestingly, despite the opportunity for students to review with notes `Automated AI` , this condition did not yield better learning outcomes.

No significant difference between Video and the learning outcome was found. Hence, we argue that the difficulty levels of the videos are comparable, so we excluded video differences from further analyses in this paper.

*5.1.2 Cognitive Load.* We analyzed cognitive load across different AI-assisted note-taking conditions ( `Automated AI` , `Intermediate AI` , and `Minimal AI` ) using repeated-measures ANOVA, focusing on dimensions of mental effort, intrinsic, extraneous, and germane cognitive load.

*Overall Mental Effort.* The repeated-measures ANOVA for overall mental effort showed no significant main effect of condition ($F(2, 58) = 1.67$, $p = 0.197$), indicating similar levels of mental effort across conditions.

*Intrinsic Cognitive Load.* For intrinsic cognitive load, measured by perceived complexity and unfamiliarity with lecture content, no significant main effect of condition was observed. These findings suggest that intrinsic load was not significantly impacted by the AI assistance levels, aligning with cognitive load theory, which posits that intrinsic load is driven primarily by the inherent complexity of the content rather than instructional conditions.

*Extraneous Cognitive Load.* Significant effects were observed for extraneous cognitive load across conditions, particularly in manual effort, AI engagement, and reading AI notes. For the effort in manual note-taking, `Minimal AI` (Mean = 3.77, SD = 1.43) required significantly more effort than `Automated AI` (Mean = 2.50, SD = 1.68, $F = 4.39$, $p = 0.017$), indicating that manual note-taking

| Cognitive Load Question | Automated AI | Intermediate AI | Minimal AI | F (Pr > F) |
|---|---|---|---|---|
| Overall: Please evaluate your mental effort in the task | 4.60 (2.08) | 4.53 (1.74) | 5.17 (1.39) | 1.6716 |
| Intrinsic: The learning content covered in this lecture was very complex | 3.03 (1.59) | 2.53 (1.17) | 3.00 (1.14) | 1.9618 |
| Intrinsic: The lecture included many terms and concepts that were unfamiliar to me | 3.13 (1.61) | 3.13 (1.63) | 3.57 (1.65) | 1.1086 |
| Extraneous: The instructor did not explain the concepts clearly in this lecture | 2.40 (1.28) | 2.30 (1.21) | 2.10 (1.03) | 0.7346 |
| Extraneous: It requires a lot of effort for me to write notes by myself. | 2.50 (1.68) | 2.90 (1.69) | 3.77 (1.43) | **4.3899*** |
| Extraneous: Engaging with the system (e.g., using the AI-generated notes) took a lot of effort from me. | 2.87 (1.50) | 3.17 (1.18) | 3.90 (1.65) | **4.3497*** |
| Extraneous: It requires a lot of effort for me to create notes with the AI's help. | 2.87 (1.53) | 2.83 (1.23) | 3.93 (1.76) | **4.7137*** |
| Extraneous: It requires a lot of effort for me to read the AI-generated notes. | 2.93 (1.66) | 3.07 (1.70) | 4.57 (1.61) | **9.2956***** |
| Germane: I am fully engaged during the lecture. | 4.33 (1.53) | 4.77 (1.52) | 4.50 (1.61) | 1.0331 |
| Germane: Taking notes (with the assistance of AI) enhanced my understanding of the lecture. | 4.93 (1.11) | 5.23 (1.04) | 4.43 (1.52) | **4.1359*** |
| Germane: Reading the AI-generated notes enhanced my understanding of the lecture. | 5.23 (1.10) | 4.83 (1.18) | 4.23 (1.50) | **4.4918*** |

Table 3. Repeated Measures ANOVA Results for Cognitive Load Questions across Conditions with Mean (SD) for Each Condition

without AI support increases extraneous load. Similarly, for engaging with the system, Minimal AI (Mean = 3.90, SD = 1.65) required significantly more engagement effort than Automated AI (Mean = 2.87, SD = 1.50, $F = 4.35$, $p = 0.017$). For reading AI-generated notes, Minimal AI (Mean = 4.57, SD = 1.61) required significantly higher reading effort than both Intermediate AI (Mean = 3.07, SD = 1.70, p=0.002) and Automated AI (Mean = 2.93, SD = 1.66, p=0.0007) (F=9.30, p<0.001).

*Germane Cognitive Load.* For germane cognitive load, significant effects were observed for both questions related to the interaction with AI. The question, *"Taking notes (with the assistance of AI) enhanced my understanding of the lecture,"* showed a significant effect of condition ($F = 4.14$, $p = 0.021$), with Intermediate AI (Mean = 5.23, SD = 1.04) significantly outperforming Minimal AI (Mean = 4.43, SD = 1.52, $p = 0.038$). Similarly, *"Reading the AI-generated notes enhanced my understanding of the lecture,"* yielded a significant effect ($F = 4.49$, $p = 0.015$), with Automated AI (Mean = 5.23, SD = 1.10) significantly higher than Minimal AI (Mean = 4.23, SD = 1.50, $p = p = 0.009$).

In summary, Automated AI is optimal for lowering extraneous load, while Intermediate AI supports germane cognitive load by using AI-generated Notes. Conversely, Minimal AI demands greater manual effort, showing higher extraneous load and challenging cognitive resources.

*5.1.3 Note-taking behavior.* We analyzed the note outcomes and note-taking behaviors across different AI-assisted note-taking conditions using repeated-measures ANOVA.

There are notable differences in average total note counts across different AI conditions, as shown in Figure 8. The `Automated AI` condition yielded a significantly higher total note count ( Mean=300.68, SD=160.52) compared to both the `Intermediate AI` condition (Mean=170.46 SD=70.21) and the `Minimal AI` condition (Mean=143.69, SD=86.93). ANOVA results confirmed a significant effect of AI condition on note volume ($F(2,60)=16.66$, $p<0.001$).

To further understand these differences, we classified users into three general note-taking patterns based on their approach to AI assistance: Fully AI Mode, where users only used AI-generated notes without adding their own; Mixed AI+Human, where users combined AI notes with their own manual entries; and Fully Manual, where users relied solely on self-generated notes. The distribution of these strategies across conditions reveals distinct behavioral tendencies. In the `Intermediate AI` condition, over half of the users adopted the Mixed Mode, integrating AI-generated notes with their manual inputs. A clear divergence emerged in the `Automated AI` condition: 10 users chose to take their own notes manually, while 8 relied solely on AI. In the `Minimal AI` condition, users predominantly engaged in fully manual note-taking, with 19 users relying exclusively on human-generated notes and 10 employing a mixed approach.

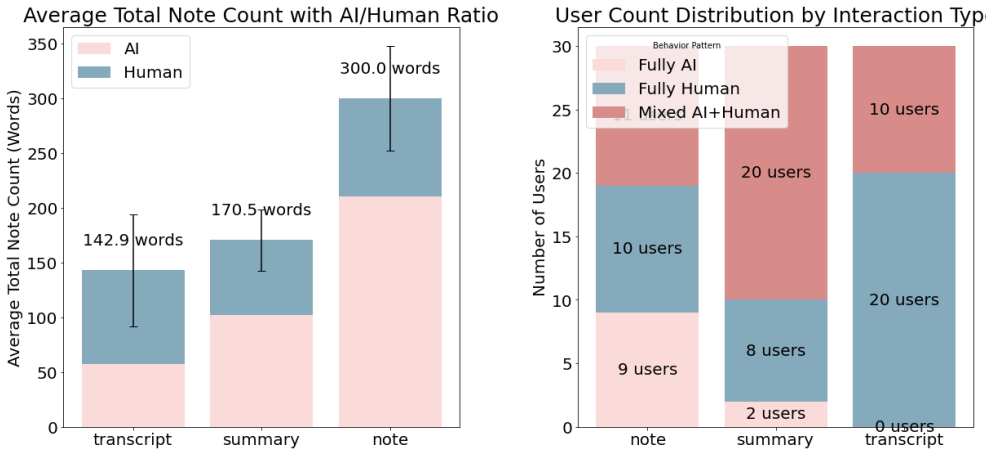


Fig. 8. **Comparison of note-taking behaviors across conditions with varying levels of AI assistance.** The left chart shows the average total note count for each condition, divided into AI-generated content and manually typed content with error bars indicating standard deviation. On the right, a stacked bar chart illustrates the distribution of users by interaction type (fully AI-generated, mixed AI-human, or fully manual) across conditions, with labels indicating the number of users per type.

Further analysis of interaction behaviors in each condition offered insights into users' note-taking engagement. Users in the `Intermediate AI` condition frequently interacted with AI-generated notes, as indicated by the highest click counts in the text editor (mean = 37.41, SD = 24.86) compared to the `Minimal AI` condition (mean = 28.14, SD = 18.47, $p=0.049$). Typing frequency was also higher in the `Intermediate AI` condition (mean = 37.59, SD = 13.33), significantly surpassing `Minimal AI` ($p=0.008$). AI content "drops" (when AI-generated text was selectively added to the editor) were also most frequent in the `Intermediate AI` condition (mean = 8.03, SD = 7.51), demonstrating a more active and curated approach than `Automated AI` ( $p<0.001$) and the `Minimal AI` condition ($p<0.001$). Interaction with initiating a query and constantly listening to the upcoming content was notably higher in the `Intermediate AI` condition (mean = 7.38, SD = 30.44), compared to that in `Automated AI` (mean = 1.93, SD=4.96), ($p<0.001$).

| Behavior | Minimal AI | Intermediate AI | Automated AI | F (Pr>F) |
|---|---|---|---|---|
| **Manual type words** | 110 (83.58) | 86.58 (56.01) | 94.46 (61. 62) | F = 1.42 |
| **Click Text Editor** | 28.14 (18.47) | 37.41 (24.86) | 32.07 (24.46) | F = 3.18 * |
| **Times for Typing words** | 30.21 (15.59) | 37.59 (13.33) | 28.97 (12.12) | F = 0.86 |
| **Drop AI Content** | 1.14 (1.94) | 8.03 (7.51) | 3.00 (2.93) | F = 4.58 * |
| **Search through query** | 0.66 (1.26) | 0.55 (1.59) | 0.52 (1.30) | F = 0.66 |
| **In-context search** | 0.24 (0.64) | 0.14 (0.44) | 0.00 (0.00) | F = 1.42 |
| **Listen upcoming content** | N/A | 7.38 (30.44) | 1.93 (4.96) | N/A |

Table 4. Summary of Note-Taking Behaviors During the Lecture Stage Across Conditions, with ANOVA Results for Each Behavior

*5.1.4 Usability.* We used repeated measures ANOVA to assess usability across different levels of AI assistance, with post-hoc Tukey HSD tests to identify specific pairwise differences between conditions.

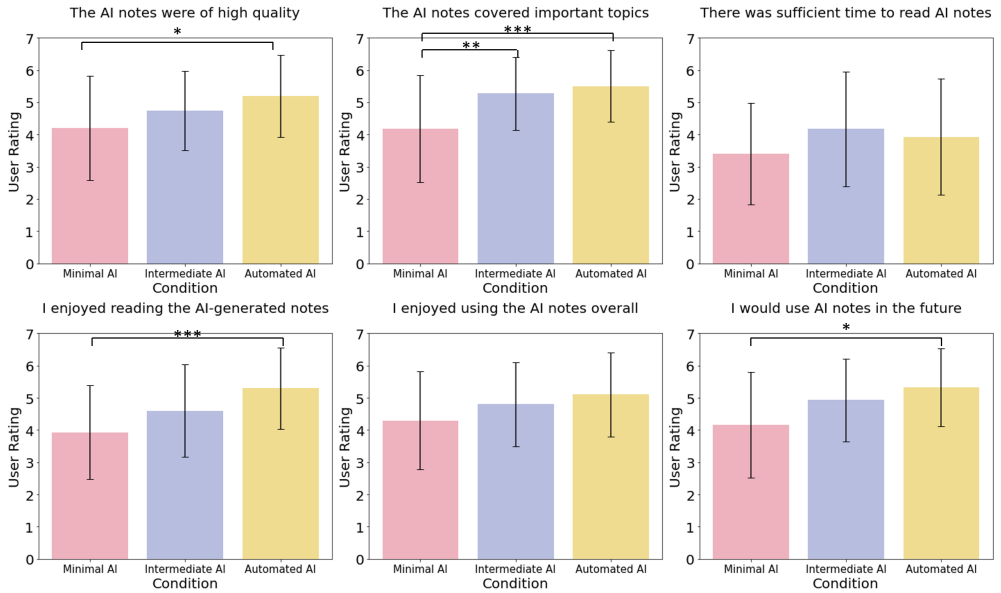


Fig. 9. Usability Rating. This figure presents the mean usability ratings across three conditions—Minimal AI , Intermediate AI , and Automated AI —for six questions assessing different aspects of the AI-generated notes. Tukey's HSD pairwise comparisons were conducted after the repeated measure ANOVA. .p-values are marked as follows: *p < 0.05, **p < 0.01, ***p<0.001. The non-significant difference was annotated as NS

In the Automated AI condition, participants rated the perceived quality of AI-generated notes significantly higher (Mean = 5.20, SD = 1.27) compared to the Minimal AI condition (Mean = 4.20, SD = 1.61), with $F = 3.95, p = 0.025$. Additionally, for coverage of important topics, Automated AI (Mean = 5.50, SD = 1.11) was rated significantly higher than Minimal AI (Mean = 4.17, SD = 1.66), with $F = 9.50 < 0.001$. These results suggest that users found Automated AI 's output of notes more effective in capturing key points. Enjoyment in reading AI-generated notes was also substantially higher in the Automated AI condition (Mean = 5.30, SD = 1.26) compared to Minimal AI (Mean = 3.93, SD = 1.46), $F = 8.18, p < 0.001$. Furthermore, the intention for future

use was notably higher in the Automated AI condition (Mean = 5.33, SD = 1.21) compared to Minimal AI (Mean = 4.17, SD = 1.64), $F = 5.62$ $= 0.006$.

In the Intermediate AI condition, participants rated coverage of important topics significantly higher (Mean = 5.27, SD = 1.14) than in Minimal AI ($p = 0.005$). Besides, Minimal AI (Transcript) condition shows a higher preference for taking manual notes (Mean = 3.80, SD = 1.56) than those in the Automated AI (Auto-Note) condition (Mean = 2.63, SD = 1.45), with $p = 0.01$, as shown by post-hoc Tukey HSD tests. This suggests that users prefer manual note-taking more when AI is seen as not usable. We didn't see other difference in the usability rating between Intermediate AI and Minimal AI or between Intermediate AI and Automated AI

In summary, Automated AI demonstrates optimal usability, significantly enhancing perceived quality, topic relevance, and enjoyment in reading AI-generated notes. Intermediate AI offers a balanced usability experience, showing improved coverage of important topics over Minimal AI , though it does not reach the high preference and engagement levels of Automated AI . Conversely, Minimal AI has lower usability ratings, with a higher preference for manual note-taking, suggesting that increased manual effort may detract from the perceived utility and enjoyment of AI-assisted notes.

### 5.2 RQ2. How Do Students Perceive AI's Helpfulness in Note-Taking?

For RQ2, we conducted a qualitative analysis to explore how students perceive the helpfulness of AI in the note-taking process, focusing on when specific AI conditions were considered useful or not, and why certain conditions were ultimately less utilized despite initial perceptions.

*5.2.1* Intermediate AI *enhanced students' learning by fostering active engagement with synchronous and digestible information.* Students found the summaries valuable for two main reasons. First, the summaries scaffolded their cognitive engagement. Beyond serving as a "constant flow" to help them stay focused on lecture content in real-time (P13, P15), they felt that interacting with summary blocks actively deepened their understanding. This involved *"quickly glance over the summaries to check if key points were there"*(P5), *"which summaries are relevant and which ones they need"*(P3), and *"trying to put it under my own structure"*(P1). As P3 said, *"I'm just gonna drag the summaries of the last few things that you talked about onto here, and then I can kind of piece that together, and then that catches me up at least a little bit more than I normally would, where I would just miss that whole chunk of information."* As a result, students reported substantial learning gains in the Intermediate AI condition (P1, P13, P15). For example, P15 mentioned that, despite the lecture topic being the least familiar to them, they achieved their best understanding of the content with the help of Intermediate AI , compared with other conditions.

The concise, digestible format of the Intermediate AI summaries makes engagement possible in real-time lectures. P2 remarked that the summaries helped them focus on the lecture by breaking down information into *"digestible chunks that were easy to put into the notes."* The granularity of the summaries matched many students' natural note-taking styles, allowing seamless integration into their notes. As students highlighted, the summaries consisted of short sentences they would typically write, enabling them to *"catch up by just dragging and dropping the summaries"* (P15) and do so *"without needing to add further details"* (P20).

*5.2.2 The* Automated AI *was perceived beneficial for its well-structured content, but it also led to feelings of disengagement and inflexibility.* Many students perceived Automated AI as valuable during lectures due to their clear and logical structure. For instance, P3 and P13 noted that the structured format *"organized and contextualized the information better than other conditions."* P6 appreciated the structured output, especially when they lost track of lecture steps, stating, *"the system provided a structured output, and it is very short, and I can read it and get on with it."* Similarly,

the structure allowed P4 to *"work at their own pace and reference it later."* Students also highlighted the value of structured notes for post-lecture review and information retrieval. For example, P12 used the notes to confirm their understanding of concepts, while P15 used them to recall connections between topics. Due to the clear structure, students *"knew exactly where they could find a piece of information in the notes"* (P11) and could *"easily find a relevant bullet point by scrolling through the notes"* (P4).

However, structured content and the seemingly complete nature of notes are perceived to lead to less cognitive engagement.P3 admitted to *"zoning out"* at times, despite recognizing that the topic required their full attention. Some students felt retained only a general impression of the course content due to over-relying on the notes, as noted by P5,*"I was struggling with answering the questions because I thought notes would help, but then I couldn't remember the steps, so I needed to look back at my notes for this question. I felt like some sort of disconnection."* Similarly, P7 thought heavily relying on the AI for note-taking *"didn't really encode or process that information,"*, making it difficult to retrieve later. Lastly, the granularity of the notes presented challenges; P2 found it difficult to work with, commenting, *"everything was displayed in a giant chunk,"* which hindered their ability to integrate information into their existing notes.

*5.2.3 Students valued manual note-taking as a learning opportunity and preferred maintaining control over their own notes.* Students value the agency they have over their notes. Most students viewed AI assistance as a *"backup"* (P20) or *"augmentation"* (P8) rather than a replacement, using it selectively for tasks like verifying information (P5), catching up on missed points (P10), and enhancing note completeness (P20). P3 even felt a higher sense of control with AI, as it allowed them to focus on note-taking by *"getting rid of busy work"*.

Although students found AI support helpful, they emphasized the importance of manual note-taking as a critical learning opportunity. Writing original notes allowed them to *"memorize better"* (P11, P12), *"avoid distraction and keep up with the lecture"* (P13), *"engage actively in the lecture instead of passively copying information"* (P15), and *"gain a deeper understanding of concepts"* (P20). Interestingly, P7 likened manual note-taking to hands-on dissection in a behavioral neuroscience class, explaining that this active engagement helped them retain details more effectively for exams.

In the `Automated AI` condition, where AI generated structured notes, some students (10/30) preferred to maintain their own notes rather than use the AI-generated content and fount it *"easier to retrieve information"* when notes followed their own structure (P4). However, some students (9/30) relied heavily on AI-generated notes in this condition. For example, P3 and P7 admitted to *"zoning out until* `Automated AI` *sent them bullet points,"*. P12 felt that *"the AI almost replaced their own notes,"* leading them to take notes less actively and lose the agency they have over their notes.

*5.2.4 Students expressed intentions of asking for real-time clarification, post-lecture review, and verification through active searching.* When students felt that the AI-generated content didn't align with what they wanted in their own notes, they often needed to search for additional information or prompt the AI to refine the output. We present findings on when these needs typically arise.

Participants reported using AI assistance to help *"capture missed points when the lecture went fast"* (P6) and to *"efficiently access specific information"* when the sheer volume of notes felt overwhelming (P4). After the lecture, students sought more information from the AI to clarify concepts they didn't fully understand or recall. For instance, P1 noted, *"I think the notes didn't really go deep into its pros and cons, so I asked it to generate more notes on this."*

Interestingly, some students used the search feature not just to fill gaps but also to *"verify the accuracy of AI-generated notes"*—even for concepts they felt they already understood. In the `Automated AI` condition, P5 searched for a specific concept to assess the AI's accuracy, but when the result didn't meet expectations, it reduced their trust in the AI. Additionally, students valued

the ability to perform *"in-contextual searches"*; as P28 noted, *"listening to the future makes sense since I can selectively get some results related to that topic constantly, which saves my time reading."*

While the search feature enabled active engagement, some students expressed concern about the time trade-off. P10 remarked, *"It took a long time to search things, and I can't waste that time during the lecture."* Additionally, students felt uncertain when searching for notes on future topics, unsure whether the instructor would actually cover them. P1 described this uncertainty: *"I'm not sure whether the teacher will really talk about it."*

*5.2.5* `Minimal AI` *has a higher cognitive load but also retains the highest level of trust.* Many students experienced the highest cognitive load during the `Minimal AI` condition. Firstly, the *"large chunks of text"* made it challenging to parse and transfer information (P13). Secondly, students felt distracted by the constant need to switch between interpreting the transcript, editing, and listening to the lecture. For example, P12 found it *"time-consuming to decipher the transcript while trying to keep up with the lecture,"* causing them to fall behind.

However, in this condition, the AI was often seen as a *"shadow companion for real-time understanding,"* helping students capture specific terminology or follow along with the lecture content as closely as possible. Since the AI only performed basic grammatical and punctuation corrections, students felt that it *"reflected the lecture content as closely as possible"* (P29) and *"did not add the cognitive burden of learning additional AI-generated content"* (P17). Compared to the other two conditions, although the cognitive load was highest here, the AI rarely introduced misunderstandings. Students noted that it *"helped them understand the material to the fullest extent"* (P9).

In contrast, the other two conditions (`Automated AI` , `Intermediate AI` ) occasionally led to AI-generated misinterpretations. Users remarked that they felt *"forced to adopt a new logic to interpret the AI content, which was impossible when I hadn't fully understood the lecture in the first place"* (P19).

## 6 DISCUSSION

This study combines quantitative and qualitative data to examine the effects of different levels of AI assistance on comprehension, cognitive load, usability, and note-taking behaviors. Our research centers around the questions of whether high levels of AI assistance harm students' cognitive engagement and comprehension and whether providing intermediate outputs can mitigate some of these negative effects. Our findings indicate that while `Automated AI` offers the highest perceived usability and lowest cognitive load, it also results in the poorest comprehension. In contrast, AI assistance in the `Intermediate AI` condition effectively mitigates the potential negative impact of AI on comprehension, with students benefiting from interacting with AI and perceiving this interaction as helpful for learning. We discuss our results with note-taking, cognitive load, and learning theories to explain the mechanisms behind this trend. We also discuss the implications of designing intermediate AI to support human cognition.

### 6.1 Unpacking the Paradox: Reduced Effort v.s. Enhanced Comprehensions

This section aims to discuss our core findings by combining quantitative and qualitative analyses and relevant theories. This discussion seeks to understand why `Intermediate AI` can promote comprehension and support cognitive processes to a certain extent, what mechanisms make it useful, and why `Automated AI` , despite seeming to ease the burden on students, does not truly aid comprehension and cognitive processes and, in some cases, even hinders them.

Our findings indicate that students in the `Intermediate AI` condition scored the highest on both post-lecture and review-modified comprehension tests, while `Automated AI` scored the lowest across both measures (Table 2. In real-time conditions, `Automated AI` performed even

worse than `Minimal AI` . The poor real-time performance of `Automated AI` may relate to the lack of encoding function of note-taking [29]. Students actively engaged in note-taking in the other two conditions, which likely enhanced encoding through greater participation. Our qualitative analysis revealed that students remained involved in encoding information themselves, with `Intermediate AI` offering a scaffold for metacognition [52, 64], since students scanned the AI-generated content, assessed its relevance, and considered how to incorporate it structurally when using AI. In the `Minimal AI` condition, students' real-time comprehension was also relatively high, likely because the transcript acted as a "shadow" assistant, helping students track information even without AI summarization, particularly when unfamiliar terms were mentioned. Transcripts also catered to students who needed more reading support rather than purely auditory input to aid comprehension [1].

A counterintuitive finding was that even after students used the notes to adjust their final scores, `Automated AI` still performed worse. This contrasts with students' intuitive expectations, as most interviewees reported feeling that structured notes provided the most complete and organized information for rapid retrieval during the review phase. We believe this finding aligns with our design rationale: when AI takes over much of the encoding process, it may effectively reduce the burden of encoding but also increase the challenge of interpreting and making sense of this information during storage, as discussed in 5. This also ties in with theories in collaborative sense-making that suggest that when the process of instantiating external representations is largely taken over by collaborators, it can make consuming encodons more difficult [47]. Our findings suggest that when note-taking is heavily replaced by AI, even if the generated content is comprehensive and high quality, it may not effectively support comprehension, aligning with prior work discussing the agency in collaborative note-taking [9].

Our findings also emphasize the impact of different levels of AI assistance on cognitive load in real-time lectures. Here, we particularly discuss the observed disconnect between the perceived low burden and usability versus students' reflections on their actual cognitive engagement after the post-test. Our data show that while overall cognitive load did not differ significantly, the perceived extraneous cognitive load was lowest in the `Automated AI` condition (Table 3. The transcript condition reported the highest extraneous load, as students either relied entirely on themselves or tried to locate specific content within the lengthy transcript. However, as cognitive load theory suggests, cognitive load is not always detrimental [31, 42]; for instance, germane cognitive load is considered beneficial for learning [31, 35]. Although students reported that interacting with AI in the `Intermediate AI` condition required effort, they perceived this effort as enhancing their understanding (germane load). Additionally, while students in the `Minimal AI` condition reported that note-taking required considerable effort, many particularly valued the interoperability of the transcript, which they felt best retained lecture information without adding another layer of information. In interviews, students noted that the low perceived load in the `Automated AI` condition likely stemmed from reduced engagement. Many took their own notes initially, but after viewing the first AI note, they felt it was sufficient and stopped recording—creating an illusion that the AI notes reflected their learning. Only during questioning did they recognize gaps in their understanding.

The implications of these findings suggest that we need to be more cautious when providing AI assistance to users, especially on tasks that require substantial cognitive effort. As users naturally prefer to minimize effort, the risk of over-reliance on AI becomes even more pronounced. This insight also calls for reconsidering how we measure the effectiveness of AI tools for cognitive tasks. Many studies use task load surveys to measure the amount of load and assume that a lower load is better [21]. However, if these tools are designed for cognitive tasks, such as learning,

brainstorming, or meetings, a lower load does not necessarily indicate better outcomes. More appropriate measurements should be considered.

### 6.2 Design Moderate AI to Support Human Cognition

This section discusses the design insights from our study, particularly on how moderate AI assistance (like `Intermediate AI` ) can effectively support students' cognitive processes.

Our study shows that `Intermediate AI` significantly promotes student understanding and maintains cognitive engagement. The success of `Intermediate AI` can be attributed to its design as a "middle layer," where AI-generated "building blocks" allow students to integrate relevant information into their notes in real-time. This "middleware" design, also seen in MeetMap's content holding area [8], encourages users to decide which AI suggestions are relevant. This approach preserves student agency: rather than passively receiving information, students are prompted to make decisions about the value and organization of AI content, a process that keeps their meta-cognitive skills active [52].

Furthermore, in cognitive tasks, we believe that a key design principle for providing moderate AI assistance lies in identifying the task's most creative and thought-intensive aspects and reserving these for the user. For example, clarifying ideas in writing [12] and keeping people intentional in meetings [49] are cognitively valuable activities that should be human-led. In contrast, more mechanical, repetitive tasks, like typing words and keeping meeting records, can be handled by AI, allowing users to focus on the meaningful and higher-level aspects of their work.

Additionally, our findings reveal the need for adaptable AI assistance that adjusts to different levels of course difficulty. Students expressed a strong preference for AI that adapts to varying content demands. They also emphasized the need for traceability, such as quick access to information sources, which enhances trust and transparency and aids in verifying AI-generated content [63].

Finally, aligning with user intent is crucial. Our system allows students to initiate searches in different ways. Some students favored quick keyword searches, while others preferred in-context content expansion for continuity. Future research could explore how AI can more naturally detect and respond to user needs through behaviors like pauses or typing cues, optimizing support in note-taking contexts.

### 6.3 Limitation

We acknowledge the limitations of this work. 1) Our sample size was limited, which also prevented us from using more complex models to explore relationships between variables. 2) The videos used in this study were only ten minutes long, which may not fully capture the impact of AI assistance over extended lectures or in real, long-term courses. Future work could address this with a larger sample, deploying a real classroom setting. Although we controlled for video difficulty and ensured the content was accessible without prior knowledge, our participants were from different academic years with varying knowledge levels. This may have affected their final scores; however, we mitigated this by treating individual differences as a random effect in our model, which likely minimized its influence. Future research could incorporate pre-and post-tests to measure learning gains more accurately.

## 7 CONCLUSION

As AI tools become integral to tasks requiring high cognitive effort, such as note-taking, questions arise about their effects on cognitive engagement. This paper explores the "AI Assistance Dilemma" in note-taking by studying the impact of different AI support levels on user engagement and comprehension through a within-subject experiment where participants (N=30) took notes during lecture videos under three conditions: `Automated AI` (high assistance with structured

notes), `Intermediate AI` (moderate assistance with real-time summary), and `Minimal AI` (low assistance with transcript); results indicate that Intermediate AI led to the highest post-test scores, while Automated AI resulted in the lowest, although participants favored Automated AI for its perceived ease and lower cognitive effort, highlighting a potential mismatch between preferred convenience and cognitive benefits. Our study provides insights into designing AI systems that maintain cognitive engagement, with implications for developing moderate AI support in cognitive tasks.

## REFERENCES


[1] Riad S Aisami. 2015. Learning styles and visual literacy for learning and performance. *Procedia-Social and Behavioral Sciences* 176 (2015), 538–545.

[2] Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. 2023. Summaries, Highlights, and Action items: Design, implementation and evaluation of an LLM-powered meeting recap system. *arXiv preprint arXiv:2307.15793* (2023).

[3] Aaron Bauer and Kenneth R Koedinger. 2007. Selection-based note-taking applications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 981–990.

[4] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[5] Dung C. Bui, Joel Myerson, and Sandra Hale. 2013. Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology* 105, 2 (2013), 299–309. https://doi.org/10.1037/a0030367

[6] Ting-Ju Chen. 2021. *Association, Reflection, Stimulation: Problem Exploration in Early Design through AI-Augmented Mind-Mapping*. Thesis. https://oaktrust.library.tamu.edu/handle/1969.1/195385 Accepted: 2022-01-27T22:18:02Z.

[7] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. Meetscript: designing transcript-based interactions to support active participation in group video meetings. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.

[8] Xinyue Chen, Nathan Yap, Xinyi Lu, Aylin Gunal, , and Xu Wang. 2025. MeetMap: generating dialogue map as cognitive scaffolds in meetings. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2025), 1–28.

[9] Jamie Costley, Matthew Courtney, and Mik Fanguy. 2022. The interaction of collaboration, note-taking completeness, and performance over 10 weeks of an online course. *The Internet and Higher Education* 52 (2022), 100831.

[10] Jamie Costley and Mik Fanguy. 2021. Collaborative note-taking affects cognitive load: the interplay of completeness and interaction. *Educational Technology Research and Development* 69 (2021), 655–671.

[11] Richard C Davis, James A Landay, Victor Chen, Jonathan Huang, Rebecca B Lee, Frances C Li, James Lin, Charles B Morrey III, Ben Schleimer, Morgan N Price, et al. 1999. NotePals: Lightweight note sharing by the group, for the group. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 338–345.

[12] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[13] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel P. Robert. 2024. Shaping Human-AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. http://arxiv.org/abs/2402.11723 arXiv:2402.11723 [cs].

[14] Jingchao Fang, Yanhao Wang, Chi-Lan Yang, Ching Liu, and Hao-Chuan Wang. 2022. Understanding the Effects of Structured Note-taking Systems for Video-based Learners in Individual and Social Learning Contexts. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (Jan. 2022), 1–21. https://doi.org/10.1145/3492840

[15] Jingchao Fang, Yanhao Wang, Chi-Lan Yang, and Hao-Chuan Wang. 2021. NoteCoStruct: Powering online learners with socially scaffolded note taking and sharing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[16] Mik Fanguy, Matthew Baldwin, Evgeniia Shmeleva, Kyungmee Lee, and Jamie Costley. 2023. How collaboration influences the effect of note-taking on writing performance and recall of contents. *Interactive Learning Environments* 31, 7 (Oct. 2023), 4057–4071. https://doi.org/10.1080/10494820.2021.1950772

[17] Abraham E. Flanigan and Scott Titsworth. 2020. The impact of digital distraction on lecture note taking and student learning. *Instructional Science* 48, 5 (Oct. 2020), 495–524. https://doi.org/10.1007/s11251-020-09517-2

[18] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–38.

[19] Nitesh Goyal, Gilly Leshed, and Susan R. Fussell. 2013. Effects of visualization and note-taking on sensemaking and analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France,

2721–2724. https://doi.org/10.1145/2470654.2481376

[20] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K Kummerfeld, and Elena L Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.

[21] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[22] Ken Hinckley, Shengdong Zhao, Raman Sarin, Patrick Baudisch, Edward Cutrell, Michael Shilman, and Desney Tan. 2007. InkSeine: In Situ search for active note taking. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 251–260.

[23] Xinying Hou, Barbara Jane Ericson, and Xu Wang. 2022. Using Adaptive Parsons Problems to Scaffold Write-Code Problems. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*. ACM, Lugano and Virtual Event Switzerland, 15–26. https://doi.org/10.1145/3501385.3543977

[24] Faria Huq, Abdus Samee, David Chuan-en Lin, Xiaodi Alice Tang, and Jeffrey P Bigham. 2024. NoTeeline: Supporting Real-Time Notetaking from Keypoints with Large Language Models. *arXiv preprint arXiv:2409.16493* (2024).

[25] Renée S Jansen, Daniel Lakens, and Wijnand A IJsselsteijn. 2017. An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review* 22 (2017), 223–233.

[26] Slava Kalyuga. 2009. Adapting levels of instructional support to optimize learning complex cognitive skills. In *Managing Cognitive Load in Adaptive Multimedia Learning*. IGI Global, 246–271.

[27] Matthew Kam, Jingtao Wang, Alastair Iles, Eric Tse, Jane Chiu, Daniel Glaser, Orna Tarshish, and John Canny. 2005. Livenotes: a system for cooperative and augmented note-taking in lectures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 531–540.

[28] Anam Ahmad Khan, Sadia Nawaz, Joshua Newn, Ryan M. Kelly, Jason M. Lodge, James Bailey, and Eduardo Velloso. 2022. To type or to speak? The effect of input modality on text understanding during note-taking. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3501974

[29] Kenneth A. Kiewra. 1989. A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review* 1, 2 (June 1989), 147–172. https://doi.org/10.1007/BF01326640

[30] Kenneth A. Kiewra, Nelson F. DuBois, David Christian, Anne McShane, Michelle Meyerhoffer, and David Roskelley. 1991. Note-taking functions and techniques. *Journal of Educational Psychology* 83, 2 (June 1991), 240–245. https://doi.org/10.1037/0022-0663.83.2.240 Publisher: American Psychological Association.

[31] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology* 8 (2017), 294028.

[32] Kenneth R. Koedinger and Vincent Aleven. 2007. Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educational Psychology Review* 19, 3 (Sept. 2007), 239–264. https://doi.org/10.1007/s10648-007-9049-0

[33] Anastasia Kuzminykh and Sean Rintel. 2020. Classification of Functional Attention in Video Meetings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376546

[34] Hui Yi Leong, Yi Fan Gao, Shuai Ji, Bora Kalaycioglu, and Uktu Pamuksuz. 2024. A GEN AI Framework for Medical Note Generation. *arXiv preprint arXiv:2410.01841* (2024).

[35] Jimmie Leppink, Fred Paas, Cees PM Van der Vleuten, Tamara Van Gog, and Jeroen JG Van Merriënboer. 2013. Development of an instrument for measuring different types of cognitive load. *Behavior research methods* 45 (2013), 1058–1072.

[36] Susan Lin, Jeremy Warner, JD Zamfirescu-Pereira, Matthew G Lee, Sauhard Jain, Shanqing Cai, Piyawat Lertvittayakumjorn, Michael Xuelin Huang, Shumin Zhai, Björn Hartmann, et al. 2024. Rambler: Supporting Writing With Speech via LLM-Assisted Gist Manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–19.

[37] Ching Liu, Chi-Lan Yang, Joseph Jay Williams, and Hao-Chuan Wang. 2019. Notestruct: Scaffolding note-taking while learning from online videos. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[38] Tamas Makany, Jonathan Kemp, and Itiel E. Dror. 2009. Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology* 40, 4 (July 2009), 619–635. https://doi.org/10.1111/j.1467-8535.2008.00906.x

[39] Bruce M McLaren, S Lim, and Kenneth R Koedinger. 2008. When and how often should worked examples be given to students? New results and a summary of the current state of research. In *Proceedings of the 30th annual conference of the cognitive science society*. 2176–2181.

[40] Andrew M Mcnutt, Chenglong Wang, Robert A Deline, and Steven M. Drucker. 2023. On the Design of AI-powered Code Assistants for Notebooks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3580940

[41] Fred GWC Paas and Jeroen JG Van Merriënboer. 1993. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors* 35, 4 (1993), 737–743.

[42] Annie Piolat, Thierry Olive, and Ronald Kellogg. 2005. Cognitive effort during note taking. *Applied Cognitive Psychology* 19 (April 2005), 291–312. https://doi.org/10.1002/acp.1086 270 citations (Crossref) [2024-01-03].

[43] Yi Ren, Yang Li, and Edward Lank. 2014. InkAnchor: enhancing informal ink-based note taking on touchscreen mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1123–1132.

[44] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. 2007. Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning* 2 (2007), 125–140.

[45] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. 2011. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and instruction* 21, 2 (2011), 267–280.

[46] Ethan Z. Rong, Mo Morgana Zhou, Ge Gao, and Zhicong Lu. 2023. Understanding Personal Data Tracking and Sensemaking Practices for Self-Directed Learning in Non-classroom and Non-computer-based Contexts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–16. https://doi.org/10.1145/3544548.3581364

[47] Daniel M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*. ACM Press, Amsterdam, The Netherlands, 269–276. https://doi.org/10.1145/169059.169209

[48] Martin Schroder. 2023. AutoScrum: Automating Project Planning Using Large Language Models. https://doi.org/10.48550/arXiv.2306.03197 arXiv:2306.03197 [cs].

[49] Ava Elizabeth Scott, Lev Tankelevitch, and Sean Rintel. 2024. Mental Models of Meeting Goals: Supporting Intentionality in Meeting Technologies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[50] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. https://doi.org/10.1145/3613904.3642414

[51] Meng Sun, Minhong Wang, Rupert Wegerif, and Jun Peng. 2022. How do students generate ideas together in scientific creativity tasks through computer-based mind mapping? *Computers & Education* 176 (Jan. 2022), 104359. https://doi.org/10.1016/j.compedu.2021.104359 20 citations (Crossref) [2024-01-03].

[52] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[53] Johanna Telenius. 2016. *Sensemaking in Meetings - Collaborative Construction of Meaning and Decisions through Epistemic Authority*. Aalto University. https://aaltodoc.aalto.fi/handle/123456789/23391 ISSN: 1799-4942 (electronic).

[54] Hsin-Ruey Tsai, Shih-Kang Chiu, and Bryan Wang. 2024. GazeNoter: Co-Piloted AR Note-Taking via Gaze Selection of LLM Suggestions to Match Users' Intentions. *arXiv preprint arXiv:2407.01161* (2024).

[55] Karthikeyan Umapathy. [n. d.]. Requirements to support Collaborative Sensemaking. ([n. d.]).

[56] Erin Walker, Nikol Rummel, and Kenneth R Koedinger. 2011. Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning* 6 (2011), 279–306.

[57] Jixuan Wang, Jingbo Yang, Haochi Zhang, Helen Lu, Marta Skreta, Mia Husić, Aryan Arbabi, Nicole Sultanum, and Michael Brudno. 2022. PhenoPad: building AI enabled note-taking interfaces for patient encounters. *NPJ digital medicine* 5, 1 (2022), 12.

[58] Ruotong Wang, Lin Qiu, Justin Cranshaw, and Amy X. Zhang. 2024. Meeting Bridges: Designing Information Artifacts that Bridge from Synchronous Meetings to Asynchronous Collaboration. http://arxiv.org/abs/2402.03259 arXiv:2402.03259 [cs].

[59] Xu Wang, Carolyn Rose, and Ken Koedinger. 2021. Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 51, 14 pages. https://doi.org/10.1145/3411764.3445045

[60] Amber E. Witherby and Sarah K. Tauber. 2019. The Current Status of Students' Note-Taking: Why and How Do Students Take Notes? *Journal of Applied Research in Memory and Cognition* 8, 2 (June 2019), 139–153. https://doi.org/10.1016/j.jarmac.2019.04.002

[61] Sarah Shi Hui Wong and Stephen Wee Hun Lim. 2023. Take notes, not photos: Mind-wandering mediates the impact of note-taking strategies on video-recorded lecture learning performance. *Journal of Experimental Psychology: Applied* 29, 1 (March 2023), 124–135. https://doi.org/10.1037/xap0000375

[62] David Wood. 2001. Scaffolding, contingent tutoring, and computer-supported learning. *International Journal of Artificial Intelligence in Education* 12, 3 (2001), 280–293.

[63] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–30.

[64] Zheng Zhang, Weirui Peng, Xinyue Chen, Luke Cao, and Toby Jia-Jun Li. 2024. LADICA: A Large Shared Display Interface for Generative AI Cognitive Assistance in Co-Located Team Collaboration. *arXiv preprint arXiv:2409.13968* (2024).

[65] Rebecca Zheng, Marina Fernández Camporro, Hugo Romat, Nathalie Henry Riche, Benjamin Bach, Fanny Chevalier, Ken Hinckley, and Nicolai Marquardt. 2021. Sketchnote components, design space dimensions, and strategies for effective visual note taking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

# A VIDEO LECTURES USED IN THE USER STUDY

| Video Title | Description | Link |
|---|---|---|
| Video 1: Humans vs. AI: Who should make the decision? | Martin Keen, Master Inventor at IBM, delves into the strengths and weaknesses of both AI and human and introduces the concept of augmented intelligence. | https://www.youtube.com/watch?v=8lo1s29ODj8 |
| Video 2: Why Large Language Models Hallucinate | Martin Keen explains the different types of "LLMs hallucinations", why they happen, and ends with recommending steps for LLM users to take to minimize their occurrence. | https://www.youtube.com/watch?v=cfqtFvWOfg0 |
| Video 3: How to Pick the Right AI Foundation Model | Martin Keen walks through his six-step approach to picking the right AI model for a project. | https://www.youtube.com/watch?v=pePAAGfh-IU |

Table 5. Collection of videos used in the user study, listed in the same order as presented in the user study.

# B POST-TESTS AND RUBRICS

For each video, we designed a post-test comprising five questions: two multiple-choice and three open-ended. Each question is worth four points. Both question types are weighted equally, as prior research has shown that well-designed multiple-choice and open-ended questions can effectively support high learning quality and scalability in assessment[59].

## B.1 Video 1: Humans vs. AI: Who should make the decision?

**[4 Points] Question 1:** In the case of using an AI algorithm to predict fraudulent transactions on credit cards, which of the following is correct?

A. AI performs well when it thinks it definitely is a fraud transaction, or definitely is not a fraud transaction.
B. AI's success rate increases as the likelihood of AI predicting the transaction to be fraud increases.
C. AI's success rate is the highest when AI does not think the transaction is fraud.
D. If the AI algorithm works well on fraud detection, it works equally well on all transactions, regardless of the likelihood of the transaction being fraudulent.
E. I don't know the answer.

**[4 Points] Question 2:** When human financial analysts make decisions on fraud transactions, explain when they will do a better job than AI, and when they will do a worse job than AI, and why?

**Rubric:**

- 1 Point: Students mention that AI performs better when its confidence score is either very low or very high.
- 1 Point: Students mention that humans perform better when AI's confidence score is around 50% or AI is uncertain.
- 2 Points: Students provide correct reasoning for why AI or humans perform better. Students mention that AI relies on consistent logic while humans bring addition contexts and information.

**Good Example (4 Points):** When the AI confidence of fraud is not too high or too low, human experts perform better. When at very low or very high confidence, AI usually does better. When an AI is certain of itself, it's highly performed and beats out humans who can lose consistency, focus, and attention. A is they don't get distracted. On the other hand, when an AI is unsure, often for cases that are complex or statistically rare, humans can outperform an AI prediction by bringing in additional information and context that can look stuff up or ask a colleague, whereas the AI sticks to.

**Bad Example (0 Point):** When there is human bias involved, whether it's positively impacted (additional info that AI cannot get) or negatively (where humans are told to be wrong or some other form of distraction), humans can do better or worse than AI accordingly, and vice versa.

**[4 Points] Question 3:** Consider using augmented intelligence to detect fraudulent transactions, which is a combination of human and AI decision-making. When the AI is around 50% confident in its prediction, the success rate of the augmented intelligence is higher than AI but lower than humans, why is that? Why is the combination worse than human intelligence alone?

**Rubric:**

- 1 Point: Students provide reasoning for why the success rate of augmented intelligence is higher than AI. Students mention that AI alone will be uncertain when the confidence level is around 50% and involving humans can add additional contexts and refinement.
- 3 Points: Students provide reasoning for why the success rate of augmented intelligence is lower than humans.
  - 1 Point: Students receive 1 point for mention the concept of "cognitive bias".
  - 2 Points: Students explain the concept well.

**Good Example (4 Points):** When the AI is 50% confident in its prediction, it is much less successful, as lower confidence rates correlate strongly with success rates in AI. For humans, this is when they are more successful than AI, as there is less of a correlation. The issue with augmented intelligence is that there can be automated bias where AI is trusted over human judgement, even when AI is less accurate for this specific scenario.

**Bad Example (0 Point):** Because human can consider more factors and AI is not good at very complex task.

**[4 Points] Question 4:** A government officer is making decisions on which applicants to a welfare program will receive the benefit based on their needs. Due to the large number of applications, the officer uses an AI algorithm that makes predictions on whether this person is in tremendous need based on the information they provide. In the case above, where the officer is using AI assistance to make decisions on which applications to a welfare program will receive the benefit based on their needs. Assuming that the AI model is predicting whether the person should receive the benefit or not, and the AI model also has a confidence score for the prediction. Which of the following will likely harm the officer's trust in the AI model? (Select all that apply.)

A. Display that the AI model has 95% accuracy in the prediction.
B. Display that the AI model has 80% accuracy in the prediction.
C. Do not display the confidence score of the AI model.
D. State in natural language that there is likelihood that the AI model will make mistakes.
E. I don't know the answer.

**Rubric:**

- 1 point for 1 correct option selected
- 2 points for 2 correct options selected
- 4 points for 3 correct options selected
- 1 point deduction for each wrong option selected
- 0 total point if "I don't know the answer" is selected

**[4 Points] Question 5:** Assuming U of M is offering an online master's degree in data science that has thousands of applicants. The department is considering using an AI algorithm to assess the student's prior knowledge and background, in order to decide whether the student needs to take a required onboarding course. What is an ideal setup to ensure the best decision making outcome? Consider the following sub-questions:

- What is augmented intelligence like in this context?
- How might we display AI's suggestions and confidence score?

**Rubric:**

- 2 Points: Students design a setup that involves a combination of humans aided by AI. They show a good understanding of augmented intelligence by combining AI and human's advantages appropriately.
- 1 Point: Students explain the concept of augmented intelligence in the context.
- 1 Point: Students discuss how AI's suggestions and confidence score should be displayed. Students apply concepts of optional display and forced display and show good understanding of the differences between them. Students provide justification for either optional display or forced display.
  - Justify optional display: Students argue that optional display prevents over-reliance on AI by allowing humans to form their own judgment upfront, reducing automation bias and improving decision accuracy.
  - Justify forced display: Students argue that forced display ensures transparency and helps humans access all relevant information upfront, reducing the risk of missing important insights from the AI.

**Good Example (4 Points):**
1. I believe that augmented intelligence would be in the form of summarized information to the human decision maker. It would save a lot of time for that person if they could just use an AI model to summarize large documents like SOP's, and entire applications at a time. However, I don't really like this idea. It is convenient, but I believe it is a little dehumanizing towards the applicant. Especially if the AI gets some sort of information wrong, it's especially devastating toward the applicant which otherwise might've had a good chance at acceptance.
2. I believe we should make them true. The video states that showing text like "The AI's answer might be wrong" lowers the trust factor for humans using the AI recommendation, but I believe this should be good for this use case. Confidence score should also be displayed as true. If the AI isn't sure, we must be able to convey that to the decisionmaker, who has the right to be skeptical about the information coming in. This exact skepticism is what allows humans to make that additional success rate jump when the confidence score in unsure, as we can leverage additional bits and pieces of content.
I believe that in this case, we should consider an optional display approach. The human decisionmaker can do through the applicant's file on their own, and then additionally ask the AI for help if needed.

**Bad Example (0 Point):** The augmented intelligence system should be able to evaluate the prior knowledge level and the extent to which the background of the candidate matches with the program, and whether or not the student should be considered for admission. The AI's suggestions should be presented along with natural reasoning based on the assessment results.

### B.2 Video 2: Why Large Language Models Hallucinate

**[4 Points] Question 1:** Which of the following is NOT a reason for hallucination?

A. There may be mistakes in the data source, e.g., when posts on Reddit (that are counterfactual) are used for training.
B. The generation method may favor metrics, such as novelty or creativity, which leads to hallucination.
C. The prompt (user input) may be vague, causing the LLMs to give untruthful answers.
D. There may not be sufficient computing resources to run a large parameter model (e.g., 200 billion parameters), leading the model to hallucinate.
E. I don't know the answer.

**[4 Points] Question 2:** An AI assistant powered by a Large Language Model (LLM) responds to users with the following statements:

A. Prompt: "When was the Great Wall of China built?"
Response: ""The Great Wall of China was built during the 20th century to protect against aerial attacks."
Correct Type: Factual contradiction
B. Prompt: "Write a positive review of the restaurant."
Response: "The food was cold, and the service was terrible."
Correct Type: Prompt contradiction
C. Prompt: "What are the symptoms of the common cold?"
Response: "The common cold can cause sneezing and coughing. It can also cause your entire body to turn green overnight."
Correct Type: Nonsensical information
D. Prompt: "What's the color of the floor of those tennis courts?"
Response: "The color is red. The color is blue."
Correct Type: Sentence contradiction

Map each hallucination to one of the following options.

- Sentence contradiction
- Prompt contradiction
- Factual contradiction
- Nonsensical information
- I don't know the answer

**Rubric:** 1 point for correctly matching each type.

**[4 Points] Question 3:** Please explain what is a sentence contradiction and what is a prompt contradiction.

**Rubric:**

- 2 Points: Students correctly explain sentence contradiction, which means an LLM generates a sentence that contradicts one of the previous sentences.
- 2 Points: Students correctly explain prompt contradiction, which means the generated sentence contradicts with the prompt that was used to generate it.

**Good Example (4 Points):** Sentence contradiction: the AI model contradicting itself within a sentence or paragraph, saying one thing and then saying another than can't exist with the first thing it said being true.
Prompt contradiction: the AI model does not generate a response to the question asked, i.e., displaying a negative review when asked for a positive one.

**Bad Example (2 Points):** Sentence contradiction is when the second sentence contradicts first sentence. Prompt contradiction is when the LLM gives a complete different answer to what the prompt asked for.

**[4 Points] Question 4:** What strategies can we use to minimize hallucination? *Hint: You can talk about how users can craft prompts to reduce hallucination; you can also explain what temperature is and how that relates to hallucination.*

**Rubric:**

- 1 Point: Students mention the strategy of providing clear and specific prompts to the system.
- 1 Point: Students correctly explain the temperature parameter.
- 1 Point: Students provide reasoning of how adjusting temperature can reduce AI's hallucination.
- 1 Point: Students mention more examples about clear and specific prompts, or the multi-shot prompting strategy, or more discussion about the temperature.

**Good Example (4 Points):** To minimize hallucinations users can use clear and specific prompts which include the context for their query. You can also alter the settings such as by lowering the temperature which minimizes the number of hallucinations. Temperature is a metric which determines the deviation between an LLMs responses given the same prompt. A high temperature will lead to highly varied and distinct responses whereas a low temperature will have deterministic responses.

**Bad Example (0 Point):** As the reasoning capabilities of models increases, the hallucinations in the model decreases. Not sure of second part.

**[4 Points] Question 5:** You are using ChatGPT to help you write poems that follow a certain rhythm. How do you leverage the multi-shot prompting strategy to reduce hallucination?

**Rubric:**

- 1 Point: Students correctly explain the multi-shot prompting strategy in the context of poem generation.
- 2 Points: Students correctly describe the examples used in the multi-shot prompting strategy.
- 1 Point: Students provide reasoning of how using this strategy can reduce AI's hallucination.

**Good Example (4 Points):** I would include in my prompt multiple poems that follow the same rhythm as the one I intend, in order to allow the LLM to recognize the pattern and establish a better context.

**Bad Example (0 Point):** use 3.5, 4o and o1 separately with the same prompts; paraphrase the prompt

### B.3 Video 3: How to Pick the Right AI Foundation Model

**[4 Points] Question 1:** Assuming you have a use case to use generative AI and are deciding which foundational models to use. Which of the following is not a step to take according to the lecture?

A. Figuring out what you want to use generative AI for.
B. Identify important properties of the models you're considering, e.g., size, costs, etc.
C. Fine-tune the foundational model using a small training set.
D. Evaluate the models on the tasks you want the models to perform.
E. I don't know the answer.

**[4 Points] Question 2:** Assume you work at the AI department of Grammarly, and you are considering using generative AI to give users suggestions to address grammar mistakes in their writing. You need to look at the model cards of a series of shortlisted models, including Llama-2, GPT-4, etc. What information do you need to look for on the model cards?

**Rubric:**

- 2 Points: Students clearly state that they will check if the model has been trained on data specifically for their purposes.
- 2 Points: Students explain their reasoning in the context.

**Good Example (4 Points):** First it is their size and cost to use it. Then to see what the model is pre-trained for, a model pre-trained for text generation or grammar revision may be better in our use.

**Bad Example (0 Point):** Common grammatical errors made by users, a repository that contains proper spellings and syntax, etc.

**[4 Points] Question 3:** When you assess the shortlisted models, you may assess their accuracy, reliability, and speed. Is there any tradeoff between these performance metrics? Please explain.

**Rubric:**

- 2 Points: Students correctly discuss the trade-off between accuracy and speed for large models.
- 2 Points: Students correctly discuss the trade-off between accuracy and speed for small models.
- 1 point deduction if students over-claim the relationship between accuracy and speed using word such as "definitely", we will deduct 1 point.

**Good Example (4 Points):** A larger model is more accurate, but can be slower and require more powerful hardware. A smaller model may run faster on less powerful hardware, but can produce less accurate output.

**Bad Example (0 Point):** The fastest cannot be the most accurate one, and reliability comes down to how much data the model is trained on as well as how accurate the dataset is, finetuning the parameters etc.

**[4 Points] Question 4:** You are considering using a foundational language model to power a chatbot embedded on the M-Den website that sells Michigan swags. The chatbot needs to address customer questions and make recommendations. How would you approach this decision making process, based on the 6-stage decision making framework proposed in the lecture?

**Rubric:**

- 2 Points: Students correctly define all the stages with definition.
- 2 Points: Students explain each stage in the context.
- 0.7 point deduction for each missing stage.

**Good Example (4 Points):**
Step 1: respond to customer questions
Step 2: Consider models like GPT3.5, GPT4o, etc.
Step 3: Assess models based on size and cost, given relatively small operation probably lean towards cheaper models
Step 4: Assess characteristics desired such as accuracy in responding customer questions and try to increase sales revenue
Step 5: Ask the model questions like how much does an item cost or if it is in stock and cross verify with actual inventory information to assess the accuracy of the model
Step 6: Determine which model is most accurate, in this use case accuracy is more important than speed because wrong information might anger customers more so than a slower but factually correct response

**Bad Example (0.7 Point):** I would find open models that are designed for customer service and see how they perform in tests. After choosing the best one, I would like to give it whatever the M-Den (R.I.P.) has in inventory, along with product descriptions, so it can give relevant and accurate answers.

**[4 Points] Question 5:** Which of the following is NOT correct on how you may deploy the selected large language model (for the goal above)?

A. I may consider deploying it on a public cloud service (e.g., Google Cloud platform), which may be cheaper than purchasing compute resources dedicated to this task.
B. Deploying it on a public cloud service is going to have better performance than deploying on local compute clusters, since the public cloud service has more powerful machines.
C. The local (on-premise) deployment may achieve better performance since we can do more fine-tuning with information on the products that are sold at M-Den.
D. Deploying on premise can help protect the security of data being generated in the process.
E. I don't know the answer.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009