

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

Jiayin Zhi, University of Chicago

Mina Lee, University of Chicago

Main theme(s): assessing and measuring outcomes / usage strategy / design strategy

Target domain(s): experiments, thinking, reasoning, measurements

Cognitive target(s): critical thinking, reasoning, creativity, learning

## Type of contribution & main idea

- Synthesis of how behavioral logs have been used for analytical purposes across different AI tools; a combined measurement approach that connects cognitive outcomes to the processes
- An AI tool for critical thinking tasks that captures interactions across document viewing, LLM interaction, and essay writing. Case study showing that improved task performance can coexist with reduced independent deliberation. Design implications for temporally-aware AI tools that adapt availability based on task stage and time constraints

## The AI tool's key characteristics

### AI Tool Characteristics:

- The AI tool is an LLM-powered chatbot embedded alongside a document viewer and a text editor in a three-panel interface
- The chatbot can answer questions about all provided documents, summarize content, verify ideas, and discuss arguments, but does not directly insert text into the essay
- The interface logs all interactions across panel: document opens and viewing durations, chatbot prompts and responses, and essay writing activity

### Design motivation:

- The three-panel design was motivated by the need to capture behavioral signals across the non-linear process of critical thinking: reading and navigating source documents, switch between sources to analyze trade-off, and communicating the reasoning
- LLM access timing was manipulated as a design variable (Early, Continuous, Late, No access) to investigate how the temporal availability of AI shapes cognitive engagement

### Task characteristics:

- The task is a critical thinking performance assessment: making a reasoned decision on a civic scenario based on diverse documents and writing an argumentative essay explaining the reasoning
- The task is goal-oriented (produce a reasoned essay) but open-ended (no single correct answer), requiring participants to weigh trade-offs, check their own bias, and synthesize arguments from multiple sources

## What you would like to discuss

How to design AI interfaces that capture informative behavioral logs without altering natural behavior. How to choose the right granularity of log-based metrics for different cognitive constructs. Whether temporally-aware AI tools can work in practice beyond controlled experiments. How others have combined task outcomes with process indicators. Approaches for validating log-based metrics as cognitive indicators.

## Interaction & outcomes

How people use and interact with the AI tool for a specific activity (first column) and what outcomes this provides, enables, or leads to (second and third column).

## Interaction design/usage strategy

The interface presents three side-by-side panels: a document viewer, an LLM chatbot, and a writing editor. Participants freely move between reading, querying, and writing. The chatbot answers questions about documents but does not insert text into the essay. LLM access timing varies by condition, shaping when AI enters the task process.

## Intermediary task outcomes

Material: evolving essay drafts and LLM chatbot conversation history. Cognitive: developing understanding of source documents, forming arguments, evaluating sources. Behavioral logs capture these processes—document viewing patterns, essay development, and LLM chatbot usage.

## Final task outcomes

Material: completed argumentative essay. Outcome measure: Essay score and Myside Bias. Key insight: having LLM access from the start improved Essay scores under time pressure, but logs revealed reduced document engagement and constrained deliberation—final outcomes alone miss concerning process patterns.

## Objective value

The 'objective' quality of the process and the outcomes—i.e. what makes them 'good' or desirable (e.g. critical thinking, understanding, new insights, learning).

AI's impact on critical thinking performance is neither uniformly positive nor negative—early access helps under time pressure but can impair it with sufficient time. When the LLM becomes available shapes how participants approach the task. Early access leads to consulting the chatbot before deeply engaging with documents, anchoring subsequent reasoning around the LLM's framing. Late or no access encourages independent engagement first.

Intermediary outcomes reveal the quality of the cognitive process, not just the final product. These are assessed through log-based metrics: aggregated counts (total documents viewed), phase-segmented counts (documents viewed during writing), and content overlap analysis (essay arguments vs. LLM responses). Quality is determined by breadth of source engagement, independence of argumentation, and iterative source consultation.

Essay score reflects argument quality; Myside Bias reflects whether multiple perspectives were considered. These are assessed through established scoring rubrics. Crucially, improved scores can coexist with reduced deliberation—combining final outcomes with process indicators is essential.

## Perceived value & UX

How the process of using the AI tool and its outcomes is experienced by users (e.g. how they experience the cognitive work they are doing).

Participants are unlikely to perceive the impact of LLM access timing on their thinking. Self-reported critical thinking showed minimal variation across conditions despite substantial performance differences. Participants with early access experienced the chatbot as helpful without noticing it shaped their subsequent reasoning.

Process shifts are largely invisible to participants. Those with early LLM access are unlikely to notice they viewed fewer documents or generated fewer independent arguments—the chatbot efficiently provided useful information, making independent engagement feel unnecessary.

Participants are unlikely to attribute essay quality differences to LLM access timing. Under time pressure, early access genuinely helps. With sufficient time, impairment goes unnoticed. This disconnect between perceived and actual cognitive impact underscores the necessity of performance assessment and behavioral logs over self-reports.

## How do people's goals interact with the AI tool's goals? Is there a tension?

**People's short-term goals:** Complete the critical thinking task effectively within the available time—read the documents, form a position, and write a well-argued essay. When an LLM is available, participants naturally use it to accomplish this efficiently: summarizing documents, verifying ideas, or checking arguments.

**People's long-term goals:** Develop and maintain critical thinking abilities—evaluating sources independently, considering multiple perspectives, and forming reasoned judgments without reliance on external tools. **Tensions:** The core tension is that the LLM helps achieve short-term task goals (especially under time pressure) through exactly the mechanisms that may undermine long-term cognitive goals. Participants are optimizing for task completion, not for practicing independent reasoning. This tension is sharpest when time is sufficient—the LLM still feels helpful, but the cognitive shortcutting it enables becomes counterproductive because participants have enough time to engage deeply on their own. Designing AI tools that navigate this tension—supporting task efficiency while preserving opportunities for independent cognitive engagement—is a key challenge.

## How do you expect people to continue using this AI tool?

Users may abandon the tool if LLM restrictions feel unnecessarily limiting, especially since cognitive benefits of delayed access are invisible to them. Sustaining engagement requires making the rationale transparent, adapting restrictions to context rather than applying them rigidly, and giving users some agency. The core challenge: tools that protect cognition create friction that conflicts with users' preference for efficiency.

## What would you like to take away from the workshop?

Feedback on our measurement approach from researchers working on other cognitive constructs. New ideas for log-based metrics from other IJT research practitioners. Connections with researchers interested in collaborating on synthesizing behavioral indicators across different AI tools and cognitive tasks.

## Key references (e.g. of main theories, empirical evidence, measurement methods etc.)

- [1] Zhi, J., Kumac, H., & Lee, M. (upcoming). 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 (CHI '26), ACM.
- [2] Collins-Thompson, K., Rieh, S. Y., Haynes, C. C., & Syrel, R. (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.
- [3] Lee, M., Liang, P., & Yang, Q. (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.
- [4] Kneijves, P., Kewenig, V., Kowaja, M., Lee, M., Hofman, J. M., et al. (2025). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. Computers & Education, 243, 105514.
- [5] Umarova, K., Wise, T., Lpu, Z., Lee, M., & Yang, Q. (2025). How Problematic Writer-AI Interactions (Rather than Reading AI) Hinder Writers' Effectiveness. arXiv preprint arXiv:2503.11915.

## How to proceed with this work/idea?

We plan to apply the combined approach to other cognitive constructs. We aim to design more log-based metrics that better capture process-level patterns, moving beyond the aggregation-segmentation spectrum. We also plan to explore how interface design choices affect the interpretability of captured logs. Based on workshop feedback, we hope to refine the space (computed vs. coded & aggregated vs. segmented) with examples from other researchers' work. Also, we would love to explore how temporally-aware AI designs might translate from controlled experiments to real-world tools.