Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools

Authors:

Pia Kreijkes[1], *Viktor Kewenig*[2]*, *Martina Kuvalja*[1]*, *Mina Lee*[2], *Sylvia Vitello*[1], *Jake M. Hofman*[2], *Abigail Sellen*[2], *Sean Rintel*[2], *Daniel G. Goldstein*[2], *David Rothschild*[2], *Lev Tankelevitch*[2], *Tim Oates*[1]

*Joint second authors

Affiliations:

[1]Cambridge University Press and Assessment
[2]Microsoft Research

Abstract

The rapid uptake of Generative AI, particularly large language models (LLMs), by students raises urgent questions about their effects on learning. We compared the impact of LLM use to that of traditional note-taking, or a combination of both, on secondary school students' reading comprehension and retention. We conducted a pre-registered, randomised controlled experiment with within- and between-participant design elements in schools. 405 students aged 14-15 studied two text passages and completed comprehension and retention tests three days later. Quantitative results demonstrated that both note-taking alone and combined with the LLM had significant positive effects on retention and comprehension compared to the LLM alone. Yet, most students preferred using the LLM over note-taking, and perceived it as more helpful. Qualitative results revealed that many students valued LLMs for making complex material more accessible and reducing cognitive load, while they appreciated note-taking for promoting deeper engagement and aiding memory. Additionally, we identified "archetypes" of prompting behaviour, offering insights into the different ways students interacted with the LLM. Overall, our findings suggest that, while note-taking promotes cognitive engagement and long-term comprehension and retention, LLMs may facilitate initial understanding and student interest. The study reveals the continued importance of traditional learning approaches, the benefits of combining AI use with traditional learning over using AI alone, and the AI skills that students need to maximise those benefits.

Main

Learners' rapid and widespread adoption of Generative Artificial Intelligence (GenAI) tools, particularly Large Language Models (LLMs), has unsettled the global educational landscape by offering

new ways for students to engage with learning materials[1;2;3;4;5;6] while also creating new challenges[7;8;9;10;11;12]. Large national surveys in the UK and US have found that a sizeable proportion of school students use GenAI tools such as OpenAI's ChatGPT[13;14]. This development raises fundamental questions about teaching and learning models. And yet, the vast majority of existing research on learning with LLMs has focused on the higher education context, leaving substantial knowledge gaps regarding effects on younger learners[15]. In addition, previous research has concentrated on second language education, mostly writing performance, as well as computing, health, and physics[15]. While such studies overall reveal positive effects of LLM use on academic performance, researchers call for caution as these might reflect the quality of LLM-produced work rather than genuine improvements in students' learning[15]. The effect of LLM use on two foundational aspects of learning – understanding and retaining information – remains critically underexplored. Knowledge stored in long-term memory is a fundamental element of cognition, forming the basis of nearly all human activity[16]. Thus, understanding the effects of LLMs on these foundations is urgently required to guide how such tools are integrated into schools, as policymakers and educators on the front-line are grappling with many unknowns. This study presents one of the first large-scale quantitative investigation into how reading comprehension and retention are affected by the use of LLMs.

Reading comprehension is the process of making sense of written materials resulting in a mental representation of the material[17]. Models of reading comprehension, such as the Construction-Integration (CI) model[18], highlight that readers need to understand a text at several levels: the surface structure (words and their syntactic relations), the textbase (propositions, which generally represent one full idea), and the situation model (inferences about the text)[17]. This multi-level structure is supported by neuroimaging studies[19;20;21;22;16]. The ability to make inferences is a key aspect of comprehension. Usually, two types of inferences are distinguished: text-based bridging inferences involve connecting information from different text locations (e.g., the current sentence with a previous sentence) and knowledge-based inferences involve connecting information in the text with prior knowledge[17]. A reader's ultimate comprehension of a text depends on complex interactions between various elements, including factors related to the reader's characteristics (e.g., decoding skills, vocabulary and linguistic knowledge, prior domain knowledge, working memory capacity, inference-making ability, knowledge of reading strategies, motivation, and goals)[23;24;25;26;27], the text itself (e.g., genre, length, word and sentence complexity, cohesion)[28;29], and the reading context (e.g., reading for leisure or academic purposes)[30;31].

Reading retention is the process of storing the comprehended content from a text in long-term memory. For learning it is necessary to not just comprehend the text at the time of reading, but also being able to remember what one has read and understood later. Retention is, in part, determined by the level and quality of information processing during encoding (i.e., the initial information acquisition while reading). According to the Levels of Processing framework[32;33], information that is processed deeply and elaborately —through semantic analysis involving meaning, inferences, and implications— can be recalled more readily. Deep processing facilitates the formation of rich, interconnected semantic networks, which provide multiple retrieval cues, and thus enhance the retrieval potential, as well as the construction of a robust schematic framework wherein specific details are meaningfully organised and related[32;34].

There are several reading strategies and learning activities that can enhance comprehension and retention as outlined by McNamara[35] and Chi[36]. Throughout the reading process, monitoring comprehension is particularly crucial, and includes strategies such as generating questions to gauge one's understanding[35]. Text-focused strategies involve interpreting the meaning of words, sentences and ideas (e.g., paraphrasing, breaking up long and complex sentence into manageable chunks, making bridging inferences to link different concepts)[35]. Strategies such as paraphrasing, selecting, and repeating are also considered active learning strategies, and these can activate prior knowledge and support the encoding, storing and assimilation of new knowledge[36]. There

are also several effective reading strategies that go beyond the text (e.g., generating questions, using self-explanations, and using external information sources)[35]. Such strategies are considered to be constructive as learners generate new ideas and integrate information more deeply through explaining, elaborating, and connecting. This involves cognitive processes such as inferring new knowledge, integrating and organising new and existing knowledge, and repairing faulty knowledge[36]. Lastly, interactive learning activities involve meaningful dialogue with a partner, including with peers or systems like intelligent tutoring agents[36;28]. Such interactions can enhance learning by providing scaffoldings, corrective feedback, as well as additional information and new perspectives. Importantly, a dialogue is only considered to be interactive if both partners make substantive contributions[36].

The integration of LLM tools into education raises the crucial question of whether their use could facilitate or undermine such learning strategies while reading. These models offer unprecedented flexibility in generating explanations, providing diverse perspectives, responding to complex questions in real-time, and adapting to individual learners' needs[37;38]. By serving as an external knowledge resource that extends beyond learners' personal knowledge and skills, LLMs can potentially enhance students' understanding and engagement with educational materials[39;40;10;41]. Furthermore, LLMs' ability to provide immediate clarifications and simplify complex concepts may help reduce cognitive load[42;43]. Thus, LLMs may be particularly useful in helping learners build understanding at multiple levels: from surface-level text comprehension and identification of key ideas, to deeper text-base representation of meanings, and ultimately to a comprehensive mental representation at the situation-model level of comprehension.

However, over-use of LLMs could lead to shallow processing, where learners passively receive information without actively engaging in deep cognitive processing or critical thinking[44;36;45;46;47]. This superficial engagement could hinder the development of comprehensive mental models, negatively affecting comprehension and long-term retention[33;48]. When learners depend excessively on LLMs for answers and explanations, they may be less inclined to employ self-explanation and elaboration strategies that are essential for comprehension and meaningful learning[35;49;42]. While LLMs can make information readily accessible, this accessibility needs to be leveraged in ways that promote, rather than substitute for, the deep cognitive processing necessary for knowledge consolidation and learning[50;51].

In order to assess the effectiveness of using LLMs as a learning tool for reading comprehension and retention, we compared it to a widely used learning activity that can facilitate many active and constructive strategies – note-taking. It is one of the most common and widely used learning activities and has been found to be an effective aid to learning while reading[52;53]. Note-taking can stimulate active processing of information and encourage the integration of new material with prior knowledge, thereby aiding comprehension as well as creating retrieval cues that aid later recall[52;54]. The impact of note-taking appears to vary depending on the depth of cognitive processing involved. It could focus readers on shallower processing, because readers might pay more attention to the surface structure and textbase but it could also enhance the situation-model by encouraging elaboration and better mental organisation[55;56;57]. Kobayashi's[52] meta-analysis supports the former as it found relatively small effects for higher-order performance tests, suggesting that the generative value of note-taking may be limited and highly dependent on the quality of the notes taken (whether they are verbatim or generative). We also compared the effectiveness of using an LLM on its own with using an LLM in conjunction with note-taking, given that it might be useful to combine the activities of querying LLMs and taking notes to facilitate learning. The two activities could potentially have complementary effects on reading comprehension and retention by drawing on their respective strengths. However, there might also be a risk of dividing attention in a way that renders both activities less effective.

To examine whether LLMs can be used as a tool to support the fundamental learning processes of reading comprehension and retention, we conducted a large-scale, **pre-registered**, randomised

controlled experiment with within- and between-participant design elements. The study involved 405 secondary school students, aged 14-15 years, and took place in seven schools in England (UK). The experiment consisted of a *learning* session and a *test* session, which were three days apart. In the *learning* session, each student was tasked with understanding and learning two text passages on a different history topic (Apartheid in South Africa and the Cuban Missile Crisis), each by using a different learning activity (learning condition) drawing on evidence-based strategies. Students were not informed that they would be tested on the passages. They were randomly assigned to one of two groups. Group 1 was exposed to conditions referred to as "LLM" (i.e., using an LLM to understand and learn a text) and "Notes" (i.e., taking notes to understand and learn a text) and Group 2 was exposed to conditions referred to as "LLM" and "LLM+Notes" (i.e., using an LLM alongside note-taking to understand and learn a text). Both learning condition and text order were randomised. The LLM functionality in the learning session was provided by a private Azure-hosted instance of OpenAI's GPT-3.5 turbo model. After each learning task, students responded to a survey about their learning experience, with both quantitative and qualitative questions.

In the *test* session, students completed a range of questions assessing different levels of comprehension and retention. Specifically, we assessed their literal retention, comprehension, and free recall. For each passage, literal retention (i.e., lower-level retention) was measured through eight short response (cued recall) and ten multiple choice (recognition) questions assessing literal information which did not require any knowledge-based inferences, and no or only minimal text-based (bridging) inferences. Comprehension (i.e., higher-level retention) was measured through three open response questions requiring bridging inferences to connect information from several different text locations as well as knowledge-based inferences. Free recall was assessed through one open response question for each text, asking students to write down everything they remembered, and thus measuring how much students retained and understood without any cueing.

Our primary aim was to quantify the impact of using an LLM on students' reading comprehension and retention. We made the choice not to have a "reading-only" control condition both because it would limit participant fatigue in responding to conditions, and on the basis that any engagement with the text beyond passive reading is likely going to lead to improved learning outcomes[35;36], setting the bar for LLM use comparatively low. Instead, we decided to compare it against the common, evidence-based learning activity of note-taking. We also explored students' learning experiences when engaging in the different learning activities, including which activity they preferred and why, as well as different "archetypes" of prompting behaviour that shed light on the learning outcomes. The results offer valuable insights for stakeholders and policy makers of the global education landscape.

Results

Our study investigated the effects of using an LLM on student learning outcomes compared to traditional note-taking in a sample of 344 students (after applying pre-registered exclusion criteria, see Methods for more information). Group 1 (LLM vs Notes conditions) had a final sample of 184 students and Group 2 (LLM vs LLM+Notes conditions) of 160 students. Among the students there were slightly more males than females, most were English native speakers, a small number of students (5.2%) received free school meals indicating socioeconomic disadvantage, and about half were taking History GCSEs (see Supplementary Table 3 for all student characteristics). Both groups showed similar prior familiarity with the three learning conditions (LLM, Notes, LLM+Notes). About half of the students regularly took notes and most reported limited prior use of LLM for learning (see Supplementary Table 4 for detailed frequencies).

Learning outcomes

We compared the impact of LLM (reference condition, used by all students) to the impact of Notes (used by students in Group 1) and LLM+Notes (used by students in Group 2) on students' literal retention, comprehension, and free recall. Traditional note-taking led to the best performance across all measures, followed by LLM+Notes, while using LLM alone resulted in the lowest scores (see Supplementary Table 5 for descriptive statistics).

Linear mixed-effects models confirmed significant differences across the conditions (see Figure 1, see Supplementary Table 6 for all model coefficients, confidence intervals and effect sizes).

For **literal retention**, we found significant main effects for both Notes ($\beta = 1.92$, $p < 0.001$, 95% CI [1.42, 2.42]) and LLM+Notes ($\beta = 0.57$, $p = 0.040$, 95% CI [0.03, 1.11]), indicating that students performed better with Notes compared to LLM and better with LLM+Notes compared to LLM.

For **comprehension**, we again found significant main effects for both Notes ($\beta = 0.95$, $p < 0.001$, 95% CI [0.62, 1.28]) and LLM+Notes ($\beta = 0.35$, $p = 0.049$, 95% CI [0.00, 0.70]), where students had better performance with Notes compared to LLM and with LLM+Notes compared to LLM.

For **free recall**, we found a significant main effect for Notes ($\beta = 1.02$, $p = 0.018$, 95% CI [0.18, 1.86]) but not for LLM+Notes ($\beta = -0.08$, $p = 0.855$, 95% CI [-0.98, 0.81]). Thus, students showed better performance with Notes compared to LLM but there was no significant difference between LLM+Notes compared to LLM. Given the non-normal distribution of free recall scores, we also conducted non-parametric versions of these tests as a robustness check, detailed in the Methods section, which corroborated these findings.

These results suggest that both note-taking conditions (either alone or with LLM) showed improved learning compared to using LLM on its own. However, the benefit of note-taking was seen across all different measures of learning, whereas the benefit of LLM+Notes was seen for literal retention and comprehension but not for free recall.

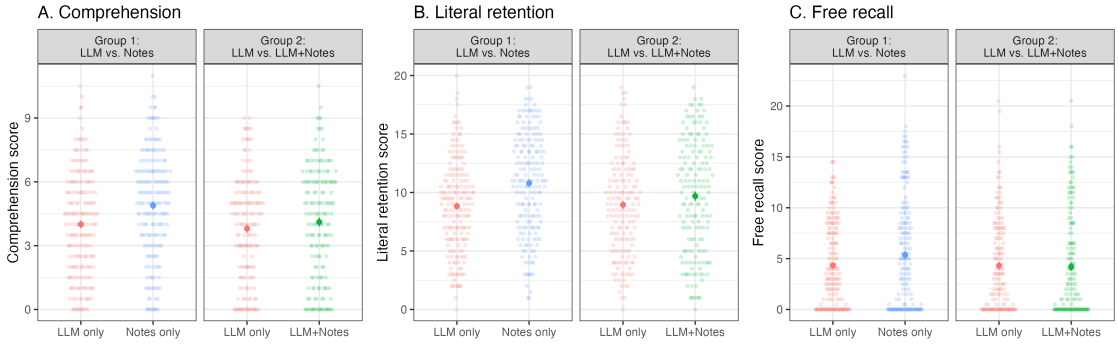

Figure 1: Distribution of test performance by condition and group for **Comprehension** (left, max 12 points; Notes: $M = 4.89$, $SD = 2.52$; LLM+Notes: $M = 4.11$, $SD = 2.65$; LLM Group 1: $M = 4.00$, $SD = 2.44$; LLM Group 2: $M = 3.80$, $SD = 2.47$), **Literal retention** (middle, max 20 points; Notes: $M = 10.8$, $SD = 4.29$; LLM+Notes: $M = 9.68$, $SD = 4.83$; LLM Group 1: $M = 8.83$, $SD = 3.96$; LLM Group 2: $M = 8.95$, $SD = 4.29$), and **Free recall** (right, max 50 points; Notes: $M = 5.36$, $SD = 5.49$; LLM Group 1: $M = 4.32$, $SD = 4.15$; LLM Group 2: $M = 4.32$, $SD = 4.63$; LLM+Notes: $M = 4.20$, $SD = 5.07$). Mean values are indicated by the two large circles within each facet, whereas the smaller points show individual students scores. Error bars indicate one standard error above and below the mean. Group 1 is shown on the left facet of each subfigure, comparing LLM (red) and Notes (blue). Group 2 is on the right facet of each plot, comparing LLM (red) and LLM+Notes (green).

Behavioural engagement

Behavioural engagement with the LLM and note-taking was quantified by the average number of queries made to the LLM, the average number of words written in students' notes as well as time spent on task. Access to notes alongside the LLM reduced students' query frequency compared to LLM-only conditions (from 9.21 to 6.02 queries in Group 2). While students wrote a similar number of words in their notepad in both Notes and LLM+Notes conditions (around 100 words), a concerning proportion (25.63%) heavily copied from LLM outputs into their notes, with some (16.25%) showing nearly complete copying (more than 90% overlap of trigrams between LLM output and notes). Additionally, students spent significantly less time on task when using only the LLM compared to conditions involving note-taking (differences of 0.80 and 1.54 minutes for Groups 1 and 2, respectively), suggesting deeper engagement when note-taking was involved. See Supplementary Table 7 for a full description of behavioural measures.

Prompting behaviour

In order to understand how students engaged with the LLM, we performed a qualitative analysis of all prompts (n = 4,929) using a hierarchical coding scheme where specific prompts were nested within overarching prompt types. Each prompt could be assigned to multiple codes. We identified four behavioural archetypes of how students worked with the LLM in relation to the task as well as two additional overarching prompt types that were not directly related to the task (see Figure 2 for the distribution of prompt types across each LLM session). For exact frequency counts of overarching prompt-types, see Supplementary Table 21 and for specific prompt types see Supplementary Table 22.

The most frequent archetype was seeking additional information and deeper understanding (2,265 prompts, as shown in the purple bars in Figure 2). The vast majority of students (90%)

used such a prompt type at least once, about 40% used this as their first prompt, and 60% as their most common prompt type (see Figure 3). These prompts primarily comprised requests for elaboration (1,479 instances) and general background information (514 instances). Examples include "how are people today affected by the apatheid" and "why did it take so long to free nelson mandela".

Information condensation (749 prompts, as shown in the teal bars in Figure 2) emerged as the second most common archetype, with 27% of students using it as their first prompt, typically requesting summaries or key ideas, such as "What are five key points from the entire text?" or "create a timeline of all the events". The third archetype, basic understanding of the text (615 prompts, green bars in Figure 2), was used by 70% of students at least once, mainly for definitions and content simplifications such as "What is a sanction?" and "explain communist". A fourth archetype, requesting direct study and memory help, was used infrequently (39 instances, red bars in Figure 2) despite students receiving no explicit instructions for such use. These ranged from asking the LLM to generate a quiz ("ask me 4 questions about the text and tell me if i get them right after my next reply") to mneumonic devices ("create me a mnemonic device on the cuban missile crisis").

Beyond these archetypes, 760 prompts focused on interacting with the LLM rather than (or in addition to) text content (blue bars in Figure 2), primarily requesting specific formats or response improvements. Examples include "can you put this into bullet points?" and "shorten the aftermath into 1 sentence". Notably, only six prompts questioned the LLM's reliability. Finally, about 10% of all interactions (501 prompts, brown bars in Figure 2) were off-topic or irrelevant (e.g., "what is the meaning to life" and "Tell me about Harry Potter"), showing that a small but potentially relevant prompt proportion was not task-focused, potentially due to low task motivation or boredom.

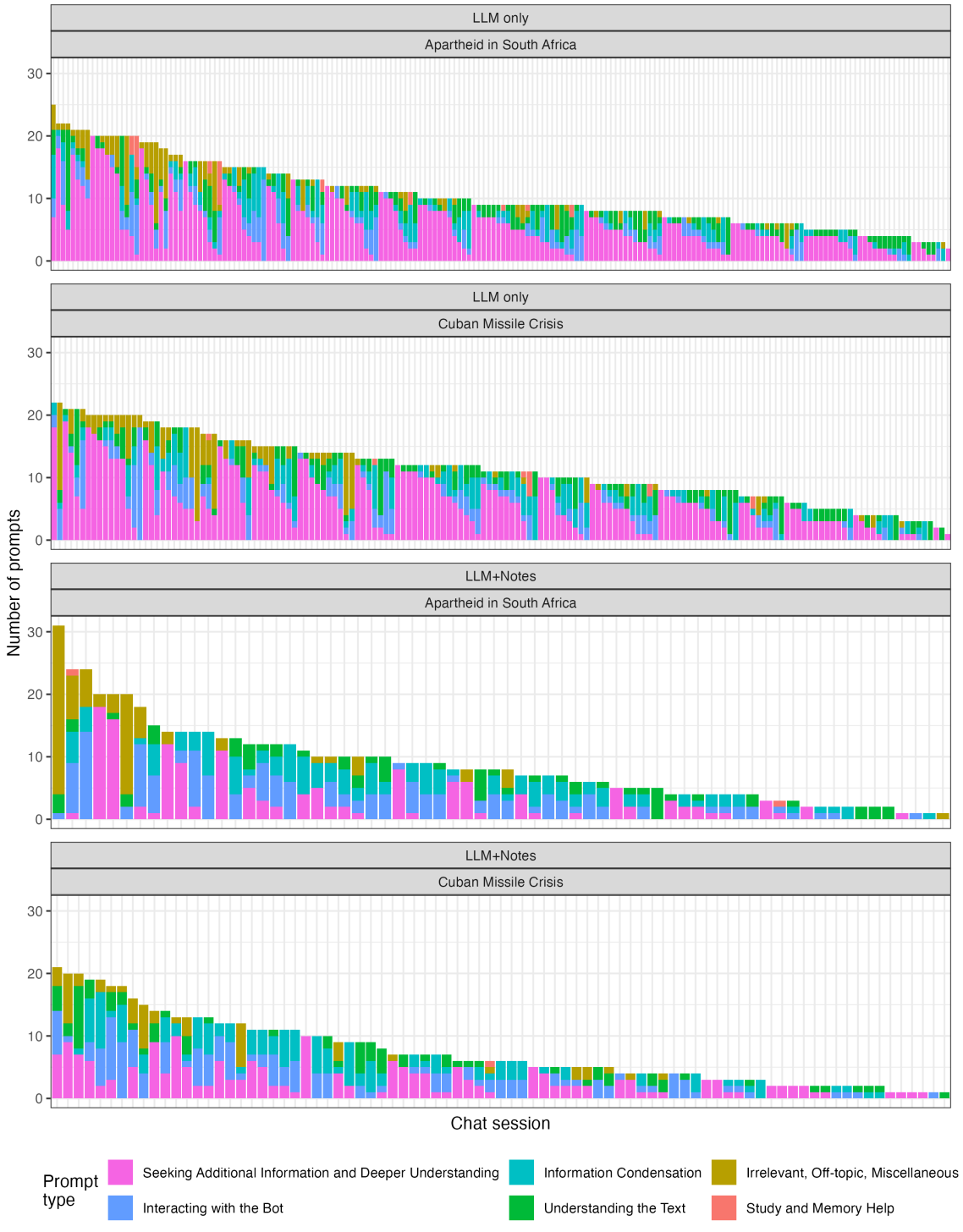

Figure 2: Distribution of prompt types across LLM sessions for different conditions and students. Each panel represents a specific combination of condition (LLM-only or LLM+Notes) and text passage (Apartheid in South Africa or Cuban Missile Crisis). Each bar shows the number of prompts within each type for an individual LLM session, with sessions sorted in descending order by the total number of prompts and ties broken by the number of prompts within each type.

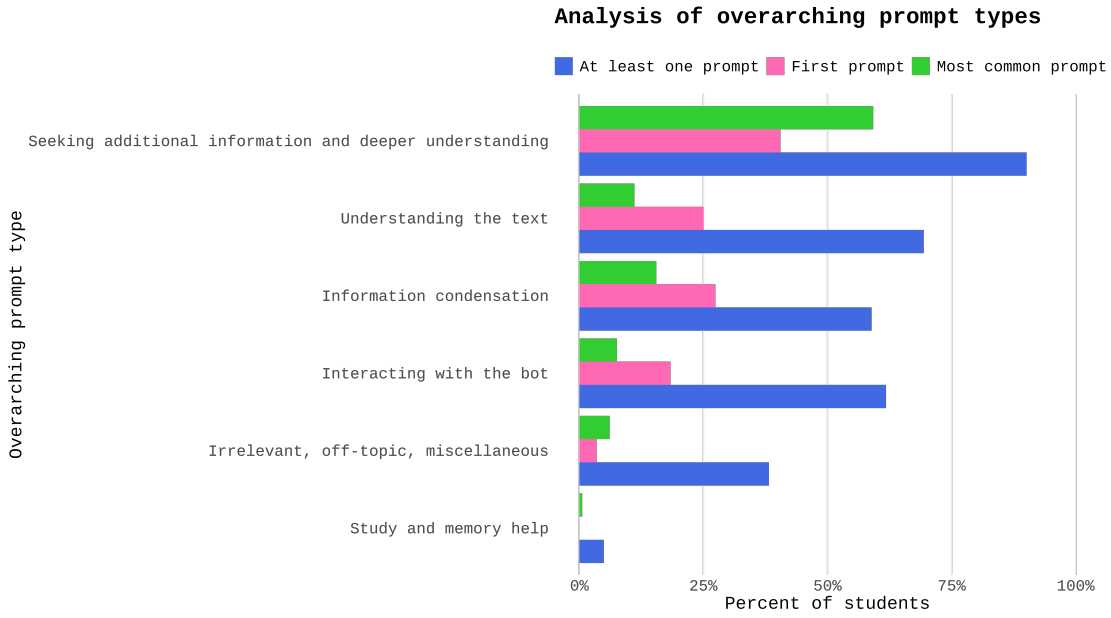

Figure 3: Distribution of student prompts across different types, showing the percentage of students who used the prompt type at least once (blue), as their most common prompt (magenta), and as their first prompt (green). Prompt types are arranged by overall frequency.

Learning experiences and perceptions

In addition to analysing students' behavioural engagement, we asked them about their learning experiences and perceptions of the different conditions. The quantitative results are summarised in Figure 4, with details of statistical tests in Supplementary Table 15. We used an adjusted p-value threshold of $0.05/18 = 0.002$ to gauge statistical significance based on the Bonferroni correction to account for multiple comparisons (n = 18).

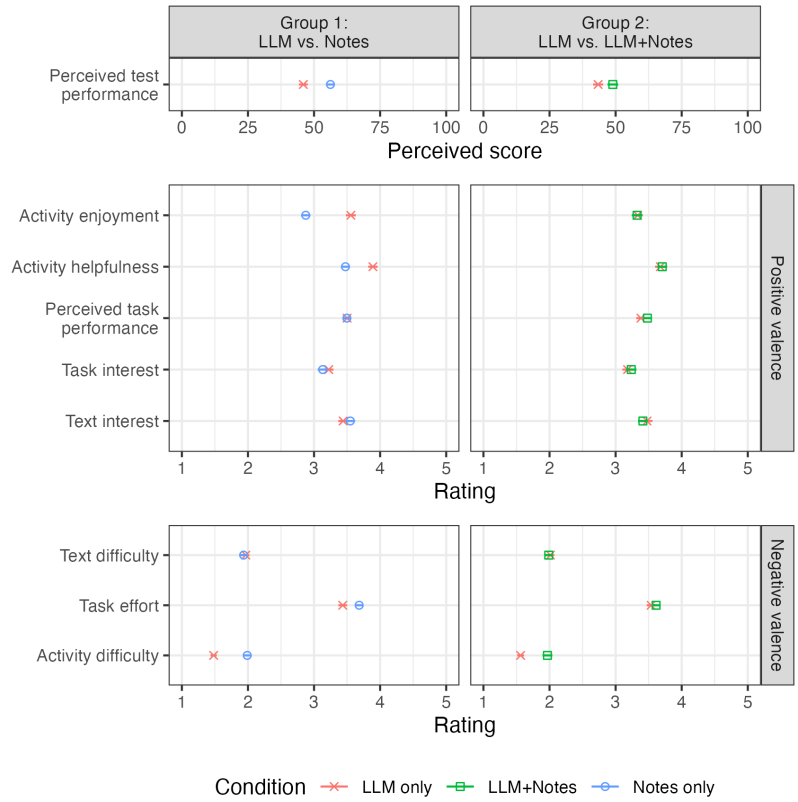

Figure 4: Differences in learning experiences and perceptions by group and condition. The top panel displays perceived test performance on a 0–100 scale, while the middle and bottom panels show ratings for measures with positive and negative valences, respectively, on a 1–5 scale. Each point represents the mean rating for a condition, with error bars indicating one standard error above and below the mean.

Contrary to actual learning outcomes, Group 1 students found the LLM more helpful, easier to use, and more enjoyable than note-taking, while reporting less effort investment. Group 2 showed similar experiences between conditions, except perceiving the LLM-only condition as less difficult than LLM+Notes. Students perceived task performance similar across conditions during learning. Following the test, students in both groups accurately reported their perceived test performance to be lower in the LLM-only conditions than in the Notes and LLM+Notes conditions.

These findings suggest that while the LLM-only condition was less effective for learning, it provided motivational benefits - particularly evident in Group 1's preferences. Importantly, these motivational benefits were maintained when combining LLM use with note-taking in Group 2.

Activity preferences

Students were asked to indicate their preferred learning activities and explain their preferences through an open response (see Table 1). In Group 1, most students preferred the LLM activity over traditional note-taking. Those students cited enhanced understanding, the LLM's ability to answer questions, and ease of the activity as their main reasons. Students favouring traditional note-taking emphasised benefits for understanding, the importance of self-generated work, and improved

memory retention. In Group 2, a substantial majority preferred the combined activity over using the LLM alone. Students preferring the combined activity noted the complementary benefits of both approaches, enhanced memory retention, and improved organisation. Those favouring the LLM-only activity emphasised its efficiency, particularly appreciating that the LLM did the work for them. This reveals an underlying tension between efficiency and depth of processing - while the LLM-only activity was perceived as more efficient, conditions involving note-taking demonstrated superior learning outcomes through deeper engagement and better retention.

Table 1: Learning activity preferences and reasons by group

Activity preference and reasons	Count	Percentage
Group 1: LLM vs Notes		
LLM over Notes	89	42.0
Notes over LLM	57	26.9
No preference	48	22.6
Not sure	18	8.5
Group 2: LLM vs LLM+Notes		
LLM over LLM+Notes	32	16.2
LLM+Notes over LLM	100	50.5
No preference	48	24.2
Not sure	18	9.1
Reasons for LLM over Notes preference		
Helps understanding	34	21.9
Answers questions	23	14.8
Easy to use	22	14.2
Quick to use	18	11.6
Provides background	18	11.6
Summarises and simplifies	17	11.0
Engaging	10	6.5
Interactive	8	5.2
Helps remember	4	2.6
Reasons for Notes over LLM preference		
Helps understanding	22	21.4
Own work	21	20.4
Aids memory	18	17.5
Helps processing	8	7.8
Unclear usage of LLM	7	6.8
Active learning	6	5.8
LLM distracts	6	5.8
Revisitable	5	4.9
Easier	4	3.9
Helps organisation	4	3.9
Reasons for LLM over LLM+Notes preference		
Does the work for you	15	50.0
Notes not necessary	5	16.7
Quicker	4	13.3
More time for questions	4	13.3
Reasons for LLM+Notes over LLM preference		
Best of both worlds	35	23.2
Helps remember	27	17.9
Helps organisation	24	15.9
Own work	21	13.9
Helps understanding	16	10.6
More helpful and easier	12	7.9
Helps process LLM output	6	4.0
More fun	4	2.6
LLM errors	3	2.0

Note: This table only includes reasons that have been mentioned by at least three students.

Future use

At the end of the learning session, students reported their intentions for future use of each activity. In Group 1, the majority of students (64.4%) indicated they would use LLMs in the future, with only 7.3% negating and 28.2% being unsure. A smaller majority of students (55.3%) planned to take notes in the future, and 10.6% did not think they would do so, while 34.1% were uncertain. In Group 2, the majority of students (59.5%) intended to use LLMs in the future, 10.4% did not and 30.1% were unsure. A similar majority (58.5%) planned to use the combined LLM+Notes activity in the future, while 14.6% did not and 26.8% were unsure.

Discussion

This study provides new insights into how the use of LLMs compares to and interacts with traditional evidence-based practices (specifically note-taking) to support students' reading comprehension, retention, and engagement. It offers important perspectives on the cognitive and motivational dynamics underlying human-AI interactions in learning, and how these interactions influence educational outcomes and perceptions. In particular, it suggests that LLM use and more traditional note-taking have complementary roles in the learning process.

In this study, we found that note-taking—whether done alone or alongside LLM usage—produced higher comprehension and retention scores compared to using an LLM alone, underscoring the importance and effectiveness of traditional active learning strategies. At the same time, students generally used LLMs constructively and perceived them as more "helpful" and preferable to note-taking. How can we reconcile these seemingly conflicting results?

One part of the answer may be that students simply have a limited metacognitive understanding of what is in fact helpful for their own learning [58;59;60], specifically in the context of GenAI [61]. In particular, they may underweight the importance of the "desirable difficulties" induced by activities such as note-taking [48]. Note-taking requires active processing of information, such as identifying important information, paraphrasing and summarising [52]. While these tasks demand cognitive effort and may not be inherently enjoyable, past research shows that the learning potential increases with the level of required cognitive engagement [62]. Having an LLM do some of the work of summarising a passage or explaining a concept may *feel* more enjoyable and efficient, but can reduce the cognitive engagement necessary for deep comprehension and long-term retention. Similar effects on LLM use on learners' affective-motivational state and mental effort were found in Deng et al.'s meta-analysis [15]. Additionally, LLMs may sometimes provide learners with distractions that are interesting, but that compete with the primary task at hand.

At the same time, our exploratory analysis of student prompts suggests that another part of the answer lies in the unique benefits LLMs provide, which may have been genuinely helpful beyond what our primary analyses captured. The vast majority of LLM use was constructive rather than distracting or reductive, with students seeking additional information and deeper understanding. Students demonstrated remarkable curiosity, asking sophisticated questions that extended beyond the immediate text. For example, in a passage about apartheid in South Africa that briefly mentions Nelson Mandela's journey from prisoner to president, one student asked, "What was Mandela's life story?" Similarly, in a passage on the Cuban Missile Crisis that assumes some background knowledge of the Cold War, another student asked, "Why was America afraid of communism?" These explorations represent a different kind of active learning opportunity that may not result from note-taking alone, underscoring the LLM's potential to expand intellectual horizons. That said, these deeper inquiries may have involved tradeoffs: they could have competed with processing the core information in the passage, reducing performance on tested items, but they likely also enhanced learning in ways not captured by our tests, which focused only on the explicit and implied content within the texts.

Taken together, our findings demonstrate the value of combining LLM use and note-taking, which was not only more effective than LLM use alone but also students' preferred activity. This raises the opportunity and challenge of how to combine traditional evidence-based strategies like note-taking with the unique benefits offered by LLMs. Rather than viewing these as competing alternatives, we should think of them as complements that when thoughtfully integrated can enhance learning outcomes in ways that neither can achieve alone. A key to doing so is leveraging input from educators and researchers in the design and use of new LLM-based tools for learning, as has been key for past hybridisation of traditional and digital approaches [63;64].

Our work suggests several such directions. First and most easily would be to separate LLM use from note-taking. Under this model, students would first independently read a text, and then interact with an LLM to further clarify and explore its content. Following this they would take notes independently, without the ability to simply copy and paste output from the LLM. This would prevent students from taking shortcuts we have observed in this study, instead encouraging them to synthesise and internalise information themselves. This is a small but likely meaningful design choice that was not obvious to us a priori, but that emerged through our work and could be tested in future research.

Second, educators could actively train and guide students to use LLMs in ways that align with active learning strategies, such as asking targeted questions to clarify specific misunderstandings, engage in critical thinking, and integrate information, without overloading them with excessive information or reducing cognitive processing [36;35]. Likewise, educators could discourage the passive consumption of automatic summaries and explanations. This aligns with the conceptualisation of AI tools as "thought partners" that support existing human cognitive processes rather than disrupting them [9]. Going beyond learning activities, by guiding students to use LLMs more effectively, educators will help students develop their metacognitive skills more generally, which will make them better prepared to use these technologies in the long-term. Furthermore, software could be configured to support these goals by limiting distracting behaviour and encouraging productive use (plausibly by capturing data and using the LLM to provide feedback or nudges to the student based on their LLM interactions).

And third, educators could leverage insights from students' interactions with the LLM to better understand what concepts they are struggling with or what they are curious about. This could be done at an individual level but could also be conducted collectively for an entire class, possibly through the use of automated tools that collect and analyse student interactions and then provide data back to the educational instructors in a privacy-protecting way to surface insights. The results could be used to tailor future lessons, activities and group discussions. For example, through analysing the prompts in our experiments, it becomes clear that students were curious about the tenets of communism and why they provoked such fear and opposition in the U.S.

This research makes several contributions to the growing field of research examining the impact of LLMs in education. While much prior work has focused on the impact of LLMs on task performance and efficiency, the present study investigated aspects that are more fundamental to learning and cognition. In addition, it examined the effects of LLMs within a large sample of secondary school students coming from different school types, rather than amongst students in higher education, who have received much more research attention thus far [15] Such populations can be difficult to reach, especially when several study sessions are involved. In designing the study, we aimed to be authentic to students' experiences in school, ensuring the findings hold practical significance. In particular, we used texts that reflect the topics and difficulty that such students might come across in the classroom, and we compared the effects of LLM use with a learning activity that is, at least until now, commonly used.

One limitation of the present study is that students received no in-depth training for the different learning activities. While we provided instructions and a demonstration video for how to interact with the LLM and take notes, students did not have an opportunity to practice. This might have

been a particular disadvantage for the LLM conditions because students were less familiar with using LLMs than note-taking and might thus not have leveraged the activity as effectively. In addition, the study might have benefited from a baseline or passive reading condition to ascertain whether using the LLM to understand and learn a text provides benefits above passive reading (that is, to gauge its effectiveness *per se*). Another limitation is that we were practically constrained to a small set of retention and comprehension questions relative to the vast number of potential questions that could have been asked, although we sampled a wide range of content. Thus, we could have underestimated students' learning overall, with the exception of the free recall questions. Furthermore, the study was limited to a single, isolated activity outside of the context of normal use throughout an entire course of study. It is possible that repeated use or use in other settings (e.g., in everyday classrooms or independently for homework, unsupervised) could yield different results. Lastly, while we consider it a strength that we used texts that were appropriate to the student sample, it is possible that LLM usage might be more beneficial for texts that students struggle with, as indicated by a few students who stated they did not know what to ask the LLM. Hence, exploring the effects of LLM use for texts that go beyond students' current capabilities could further expand our understanding of potential applications.

It is crucial for future research to explore which ways of interacting with LLMs most effectively enhance learning outcomes. Future research must also explore the long-term consequences of LLM integration in learning contexts, particularly its impact on reading skills, independent problem-solving, and metacognition. Additionally, it will become vital to understand how these tools influence societal perceptions of effort, expertise, and achievement. The evolving role of LLMs and generative AI technology may shift the definition of essential expertise and change the landscape of necessary competencies across various fields[8]. Moving forward, it is vital for educators and society to identify which core skills remain indispensable in this new environment and to develop pedagogical strategies that ensure their preservation and growth[9]. This research marks only the beginning of understanding how to effectively use LLMs to complement existing activities and tools while maintaining students' cognitive engagement.

In summary, this study provides one of the first large-scale quantitative evidence on the effects of LLMs on reading comprehension and retention. Our findings reaffirm the importance of traditional strategies like note-taking, which foster deep cognitive engagement and strong learning outcomes. At the same time, LLMs introduce new possibilities for learning—offering opportunities to clarify, explore, and contextualise material—but these tools must be used with proper guidance aimed at enhancing, rather than bypassing, active learning. Rather than viewing these tools as a disruption to be resisted, educators and researchers have an opportunity to proactively shape their use to maximise learning potential. By doing so, we can prepare students to thrive in an AI-integrated world while preserving the focus, depth, and curiosity that define meaningful education.

Materials and Methods

This study comprised two stages: a piloting stage and a main study. The purpose of the piloting stage was to test the tasks and proposed procedures in the school context and amend them as appropriate. The methods and findings reported here are a part of the main study, which took place between March and July 2024.

Participants

Participants were 405 Year 10 students (aged 14-15 years) from seven secondary schools in England. Based on our exclusion criteria (see Supplementary Section 1.1), we retained 344 students for analysis. We made efforts to recruit 600 students but were unable to do so as we could not find enough schools before the start of the summer holidays. Recruitment methods included emailing

school headteachers in several counties and asking participating schools to contact other schools. The final school sample included three non-selective state schools, two grammar schools (one all girls, one all boys) and two independent schools, located in three different counties.

Once a school agreed to participate, all Year 10 students were invited to take part through the school's project lead. Information sheets were shared with students and their parents/guardians, after which both were asked to provide their informed written consent using an online Microsoft form. This study was conducted in line with the British Educational Research Association's[65] ethical guidelines. Ethical approval was provided by the research ethics committees of the researchers' institutions.

Experimental design and procedure

The study was a pre-registered randomised controlled experiment with within- and between-participant design elements, as illustrated in Figure 5. Conducted over two sessions spaced three days apart, the experiment consisted of a learning session followed by a test session.

Learning Session: In the learning session, students were tasked with understanding and learning two text passages on different history topics (Passage A and Passage B). Each passage was studied using a specific active learning activity (condition). The three conditions were:

- **LLM:** Students were asked to use an LLM chatbot we created to help them understand and learn the passage.
- **Notes:** Students were asked to take notes to help them understand and learn the passage.
- **LLM+Notes:** Students were asked to use our LLM chatbot as well as take notes to help them understand and learn the passage.

Students were randomly assigned to one of two groups:

- **Group 1:** Exposed to the *LLM* and *Notes* conditions.
- **Group 2:** Exposed to the *LLM* and *LLM+Notes* conditions.

Randomisation assigned 184 students to Group 1 (53.5%) and 160 to Group 2 (46.5%). The order of conditions and passages was randomised. During this session, students also completed survey questions about their learning experiences.

Test Session: In the test session, students answered comprehension and retention questions about the two passages (with passage order randomised) and completed survey questions regarding their general characteristics.

Timing: Students spent a mean of approximately 35 minutes on the learning session and 30 minutes on the test session.

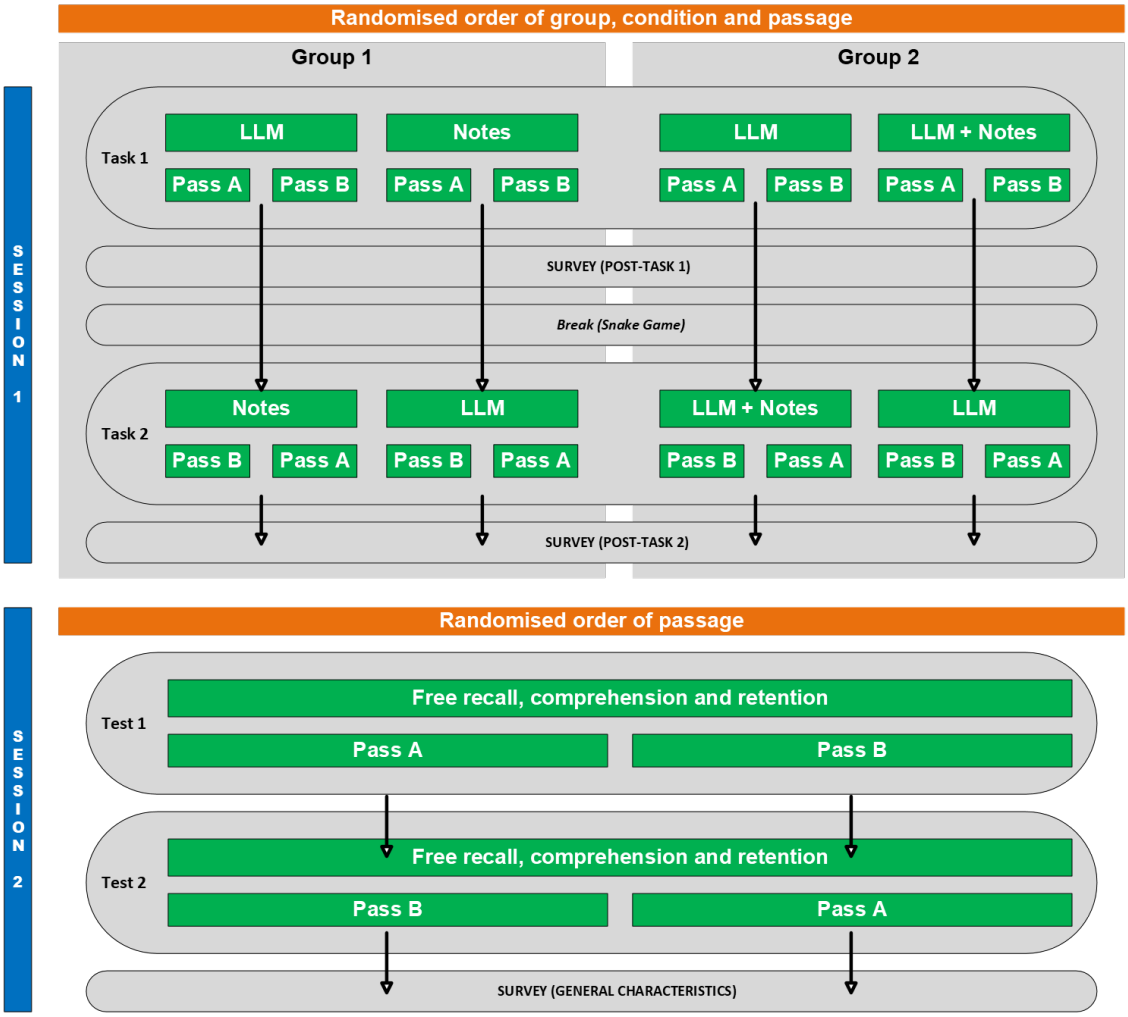

Figure 5: Study design illustrating the activities and their order during Session 1 and 2.

Setup and system

Both sessions took place in schools during regular school hours. Groups of students participated simultaneously in classrooms, with each student completing the sessions on an individual laptop or computer. At the start of each session, the experimenter or teacher read out a script with introductory instructions. They also monitored students during the entire session and answered their questions.

The experiment was a web app hosted on github.com that students accessed via the browser. For the LLM functionality in Session 1, the app made backend calls to private Azure Functions that accessed an Azure-hosted instance of OpenAI's GPT-3.5 turbo model. The LLM interactions were limited to Azure and did not go back to OpenAI. Participants could issue a maximum of 20 prompts. The LLM was customised with a meta-prompt that was not visible to students ("You are an AI chat bot that helps students read and comprehend the following passage: <text> Students can use this tool to define unfamiliar words, explain concepts, or summarise key points of the passage."). Figure 6 illustrates the task screen for the LLM+Notes condition. For the Notes and

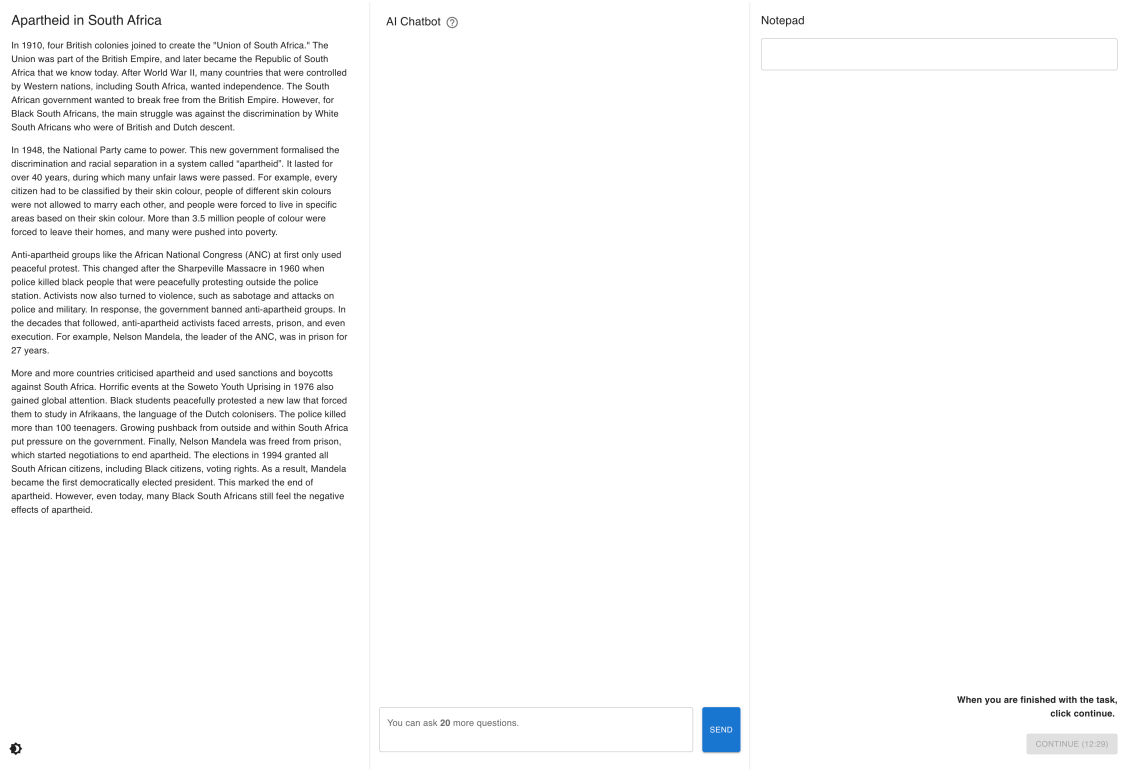

Figure 6: Example task screen for the LLM+Notes condition.

the LLM conditions, only the notepad or chatbot was displayed, respectively.

Learning task and materials (Session 1)

In the learning session, students read two passages on a history topic, each with a different learning activity. They were asked to understand and learn the content of the texts as best as they could. Notably, students had not been told that they would be tested on the materials. For each task, they first received instructions (see Supplementary Section 2.6 about the value of active reading, what it involves, and how the given reading activity might support active reading). They then received more detailed task instructions describing specific strategies, which were followed by a video demonstration of the task and interface. The suggested strategies were based on the active reading and comprehension literature[29;35;36;66]. The content and wording of the instructions for the three conditions were kept as similar as possible. Once the task started, students needed to remain on the task page for 10 (minimum) to 15 (maximum) minutes.

Each student read two expository text passages. Each passage covered a single topic which was included in at least one of the UK exam boards' GCSE History specifications: Apartheid in South Africa (Passage A) and The Cuban Missile Crisis (Passage B). The passages were adapted from two OpenStax textbooks (World History, Volume 2: from 1400; U.S. History). Substantial adaptations were made to ensure that the content and language difficulty as well as text features were comparable and appropriate for Year 10 students. Passages A and B had four paragraphs each and were nearly equal length (386 and 385 words), average word length (5.3 and 4.8 characters), word complexity (i.e., the average position of the words in the 10,000 most frequent English words list, 1986 and 1927), number of sentences (both 26) and CEFR level (both C1 – upper intermediate).

Table 2: Question types and scoring for literal retention, comprehension, and free recall

Outcome	Question Type (N Questions per Text)	Scoring	Maximum score
Literal retention	Short response - Cued recall (8)	For each literal piece of information: 0 - missing, incorrect or irrelevant 0.5 - incomplete or partially correct 1 - correct	10
	Multiple choice with four response options - Recognition (10)	0 – missing or incorrect 1 – correct	10
Comprehension	Short response - Cued recall (3)	For each idea: 0 - missing, incorrect or irrelevant 0.5 - incomplete or partially correct 1 - correct	12
Free recall	Open response (1)	For each literal piece of information/idea: 0 - incorrect or irrelevant 0.5 - incomplete or partially correct 1 - correct	50

Note: Two of the eight "Short response - Cued recall" questions for literal retention are worth two points each.

We divided each passage into 50 main ideas to ensure comparability and to aid scoring.

Test task and materials (Session 2)

In the test session, students were told that they would answer some questions about the passages they read in Session 1 as well as some general questions about the task and themselves. For each passage, there were 22 test questions assessing literal retention, comprehension and free recall. Table2 provides an overview of how the different constructs were assessed. As pre-registered, we used a single literal retention score, which was the sum of the short response and multiple-choice scores. The question order for both passages was free response, comprehension, literal retention (cued recall) and, finally, literal retention (recognition). Students had to spend at least three minutes and a maximum of five minutes on the free-recall questions. Questions were carefully sequenced and separated by screens where needed to avoid that previous questions would provide cues for later questions. Example questions can be found in Supplementary Table 11.

Literal retention questions required literal recall or recognition of information from the passage to provide a correct response. In order to succeed, students did not need background knowledge beyond understanding the vocabulary used in the passage. They did not need to make any knowledge-based inferences (elaborations), and no or only minimal text-based (bridging) inferences, such as connecting two consecutive sentences. Accordingly, literal retention questions targeted the surface and textbase level of representation.

In contrast, comprehension questions probed for deeper comprehension as they required students to make bridging inferences to connect information from several different locations in the text. Participants needed to make knowledge-based inferences to earn more points, inferring information that was implied but not explicitly stated. Accordingly, comprehension questions targeted the situation-model level of representation.

The short response and open response questions were scored by three independent raters who were PhD students in Education and/or Psychology who were blind to condition. They were trained to use a scoring scheme that provided general instructions, rules, and detailed explanations and examples for each question. As part of the training, and to demonstrate consistent and accurate use of the scheme, raters scored responses from 25 students and received feedback. Each rater then independently scored the full set of responses, including the questions for *both* passages, from approximately 140 students.

To assess inter-rater reliability, the full set of responses from 35 students (approximately 10% of the sample) was scored by all three raters. Reliability was evaluated using the intraclass-correlation coefficient (ICC) with a two-way model[67]. We measured absolute agreement and applied the single

measure approach as we ultimately used scores from a single rater for all but the 35 students in the reliability sample. For those students, we used the median of the three ratings in subsequent analyses. The inter-rater reliabilities for the combined cued-recall retention scores (one for Passage A and one for Passage B), the combined comprehension scores, and the free recall scores ranged between .97 and .99, indicating excellent reliability[67]. The lower bounds of the 95% confidence intervals were all above the .90 threshold for excellent reliability (see Supplementary Table 12).

Survey questions

All questions and response scales can be found in Supplementary Section 2.9. After each task in Session 1, students were asked to self-report on: the difficulty of the text and their familiarity with, and interest in, the topic; enjoyment, difficulty, and helpfulness of the learning activity, and likelihood of its future use; and the overall interest in the task, effort expenditure, and perceived task performance. Students were also asked to indicate whether they preferred any of the learning activities and why, whether they had ever used AI chatbots and if so, with what frequency, and, lastly, how often they had used these learning activities when reading a text for school.

After each test in Session 2, students were asked to rate their perceived test performance. At the end of the session, they were asked to indicate whether they had engaged in any learning related to the two texts in between sessions. Students were also asked to report their gender, their English language status, and whether they were taking GCSE History.

In addition, Free School Meals (FSM) eligibility data was obtained from schools as a measure of student socioeconomic disadvantage[68]. This is because eligibility for FSM is typically based on family income and other socioeconomic factors.

Analytic strategies

We did not deviate from our pre-registered analyses other than described here. First, we extended analyses to conduct qualitative analyses exploring why students preferred one learning activity over another. Second, while we initially planned to explore interaction effects between learning conditions and Gender, EAL, FSM, History GCSE, and School type, we did not do so given our smaller than planned sample size.

Quantitative analyses were run with Python 3.11 and R 4.4.2. We used a significance level of 0.05 (two-tailed) for all analyses. Effect sizes were estimated using Cohen's d, calculated as the mean difference divided by the standard deviation of paired differences for each variable.

Estimation of condition effects on text comprehension and retention

Missing data handling There were no missing data on the dependent variables because participants were excluded if they did not complete both tests (see exclusion criteria) and because any missing responses on individual questions were scored as 0 points. Missingness in covariates was minimal and only occurred for the variables Gender, EAL and History GCSE (5.23%, 1.16% and 1.16%, respectively). Missing data were handled using multiple imputation by chained equations (MICE) using the 'mice' package. Models were fitted on five imputed datasets and the results were pooled for combined estimates.

Mixed-effects regression We ran three linear mixed-effects regression models using the 'lme4' package, one for each outcome (i.e., literal retention, comprehension, free recall), where students were modelled as a random effect. Note that we pre-registered the regression for free recall as a secondary analysis but we are reporting it alongside the other outcomes for simplicity. The regression specification was as follows:

$$Y_{ij} = \beta_0 + \beta_1 \text{Condition}_{ij} + \beta_2 \text{Group}_{ij} + \beta_3 \text{School}_{ij} + \beta_4 \text{Text}_{ij} + \beta_5 \text{Task\_Order}_{ij} + \beta_6 \text{Test\_Order}_{ij} + \beta_7 \text{Gender}_{ij} + \beta_8 \text{FSM}_{ij} + \beta_9 \text{EAL}_{ij} + \beta_{10} \text{History}_{ij} + u_i j + \epsilon_{ij}$$

Where:

- Y_{ij} represents the outcome for student i in condition j.
- β_0 represents the intercept of the model.
- β_1 to β_{10} represent the coefficients for the fixed effects:
  - **Condition**: A categorical variable with three levels (0 = LLM, 1 = Notes, 2 = LLM+Notes).
  - **Group**: A binary variable indicating group membership.
  - **School**: A categorical variable with seven levels indicating school membership.
  - **Text**: A binary variable indicating which text student i studied in condition j.
  - **Task order**: A binary variable indicating whether student i did condition j first or second.
  - **Test order**: A binary variable indicating whether the text was tested first or second.
  - **Gender**: A categorical variable with four levels (0 = female, 1 = male, 2 = other, 3 = prefer not to say).
  - **FSM**: A binary variable indicating whether the student received free school meals or not.
  - **EAL**: A categorical variable indicating students' English language status (0 = first language, 1 = bilingual, 2 = other)
  - **History**: A binary variable indicating whether or not students take History GCSEs.
- u_{ij} represents the random intercept for each student.
- ϵ_{ij} represents the error term for student i in condition j.

As depicted in Figure 1, free recall scores were non-normally distributed, so we ran additional non-parametric permutation tests. Specifically, we used the 'infer' package in R to conduct paired permutation tests at the student level. These tests compared free recall scores between the LLM and Notes conditions in Group 1, and between the LLM and LLM+Notes conditions in Group 2. For each student, we calculated the difference between their two scores and averaged these differences across students. This test statistic was compared to a null distribution, generated by repeatedly randomising the signs of within-student differences and computing means. The process was repeated across all instances of imputed data, and the results were summarised by taking the median p-value across instances to yield a pooled p-value. Doing so gives similar findings to the mixed effects model: in Group 1 we find a significant difference for free recall between the Notes and LLM conditions ($p = 0.02$), but do not find evidence for a significant difference in free recall for Group 2 between the LLM+Notes vs. LLM conditions ($p = 0.80$).

Qualitative exploration of student prompts

To provide potential explanations for the effects of the LLM condition on reading comprehension and retention, we sought to understand what kind of prompts students made when using the LLM in planned exploratory analyses. The LLM prompts were analysed using a hierarchical coding scheme through GPT-4 in an automated Python script accessing the Azure OpenAI's API (deployment dated 2024-06-01). Temperature was set to 0 for deterministic outputs with a narrow sampling range (top-p=0.1) to ensure consistent classifications. The model was provided with detailed instructions and examples for each category, along with both texts that students were studying. Each prompt could receive multiple sub-codes.

The hierarchical coding scheme was developed through several iterations. The initial version was deductively and inductively developed by a researcher using active reading literature, students' task instructions, and piloting work. This scheme was expanded based on the API's suggestions and the API was then asked to code the data using the coding scheme. The researchers then iteratively refined the coding scheme based on checking portions of the API output. They merged, deleted, and added codes as needed and adapted code descriptions and examples to improve the quality of the API output. Finally, one of the researchers manually checked the API output for 500 prompts (approximately 10% of the data) and found an error rate of 5.6%. This was deemed to be an acceptable level. The assigned codes for these 500 prompts were adjusted where necessary, and the rest of the API output was left as it was. The final coding schemes for student prompts can be found in Supplementary Table 20.

Quantitative exploration of students' learning experience

As planned we explored a range of variables capturing students' learning experiences. More specifically, we compared students' learning experiences when using LLM vs. Notes and LLM vs. LLM+Notes using paired t-tests. We applied Bonferroni corrections to adjust for multiple comparisons. The t-tests were conducted using the 'tidyverse' package.

Qualitative exploration of students' activity preferences

We explored students' open response explanations for preferring one learning activity over another. The explanations were analysed by two of the authors with help from the API described above. Four preference groups were separately analysed:

1. LLM over Notes,
2. Notes over LLM,
3. LLM over LLM+Notes, and
4. LLM+Notes over LLM.

Each preference group had its own coding scheme which only included explanations for preferring the favoured activity over the non-favoured activity (i.e., benefits of note-taking were not coded if the student preferred the LLM over Notes). The initial schemes were developed by manually and deductively coding approximately 30% of responses of each preference group. Several codes could be applied to each response. The initial coding schemes, including the category label, description and examples were provided to the API alongside the data and general coding instructions. The API did not suggest any further helpful codes. The researchers then iteratively refined the coding schemes by manually checking portions of the API output. They merged, deleted, and added codes as well as refined code descriptions and examples before the API analysis was rerun. This process was repeated until both researchers were satisfied with the coding schemes. Due to the

small number of responses that had to be coded ($n = 278$), one researcher checked the entire API output and made adjustments where necessary. The final coding schemes for activity preferences can be found in Supplementary Section 2.11.

Data availability

All quantitative data will be made available upon publication. We will not provide the following qualitative data as that would risk sharing identifiable information: Students' LLM interactions (only the applied codes will be shared), students' notes, students' activity preferences (only applied codes will be shared).

Code availability

The corresponding code will be shared upon publication.

Ethics declarations

Competing interests

Some of the authors conduct research at a company that invests in generative AI and develops technology using generative AI models as a core component. The other authors are part of a publishing, assessment and learning organisation which increasingly uses AI in developing and operating assessment and learning products and services. However, this work is not connected to any specific product or monetisation efforts for either organisation.

Acknowledgements

We thank Dr Tom Benton and Dr Matthew Carroll for their valuable advice on the analyses conducted in this study.

Supplementary Material

Table of Contents

Supplementary Information

Supplementary Tables

- Specific instructions by Condition
- Test Questions
- Inter-rater Reliability Results
- Survey Questions and Response Scales
- Survey Questions and Response Scales (session 2)
- Learning Experiences and Perceptions
- Coding Scheme Activity Preferences
- Coding scheme: LLM over Notes preferences
- Coding scheme: Notes over LLM preferences
- Coding scheme: LLM+Notes over LLM preferences
- Coding Scheme Prompt Interactions
- Frequencies of Prompt Types

References

[1] Cecilia Ka Yuk Chan. A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, 20(1):38, July 2023. ISSN 2365-9440. doi: 10.1186/s41239-023-00408-3. URL https://doi.org/10.1186/s41239-023-00408-3.

[2] Abdulhadi Shoufan. Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access*, 11:38805–38818, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3268224. URL https://ieeexplore.ieee.org/document/10105236/?arnumber=10105236. Conference Name: IEEE Access.

[3] K. Aleksić-Maslać, F. Borović, and Z. Biočina. PERCEPTION AND USAGE OF CHAT GPT IN THE EDUCATION SYSTEM. *INTED2024 Proceedings*, pages 1842–1848, 2024. ISSN 2340-1079. doi: 10.21125/inted.2024.0511. URL https://library.iated.org/view/ALEKSICMASLAC2024PER. Conference Name: 18th International Technology, Education and Development Conference ISBN: 9788409592159 Meeting Name: 18th International Technology, Education and Development Conference Place: Valencia, Spain Publisher: IATED.

[4] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction*, February 2022. ISSN 1073-0516. doi: 10.1145/3511599. URL https://dl.acm.org/doi/10.1145/3511599. Just Accepted.

[5] Heather Johnston, Rebecca F. Wells, Elizabeth M. Shanks, Timothy Boey, and Bryony N. Parsons. Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, 20(1):2, February 2024. ISSN 1833-2595. doi: 10.1007/s40979-024-00149-4. URL https://doi.org/10.1007/s40979-024-00149-4.

[6] Duong Hoai Lan and Tran Minh Tung. Analyzing the Impact of Chat-GPT Usage by University Students in Vietnam. *Migration Letters*, 20(S10):259–268, November 2023. ISSN 1741-8992. doi: 10.59670/ml.v20iS10.5134. URL https://migrationletters.com/index.php/ml/article/view/5134. Number: S10.

[7] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 2023.

[8] Stefan E. Huber, Kristian Kiili, Steve Nebel, Richard M. Ryan, Michael Sailer, and Manuel Ninaus. Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning. *Educational Psychology Review*, 36(1):25, February 2024. ISSN 1573-336X. doi: 10.1007/s10648-024-09868-z. URL https://doi.org/10.1007/s10648-024-09868-z.

[9] Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M. Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, Hanaa Albanna, Mousa Ahmad Albashrawi, Adil S. Al-Busaidi, Janarthanan Balakrishnan, Yves Barlette, Sriparna Basu, Indranil Bose, Laurence Brooks, Dimitrios Buhalis, Lemuria Carter, Soumyadeb Chowdhury, Tom Crick, Scott W. Cunningham, Gareth H. Davies, Robert M. Davison, Rahul Dé, Denis Dennehy, Yanqing Duan, Rameshwar Dubey, Rohita Dwivedi, John S. Edwards, Carlos Flavián, Robin Gauld, Varun Grover, Mei-Chih Hu, Marijn Janssen, Paul Jones, Iris Junglas, Sangeeta Khorana, Sascha Kraus, Kai R. Larsen, Paul Latreille, Sven Laumer, F. Tegwen Malik, Abbas Mardani, Marcello Mariani, Sunil Mithas, Emmanuel Mogaji, Jeretta Horn Nord, Siobhan O'Connor, Fevzi Okumus, Margherita Pagani, Neeraj Pandey, Savvas Papagiannidis, Ilias O. Pappas, Nishith Pathak, Jan Pries-Heje, Ramakrishnan Raman, Nripendra P. Rana, Sven-Volker Rehm, Samuel Ribeiro-Navarrete, Alexander Richter, Frantz Rowe, Suprateek Sarker, Bernd Carsten Stahl, Manoj Kumar Tiwari, Wil van der Aalst, Viswanath Venkatesh, Giampaolo Viglia, Michael Wade, Paul Walton, Jochen Wirtz, and Ryan Wright. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71:102642, August 2023. ISSN 0268-4012. doi: 10.1016/j.ijinfomgt.2023.102642. URL https://www.sciencedirect.com/science/article/pii/S0268401223000233.

[10] Jun-Jie Zhu, Jinyue Jiang, Meiqi Yang, and Zhiyong Jason Ren. ChatGPT and Environmental Research. *Environmental Science & Technology*, 57(46):17667–17670, November 2023. ISSN 0013-936X. doi: 10.1021/acs.est.3c01818. URL https://doi.org/10.1021/acs.est.3c01818. Publisher: American Chemical Society.

[11] Alex Barrett and Austin Pack. Not quite eye to A.I.: student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education*, 20(1):59, November 2023. ISSN 2365-9440. doi: 10.1186/s41239-023-00427-0. URL https://doi.org/10.1186/s41239-023-00427-0.

[12] Aiste Steponenaite and Basel Barakat. Plagiarism in AI Empowered World. In Margherita Antona and Constantine Stephanidis, editors, *Universal Access in Human-Computer Interaction*, pages 434–442, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-35897-5. doi: 10.1007/978-3-031-35897-5_31.

[13] Ofcom. Online nation 2024 report. Technical report, Ofcom, November 2024. URL https://www.ofcom.org.uk/media-use-and-attitudes/online-habits/online-nation/.

[14] Walton Family Foundation. Teachers and Students Embrace ChatGPT for Education. Technical report, Walton Family Foundation, March 2023. URL https://www.waltonfamilyfoundation.org/learning/teachers-and-students-embrace-chatgpt-for-education. Section: Learning.

[15] Ruiqi Deng, Maoli Jiang, Xinlu Yu, Yuyan Lu, and Shasha Liu. Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies. *Computers Education*, 227:105224, 2025. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2024.105224. URL https://www.sciencedirect.com/science/article/pii/S0360131524002380.

[16] Jeffrey R. Binder and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11):527–536, November 2011. ISSN 1879-307X. doi: 10.1016/j.tics.2011.10.001.

[17] Danielle S. McNamara and Joe Magliano. Toward a comprehensive model of comprehension. In *The psychology of learning and motivation, Vol. 51*, The psychology of learning and motivation, pages 297–384. Elsevier Academic Press, San Diego, CA, US, 2009. ISBN 978-0-12-374489-0. doi: 10.1016/S0079-7421(09)51009-2.

[18] Walter Kintsch. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182, 1988. ISSN 1939-1471. doi: 10.1037/0033-295X.95.2.163. Place: US Publisher: American Psychological Association.

[19] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, May 2007. ISSN 1471-0048. doi: 10.1038/nrn2113. URL https://www.nature.com/articles/nrn2113. Publisher: Nature Publishing Group.

[20] Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312, May 2024. ISSN 1471-0048. doi: 10.1038/s41583-024-00802-4. URL https://www.nature.com/articles/s41583-024-00802-4. Publisher: Nature Publishing Group.

[21] Rolf A. Zwaan and Gabriel A. Radvansky. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185, 1998. ISSN 1939-1455. doi: 10.1037/0033-2909.123.2.162. Place: US Publisher: American Psychological Association.

[22] Junhua Ding, Keliang Chen, Haoming Liu, Lin Huang, Yan Chen, Yingru Lv, Qing Yang, Qihao Guo, Zaizhu Han, and Matthew A. Lambon Ralph. A unified neurocognitive model of semantics language social behaviour and face recognition in semantic dementia. *Nature Communications*, 11(1):2595, May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16089-9. URL https://www.nature.com/articles/s41467-020-16089-9. Publisher: Nature Publishing Group.

[23] Kate Cain and Jane Oakhill. Reading Comprehension Difficulties: Correlates, Causes, and Consequences. In *Children's comprehension problems in oral and written language: A cognitive perspective*, Challenges in language and literacy, pages 41–75. The Guilford Press, New York, NY, US, 2007. ISBN 978-1-59385-443-0.

[24] Meredyth Daneman and Patricia A. Carpenter. Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, 19(4):450–466, 1980. ISSN 0022-5371. doi: 10.1016/S0022-5371(80)90312-6. Place: Netherlands Publisher: Elsevier Science.

[25] Charles A. Perfetti, Nicole Landi, and Jane Oakhill. The Acquisition of Reading Comprehension Skill. In *The science of reading: A handbook*, Blackwell handbooks of developmental psychology, pages 227–247. Blackwell Publishing, Malden, 2005. ISBN 978-1-4051-1488-2. doi: 10.1002/9780470757642.ch13.

[26] Jane V. Oakhill, Molly S. Berenhaus, and Kate Cain. Children's reading comprehension and comprehension difficulties. In *The Oxford handbook of reading*, Oxford library of psychology, pages 344–360. Oxford University Press, New York, NY, US, 2015. ISBN 978-0-19-932457-6. doi: 10.1093/oxfordhb/9780199324576.001.0001.

[27] Keith E. Stanovich. Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(4):360–407, 1986. ISSN 1936-2722. doi: 10.1598/RRQ.21.4.1. Place: US Publisher: International Reading Association.

[28] A. C. Graesser, M. Singer, and T. Trabasso. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371–395, July 1994. ISSN 0033-295X. doi: 10.1037/0033-295x.101.3.371.

[29] Danielle S. McNamara, Irwin B. Levinstein, and Chutima Boonthum. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2):222–233, May 2004. ISSN 1532-5970. doi: 10.3758/BF03195567. URL `https://doi.org/10.3758/BF03195567`.

[30] John T. Guthrie and Allan Wigfield. Engagement and motivation in reading. In *Handbook of reading research, Vol. III*, pages 403–422. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 2000. ISBN 978-0-8058-2398-1 978-0-8058-2399-8.

[31] Tracy Linderholm, Sandra Virtue, Yuhtsuen Tzeng, and Paul van den Broek. Fluctuations in the Availability of Information During Reading: Capturing Cognitive Processes Using the Landscape Model. pages 165–186. December 2018. ISBN 978-1-315-04610-5. doi: 10.4324/9781315046105-5.

[32] Fergus I. M. Craik. Levels of processing: Past, present . . . and future? *Memory*, 10(5-6): 305–318, 2002. ISSN 1464-0686. doi: 10.1080/09658210244000135. Place: United Kingdom Publisher: Taylor & Francis.

[33] Fergus I. M. Craik and Endel Tulving. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104(3):268–294, 1975. ISSN 1939-2222. doi: 10.1037/0096-3445.104.3.268. Place: US Publisher: American Psychological Association.

[34] John R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295, June 1983. ISSN 0022-5371. doi: 10.1016/S0022-5371(83)90201-3. URL `https://www.sciencedirect.com/science/article/pii/S0022537183902013`.

[35] Danielle S. McNamara, editor. *Reading comprehension strategies: Theories, interventions, and technologies*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, 2007.

[36] Michelene T. H. Chi. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science*, 1(1):73–105, 2009. ISSN 1756-8765. doi: 10.1111/j.1756-8765.2008.01005.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2008.01005.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1756-8765.2008.01005.x.

[37] Rose Luckin, Wayne Holmes, and Laurie B Forcier. Intelligence Unleashed: An argument for AI in Education. Technical report, Open Ideas at Pearson / UCL, 2016. URL https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/innovation/Intelligence-Unleashed-Publication.pdf.

[38] Wayne Holmes, Maya Bialik, and Charles Fadel. *Artificial Intelligence in Education. Promise and Implications for Teaching and Learning.* March 2019. ISBN 978-1-79429-370-0.

[39] Margherita Bernabei, Silvia Colabianchi, Andrea Falegnami, and Francesco Costantino. Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence*, 5:100172, October 2023. doi: 10.1016/j.caeai.2023.100172.

[40] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*, pages 27–43, Lugano and Virtual Event Switzerland, August 2022. ACM. ISBN 978-1-4503-9194-8. doi: 10.1145/3501385.3543957. URL https://dl.acm.org/doi/10.1145/3501385.3543957.

[41] Harsh Kumar, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. Math Education With Large Language Models: Peril or Promise? 2023.

[42] John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2):261–292, 2019. ISSN 1573-336X. doi: 10.1007/s10648-019-09465-5. Place: Germany Publisher: Springer.

[43] Richard E. Mayer. Should There Be a Three-Strikes Rule Against Pure Discovery Learning? *American Psychologist*, 59(1):14–19, 2004. ISSN 1935-990X. doi: 10.1037/0003-066X.59.1.14. Place: US Publisher: American Psychological Association.

[44] Fergus I. M. Craik and Robert S. Lockhart. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6):671–684, December 1972. ISSN 0022-5371. doi: 10.1016/S0022-5371(72)80001-X. URL https://www.sciencedirect.com/science/article/pii/S002253717280001X.

[45] Xiaoming Zhai, Matthew Nyaaba, and Wenchao Ma. Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science?, January 2024. URL http://arxiv.org/abs/2401.15081. arXiv:2401.15081.

[46] Faycal Farhi, Riadh Jeljeli, Ibtehal Aburezeq, Fawzi Fayez Dweikat, Samer Ali Al-shami, and Radouane Slamene. Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Computers and Education: Artificial Intelligence*, 5:100180, January 2023. ISSN 2666-920X. doi: 10.1016/j.caeai.2023.100180. URL https://www.sciencedirect.com/science/article/pii/S2666920X23000590.

[47] Hao Yu and Yunyun Guo. Generative artificial intelligence empowers educational reform: current status, issues, and prospects. *Frontiers in Education*, 8:1183162, June 2023. ISSN 2504-284X. doi: 10.3389/feduc.2023.1183162. URL https://www.frontiersin.org/articles/10.3389/feduc.2023.1183162/full.

[48] Elizabeth Ligon Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society*, pages 56–64. Worth Publishers, New York, NY, US, 2011. ISBN 978-1-4292-3043-8.

[49] Michelene Chi, Stephanie Siler, Heisawn Jeong, Takashi Yamauchi, and Robert Hausmann. Learning from human tutoring. *Cognitive Science*, 25:471–533, July 2001. doi: 10.1016/S0364-0213(01)00044-1.

[50] Alvaro Pascual-Leone, Amir Amedi, Felipe Fregni, and Lotfi B. Merabet. The plastic human brain cortex. *Annual Review of Neuroscience*, 28:377–401, 2005. ISSN 0147-006X. doi: 10.1146/annurev.neuro.27.070203.144216.

[51] S. Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, April 2001. ISSN 0010-0277. doi: 10.1016/s0010-0277(00)00123-2.

[52] Keiichi Kobayashi. What limits the encoding eVect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology*, 2005.

[53] Kenneth A. Kiewra. A review of note-taking: The encoding storage paradigm and beyond. *Educational Psychology Review*, 1(2):147–172, 1989. ISSN 1573-336X. doi: 10.1007/BF01326640. Place: Germany Publisher: Springer.

[54] Kenneth A. Kiewra. Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist*, 20(1):23–32, 1985. ISSN 1532-6985. doi: 10.1207/s15326985ep2001_4. Place: US Publisher: Lawrence Erlbaum.

[55] Mark Bohay, Daniel P. Blakely, Andrea K. Tamplin, and Gabriel A. Radvansky. Note taking, review, memory, and comprehension. *The American Journal of Psychology*, 124(1):63–73, 2011. ISSN 0002-9556. doi: 10.5406/amerjpsyc.124.1.0063.

[56] Dung C. Bui and Joel Myerson. The role of working memory abilities in lecture note-taking. *Learning and Individual Differences*, 33:12–22, 2014. ISSN 1873-3425. doi: 10.1016/j.lindif.2014.05.002. Place: Netherlands Publisher: Elsevier Science.

[57] Ralf Rummer, Judith Schweppe, Kathleen Gerst, and Simon Wagner. Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology. Applied*, 23(3):293–300, September 2017. ISSN 1939-2192. doi: 10.1037/xap0000134.

[58] Lisa Geraci, Nikhil Kurpad, Rachel Tirso, Kathryn N. Gray, and Yuxiang Wang. Metacognitive errors in the classroom: The role of variability of past performance on exam prediction accuracy. *Metacognition and Learning*, 2022. doi: 10.1007/s11409-022-09326-7. URL `https://doi.org/10.1007/s11409-022-09326-7`. Advance online publication.

[59] Robert A. Bjork, John Dunlosky, and Nate Kornell. Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, 64(1):417–444, January 2013. ISSN 0066-4308, 1545-2085. doi: 10.1146/annurev-psych-113011-143823. URL `https://www.annualreviews.org/doi/10.1146/annurev-psych-113011-143823`.

[60] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, Dec 1999. doi: 10.1037//0022-3514.77.6.1121.

[61] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642902. URL `https://doi.org/10.1145/3613904.3642902`.

[62] Axel Grund, Stefan Fries, Matthias Nückles, Alexander Renkl, and Julian Roelle. When is Learning "Effortful"? Scrutinizing the Concept of Mental Effort in Cognitively Oriented Research from a Motivational Perspective. *Educational Psychology Review*, 36(1):11, March 2024. ISSN 1040-726X, 1573-336X. doi: 10.1007/s10648-024-09852-7. URL https://link.springer.com/10.1007/s10648-024-09852-7.

[63] Louise Starkey. A review of research exploring teacher preparation for the digital age. *Cambridge Journal of Education*, 50(1):37–56, 2020. doi: 10.1080/0305764X.2019.1625867.

[64] Honghong Wang and Weiping Shi. Practical approaches to integrated values education for foreign language majors. *Foreign Language World*, (6):38–45, 2021.

[65] British Educational Research Association. Ethical Guidelines for Educational Research, fourth edition, 2018. URL https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018.

[66] P. David Pearson, Laura R. Roehler, Janice A. Dole, and Gerald G. Duffy. Developing expertise in reading comprehension: What should be taught? How should it be taught? Technical Report 512, University of Illinois Urbana-Champaign Center for the Study of Reading, 1990. URL https://hdl.handle.net/2142/17648. Publisher: Champaign, Ill. : University of Illinois at Urbana-Champaign, Center for the Study of Reading.

[67] Terry K Koo and Mae Y Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. 2016.

[68] Chris Taylor. The reliability of free school meal eligibility as a measure of socio-economic disadvantage: Evidence from the millennium cohort study in wales. *British Journal of Educational Studies*, 66(1):29–51, 2018. doi: 10.1080/00071005.2017.1330464.

1 Supplementary Information

1.1 Participant Exclusion Criteria

Participants (n=61) were excluded for the following reasons:

1. Did not take part in Session 2 (n=36)
2. Did not complete both tasks in Session 1 (and/or withdrew intentionally) (n=2)
3. Stopped Session 2 before attempting all comprehension and retention questions (n=8)
4. Completed Session 2 in 10 minutes or less (n=1)
5. Reported substantially different prior knowledge of the two topics (3-point difference on a 5-point Likert-scale item) (n=13)
6. Cheated during a session (as observed by researcher, including opening a different browser to look up answers, copying answers from others, continuing conversation with neighbours). Responses of suspicious students were scanned and compared with that of other students in the same group. If suspicion confirmed based on responses (e.g., high overlap with a student), these were excluded (n=1)

2 Supplementary Tables

2.1 Student Characteristics

Table 3: Student characteristics by group and overall totals (after exclusion, N = 344)

Characteristic	**Group 1** N students (%)	**Group 2** N students (%)	**Total** N students (%)
Male	102 (29.7%)	78 (22.7%)	180 (52.3%)
Female	57 (16.6%)	63 (18.3%)	120 (34.9%)
Other	1 (0.3%)	1 (0.3%)	2 (0.6%)
Prefer not to say	2 (0.6%)	0 (0.0%)	2 (0.6%)
FSM_Yes	9 (2.6%)	10 (2.9%)	19 (5.5%)
FSM_No	160 (46.5%)	163 (47.4%)	323 (93.9%)
EAL_Yes	130 (37.8%)	117 (34.0%)	247 (71.8%)
EAL_Other Language	2 (0.6%)	3 (0.9%)	5 (1.5%)
EAL_Bilingual	35 (10.2%)	29 (8.4%)	64 (18.6%)
History_Yes	99 (28.8%)	80 (23.3%)	179 (52.0%)
History_No	81 (23.5%)	58 (16.9%)	139 (40.4%)

2.2 Familiarity with Learning Activities

Table 4: Frequencies of prior learning activity use

Activity and frequency	Group 1 N students (%)	Group 2 N students (%)
Note-taking for learning		
Never	7 (3.8%)	6 (3.8%)
Rarely	34 (18.5%)	25 (15.6%)
Sometimes	47 (25.5%)	44 (27.5%)
Often	69 (37.5%)	70 (43.8%)
Always	22 (12.0%)	17 (10.6%)
LLM use for learning		
Never	32 (25.6%)	19 (18.1%)
Rarely	45 (36.0%)	44 (41.9%)
Sometimes	29 (23.2%)	26 (24.8%)
Often	15 (12.0%)	15 (14.3%)
Always	4 (3.2%)	1 (1.0%)
LLM + Notes for learning		
Never	-	1 (1.6%)
Rarely	-	31 (48.4%)
Sometimes	-	23 (35.9%)
Often	-	8 (12.5%)
Always	-	1 (1.6%)
Prior LLM use		
Yes	125 (70.2%)	105 (64.0%)
No	53 (29.8%)	59 (36.0%)
Frequency of LLM use amongst users		
Less than once a week	74 (59.2%)	68 (64.8%)
One or two days a week	28 (22.4%)	33 (31.4%)
Three to five days a week	11 (8.8%)	5 (4.8%)
Most days of the week	12 (9.6%)	1 (1.0%)

2.3 Descriptive Statistics

Table 5: Descriptive statistics for comprehension, literal retention, and free recall across conditions.

Measure	Condition	Mean (M)	Standard Deviation (SD)
Comprehension (max 12 points)	Notes	4.89	2.52
	LLM + Notes	4.11	2.65
	LLM only (Group 1)	4.00	2.44
	LLM only (Group 2)	3.80	2.47
Literal retention (max 20 points)	Notes	10.8	4.29
	LLM + Notes	9.68	4.83
	LLM only (Group 1)	8.83	3.96
	LLM only (Group 2)	8.95	4.29
Free recall (max 50 points)	Notes	5.36	5.49
	LLM Group 1	4.32	4.15
	LLM Group 2	4.32	4.63
	LLM + Notes	4.20	5.07

2.4 Mixed Effects Regression Results

Table 6: Model coefficients for literal retention, comprehension, and free recall

Term	Estimate	Std. Error	95% CI	Statistic	df	p-value	d
Literal retention							
Intercept	8.2429	0.7966	[6.68, 9.81]	10.3476	489.3004	7.95×10^{-23}	-
Condition LLM_notes	0.5668	0.2752	[0.03, 1.11]	2.0597	660.4521	0.0398	0.132
Condition notes	1.9188	0.2559	[1.42, 2.42]	7.4974	663.2789	2.09×10^{-13}	0.443
Group 1	-0.6147	0.4155	[-1.43, 0.20]	-1.4793	661.9230	0.1395	-0.143
school_id S03	-0.8645	0.5993	[-2.04, 0.31]	-1.4424	638.7162	0.1497	-0.198
school_id S01	-1.9789	0.8005	[-3.55, -0.41]	-2.4720	657.4886	0.0137	-0.465
school_id S05	-0.3908	0.8562	[-2.07, 1.29]	-0.4564	612.9203	0.6483	-0.094
school_id S02	1.2932	0.5514	[0.21, 2.37]	2.3452	643.8234	0.0193	0.299
school_id S07	2.7561	1.1408	[0.52, 4.99]	2.4160	663.8251	0.0160	0.623
school_id S04	-4.7045	0.8102	[-6.29, -3.12]	-5.8067	641.0030	1.00×10^{-8}	-1.075
Text Cuba	1.5218	0.1880	[1.15, 1.89]	8.0952	663.5151	2.74×10^{-15}	0.351
Task_order 0	0.2310	0.1880	[-0.14, 0.60]	1.2283	659.9704	0.2198	0.052
Test_order 0	0.5186	0.1875	[0.15, 0.89]	2.7656	663.7540	0.0058	0.119
Gender (Male)	0.8396	0.4609	[-0.06, 1.74]	1.8217	335.9448	0.0694	0.193
Gender (Other)	1.1737	1.5839	[-1.93, 4.28]	0.7410	187.9029	0.4596	0.228
Gender (Prefer not to say)	1.7770	1.4362	[-1.04, 4.59]	1.2373	474.9248	0.2166	0.226
FSM (Yes)	-0.9135	0.8574	[-2.59, 0.77]	-1.0654	653.1653	0.2871	-0.207
EAL (Bilingual)	0.4650	0.4780	[-0.47, 1.40]	0.9728	645.1354	0.3310	0.116
EAL (Other)	-0.3369	1.6161	[-3.50, 2.83]	-0.2085	660.9281	0.8349	-0.027
History (No)	-1.5365	0.3832	[-2.29, -0.79]	-4.0095	641.2946	6.80×10^{-5}	-0.351
Comprehension							
Intercept	4.0264	0.4409	[3.16, 4.89]	9.1318	638.9518	8.77×10^{-19}	-
Condition LLM_notes	0.3533	0.1785	[0.00, 0.70]	1.9792	655.5471	0.0482	0.142
Condition notes	0.9500	0.1658	[0.62, 1.28]	5.7306	662.6375	1.52×10^{-8}	0.382
Group 1	-0.0735	0.2395	[-0.54, 0.40]	-0.3068	657.2449	0.7591	-0.033
school_id S03	-0.9749	0.3320	[-1.63, -0.32]	-2.9365	655.1779	0.0034	-0.399
school_id S01	-1.9371	0.4438	[-2.81, -1.07]	-4.3645	662.1221	1.48×10^{-5}	-0.783
school_id S05	-0.3167	0.4735	[-1.24, 0.61]	-0.6688	648.4704	0.5039	-0.142
school_id S02	0.5254	0.3052	[-0.07, 1.12]	1.7215	659.5381	0.0856	0.201
school_id S07	0.9683	0.6335	[-0.27, 2.21]	1.5284	663.5186	0.1269	0.377
school_id S04	-2.9725	0.4493	[-3.85, -2.09]	-6.6154	651.4740	7.74×10^{-11}	-1.192
Text Cuba	-0.6057	0.1218	[-0.84, -0.37]	-4.9727	662.4076	8.42×10^{-7}	-0.245
Task_order 0	0.0428	0.1219	[-0.20, 0.28]	0.3508	657.5431	0.7258	0.015
Test_order 0	0.6679	0.1215	[0.43, 0.91]	5.4958	662.7896	5.55×10^{-8}	0.266
Gender (Male)	0.2287	0.2517	[-0.26, 0.72]	0.9086	542.3928	0.3640	0.078
Gender (Other)	0.0375	0.9339	[-1.79, 1.87]	0.0401	102.4863	0.9681	0.574
Gender (Prefer not to say)	1.5360	0.9257	[-0.28, 3.35]	1.6593	68.4482	0.1016	0.006
FSM (Yes)	-0.6056	0.4786	[-1.54, 0.33]	-1.2655	626.0565	0.2062	-0.236
EAL (Bilingual)	0.5813	0.2649	[0.06, 1.10]	2.1943	655.2427	0.0286	0.228
EAL (Other)	-0.2195	0.9140	[-2.01, 1.57]	-0.2402	556.3704	0.8103	-0.103
History (No)	-0.6719	0.2138	[-1.09, -0.25]	-3.1423	613.1612	0.0018	-0.262
Free recall							
Intercept	4.4052	0.8507	[2.74, 6.08]	5.1786	662.4966	2.97×10^{-7}	-
Condition LLM_notes	-0.0847	0.4590	[-0.98, 0.81]	-0.1846	661.9195	0.8536	-0.015
Condition notes	1.0185	0.4269	[0.18, 1.86]	2.3856	663.2739	0.0173	0.211
Group 1	-0.2703	0.4958	[-1.24, 0.70]	-0.5452	662.0547	0.5858	-0.058
school_id S03	-0.4702	0.6185	[-1.68, 0.74]	-0.7603	663.5556	0.4474	-0.086
school_id S01	-0.9612	0.8290	[-2.59, 0.66]	-1.1595	660.3122	0.2467	-0.189
school_id S05	2.1564	0.8819	[0.43, 3.89]	2.4452	662.7977	0.0147	0.459
school_id S02	2.7874	0.5687	[1.67, 3.90]	4.9012	663.9081	1.20×10^{-6}	0.578
school_id S07	2.2260	1.1824	[-0.09, 4.54]	1.8827	663.2415	0.0602	0.459
school_id S04	-2.3075	0.8366	[-3.95, -0.67]	-2.7583	663.2134	0.0060	-0.468
Text Cuba	-0.1187	0.3137	[-0.73, 0.50]	-0.3783	662.8799	0.7053	-0.027
Task_order 0	-0.1370	0.3134	[-0.75, 0.48]	-0.4372	662.9483	0.6621	-0.029
Test_order 0	-0.3089	0.3130	[-0.92, 0.31]	-0.9870	663.8172	0.3240	-0.062
Gender (Male)	0.7972	0.4653	[-0.11, 1.71]	1.7133	662.1998	0.0871	0.178
Gender (Other)	1.5025	1.6550	[-1.74, 4.75]	0.9079	586.1239	0.3643	0.336
Gender (Prefer not to say)	-0.7067	1.7223	[-4.08, 2.67]	-0.4103	284.0426	0.6819	-0.249
FSM (Yes)	-0.0013	0.8884	[-1.74, 1.74]	-0.0014	660.6054	0.9886	0.016
EAL (Bilingual)	-0.4993	0.4958	[-1.47, 0.47]	-1.0070	644.7815	0.3143	-0.104
EAL (Other)	-0.7021	1.6974	[-4.03, 2.62]	-0.4137	647.6784	0.6793	-0.157
History (No)	-1.0261	0.3967	[-1.80, -0.25]	-2.5868	658.8462	0.0099	-0.210

2.5 Behavioural Engagement

Table 7: Behavioural engagement with the LLM and note-taking, including queries made, words in notes, and time on task. Significant differences in time spent on tasks are highlighted for comparison between conditions.

Measure	**Condition**	**Mean (M)**	**Standard Deviation (SD)**
Number of queries	Group 1 (LLM + Notes)	10.98	6.46
	Group 2 (LLM only)	9.21	5.72
	Group 2 (LLM + Notes)	6.02	4.64
Words in notes	Group 1 (Notes)	100.74	115.63
	Group 2 (LLM + Notes)	103.83	158.24
Trigram overlap (%)	Substantial overlap ($\geq 70\%$)		25.63%
Trigram overlap (%)	High overlap ($\geq 90\%$)		16.25%
Time on task (minutes)	Group 1 (LLM)	-0.80	95% CI [-1.15, -0.46], $d = -0.34$
	Group 1 (Notes)	10-15 range	-
	Group 2 (LLM only)	-1.54	95% CI [-1.91, -1.17], $d = -0.66$
	Group 2 (LLM + Notes)	10-15 range	-

2.6 Student Task Instructions

Table 8: Introduction to active reading (common across all conditions)

When you are trying to learn and understand a text, **active reading** can be a useful strategy. It can help you to process the information more deeply and thus to learn better. Active reading involves:

- figuring out what the main ideas and concepts in the text are,
- what they mean,
- how they relate to each other, and
- asking questions about the information and then trying to answer them.

Table 9: Learning activity introduction by condition

Condition	**Activity introduction**
Notes	Your task is to try to understand and learn a history text. To do so, please actively **read the text** and **take notes** to help you. Taking notes is an important part of active reading. It is not about copying a lot of information from the text. Instead, find the key information in a section, think about what it means, and note it down in your own words.
LLM	Your task is to try to understand and learn a history text. To do so, please actively **read the text** and **use an AI chatbot** to help you. Having a conversation with the AI chatbot might help you to read more actively. You can ask different questions about the text to help you understand what happened. It may also help you to identify and understand key information.
LLM+Notes	Your task is to try to understand and learn a history text. To do so, please actively **read the text**, **use an AI chatbot**, and **take notes** to help you. Having a conversation with the AI chatbot might help you to read more actively. You can ask different questions about the text to help you understand what happened. It may also help you to identify and understand key information. Taking notes is also important for active reading. It is not about copying a lot of information from the text. Instead, find the key information in a section, think about what it means, and note it down in your own words.

Table 10: Specific instructions by condition

Condition	Specific instructions
Notes	Actively **read the text** and **take notes** as you go along. Even if you think you understand everything, try doing so as best as you can. Think about the following things and note them down to help you: • The meaning of important words and concepts • The meaning of complex sentences • The key points or ideas, such as the dates, places, people and events • The connections between places, people and events • What happened, and why and how it happened • Similarities and differences between ideas and concepts • Your understanding of the text
LLM	Actively **read the text** and **use the AI chatbot** as you go along. Even if you think you understand everything, try doing so as best as you can. Think about the following things and use the AI chatbot to help you. For example, you can use it to: • Explain the meaning of important words and concepts • Rephrase or simplify complex sentences and explain them • Summarise the text and identify the key points or ideas, such as the dates, places, people and events • Clarify information you don't understand • Explain the connections between places, people and events • Explain what happened, and why and how it happened • Identify similarities and differences between ideas and concepts • Check your understanding of the text You can also: • Ask the AI chatbot for more explanation if you do not understand its response or think that something might not be quite right • Ask follow-up questions • Ask it to use bullet points, make its answer shorter, or use simpler language
LLM+Notes	Actively **read the text**, **use the AI chatbot** and **take notes** as you go along. Even if you think you understand everything, try doing so as best as you can. Think about the following things, and use the AI chatbot and take notes to help you. For example, you can use the AI chatbot to: • Explain the meaning of important words and concepts • Rephrase or simplify complex sentences and explain them • Summarise the text and identify the key points or ideas, such as the dates, places, people and events • Clarify information you don't understand • Explain the connections between places, people and events • Explain what happened, and why and how it happened • Identify similarities and differences between ideas and concepts • Check your understanding of the text You can also: • Ask the AI chatbot for more explanation if you do not understand its response or think that something might not be quite right • Ask follow-up questions • Ask it to use bullet points, make its answer shorter, or use simpler language

2.7 Test Questions

Table 11: Example questions for literal retention, comprehension, and free recall

Construct Item type	**Example question**
Literal retention	
Short response	What horrific event happened at the Soweto Youth Uprising in 1976? (Passage A) Why did US President Kennedy avoid the term "blockade" when announcing the naval action around Cuba? (Passage B)
Multiple choice	What led to violent anti-apartheid protests? (Passage A) 1) Police forcefully segregating people. 2) Police arresting Nelson Mandela. 3) Police killing Black civilians. 4) Police implementing strict curfews. How did the US government discover the presence of Soviet missiles in Cuba? (Passage B) 1) A Cuban informant told them about the missiles. 2) The Cuban government made threats to employ the missiles. 3) The US Navy intercepted a Soviet ship carrying the missiles. 4) A US plane captured photos of the missiles.
Comprehension	
Short response	Explain the role that Nelson Mandela played during apartheid and its eventual end. You only need to write a short paragraph. (Passage A) Explain the role of the Soviet Union in the Cuban Missile Crisis. You only need to write a short paragraph. (Passage B)
Free recall	
Open response	Write down everything you remember from the text "[title]". Try to include as many details as possible. For example, think about what happened, why and how, when, where, and who was involved. You can write in full sentences or bullet points.

2.8 Inter-rater Reliability Results

Table 12: Inter-coder reliability

Item	**ICC (A,1)**	**p-value**	**95% CI**	**Item**	**ICC (A,1)**	**p-value**	**95% CI**
1	0.867	3.08×10^{-24}	[0.781, 0.925]	15	0.923	2.17×10^{-32}	[0.871, 0.958]
2	0.918	5.77×10^{-32}	[0.863, 0.955]	16	0.989	1.29×10^{-61}	[0.980, 0.994]
3	0.967	1.30×10^{-45}	[0.943, 0.982]	17	0.962	8.52×10^{-43}	[0.935, 0.979]
4	0.911	1.38×10^{-30}	[0.851, 0.951]	18	0.961	4.95×10^{-42}	[0.933, 0.979]
5	0.891	1.92×10^{-27}	[0.819, 0.939]	19	0.938	7.34×10^{-36}	[0.895, 0.966]
6	1.000	NaN	[NaN, NaN]	20	0.963	8.25×10^{-44}	[0.936, 0.980]
7	0.951	2.65×10^{-39}	[0.916, 0.973]	21	0.859	3.92×10^{-24}	[0.770, 0.921]
8	0.936	2.38×10^{-33}	[0.891, 0.965]	22	0.893	3.34×10^{-27}	[0.822, 0.940]
9	0.930	9.00×10^{-31}	[0.880, 0.962]	23	0.953	2.93×10^{-25}	[0.912, 0.976]
10	0.954	1.88×10^{-39}	[0.921, 0.975]	24	0.971	9.27×10^{-33}	[0.947, 0.985]
11	0.920	1.89×10^{-30}	[0.864, 0.956]	25	0.959	3.71×10^{-39}	[0.928, 0.978]
12	0.969	5.35×10^{-40}	[0.946, 0.984]	26	0.988	1.02×10^{-60}	[0.980, 0.994]
13	0.959	6.30×10^{-42}	[0.930, 0.978]	27	0.968	4.23×10^{-38}	[0.943, 0.983]
14	0.927	2.80×10^{-33}	[0.877, 0.960]	28	0.983	7.93×10^{-56}	[0.971, 0.991]

2.9 Survey Questions and Response Scales

Table 13: Survey questions and response scales - Session 1

Variable	**Question and response scale**
Text difficulty	How difficult to understand did you find the text on [Passage title]? (Not at all difficult, Not very difficult, Somewhat difficult, Quite difficult, Very difficult)
Topic familiarity	How much did you already know about [Passage title] before starting the task? (Nothing at all, Not very much, A moderate amount, Quite a bit, Very much)
Topic interest	How interesting was the text on [Passage title]? (Not at all interesting, Not very interesting, Somewhat interesting, Quite interesting, Very interesting)
Activity enjoyment	How enjoyable was learning the text with the help of [activity]? (Not at all enjoyable, Not very enjoyable, Somewhat enjoyable, Quite enjoyable, Very enjoyable)
Activity difficulty	Overall, how difficult did you find the [activity]? (Not at all difficult, Not very difficult, Somewhat difficult, Quite difficult, Very difficult)
Activity helpfulness	How helpful was [activity] for understanding and learning the text? (Not at all helpful, Not very helpful, Somewhat helpful, Quite helpful, Very helpful)
Activity future use	Would you use a similar approach ([activity]) to understand and learn a text in the future? (Yes, No, I am not sure)
Task interest	How interesting was this task overall? (Not at all interesting, Not very interesting, Somewhat interesting, Quite interesting, Very interesting)
Task effort	How much effort did you put into understanding and learning the text on [Passage title]? (No effort at all, Only a little bit of effort, Some effort, Quite a bit of effort, A lot of effort)
Perceived task performance	How well do you think you did on the task? (Not at all well, Not very well, Somewhat well, Quite well, Very well)
Activity preference	**Group 1:** Which of the two learning approaches of this study did you prefer (note-taking or AI chatbot)? (I preferred learning by note-taking, I preferred learning with the help of the AI chatbot, I had no preference, I am not sure) **Group 2:** Which of the two learning approaches of this study did you prefer (AI chatbot only or AI chatbot with note-taking)? (I preferred learning only with the help of the AI chatbot, I preferred learning with the help of the AI chatbot and by taking notes simultaneously, I had no preference, I am not sure)
Reason for preference	Can you tell us why you preferred this approach? [Open response]
Prior LLM use	Have you ever used an AI chatbot (such as ChatGPT, Microsoft Bing, and Google Bard AI) before this study? (Yes, No)
LLM use frequency	How often do you use an AI chatbot (approximately)? (Less than once a week, One or two days a week, Three to five days a week, Most days of the week)
Notes for learning frequency	How often do you take notes when reading a text for schoolwork, such as to prepare for a lesson or a test? (Never, Rarely, Sometimes, Often, Always)
LLM for learning frequency	How often do you use an AI chatbot when reading a text for schoolwork, such as to prepare for a lesson or a test? (Never, Rarely, Sometimes, Often, Always)
LLM+Notes for learning frequency	Group 2 only: How often do you use the two approaches (using an AI chatbot and taking notes) at the same time when reading a text for schoolwork? (Never, Rarely, Sometimes, Often, Always)

Table 14: Survey questions and response scales - Session 2

Variable	Item and response categories
Perceived test performance	If all the questions on [Passage title] combined were worth a maximum of 100 points, how many points do you think you would have (approximately) scored? [Open response]
Learning in between sessions	Have you done anything between the first session and today's session to further explore or understand the topics of the two texts? That could include looking up information online, taking notes after the session or discussing the topic with others. If so, please provide as much detail as you can about what you have done. [Open response]
Gender	What is your gender? [Open response]
EAL	Which language do you feel most comfortable speaking and communicating in? (English, A language other than English, Equally English and another language)
History	Are you taking GCSE History? (Yes, No)

2.10 Learning Experiences and Perceptions

Table 15: Differences in learning experiences and perceptions between conditions (for Group 1 and Group 2)

	Group 1: LLM vs Notes					**Group 2: LLM vs LLM+Notes**				
Variable	**Diff.**	**t(df)**	**p**	**95% CI**	**d**	**Diff.**	**t(df)**	**p**	**95% CI**	**d**
Activity helpfulness	**0.41**	**4.38(181)**	<0.001	**[0.22, 0.59]**	**0.33**	-0.03	-0.35(157)	0.724	[-0.21, 0.15]	-0.03
Activity difficulty	**-0.51**	**-7.00(181)**	<0.001	**[-0.66, -0.37]**	**-0.52**	**-0.41**	**-4.99(159)**	<0.001	**[-0.57, -0.25]**	**-0.40**
Task effort	**-0.25**	**-3.53(182)**	0.001	**[-0.38, -0.11]**	**-0.26**	-0.08	-1.03(159)	0.305	[-0.22, 0.07]	-0.08
Activity enjoyment	**0.68**	**6.50(181)**	<0.001	**[0.47, 0.89]**	**0.48**	0.00	0.00(158)	1.000	[-0.16, 0.16]	0.00
Text interest	-0.11	-1.38(183)	0.170	[-0.26, 0.05]	-0.10	0.06	0.79(159)	0.428	[-0.09, 0.22]	0.06
Text difficulty	0.03	0.50(183)	0.621	[-0.10, 0.16]	0.04	0.03	0.41(159)	0.684	[-0.10, 0.15]	0.03
Task interest	0.09	1.01(183)	0.315	[-0.09, 0.27]	0.07	-0.06	-0.79(159)	0.430	[-0.20, 0.08]	-0.06
Perceived task performance	0.00	0.00(182)	1.000	[-0.14, 0.14]	0.00	-0.11	-1.45(158)	0.150	[-0.25, 0.04]	-0.12
Perceived test performance	**-9.66**	**-5.53(177)**	<0.001	**[-13.11, -6.22]**	**-0.42**	**-6.80**	**-3.55(143)**	0.001	**[-10.59, -3.02]**	**-0.30**

2.11 Coding Scheme Activity Preferences

Table 16: Coding scheme: LLM over LLM+Notes preferences

Code	**Description**	**Examples**
LLM alone is quicker	Using the LLM alone is quicker than also taking notes, which takes time.	"It took less time to use the LLM", "Notes take too much time."
Both together not necessary	Notes are not necessary when the LLM already explains the text.	"The note-taking seemed unnecessary as the bot already helped explain", "Using one sort of meant I didn't need the other."
LLM does the work for you	If you use the LLM alone, you don't have to do the work yourself. The task becomes easier if you don't have to take notes.	"Didn't have to do any work", "Clarify any information I didn't know immediately without having to scour the text", "It was difficult to take notes at the same time as using the chatbot."
Note-taking reduces question time	Note-taking takes away time from asking the LLM questions or understanding the text.	"I didn't have enough time to ask as many questions when taking notes", "I had more time to understand the text."
LLM does not support note-taking	LLM does not make note-taking easier.	"Not as useful for making note-taking easier."

Table 17: Coding scheme: LLM over Notes preferences

Code	Description	Examples
LLM is quick	LLM is quick and saves time.	"Less time-consuming", "Much quicker."
LLM is easy	LLM is easy and requires little effort compared to note-taking, which takes more effort and is more difficult.	"More simple", "It was easier."
LLM is (inter)active	LLM is an interactive or active learning activity.	"Actively engaging with the bot", "Felt more interactive."
LLM is emotionally engaging	LLM is more fun, enjoyable, and interesting.	"Enjoyed reading its responses", "More fun to use."
LLM helps you focus	LLM helps you focus on the text.	"Allowed me to focus more on the text."
LLM helps you understand	LLM helps understanding and helps you check your understanding.	"It gives you a better understanding", "I could confirm anything I was unsure of to ensure I understood it."
LLM helps you learn	LLM supports learning.	"The AI helped me to learn more efficiently", "I was able to understand and learn the text a lot easier and quicker at a higher level."
LLM answers questions	LLM is helpful for understanding because it can answer questions and explain what you don't understand.	"Ask any relevant questions", "If I had a question, it could answer it."
LLM can provide background and additional information	LLM is helpful for understanding because it provides background information and can elaborate on what happens.	"I was given more background", "It gives me the full context."
LLM can summarise and simplify information	LLM is helpful for understanding because it can simplify and rephrase information as well as summarise.	"It puts it in a simpler way and form", "I can ask the AI chatbot to rephrase key points", "It can summarise key points."
LLM helps you remember	LLM helps you to remember the information in the text.	"It has stuck in my head more", "Giving me prompt questions, mnemonics, etc., which helped me remember", "Took less time to memorise than note-taking."

Table 18: Coding scheme: Notes over LLM preferences

Code	Description	Examples
Notes help you remember better	Note-taking helps you to remember information because you are physically writing it down. LLM does not help you remember as well.	"I can remember things better when I write them down", "More helpful for developing recall", "I learned more with note-taking", "Just gave more background, rather than consolidating the knowledge."
Notes help you understand	Note-taking helps you to understand better and check your understanding.	"It was easier for me to understand what I was reading", "I was understanding it more", "Test what you have learned by paraphrasing."
Note-taking is active	Note-taking is more active.	"Better active reading", "Allows me to actively engage."
Notes are your own work	Note-taking means that you do the work yourself. You do the thinking and can use your own words and capture your own views.	"You have to personally analyse it", "I could condense the information into my own words", "Made me think for myself", "It is your view on the matter you are looking at", "Allows me to feel proud of my work in the future."
Notes help you process information	Note-taking helps you process the information.	"I was able to break down and process the text", "Summarising the second text myself helped me to process the information."
Notes help you learn	Notes help you to learn, capture what you have learned, or test what you have learned.	"I am able to write down my own knowledge of what I had learned", "I could actually learn the information rather than being told it."
Notes can be revisited	Notes can be more easily revisited than the LLM output. You can easily access what you have learned or thought so far.	"I can come back to these notes at a later date if I am doing revision", "Note-taking gives you something better to look back on in future."
Notes are easier	Note-taking is easier than using the LLM.	"Easier to summarise", "IDK, easier."
Notes help with organisation	Notes help you to organise the information and thoughts and break it down into smaller parts to aid clarity.	"It is easy to organise my notes", "It is easier to keep track of your train of thoughts", "Helped me to break down the text into smaller chunks."
LLM is distracting and provides too much information	LLM is distracting as you may ask questions that are not relevant or focus on things that are not important. LLM provides too much information, which can be overwhelming or confusing.	"I found myself easily distracted by the AI and was more tempted to ask random questions", "It's not clear as it gives too much information."
LLM is repetitive and boring	LLM is boring and repetitive as it restates the information many times.	"It felt that it was just repeating itself."
Not sure what to ask the bot	The LLM is not needed because everything is understood, or one does not know how to use it and what kind of questions to ask.	"I struggled to think of questions to ask the AI", "The text was very easy therefore didn't feel the need to ask many questions."

Table 19: Coding scheme: LLM+Notes over LLM preferences

Code	Description	Examples
Both together are more enjoyable	Using LLM and notes together is more fun and enjoyable, whereas LLM alone can be boring.	"I enjoy using both at the same time", "If I had to use the chatbot and ask it 20 questions, I would be very bored."
Both together combine the best of both worlds	LLM and notes can be used in complementary ways to get the best of both, such as doing the work yourself and then using the LLM when you are unsure or stuck.	"It was easier to have my key notes summarised as well as text with more detail", "It allowed me to note down the crucial parts of the event in a way that I can understand it and also get help from the AI chatbot on anything that isn't clear."
Both together are more helpful and easier	General statements about the strategy being more helpful, better, or easier for understanding and learning.	"Most helpful and easy to learn", "Because I find it easier to remember and learn this way."
Notes help you process and understand the information from the LLM	Notes help you process and understand the information given by the LLM.	"In order for me to process this, I find note-taking at the same time very helpful."
Notes help with organisation	LLM provides information, but notes are needed to organise and structure ideas. The notes are also more focused and accessible.	"If I am only using the chatbot, then I have to scroll up to find what I am looking for", "It was easier to keep track of things and go back over them."
Notes are your own work	Taking notes means you do actual work and can capture your own thoughts rather than just reading output.	"It meant I was doing actual work."
Notes help you remember	Notes help to remember the information.	"I like to write out information as I think it helps me remember it better."
Notes help you understand	Note-taking helps you to understand better and to check your understanding.	"Simplifying it on paper made it easier to understand and remember."
Notes help you learn	Notes help you to learn, capture what you have learned, or test what you have learned.	"You learn more", "You can simplify what you have learnt in the notes."
LLM can provide bad answers	LLM does not always answer questions well and sometimes not at all. LLM can be harmful.	"Some of the questions I had for the bot were not answered explicitly."
LLM not always available	One needs to know how to take notes as LLMs might not always be available.	"You will not get an AI chatbot at all times."
Not sure what to ask the bot	The LLM is not needed because everything is understood, or one does not know how to use it or what kind of questions to ask.	"I wasn't sure what I was supposed to say to the bot. It was just kinda irritating."

2.12 Coding Scheme Prompt Interactions

For the full prompt coding scheme, please refer to tabular file 'PromptCoding.xslx'

Table 20: Prompt Coding Scheme

Overarching Code	Sub-code	Description and Examples
Information condensation	Summarise	The student asks the bot to summarise the entire text or a specific text selection. Examples: "Help me to summarise this paragraph", "Summarise the text", "Give me a summary of the first paragraph", "Tell me what this text is about."
	Take notes	The student asks the bot to take notes about the text as a whole or a specific paragraph. Examples: "Make notes for the first paragraph."
	Identify key ideas	The student asks the bot to identify the key ideas or takeaway messages from the text, including key dates, places, people, and events. Examples: "What are the main points?", "Give me all the important dates", "What's the takeaway message?"
	Create timeline	The student asks the bot to create a timeline of events described in the text. Examples: "Put the important dates into chronological order", "Give me a timeline of the events."
Understanding the text	Define a word or concept	The student asks the bot to define or explain a specific word or concept from the text. They request help to understand terminology but do not ask for factual information beyond that. Examples: "What does apartheid mean?", "What is a colony?", "What is a missile?", "I don't know what a blockade is."
	Simplify or explain difficult sentences	The student asks the bot to simplify or explain the provided passage or a specific selection of the passage. Examples: "Explain this in simple words", "Make the text simpler", "What does this sentence mean?", "Simplify this text."
	Checking understanding	The student explains their understanding and seeks confirmation from the bot. Examples: "The US did not like Cuba because they thought that Castro was a communist, right?", "So it was one officer that prevented the whole war?"
Seeking additional information and deeper understanding	General background	The student asks for background information on a place, time, or person mentioned in the text to provide context—information that is not too central for understanding the text but could be relevant. Examples: "Who was Kennedy?", "What was Mandela famous for?", "Tell me more about Cuba", "How many British colonies were there in Africa?", "Where were the Turkish missiles located?"

Continued on next page

Overarching Code	Sub-code	Description and Examples
	Elaboration and deeper understanding	The student asks for more details about an event, such as why it happened, who was involved, and the outcome. Examples: "Why did the US not like Castro?", "Why did the exiles invade Cuba?", "How did black people feel during apartheid?"
	Ask for examples or analogies	The student requests examples or analogies to better understand a concept or event. Examples: "What are examples of how apartheid affected daily life?", "Is there an analogy that explains the Cold War tensions?", "What unfair laws were passed?", "What were some of the boycotts?"
Seeking additional information and deeper understanding	Ask for contrasts or comparisons	The student asks the bot to compare or contrast concepts, events, or figures. Examples: "How is apartheid different from segregation in the US?", "Compare Kennedy and Khrushchev's leadership styles."
	Critical analysis or evaluation	The student requests the bot to critically analyze or evaluate an action, situation, decision, or statement. Examples: "What are the strengths and weaknesses of Kennedy's decision?", "Evaluate the effectiveness of the blockade."
	Implications and significance	The student inquires about the broader implications, relevance, or consequences of information in the text. Examples: "What were the long-term effects of the crisis?", "What is the situation like now?", "Why should I care or learn about this?"
Study and memory help	Study and memory help	The student asks for assistance to learn and remember the text, including requests to be quizzed on the content. Examples: "Make a mnemonic", "Write four questions about the text", "How can I remember this better?"
Interacting with the Bot	Request specific format or length	The student requests that the bot provides its response in a specific format or length. Examples: "Summarize the main points in bullet points.", "Can you create a chart of the different policies?", "Use only a few words", "Make it short."
	Request improvement	The student asks the bot to improve its response or restate it in a simpler or shorter way rather than asking for simplifications of the provided passage. Examples: "I don't understand what you said", "Explain that again but shorter", "What do you mean?", "Simpler please.", "Can you write that in simpler terms?", "Make the summary shorter."
	Relational language	The student engages in casual, polite conversation that is unrelated to the text. Examples: "How are you?", "Thank you", "Hello."

Continued on next page

Overarching Code	Sub-code	Description and Examples
	Checking source and trustworthiness	The student inquires about the sources or questions the accuracy of information. Examples: "What are your sources?", "Why should I believe you?", "I think your answer is wrong."
	Pasting text without specific request	The student pastes text directly from the provided passages without framing it as a specific question or request. Examples: "Nelson Mandela", "In 1910, four British colonies joined to create the Union of South Africa.", "Missile."
Irrelevant, off-topic, miscellaneous	Irrelevant to text	The student asks a question unrelated to the text or its background. Examples: "Who is Che Guevara?", "What is the song Abraxas?"
	Miscellaneous	Use this code for segments that don't fit any other codes. Use this as a last resort.
Irrelevant, Off-topic, Miscellaneous	Nonsensical input	The student types nonsensical characters, symbols, or text that does not form coherent words or sentences. Examples: "asdfgh", ".", "123", "???"

2.13 Frequency of Prompt Types

Table 21: Frequencies of overarching prompt types

Overarching prompt type	Frequency
Archetype	
Seeking additional information and deeper understanding	2265
Information condensation	749
Understanding the text	615
Study and memory help	39
Other	
Interacting with the bot	760
Irrelevant, off-topic, miscellaneous	501

Table 22: Frequencies of specific prompt types

Overarching prompt type	**Specific prompt type**	**Frequency**
Seeking additional information and deeper understanding	Elaboration and deeper understanding	1479
Information condensation	Summarise	588
Seeking additional information and deeper understanding	General background	514
Understanding the text	Define a word or concept	463
Interacting with the Bot	Request specific format or length	430
Irrelevant, Off-topic, Miscellaneous	Irrelevant to text	296
Understanding the text	Simplify or explain difficult sentences	126
Seeking additional information and deeper understanding	Implications and significance	119
Information condensation	Identify key ideas	114
Interacting with the bot	Request improvement	113
Interacting with the bot	Pasting text without specific request	106
Interacting with the bot	Relational language	105
Irrelevant, off-topic, miscellaneous	Nonsensical input	109
Irrelevant, off-topic, miscellaneous	Miscellaneous	96
Seeking additional information and deeper understanding	Ask for examples or analogies	66
Seeking additional information and deeper understanding	Critical analysis or evaluation	54
Study and memory help	Study and memory help	39
Seeking additional information and deeper understanding	Ask for contrasts or comparisons	31
Understanding the text	Checking understanding	26
Information condensation	Take notes	26
Information condensation	Create timeline	21
Interacting with the bot	Checking source and trustworthiness	6

Note: This table only includes prompt types that have been used at least three times by students.