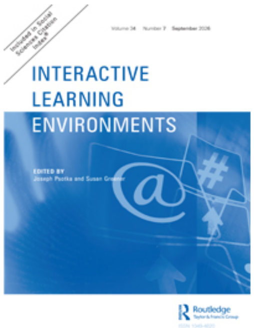

Interactive Learning Environments

ISSN: 1049-4820 (Print) 1744-5191 (Online) Journal homepage: www.tandfonline.com/journals/nile20

Comparing structural and conversational AI summarisation for time-constrained academic reading support among graduate students

Muhan Xu, Xiaomei Li, Te Lan, Xiaoke Zeng, Tianhong Wang, Sixiong Sheng, Xiaolin Deng, Jingjing Sun, Rebecca Fiebrink, Nick Bryan-Kinns & Mick Grierson

To cite this article: Muhan Xu, Xiaomei Li, Te Lan, Xiaoke Zeng, Tianhong Wang, Sixiong Sheng, Xiaolin Deng, Jingjing Sun, Rebecca Fiebrink, Nick Bryan-Kinns & Mick Grierson (01 Mar 2026): Comparing structural and conversational AI summarisation for time-constrained academic reading support among graduate students, Interactive Learning Environments, DOI: 10.1080/10494820.2025.2604649

To link to this article: https://doi.org/10.1080/10494820.2025.2604649

Published online: 01 Mar 2026.

INTERACTIVE LEARNING ENVIRONMENTS
https://doi.org/10.1080/10494820.2025.2604649

Comparing structural and conversational AI summarisation for time-constrained academic reading support among graduate students

Muhan Xu [a,b], Xiaomei Li[c], Te Lan[a], Xiaoke Zeng[d], Tianhong Wang[e], Sixiong Sheng[b,f], Xiaolin Deng[b], Jingjing Sun[g], Rebecca Fiebrink[a], Nick Bryan-Kinns[a] and Mick Grierson[a]

[a]Creative Computing Institute, University of the Arts London, London, UK; [b]Ponder AI, Singapore, Singapore; [c]College of Design and Innovation, Tongji University, Shanghai, People's Republic of China; [d]School of Creative Media, City University of Hong Kong, Hong Kong, People's Republic of China; [e]School of Psychological and Cognitive Sciences, Peking University, Beijing, People's Republic of China; [f]Judge Business School, University of Cambridge, Cambridge, UK; [g]Dyson School of Design Engineering, Imperial College London, London, UK

ABSTRACT

Academic reading can be challenging for students due to the complex structure and language of academic texts. Emerging AI-assisted approaches can facilitate academic reading, but how different AI-assisted approaches impact students requires further research. To investigate this, we conducted a user study ($n = 29$) where participants completed paper categorisation tasks under three conditions: (1) with conversational AI summarisation, (2) with structural AI summarisation, and (3) without AI support. Our findings indicate that structural AI summarisation helps students navigate text structure and locate information, while demonstrating a more positive user experience and lower perceived cognitive load compared to conversational AI summarisation. However, concerns emerged, including time-consuming verification and over-reliance. We build on our results to suggest design opportunities for future AI-assisted reading support tools: (1) combining structural overviews with conversational flexibility, (2) aligning summarisation with reading objectives through intent clarification, (3) embedding verification features that maintain reading flow, and (4) shifting from direct answer provision to guided questioning that promotes critical engagement. Our findings are bounded by 10-minute paper categorisation tasks performed by HCI graduate students across two LLM implementations. Our paper contributes empirical findings comparing AI summarisation approaches and suggests design opportunities that support interactive and critically engaged academic reading.

ARTICLE HISTORY
Received 22 July 2025
Accepted 6 December 2025

KEYWORDS
AI summarisation; academic reading; structural summarisation; conversational AI; empirical study

1. Introduction

Understanding academic reading can be challenging for students, as influenced by text structure (Kendeou & Van Den Broek, 2007; Kintsch & Van Dijk, 1978; Peregoy & Boyle, 2000), long sentences (Dubay, 2004), and background knowledge (Ahmadi et al., 2013; Cromley & Azevedo, 2007; Kintsch & Van Dijk, 1978; Yi, 2014). Large Language Models (LLMs) offer potential solutions to support academic reading through AI-powered summarisation tools. Two distinct approaches have emerged: (1) Structural AI summarisation, such as mind maps and bullet points, which we define as organising summarised content into structured formats that facilitate information processing and comprehension (Dagdelen et al., 2024; Liu et al., 2024a; Liu et al., 2024b). (2) Conversational AI summarisation, such as ChatGPT, can generate tailored summaries through an interactive dialogue with user inputs (Skjuve et al., 2023; Yang et al., n.d.). However, these approaches differ in their technical requirements and interaction patterns, leaving uncertainty about when and how students should employ each approach for optimal academic reading support.

As AI-assisted reading tools continue to develop, their educational impact requires further research. Previous research has highlighted benefits such as enhanced literature comprehension (Gu et al., 2024; Wang et al., 2024), but raised concerns, including omitting key information (Gu et al., 2024; Salleh, 2023) and

CONTACT Muhan Xu m.xu0920221@arts.ac.uk

hallucination that generates plausible yet unfaithful content (Farquhar et al., 2024; Huang et al., 2025; Ji et al., 2023). Recent neurocognitive research (Kosmyna et al., 2025) found that while LLM assistance initially reduces cognitive load, extended use over four months results in weaker learning outcomes compared to brain-only conditions across neural, linguistic, and scoring dimensions. Additionally, educators have expressed concerns that students increasingly seek quick solutions over deep thinking, with AI tools potentially amplifying superficial learning approaches (Malik et al., 2025). Effective use of AI-assisted tools requires prompt engineering skills that demand learning and experience (Liu et al., 2023; Wang et al., 2024; Zamfirescu-Pereira et al., 2023). Subramonyam et al. (Subramonyam et al., 2024) term “the gulf of envisioning” as the cognitive distance between user intentions and effective prompt formulation, encompassing capability, instruction, and intentionality gaps.

Given the potential benefits and concerns of LLMs, understanding how different AI-assisted approaches impact academic reading can be important. Therefore, our study examined how structural and conversational AI summarisation impact academic reading experiences among research students and suggested design opportunities for future development. We conducted a within-subject user study ($n = 29$) comparing three conditions: (1) reading with conversational AI summarisation, (2) structural AI summarisation, and (3) without AI summarisation.

Our results indicate that structural AI summarisation helps navigate text structure while reducing cognitive load, whereas conversational AI summarisation provides more flexible support. However, improved user experience did not translate to better categorisation accuracy, and concerns emerged, including over-reliance on AI outputs and time-consuming verification processes. We suggest four design opportunities for future AI-assisted reading tools based on these findings: (1) combining structural overviews with conversational flexibility, (2) aligning summarisation with reading objectives through intent clarification, (3) embedding verification features that maintain reading flow, and (4) shifting from direct answer provision to guided questioning that promotes critical engagement. Our findings contribute empirical evidence for comparing structural and conversational AI summarisation approaches, and suggest design opportunities that promote interactive, personalised, and critically engaged academic reading.

2. Related work

2.1. Academic reading

Academic reading involves understanding and critically evaluating scientific content (Le & Huynh, 2022; Rocha et al., 2021). Understanding academic literature requires interpreting the text’s explicit and underlying meanings to form a mental representation (Ahmadi et al., 2013; Anne Britt et al., 2014; Peregoy & Boyle, 2000). It is a complex process that is influenced by text structure (Anne Britt et al., 2014; Kendeou & Van Den Broek, 2007; Peregoy & Boyle, 2000), cognitive load (DeStefano & LeFevre, 2007; Kirkland & Saunders, 1991; Zumbach & Mohraz, 2008), long and complex sentences (Dubay, 2004), background knowledge (Ahmadi et al., 2013; Cromley & Azevedo, 2007; Yi, 2014), writing styles (Biber & Gray, 2010; Hayot, 2014), and task difficulties (Bjork & Bjork, 2011) and goals (Rayner et al., 2016). Previous research has highlighted the need for more easy-to-understand summaries tailored to different English reading skills (Wan et al., 2010).

2.2. AI-assisted reading tools

LLMs have created new opportunities for AI-assisted reading tools through their capabilities in text summarisation (August et al., 2023; Aydın & Karaarslan, 2022; Gu et al., 2024; Wang et al., 2024) and concept explanation (Kohnke et al., 2023). Meta-analysis of 44 experimental studies (Wang & Fan, 2025) showed ChatGPT has large positive impacts on learning performance and moderate effects on learning perception and higher-order thinking. Recent research has demonstrated LLMs’ potential and limitations across various academic reading tasks, including literature reviews (Aydın & Karaarslan, 2022; Van Dis et al., 2023; Wang et al., 2024) and reading comprehension (Gu et al., 2024; Wang et al., 2024). Wang et al. (Wang et al., 2024) found that LLMs can enhance young scholars’ literature understanding, but cautioned that young scholars might overlook ethical risks, recommending limitation-based training.

However, AI-assisted reading tools can introduce errors, such as information omission that excludes crucial details (Gu et al., 2024; Salleh, 2023) and hallucinated that generate plausible yet factually incorrect claims that are not supported by source content (Farquhar et al., 2024; Huang et al., 2025; Ji et al., 2023). Huang et al. (Huang et al., 2025) categorise LLM hallucinations into factuality errors for factual contradiction and factual fabrication, and faithfulness failures covering interaction, context, and logical inconsistencies. Additionally, prompt engineering is another challenge that requires learning and experience for effective use (Liu et al., 2023; Wang et al., 2024; Zamfirescu-Pereira et al., 2023). Lee and Palmer (Lee & Palmer, 2025) indicated that prompt engineering currently lacks standardised evaluation metrics for success and transferable prompt engineering frameworks. Subramonyam et al. (Subramonyam et al., 2024) theorised users' cognitive challenge in translating goals into effective prompts as "the gulf of envisioning", comprising three gaps: capability (what tasks LLMs perform), instruction (expressing intentions), and intentionality (expectations and output evaluation).

AI-assisted tools' educational effectiveness depends on the implementation approach (Weidlich et al., 2025). AI-assisted tools can function as cognitive scaffolding (Wood et al., 1976), a temporary instructional support that enables learners to perform tasks beyond their independent capabilities (Dhillon et al., 2024). However, effective scaffolding within learners' Zone of Proximal Development (Cole et al., 1978) requires appropriate support levels that challenge without overwhelming, and gradual fading as competence develops. This theoretical framework helps explain why recent computer-assisted learning research (Weidlich et al., 2025) found that ChatGPT can either support learning through productive tutoring interactions or harm learning by enabling cognitive offloading that reduces engagement with essential learning tasks. The ease of obtaining LLM answers may compromise germane cognitive load necessary for deep learning processes, suggesting reduced cognitive effort does not translate to enhanced learning outcomes (Stadler et al., 2024). Recent neurocognitive research (Kosmyna et al., 2025) supports this concern: LLM users showed weaker neural connectivity, lower essay ownership, and reduced performance compared to brain-only groups across neural, linguistic, and scoring measures during essay writing tasks.

Therefore, there is growing interest (Long et al., 2023) in developing "AI literacy" to encourage critical AI use (Long & Magerko, 2020; Malik et al., 2025; Park & Ahn, 2024; Walter, 2024; Wen et al., 2025; Yeh, 2025). Higher education faculty suggest that the lack of AI literacy and institutional unpreparedness for ethics, privacy, and assessment represent key implementation challenges (Mah & Groß, 2024). Prior research has addressed these limitations through system transparency (Kizilcec, 2016; Wu et al., 2022) with features including confidence scores (Xiong et al., 2024; Xu et al., 2024) and context-rich explanations (Zavolokina et al., 2024), which facilitate users to verify AI responses (Yun et al., 2025) and assess output reliability. Gu et al. (Gu et al., 2024) proposed a text-rendering approach that de-emphasises text, which supports users in recovering from AI summarisation errors, including information omission. However, its adaptability across various document types requires further investigation. These challenges underscore the need for careful implementation of AI-assisted reading tools in academic settings.

2.3. Structural and conversational AI summarisation

As AI-assisted reading tools evolve, conversational and structural summarisation are emerging as key approaches. Conversational AI can generate tailored summaries through an interactive, query-based process (Skjuve et al., 2023; Yang et al., n.d.). Tools such as ChatGPT enable users to interact with documents using natural language queries, facilitating efficient information extraction and comprehension (Yang et al., 2024). The flexibility of conversational systems can support various academic tasks, including writing (Lee et al., 2024) and experiment design (Dowling & Lucey, 2023).

Structural AI summarisation organises summarised content into structured formats that facilitate information processing and has gathered research attention. Liu et al. (Liu et al., 2024a) identify that users require LLM output constraints at two distinct levels: low-level constraints that follow structured output formats and appropriate length, and high-level constraints that enforce semantic and stylistic guidelines while preventing hallucination. Prior research further demonstrated that different structural formats serve distinct purposes: mind maps support intuitive understanding of complex information (Chaudhri et al., 2013), while bullet points facilitate efficient content selection and planning (Radensky et al., 2024). Users express preferences for clearly structured summaries with bulleted information (Ellen et al., 2014), and structured formats such as lists and flowcharts are increasingly important for LLM integration in both human comprehension and software development (Liu et al., 2024b).

Prior research implemented structural approaches with varied strategies. Addressing the limitations of linear conversational structures that produce long-winded responses, Graphologue (Jiang et al., 2023) transforms LLM outputs into interactive graphical diagrams, enabling non-linear exploration and flexible comprehension of information. Similarly, Suh et al. (Suh et al., 2023) developed interfaces combining canvas and hierarchy views that support multilevel exploration by enabling users to navigate and organise information across different abstraction levels. Dagdelen et al. (Dagdelen et al., 2024) further demonstrate that LLMs can extract structured knowledge from scientific texts, capturing hierarchical entity relationships in simple sentences or structured formats. However, LLMs face challenges in processing large amounts of structured documents (Chen et al. 2024; Edge et al. 2024) and retrieving relevant information accurately (Gao et al., 2024). Given the potential and challenges of conversational and structural AI summarisation, it is valuable to investigate how these approaches differently impact students' academic reading experiences.

3. Methodology

We conducted a user study with 29 students to understand the benefits, concerns, and design opportunities of AI-assisted academic reading tools. Participants completed paper categorisation tasks under study under three conditions: conversational AI summarisation (ChatGPT), structural AI summarisation (Linfo AI), and without AI summarisation (Chrome PDF reader). Quantitative data measured user experience, cognitive load, tool preferences, reading understanding, comprehension, and speed. Qualitative data were collected through think-aloud protocols and interviews to understand participants' internal thinking and suggestions on tool design.

3.1. Conditions

Our user study compared three conditions: conversational AI summarisation (ChatGPT), structural AI summarisation (Linfo AI), and a No-AI baseline (Chrome PDF reader). We selected ChatGPT (GPT-4o) for its large user base, with 2.6 billion monthly visits (Similarweb, 2024). Linfo AI Chrome plugin (v0.1.9) was chosen because it provides structural summarisation without conversational features, enabling clear differentiation between approaches (see Appendix A for configuration details). Linfo AI (Figure 1) includes three features: a Snapshot providing 50-word summaries, an outline/mind-map view for structural summarisation, and a click-to-source feature locating summaries with the original text (linfo.ai, 2024). Table 1 details the supported user actions across both AI tools. The No-AI baseline used the built-in Chrome PDF reader, providing basic reading features, including text search and zooming. To ensure experimental consistency, participants used a pre-configured test computer with pre-uploaded texts, were restricted to the three designated tools, and conducted all interactions in English only.

The two AI conditions involved different LLMs (GPT-4o for ChatGPT versus Claude 3 Haiku for Linfo AI), reflecting developers' deliberate model choices aligned with each interface's intended use. To quantify semantic differences between model outputs, we applied SBERT (all-mpnet-base-v2) to encode both models' responses to identical Linfo structural summarisation prompts, computing semantic similarity through cosine similarity of contextual embeddings. Cross-model analysis (Table 2) showed initial evidence of high semantic similarity ($M = 0.93$, SD = 0.02) despite GPT-4o producing 48% longer responses on average. Baseline comparisons (Table 3) revealed that within-model similarities across different papers were substantially lower (GPT: $M = 0.49$, SD = 0.09; Claude: $M = 0.43$, SD = 0.09), demonstrating that the observed cross-model similarity ($M = 0.93$) represents meaningful semantic similarity rather than typical similarity between academic summaries. This pattern suggests that our LLM choice primarily affects presentation style rather than core content quality, with both models extracting equivalent semantic information from source materials. Therefore, the user performance differences may reflect interface design rather than underlying model capabilities, supporting our research focus on interaction approaches over raw model performance.

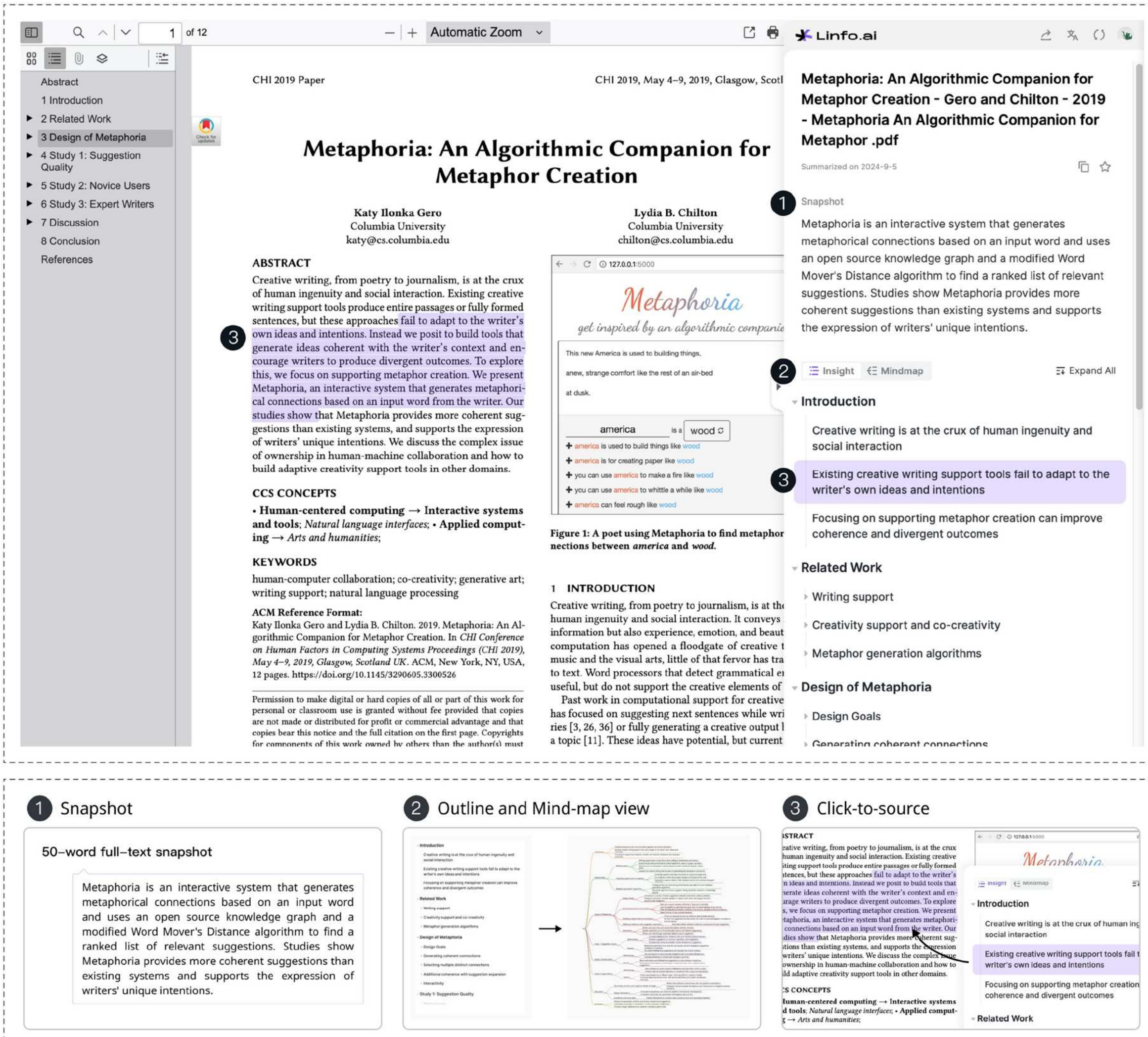

Figure 1. Linfo AI Chrome plugin appears as a side panel next to the paper. Including three features: (1) a 50-word snapshot summary, (2) an outline/mind-map view summary, and (3) a click-to-source feature for locating relevant text.

Table 1. Comparison of supported user actions between Linfo AI and ChatGPT, showing key interaction differences between structural and conversational AI summarisation approaches.

Supported User Action	Linfo AI	ChatGPT
Summary Acquisition	Automatic upon document load	User-initiated through natural language prompts
Summary Content	50-word snapshots, bullet points, and mind maps	Tailored responses based on user queries
Summary Structure	Structured mind map and bullet point	Paragraph-based conversational text
Summary Verification	Click-to-source linking for direct verification	Required manual cross-referencing
Summary Depth	Predetermined template-defined depth levels	User-controlled through iterative prompting
Follow-up Questioning	Static presentation, no conversational capability	Multi-round dialogue support

3.2. Tasks

Our user study involved paper categorisation tasks under three conditions, informed by two pilot studies. The first pilot study used GRE reading tasks but revealed two limitations. Participants achieved fully correct scores (10/10) by copying questions directly into ChatGPT, and the 400-word passages did not reflect typical academic paper lengths (5000–12,000 words). To address these issues, the second pilot introduced paper categorisation tasks using longer academic texts. When participants attempted to copy questions into ChatGPT during the second pilot, their average scores dropped (4.67/10), indicating that longer texts reduced the effectiveness of simple copy–paste strategies. The final task design, which used paper

Table 2. Comparison of GPT-4o and Claude 3 Haiku structural summarisation outputs. Word counts comparing each model's responses to identical structural summarisation prompts across six papers from the user study. Semantic similarity metric measured by the SBERT model. Longer responses in bold.

Paper	GPT Word Count	Claude Word Count	Semantic Similarity
Paper 1	**766**	506	0.892
Paper 2	**1136**	615	0.952
Paper 3	**1174**	605	0.930
Paper 4	460	**638**	0.937
Paper 5	**737**	587	0.951
Paper 6	**773**	458	0.943

Table 3. Semantic similarity baselines for model output comparisons. Within-model semantic similarities compare the same model's outputs across different papers. While cross-model semantic similarities are compared between both model outputs using the same papers. Semantic similarity metric measured by the SBERT model. Best responses in bold.

Model (s)	Mean	SD	Median	Min	Max	Pairs
GPT-4o	0.492	0.088	0.499	0.313	0.611	15
Claude 3 Haiku	0.426	0.086	0.434	0.253	0.557	15
GPT x Claude	**0.934**	**0.020**	**0.940**	**0.892**	**0.952**	6

categorisation with long academic texts, ensured fair comparisons across conditions and encouraged participants to perform gist categorisation supported by AI tools within time constraints.

Participants completed paper categorisation tasks that reflect the coding process in systematic literature reviews, where researchers assess paper relevance and assign categorical codes. In each condition, participants read two academic papers and selected one or two categories from four available options within a 10-minute time limit (Figure 2). We used multiple-choice categories to reflect that academic papers often cover multiple topics. This 10-minute constraint served two purposes: reflecting common academic time constraints (Carson et al., 2013; Yi, 2014) and controlling total session duration to reduce participant fatigue in our three-condition within-subjects design.

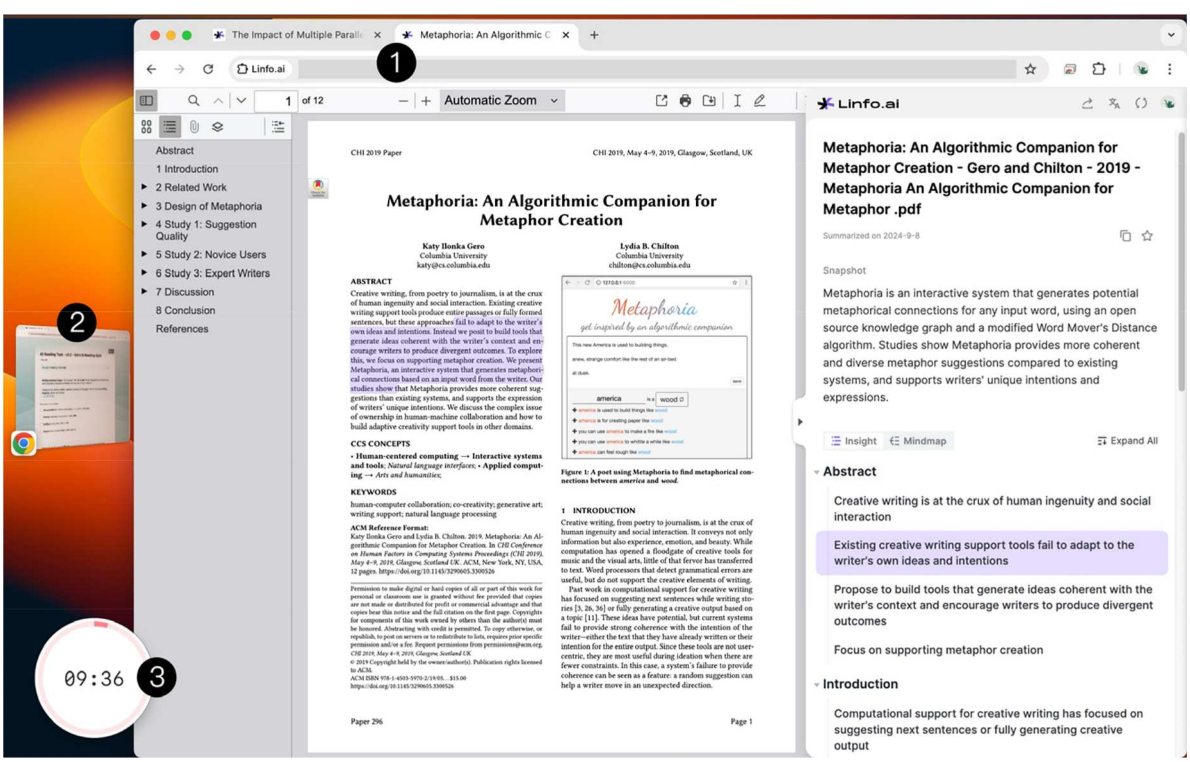

Figure 2. A screenshot of the desktop layout during user study. Including (1) a Chrome browser showing two papers, (2) a questionnaire for task answers, and (3) a 10-minute task timer.

We measured reading performance by comparing participant categorisations against prior systematic review results, scoring 20 points for correct categorisation and 10 points for partially correct responses. Papers were distributed to three conditions in a balanced order using a 3×3 Latin Square Design. We employed a within-subjects design, allowing participants to experience and compare all three conditions while controlling for individual differences.

We selected six papers from a systematic review paper (Lee et al., 2024) that annotated 115 writing assistant papers. Papers were chosen based on the Writing Stage dimension and its four codes: Drafting, Idea generation, Planning, and Revision. From 45 initially screened papers, we chose seven CHI papers (2019–2021) as CHI is the premier HCI conference relevant to our participants' field, ensuring content relevancy, research quality and consistent layout to control formatting variations across venues. We excluded one paper because its category was identifiable from the metadata. The final six papers ranged from 8500–14,000 words.

3.3. Procedure

Our user study consisted of six steps (Figure 3): (1) information consent and application questionnaire, (2) a pre-study questionnaire, (3) an introduction with an example task and quiz, (4) three paper categorisation tasks with post-task questionnaires and breaks, (5) a post-study questionnaire and break, and (6) a 30-minute post-study interview. Participants completed all tasks remotely through a pre-configured MacBook Pro, with Microsoft Teams used for screen recording and transcription. The study lasted around 100 minutes and was conducted from June to July 2024 with University Research Ethics Committee approval.

Participants categorised two papers within a 10-minute limit for each condition: conversational, structural, and without AI summarisation (Appendix B). To ensure adequate understanding, participants received an introduction that included procedure walkthroughs, example tasks and quizzes, and introductions to tool features. Participants were encouraged to think aloud (Nielsen, 1994), verbalising their thought processes during tasks. After each condition, participants completed post-task questionnaires measuring user experience through UEQ-S (Schrepp et al., 2017), cognitive load through NASA-TLX (Hart & Staveland, 1988), and reading understanding through Modified Bloom's Taxonomy (MBT) (Appendix C). We developed a four-item, 5-point Likert scale based on Bloom's Taxonomy's "Understanding" category (Anderson & Krathwohl, 2001), focusing on Exemplifying, Classifying, Summarising, and Comparing. As participants primarily categorised rather than critically evaluated papers, we excluded three concepts that required deeper analytical processes than our time-constrained categorisation task demanded. Three excluded concepts are: Interpreting

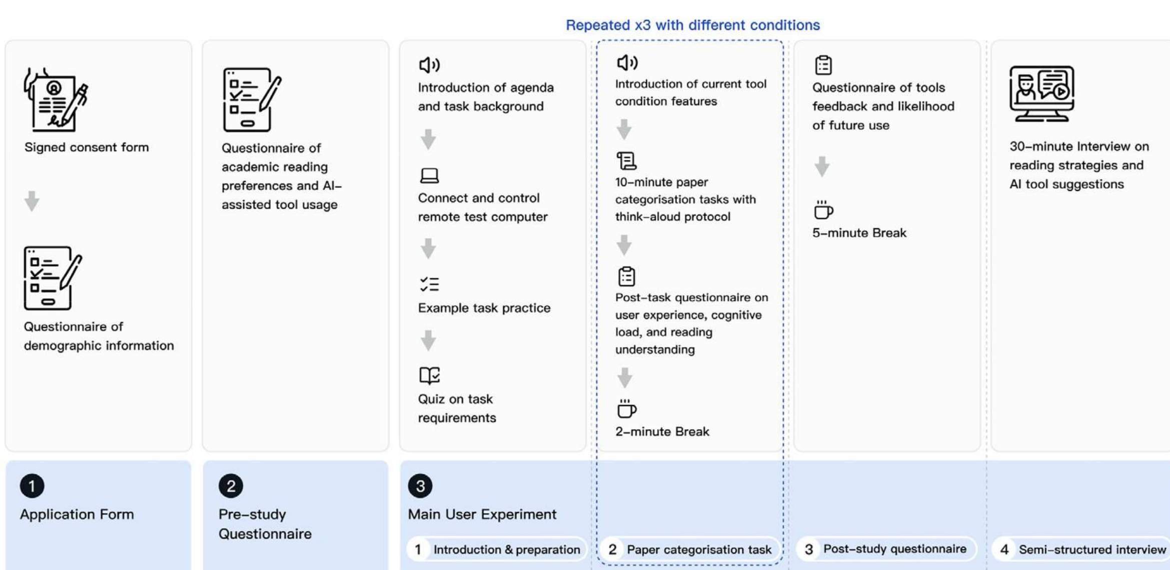

Figure 3. User study procedure in six parts and three stages, showing actions taken by participants.

(paraphrasing across representations), Inferring (drawing logical conclusions), and Explaining (constructing cause-and-effect models). However, our selection process should be considered an initial attempt requiring further validation. The full question statements are available in Appendix C.3.

After completing all tasks, participants completed a post-study questionnaire (Appendix D). Participants ranked the tool's helpfulness for task completion, provided feedback on tool features, and indicated their likelihood of future use. Finally, we conducted a 30-minute post-study semi-structured interview. We gathered qualitative feedback regarding participants' reactions to their task scores, reading strategies employed across three conditions, and suggestions for future tool improvement.

3.4. Participants

The user study included 29 valid responses (Table 4). Participants were recruited through the university Slack channel and WeChat groups. Recruitment was limited to HCI graduates and PhD students, ensuring familiarity with HCI research papers used in the paper categorisation task and reducing domain knowledge bias. All participants attended universities where English is the primary language and self-rated their English reading proficiency above the "Beginner" level. We manually examined all responses and excluded one participant for suspicious response patterns. Participants received a 20 GBP Amazon gift card as compensation.

3.5. Data analysis

We used a mixed-methods approach to analyse AI-assisted academic reading experiences. Quantitative data, where applicable, were analysed with repeated measures ANOVA, with post-hoc pairwise comparisons using Tukey's test. Post-hoc power analysis was conducted using G*Power 3.1.9.6 to assess whether near-significant results reflected inadequate statistical power or genuine marginal effects. For qualitative data analysis, we employed open coding (Thomas, 2003; Williams & Moser, 2019), which involved reviewing data without predefined categories. The coding process involved four steps. Initially, the first four authors independently coded transcripts. Secondly, authors swapped and cross-reviewed, ensuring that at least two authors reviewed each transcript. Thirdly, we conducted multiple discussion sessions to refine the coding approach

Table 4. Demographics of the participants in the user study.

ID	Age	Gender	Current Student Status	English reading proficiency
P1	27	Female	PhD student	Native/Fluent
P2	27	Female	PhD student	Intermediate
P3	30	Female	PhD student	Intermediate
P4	26	Male	Graduate student	Native/Fluent
P5	30	Male	PhD student	Intermediate
P6	23	Male	PhD student	Native/Fluent
P7	27	Female	PhD student	Intermediate
P8	28	Male	PhD student	Advanced
P9	31	Female	PhD student	Advanced
P10	26	Female	Graduate student	Advanced
P11	26	Male	Graduate student	Intermediate
P12	26	Male	PhD student	Intermediate
P13	26	Male	PhD student	Advanced
P14	24	Male	Graduate student	Advanced
P15	25	Male	PhD student	Advanced
P16	28	Male	PhD student	Intermediate
P17	28	Female	PhD student	Advanced
P18	23	Female	Graduate student	Intermediate
P19	26	Female	Graduate student	Advanced
P20	26	Female	PhD student	Advanced
P21	27	Female	PhD student	Intermediate
P22	30	Female	PhD student	Advanced
P23	29	Female	PhD student	Intermediate
P24	22	Female	Graduate student	Advanced
P25	31	Male	PhD student	Advanced
P26	27	Male	PhD student	Advanced
P27	26	Non-binary	Graduate student	Native/Fluent
P28	23	Female	Graduate student	Advanced
P29	27	Female	PhD student	Intermediate

and categorised codes into themes and patterns. Finally, the first author categorised all the codes, after which the second author reviewed the final categorisation. Coding disagreements were discussed and resolved through consensus. Inter-rater reliability was assessed using Krippendorff's Alpha ($\alpha = 0.714$), indicating moderate reliability for exploratory qualitative research.

4. Results

This section presents our quantitative and qualitative results regarding AI-assisted reading tools' benefits, concerns, and design challenges from our user study ($n = 29$). We collected quantitative data on participants' user experience, cognitive load, tool preferences, reading understanding (Table 5), comprehension, and speed. Qualitative analysis of interviews and think-aloud protocols identified three categories (benefits, concerns, and design challenges) with 11 themes and 41 sub-themes (Figure 4), derived from 1195 initial codes.

4.1. Benefits of AI-assisted reading tools

4.1.1. User experience differences

User experience was assessed using the UEQ-S scale across all three conditions (Figure 5a). Mauchly's test confirmed that the assumption of sphericity was satisfied ($W = 0.98$, $\chi^2(2) = 0.54$, $p = 0.76$), eliminating the need for degrees of freedom corrections. The analysis revealed differences in user experience ratings across conditions ($F(2,56) = 53.78$, $p < 0.001$, $\eta_p^2 = 0.66$, $\omega^2 = 0.64$). According to Cohen (2013), both effect size measures, partial eta squared ($\eta_p^2 = 0.66$) and the more conservative omega squared ($\omega^2 = 0.64$), exceed the 0.14 threshold for large effects. This large effect size, accounting for approximately 64-66% of variance in user experience scores, suggests that AI-assisted conditions have a substantial impact on how students perceive and engage with reading support tools. Descriptively, Linfo AI received the highest mean ratings ($M = 1.16$, SD = 0.81), followed by ChatGPT ($M = 0.67$, SD = 0.80), while the No-AI baseline scored lowest ($M = -1.00$, SD = 0.85). Post-hoc pairwise comparisons with Holm–Bonferroni correction revealed significant differences between the AI conditions and the baseline. Both AI conditions significantly outperformed the No-AI baseline (ChatGPT: mean difference = −1.67, $p < 0.001$; Linfo AI: mean difference = −2.15, $p < 0.001$), representing meaningful improvements of 1.67 and 2.15 scale points, respectively. The comparison between ChatGPT and Linfo AI did not reach statistical significance after correction (mean difference = −0.48, $p = 0.086$).

NASA-TLX results (Figure 5b) examined the impact of AI assistance on perceived cognitive workload. Mauchly's test indicated sphericity was met ($W = 1.00$, $\chi^2(2) = 0.12$, $p = 0.940$). ANOVA results showed differences in perceived workload across conditions ($F(2,56) = 11.06$, $p < 0.001$, $\eta_p^2 = 0.28$, $\omega^2 = 0.25$), with a large effect size. Post-hoc comparisons with Holm–Bonferroni correction showed that both AI conditions

Table 5. UEQ-S, NASA-TLX, and Modified Bloom's Taxonomy (MBT) statistic results. We calculated the Mean (*M*) and Standard Deviation (SD) for each scale and condition, along with the ANOVA results. The user experience measures with UEQ-S, cognitive load measures with NASA-TLX, and the cognitive process of "understanding" measures with MBT. "*LIB*" stands for "*Lower Is Bette*", with best results in bold.

Metric / scales	No-AI (SD)	ChatGPT (SD)	Linfo AI (SD)	*F*	*P*	η_p^2
UEQ-S: Pragmatic Quality	-0.65 (0.98)	0.89 (1.05)	**1.22 (1.12)**	*F*(2,56)=23.20	<0.001***	0.45
UEQ-S: Hedonic Quality	-1.35 (1.08)	0.46 (0.83)	**1.09 (0.71)**	*F*(2,56)=63.24	<0.001***	0.69
UEQ-S: Overall	-1.00 (0.85)	0.67 (0.80)	**1.16 (0.81)**	*F*(2,56)=53.78	<0.001***	0.66
NASA-TLX: Mental (LIB)	3.86 (0.79)	3.10 (0.86)	**3.00 (1.04)**	*F*(2,56)=10.24	<0.001***	0.27
NASA-TLX: Physical (LIB)	2.83 (1.10)	2.34 (0.97)	**2.10 (1.01)**	*F*(2,56)=6.47	0.003**	0.19
NASA-TLX: Time (LIB)	3.55 (0.99)	3.04 (0.92)	**2.86 (1.19)**	*F*(2,54)=3.10	0.053	0.10
NASA-TLX: Performance (LIB)	3.34 (0.81)	2.86 (0.83)	**2.83 (0.89)**	*F*(2,56)=4.66	0.013*	0.14
NASA-TLX: Load (LIB)	3.79 (0.77)	**2.90 (0.98)**	2.96 (0.88)	*F*(2,54)=11.71	<0.001***	0.30
NASA-TLX: Frustration (LIB)	2.86 (0.99)	2.72 (1.16)	**2.45 (1.06)**	*F*(2,56)=1.56	0.218	0.05
NASA-TLX: Overall (LIB)	3.37 (0.58)	2.83 (0.63)	**2.70 (0.71)**	*F*(2,56)=11.06	<0.001***	0.28
MBT: Exemplifying	3.17 (0.85)	**3.72 (0. 92)**	3.45 (0.95)	*F*(2,56)=3.83	0.028*	0.12
MBT: Classifying	3.00 (1.07)	3.48 (0.83)	**3.55 (0.99)**	*F*(2,56)=3.66	0.032*	0.12
MBT: Summarising	2.52 (0.87)	3.10 (1.05)	**3.52 (0.91)**	*F*(2,56)=8.66	<0.001***	0.24
MBT: Comparing	2.72 (0.84)	3.03 (0.94)	**3.34 (0.90)**	*F*(2,56)=3.58	0.035*	0.11
MBT: Overall	2.85 (0.74)	3.34 (0.69)	**3.47 (0.68)**	*F*(2,56)=7.51	0.002**	0.21

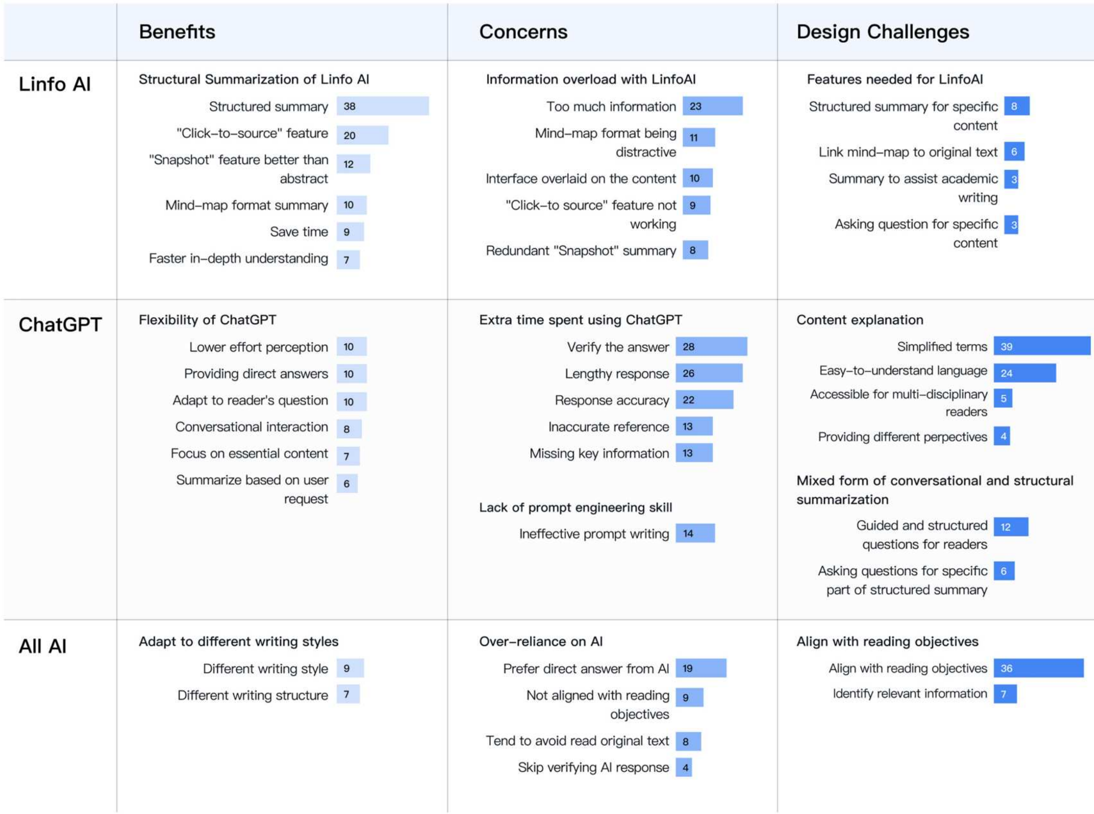

Figure 4. Number of occurrences of each code applied to the user study transcript. Codes are organised into 11 themes (boldface titles of groups) and 41 sub-themes, which in turn fall into three high-level categories (Benefits, Concerns, and Design Opportunities).

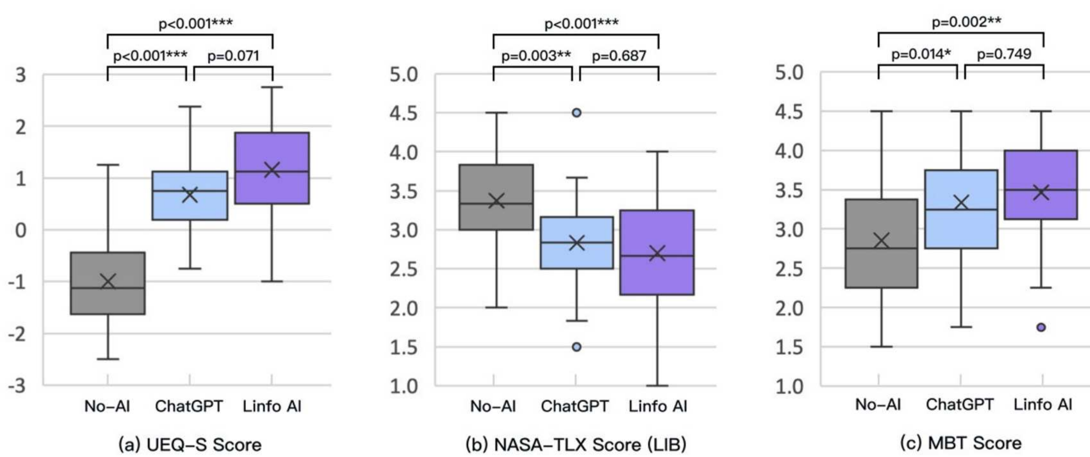

Figure 5. Box plots comparing participant-level mean scores of UEQ-S (a), NASA-TLX (b), and MBT (c) for No-AI baseline, ChatGPT, and Linfo AI conditions. Significant differences ($p < 0.050$*, $p < 0.01$**, $p < 0.001$***) from Posthoc Tukey's tests are indicated.

significantly reduced perceived workload compared to the No-AI baseline (ChatGPT: mean difference = 0.54, $p = 0.004$; Linfo AI: mean difference = 0.67, $p < 0.001$). The difference between the two AI conditions did not reach significance after correction (mean difference = 0.13, $p = 1.000$). Furthermore, post-hoc power analysis for the temporal demand subscale showed adequate statistical power of 0.977 (1-β err prob), indicating that the near-significant result ($p = 0.053$) reflected a marginal effect size rather than insufficient sample size.

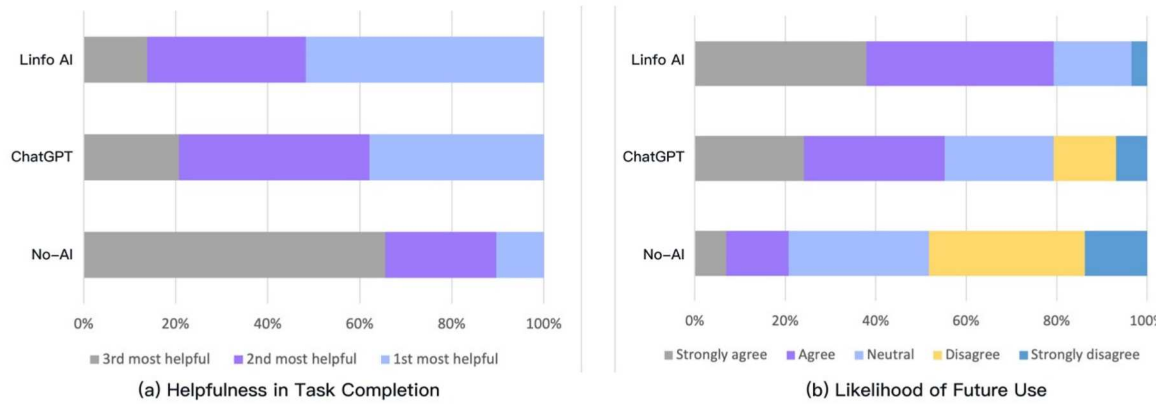

Figure 6. Perceived helpfulness for task completion (a) and likelihood of future use (b) across No-AI baseline, ChatGPT, and Linfo AI conditions.

Participants showed distinct patterns in both perceived helpfulness and likelihood of future use across the three tool conditions. For perceived helpfulness in task completion (Figure 6a), 15 of 29 participants ranked Linfo AI highest. For likelihood of future use (Figure 6b), Mauchly's test indicated borderline satisfaction of the sphericity assumption ($W = 0.83$, $\chi^2(2) = 4.99$, $p = 0.083$), so Greenhouse-Geisser corrections were applied as a conservative approach ($\varepsilon = 0.856$). The corrected analysis showed a significant main effect of tool conditions ($F(1.711, 47.920) = 13.63$, $p < 0.001$, $\eta_p^2 = 0.33$, $\omega^2 = 0.30$). Post-hoc comparisons using Holm–Bonferroni showed only the strongest effects survived correction: No-AI versus Linfo AI (mean difference = −1.448, $p < 0.001$). While No-AI versus ChatGPT ($p = 0.041$) and ChatGPT versus Linfo AI ($p = 0.050$) did not reach significance after correction, post-hoc power analyses confirmed adequate statistical power of 0.72 and 0.69 (1-β err prob), respectively, indicating these reflect moderate effects rather than insufficient sample size.

4.1.2. Reading understanding

We applied Modified Bloom's Taxonomy (MBT) to measure participants' understanding processes across conditions. Mauchly's test confirmed sphericity ($W = 0.98$, $\chi^2(2) = 0.52$, $p = 0.772$), requiring no corrections. The ANOVA analysis showed significant differences in perceived understanding processes across conditions ($F(2,56) = 7.51$, $p < 0.001$, $\eta_p^2 = 0.21$, $\omega^2 = 0.18$), indicating a large effect size. Internal consistency analysis showed condition-dependent reliability: No-AI baseline demonstrated good reliability (Cronbach's $\alpha = 0.82$, McDonald's $\omega = 0.83$), while both AI conditions showed acceptable reliability (ChatGPT: $\alpha = 0.72$, $\omega = 0.71$; Linfo AI: $\alpha = 0.71$, $\omega = 0.72$). Post-hoc pairwise comparisons with Holm–Bonferroni correction showed that only the comparison between No-AI and Linfo AI survived the stringent correction (mean difference = −0.61, $p = 0.002$). The comparison between ChatGPT and No-AI baseline did not reach statistical significance after correction (mean difference = 0.49, $p = 0.016$). Results further showed ChatGPT excelled in Exemplifying ($M = 3.72$, SD = 0.16), while Linfo AI performed best in Summarising ($M = 3.52$, SD = 0.17) and Comparing ($M = 3.34$, SD = 0.17).

All conditions achieved similar paper categorisation scores and reading speed. Paper categorisation scores (Figure 7a) showed minimal differences across conditions. Linfo AI achieved the highest mean score ($M = 19.66$, SD = 11.17), followed by No-AI baseline ($M = 17.93$, SD = 12.07) and ChatGPT ($M = 17.59$, SD = 9.12). Task completion times (Figure 7b) averaged 9.71 minutes (SD = 1.55) for all conditions with no significant differences.

4.1.3. Structural summarisation in Linfo AI

The benefits were twofold. Firstly, 17/29 participants reported that structural AI summarisation facilitates their understanding of text structures and relationships. P4 mentioned that Linfo AI clarifies "*the logical relationship within the paper*," while P15 noted it helps "*extract relationships between concepts*". Secondly, structural summarisation helped nine participants locate information efficiently, enabled step-by-step comprehension (P9, 11, 16), and accelerated thinking for two others (P13, 24). As P24 explained: "*Linfo AI lists the*

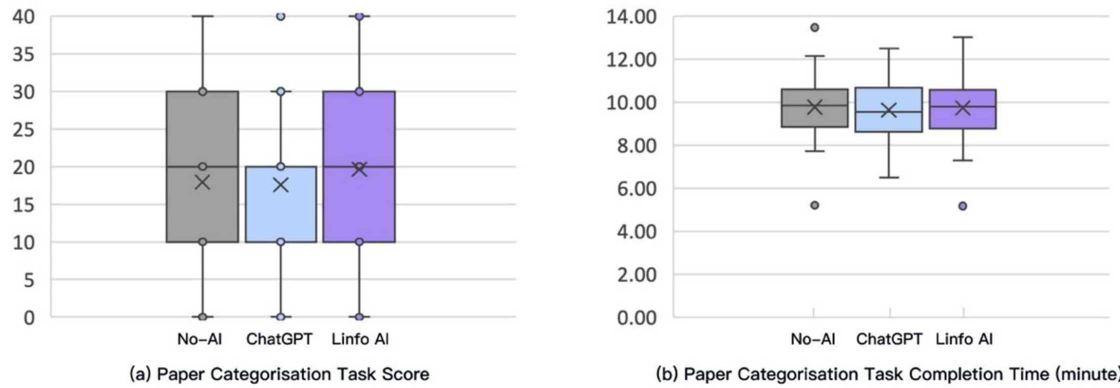

Figure 7. Scores (a) and completion times (b) of the paper categorisation task, comparing across three conditions.

structure of the paper […] lets me know where to look for things. Otherwise, […] would be time-consuming." Eight participants reported that reading with Linfo AI was similar to the No-AI baseline, relying on text structure (P11, 17, 19) and keywords (P26). However, P16 expressed a preference for manually developing mind maps for comprehension, stating, "*If mind maps are generated all at once, I don't think it would be beneficial.*"

4.1.4. Click-to-source verification in Linfo AI

Nine participants expressed concerns about errors, distortion, and oversimplification in AI outputs. Thirteen participants found Linfo AI's click-to-source feature helpful for locating and verifying key information by navigating directly to the source text. As P15 noted: "*It gives me instant access to the corresponding text, which is very helpful […] I can understand it in context.*" However, four participants (P17, 18, 21, 22) reported that Linfo AI sometimes failed to highlight the correct relevant text, limiting the feature's reliability.

4.1.5. Flexibility of ChatGPT

While three participants (P5, 10, 21) found Linfo AI less adaptable, six participants preferred ChatGPT's flexibility (P5, 6, 11, 16, 25, 17). Eight participants valued ChatGPT's ability to provide direct answers, customise responses (P5, 17, 25), ask follow-up questions (P5), and constrain summaries (P17). As P5 noted, "*ChatGPT is responsive […] I can ask more in-depth questions and explain it.*" Screen recordings showed participants used ChatGPT for summaries (18/29), explanations (14/29), paper categorisation (12/29), and role-playing as a researcher (3/29). However, participants encountered challenges with prompt writing (P13, 19, 25, 27), found it time-consuming (P1, 12, 19, 28), and required repeated clarifications (P14).

4.2. Concerns about AI-assisted reading tools

4.2.1. Time required to verify AI responses

We observed varied attitudes towards verifying AI responses, ranging from reliance to scepticism. Notably, 12 of 29 participants copied task questions directly into ChatGPT for immediate answers. Screen recordings revealed that nine participants skipped verifying ChatGPT responses entirely. P14 described using ChatGPT as "*an easy way of reading*", while P1 and P9 reported reduced motivation to engage with the source text after receiving ChatGPT's answers. Meanwhile, ChatGPT condition showed the lowest average categorisation scores (Figure 7a). Nine participants expressed concerns about errors, distortion, and oversimplification in AI outputs. P12 stressed that even minor AI errors could undermine trust and invalidate prior results.

In contrast, P15 achieved fully correct categorisation scores by treating ChatGPT interactions as "*peer review*," actively challenging the AI for robust evidence and clearer claims. However, four participants (P8, 12, 25, 28) reported verifying AI responses as more time-consuming than reading independently.

4.2.2. Information overload

While seventeen participants found Linfo AI's structural summarisation helpful, six others (P3, 16, 17, 21, 22, 29) reported information overload. P16 stated: "*The mind map […] has too much information […] defeats the* *purpose of the mind map*." In response to information overload, participants requested adjustable summarisation detail levels (P6, 17, 26) and expandable mind maps that "*grow*" as users read further in-depth (P15).

4.3. Design challenges of AI-assisted reading tools

4.3.1. Summarisation that aligns with reading objectives

Four participants found Linfo AI's predetermined summaries too general (P5, 7, 10, 21), with P10 suggesting AI should "*capture information based on provided research context*" for more relevant outputs. Twelve participants expressed needs for AI summaries to align with reading objectives, including identifying research gaps (P10, 25, 26), sourcing references for literature reviews (P9, 15, 16, 25, 28) and experiment design (P24, 25). As P25 stated, "*different (reading) objectives change how I use this tool*". Additionally, eight participants requested features for efficient information recall during writing tasks, including paper relevance indicators (P8, 10, 24) and quote location functionality (P21) based on their writing content. However, participants raised concerns about losing valuable information during the relevancy assessment (P20, 28) and privacy risks in understanding user intentions (P27).

4.3.2. Guidance during the reading process

Six participants (P5, 9, 10, 15, 20, 29) expressed interest in AI tools that guide them throughout reading, while five (P10, 13, 26, 28, 29) reported challenges in formulating questions for AI. P15 emphasised the need for reflective questioning to rationalise the reading process, explaining: "*I would like a supportive and constant questioning experience that keeps asking me why I am doing certain things.*" Participants suggested follow-up question features that provide "*useful reflections*" (P9) and support detailed reading (P10). P22 suggested reading assistance tailored to different reader roles, from helping new PhD students understand topics to guiding PhD candidates in literature searches.

5. Discussion

Our mixed-method user study on AI-assisted academic reading for students revealed benefits, concerns, and design opportunities. Our results indicated that structural summarisation facilitated navigating text structure, locating information, and lowering cognitive load. Conversational AI offers flexible support through direct answers and follow-up questions. We based our findings on suggesting design opportunities for AI-assisted academic reading tools.

5.1. Improved experience with maintained performance

Our study suggested that positive user experience may not translate to better performance in time-constrained paper categorisation tasks. Participants in AI-assisted conditions reported higher user experience scores and lower cognitive load compared to No-AI baselines. However, no significant differences were observed in reading speed or comprehension scores compared to No-AI baselines. This maintained performance aligns with prior cognitive studies showing that reduced cognitive load does not universally enhance learning (Kosmyna et al., 2025), as students shift from active critical reasoning to passive oversight, compromising deep understanding and schema development (Stadler et al., 2024). Meanwhile, known LLM limitations in hallucination (Farquhar et al., 2024; Huang et al., 2025; Ji et al., 2023) and information omission (Gu et al., 2024; Salleh, 2023) may affect comprehension despite positive user perceptions. Therefore, future research should balance user satisfaction with learning effectiveness, designing reading support that increases educational gains without degrading the positive user experience observed in this study.

5.2. Combining structural and conversational summarisation

Our results suggest distinct benefits and trade-offs between structural and conversational AI summarisation approaches. Structural summarisation was found helpful by 17 of 29 participants for understanding text structure and relationships, and quantitative data showed reduced cognitive load with a more positive user experience. However, six participants reported information overload from dense detail, and three found predetermined summaries were too general. Meanwhile, conversational summarisation provided flexibility valued by eight participants, but four struggled with prompt engineering, a skill that demands training and experience (Liu et al., 2023; Zamfirescu-Pereira et al., 2023). Recent research on LLMs implementation in higher education (Park & Ahn, 2024; Wang et al., 2024) further identifies multiple risks, including difficult prompt engineering and user expertise gaps, insufficient limitation-focused training, and trade-offs between response speed and accuracy, that can negatively impact student learning experiences.

Our analysis indicates that both approaches rely on LLMs but vary in user controls. Structural summarisation provided cognitive scaffolding (Dhillon et al., 2024; Wood et al., 1976) through expert-defined templates, reducing prompt engineering demands but producing predetermined outputs. Conversational AI enabled personalised interactions through user-constructed prompts, but demanded greater prompt engineering expertise (Liu et al., 2023; Zamfirescu-Pereira et al., 2023). This distinction aligns with scaffolding theory's emphasis on appropriate support within learners' Zone of Proximal Development (Cole et al., 1978): structural summarisation may provide less flexible support, while conversational approaches may provide insufficient guidance. Recent research on context-based interaction modes (Gao et al., 2024b) suggests providing predefined rules for specific tasks while adapting to user intent through inferring implicit queries. Therefore, the observed complementary patterns suggest potential value for future research investigating hybrid systems, balancing pre-defined templates consistency with conversational flexibility. For example, provide structural overviews through mind maps while enabling conversational interactions for specific sections. However, these design recommendations remain speculative, as our study examined these approaches separately. Future research should address managing multiple interaction modes while accommodating varying prompt engineering abilities among users.

5.3. Intent clarification for academic reading objectives

Our results suggest a mismatch between diverse user reading objectives and current AI tool capabilities for supporting intent clarification. 12 of 29 participants explicitly requested AI summaries aligned with specific reading objectives, including research gap identification, literature reviews, and experiment design. However, three participants found Linfo AI's summaries too generic for their specific needs, suggesting expert-designed templates may fail without understanding user intent. Additionally, four participants struggled with formulating prompts for ChatGPT, requiring repeated clarifications. These findings exemplify the "capability gap" within Subramonyam et al.'s "gulf of envisioning" framework (Subramonyam et al., 2024): users lack mental models of what tasks LLMs can perform and how to specify procedural intentions for varied reading objectives.

These intent clarification challenges suggest opportunities for domain-specific prompting strategies in academic reading tools. Domain-specific prompting strategies (Subramonyam et al., 2024) involve custom approaches that steer LLM outputs towards usable end goals while minimising ambiguity, allowing users to build task-specific mental models. Future research could implement semi-structured templates that guide users through intent clarification via open-ended questions before generating summaries, similar to conversational search systems that use clarifying questions to resolve query ambiguity (Wang et al., 2023). Additionally, proactive prompt suggestions (Park & Ahn, 2024; Subramonyam et al., 2024) could help users clarify reading objectives more precisely, enabling task-specific outputs. For example, prompts for objective summaries focused on evidence for literature review tasks or towards creative analysis exploring multiple perspectives for research ideation (Yun et al., 2025). However, future research should address challenges in accurately capturing readers' interests, context, and tasks (Gu et al., 2024; Zhang et al., 2025).

5.4. Embedded verification for AI content reliability

Our results suggest both the necessity and challenges of supporting the verification process in AI-assisted reading tools. Nine participants expressed concerns about errors and oversimplification in AI outputs, with P12 emphasising that even minor errors could undermine trust and invalidate prior results. These concerns align with known LLM limitations, including hallucinated references, where AI-generated citations may not align with actual source content (Shen et al., 2023). While thirteen participants found Linfo AI's click-to-source feature helpful for verifying information with source text, four participants reported that verifying AI responses was more time-consuming than reading independently. Critically, participants who actively verified AI outputs, such as P15, achieved fully correct categorisation scores, suggesting that verification processes could increase the performance. These findings suggest the challenge of designing verification mechanisms that maintain accuracy without disrupting reading workflows.

Based on our findings, we suggest embedded verification design recommendations. Prior research demonstrated that the values of system transparency (Kizilcec, 2016; Wu et al., 2022) and confidence scores (Xiong et al., 2024; Xu et al., 2024) could help users assess AI reliability. A recent study (Yun et al., 2025) observed participants build confidence by cross-checking outputs against multiple sources, suggesting the need for tracing AI-generated text back to its source evidence (Huang & Chang, 2024). We propose that future AI-assisted reading tools could explore designs that support reliability assessment without disrupting users' natural reading flow (Zavolokina et al., 2024). One possible approach might involve colour-coded click-to-source verification (Figure 8), where coloured underlines could indicate the reliability of AI-generated content relative to the source text. Then, users could click on the underlined text to instantly scroll to the source text for immediate verification. Such features would need to balance informativeness with cognitive load, as overly complex visual systems could potentially be too intrusive to be ignored or disabled by users (Zavolokina et al., 2024). Future research should empirically test whether embedded verification could improve performance without degrading user experience, including controlled studies comparing verification-support systems that could examine both comprehension scores and user satisfaction across different interface designs. Specific research directions include comparing verification visual cues, investigating optimal reliability thresholds for colour-coding, and examining how verification workflow integration affects user behaviours and comprehension performance in academic reading contexts.

5.5. From direct answers to guided questions

Participants showed concerning over-reliance patterns in AI verification. Twelve participants copied task questions directly into ChatGPT for immediate answers, while nine skipped verifications entirely. This pattern persisted despite pilot study modifications, where participants initially achieved fully correct scores (10/10) by copying questions into ChatGPT without adequate reading. In contrast, P15 achieved fully correct categorisation scores by actively challenging AI outputs for more robust evidence, rather than passively accepting them. This successful critical engagement strategy aligned with the interests of six participants in AI tools that guide them throughout reading, although five struggled with question formulation. Malik et al. (Malik et al., 2025) revealed educator concerns that students increasingly seek quick solutions over deep thinking, with AI tools potentially amplifying superficial learning. These concerns

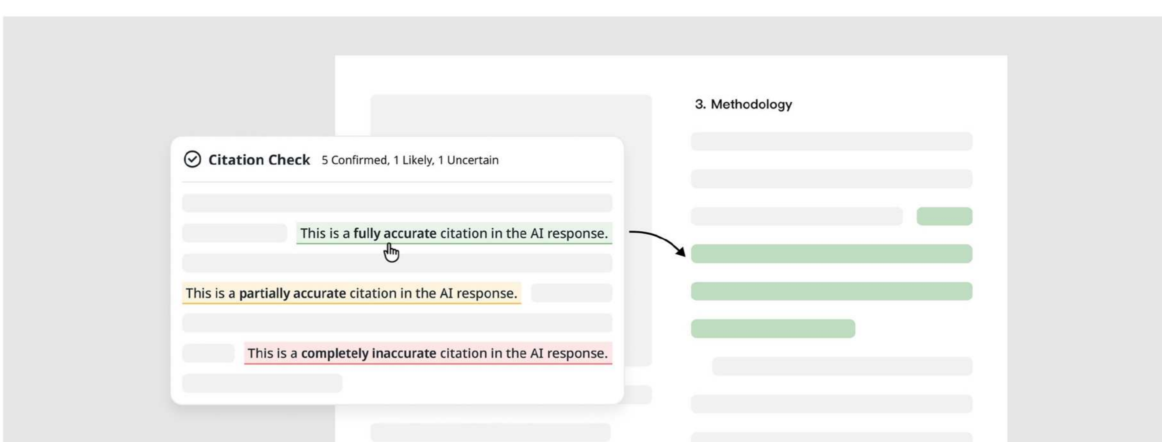

Figure 8. Mock-up interface for colour-coded click-to-source verification. AI-generated text is colour-underlined by confidence scores, while clicking any underline instantly scrolls to the source text, supporting quick and non-intrusive verification.

reflect rich research in developing "AI literacy" to encourage critical AI use (Long et al., 2023; Long & Magerko, 2020; Walter, 2024; Wen et al., 2025).

Question-guided interfaces emerge as a potential solution to address the over-reliance patterns and maintained performance observed in our study. Prior research (August et al., 2023; Chaudhri et al., 2013) demonstrated that AI-generated questions can support reading comprehension and information retention, with "intended inconveniences" (Park & Ahn, 2024) cultivating critical thinking through iterative trial-and-error processes. Meta-analysis evidence (Wang & Fan, 2025) shows ChatGPT functions effectively as an intelligent tutor for enhancing higher-order thinking skills. Research on integrating AI tools with inquiry-based learning shows potential for more interactive, student-centred learning environments that align AI capabilities with pedagogical demands (Yeh, 2025). Building on this evidence, we propose a question-guided interface (Figure 9) where users click highlighted text to receive AI-generated questions in margins, supported by optional explanations. However, empirical validation remains necessary to evaluate whether such interfaces could reduce over-reliance behaviours, foster critical reading, and ultimately improve comprehension scores compared to direct answer provision. Future research could further compare different questioning strategies, investigate optimal timing for question presentation, and examine how question complexity affects both engagement and comprehension in academic reading tasks.

5.6. Limitations

Five limitations should be noted. First, this study examined categorisation tasks within a 10-minute timeframe to control methodological variables and reduce participant fatigue in the within-subjects design. Academic reading typically takes longer than our 10-minute constraint, involving synthesis, evaluation, and integration processes not captured in category-selection tasks. The time constraint may have limited opportunities for iterative dialogue in the conversational AI condition. This research boundary indicates that the findings characterise gist extraction and quick assessment rather than comprehensive academic reading processes. Future work should extend experimental timeframes to investigate deeper reading processes such as inference-making, explanation, and evaluation.

Second, while semantic similarity testing (see Section 1) suggested comparable content extraction across two LLMs, our study compared two AI conditions using developers' default model choices. Future work should isolate interface effects by testing identical models across different systems. Third, remote testing lacked temporal controls for circadian rhythm effects on cognitive performance. Future work should standardise testing times.

Fourth, our study examined participants' self-reported reading understanding processes through a Modified Bloom's Taxonomy (MBT). As an initial measurement approach, MBT's psychometric properties require further validation. While MBT results showed a main effect across conditions, the self-reported nature of MBT captures perceived understanding processes rather than objective comprehension

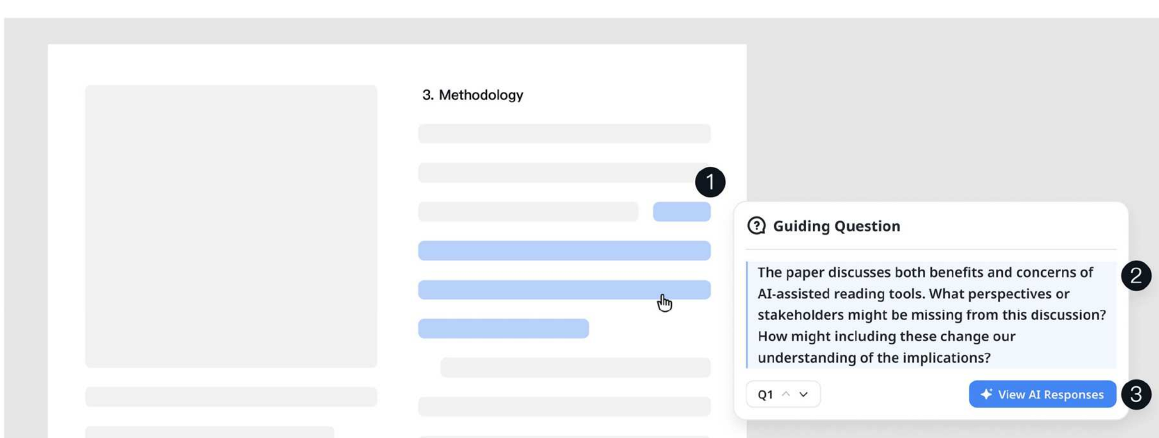

Figure 9. Mock-up interface for AI-guided reading questions. Clicking a highlighted sentence (1) reveals guided questions (2) for critical evaluation of the paper's claims and view AI response (3) for additional support.

performance. Future research should incorporate transfer tasks to assess whether perceived understanding reflects objective learning outcomes. Finally, our recruitment targeted HCI graduate students from English-medium universities (UK: $n = 21$, US: $n = 6$, Netherlands: $n = 1$, China: $n = 1$) through university channels, controlled for domain knowledge and technology familiarity. This targeted sampling defines a boundary for the findings, as this research does not address how tool effectiveness varies by AI literacy, language proficiency, disciplinary background, or academic experience level. Future research should recruit broader participant samples and include moderator analyses to improve generalisability and validity.

6. Conclusion

AI-assisted academic reading tools present both benefits and challenges for students. Our study suggests that structural summarisation improves text structure navigation, facilitates information location, and improves user experience while reducing cognitive load. Conversational AI provides flexible support through direct answers and follow-up questions. However, participants' varied attitudes towards verifying AI responses ranged from over-reliance to time-consuming verification. Notably, improved user experience with AI tools did not improve categorisation accuracy, suggesting positive experiences may not translate to educational gains in time-constrained academic tasks. These findings should be interpreted within the boundaries of 10-minute gist categorisation tasks, the use of different LLMs across conditions, and our HCI graduate student sample.

Based on our findings, we suggest four design opportunities for future AI-assisted reading tools. First, we propose hybrid approaches that combine structural overviews with conversational flexibility. Second, we propose bridging the intent clarification gap through guided user interaction for diverse reading objectives. Third, we recommend embedded verification features to resolve the tension between reliability assessment and reading flow disruption. Fourth, we suggest shifting from direct answer provision to guided questioning for addressing over-reliance concerns while fostering critical engagement. We hope our work inspires future designs of new interactions and promotes the cautious use of AI-assisted reading tools.

Acknowledgements

The authors thank Junyi Zeng, Xin Wang, and Tianxiao Wang for their support in data analysis. They also want to thank all participants for their valuable time and all reviewers for their constructive feedback. Claude 4.0 Sonnet was used during manuscript preparation to enhance the clarity and fluency of the text. All content was reviewed and verified by authors for accuracy and integrity. M.X. conceptualised the study. M.G., N.B., and R.F. provided supervision and guidance. M.X., T.W., S.S., and J.S. designed the methodology. M.X., X.L., and T.L. collected the data. M.X., X.L., T.L., and X.Z. performed qualitative data analysis. M.X. and T.W. conducted quantitative data analysis. M.X. wrote the original draft and handled project administration. M.X. and T.L. reviewed and edited the manuscript. M.X. and X.D. prepared the figures.

Disclosure statement

In accordance with Taylor & Francis policy and our ethical obligations as researchers, we report that M.X., S.S., and X.D. are employed by the company responsible for developing one of the products Linfo AI evaluated in this study and participated in its design and implementation. We have disclosed those interests fully to Taylor & Francis, and we have in place an approved plan for managing any potential conflicts arising from this involvement.

Notes on contributors

Muhan Xu is a second-year PhD student at the Creative Computing Institute, University of the Arts London. His research focuses on building trustworthy AI systems to support academic reading, with an emphasis on designing transparent, verifiable, and user-adaptive tools. He holds an MSc in Creative Computing from the Creative Computing Institute and a BA in Product Design from Central Saint Martins, both at UAL. Previously, he co-founded Linfo AI and Tiamat AI, leading product teams to develop AI summarisation and image generation tools, which collectively reached over 500,000 users and secured multimillion-dollar funding. He has published work in venues including Frontiers in Computer Science – Human-Media Interaction and BCS HCI. Additionally, he has taught at universities including the Creative Computing Institute at UAL, Tongji University, the Central Academy of Fine Arts, and Beihang University.

Xiaomei Li, Postdoctoral Researcher in AI and Data Design at Tongji University, Ph.D. in Design from Central Academy of Fine Arts. Research areas: Design AI, generative AI tool development, design education innovation. Published 15 academic papers, leading 2 national grants. Awards include China Design Intelligence Award and Beijing Winter Olympics Outstanding Team Award. Focused on integrating AI technology in creative design and educational applications.

Te Lan holds a MSc in Data Science and AI from Creative Computing Institute at University of the Arts London, where he focused on Machine Learning and User Experience studies. His academic interests include trust, explainability and transparency in Machine Learning experience and human-AI collaboration. He currently collaborates with startups to apply user-centered design principles and machine learning techniques in a real-world setting.

Xiaoke Zeng focuses on creative applications of XR and AIGC, with cross-disciplinary experience as a researcher, designer and director. His work has been published in leading conferences including ACM CC, ACM CSCW, and ISEA. He has facilitated several GenAI-related projects and exhibitions, and his current research centres on GenAI for human well-being, education, and gamification, aiming to bridge technological innovation with societal impact.

Tianhong Wang is a Lecturer in Psychology at the School of Philosophy, Anhui University. She received her Ph.D. in Psychology from the School of Psychological and Cognitive Sciences at Peking University in January 2025, and her B.S. in Psychology from the School of Psychology and Cognitive Science at East China Normal University in 2019. She previously served as a Research Assistant at the Institute for AI Industry Research at Tsinghua University, where she conducted interdisciplinary research in human–computer interaction. She serves as a reviewer for academic journals including Humanities and Social Sciences Communications. Her research interests focus on human-machine collaboration, social psychology, and judgment and decision-making. She has published articles in journals such as the Journal of Experimental Psychology: Applied, Technology in Society, and Acta Psychologica Sinica. She has presented her research at high-level international conferences, including the International Congress of Psychology (ICP). She was selected as a Global Shaper by the World Economic Forum, recognising her commitment to addressing social challenges and her potential to drive positive change in her field and community.

Sixiong (Simon) Sheng is the CEO of Ponder AI Limited (ponder.ing, formerly ResearchFlow/Linfo.AI). He holds two master's degrees – an MRes and an MSc in Creative Computing – from the Creative Computing Institute at the University of the Arts London. He is currently pursuing a master's degree at Cambridge Judge Business School. Simon also earned a BA in Product Design from Central Saint Martins, University of the Arts London. His research focuses on Human–Computer Interaction (HCI) and Generative AI, with a current emphasis on enhancing cognitive processing in deep knowledge work through the use of Generative AI (LLMs).

Xiaolin Deng is the Co-founder and COO of Ponder AI Limited (ponder.ing, formerly ResearchFlow/Linfo.AI). She holds dual Master's degrees in Data Science & Artificial Intelligence (MSc) and Creative Computing (MRes) from the Creative Computing Institute at the University of the Arts London (UAL), as well as a BA in Design from the China Academy of Art. She is the co-author of The Future of AIGC in Design, a publication frequently cited in the fields of artificial intelligence and creative industries. Her research interests include human–computer interaction, generative AI, and computational creativity, with a current focus on the application of multimodal models in creative and knowledge-intensive contexts. She has been involved in the development of AI-supported systems for design and research, contributing to a broader understanding of how emerging technologies can facilitate interdisciplinary innovation.

Jingjing Sun is a second-year PhD student at Imperial College London focusing on designing playful AI music interactions for general wellbeing. Her research interests include Human-AI co-creation, playful design, and digital wellbeing. She holds an MFA in Design and Technology from Parsons School of Design and a B.S. in Information Science and Technology from Penn State University. Before joining Imperial, she worked as a research assistant at Tsinghua University, Lenovo Research, and Penn State University, conducting HCI-related user research and wellbeing studies.

Rebecca Fiebrink I am a Professor of Creative Computing at the UAL Creative Computing Institute. My students, research assistants, and I work on a variety of projects developing new technologies to enable new forms of human expression, creativity, and embodied interaction. Much of my current research combines techniques from human–computer interaction and machine learning to allow people to apply AI more effectively to new problems, such as the design of new digital musical instruments, and gestural interfaces for gaming and accessibility. I am also involved in projects developing rich interactive technologies for digital humanities scholarship, exploring ways that machine learning can be used and appropriated to reveal and challenge patterns of bias and inequality, and advancing machine learning education. My research frequently includes the creation of software tools for use in creative practice and teaching. In 2008, I developed the popular Wekinator tool for real-time interactive machine learning, which enables creative practitioners to develop new real-time, gestural interactions through providing demonstrations. Since then, my research collaborators and I have released a number of other tools for applying embodied and interactive machine learning in various environments, for instance RAPID-MIX for creative industries developers, Sound Control for music teachers and students, InteractML for game and VR designers, and MIMIC for creative coding students.

Nick Bryan-Kinns is Professor of Creative Computing at the Creative Computing Institute, University of the Arts London. He is a Fellow of the Royal Society of Arts and the British Computer Society, and Association of Computing Machinery Senior Member. Bryan-Kinns has published award winning papers on his extensively funded research into Human Centred AI, explainable AI, AI Music, cross-cultural design, mutual engagement, interactive art, and tangible interfaces.

Professor ***Mick Grierson*** is a Professor of Computer Science who has worked at the intersection of AI and the Creative Industries for 25 years, raising approx. £15 million from governments and foundations, including for the UK's first Creative Deep Learning research project, MIMIC (2018), in partnership with Google Magenta. He is founder of the discipline of Creative Computing and co-founder of UAL Creative Computing Institute (CCI). His research lab produced some of the first Creative AI research, including the first commercial video artwork using Deep Learning (AutoEncoding Bladerunner, 2016), the first real-time interactive video generation system using Deep Learning (Learning to See, 2019), and the first professional quality generative audio system (MAGNet, 2016). This was used by Massive Attack in 2018 to create the first CD quality neural synthesis full-length work, Mezzanine Vs MAGNET, which toured for 5 years. He co-founded music technology startups in AI and generative music, including Bronze Format Limited in 2010. He led the first BSc and MSc Creative Computing programmes in the UK (2007/8), and the first Massive Open Online Course in Creative Programming in the world (2013), attracting 250,000 learners.

Data availability statement

The data that support the findings of this study are available from the corresponding author, M.X., upon reasonable request.

ORCID

Muhan Xu http://orcid.org/0009-0004-9692-3148

References

Ahmadi, M. R., Ismail, H. N., & Abdullah, M. K. K. (2013). The importance of metacognitive Reading strategy awareness in Reading comprehension. *English Language Teaching*, *6*(10), 235. https://doi.org/10.5539/elt.v6n10p235

Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman.

Anne Britt, M., Richter, T., & Rouet, J.-F. (2014). Scientific literacy: The role of goal-directed Reading and evaluation in understanding scientific information. *Educational Psychologist*, *49*(2), 104–122. https://doi.org/10.1080/00461520.2014.916217

August, T., Wang, L. L., Bragg, J., Hearst, M. A., Head, A., & Lo, K. (2023). Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans. Comput.-Hum. Interact*, *30*(5), 74:1-74:38. https://doi.org/10.1145/3589955

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. In Ö. Aydın (Ed.), *Emerging Computer Technologies*, *2*, 22–31.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

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* 2, 59–68. Retrieved September 11, 2024 from https://burrell.edu/wp-content/uploads/2020/09/EBjorkRBjork_FABBSchapter2014-2nd-ed._WithCoverPage.pdf.

Carson, L., Bartneck, C., & Voges, K. (2013). Over-Competitiveness in academia: A literature review. *Disruptive Science and Technology*, *1*(4), 183–190. https://doi.org/10.1089/dst.2013.0013

Chaudhri, V. K., Cheng, B., Overtholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., & Gunning, D. (2013). Inquire biology: A textbook that answers questions. *AI Magazine*, *34*(3), 55–72. https://doi.org/10.1609/aimag.v34i3.2486

Chen, L., Yan, F., Zhong, Y., Chen, S., Jie, Z., & Ma, L. (2024). *MindBench: A comprehensive benchmark for mind map structure recognition and analysis.* Retrieved September 10, 2024 from http://arxiv.org/abs/2407.02842.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. (1978). *Mind in society. Mind in society the development of higher psychological processes*. Harvard University Press. Retrieved October 5, 2025 from https://www.academia.edu/download/40011524/vygotskysummary_missing.pdf.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of Reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418. https://doi.org/10.1038/s41467-024-45563-x

DeStefano, D., & LeFevre, J.-A. (2007). Cognitive load in hypertext Reading: A review. *Computers in Human Behavior*, *23*(3), 1616–1641. https://doi.org/10.1016/j.chb.2005.08.012

Dhillon, P. S., Molaei, S., Li, J., Golub, M., Zheng, S., & Robert, L. P. (2024). Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–18). https://doi.org/10.1145/3613904.3642134

Dowling, M. M., & Lucey, B. M. (2023). ChatGPT for (finance) research: The bananarama conjecture. *Finance Research Letters*, *53*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Dubay, W. (2004). The principles of readability. CA 92627949: 631–3309.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From local to global: A graph RAG approach to query-focused summarization. https://doi.org/10.48550/arXiv.2404.16130

Ellen, M. E., Lavis, J. N., Wilson, M. G., Grimshaw, J., Brian Haynes, R., Ouimet, M., Raina, P., & Gruen, R. (2014). Health system decision makers' feedback on summaries and tools supporting the use of systematic reviews: A qualitative study. *Evidence & Policy*, *10*(3), 337–359. https://doi.org/10.1332/174426514X672362

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

Gao, J., Gebreegziabher, S. A., Choo, K. T. W., Li, T. J.-J., Perrault, S. T., & Malone, T. W. (2024a). A taxonomy for human-LLM interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing* Systems (pp. 1–11). https://doi.org/10.1145/3613905.3650786

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024b). *Retrieval-augmented generation for large language models: A Survey*. https://doi.org/10.48550/arXiv.2312.10997

Gu, Z., Arawjo, I., Li, K., Kummerfeld, J. K., & Glassman, E. L. (2024). An AI-resilient text rendering technique for reading and skimming documents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–22). https://doi.org/10.1145/3613904.3642699

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 139–183). https://doi.org/10.1016/S0166-4115(08)62386-9

Hayot, E. (2014). *The elements of academic style: Writing for the humanities*. Columbia University Press.

Huang, J., & Chang, K. (2024). Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 464–473). https://doi.org/10.18653/v1/2024.findings-naacl.31

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 248. 1-248:38. https://doi.org/10.1145/3571730

Jiang, P., Rayan, J., Dow, S. P., & Xia, H. (2023). Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–20). https://doi.org/10.1145/3586183.3606737

Kendeou, P., & Van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during Reading of scientific texts. *Memory & Cognition*, *35*(7), 1567–1577. https://doi.org/10.3758/BF03193491

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, *25*(1), 105–121. https://doi.org/10.2307/3587030

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 2390–2395). https://doi.org/10.1145/2858036.2858402

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task*. . https://doi.org/10.48550/arXiv.2506.08872

Le, V. H., & Huynh, N. V. T. (2022). Critical reading course: Problems and reflections from Vietnamese English-majored students' Perspectives. In *2022 the 4th International Conference on Modern Educational Technology (ICMET)* (pp. 118–126). https://doi.org/10.1145/3543407.3543427

Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education*, *22*(1), 7. https://doi.org/10.1186/s41239-025-00503-7

Lee, M., Gero, K. I., Chung, J. J. Y., Shum, S. B., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E. A., August, T., Bhat, A., Choksi, M. Z., Dutta, S., Guo, J. L. C., Hoque, M. N., Kim, Y., Knight, S., Neshaei, S. P., … Siangliulue, P. (2024). A design space for intelligent and interactive writing assistants. https://doi.org/10.1145/3613904.3642697

linfo.ai. (2024). *Revolutionizing web reading with Linfo AI: A Deep dive into the Chrome Plugin*. Retrieved September 10, 2024 from https://blog.linfo.ai/post/revolutionizing-web-reading-with-linfo-ai-a-deep-dive-into-the-chrome-plugin/.

Liu, M. X., Liu, F., Fiannaca, A. J., Koo, T., Dixon, L., Terry, M., & Cai, C. J. (2024a). "We need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1–9). https://doi.org/10.1145/3613905.3650756

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Liu, Y., Li, D., Wang, K., Xiong, Z., Shi, F., Wang, J., Li, B., & Hang, B. (2024b). Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. *Information Processing & Management*, *61*(5), 103809. https://doi.org/10.1016/j.ipm.2024.103809

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3313831.3376727

Long, D., Roberts, J., Magerko, B., Holstein, K., DiPaola, D., & Martin, F. (2023). AI literacy: finding common threads between education, design, policy, and explainability. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)* (pp. 1–6). https://doi.org/10.1145/3544549.3573808

Mah, D.-K., & Groß, N. (2024). Artificial intelligence in higher education: Exploring faculty use, self-efficacy, distinct profiles, and professional development needs. *International Journal of Educational Technology in Higher Education*, *21*(1), 58. https://doi.org/10.1186/s41239-024-00490-1

Malik, A., Laeeq Khan, M., Hussain, K., Qadir, J., & Tarhini, A. (2025). AI in higher education: Unveiling academicians' perspectives on teaching, research, and ethics in the age of ChatGPT. *Interactive Learning Environments*, *33*(3), 2390–2406. https://doi.org/10.1080/10494820.2024.2409407

Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.

Park, H., & Ahn, D. (2024). The promise and peril of ChatGPT in higher education: Opportunities, challenges, and design implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–21). https://doi.org/10.1145/3613904.3642785

Peregoy, S. F., & Boyle, O. F. (2000). English learners reading English: What We know, what we need to know. *Theory Into Practice*, *39*(4), 237–247. https://doi.org/10.1207/s15430421tip3904_7

Radensky, M., Weld, D. S., Chang, J. C., Siangliulue, P., & Bragg, J. (2024). Let's get to the point: LLM-supported planning, drafting, and revising of research-paper blog posts. Retrieved September 10, 2024 from http://arxiv.org/abs/2406.10370.

Rayner, K., Schotter, E. R., Masson, M. E. J., Potter, M. C., & Treiman, R. (2016). So much to read, So little time: How Do We read, and Can speed Reading help? *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *17*(1), 4–34. https://doi.org/10.1177/1529100615623267

Rocha, M. A. M., Nacenta, M. A., & Ye, J. (2021). An exploratory study on academic reading contexts, technology, and strategies. In *8th Mexican Conference on Human-Computer Interaction* (pp. 1–10). https://doi.org/10.1145/3492724.3492727

Salleh, H. M. (2023). Errors of commission and omission in artificial intelligence: Contextual biases and voids of ChatGPT as a research assistant. *Digital Economy and Sustainable Development*, *1*(1), 14. https://doi.org/10.1007/s44265-023-00015-0

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models Are double-edged swords. *Radiology*, *307*(2), e230163. https://doi.org/10.1148/radiol.230163

Similarweb. (2024). chatgpt.com traffic analytics, ranking & audience [August 2024] Similarweb. Retrieved September 9, 2024 from https://www.similarweb.com/website/chatgpt.com/#overview.

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The user experience of ChatGPT: Findings from a questionnaire study of early users. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–10). https://doi.org/10.1145/3571884.3597144

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*, 108386. https://doi.org/10.1016/j.chb.2024.108386

Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., & Seifert, C. (2024). Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with LLMs. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3642754

Suh, S., Min, B., Palani, S., & Xia, H. (2023). Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (pp. 1–18). .https://doi.org/10.1145/3586183.3606756

Thomas, D. A. (2003). *A general inductive approach for qualitative data analysis*.

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226.

Walter, Y. (2024). Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, *21*(1), 15. https://doi.org/10.1186/s41239-024-00448-3

Wan, X., Li, H., & Xiao, J. (2010). EUSUM: Extracting easy-to-understand english summaries for non-native readers. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval: 491–498. https://doi.org/10.1145/1835449.1835532

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 621. https://doi.org/10.1057/s41599-025-04787-y

Wang, J., Hu, H., Wang, Z., Yan, S., Sheng, Y., & He, D. (2024). Evaluating large language models on academic literature understanding and review: An empirical study among early-stage scholars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3613904.3641917

Wang, Z., Tu, Y., Rosset, C., Craswell, N., Wu, M., & Ai, Q. (2023). Zero-shot clarifying question generation for conversational search. In *Proceedings of the ACM Web Conference 2023* (pp. 3288–3298). https://doi.org/10.1145/3543507.3583420

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), e70105. https://doi.org/10.1111/jcal.70105

Wen, H., Lin, X., Liu, R., & Su, C. (2025). Enhancing college students' AI literacy through human-AI co-creation: A quantitative study. *Interactive Learning Environments*, 1–19. https://doi.org/10.1080/10494820.2025.2495733

Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. International management review. Retrieved September 9, 2024 from https://www.semanticscholar.org/paper/The-Art-of-Coding-and-Thematic-Exploration-in-Williams-Moser/c0a0c26ac41cb8beb337834e6c1e2f35b91d071d.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100.

Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems* (pp. 1–22). https://doi.org/10.1145/3491102.3517582

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). *Can LLMs Express their uncertainty? An empirical evaluation of confidence elicitation in LLMs*. https://doi.org/10.48550/arXiv.2306.13063

Xu, T., Wu, S., Diao, S., Liu, X., Wang, X., Chen, Y., & Gao, J. (2024). SaySelf: Teaching LLMs to express confidence with self-reflective rationales. https://doi.org/10.48550/arXiv.2405.20974

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. ACM Transactions on Knowledge Discovery from Data: 3649506. https://doi.org/10.1145/3649506

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (n.d.) *Exploring the limits of ChatGPT for query or aspect-based text summarization.*

Yeh, H.-C. (2025). The synergy of generative AI and inquiry-based learning: Transforming the landscape of English teaching and learning. *Interactive Learning Environments*, *33*(1), 88–102. https://doi.org/10.1080/10494820.2024.2335491

Yi, J. S. (2014). QnDReview: read 100 CHI papers in 7 hours. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (pp. 805–814). https://doi.org/10.1145/2559206.2578884

Yun, B., Feng, D., Chen, A. S., Nikzad, A., & Salehi, N. (2025). Generative AI in knowledge work: Design implications for data navigation and decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). . https://doi.org/10.1145/3706598.3713337

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). https://doi.org/10.1145/3544548.3581388

Zavolokina, L., Sprenkamp, K., Katashinskaya, Z., Jones, D. G., & Schwabe, G. (2024). Think fast, think slow, think critical: designing an automated propaganda detection tool. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–24). https://doi.org/10.1145/3613904.3642805

Zhang, L., Liu, P., Henriksboe, M. T. O., Lauvrak, E. W., Gulla, J. A., & Ramampiaro, H. (2025). PersonalSum: A user-subjective guided personalized summarization dataset for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)* (pp. 99333–99346).

Zumbach, J., & Mohraz, M. (2008). Cognitive load in hypermedia Reading comprehension: Influence of text type and linearity. *Computers in Human Behavior*, *24*(3), 875–887. https://doi.org/10.1016/j.chb.2007.02.015

Appendices

A. Configuration details for Linfo AI

Prompt for Linfo AI structural summarisation was: "Your task is to transform the content into a complete comprehensive structured mind map. This mind map should offer a thorough understanding of the essay's core ideas. Please follow this formatting structure of Markdown: Use # (h1) for the title of the content. Use ## (h2) for sub-categories. Ensure there are ample sub-categories for clarity. Use ### (h3) for detailed statements summarised from the essay. Please use legit Markdown format. Please make the outcome coherent, and insightful. Strictly follow the original language of the article above. Use the language of the majority of the article." Linfo AI used Claude 3.5 Haiku through an API call with the temperature set to 0.

B. Paper categorisation task question

Writing Assistant Paper: Design and Evaluation of a Social Media Writing Support Tool for People with Dyslexia Categorise the writing assistant systems mentioned in the paper into the following **writing stage(s)**. (Select one or two)

A. **Idea generation:** Brainstorming and developing concepts or content themes
B. **Planning:** Organising the structure and content outline
C. **Drafting:** Composing the written content
D. **Revision:** Reviewing and refining the written material

C. Post task questionnaire

1. Please choose the term that best describes your reading experience with Mind-map-based AI support. There are no right or wrong answers, only your personal opinion counts. (7-point Likert Scale).
   Obstructive (1) – Supportive (7)
   Complicated (1) – Easy (7)
   Inefficient (1) – Efficient (7)
   Confusing (1) – Clear (7)
   Boring (1) – Exciting (7)
   Not interesting (1) – Interesting (7)
   Conventional (1) – Inventive (7)
   Usual (1) – Leading edge (7)
2. Considering your recent academic reading experience, please answer the following questions (5-point Likert scale: 1 = very low, 5 = very high).
   How mentally demanding was the task?
   How physically demanding was the task?
   How hurried or rushed was the pace of the task?
   How successful do you think you were in accomplishing the task?
   How hard did you have to work to accomplish your level of performance?
   How insecure, discouraged, irritated, stressed, and annoyed were you during the task?
3. Please indicate your level of agreement with the following statements (5-point Likert scale: 1 = strongly disagree, 5 = strongly agree).
   I can effectively find specific examples in the reading material to support my chosen writing stage(s).
   I can effectively identify the writing stage(s) that the reading material belongs to.
   I can effectively summarise a single statement that represents the reading material.
   I can effectively detect similarities and differences between the reading material and different writing stage(s).

D. Post study questionnaire

1. Please rank the three conditions from most to least helpful for answering the reading questions:
   A. Non-AI Support
   B. Conversational AI Support (ChatGPT)
   C. Structural AI Support (Linfo AI)
2. What did you like the most about the condition you found the most helpful?
3. Are there any features missing that you would like to see in the condition you found the most helpful?
4. Considering your recent experience completing the reading task, please indicate the extent to which you agree with the following statements (5-point Likert Scale).
   1. I would like to use **Non-AI support** for academic reading in the future.
   2. I would like to use **conversational AI support** for academic reading in the future.
   3. I would like to use **structural AI support** for academic reading in the future.