Computers & Education 243 (2026) 105514

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

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

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

[a] *Cambridge University Press & Assessment, Shaftesbury Rd, Cambridge, CB2 8EA, UK*
[b] *Microsoft Research, 21 Station Rd, Cambridge, CB1 2FB, UK*
[c] *Microsoft Research, 300 Lafayette St, New York, NY, 10012, USA*

ARTICLE INFO

Keywords:
Generative artificial intelligence
Note-taking
Comprehension
Retention
Secondary education

ABSTRACT

Students' rapid uptake of Generative Artificial Intelligence tools, particularly large language models (LLMs), 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 in England. 405 students, aged 14–15 years, studied two text passages and completed comprehension and retention tests three days later. Quantitative results demonstrated that both note-taking alone and combined with LLM use had significant positive effects on retention and comprehension compared to using 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 the LLM 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 activities, the benefits of combining LLM use with traditional learning over using LLMs alone, and the AI skills that students need to maximise those benefits.

1. Introduction

Students' 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 (Aleksić-Maslać et al., 2024; Chan, 2023; Johnston et al., 2024; Lan & Tung, 2023; Shoufan, 2023; Singh et al., 2022) while also creating new challenges (Barrett & Pack, 2023; Dwivedi et al., 2023; Huber et al., 2024; Kasneci et al., 2023; Steponenaite & Barakat,

[*] Corresponding author.
E-mail address: martina.kuvalja@cambridge.com (M. Kuvalja).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compedu.2025.105514
Received 3 February 2025; Received in revised form 20 November 2025; Accepted 21 November 2025
Available online 24 November 2025

2023; Zhu et al., 2023). Accordingly, research has begun to explore the effects of LLM use in education, and a recent review of experimental studies suggests that using LLMs can enhance students' academic performance (Deng et al., 2025). While this seems encouraging, the review cautions that these improvements may not reflect actual gains in students' learning, but rather the quality of work generated by LLMs. Existing studies have tended to focus on performance outcomes such as the quality of writing (Meyer et al., 2024) or task resolution (Urban, 2024), and several studies have allowed participants to use LLMs during the assessments of academic performance (see Deng et al., 2025). Consequently, there are substantial gaps in our understanding of how the use of LLMs affects genuine learning.

The effect of LLM use on two foundational aspects of learning – understanding and retaining information – is underexplored. Knowledge stored in long-term memory is a fundamental element of cognition, forming the basis of nearly all human activity (Binder & Desai, 2011). Understanding the effects of LLM use on these foundations is thus urgently required to guide how such tools are integrated into schools, as policymakers and educators are grappling with many unknowns. Accordingly, we conducted one of the first large-scale empirical studies examining how LLM use affects students' comprehension and retention. We specifically focused on reading comprehension and retention given the central role of text-based learning (e.g., textbooks, worksheets) in education. To evaluate the effectiveness of the activity, we compared LLM use with the common, evidence-based learning activity of note-taking, as well as a combination of both. At the same time, our goal in this study was not to prescribe optimal usage but to compare LLM-supported learning and note-taking as they naturally unfolded (i.e., without use being prescribed by the researcher) under the same instructions. Just as note-taking varies in quality and depth, so too does LLM use. We therefore see this study as a necessary first step in documenting naturalistic use, which can inform more controlled, usage-specific manipulations in future work. To shed light on the learning outcomes, we also explored students' task engagement when using the different learning activities as well as what types of LLM prompts they used. Our study focused on secondary school students because it is an important phase of education where learners are developing their independence, shaping the way in which they will learn in the future. A sizeable proportion of secondary school students already use GenAI tools such as OpenAI's ChatGPT (Ofcom, 2024; Walton Family Foundation, 2023). Understanding the effects of LLM use during this formative time seems particularly urgent. Finally, this research will also add a necessary perspective to the literature on LLM use in education, which has mostly been concerned with higher education (see Deng et al., 2025).

2. Literature review

2.1. Reading comprehension and retention

Reading comprehension is the process of making sense of written materials resulting in a mental representation of the material (McNamara & Magliano, 2009). Models of reading comprehension, such as the Construction-Integration (CI) model (Kintsch, 1988), 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) (McNamara & Magliano, 2009). This multi-level structure is supported by neuroimaging studies (Ding et al., 2020; Fedorenko et al., 2024; Hickok & Poeppel, 2007; Zwaan & Radvansky, 1998). 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 (McNamara & Magliano, 2009). 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) (Cain & Oakhill, 2007; Daneman & Carpenter, 1980; Oakhill et al., 2015; Perfetti et al., 2005; Stanovich, 1986), the text itself (e.g., genre, length, word and sentence complexity, cohesion) (Graesser et al., 1994; McNamara et al., 2004), and the reading context (e.g., reading for leisure or academic purposes) (Guthrie & Wigfield, 2000; Linderholm et al., 2018, pp. 165–186).

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 to be 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 (Craik, 2002; Craik & Tulving, 1975), 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 (Anderson, 1983; Craik, 2002).

2.2. Learning approaches supporting comprehension and retention

There are several learning approaches, and more specifically reading strategies, that can enhance comprehension and retention as outlined by McNamara (2007) and Chi (2009). Throughout the reading process, monitoring comprehension is particularly crucial, and includes strategies such as generating questions to gauge one's understanding (McNamara, 2007). Text-focused strategies involve interpreting the meaning of words, sentences, and ideas (e.g., paraphrasing, breaking up long and complex sentences into manageable chunks, making bridging inferences to link different concepts) (McNamara, 2007). Learning activities in which the learners receive information without doing anything with it (e.g., just listening or reading without engagement) are considered *passive*. On the other hand, strategies such as paraphrasing, selecting, and repeating are considered *active* learning strategies, and these can activate prior knowledge and support the encoding, storing and assimilation of new knowledge (Chi, 2009). There are also several effective reading strategies that go beyond the text (e.g., generating questions, using self-explanations, and using external information sources) (McNamara, 2007). 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 (Chi, 2009). Lastly, *interactive* learning involves meaningful dialogue with a partner, such as peers, teachers, or intelligent tutoring systems (Chi, 2009; Graesser et al., 1994). Its characteristics include dialoguing with a partner who is viewed as 'a more knowledgeable other' (Daniels, 2001; Vygotsky, 1978) and may explain, provide corrective feedback and scaffold, or where both partners make substantive contributions. Such interactions may enhance learning by giving learners additional information and access to different perspectives. Exchange in which only one partner makes all substantive contributions while the other mostly listens is not considered interactive learning (Chi, 2009). Please note that these categories are not strictly mutually exclusive. For example, a truly interactive activity will typically also be constructive and active, but with the added layer of co-construction between partners. Finally, these categories are viewed as hierarchically organised such that interactive subsumes constructive, and constructive subsumes active processes (Chi, 2009).

2.2.1. LLM use for comprehension-focused reading

The integration of LLM tools into education raises the crucial question of whether their use could facilitate or undermine such active, constructive, and interactive strategies during reading (henceforth collectively referred to as *comprehension-focused reading*) and thereby enhance or diminish learning. Due to the capability of LLMs to transform texts, they may support active, text-focused strategies such as paraphrasing and providing definitions of words, and thereby help students to understand the language and meaning of the text itself (i.e., the surface structure and textbase). Furthermore, LLMs' ability to provide immediate clarifications and simplify complex concepts may help reduce cognitive load (Mayer, 2004; Sweller et al., 2019), which could allow students to spend more time and effort on constructive strategies. LLMs may also support constructive strategies more directly, such as by making connections between concepts and elaborating on information, as they offer unprecedented flexibility in generating explanations, providing diverse perspectives, responding to complex questions in real time, and adapting to individual learners' needs (Holmes et al., 2019; Luckin et al., 2016). 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 (Bernabei et al., 2023; Kumar et al., 2023; Sarsa et al., 2022; Zhu et al., 2023), including understanding that goes beyond the literal text (i.e., the situation model). Note that the generation of questions (in this case LLM prompts) itself can be considered as a constructive process if the questions encourage students to think beyond the immediate text (Chi, 2009).

LLMs also have the potential to support interactive learning, which, depends on the contributions that the dialoguing partners make. For example, an interactive learner may respond to the helpful input from the LLM in a substantive or meaningful way by asking a related question, sharing their own perspective, or bringing in a new idea to their interaction with the LLM. Overall, given these various ways in which LLMs may support the reading process, LLMs may be useful in helping learners build understanding at multiple levels: from surface-level text comprehension to deeper text-base representations of meanings and key ideas, and ultimately to a comprehensive mental representation at the situation-model level of comprehension.

And yet, using an LLM for comprehension-focused reading, even if it is with the intention to understand and learn a text, has the inherent risk of offloading the thinking process to the LLM. 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 (Chi, 2000; McNamara, 2007; Sweller et al., 2019). This would thus reflect an individual dialogue where only one partner (i.e., the LLM) makes all substantive contributions (Chi, 2009). Indeed, a recent review found that LLM use can lead to a reduction in mental effort (Deng et al., 2025), and over-use of LLMs could lead to shallow processing, where learners passively receive information without actively engaging in deep cognitive processing or critical thinking (Chi, 2009; Craik & Lockhart, 1972; Farhi et al., 2023; Lee et al., 2025; Zhai et al., 2024). This superficial engagement could hinder the development of comprehensive mental models, negatively affecting comprehension and long-term retention (Bjork & Bjork, 2011; Craik & Tulving, 1975). Hence, 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 (Dehaene & Naccache, 2001; Pascual-Leone et al., 2005).

2.2.2. Note-taking for comprehension-focused reading

In order to assess the effectiveness of using LLMs as a learning tool for reading comprehension and retention, it is useful to compare 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 (Kiewra, 1989; Kobayashi, 2005). 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 (Kiewra, 1985; Kobayashi, 2005). 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 (Bohay et al., 2011; Bui & Myerson, 2014; Rummer et al., 2017). The former is supported by a meta-analysis that 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) (Kobayashi, 2005).

2.3. Student task engagement

This study focuses on the effects of LLM use on the outcomes of learning, and central to learning is students' engagement in the learning activity. Student engagement has been described as "the holy grail of learning" because of its links to positive learning outcomes (Sinatra et al., 2015, p. 1). The way students engage with learning, including when using an LLM, can help us understand both the effectiveness of these tools and students' future willingness to use them. While engagement is defined and measured differently across studies, it is widely recognised as a multidimensional construct comprising emotional/affective, cognitive, and behavioural components (for reviews, see Salmela-Aro et al., 2021; Wong et al., 2024). It can be understood on multiple levels ranging from the macrolevel (e.g., engagement in a class, course, school, or community) to the microlevel (e.g., an individual's momentary engagement in a learning activity) (Sinatra et al., 2015), with the latter being of particular interest for this study.

Wong et al. (2024) systematically reviewed how the three engagement dimensions are conceptualised in the literature. They found that emotional engagement refers to students' affective involvement in learning, including positive and energised emotional responses toward or during a learning activity. Cognitive task engagement has been described in terms of students' motivational engagement, self-regulatory engagement as well as effortful engagement, and as such comprises a range of different facets. Behavioural engagement, on the other hand, refers to observable participatory and effortful engagement. Effortful engagement, therefore, can be understood as both a cognitive and behavioural construct.

Evidence is emerging that the use of LLMs may enhance emotional engagement in learning activities. The review by Deng et al. (2025) showed that the use of ChatGPT has significant positive effects on students' affective-motivational states. At the same time, they caution that this might be a temporary effect due to the novelty of the tool. The same review also found that the use of ChatGPT significantly reduced students' mental effort, indicating that cognitive engagement with the learning task might be negatively affected, with one study suggesting that this might come at the expense of deeper learning (Stadler et al., 2024). There is also some evidence that interacting with GenAI might affect students' behavioural engagement. For example, one study found that self-regulated learning training supported by GenAI led to greater behavioural engagement during reading - such as increased reading time - compared to training without GenAI support (Pan et al., 2025).

2.4. The present study

There is limited understanding of how LLM use impacts students' genuine learning, beyond the task performance achieved through LLM assistance (Deng et al., 2025). Integrating LLMs into education without understanding such effects raises serious concerns about students' knowledge development both in the short- and long-term. This study aimed to address this research gap by examining whether LLMs can be used as a tool to support comprehension-focused reading and thus the fundamental learning processes of reading comprehension and retention. We conducted a large-scale, pre-registered,[2] 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). These students are at a crucial point in the UK educational system as they begin preparing for their General Certificates of Secondary Education (GCSE) — exams that shape their future academic and career trajectories.

Our primary objective was to quantify the impact of LLM use aimed at comprehension-focused reading, involving active, constructive or interactive strategies, on students' reading comprehension and retention. More specifically, we assessed learning by testing students' literal retention (i.e., lower-level retention of information in the text, requiring no knowledge-based inferences, and no or only minimal bridging inferences), comprehension (i.e., higher-level retention requiring bridging inferences and knowledge-based inferences) as well as free recall (i.e., any literal retention and comprehension without cueing). We compared LLM use against the common, evidence-based learning activity of note-taking, as well as a combination of LLM use and note-taking. In practice, it might be useful to combine the activities of prompting 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. We did not include an additional "reading-only" control condition due to (i) concerns about insufficient statistical power given constraints on the number of participants we could recruit, (ii) the need to limit participant fatigue from responding to multiple conditions, and (iii) the assumption that any engagement with the text beyond passive reading is likely going to lead to improved learning outcomes (Chi, 2009; McNamara, 2007), thereby setting a comparatively low bar for evaluating the effectiveness of LLM use.

To shed light on the learning outcomes and gain insights into potential future use of LLMs for learning, we additionally explored students' emotional, cognitive and behavioural engagement with the different learning activities, including which one they preferred and why. Lastly, we explored different "archetypes" of students' prompting behaviour. We did not restrict students' use of the LLM to a specific learning strategy but instead provided them with suggestions that related to active and constructive strategies, and which could be used in interactive ways. Giving students the freedom over how to use the LLM reflects how most students currently interact with LLMs and, therefore, increases ecological validity. Analysing students' prompts also enabled us to explore the behavioural

[2] https://aspredicted.org/K4H_PHV: The preregistration specified the main research questions (research question 1), study design, dependent variables and their measurement, exclusion criteria, and the main analysis plan (two mixed-effects regression models, one for comprehension and one for retention; all independent variables were specified). The preregistration also included secondary analyses (e.g., regression for free recall), variables collected for exploratory purposes (e.g., engagement variables; relevant for research question 2), and qualitative analysis of students' LLM queries (relevant for research question 3). Any deviations from the preregistration are declared as such.

archetypes of how students engaged with the LLM in relation to the task, which may help us in interpreting the quantitative findings. Please note that analysing the interactive nature of students' LLM conversations (i.e., dialogue analysis) was outside the scope of this study.

The current study addressed the following research questions.

1) How does using an LLM for comprehension-focused reading affect students' text comprehension and retention compared to the more traditional learning activity of note-taking as well as compared to using an LLM alongside note-taking?
2) How does using an LLM for comprehension-focused reading affect students' emotional, cognitive, and behavioural engagement compared to the more traditional learning activity of note-taking as well as compared to using an LLM alongside note-taking?
3) What types of prompts do students use when interacting with the LLM?

The results are expected to provide valuable insights for education stakeholders and policymakers on thoughtfully integrating AI tools into schools to enhance fundamental learning processes.

3. 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 part of the main study, which took place between March and July 2024.

3.1. Participants

Participants were 405 Year 10 students,[3] aged 14–15 years, from seven secondary schools in England. Based on our exclusion criteria (see Appendix A), we retained 344 students for analysis. 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 (British Educational Research Association, 2018) ethical guidelines. Ethical approval was provided by the research ethics committees of the researchers' institutions.

3.2. Experimental design

The study was a randomised controlled experiment with within- and between-participant design elements, as illustrated in Fig. 1. The study was conducted over two sessions, a learning session and a test session, which were spaced three days apart. We chose this time gap in order to tap into long-term memory while avoiding floor effects. A gap of three days falls into the period of transitional long-term memory (lasting from 12 h to 7 days), which is characterised by relative stability of memory, and after which the rate of forgetting increases (Radvansky et al., 2022).

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 learning activity (condition). There were three conditions.

- **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 in which the LLM condition served as the reference point across both groups.

- **Group 1 (184 students, 53.5 %):** Participated in the *LLM* and *Notes* conditions.
- **Group 2 (160, 46.5 %):** Participated in the *LLM* and *LLM + Notes* conditions.

The order of conditions and passages in the learning session were randomised for each group to minimise order effects (Charness et al., 2012). Students completed survey questions about their task engagement immediately after each learning task to ensure the experience was still salient in their minds. To provide students with a short break between tasks, they played a 1-min game of Snake.

In the *test session*, students answered comprehension and retention questions about the two passages. The order of the passage was again randomised to minimise order effects. In between tests, students played a 1-min game of Snake. At the end of the test session, students completed survey questions regarding their general characteristics as described further below.

[3] 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.

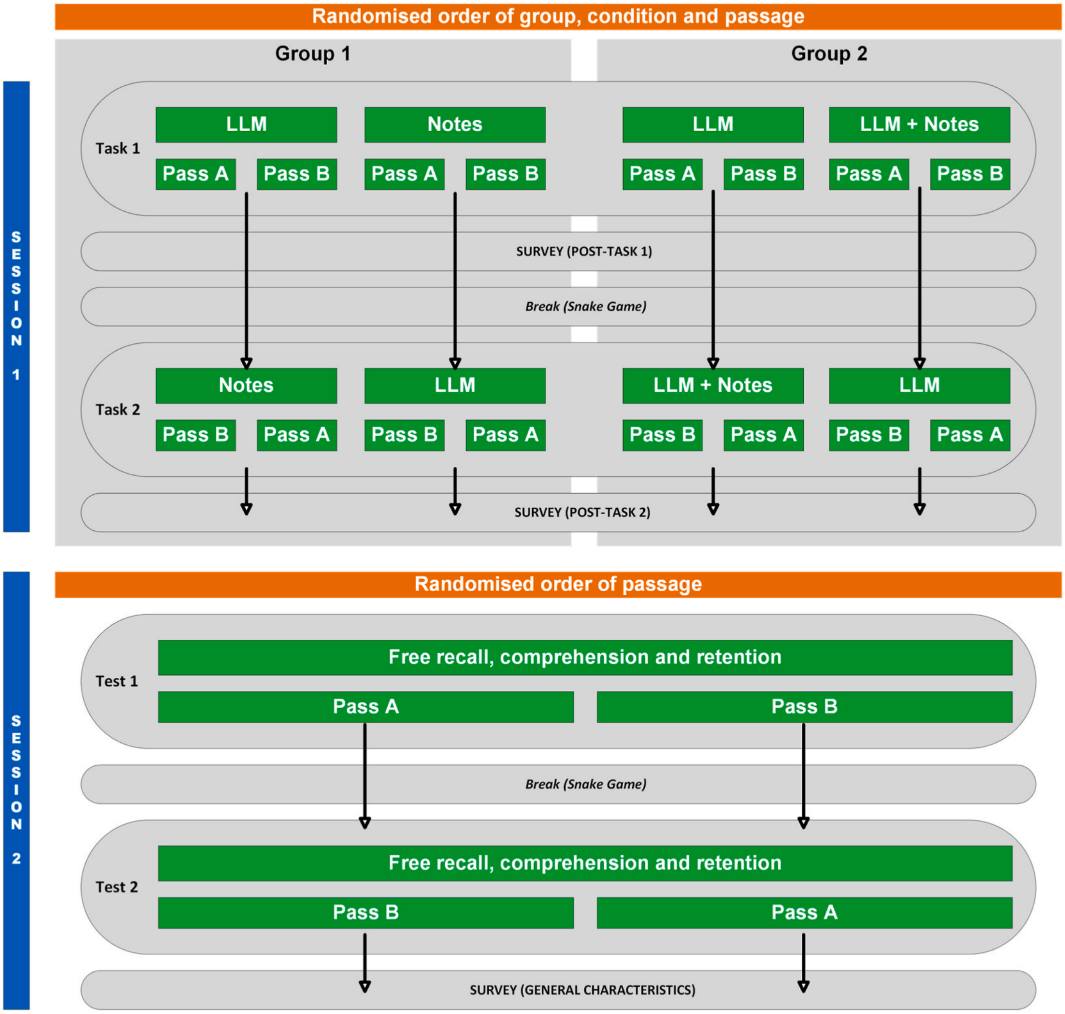


Fig. 1. Study design illustrating the activities and their order during Session 1 and 2.

Each student experienced two of the three conditions, with the LLM condition serving as a common reference point across both groups. This design enabled us to examine the effects of LLM compared to Notes and of LLM compared to LLM + Notes using a within-participant comparison. Within-participant comparisons provide a substantial increase in statistical power and their internal validity does not depend on the success of random assignment (Charness et al., 2012). This design also allowed us to make an indirect comparison of Notes and LLM + Notes in order to test whether either of these conditions differed in their relative advantage over the LLM-only condition, by comparing their coefficients. However, because the study was not designed for a direct contrast between Notes and LLM + Notes, these exploratory comparisons should be interpreted with caution due to possible carryover or interaction effects across the conditions within each group. We limit these analyses to objective learning outcomes and do not report such comparisons for subjective judgments, which are inherently relative because participants evaluate each condition in comparison to the other.

This design strategically balances the amount of information we could collect while maximally controlling for sampling variability and contextual confounds (associated with distinct participant groups) as well as working within practical constraints of conducting research with a limited pool of students (e.g., access to students, time commitments of students and schools, participant fatigue in responding to multiple conditions).

3.3. *Setup and system*

All 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 researcher or teacher read out a script with introductory instructions. They also monitored students during the entire session and answered their technical and procedural questions.

The experiment was carried out on a web app hosted on github.com that students accessed via a web browser. For the LLM functionality in Session 1 (learning), 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.[4] Fig. 2 illustrates the task screen for the LLM + Notes condition. For the Notes-only and the LLM-only conditions, only the notepad or chatbot was displayed, respectively.

3.4. Procedure

3.4.1. Learning session (session 1)

In the learning session, students read two passages on a history topic, each using 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 about the value of active reading and what it involves (see Supplementary Table 1) as well as about how the given reading activity might support active reading (see Supplementary Table 2). They then received more detailed task instructions describing specific comprehension-focused reading strategies (see Supplementary Table 3), which were followed by a video demonstration of the task and interface. The suggested strategies were based on the text comprehension literature (e.g., Chi, 2009; McNamara, 2007; McNamara et al., 2004; Pearson et al., 1990, p. 512), and included both active, text-based strategies (e.g., understanding the meaning of key words and difficult sentences, identifying key ideas) and constructive strategies (e.g., making connections, thinking about differences and similarities between concepts). The content and wording of the instructions for the three conditions were kept as similar as possible. Note that whether students use the LLM in an interactive way as defined by Chi (2009) depends on their own contributions to the conversation beyond asking questions. Once the task started, students needed to remain on the task page for 10 (minimum) to 15 (maximum) minutes. This was followed by a survey assessing task engagement and a short break (i.e., Snake game), before the procedure repeated for the second learning task.

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 syllabuses: 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 the text features were comparable and appropriate for Year 10 students. Passages A and B had four paragraphs each and were highly similar with regards to length (386 and 385 words), average word length (5.3 and 4.8 characters), word complexity[5] (1986 and 1927), number of sentences (both 26) and CEFR level (both C1 – upper intermediate). To aid comparability, we also adapted the passages so that they each contained 50 main idea units, or core propositions.[6] To check whether these comparability manipulations were successful, we asked students to rate their perceived text difficulty and topic interest after each learning task, using a single item with a 5-point Likert-Scale (for the items, see Supplementary Table 4). There were no significant differences in reported text difficulty (mean difference = −0.06, *t (df)* = −1.42 (343), *p* = .158) or topic interest (mean difference = −0.01, *t (df)* = −0.10 (343), *p* = .917) between the two texts.

3.4.2. Test session (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. The passage order was randomised, and for each passage, there were 22 test questions assessing literal retention, comprehension and free recall. Table 1 provides an overview of how the different constructs were assessed, and how responses were scored, which is further described in the Analysis section. The question order for both passages was free recall, comprehension, literal retention (cued recall) and, finally, literal retention (recognition). Students had to spend at least 3 min and a maximum of 5 min on the free-recall questions. Questions were carefully sequenced and on separate pages where needed to prevent previous questions providing cues for later questions. Example test questions can be found in Supplementary Table 5. At the end of the session, students completed a survey about their general characteristics.

On average, students spent approximately 35 min on the learning session and 30 min on the test session.

3.5. Measures

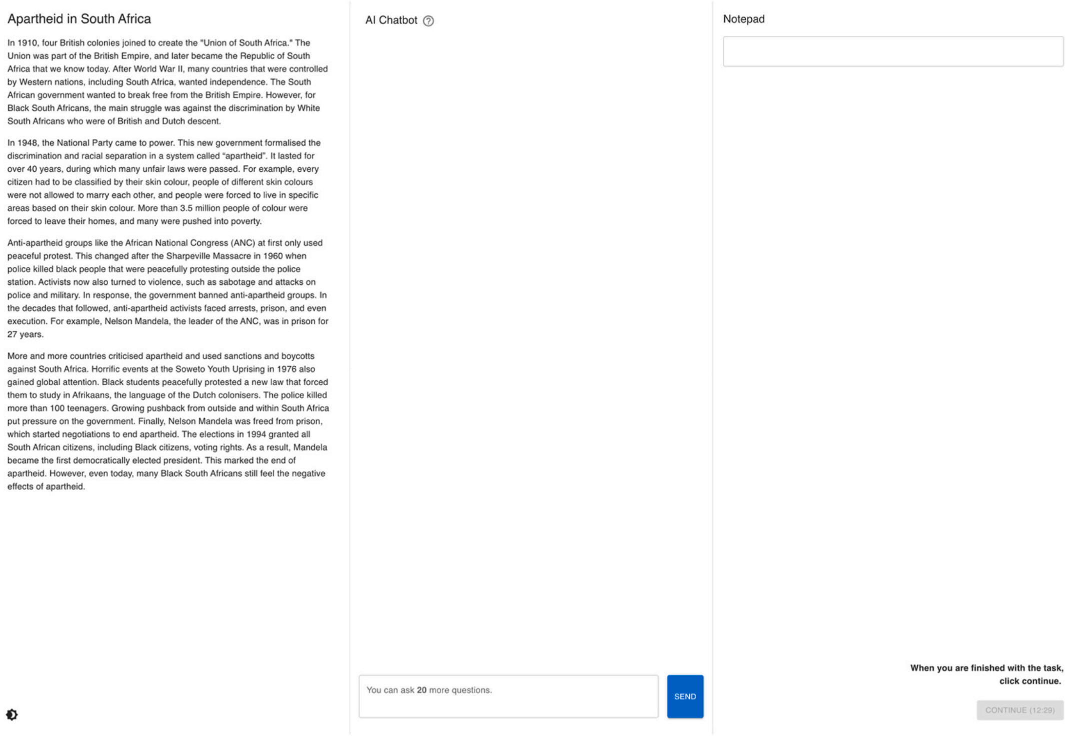
3.5.1. Comprehension and retention outcomes

Literal retention questions (short response and multiple choice) required literal recall or recognition of information from the passage to provide a correct response. Students did not need background knowledge beyond understanding the vocabulary used in the passage in order to succeed. 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.

[4] "You are an AI chat bot that helps students read and comprehend the following passage: Students can use this tool to define unfamiliar words, explain concepts, or summarise key points of the passage."

[5] I.e., the average position of the words in the 10,000 most frequent English words list.

[6] A proposition consists of a predicate and argument(s), and generally represents one complete idea. One sentence can have one or more core propositions, as well as subpropositions. We focused on the core propositions. For example, we counted two idea units for the sentence "The US was a capitalist country and they feared that Castro supported communism."



Apartheid in South Africa

In 1910, four British colonies joined to create the "Union of South Africa." The Union was part of the British Empire, and later became the Republic of South Africa that we know today. After World War II, many countries that were controlled by Western nations, including South Africa, wanted independence. The South African government wanted to break free from the British Empire. However, for Black South Africans, the main struggle was against the discrimination by White South Africans who were of British and Dutch descent.

In 1948, the National Party came to power. This new government formalised the discrimination and racial separation in a system called “apartheid”. It lasted for over 40 years, during which many unfair laws were passed. For example, every citizen had to be classified by their skin colour, people of different skin colours were not allowed to marry each other, and people were forced to live in specific areas based on their skin colour. More than 3.5 million people of colour were forced to leave their homes, and many were pushed into poverty.

Anti-apartheid groups like the African National Congress (ANC) at first only used peaceful protest. This changed after the Sharpeville Massacre in 1960 when police killed black people that were peacefully protesting outside the police station. Activists now also turned to violence, such as sabotage and attacks on police and military. In response, the government banned anti-apartheid groups. In the decades that followed, anti-apartheid activists faced arrests, prison, and even execution. For example, Nelson Mandela, the leader of the ANC, was in prison for 27 years.

More and more countries criticised apartheid and used sanctions and boycotts against South Africa. Horrific events at the Soweto Youth Uprising in 1976 also gained global attention. Black students peacefully protested a new law that forced them to study in Afrikaans, the language of the Dutch colonisers. The police killed more than 100 teenagers. Growing pushback from outside and within South Africa put pressure on the government. Finally, Nelson Mandela was freed from prison, which started negotiations to end apartheid. The elections in 1994 granted all South African citizens, including Black citizens, voting rights. As a result, Mandela became the first democratically elected president. This marked the end of apartheid. However, even today, many Black South Africans still feel the negative effects of apartheid.

You can ask **20** more questions.

When you are finished with the task, click continue.

Fig. 2. Example task screen for the LLM + Notes condition.

Table 1
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.

In contrast, comprehension questions (short response) 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, inferring information that was implied but not explicitly stated. Accordingly, comprehension questions targeted the situation-model level of representation.

The free recall question (open response) captured both literal retention and comprehension aspects.

We assessed the reliability of the multiple-choice questions using McDonald’s omega (total). Results indicate acceptable internal consistency for the Cuba questions ($\omega = 0.78$) but lower internal consistency for the South Africa (SA) questions ($\omega = 0.56$). Based on students’ mean performance across all groups and conditions, the difficulty level of literal retention questions was significantly higher for SA compared to Cuba, and that of comprehension questions was significantly lower for SA compared to Cuba. There was no significant difference for free recall. Please see Supplementary Table 6 for *t*-test results. Accordingly, we controlled for the text in the regression analyses.

3.5.2. Student task engagement

All survey items and response scales can be found in Supplementary Table 4. Unless otherwise stated, task engagement was assessed through self-report following each learning task in Session 1, whereby each specific variable was assessed using a single item with a 5-point Likert-Scale.

Emotional engagement. Emotional engagement was assessed through two variables, namely *activity enjoyment* and *task interest*.

Cognitive engagement. Cognitive engagement was assessed through five variables, namely *activity difficulty, task effort, activity helpfulness, perceived task performance*, and *activity preference*. Activity preference was assessed at the end of Session 1 using one item that asked students which of the two learning activities they preferred, if any. To gain a deeper insight into students' preferences, they were then asked to explain why they preferred the learning activity over another in an open response.

Behavioural engagement. Behavioural engagement was primarily assessed through process data that was automatically logged during the learning activities (see also Schiller et al., 2024). For conditions involving the LLM, this included the *number of LLM prompts*. For conditions involving note-taking, this included the *number of words* written in the notepad. For students in the LLM + Notes condition, we additionally analysed students' *copying behaviour*, that is how much text they copied from the LLM output into the notepad to gauge whether their notes reflected their own efforts or were primarily based on the LLM output. We used the 'textreuse' package in R (Mullen, 2020) to analyse the proportion of trigrams (i.e., overlapping three-word sequences) that appeared both in the LLM output and in students' written notes. For all conditions, *time on task* was also recorded. At the end of Session 1, *activity future use intentions* (Yes, No, or I am not sure) were assessed using a single item for each learning activity they were exposed to.

3.5.3. Student characteristics

All survey items and response scales can be found in Supplementary Table 4.

Control variables. At the end of Session 2, students were asked to report their *gender*, whether English was an additional language for them (*EAL*), 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 (Taylor, 2018). This is because eligibility for FSM is typically based on family income and other socioeconomic factors. These variables were controlled for in the main analyses.

Exclusion criteria. Students were asked to indicate in an open response whether they had engaged in any text-related *learning in between sessions*. Similarly, students reported their *topic familiarity* in Session 1. For more information, please see the exclusion criteria (Appendix A).

Familiarity with learning activities. In Session 1, to understand to what extent students have engaged in the different learning activities in the past, we asked them to report their *familiarity with the learning tasks*. Students indicated whether or not they ever used an LLM and if so, how frequently. In addition, students were asked to indicate how often they used the activities they were exposed to when reading a text for schoolwork, using single items with a 5-point Likert-Scale (Never to Always).

3.6. Analysis

We did not deviate from our pre-registered analyses other than described here. First, we extended the quantitative analyses to conduct simultaneous tests for general linear hypotheses comparing the regression coefficients for Notes and LLM + Notes. Second, we conducted additional qualitative analyses exploring why students preferred one learning activity over another. Third, we did not examine initially planned interaction effects between learning conditions and student characteristics given our smaller than planned sample size. Quantitative analyses were run with Python 3.11 (Python Software Foundation, 2022) and R 4.4.2 (R Core Team, 2024). 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.

3.6.1. Scoring of learning outcomes

The short 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 (Koo & Li, 2016). 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 0.97 and 0.99, indicating excellent reliability (Koo & Li, 2016). The lower bounds of the 95 % confidence intervals were all above the 0.90 threshold for excellent reliability (see Supplementary Table 7).

3.6.2. Estimation of condition effects on text comprehension and retention

Missing data handling. There were no missing data in the dependent variables dataset that was analysed 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 incorrect (given 0 points). Missingness in covariates was minimal and only occurred for the variables Gender, EAL and History (5.23 %, 1.16 % and 1.16 %, respectively). Missing data were handled using multiple imputation by chained equations (MICE) using the ‘mice’ package in R (van Buuren & Groothuis-Oudshoorn, 2011). 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 in R (Bates et al., 2015), 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. As pre-registered, we used a single literal retention score, which was the sum of the short response and multiple-choice scores. The regression specification was as follows:

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

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.

The LLM condition was used as the reference group for the regression, so that the coefficients of Notes and LLM + Notes provide an estimate of their effectiveness compared to LLM. We also statistically compared the size of the coefficients of Notes and LLM + Notes — a between-group comparison — to gauge whether one of these learning conditions had a relatively larger benefit on learning outcomes compared to LLM use. To do so, we used the simultaneous tests for general linear hypotheses from the ‘multcomp’ package in R (Hothorn et al., 2008).

As depicted in Fig. 3 in the Results, free recall scores were non-normally distributed, so we ran additional non-parametric permutation tests. Specifically, we used the ‘infer’ package in R (Couch et al., 2021) 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

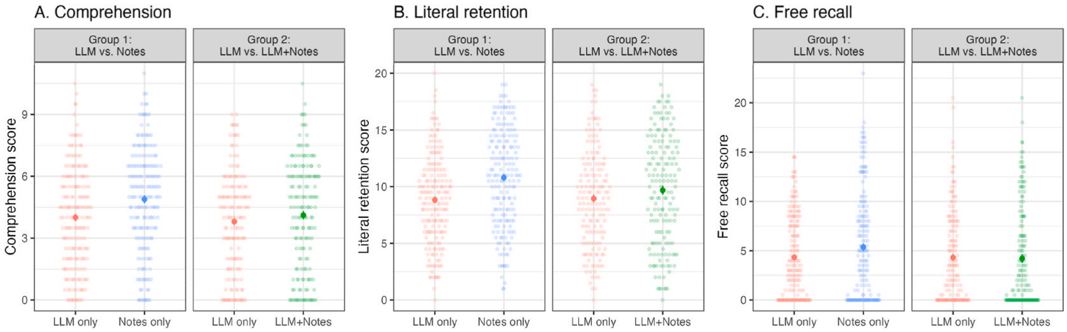


Fig. 3. Distribution of test performance by condition and group for comprehension, literal retention, and free recall. Mean values are indicated by the two large circles within each facet, whereas the smaller points show individual students’ scores. Error bars indicate one standard error above and below the mean. Group 1 is shown on the left facet of each plot, comparing LLM (red) and Notes (blue). Group 2 is on the right facet of each plot, comparing LLM (red) and LLM + Notes (green). **Comprehension** (left, max 12 points): **Group 1** - LLM: $M = 4.00$ ($SD = 2.44$), Notes: $M = 4.89$ ($SD = 2.52$); **Group 2** - LLM: $M = 3.80$ ($SD = 2.47$), LLM + Notes: $M = 4.11$ ($SD = 2.65$). **Literal retention** (middle, max 20 points): **Group 1** - LLM: $M = 8.83$ ($SD = 3.96$), Notes: $M = 10.80$ ($SD = 4.29$); **Group 2** - LLM: $M = 8.95$ ($SD = 4.29$), LLM + Notes: $M = 9.68$ ($SD = 4.83$). **Free recall** (right, max 50 points): **Group 1** - LLM: $M = 4.32$ ($SD = 4.15$), Notes: $M = 5.36$ ($SD = 5.49$); **Group 2** - LLM: $M = 4.32$ ($SD = 4.63$), LLM + Notes: $M = 4.20$ ($SD = 5.07$). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

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. This resulted in similar findings to the mixed effects model: in Group 1 we found a significant difference for free recall between the Notes and LLM conditions ($p = .02$), but did not find evidence for a significant difference in free recall for Group 2 between the LLM + Notes vs. LLM conditions ($p = .80$).

3.6.3. Quantitative exploration of students' task engagement

As planned, we analysed whether there were any differences in task engagement between the conditions. More specifically, we compared the ratings for the LLM vs. Notes conditions and the ratings for the LLM vs. LLM + Notes conditions, using paired *t*-tests for emotional and cognitive engagement variables that used 5-point Likert-Scales. We also conducted a *t*-test for time on task as an indicator of behavioural engagement. We applied Bonferroni corrections to adjust for multiple comparisons. The t-tests were conducted using base R.

3.6.4. Post-hoc power analyses

To understand the sensitivity of our design, we conducted post-hoc power analyses, assuming a modest Cohen's d of 0.20 throughout (for detailed results, see Supplements Section 6). Using such a benchmark is more meaningful than using observed effect sizes as argued by Dziak et al. (2020), who refer to this approach as a "partially post-hoc power" analysis. For learning outcomes, we ran simulation–based power analyses (100 Monte-Carlo replications, $\alpha = .05$) with the 'simr' package in R (Green & MacLeod, 2016) on the fitted mixed-effects models used in the main analyses. Results show that the pre-registered within-participant contrasts (LLM vs. Notes; LLM vs. LLM + Notes) were sufficiently powered (i.e., all above 80 %) on the outcome *literal retention* and *comprehension* but somewhat underpowered on *free recall* (64 % and 63 %, respectively). The exploratory between-participant comparison of Notes vs. LLM + Notes was slightly under-powered for *literal retention* (76 %) and severely under-powered for *comprehension* (41 %) and *free recall* (43 %).

We used the 'pwr' package in R (Champely et al., 2020) for post-hoc paired power analyses for student task engagement across the two pre-registered within-subject comparisons (LLM vs. Notes; LLM vs. LLM + Notes). Power estimates ranged from 70.5 to 77.0 %, and thus fell short of the conventional 80 % benchmark, leaving a risk that true but small differences in engagement may have gone undetected.

3.6.5. Qualitative exploration of students' activity preferences

We explored students' open response explanations for preferring one learning activity over another to gain further insight into their task engagement. The explanations were qualitatively analysed by two of the authors with the support of 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.

Four preference groups were separately analysed.

1. LLM over Notes
2. Notes over LLM
3. LLM over LLM + Notes
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). One of the authors developed the initial schemes 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. Two of the authors then iteratively refined the coding schemes by manually checking portions of the API output until a concern was detected. For example, if the API frequently assigned code A to a preference explanation where the human coder found code B more appropriate, the code descriptions for A and B were refined to clarify the difference. Depending on the concern, 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 authors were satisfied with the coding schemes. Due to the small number of responses that had to be coded ($n = 278$), one author checked the entire API output against the coding schemes and made adjustments when other codes seemed more appropriate. The final coding schemes for activity preferences can be found in Supplementary Section 7.

3.6.6. Qualitative exploration of student prompts

To better understand if students used the LLM aligned with active and constructive reading strategies, and thereby to provide potential explanations for the effects of the LLM conditions 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 researcher-developed hierarchical coding scheme through the API described above. The API was provided with detailed instructions and examples for each category, along with both texts that students were studying. Each prompt could be coded with multiple codes.

The hierarchical coding scheme was developed through several iterations. The initial version was deductively and inductively developed by one of the authors and was grounded in key literature (see Literature review), the students' task instructions (which included suggestions based on active and constructive learning), and piloting work. We intentionally did not constrain coding to definitions of a specific learning framework (e.g., active, constructive) in order to remain open to the types of behavioural prompting that the data might contain. This scheme was expanded based on the API's suggestions and the API was then asked to code the data using the coding scheme. Two authors then iteratively refined the coding scheme based on checking portions of the API output until concerns were identified (also see preference analysis). They merged, deleted, and added codes as needed and adapted code descriptions and examples to improve the quality of the API output. This iterative process continued until both authors were satisfied with the coding scheme and how the API applied codes. Finally, after a shared understanding of the codes had been created between the two authors, one of them manually checked the API output for 500 prompts (approximately 10 % of the data). More specifically, they checked the code that was assigned to a given prompt against the coding scheme, and evaluated whether or not the most appropriate code had been assigned. If another code seemed more appropriate, this was considered an error. The total error rate for the 500 prompts was 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 12.

4. Results

4.1. Student characteristics

The final sample, after applying exclusion criteria, consisted of 344 students, of which 184 students belonged to Group 1 (LLM vs. Notes conditions) and 160 students to Group 2 (LLM vs. LLM + Notes conditions). Their characteristics are reported in Table 2. In addition, both groups showed similar prior familiarity with the three learning conditions (LLM, Notes, LLM + Notes). About half of the students often or always took notes for learning compared to about 15 % for LLM use or LLM use alongside note-taking for learning (Group 2 only), indicating that students were overall more familiar with note-taking than LLM use for learning (see Supplementary Table 13 for detailed frequencies).

4.2. Comprehension and retention outcomes

To answer research question 1, 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. Descriptive statistics indicate that note-taking led to the best performance across all measures, followed by LLM + Notes, while using LLM alone resulted in the lowest scores (see Fig. 3).

Linear mixed-effects models confirmed significant differences across the conditions (see Table 3). For the complete regression results, including all covariates, see Supplementary Table 14). Specifically, students performed significantly better with Notes compared to LLM and with LLM + Notes compared to LLM for both literal retention and comprehension. For free recall, students showed significantly 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. We also compared the coefficients of Notes and LLM + Notes for literal retention and comprehension. The results indicate that the effect of Notes was significantly larger than that of LLM + Notes, using LLM as the common reference group, for both literal retention (Estimated mean difference = 1.35, $SE = 0.38$, $z = 3.60$, $p < .001$) and comprehension (Estimated mean difference = 0.59, $SE = 0.24$, $z = 2.43$, $p = .015$).

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 over LLM use tended to be stronger and 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.

Table 2
Student characteristics by group and overall totals (after exclusion, N = 344).

Characteristic	Group 1	Group 2	Total
	N students (%)	N students (%)	N students (%)
Male	102 (29.7 %)	78 (22.7 %)	180 (52.3 %)
Female	57 (16.6 %)	63 (18.3 %)	120 (34.9 %)
Other gender	1 (0.3 %)	1 (0.3 %)	2 (0.6 %)
Prefer not to say gender	2 (0.6 %)	0 (0.0 %)	2 (0.6 %)
Receives FSM	9 (2.6 %)	10 (2.9 %)	19 (5.5 %)
No FSM	160 (46.5 %)	163 (47.4 %)	323 (93.9 %)
English first language	130 (37.8 %)	117 (34.0 %)	247 (71.8 %)
Other first language	2 (0.6 %)	3 (0.9 %)	5 (1.5 %)
Bilingual	35 (10.2 %)	29 (8.4 %)	64 (18.6 %)
GCSE History	99 (28.8 %)	80 (23.3 %)	179 (52.0 %)
No GCSE History	81 (23.5 %)	58 (16.9 %)	139 (40.4 %)

Note: FSM = Free school meals.

Table 3
Effects of Notes and LLM + Notes conditions, relative to the LLM condition, on literal retention, comprehension, and free recall.

Condition	*β*	SE	95 % CI	*t*	*p*	*d*
Literal retention						
Notes	1.92	0.26	[1.42, 2.42]	7.50	<0.001	0.44
LLM + Notes	0.57	0.28	[0.03, 1.11]	2.06	0.040	0.13
Comprehension						
Notes	0.95	0.17	[0.62, 1.28]	5.73	<0.001	0.38
LLM + Notes	0.35	0.18	[0.00, 0.70]	1.98	0.048	0.14
Free recall						
Notes	1.02	0.43	[0.18, 1.86]	2.39	0.017	0.21
LLM + Notes	−0.08	0.46	[-0.98, 0.81]	−0.18	0.854	−0.02

4.3. *Student task engagement*

To answer research question 2, we explored how students' task engagement compared when using the LLM on its own compared to taking notes (Group 1) or using the LLM alongside note-taking (Group 2). The quantitative results for emotional, cognitive and behavioural (time on task only) engagement are summarised in Fig. 4, with details of statistical tests in Appendix B). We used an adjusted p-value threshold of 0.05/14 = 0.004 to gauge statistical significance based on the Bonferroni correction to account for multiple comparisons (n = 14).

4.3.1. *Emotional engagement*

Students' task interest did not differ when using the LLM on its own compared to when taking notes or when using the LLM alongside notes. However, students reported higher activity enjoyment when using the LLM compared to note-taking (but not LLM alongside note-taking).

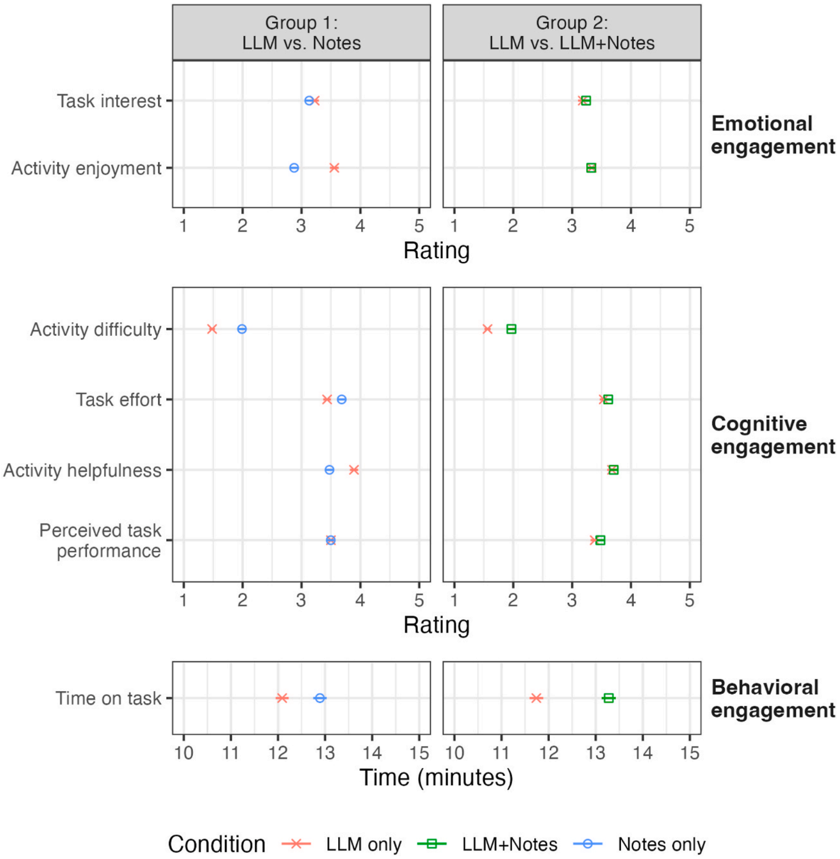


Fig. 4. Differences in student engagement by group and condition. The top and middle panels show ratings for emotional and cognitive engagement measures, respectively, on a 1–5 scale. The bottom panel displays time on task, a behavioural engagement measure, in minutes. Each point represents the mean for a condition, with error bars indicating one standard error above and below the mean.

4.3.2. Cognitive engagement

Students perceived the LLM-only condition as less difficult compared to Notes and LLM + Notes. Similarly, students reported less effort investment in the LLM-only condition compared to Notes, but not LLM + Notes. Contrary to performance outcomes, students thought that the LLM was more helpful for understanding and learning the text compared to note-taking. Again, such differences were not found between the LLM-only and LLM + Notes conditions. There were no significant differences in perceived task performance.

Students were asked to indicate their preferred learning activities and explain their preferences through an open response (see Table 4). In Group 1, most students preferred the LLM activity over 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 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 activities, 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.

4.3.3. Behavioural engagement

Students had to stay on task for at least 10 min and a maximum of 15 min. Students spent significantly less time on task when using

Table 4
Learning activity preferences and reasons by group.

Activity preference	N students	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.

only the LLM compared to note-taking (Group 1) and LLM use alongside note-taking (Group 2), indicating a reduction in behavioural engagement when the LLM was available (see Fig. 4 and Appendix B). As an additional check on the possibility that differences in behavioural engagement might explain the findings, we examined total time-on-task across conditions. Including the time-on-task variable did not change the experimental condition effects. While students in the note-taking and LLM + Notes conditions spent slightly longer on the task (≈1–1.5 min), linear mixed-effects models indicated that time-on-task did not significantly predict performance on any outcome measure (retention: $p = .237$; comprehension: $p = .348$; free recall: $p = .559$). This suggests that the observed effects are not attributable to engagement duration.

Process data for Group 2 shows that students' prompt frequency was lower ($M = 6.02$, $SD = 4.64$) when having access to notes alongside the LLM compared to when using the LLM alone ($M = 9.21$, $SD = 5.72$). Students in Group 1 made a similar number of prompts when using the LLM on its own ($M = 10.98$, $SD = 6.46$). While students wrote a similar number of words in their notepad in both Notes ($M = 100.74$, $SD = 115.63$, Group 1) and LLM + Notes conditions ($M = 103.83$, $SD = 158.24$, Group 2), a large number of students in the LLM + Notes condition copied heavily from the LLM output into the notepad. Specifically, 25.63 % of students had a substantial overlap of more than 70 % of trigrams, and 16.25 % showed nearly complete copying (more than 90 % overlap of trigrams). This indicates that students' own note-taking efforts were reduced, and that the effectiveness of note-taking may be diminished.

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 unsure. 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.

4.3.4. *Summary of student task engagement results*

Overall, the student engagement findings indicate that while note-taking was more effective than LLM use, it was less favourably experienced across different emotional, cognitive and behavioural dimensions. This is particularly evident in lower perceived enjoyment and helpfulness, higher perceived difficulty, and overall lower preference and reported future use. At the same time, there are indicators that students might invest more effort and behaviourally engage with the learning content for a longer time when taking notes compared to when using the LLM. Importantly, while LLM + Notes was also more effective than LLM-only, students reported similar emotional engagement, and the majority of students preferred the combined activity over using the LLM alone.

4.4. *The types of LLM prompts*

To answer research question 3, we analysed what type of prompts students submitted to the LLM during the learning tasks. The qualitative analysis of all prompts (n = 4929) revealed four behavioural archetypes of how students worked with the LLM in relation to the task, namely 'seeking additional information and deeper understanding', 'information condensation', 'understanding the text', and 'study and memory help'. An additional overarching prompt type that captured learners' interaction with the LLM's functionality. Finally, only one type was identified not to be directly related to the task, specifically 'irrelevant, off-topic, miscellaneous' (see Appendix C for the distribution of overarching prompt types across each LLM session). For all specific prompt types with frequencies, see Appendix D.

The most frequent archetype was 'seeking additional information and deeper understanding' (2265 prompts). 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

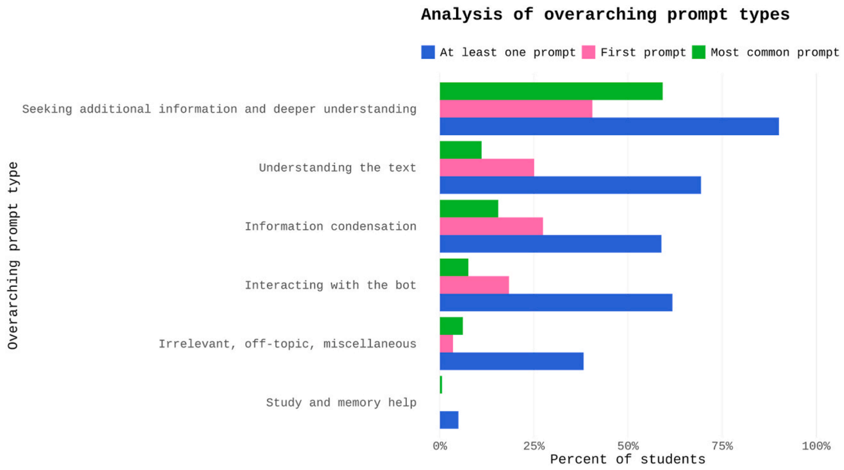


Fig. 5. Distribution of student prompts across different types, showing the percentage of students who used the prompt type at least once (blue), as their most common prompt (magenta), and as their first prompt (green). Prompt types are arranged by overall frequency. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

prompt (see Fig. 5). These prompts primarily comprised requests for elaboration (1479 instances) and general background information (514 instances). Examples include "how are people today affected by the apa[r]theid" and "why did it take so long to free nelson mandela".

'Information condensation' (749 prompts) 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, 'understanding the text' (615 prompts), refers to students seeking a basic understanding of the text and 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) but 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 mnemonic devices ("create me a mnemonic device on the cuban missile crisis").

Beyond these archetypes, 760 prompts focused on interacting with the LLM's functionality rather than (or in addition to) text content, 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 trustworthiness. Finally, about 10 % of all interactions (501 prompts) 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.

5. Discussion

This study provides new insights into how the use of LLMs for comprehension-focused reading compares to and interacts with the more traditional, evidence-based practice of note-taking in supporting 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 may influence educational outcomes. Specifically, we found that note-taking — whether done alone or alongside LLM usage — led to higher comprehension and retention scores compared to using an LLM alone, underscoring the importance and effectiveness of more traditional learning activities. This is interesting because our exploratory prompt analysis showed that, in most cases, the LLM was used in a variety of task-relevant ways by learners, including in ways that align with frameworks of reading comprehension and learning. For example, learners' prompts targeted different levels of the text (McNamara & Magliano, 2009). Three prompt types, in particular - 'Understanding the text', 'Information Condensation' and 'Seeking additional information and deeper understanding' may have supported learners in processing the surface structure of the text (words and grammar), the textbase (ideas within the text) and developing a situational model of text.

In addition, in relation to Chi's learning framework (2009), it is clear that 'Seeking additional information and deeper understanding' and 'Understanding the text' archetypes reflect active processing of the text (Chi, 2009). The learners had to have engaged with the text directly in order to have selected which particular information they wished to ask the LLM about and then to have constructed a question/prompt for the LLM about it. Similarly, other prompt archetypes may also involve active processing at some level and have the potential to have been used as part of a more constructive process. However, it was not possible, with the current data, to determine whether this was the case, as this would depend on how the learners engaged with the text in the first place, why they asked the LLM for that particular information, and how they then engaged with the LLM's response to their prompt; this data is not captured within the prompt data. For example, 'Information condensation' may have involved more passive engagement if learners relied on the LLM to summarise content rather than generating their own understanding.

Given these seemingly useful ways of using the LLM, what could explain why the LLM condition was outperformed by note-taking? One possible explanation may lie in the precise ways that learners used the LLM. As noted, the prompt data could not tell us how constructive the learners had been with the LLM, but the prompt archetypes suggested that learners may have viewed the LLM as 'a knowledgeable other' and used it primarily to receive information about the text. A different explanation of the advantage of note-taking for learning outcomes might lie in the way comprehension and retention were assessed. Our assessment measured only comprehension and retention of the given texts, it did not capture broader learning. Thus, the differences between conditions could reflect complementary outcomes: note-taking supported mastery of the assigned material, while LLM use facilitated both engagement with the texts and intellectual exploration beyond them. For example, in a passage about apartheid in South Africa that mentions Nelson Mandela's involvement in the anti-apartheid movement, one student asked, "What was Mandela's life story?" Similarly, in a passage on the Cuban Missile Crisis, 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.

Furthermore, students perceived the LLM as more helpful and preferable to note-taking alone, which again seem to contradict the main finding demonstrating more positive learning effects of note taking. One possible explanation may be that students simply have limited metacognitive knowledge about what is in fact helpful for their own learning (Bjork et al., 2013; Geraci et al., 2022; Kruger & Dunning, 1999), specifically in the context of GenAI (Tankelevitch et al., 2024). They may also underweight the importance of the "desirable difficulties" induced by activities such as note-taking (Bjork & Bjork, 2011). Note-taking requires active processing of information, such as identifying important information, paraphrasing, and summarising (Kobayashi, 2005). 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 (Grund et al., 2024). 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, as indicated by students' reports of task difficulty and effort as well time spent on task. Similar effects of LLM use on learners' affective-motivational state and mental effort were found in Deng et al.'s meta-analysis (Deng et al., 2025).

Overall, our findings demonstrate the value of using note-taking and LLM + note-taking over LLM alone for comprehension-focused reading. Although the combined activity had a comparatively smaller benefit over LLM use alone than note-taking alone, it did not suffer from the same negative effects on emotional engagement. Students preferred LLM + Notes over LLM use alone, and LLM use appeared to encourage exploration of ideas that extended beyond the given text (but still relevant to the text). As such, the propensity to engage with this learning activity might be increased. This highlights the opportunity and challenge of how to combine traditional evidence-based activities like note-taking with the unique benefits offered by LLMs. Rather than viewing these as competing alternatives, we could think of them as complementary that, when thoughtfully integrated, may 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 (Starkey, 2020; Wang & Shi, 2021).

5.1. Pedagogical implications

Our work suggests several practical implications. The first and easiest approach would be to separate LLM use from note-taking, given that we observed that many students simply copied and pasted LLM output into their notepads. Students would first read a text and interact with an LLM to clarify and explore its content. Following this they would take notes independently, without the ability to simply copy and paste output from the LLM. Preventing this shortcut, which might have reduced the effectiveness of the combined activity, might encourage students to synthesise and internalise information themselves. Alternatively, if the LLM and note-taking are used simultaneously, pasting into the notepad could be blocked. This is a small but likely meaningful design choice that emerged through our work and could be tested in future research.

Second, our findings indicate that secondary school students require more comprehensive training than the instructions we provided to use LLMs in ways that may benefit their learning on a par with or beyond traditional note-taking. Educators are well positioned to work closely with students, actively teaching and guiding them to use LLMs in ways that are goal-focused (e.g., understanding immediate text content or broadening their understanding) and align with active and constructive learning strategies (e.g., clarifying specific misunderstandings, engaging in critical thinking, and integrating information) (Chi, 2009; McNamara, 2007). Given that our data suggests a reduction in students' effort expenditure when using the LLM, it seems crucial to guide students to use the LLM in ways that support and amplify rather than reduce their cognitive processing. While not directly examined in this study, the literature suggests that encouraging students to engage in interactive dialogue rather than relying on the passive consumption of automatic summaries and explanations is particularly important. This aligns with the conceptualisation of AI tools as "thought partners'' that support existing human cognitive processes rather than disrupt them (Dwivedi et al., 2023). Furthermore, software could be configured to support these goals by limiting distracting behaviour, such as asking irrelevant questions which constituted about 10 % of prompts, and encouraging productive use (plausibly by capturing data and using the LLM to provide feedback 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. For example, through analysing the prompts in our experiments, it became clear that students were curious about the tenets of communism and why they provoked such fear and opposition in the U.S. Teachers could use such insights at an individual level and 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.

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 (see Deng et al., 2025), 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 (Deng et al., 2025). 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.

5.2. Limitations and future directions

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.

We compared LLM use with note-taking and LLM use alongside note-taking but did not include an additional "reading-only" control condition. As explained in the introduction, this was due to concerns about insufficient statistical power given constraints on the number of participants we could recruit and in order to limit participant fatigue in responding to conditions (for the same reasons, the study did not include a third group that experienced both the Notes and LLM + Notes conditions). We chose a strong baseline (i.e., note-taking) instead, because most activities going beyond passive reading are likely to be more effective than passive reading, so that the comparison of a frequently used and effective learning activity seemed more informative. And yet, only about 50 % of students reported that they 'often' or 'always' used note-taking for learning, leaving a substantial number of students that did not frequently engage in note-taking. As such, the study would have benefited from a passive reading condition to gauge the effectiveness of LLM use *per se*.

In addition, although our analyses tested whether Notes versus LLM + Notes had a greater advantage over the LLM-only condition, we could not conduct a direct comparison of those two conditions. This was because the design was not optimised for this analysis (see Methods). Although the design allowed us to conduct an indirect comparison of coefficients between these two conditions, we recognise that these comparison outcomes still need to be considered with caution because of the potential interaction effects (each group could have experienced the LLM differently). This is particularly a concern for subjective measures, for which we refrained from conducting any between-group comparisons.

While human input was essential in the exploratory qualitative analyses, we did not assess inter-rater agreement through independent double coding. We instead chose verifying the API output to balance efficiency and accuracy. Although independent coding would be more rigorous, we believe this approach does not significantly threaten our findings' validity. The careful development of the coding schemes involved the creation of a shared understanding of the codes by two authors, with which the API coding was aligned over several iterations. This ensured that the API coded the data largely as we intended as evidenced by the low error rate in the final prompt analysis. Importantly, we checked, and when necessary adjusted, the entire API output for the preference analysis.

Furthermore, while our qualitative coding scheme captured diverse patterns of LLM use, it did not classify each prompt according to Chi's (2009) active–constructive–interactive framework. The primary reason was that the framework is best applied across multiple conversational turns rather than at the level of single prompts. We made a deliberate methodological choice to analyse at single-prompt level in order to identify the kinds of information students sought from the LLM, rather than focusing on interactional dynamics. This approach ensured that the analysis remained aligned with the exploratory goals of the study. As a result, however, our findings cannot fully disentangle whether outcomes were driven by more interactive/active/constructive uses of the LLM. The implication is that our qualitative analysis primarily documents naturalistic patterns of use rather than determining which kinds of LLM use are most beneficial.

Single items were used to assess students' self-reported task engagement, which limits variation and may reduce the reliability of the items compared to multi-item measures (Allen et al., 2022). We used single items to reduce the time spent on surveys, to limit participant fatigue, and allow more time on the learning task. More comprehensive measures would have been preferable, but we believe that the current items have been informative and are not a significant threat to the validity of our findings. In fact, the use of single items to assess perceived task difficulty, interest, and similar constructs is not uncommon (e.g., Nuutila et al., 2021; Robinson, 2001).

Furthermore, the study was limited to a single, isolated learning activity outside of the context of normal use throughout an entire course of study. Repeated use, particularly repeated LLM use, or use in other settings (e.g., in everyday classrooms or independently for homework, unsupervised) could yield different results. 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.

Lastly, it should be acknowledged that the LLM model used (OpenAI's GPT-3.5 turbo model) is not as advanced as some of the more recent models. It is crucial for future research to explore which ways of interacting with LLMs most effectively enhance learning outcomes. For example, studies could compare specific prompt types and LLM uses (e.g., information condensation vs. elaboration and deeper understanding) by creating experimental conditions that are restricted to each use (Weidlich et al., 2025). Alternatively, with a large enough sample size, studies could use profile analysis to identify naturally occurring prompt patterns and link these to learning outcomes. It will also be important to analyse the nature of students' conversations or dialogues with LLMs, in addition to the types of individual prompts. This goes beyond the scope of the current study, but a dialogue analysis would reveal to which extent the conversations are of an interactive (i.e., both partners meaningfully contribute to the conversation) or individual (i.e., only one partner meaningfully contributes to the conversation) nature (see Chi, 2009). Such in-depth qualitative analyses of prompting behaviour are also needed to review theories of learning, so that they reflect the changing learning environment.

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 (Huber et al., 2024). 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 (Dwivedi et al., 2023). This research marks only the beginning of understanding how to effectively use LLMs to complement existing activities and tools while maintaining students' cognitive engagement. It also highlights which LLM uses to consider exploring in future experimental work.

5.3. Conclusions

In summary, this is one of the first large-scale quantitative studies providing evidence on the effects of LLM use on reading comprehension and retention. Our findings reaffirm the importance of traditional learning activities like note-taking, which foster deep cognitive engagement and enhanced learning outcomes. At the same time, LLMs introduce new opportunities for learning — especially for comprehension-focused reading — by helping students clarify, explore, and contextualise material, and by enhancing emotional engagement. However, to be effective, these tools must be used with appropriate guidance that supports cognitive engagement rather than bypasses it. Rather than viewing these tools as a disruption to be resisted, educators and researchers have the 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.

CRediT authorship contribution statement

Pia Kreijkes: Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Viktor Kewenig:** Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Martina Kuvalja:** Writing – original draft, Visualization, Resources, Methodology, Investigation, Formal analysis. **Mina Lee:** Writing – review & editing, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Jake M. Hofman:** Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Sylvia Vitello:** Writing – review & editing, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Abigail Sellen:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Sean Rintel:** Writing – review & editing, Methodology, Conceptualization. **Daniel G. Goldstein:** Writing – review & editing, Methodology, Conceptualization. **David Rothschild:** Writing – review & editing, Methodology, Conceptualization. **Lev Tankelevitch:** Writing – review & editing, Methodology, Conceptualization. **Tim Oates:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Code availability

The corresponding code can be found in the Open Science Framework at: https://osf.io/56w9j/?view_only=3540e72af04b4f71984915f6c98998c9.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Microsoft Copilot and Open AI’s ChatGPT in order to improve the readability and language of some sentences. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of 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.

Acknowledgments

We thank Dr Tom Benton and Dr Matthew Carroll from Cambridge University Press & Assessment for their valuable advice on the analyses conducted in this study, as well as Elinor Lieber for supporting us with data collection. We also thank Abigail Agyemang, Caroline Breeden, and Megan Ennion for scoring participants’ test responses.

Appendix A. Participant exclusion criteria

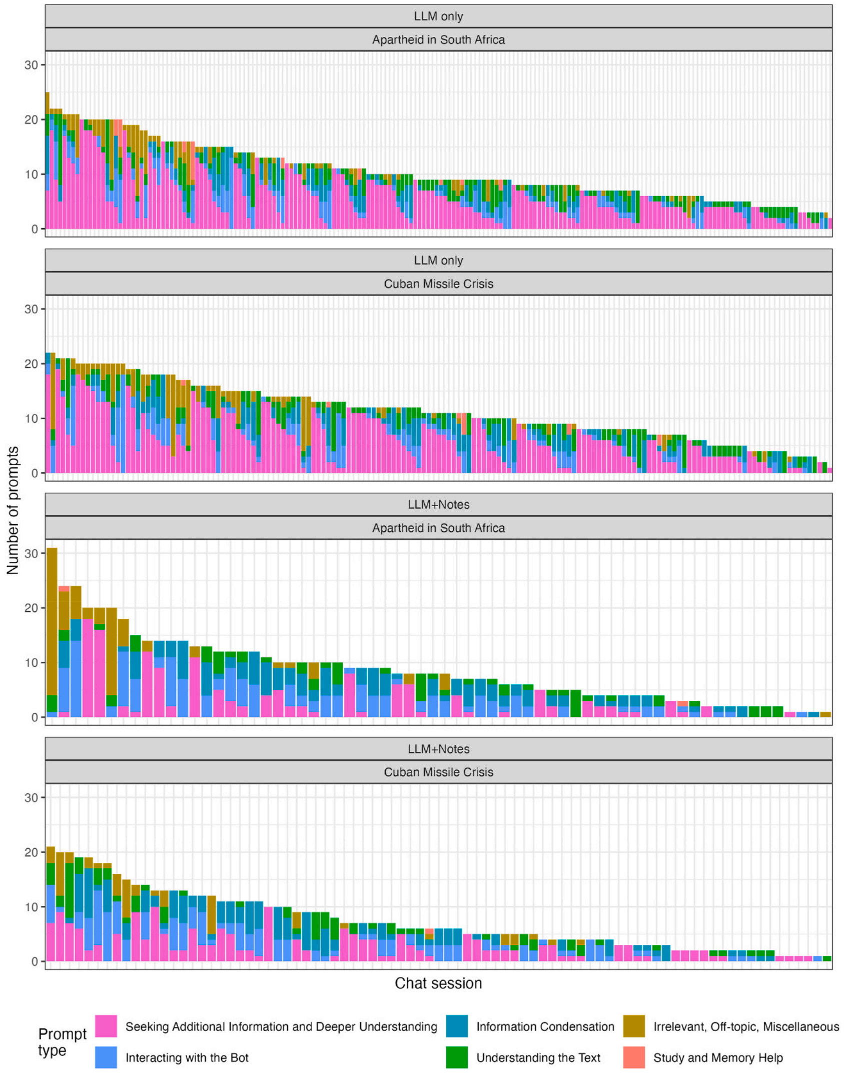
Participants (n = 61) were excluded for the following reasons.

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 min 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)

Appendix B. Differences in student task engagement between conditions (for Group 1 and Group 2)

Variable	Group 1: LLM vs. Notes					Group 2: LLM vs. LLM + Notes				
	Diff.	*t(df)*	*p*	95 % CI	*d*	Diff.	*t(df)*	*p*	95 % CI	*d*
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
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
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 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
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
Time on task	**−0.80**	**−4.62(183)**	**< 0.001**	**[-1.15, -0.46]**	**−0.34**	**−1.54**	**−8.26(159)**	**< 0.001**	**[-1.91, -1.17]**	**−0.66**

Appendix C. Distribution of overarching prompt types across LLM sessions for different conditions and students Specific prompt types with frequencies



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.

Appendix D. Specific prompt types with frequencies

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

Note. This table only includes prompt types that have been used at least three times by students. Prompts are ordered by most frequent overarching prompt types (starting with task-relevant prompts) and within that by most frequent specific prompt type.

Appendix E. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2025.105514.

Data availability

All quantitative data can be found in the Open Science Framework at: https://osf.io/56w9j/?view_only=3540e72af04b4f71984915f6c98998c9. We did not provide the following qualitative data as that would risk sharing identifiable information: Students' LLM interactions (only the applied codes are being shared), students' notes, students' activity preferences (only applied codes are being shared).

References

Aleksić-Maslać, K., Borović, F., & Biočina, Z. (2024). Perception and usage of ChatGPT in the education system. In *INTED2024 proceedings* (pp. 1842–1848). https://doi.org/10.21125/inted.2024.0511

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment, 38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*(3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3

Barrett, A., & Pack, A. (2023). 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. https://doi.org/10.1186/s41239-023-00427-0

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bernabei, M., Colabianchi, S., Falegnami, A., & Costantino, F. (2023). 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*, Article 100172. https://doi.org/10.1016/j.caeai.2023.100172

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Bjork, E. L., & Bjork, R. A. (2011). 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* (pp. 56–64). Worth Publishers.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Bohay, M., Blakely, D. P., Tamplin, A. K., & Radvansky, G. A. (2011). Note taking, review, memory, and comprehension. *American Journal of Psychology, 124*(1), 63–73. https://doi.org/10.5406/amerjpsyc.124.1.0063

British Educational Research Association. (2018). In *Ethical Guidelines for educational research* (4th ed.) https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018.

Bui, D. C., & Myerson, J. (2014). The role of working memory abilities in lecture note-taking. *Learning and Individual Differences, 33*, 12–22. https://doi.org/10.1016/j.lindif.2014.05.002

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41–75). The Guilford Press.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). Pwr: Basic functions for power analysis. https://CRAN.R-project.org/package=pwr.

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. *Advances in instructional psychology: Educational design and cognitive science, 5*, 161–238. Lawrence Erlbaum Associates Publishers.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Couch, S. P., Bray, A. P., Ismay, C., Chasnovski, E., Baumer, B. S., & Çetinkaya-Rundel, M. (2021). Infer: An R package for tidyverse-friendly statistical inference. *Journal of Open Source Software, 6*(65), 3661. https://doi.org/10.21105/joss.03661

Craik, F. I. M. (2002). Levels of processing: Past, present . . . And future? *Memory, 10*(5–6), 305–318. https://doi.org/10.1080/09658210244000135

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daniels, H. (2001). *Vygotsky and pedagogy*. Routledge.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37. https://doi.org/10.1016/s0010-0277(00)00123-2

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Ding, J., Chen, K., Liu, H., Huang, L., Chen, Y., Lv, Y., Yang, Q., Guo, Q., Han, Z., & Lambon Ralph, M. A. (2020). A unified neurocognitive model of semantics language social behaviour and face recognition in semantic dementia. *Nature Communications, 11*(1), 2595. https://doi.org/10.1038/s41467-020-16089-9

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). 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*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Dziak, J. J., Dierker, L. C., & Abar, B. (2020). The interpretation of statistical power after the data have been gathered. *Current Psychology, 39*(3), 870–877. https://doi.org/10.1007/s12144-018-0018-1

Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F. F., Al-shami, S. A., & Slamene, R. (2023). Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Computers and Education: Artificial Intelligence, 5*, Article 100180. https://doi.org/10.1016/j.caeai.2023.100180

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience, 25*(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Geraci, L., Kurpad, N., Tirso, R., Gray, K. N., & Wang, Y. (2022). Metacognitive errors in the classroom: The role of variability of past performance on exam prediction accuracy. *Metacognition and Learning*. https://doi.org/10.1007/s11409-022-09326-7

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395. https://doi.org/10.1037/0033-295x.101.3.371

Green, P., & MacLeod, C. J. (2016). Simr: An r package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grund, A., Fries, S., Nückles, M., Renkl, A., & Roelle, J. (2024). When is learning "Effortful"? Scrutinizing the concept of mental effort in cognitively oriented research from a motivational perspective. *Educational Psychology Review, 36*(1), 11. https://doi.org/10.1007/s10648-024-09852-7

Guthrie, J. T., & Wigfield, A. (2000). Engagement and motivation in reading. In *Handbook of reading research* (Vol. III, pp. 403–422). Lawrence Erlbaum Associates Publishers.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education. In *Promise and implications for teaching and learning* (1st ed.). Center for Curriculum Redesign.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50*(3), 346–363.

Huber, S. E., Kiili, K., Nebel, S., Ryan, R. M., Sailer, M., & Ninaus, M. (2024). Leveraging the potential of large Language models in education through playful and game-based learning. *Educational Psychology Review, 36*(1), 25. https://doi.org/10.1007/s10648-024-09868-z

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity, 20*(1), 2. https://doi.org/10.1007/s40979-024-00149-4

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*. https://doi.org/10.1016/j.lindif.2023.102274

Kiewra, K. A. (1985). Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist, 20*(1), 23–32. https://doi.org/10.1207/s15326985ep2001_4

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

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kobayashi, K. (2005). What limits the encoding effect of note-taking? A meta-analytic examination. *Contemporary Educational Psychology, 30*(2), 242–262. https://doi.org/10.1016/j.cedpsych.2004.10.001

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kruger, J., & Dunning, D. (1999). 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. https://doi.org/10.1037//0022-3514.77.6.1121

Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Math education with large Language models: Peril or promise? Preprint. *SSRN*.

Lan, D. H., & Tung, T. M. (2023). Analyzing the impact of Chat-GPT usage by university students in Vietnam. *Migration Letters, 20*(S10), 259–268. https://doi.org/10.59670/ml.v20iS10.5134

Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). April. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 CHI conference on human factors in computing systems* (pp. 1–22). New York, NY, USA: Association for Computing Machinery.

Linderholm, T., Virtue, S., Tzeng, Y., & Van den Broek, P. (2018). Fluctuations in the availability of information during reading: Capturing cognitive processes using the landscape model. *Accessibility in Text and Discourse Processing* (pp. 165–186). Routledge.

Luckin, R., Holmes, W., & Forcier, L. B. (2016). *Intelligence Unleashed: An argument for AI in Education*. Open Ideas at Pearson/UCL. https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/innovation/Intelligence-Unleashed-Publication.pdf.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19. https://doi.org/10.1037/0003-066X.59.1.14

McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theories, interventions, and technologies*. Lawrence Erlbaum Associates Publishers.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers, 36*(2), 222–233. https://doi.org/10.3758/BF03195567

McNamara, D. S., & Magliano, J. (2009). Toward a Comprehensive Model of Comprehension. In. *Psychology of Learning and Motivation, 51*, 297–384. https://doi.org/10.1016/S0079-7421(09)51009-2

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mullen, L. (2020). Textreuse: Detect text reuse and document similarity. https://cran.r-project.org/package=textreuse.

Nuutila, K., Tapola, A., Tuominen, H., Molnár, G., & Niemivirta, M. (2021). Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task engagement. *Learning and Individual Differences, 92*, Article 102090. https://doi.org/10.1016/j.lindif.2021.102090

Oakhill, J. V., Berenhaus, M. S., & Cain, K. (2015). Children's reading comprehension and comprehension difficulties. In *The Oxford handbook of reading* (pp. 344–360). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.001.0001.

Ofcom. (2024). Online nation 2024 report. www.ofcom.org.uk.Ofcom. https://www.ofcom.org.uk/media-use-and-attitudes/online-habits/online-nation/.

Pan, M., Lai, C., & Guo, K. (2025). Effects of GenAI-empowered interactive support on university EFL students' self-regulated strategy use and engagement in reading. *The Internet and Higher Education, 65*, Article 100991. https://doi.org/10.1016/j.iheduc.2024.100991

Pascual-Leone, A., Amedi, A., Fregni, F., & Merabet, L. B. (2005). The plastic human brain cortex. *Annual Review of Neuroscience, 28*, 377–401. https://doi.org/10.1146/annurev.neuro.27.070203.144216

Pearson, P. D., Roehler, L. R., Dole, J. A., & Duffy, G. G. (1990). *Developing expertise in reading comprehension: What should be taught? How should it be taught?* University of Illinois Urbana-Champaign Center for the Study of Reading. https://hdl.handle.net/2142/17648.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In *The science of reading: A handbook* (pp. 227–247). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch13.

Python Software Foundation. (2022). Python programming language (Version 3.11). https://www.python.org/.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(11), 1698–1723. https://doi.org/10.1037/xlm0001110

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied, 23*(3), 293–300. https://doi.org/10.1037/xap0000134

Salmela-Aro, K., Tang, X., & Symonds, J. (2021). Student engagement in adolescence: A scoping review of longitudinal studies 2010–2020. *Journal of Research on Adolescence, 31*, 256–272. https://doi.org/10.1111/jora.12619

Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large Language models. In *Proceedings of the 2022 ACM conference on international computing Education research* (Vol. 1, pp. 27–43). https://doi.org/10.1145/3501385.3543957

Schiller, R., Fleckenstein, J., Mertens, U., Horbach, A., & Meyer, J. (2024). Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education, 223*, Article 105163. https://doi.org/10.1016/j.compedu.2024.105163

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access, 11*, 38805–38818. https://doi.org/10.1109/ACCESS.2023.3268224

Sinatra, G. M., Heddy, & Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educational Psychologist, 50*(1), 1–13. https://doi.org/10.1080/00461520.2014.1002924

Singh, N., Bernal, G., Savchenko, D., & Glassman, E. L. (2022). Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction*. https://doi.org/10.1145/3511599

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, Article 108386. https://doi.org/10.1016/j.chb.2024.108386

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

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

Steponenaite, A., & Barakat, B. (2023). Plagiarism in AI empowered world. In M. Antona, & C. Stephanidis (Eds.), *Universal access in human-computer interaction* (pp. 434–442). Nature Switzerland: Springer. https://doi.org/10.1007/978-3-031-35897-5_31.

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3613904.3642902

Taylor, C. (2018). 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. https://doi.org/10.1080/00071005.2017.1330464

Urban, M. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education, 215*. https://doi.org/10.1016/j.compedu.2024.105031

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vygotsky, L. S. (1978). In M. Cole, V. John-Steiner, S. Scribner, E. Souberman, & Trans (Eds.), *Mind in society: The development of higher psychological processes*. Harvard University Press.

Walton Family Foundation. (2023). Teachers and students embrace ChatGPT for education. In *Walton family foundation*. Walton Family Foundation. https://www.waltonfamilyfoundation.org/learning/teachers-and-students-embrace-chatgpt-for-education.

Wang, H., & Shi, W. (2021). Practical approaches to integrated values education for foreign language majors. *Foreign Language World, 6*, 38–45.

Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning, 41*(5). https://doi.org/10.1111/jcal.70105.

Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology, 116*(1), 48–75. https://doi.org/10.1037/edu0000833

Zhai, X., Nyaaba, M., & Ma, W. (2024). Can generative AI and ChatGPT outperform humans on cognitive-demanding problem-solving tasks in science? *arXiv*. http://arxiv.org/abs/2401.15081.

Zhu, J.-J., Jiang, J., Yang, M., & Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology, 57*(46), 17667–17670. https://doi.org/10.1021/acs.est.3c01818

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162