

Combining Structured Tasks and Behavioral Logs for Measuring AI’s Impact on Cognition

Jiayin Zhi

jzhi@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Mina Lee

mnlee@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Abstract

Generative AI is transforming how people read, write, and think, but measuring its impact on cognition remains challenging. We highlight and synthesize a combined approach using two complementary measurement layers: structured tasks with established metrics that capture cognitive outcomes, and behavioral interaction logs that reveal the intermediate pathways of how people engage with AI during the task. We categorize how each method has been used in prior work, and illustrate the advantage of combining them through a case study investigating AI’s impact on critical thinking, where the two layers produced insights that neither would have provided alone, pointing to specific design implications for AI tools that enhance outcomes and protect human cognition. We point to future work on expanding this approach across cognitive constructs, domains, and tool types—toward a principled, transferable framework for measuring AI’s impact on human cognition.

1 Introduction

Generative AI (GenAI) tools are increasingly integrated in people’s cognitive tasks, such as creative ideation, problem-solving, and critical thinking [25, 26, 29]. As adoption grows, so does attention to the potential benefits and risks of AI for human cognition. A growing body of empirical research is investigating these effects across cognitive tasks. A critical challenge is: how do we measure AI’s impact on human cognition?

One common approach facilitating empirical investigation is to use structured tasks with established metrics—validated tasks designed to assess a targeted cognitive construct, with scoring methods grounded in relevant literature. For example, for investigating AI’s impact on creativity, Lee and Chung [19] compared participants using ChatGPT, conventional web search, or no external aid on custom problem-solving tasks—repurposing unused household items or designing an innovative dining table. Participants’ ideas were rated on originality and fluency—established metrics grounded in creativity literature. Such metrics capture task outcomes, and comparing them across conditions can provide direct evidence of AI’s impact on a targeted cognitive ability. Yet they tell us less about the cognitive processes during the task, such as how people engage with the AI, deliberate, and make decisions along the way.

Understanding these underlying cognitive processes matters for two reasons. First, task outcomes alone can obscure critical insights: given AI’s capabilities, outcomes may appear beneficial even when people are passively relying on AI. Uncovering how people engage with AI during cognitive tasks helps distinguish productive collaboration from passive reliance. Second, such understanding can yield more actionable insights for design: they can reveal not just whether a cognitive outcome is affected, but where in the workflow

and through what mechanisms, pointing to design opportunities for AI tools that protect and augment human cognition.

A complement to structured tasks is behavioral interaction logs—timestamped records of user actions within a digital environment, such as keystroke, cursor, and navigation events, collected as a byproduct of the interactive system. Researchers across HCI, educational psychology, and information science have combined task-based assessments with such process data to study cognition in various contexts: for example, combining learning outcome tests with search behavior logs to examine how people learn through information seeking [6, 11].

In this paper, we highlight the value of this combined approach for measuring AI’s impact on cognition. We synthesize how each approach has been used in prior work, discuss their relative strengths and limitations, and illustrate the advantage of combining the two complementary measurement layers. The foundational layer, structured tasks with established metrics, captures cognitive outcomes: it tells us whether and how a targeted cognitive ability was affected (Section 2). Built on top of this, behavioral interaction logs, captures what people actually did during the task in their interaction with the study interface. In particular, we focus on log-based metrics that can be automatically computed, offering a scalable and unobtrusive complement to task outcomes (Section 3). While structured tasks tell us whether a cognitive outcome was affected, behavioral logs show us the intermediate pathways. Together, the two layers connect cognitive outcomes to the processes that produced them, enabling researchers and designers to identify not just that cognition was affected, but how and where to intervene.

We illustrate this combined approach through a case study investigating AI’s impact on critical thinking (Section 4). We show that the two layers produced insights that neither would have provided alone, connecting what happened to critical thinking task performance with how it happened during interaction, and pointing to specific design implications for AI tools. We point to future work on expanding this combined approach across other cognitive constructs, domains, and tool types, and on addressing limitations of log-based metrics through complementary methods with qualitative insights (Section 5)—toward a principled, transferable framework for measuring AI’s impact on human cognition, one that can help build understanding of how AI affects specific cognitive abilities and the processes underlying those effects, and translate empirical findings into design strategies that protect and augment cognition.

2 Structured Tasks with Established Metrics

Structured tasks with established metrics have been widely used to assess the impact of AI on specific cognitive constructs such as

Approach	Description	What It Captures	Example	Risk
Intermediate outcomes	Outcomes collected at different stages of a task	Snapshots of cognitive progress at different task stages	Task questions requiring written responses, each reflecting a particular cognitive process in Bloom’s taxonomy, from understanding to evaluating [2, 11].	Can miss the pathways between snapshots, and assessments may interrupt natural, non-linear cognitive pathways.
Self-reports	Surveys or interviews capturing participants’ perceptions	Perceived cognition across various dimensions	Perceived ability (e.g., “I can distinguish if I interact with an AI or a real human” [4]); cognitive processes (e.g., “I examine the logical strength of the underlying reason” [22, 27]).	Participants tend to overestimate their abilities [5] and show poor precision when assessing their own performance [7].
Embodied sensing	Physiological sensing technologies	Moment-to-moment cognitive states	Electroencephalography to measure cognitive load accumulation during AI-assisted writing [16].	Largely confined to lab settings; cost and privacy challenges for large-scale deployment.
Behavioral interaction logs	Timestamped records of user actions within a digital environment	What people actually did during the task	Keystrokes, clicks, document navigation, AI prompts [20, 28]	Observable behavior does not always reflect intent.

Table 1: Approaches in prior works for capturing cognitive processes during human-AI interaction, with descriptions, examples, and potential risks. Among these, behavioral interaction logs offer a distinct advantage: they are low-cost, scalable, and require no extra equipment beyond the interactive system itself [20, 28].

creativity [18, 19], comprehension [17], and critical thinking [3, 8]. Here, we offer a categorization of the range of such tasks.

Standardized Cognitive Tasks. Standardized cognitive tasks are validated, off-the-shelf instruments with established scoring rubrics. For example, the Alternate Uses Task (AUT) [12] measures divergent thinking by asking participants to generate novel uses for everyday objects. Another example is the Remote Associates Test (RAT) [14] that assesses convergent thinking by asking participants to find a word connecting three seemingly unrelated words.

Custom Tasks with Established Metrics. When standardized tasks risk potential biases from AI models’ prior exposure to the tasks or may not fit the research context, researchers have designed study-specific tasks while retaining established metrics and scoring methods. For instance, Lee and Chung [19] designed novel problem-solving tasks—such as repurposing unused household items or designing an innovative dining table—deliberately avoiding publicly available tests like the AUT, but had external judges rate outputs on originality and appropriateness, established metrics from the creativity literature. Kreijkes et al. [17] designed custom reading comprehension tasks using cued recall, multiple choice, and free recall questions targeting different levels of text understanding. In both cases, the tasks were custom-designed but the measurement approaches drew on established metrics and scoring methods.

Performance Assessment. Some cognitive constructs, like critical thinking, are characterized by multiple facets and involve several interacting cognitive abilities, making them difficult to capture with a single standardized test or isolated task. Critical thinking is the ability to reason through diverse and sometimes conflicting information to reach reasoned decisions [3, 23]. In practice, this unfolds as a non-linear process, involving distilling arguments from sources, switching between sources to weigh trade-offs, checking one’s own bias, and communicating the deliberative process—each representing a distinct facet of critical thinking. Performance assessments [9, 10, 15], such as the iPAL (International Performance Assessment of Learning) framework [3, 23], address this complexity by tasking participants to make a reasoned decision for real-world scenarios with a set of documents, capturing this non-linear process within a single task. The framework has demonstrated content

validity, showing that argumentation, source analysis, and written communication emerge as distinct measurable facets aligned with these theoretical components of critical thinking; and construct validity, confirming that the written essays capture key aspects of the underlying deliberative process [8]. We use this framework in our case study (Section 4).

3 Behavioral Logs for Analytic Purposes

Compared to the alternative approaches summarized in Table 1, behavioral interaction logs offer a distinct advantage for capturing cognitive processes: they are low-cost, scalable, and require no extra equipment beyond the interactive system itself [20, 28]. In this section, we illustrate how behavioral logs have been collected and used for analytic purposes through examples of empirical studies across AI writing assistants, AI chatbots, and AI-powered search tools. These tools all involve reading and writing as part of the interaction process—the fundamental activities through which people engage with information and that serve as windows into human cognition. Table 2 provides a summary of the examples discussed.

AI Writing Assistants (popup suggestion). In the CoAuthor platform [20], users press a key to request multiple AI-generated text continuations, which appear in a popup menu to browse, accept, or dismiss. The platform records every text, cursor, and suggestion-related events. Each character in the final text is traced back to either the writer or the model based on keystroke-level provenance. Padmakumar and He [21] used this infrastructure to study whether co-writing with LLMs makes different writers’ essays more similar to each other. They compared essays written solo, with a base model, and with a feedback-tuned model in an open-ended essay writing task. Log-derived metrics such as acceptance rates and model-written text fractions served as manipulation checks confirming comparable engagement across conditions, while the character-level attribution served as the primary analytical outcome, allowing the researchers to split each essay into its user-written and model-written portions and measure content similarity within each portion separately. They found that the increased similarity was driven entirely by the model-contributed text, while user-written portions remained as diverse as solo writing. This

Tool Type	Task Type	Key Metric	How the Metric was Computed	Analytical Role
Popup AI writing assistant [21]	Open-ended essay writing task	Content similarity (user portion)	Cosine similarity across essays using only writer-attributed characters	Primary
		Content similarity (model portion)	Cosine similarity across essays using only model-attributed characters	Primary
Inline AI writing assistant [1]	Open-ended essay writing task	Acceptance rate	Proportion of accepted AI suggestions	Primary
		AI reliance	AI-originated characters / total characters via longest common subsequence	Primary
AI chatbot alongside text and notepad [17]	Structured reading comprehension task	Prompt frequency	Number of prompts per session	Exploratory
		Time on task	Total time spent on the task	Exploratory
AI conversational search [24]	Open-ended search task	Number of topics	Unique topics counted across queries	Primary
AI chatbot alongside document viewer and text editor [30]	Structured critical thinking performance assessment task	Argument overlap	Proportion of essay arguments overlapping with LLM responses	Exploratory
		Document viewing count	Number of unique documents viewed before and after participants start writing	Exploratory

Table 2: Examples of behavioral log-based metrics (computed from raw logs) across AI-augmented tools discussed in Sections 3 and 4. Researcher-coded metrics such as prompt archetypes [17] and query cognitive level [24] are discussed in the main text but excluded here as they require human interpretation beyond log processing.

suggested that co-writing with AI may reshape what appears on the page without altering the writer’s own generative thinking.

AI Writing Assistants (inline suggestion). Agarwal et al. [1] studied whether AI suggestions homogenize writing styles across cultures using an inline text editor where word-level suggestions appear as grey text as the writer types, which users accept or ignore. Comparing essays written by American and Indian participants with and without AI in an open-ended writing task, they derived four behavioral metrics from the logs: suggestion acceptance rate (proportion of suggestions accepted), AI reliance (proportion of the final text originating from AI, computed via longest common subsequence between accepted suggestions and the final text), suggestion modification (whether accepted suggestions were later edited), and writing productivity (words written per second). These metrics served as the primary outcomes for analysis, revealing that Indian participants accepted more suggestions, relied more heavily on AI, and modified accepted suggestions more. These behavioral differences suggest that AI suggestions may mediate the composing process differently across cultures.

AI Chatbots alongside Text Passage and Notepad. Beyond analyzing chat transcripts alone, prior work has leveraged timestamped interaction data to capture cognitive processes. Kreijkes et al. [17] studied whether using an LLM chatbot for reading comprehension affects students’ learning compared to note-taking. The interface presented a text passage, an AI chatbot, and a notepad. In this structured reading comprehension task, the main analysis used test scores to compare retention and comprehension across conditions, finding that both note-taking alone and LLM combined with note-taking outperformed LLM alone. For exploration, log-based metrics such as prompt frequency and time-on-task showed that

students using the LLM alone spent slightly less time on the task, suggesting less sustained engagement with the reading material, while students with access to both tools prompted the LLM less frequently, suggesting greater engagement with note-taking. The researchers also classified prompts into behavioral archetypes—such as seeking deeper understanding, requesting summaries, and asking for definitions—to characterize how students engaged with the LLM. Note that such researcher-coded indicators differ from the automatically computed log-based metrics that are the focus of this section; we return to this distinction in Section 5.

AI Search. Search log analysis has established that behavioral signals, such as documents clicked, time spent assessing search result pages, and time spent viewing each document, can be meaningfully correlated with cognitive learning outcomes [6]. As search tools increasingly integrate AI, for example, Singh et al. [24] study whether metacognitive prompts can enhance critical thinking when students search with an AI conversational search tool. In this open-ended search task, participants searched a topic and took notes using either a standard notepad or one augmented with metacognitive prompts. From the search sessions, the researchers derived both log-based metrics—search duration, number of queries, number of sources clicked, and number of unique topics explored—and researcher-coded indicators based on qualitative analysis of screen recordings: the cognitive level of each query (receptive versus critical), persistent inquiry (whether students pursued ideas through follow-up queries), source engagement (whether students gathered information from cited sources), and independent thinking (whether students added their own inputs beyond AI-generated content). Together, these log-based metrics and researcher-coded indicators revealed that participants receiving metacognitive prompts explored more topics and demonstrated greater persistent inquiry,

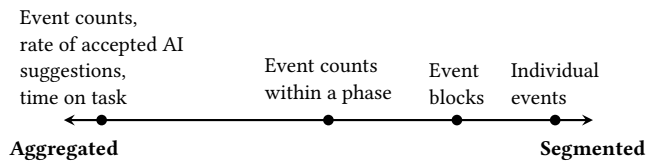


Figure 1: The aggregation-segmentation spectrum of log-based metrics: the extent to which a metric preserves temporal and sequential information from the task process. Fully aggregated metrics collapse the entire task into summary statistics, while fully segmented metrics retain individual events. Examples of metrics illustrate different granularities along the spectrum.

suggesting active questioning and deeper cognitive engagement with information.

Takeaways. In these examples (Table 2), log-based metrics served different analytical roles. In open-ended tasks, where no single correct answer exists, they served as the primary analytical outcome. These metrics mostly fall toward the aggregated end of the spectrum (Figure 1), collapsing raw event streams into summary statistics such as event counts, proportion of accepted AI suggestions, or the AI’s share of final text. While effective for answering targeted research questions, such aggregation can compress rich sequential and temporal information, potentially overlooking intermediate cognitive processes such as how engagement patterns shifted over the course of the task. When structured tasks with established metrics provide the primary outcome measure, as in Kreijkes et al. [17], logs can play a more exploratory role—still characterizing engagement patterns, but with the added advantage that validated task outcomes provide an anchor for interpreting what those patterns mean. In both cases, behavioral logs can help explain how and why cognitive outcomes were affected; the difference lies in whether they carry the primary analytical burden or complement a separate outcome measure. This points to a natural complementarity: structured tasks establish what cognitive outcome was affected, while behavioral logs help explain how and why.

Moving beyond aggregated metrics, those that preserve more temporal structure—such as event counts within a task phase or event blocks—can better capture granular process-level patterns, though the examples reviewed here largely cluster toward the aggregated end. In the next section, we illustrate this combined approach through a case study on critical thinking, where log-based metrics at intermediate granularities, providing process-level insights alongside structured task outcomes.

4 Case Study: AI’s Impact on Critical Thinking

We illustrate the combined approach through our recent study investigating the effects of LLM use on critical thinking under time constraints [30].¹ This case study shows how the two measurement layers—performance assessment and behavioral interaction logs—together provide a more comprehensive picture of how AI affects a

cognitive ability and the processes underlying that effect, yielding insights that neither layer would have provided alone.

Study Design. We designed a 4×2 between-subjects experiment ($n = 393$) manipulating two types of time constraints: LLM access timing—whether the LLM chatbot was available only at the beginning of the task (Early), throughout (Continuous), only near the end (Late), or not at all (No LLM access)—and time availability (Insufficient or Sufficient time). Using the critical thinking performance assessment task [3, 8], participants were asked to make a reasoned decision based on a curated set of documents of varying characteristics, and to write an argumentative essay for their reasoning.

AI Chatbot alongside Document Viewer and Text Editor. A custom web interface presented three side-by-side panels—a document viewer, an LLM chatbot, and a text editor. This interface captured keystroke, cursor, and navigation events across all three panels, enabling analysis of how participants moved between reading, using the LLM, and writing throughout the task, yielding log-based metrics at intermediate granularities along the aggregation-segmentation spectrum (Figure 1).

What Performance Assessment Revealed. The layer captured critical thinking primarily through the outcome measures of Essay score (primarily based on number of valid arguments) and Myside Bias score (the imbalance between pro and con arguments, reflecting whether participants considered multiple perspectives). These measures revealed a striking temporal reversal: having LLM access from the start (Early, Continuous LLM access) enhanced Essay performance with insufficient time but impaired it with sufficient time, while working independently first (Late, No LLM access) showed the opposite pattern. Notably, self-reported assessments of critical thinking showed minimal variation across conditions, highlighting the necessity of performance assessment to capture effects on cognition that participants themselves cannot perceive.

What Behavioral Logs Added. The second layer of behavioral logs revealed mechanisms underlying the performance patterns. Copying behavior analysis revealed that direct copying of LLM responses was rare across all conditions, suggesting that the influence of having LLM access from the start operates not through direct content adoption but through subtler shifts in engagement. Argument overlap analysis between participants’ essays and the LLM responses they received showed that participants with Early and Continuous LLM access had minimal increase in non-overlapping arguments from insufficient to sufficient time, suggesting that having LLM access from the start may limit further argumentation. Document viewing counts—measured as unique documents clicked before and after participants started writing—showed that participants having LLM access from the start viewed fewer documents during writing, particularly under sufficient time, indicating less iterative consultation of source documents and greater anchoring to LLM-provided initial framing. Together, these behavioral signals revealed how early AI availability can shape the trajectory of thinking by narrowing document engagement and limiting further deliberation, even without direct copying of AI content.

Combined Insights for Design. The two layers together yielded design implications that neither alone could support. Performance assessment alone showed that having LLM access from the start

¹This study was conducted independently and accepted to CHI ’26. We use it here as a case study to demonstrate the combined approach.

improved performance under time pressure but impaired it with sufficient time, suggesting broadly that independent work before LLM use is preferable. But behavioral logs revealed that reduced deliberation and reliance on AI framing occurred across conditions: even when having LLM access from the start improved task performance under time pressure, the underlying pattern of narrowed document engagement and constrained deliberation remained concerning. Task performance alone would miss these patterns. Knowing *where and why* these shifts occur, designers can target interventions accordingly: for example, prompting users to consult sources not covered in the AI response before finalizing arguments, or reminding users to attempt further deliberation after receiving an AI response.

5 Implications and Future Work

We discuss considerations for designing and using log-based metrics, complementary methods, and directions for future work.

Navigating the Aggregation–Segmentation Spectrum for Designing Log-Based Metrics. A crucial consideration for designing log-based metrics is that the same observable action can reflect different cognitive states. In our case study, a participant who viewed fewer documents during writing could be anchoring to the LLM’s framing, or could have already read carefully and felt no need to revisit sources. However, counting document views separately before and after participants started writing—as we did in our case study—provides more process-level insight than the aggregated count alone, revealing whether participants returned to sources during writing or only consulted them beforehand. Similarly, a participant who spent more time on the task could be deeply engaged or simply distracted, but segmenting time-on-task by activity phase could distinguish sustained reading from idle periods. These examples suggest that moving along the spectrum toward more segmented metrics (Figure 1) can help disambiguate behavioral signals, though at the cost of increased analytical complexity. Choosing where to operate on this spectrum is a key design decision when constructing log-based metrics for studying cognitive processes.

Expanding the Space of Log-Based Metrics. The log-based metrics reviewed in this paper reflect a subset of how behavioral logs can be used for analytical purposes. Many signals remain underexplored: for instance, revision patterns in writing, transitions between AI-generated and self-authored content, or temporal rhythms of engagement such as bursts of activity followed by pauses that may indicate reflection. Different design choices in the pipeline from raw logs to metrics—what to record, how to segment events, what to compute—can surface different cognitive processes from the same data. Future work can map this space more systematically, developing principled approaches for deriving metrics aligned with cognitive constructs of interest.

Extending the Combined Approach. We illustrate this combined approach through critical thinking using one type of performance assessment, but see several directions for extending it:

- (1) transferring the approach to other cognitive constructs such as creativity or scientific reasoning, adapting the specific metrics and interpretation to each construct and tool context;

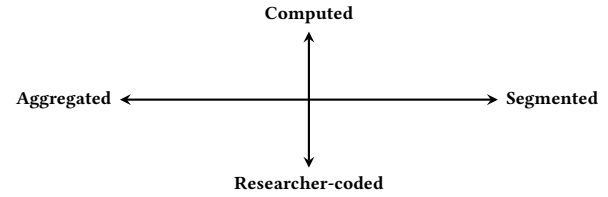


Figure 2: Two-dimensional space for characterizing indicators used to study cognitive processes during human-AI interaction. The horizontal axis represents the extent to which temporal information is preserved; the vertical axis represents how the indicator is derived.

- (2) extending to longitudinal studies examining whether repeated AI use affects cognitive development over time, where behavioral logs could track evolving patterns of engagement and offloading across sessions;
- (3) investigating how interface design decisions shape what behavioral signals are available, and how to design AI tool interfaces that generate informative logs without introducing noise or altering natural behavior.
- (4) developing new metrics and analytical approaches as AI tools become more complex and multi-modal.

Validation Challenges. When log-based metrics are used alongside structured tasks, the tasks ground interpretation against validated cognitive outcomes. When such tasks are unavailable, however, log-based metrics may need to serve as direct indicators of cognitive constructs, requiring careful validation [13]. This challenge intensifies in AI-augmented settings, where the AI itself shapes behavioral patterns. Developing systematic validation approaches for behavioral metrics in human-AI interaction remains an important open challenge.

Complementing Log-Based Metrics with Qualitative Insights. The log-based metrics in Table 2 can be automatically computed, falling along the aggregation–segmentation spectrum (Figure 1). A complementary approach is researcher-coded analysis, where human interpretation is applied to interaction data to capture cognitive processes that computed metrics may miss. For example, in Section 3, Kreijkes et al. [17] classified prompts into archetypes and Singh et al. [24] coded the cognitive level of each query—analyses that require qualitative judgment beyond log processing. In recent work, Umarova et al. [28] took a step towards bridging these approaches: translating qualitative descriptions of key cognitive interactions—such as “mindless echoing,” where writers merely rephrase AI output without contributing new ideas—into search queries that can detect these interactions across large log datasets.

These two dimensions—how an indicator is derived (computed versus researcher-coded) and how much temporal information it preserves (aggregated versus segmented)—form a two-dimensional space for characterizing indicators used to study cognitive processes during human-AI interaction (Figure 2). The log-based metrics reviewed in this paper occupy two quadrants of this space. Future work can expand into other quadrants, developing computed metrics that preserve more temporal structure, systematic coding schemes that scale, and hybrid approaches that bridge qualitative insight and computational scalability.

References

- [1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [2] Patricia Armstrong. 2010. Bloom's taxonomy. *Vanderbilt University Center for Teaching* 12, 05 (2010), 2023.
- [3] Henry I Braun, Richard J Shavelson, Olga Zlatkin-Troitschanskaia, and Katrina Borowiec. 2020. Performance assessment of critical thinking: Conceptualization, design, and implementation. In *Frontiers in Education*, Vol. 5. Frontiers Media SA, 156.
- [4] Astrid Carolus, Martin J Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAIIS-Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100014.
- [5] James S Cole and Robert M Gonyea. 2010. Accuracy of self-reported SAT and ACT test scores: Implications for research. *Research in Higher Education* 51, 4 (2010), 305–319.
- [6] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM conference on human information interaction and retrieval*. 163–172.
- [7] Kit S Double. 2025. Survey measures of metacognitive monitoring are often false. *Behavior Research Methods* 57, 3 (2025), 97.
- [8] Blake D. Ebright-Jones and Kai S. Cortina. 2025. Critical Thinking in College - Can We Teach It? *Journal of Education* 2025, 2 (April 2025), 145–164. <https://doi.org/10.3262/zp2502145>
- [9] Robert Hugh Ennis and Eric Edward Weir. 1985. *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Midwest Publications.
- [10] Peter A Facione. 1990. The California Critical Thinking Skills Test—College Level. Technical Report# 1. Experimental Validation and Content Validity. (1990).
- [11] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 22–31.
- [12] Joy P Guilford. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1, 1 (1967), 3–14.
- [13] Carolin Hahnel, Ulf Kroehne, Frank Goldhammer, Cornelia Schoor, Nina Mahlow, and Cordula Artelt. 2019. Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology* 89, 3 (2019), 524–537.
- [14] Trey Hedden, Gary Lautenschlager, and Denise C Park. 2005. Contributions of processing ability and knowledge to verbal memory tasks across the adult life-span. *The Quarterly Journal of Experimental Psychology Section A* 58, 1 (2005), 169–190.
- [15] Stephen Klein, Roger Benjamin, Richard Shavelson, and Roger Bolus. 2007. The collegiate learning assessment: Facts and fantasies. *Evaluation review* 31, 5 (2007), 415–439.
- [16] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872* 4 (2025).
- [17] Pia Kreijkes, Viktor Kewenig, Martina Kuvajla, Mina Lee, Jake M. Hofman, Sylvia Vitello, Abigail Sellen, Sean Rintel, Daniel G. Goldstein, David Rothschild, Lev Tankelevitch, and Tim Oates. 2026. Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers Education* 243 (2026), 105514. [doi:10.1016/j.compedu.2025.105514](https://doi.org/10.1016/j.compedu.2025.105514)
- [18] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [19] Byung Cheol Lee and Jaeyeon Chung. 2024. An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour* 8, 10 (2024), 1906–1914.
- [20] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.
- [21] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity?. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=Feiz5HtCD0>
- [22] Rita Payan-Carreira, Ana Sacau-Fontenla, Hugo Rebelo, Luis Sebastião, and Dimitris Pnevmatikos. 2022. Development and validation of a critical thinking assessment-scale short form. *Education sciences* 12, 12 (2022), 938.
- [23] Richard J Shavelson, Olga Zlatkin-Troitschanskaia, Klaus Beck, Susanne Schmidt, and Julian P Marino. 2019. Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing* 19, 4 (2019), 337–362.
- [24] Anjali Singh, Zhitong Guan, and Soo Young Rieh. 2025. Enhancing critical thinking in generative AI search with metacognitive prompts. *Proceedings of the Association for Information Science and Technology* 62, 1 (2025), 672–684.
- [25] Anjali Singh, Karan Taneja, Zhitong Guan, and Avijit Ghosh. 2025. Protecting human cognition in the age of AI. *arXiv preprint arXiv:2502.12447* (2025).
- [26] Lev Tankelevitch, Elena L Glassman, Jessica He, Aniket Kittur, Mina Lee, Srishti Palani, Advait Sarkar, Gonzalo Ramos, Yvonne Rogers, and Hari Subramonyam. 2025. Understanding, Protecting, and Augmenting Human Cognition with Generative AI: A Synthesis of the CHI 2025 Tools for Thought Workshop. *arXiv preprint arXiv:2508.21036* (2025).
- [27] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate chatbots to facilitate critical thinking on youtube: Social identity and conversational style make a difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.
- [28] Khonzoda Umarova, Talia Wise, Zhuoer Lyu, Mina Lee, and Qian Yang. 2025. How Problematic Writer-AI Interactions (Rather than Problematic AI) Hinder Writers' Idea Generation. *arXiv preprint arXiv:2503.11915* (2025).
- [29] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (2024), 2293–2303.
- [30] Jiayin Zhi, Harsh Kumar, and Mina Lee. 2026. Investigating the Effects of LLM Use on Critical Thinking Under Time Constraints: Access Timing and Time Availability. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (Upcoming) (CHI '26)*. ACM, New York, NY, USA, 1–21. [doi:10.1145/3772318.3791796](https://doi.org/10.1145/3772318.3791796)