

Bridging the AI-Memory Gap: Exploring Salient Cues to Support Users' Source Monitoring when Writing with AI

Tim Zindulka

tim.zindulka@uni-bayreuth.de
University of Bayreuth
Bayreuth, Germany

Sven Goller

sven.goller@uni-bayreuth.de
University of Bayreuth
Bayreuth, Germany

Daniel Buschek

daniel.buschek@uni-bayreuth.de
University of Bayreuth
Bayreuth, Germany

Abstract

Accurately remembering whether one's work was produced alone or in collaboration with AI is a key cognitive ability that helps people assess their own capabilities, calibrate confidence, and indicate ownership. While some contemporary AI writing assistants track contributions (e.g., via persistent prompting histories), they largely fail to support users' memory in the actual process of writing. Moreover, these systems are designed to arrive at a polished text as quickly and seamlessly as possible, thereby foregoing memory cues that would otherwise be generated during writing.

In this paper, we propose embedding salient cues into intelligent writing interfaces to help people better internalise the source of contributions when co-creating with AI. To this end, we conceptually transfer several cues from the literature on source memory to the new context of AI writing assistance. Finally, we discuss broader implications for AI-assisted writing, and the trade-off between designing AI to support the process rather than focusing on the outcome.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Text input**; • **Computing methodologies** → **Natural language processing**.

Keywords

Writing assistance, Large language models, Human-AI interaction, Memory, Source Memory, Cognition

ACM Reference Format:

Tim Zindulka, Sven Goller, and Daniel Buschek. 2026. Bridging the AI-Memory Gap: Exploring Salient Cues to Support Users' Source Monitoring when Writing with AI. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Barcelona, Spain. ACM, New York, NY, USA, 5 pages. <https://doi.org/XX.XXXX/XXXXXXXX.XXXXXXX>

1 Introduction

Writing enables people to concretise, store, and communicate otherwise transient thoughts. Beyond externalising, writing is also a tool for sensemaking and learning as it promotes durable memory traces through generative processing and elaboration. As such, *writing is thinking* and is pivotal to many cognitive processes, including reasoning, comprehension, and memory formation.

Today, writers increasingly integrate Large Language Models (LLMs) into their workflows for tasks such as polishing drafts, rephrasing sentences, and generating ideas. Prior work suggests that such assistance can improve efficiency and lower barriers to production (e.g., see [3, 11, 13]). At the same time, the growing prevalence of AI-supported writing raises questions concerning authorship, agency, and the metacognitive demands and potential risks associated with LLM use [16]. Indeed, emerging evidence suggests that AI can induce false recollections [14], and shift opinions [7]. These findings motivate closer examination of the cognitive benefits users may forgo when outsourcing parts of writing to AI systems. They also contribute to recurring debates about AI's role in human cognition and how cognitive processes can be protected and augmented [15, 17].

In this short paper, we focus on memory – specifically source memory, which is an integral part of cognition. Source memory describes *where*, *when*, and *how* a given item was acquired [8, 12]. In the context of writing with AI, source misattribution may affect perceived authorship, identity, learning, accountability, and trust. Users may misremember AI-generated errors as their own, attribute misplaced reliability to AI content recalled as self-produced, or forfeit deserved credit by mislabelling their own ideas as AI-generated.

Although intelligent writing tools can log contribution histories (cf. [6]), supporting users' memory directly may be more consequential. Memory shapes cognitive processes, whereas external records require an explicit lookup. Once misattributions are integrated into memory, users may also be unlikely to verify them due to confidence in their own memory (cf. [20]).

In recent empirical work, we examined how AI assistance affects source memory [20]. In a pre-registered experiment, participants generated and elaborated on ideas either unaided or with a chatbot. One week later, they were asked to attribute each idea and text to its source (noAI vs. withAI). Our results revealed a systematic source memory gap: After AI use, the odds of correct attribution decreased significantly. The decline was most pronounced in mixed human-AI workflows, where either the idea or the elaboration had been generated with AI.

Building on this prior work, we identify the absence of *salient memory cues* as a driving factor for the observed source memory gap. The Source Monitoring Framework (SMF) [8, 9] posits that such cues are essential for accurate source attribution, yet current AI writing systems often minimise them in favour of arriving at a polished text as quickly and seamlessly as possible. We next conceptually apply the SMF to the design of AI-supported writing systems and examine opportunities to (re)introduce these missing cues to preserve source memory when writing with AI as a “tool for thought.”



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).

ACM ISBN XXX-X-XXXX-XXXX-X/XX/XX

<https://doi.org/XX.XXXX/XXXXXXXX.XXXXXXX>

2 Designing Salient Memory Cues for AI-supported Writing

The Source Monitoring Framework is the dominant theoretical account for source memory. It describes how people infer the origin of mental contents by evaluating qualitative characteristics of the memory trace [8, 9]. These *cues* are categorised as sensory/perceptual information, contextual (spatial and temporal) information, semantic detail, affect, and cognitive operations. When they lack distinctiveness, individuals rely on heuristic cues, which increases the likelihood of misattribution.

In AI-supported writing, these cues are often weak or missing, making it more likely for source confusion to occur [20]. We therefore propose to explore how salient cues can be deliberately integrated through interface and interaction design to facilitate more accurate source memory. Note that the proposed design ideas can overlap, contradict, or complement each other and might not fit into one specific category only.

2.1 Sensory & Perception

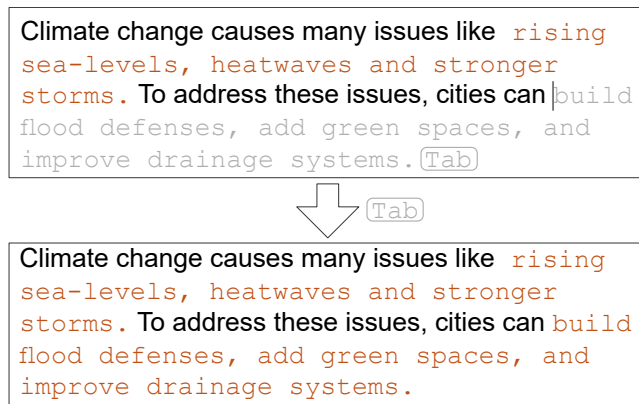


Figure 1: Sensory and perceptual cues could be resurfaced by highlighting AI-generated content via distinct fonts or colours. Here, accepting an AI suggestion remains visible in the UI.

Sensory and perceptual cues refer to modality-specific surface features of a memory trace, such as visual appearance (e.g., colours, shapes), sounds (e.g., voices, accents) or linguistic style (e.g., tone, phrasing). For example, a writer may recall drafting the bullet points for a paragraph on a napkin in blue ink.

In AI-supported writing, such cues are often weak because AI-generated text is typically rendered identically to user-authored text, that is, in the same interface, font, and style, with minimal visual distinctions between manual user input, prompts, and system output. In many prevalent interfaces (e.g., chatbots such as ChatGPT), the token-streaming animation can even mimic human typing behaviour [19], rendering both the generative process and the final product *perceptually homogeneous* (see also [2, 10]). Additionally, language generation models are *designed to align* their linguistic style closely with user input, further reducing distinctiveness.

We propose three concrete opportunities to resurface sensory and perceptual memory cues during interaction with generative AI

for writing: First, human and AI contributions could be designed to remain visually distinguishable throughout the writing process. For example, AI-generated text inserted into a document could be marked using a distinct colour, font, or other styling options (Figure 1). If such text is later revised by the user, or if only AI-generated ideas (e.g., from an ideation phase) are incorporated into the final draft, annotations could persist to signal the origin of these contributions across stages of writing. These markers could either remain visible in the final document to provide transparency and context, or be toggled on and off, appearing only in an “editor view” during active writing, similar to comment or markup modes in LaTeX editors.

Second, additional sensory cues may be introduced at specific *moments of interaction*. For instance, inserting AI-generated text into a document could be accompanied by a brief auditory signal (e.g., a jingle), or a subtle but unique animation reinforcing the perceptual distinctiveness of the event.

Third, when AI output is not intended for direct inclusion in the final text (e.g., during non-diegetic interactions such as prompting or ideation), the system could adopt a deliberately distinct linguistic tone, to preserve linguistic cues without interfering with the user’s stylistic goals in the final document.

2.2 Contextual

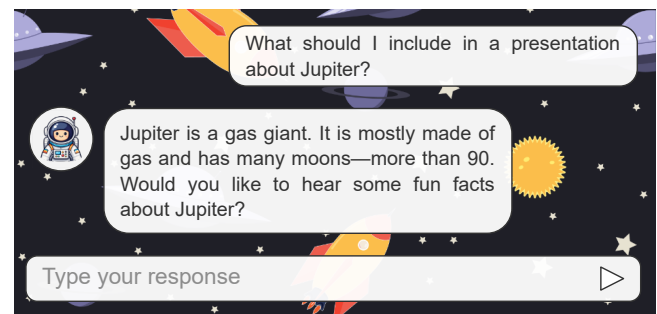


Figure 2: Enriching the often sterile user interfaces of AI tools with varying environments could help users distinguish between conversations better. Themes could either be selected randomly, chosen by the user or adapt to the topic of the conversation as shown here.

Contextual cues capture spatial and temporal information to describe under which circumstances content was produced. For example, one might come up with an idea while walking down a specific street, draft a section of text in a busy coffee shop, or notice a black cat sitting nearby immediately after completing a paragraph. Even if incidental, such environmental details can later serve as anchors for source monitoring.

This is not considered in design: UIs for generative AI typically use one consistent layout, typography, and interaction pattern across tasks and sessions.

To improve the opportunity for contextual cues in this UI environment, AI writing interfaces could deliberately vary across interactions. For instance, shifts in visual themes, layout structure,

AI avatars or ambient interface elements could enrich distinct conversations without interfering with the workflow (Figure 2).

2.3 Semantic Detail



Figure 3: To resurface semantic detail, intelligent writing interfaces could present the generated content in varying levels of semantic abstraction, for example as an illustration as shown here. These cues could be shown when the text is generated, the moment it is accepted, when hovering over the AI generated content or persistently in the UI.

Semantic detail refers to the conceptual meaning and factual content of information, independent of perceptual or contextual features. For example, one might remember the central argument of a paragraph, a specific claim that was made, or the logical structure of an explanation, without remembering every single sentence or word in their own writing.

AI-assisted writing may attenuate the encoding of semantic detail, as LLMs typically generate polished prose rather than abstract representations that users would otherwise need to articulate themselves. Since people generally think in concepts rather than fully formed sentences, bypassing this translation step (cf. [4]) may reduce deeper semantic processing and bypass people “making sense” of their thoughts and writing.

While the full cognitive benefits of translating abstract thought into written language may not be easily replicated, semantic detail could at least be made more explicit and perceptually salient. For example, an AI-infused writing interface might extract and present the core meaning of generated text in condensed form, such as a brief summary or illustration (Figure 3). These cues could be introduced at multiple stages of the process, either directly at the moment of generation, when the content is pasted into the draft document, or at any point in between.

2.4 Affect

Affect refers to the affective tone and intensity associated with the information and its generation. For example, text written during moments of happiness or used as a way of coping with strong emotions often carries affective cues that become embedded in the memory trace.

In AI-supported writing, affective cues are often attenuated because generated text is typically emotionally neutral and lacks personal markers of engagement.

Here, we deliberately limit the discussion to professional or informational contexts and do not address emotionally expressive domains such as diary writing or affective computing applications. Within this scope, we identify one actionable design lever: friction. Generative systems may reduce feelings of being stuck and mitigate writer’s block [5]. However, research on desirable difficulty indicates that effortful struggle can support learning [1]. For tools for thought, it remains an open research question of how to determine the appropriate level of friction to benefit from AI without removing productive, valued “struggle” (cf. [18]).

2.5 Cognitive Operations

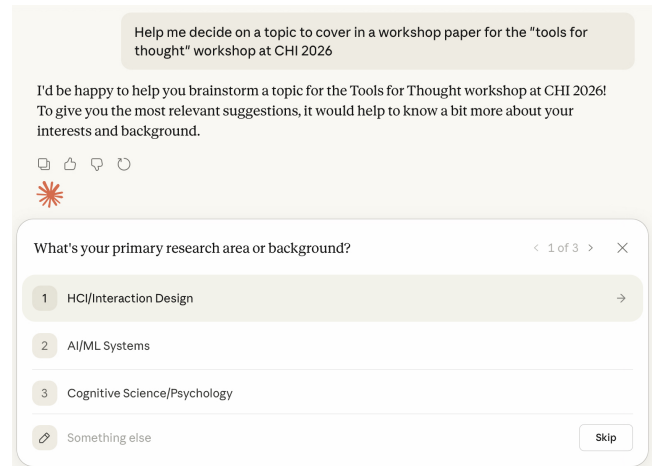


Figure 4: Industry example: Claude AI (claude.ai) features an interface element that requires cognitive operations from the user.

Cognitive operation cues capture evidence of mental effort involved in generating the memory. For example, after writing a CHI paper, an author might remember the process of “trying to figure it out” or making sense of the findings.

Many AI systems are designed to provide fast and seamless outputs that bypass the user’s own cognitive effort. As a result, content may appear without the intermediate cognitive work that supports source memory.

We propose two interaction opportunities to resurface cognitive operation cues: First, a human-in-the-loop approach to the “text generation” step could deliberately engage users at cognitively meaningful decision points. For instance, some contemporary conversational interfaces already request clarification when prompts are ambiguous (Figure 4). Also see the mockup by Tankelevitch et al. [16]. Building on this, future systems could involve users at moments of high cognitive value, for example when translating abstract thoughts into text (cf. Section 2.3), or at user-defined check-points similar to a “debug mode” in programming (e.g. expressed as meta-prompts such as “Stop and ask whenever you need to generate a character name”).

Second, cognitive operation cues could be supported by making generative processes more transparent. Some reasoning-oriented models already partially share their “thinking” process with users.

This provides a concrete starting point in the UI design to introduce such cues.

3 Discussion

We employed the SMF as a theoretical lens for the design of AI-assisted writing systems that support source memory. Although we described salient cues for writing with generative AI, they may be applicable more broadly to interaction contexts that integrate externally provided and internally generated content, particularly in distributed cognitive work, where they can strengthen source monitoring over time. Next, we discuss further opportunities for extending this line of research and reflect on challenges and limitations.

3.1 Opportunities

Theory Transfer as a Design Strategy. Our work illustrates that established cognitive frameworks can be systematically translated into concrete interaction design decisions for co-creative AI systems. To this end, we applied the SMF to AI-assisted writing by identifying how current interactions often lack salient cues that support source memory, and by outlining opportunities to deliberately reintroduce such cues in order to strengthen source memory.

More broadly, revisiting additional principles from cognitive psychology may prove valuable for the emerging context of writing and thinking with AI, and beyond.

Designing for Cognition Without Sacrificing Efficiency. Importantly, many of the identified salient cues can be implemented with minimal disruption to existing designs. While some might create (desired) friction, this highlights that designing for cognition does not necessarily mean sacrificing efficiency-oriented UI designs, which for many people is a core motivation for using generative AI. In some cases, memory cues may even enhance efficiency by reducing the need to revisit chat histories or reconstruct authorship and decision processes. Designing for cognition can therefore complement productivity goals; however, empirical research is needed to determine which design choices translate to measurable improvements.

Enriching Interaction with Generative AI Beyond Memory. While we focused on source memory, enriching interaction with generative AI systems with salient cues might bring broader benefits. First, such cues might improve metacognitive awareness by making users think more explicitly about when and why they rely on AI assistance. Second, they might improve interpretability by surfacing contextual information about how content was generated (e.g., what was the author thinking or feeling during creation). Additionally, by embedding salient cues into the user’s memory trace, such designs may facilitate reflection beyond the moment of interaction.

Taken together, future work could examine how such cues that foreground the process might contribute to more transparency, reflective, and cognitively engaged human-AI interaction, beyond providing source memory support.

Operationalising and Measuring Cognitive Effects. A recurring question concerns how cognition can be meaningfully measured in human-AI interaction. In prior work, we examined users’ ability to remember whether ideas and elaborations were generated

independently or in collaboration with AI, observing a significant impact of AI use on source memory [20]. Since source memory is an integral component of cognition, measuring it as a “cognitive byproduct” may support evaluations of “Tools for Thought”: Even when supporting memory is not the primary design objective, differences in source memory performance may reveal how a system shapes cognition. For example, a study comparing two tool designs for a given task could assess and contrast users’ source memory to better understand how each design influences cognitive processes.

3.2 Challenges & Limitations

While the proposed design directions are promising, they also raise important challenges.

Distraction and Cognitive Overload. Salient cues may introduce additional visual or cognitive complexity. Particularly in cognitively demanding writing tasks, additional signals may compete for attention rather than support encoding. Careful consideration of modality, granularity, and timing is therefore required to ensure that cues remain supportive rather than intrusive.

Generating False Memory Traces. Prior work has shown that AI systems can induce false memories [14]. Embedding salient cues into human-AI interactions may similarly reinforce inaccurate source attributions, as generative AI can produce rich memory cues that, if not carefully designed, might feel self-generated. Design choices that amplify contextual or perceptual cues therefore require careful calibration. For example, generating images to add semantic detail (cf. Section 2.3) may increase self-attribution, as a plausible visual trace could feel like an experienced memory.

Empirical Validation and Generalisability. Finally, the effectiveness and tradeoffs of embedding salient cues must be empirically evaluated across cue types, interaction timings, writing tasks and domains, and longitudinal use. As some cues can be viewed both as an opportunity (Section 3.1) and a limitation (Section 3.2), more research is needed to determine whether such interventions produce durable cognitive benefits without unintended side effects.

4 Conclusion

In this workshop paper, we explore how salient cues can enrich the co-creative process of writing with AI, to support users’ memory of what they created with or without AI. Transferring concepts from the Source Monitoring Framework, we identify concrete opportunities for designing mnemonically supportive AI-infused writing interfaces. However, further research is needed to evaluate the effectiveness and potential trade-offs of embedding memory traces with (multiple) such cues, introduced at different moments of interaction and across diverse writing tasks. Future work should also examine the broader societal implications of memory adaptations associated with the use of generative AI tools for writing and thinking, including their effects on authorship, accountability, and the development of cognitive skills.

Acknowledgments

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 525037874.

References

- [1] Robert Bjork and Elizabeth Bjork. 2020. Desirable Difficulties in Theory and Practice. *Journal of Applied Research in Memory and Cognition* 9 (12 2020), 475–479. <https://doi.org/10.1016/j.jarmac.2020.09.003>
- [2] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. <https://doi.org/10.1145/3411764.3445372>
- [3] Allison E. Connell Pensky, Jordan H. Usdan, and Harley Chang. 2025. Generative AI's Impact on Graduate Student Professional Writing Productivity and Quality. *International Journal of Artificial Intelligence in Education* 35, 6 (2025), 4057–4082. <https://doi.org/10.1007/s40593-025-00528-z>
- [4] Linda Flower and John Hayes. 2004. A Cognitive Process Theory of Writing. *College Composition and Communication* 32 (01 2004). <https://doi.org/10.2307/356600>
- [5] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social Dynamics of AI Support in Creative Writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 245, 15 pages. <https://doi.org/10.1145/3544548.3580782>
- [6] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D. Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark Effect: Supporting Provenance and Transparent Use of Large Language Models in Writing with Interactive Visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1045, 15 pages. <https://doi.org/10.1145/3613904.3641895>
- [7] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. <https://doi.org/10.1145/3544548.3581196>
- [8] M. K. Johnson, S. Hashtroudi, and D. S. Lindsay. 1993. Source monitoring. *Psychological Bulletin* 114, 1 (July 1993), 3–28. <https://doi.org/10.1037/0033-2909.114.1.3> Review; Research Support, U.S. Gov't, P.H.S..
- [9] Marcia K. Johnson and Carol L. Raye. 1981. Reality monitoring. *Psychological Review* 88, 1 (1981), 67–85. <https://doi.org/10.1037/0033-295X.88.1.67>
- [10] 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* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. <https://doi.org/10.1145/3491102.3502030>
- [11] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 1048, 25 pages. <https://doi.org/10.1145/3613904.3642625>
- [12] Karen J Mitchell and Marcia K Johnson. 2009. Source monitoring 15 years later: what have we learned from fMRI about the neural mechanisms of source memory? *Psychological bulletin* 135, 4 (2009), 638.
- [13] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. <https://doi.org/10.1126/science.adh2586> arXiv:<https://www.science.org/doi/pdf/10.1126/science.adh2586>
- [14] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 538, 20 pages. <https://doi.org/10.1145/3706598.3713697>
- [15] 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* (August 2025). <https://www.microsoft.com/en-us/research/publication/understanding-protecting-and-augmenting-human-cognition-with-generative-ai-a-synthesis-of-the-chi-2025-tools-for-thought-workshop/>
- [16] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 680, 24 pages. <https://doi.org/10.1145/3613904.3642902>
- [17] Zelun Tony Zhang, Nick von Felten, Leon Reicherts, Lev Tankelevitch, Zhitong Guan, Sean Rintel, Yue Fu, Jessica He, Kenneth Holstein, Advait Sarkar, Gonzalo Ramos, Anuschka Schmitt, Anjali Singh, Haotian Li, Srishti Palani, and Peter Dalsgaard. 2026. Tools for Thought: Understanding, Protecting, and Augmenting Human Cognition with Generative AI—From Vision to Implementation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (CHI EA '26) (Barcelona, Spain). ACM. <https://doi.org/10.1145/3772363.3778733>
- [18] D. Zhou and Sarah Sterman. 2023. Creative struggle: Arguing for the value of difficulty in supporting ownership and self-expression in creative writing. In *Proceedings of the Second Workshop on Intelligent and Interactive Writing Assistants (In2Writing)*.
- [19] Jijie Zhou and Yuhuan Hu. 2024. Beyond Words: Infusing Conversational Agents with Human-like Typing Behaviors. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) (CUI '24). Association for Computing Machinery, New York, NY, USA, Article 24, 12 pages. <https://doi.org/10.1145/3640794.3665560>
- [20] Tim Zindulka, Sven Goller, Daniela Fernandes, Robin Welsch, and Daniel Buschek. 2026. The AI Memory Gap: Users Misremember What They Created With AI or Without. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Barcelona, Spain) (CHI '26). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3772318.3791494>