

Felt but Not Seen: Design Patterns from Building a Metacognitive Writing Tool

Sergio Abraham, Auster Center for Applied Innovation and Research, Tufts University
Filip Čučkov, Auster Center for Applied Innovation and Research, Tufts University

Main theme(s): Design strategy / Adoption

Target domain(s): Professional writing, knowledge work

Cognitive 'target'(s): Metacognition, critical thinking, cognitive ownership

Type of contribution & main idea

Type: Proposal of a new approach to design AI tools (i.e. a JTJ)

Main idea: Two design patterns for cognitive-state-responsive AI: presence without visibility (behavioral observation, imperceptible until engaged) and moments over conversations (atomic interactions, invisible context). AI as metacognitive extension: felt but not seen.

The AI tool's key characteristics

Context: Professional and creative writing (essays, reports, arguments). Tasks are goal-oriented but cognitively ill-defined: the writer discovers what they think through writing. Users are people who write to think, not just to produce output.

Role: Metacognitive extension. Not a thought partner (does not co-create), not an assistant (generates nothing), not a provocateur (does not argue). It observes how the user writes (hesitation, deletion, avoidance) and surfaces reflections that trigger the user's own thinking.

Thinking supported: Metacognition, critical self-evaluation, cognitive ownership. Constrained by design: atomic interactions (3 turns max), ambient entry points (not notifications), one-message limit, invisible context accumulation.

Rationale: Our early prototypes improved text quality while reducing cognitive ownership. These constraints are deliberate: less engagement, less generation, less visibility. Protecting the thinking process over perceived productivity.

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).

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).

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).

Interaction design/usage strategy

The proposed design departs from conversational AI. Instead of analyzing text, the AI observes behavioral writing patterns (hesitation, deletion, avoidance) and surfaces reflections through ambient entry points designed as environmental states, not notifications. Each interaction is a constrained "moment" AI users respond once, AI closes). No threading, no visible history. The user's cognitive rhythm determines when moments occur.

Well-timed reflections could lead to more self-aware writing: pausing to evaluate reasoning rather than pushing through on autopilot. Atomic interactions prevent offloading onto extended AI dialogue. Could be assessed by comparing self-explanation quality and revision depth with and without the tool. Could be negative if reflections interrupt flow at the wrong moment.

Writers who value their thinking process experience moments as welcome interruptions. Those seeking productivity experience them as friction. Could be assessed by engagement patterns (respond vs. dismiss) and post-session self-reports on perceived value.

Intermediary task outcomes

Cognitive outcomes: the writer becomes aware of their own cognitive state (e.g., recognizing they are avoiding a section, or that an argument isn't clear to them yet). This metacognitive awareness is the primary intermediary outcome.

Material outcomes: revised arguments, restructured sections, articulated reasoning. All material outcomes are produced entirely by the writer; not the AI. The AI produces no artifacts; it triggers the cognitive events that lead the writer to produce their own.

Metacognitive awareness (recognizing one's own cognitive state) is a precursor to self-regulated thinking. Reflections that trigger this awareness are valuable because they are the user's own insights, not the AI's, so they integrate into understanding rather than sitting alongside it as external input. Could be assessed through think-aloud protocols measuring frequency and depth of self-monitoring statements during writing.

Users may not recognize metacognitive awareness as a valuable intermediary outcome. A writer who pauses to reconsider an argument rarely thinks "the tool just helped me." Value is clearest when a reflection resonates with knowledge the writer was already sensing but hadn't articulated: recognition, not instruction. Under deadline pressure, even valuable reflections may feel like unwelcome interruptions. When reflections miss, cost is low: the moment is brief and leaves no residue.

Final task outcomes

Material outcome: a written artifact (essay, report, argument) produced entirely by the writer. Cognitive outcome: the writer can fully explain, defend, and build upon every choice in the text. This cognitive ownership is the primary quality metric, not output polish. The AI contributed nothing to the material outcome; its contribution was to the cognitive outcome through the intermediary metacognitive events.

Cognitive ownership: can the writer explain, defend, and build upon every element? Self-produced reasoning is better retained and understood than passively received content. Could be assessed through self-explanation tasks, argument defense interviews, or delayed recall tests. Standard text quality metrics (coherence, argumentation) apply but are insufficient alone for measuring cognitive outcomes.

The tool produced nothing visible, so users may not attribute improved understanding to it. Whether it felt "worth it" depends on whether the writer values the cognitive work itself or only the artifact. Writers who can defend every choice in their text may recognize that value retrospectively, but only after repeated use with and without the tool.

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

(1) Short-term: produce a written artifact within a deadline. Users may ask the AI to help them write faster or better, but the tool does neither directly.

(2) Long-term: develop as a thinker, maintain authentic voice, build expertise that goes beyond any single text.

(3) These goals conflict. Writers under deadline pressure need to reconcile the tool's cognitive friction (no generation, no shortcuts, constrained interactions) with time constraints.

The tool bets that the cognitive investment pays off in deeper understanding, but this is hardest to accept precisely when the short-term pressure is highest. Users who approach writing primarily as production may find the tool frustrating. Users who also value learning through writing may tolerate the friction because it aligns with their long-term goals. Users expect AI to reduce effort. This tool redistributes it: less effort managing AI interaction, more effort on own thinking. Users who value their thinking process engage readily. Those seeking efficiency find it frustrating. Equity implication: may serve cognitively privileged users best.

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

Users may abandon the tool early if their first reflection is poorly timed or irrelevant, since the tool has no visible features to demonstrate value otherwise. The configurable entry points (visual, auditory, invocation) help: users who find ambient signals intrusive can switch to self-initiated invocation, reducing friction without losing the tool. Long-term retention may be stronger than initial adoption because the tool improves as context accumulates, and users who stay long enough to experience deeper reflections are more likely to value the cognitive investment.

What you would like to discuss

How to detect cognitive patterns from how people interact with digital devices. What behavioral signals are reliable across domains, and what sensing approaches exist beyond the writing-specific signals we describe?

Whether augmented cognition is something the general population would seek out, or whether it is inherently niche. Is cognitive augmentation a viable design direction at scale?

What would you like to take away from the workshop?

Finding collaborators with expertise in measuring cognitive outcomes and in behavioral signal detection across domains. Getting feedback on whether the design patterns resonate with researchers working on other JTJ contexts beyond writing. Potentially starting a collaborative project in the space.

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

[1] Engelbart, D. C. (1962). Augmenting Human Intellect: A Conceptual Framework. Stanford Research Institute. [2] Brynjolfsson, E. (2022). The Turing Trap: The Promise and Peril of Human-Like Artificial Intelligence. *Daedalus*, 151(2), 272-287. [3] Sarkar, A. (2023). AI Should Challenge, Not Obey. *Communications of the ACM*, 66(3), 30-32. [4] Tankelevitch, L., Glasman, E. L., He, J., Ritter, A., Lee, M., Palmi, S., Sarkar, A., Ramon, G., Rogers, Y., & Subramonyam, H. (2023). Understanding, Protecting, and Augmenting Human Cognition with Generative AI: A Synthesis of the CHI 2023 Tools for Thought Workshop. *CHI '23 Extended Abstracts*. [5] Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science*, 333(6043), 776-779. [6] Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, 32(4), 365-387. [7] Bakker, C., Houten, D., & Salkes, T. (2015). Peripheral Interaction: Characteristics and Considerations. *Personal and Ubiquitous Computing*, 19(1), 239-254. [8] Dool, A. R., & Houten, C. P. (2024). Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Concepts. *Science Advances*, 10(28). [9] Melamed, S., & Van, J. H. (2023). Experimental Evidence of the Effects of Large Language Models versus Web Search on Depth of Learning. *PLoS ONE*, 18(10), 1-15. [10] Sun, S., Li, A. Z., Foo, M. D., Zhou, J., & Lu, J. G. (2023). How and For Whom Using Generative AI Affects Creativity: A Field Experiment. *Journal of Applied Psychology*, 113(1), 1-15. [11] Shao, Y., Zope, H., Jiang, T., Pei, J., Nguyen, D., Brynjolfsson, E., & Yang, D. (2023). Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. *Workforce*. [arXiv:2306.06376](https://arxiv.org/abs/2306.06376).

How to proceed with this work/idea?

I want to move beyond my own prototyping experiments toward controlled group studies that can validate whether the design patterns produce measurable cognitive outcomes. I want to explore whether cognitive-state-responsive design applies beyond writing to other areas of thought (decision-making, learning, creative work), which would require identifying new behavioral signals and sensing approaches for each domain.