# Beyond Frictionless: Designing Productive Friction into Augmented Reading Interfaces

Anukriti Kumar
anukumar@uw.edu
University of Washington
Seattle, WA, USA

## Abstract

Current augmented reading systems largely optimize for speed and ease, yet cognitive science suggests that strategic difficulty ("desirable difficulties") can deepen comprehension and long-term retention. We argue that a critical gap remains in how augmented reading systems are evaluated: existing systems such as ScholarPhi, Scim, and Paper Plain [3, 14, 19] commonly report task time and subjective experience, while rarely measuring situation-model comprehension, metacognitive calibration, or delayed retention. This is a position paper: we do not report an implemented system or empirical results, but outline a design framework and evaluation agenda for studying deep comprehension in augmented reading. Drawing on Kintsch's Construction-Integration model and Bjork's desirable difficulties framework [5, 24], we propose three interface mechanisms that introduce *productive friction* in targeted moments: active recall prompts, elaborative interrogation, and generate-first interactions. We outline an evaluation blueprint that adapts validated comprehension instruments, adds metacognitive calibration measures, and includes delayed retention testing (e.g., one-week follow-up). We challenge the field's implicit assumption that frictionless reading optimizes learning, and argue for treating the speed-depth trade-off as an explicit, measurable design variable.

## CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; *Interactive systems and tools.*

## Keywords

reading interfaces, desirable difficulties, comprehension, situation model, productive friction

## 1 Introduction

Augmented reading interfaces have made remarkable progress helping researchers navigate, skim, and extract information from scientific papers. ScholarPhi provides just-in-time definitions for technical terms [19]. Scim uses AI-generated highlights to accelerate skimming [14], Paper Plain simplifies medical text for non-expert readers [3], and GP-TSM modulates word opacity to support multi-level reading [17]. These systems share a common design philosophy: reading should be faster and easier. Industry deployments from Adobe, Apple, and Google similarly emphasize instant access to summaries, simplified text, and effortless navigation [1, 2, 16].

But cognitive science reveals a fundamental tension between processing speed and depth of understanding. Meta-analyses show that screen reading, with its frictionless scrolling and instant access to information, often yields lower comprehension than paper reading for complex texts [11]. One explanation is that digital reading breeds metacognitive overconfidence: readers feel they understand material better than they actually do, creating an illusion of learning [7]. When tools provide instant answers and summaries, readers may become passive consumers rather than active thinkers.

Research on desirable difficulties demonstrates that conditions requiring more mental effort, including retrieval practice, generation tasks, and spaced repetition, produce stronger retention and deeper understanding than passive study, even when slowing initial performance [4, 5]. The very struggle that learners instinctively avoid is often what produces durable learning. A meta-analysis of pre-question effects found that answering questions before reading improved post-test performance with an average effect size of g = 0.54 across 97 studies, even when initial attempts were incorrect [35]. Productive failure research with 133 ninth-graders found that solving problems without instruction before receiving teacher-led consolidation significantly outperformed direct instruction on both well-structured and complex problems [21].

Yet current augmented reading tools have not fully explored this tension. Analysis of representative recent systems [3, 14, 17, 19] suggests a recurring pattern: studies rarely measure both textbase- and situation-model comprehension, include delayed retention tests, or assess metacognitive calibration. ScholarPhi (N=27) measured task time and self-reported ease [19]. Scim (N=31) measured information location speed [14]. Paper Plain (N=24) used a non-inferiority test showing comprehension was not worse, not that it was better [3]. GP-TSM (N=18) measured immediate GRE scores but not delayed retention [17].

Kintsch's Construction-Integration model distinguishes three levels of mental representation formed during reading, including surface form (exact wording), textbase (propositional meaning), and situation model (integrated mental model of the described situation), each with validated assessment methods and distinct forgetting curves [25]. Surface form vanishes within one hour, textbase drops around seven days, and situation models persist for months [13]. Critically, measures improving textbase retrieval can impair situation model construction [24, 28]. Yet augmented reading studies test immediately and rarely differentiate comprehension levels.

We argue that the augmented reading community should explore the design space of productive friction: interfaces that strategically slow readers down to enhance deep comprehension and long-term retention. This position paper makes four claims and outlines a corresponding research agenda. First, we synthesize Kintsch [25] comprehension levels with desirable difficulties [5], arguing that current interfaces optimize textbase retrieval at the expense of situation model construction. Second, we propose three concrete mechanisms for productive friction grounded in learning science with documented effect sizes: active recall prompts, elaborative

interrogation, and generate-first interactions. Third, we outline a comprehensive evaluation methodology combining multi-level comprehension instruments (including situation-model probes), metacognitive calibration, and delayed retention testing (e.g., one-week follow-up). Fourth, we articulate testable empirical predictions, including a crossover interaction where frictionless reading shows immediate advantages that reverse at delayed testing.

## 2 Theoretical Foundation and Prior Works

Our work builds on two foundational theories from cognitive science: Kintsch's three-level model of text comprehension and Bjork's framework of desirable difficulties. Understanding how these theories interact reveals why current reading interfaces may inadvertently undermine deep learning.

**Kintsch's Three-Level Model.** The Construction-Integration model defines three mental representations [25, 31]. The surface code captures verbatim wording, decaying to chance within one hour. The textbase represents propositional meaning, persisting roughly seven days. The situation model integrates text with prior knowledge, remaining robust for months [13]. Assessment uses a recognition paradigm: participants judge whether sentences appeared in the text using original, paraphrase, meaning-change, and situation-change probes [36]. Signal detection sensitivity (d') between probe types indexes each level.

Kintsch's critical insight is that "text memory" differs from "learning from text"; interventions improving one can impair the other [24]. Mannes and Kintsch [28] demonstrated that advance organizers matching text structure improved sentence verification but worsened inference and problem-solving, while mismatching organizers showed the reverse. Interfaces optimizing for easy information access may undermine the comprehension level that matters most for learning.

**Bjork's Desirable Difficulties.** The framework identifies mechanisms, such as spacing, interleaving, generation, and varying conditions, that slow initial performance while building durable knowledge [4, 5]. The generation effect is most relevant: producing answers before seeing them creates stronger memory than passive reading, even when the generated answers are wrong. Kornell et al. [27] showed that participants generating incorrect guesses (97% error rate) who then received feedback recalled targets more often than those given twice as much study time. Meta-analysis found g = 0.54 for pre-question effects [35]. Prior productive failure studies demonstrated that solving problems without instruction before receiving teaching outperformed direct instruction [21, 22].

Importantly, perceptual disfluency (hard-to-read fonts) is unreliable. Meta-analysis of 25 studies found d = -0.01 for recall and d = 0.03 for transfer, essentially null effects [39]. Cognitive generation (requiring active production) produces robust benefits while perceptual friction does not.

**Preference-Performance Paradox.** Users prefer worse performing systems approximately 25% of the time [33]. Bjork and Bjork [5] showed that participants performing better with interleaving "almost uniformly said blocking helped them learn better." Prior analysis of Cognitive Tutor data revealed students frequently used hints to obtain answers without learning [26]. User satisfaction ratings will favor frictionless designs, making objective comprehension measurement essential.

The closest prior work is iSTART, which teaches reading strategies through animated agents [29]. Learners generate self-explanations before receiving LSA-scored feedback. A study of 300 students found reliable advantages in science comprehension one week post-training [32]. However, iSTART is a training system wherein readers complete sessions and then return to normal text, rather than integrating friction into the reading interface itself.

## 3 Interface Design

Three potential mechanisms of inducing productive friction are proposed below:

**Active Recall Prompts.** The interface could pause readers at section endings with: "Before proceeding, summarize the key point of this section in your own words." Readers type a brief response (e.g., a few sentences) before the next section becomes visible. Responses are not graded during reading but could be scored post-hoc using semantic similarity methods (e.g., LSA-style approaches) for paraphrasing versus verbatim copying. This implements retrieval practice (g = 0.54) [35], strengthening memory consolidation [23] and building situation models by forcing integration [29]. We would predict higher situation-model scores at a 7-day delay despite potential immediate costs and improved metacognitive calibration.

**Elaborative Interrogation.** At high-importance claims (3-5 per paper), the interface highlights the sentence and asks: "Why is this true?" or "How does this connect to what you know?" Readers formulate explanations before optionally revealing AI-generated expert answers. We could use a similarity-based heuristic between reader and expert explanations to determine whether to show the answer immediately (threshold to be established in piloting) or encourage revision. Self-explanation prompts deeper processing than reading exposition [10], activating prior knowledge [38] and building causal models [9]. Meta-analysis confirms benefits for conceptual learning [12]. We predict higher scores on inference questions and transfer tasks.

**Generate-First, Reveal-Second.** When readers encounter unfamiliar terms (identified via domain glossaries or TF-IDF), the interface initially hides definitions. Readers click "Generate" and attempt to infer meaning from context. After an attempt (or a short tunable delay), "Reveal" becomes available to access AI-generated definitions. A "Skip" option logs usage. Comparison views show reader inferences alongside expert definitions. This leverages productive failure; attempting problems before instruction improves learning [22]. Brief struggle followed by confirmation triggers prediction error signals enhancing encoding [6] and makes information more memorable through surprise [27]. We predict higher delayed vocabulary retention and improved context-based inference skills.

These mechanisms share a structure: effortful cognitive processing before information support. This contrasts with current systems providing instant access. They should be mode-based, wherein deep-mode enables all prompts for learning goals; skim-mode minimizes friction for rapid information seeking.

## 4 Evaluation Methodology

As a plausible first study, we outline a within-subjects design with approximately 40-60 participants (final sample size to be determined via power analysis and piloting). Advanced undergraduate and graduate STEM students would read three matched papers from the same conference proceedings under three conditions: baseline (standard PDF), frictionless augmentation (instant AI summaries/highlights/definitions, similar to existing systems), and productive friction (our three mechanisms). Papers would be matched using Coh-Metrix readability indices [30], with Latin square counterbalancing ensuring each paper appears equally across conditions. Sessions would be separated by a minimum of 24 hours. The task instruction would be: "Read as you would if preparing to discuss with your research group. You'll be tested immediately and in one week."

Comprehension would be assessed at three levels. Surface recall would include verbatim sentence recognition and factual details. Textbase comprehension would be measured through paraphrase recognition [36], single-sentence inferences, and fill-in-the-blank items requiring propositional understanding. Situation model comprehension would be assessed using deep cloze tests requiring global inferences [20], cross-sentence inference questions, transfer tasks applying principles to novel scenarios, keyword sorting scored against expert clustering, and questions requiring prior knowledge integration. Automated scoring could optionally use semantic similarity methods (e.g., LSA-style approaches) to support scalable analysis of free-response explanations, following prior work linking such representations to comprehension outcomes [15], and scoring self-explanations for paraphrasing, bridging inferences, and elaborative inferences.

**Metacognitive Calibration.** After completing questions, participants would provide text-level Judgments of Learning (0-100 scale) and item-level confidence (1-5). Absolute calibration (bias) would be computed as the difference between predicted and actual performance. Relative calibration (resolution) would be measured as the Gamma correlation between confidence and accuracy. We predict productive friction improves both by forcing readers to confront understanding gaps.

**Delayed Retention.** One week after each session, participants complete surprise parallel tests. Based on Fisher and Radvansky [13], surface form should be at chance across conditions, textbase should show substantial forgetting, and situation model is the key differentiator. Critical prediction: crossover interaction where frictionless reading shows immediate advantage but productive friction shows delayed advantage.

**Secondary Measures.** Secondary measures would include time on task, scrolling patterns, re-reading detected via backward scrolls, and use of skip options. Subjective measures would include NASA-TLX cognitive load [18], perceived helpfulness ratings (Likert scales), and open-ended feedback. We would predict preference-performance dissociation: participants may prefer frictionless reading despite better learning with friction.

This evaluation agenda would differ from much prior work in four ways: (1) differentiated comprehension levels using validated instruments, (2) delayed testing at 7 days, (3) metacognitive calibration measurement, and (4) inclusion of established comprehension instruments (e.g., deep cloze) and optional automated analysis techniques for free-response data.

## 5 Expected Implications

**Hypotheses.** We articulate four testable hypotheses that future empirical work can evaluate. First, readers in the productive friction condition may score significantly higher on situation-model questions at delayed testing ($d > 0.50$), though potentially lower on immediate textbase questions due to time costs. This prediction follows directly from the generation effect and retrieval practice findings, which consistently show that initial performance costs are offset by superior delayed performance. Second, productive friction should yield more accurate judgments of learning, evidenced by smaller absolute bias (predicted minus actual performance closer to zero) and higher Gamma correlation between confidence and accuracy. This follows from the finding that attempting retrieval reveals knowledge gaps, improving metacognitive monitoring [37].

Third, the advantage of productive friction should grow with delay, producing a crossover interaction where frictionless reading shows higher immediate scores, but friction surpasses it at 7 days. This pattern is the signature of desirable difficulties, wherein immediate performance suffers while long-term retention improves [5]. The interaction should be especially pronounced for situation-model questions, as these are the measures most sensitive to durable learning. Fourth, despite showing better learning outcomes, participants should rate frictionless reading as more helpful and less effortful, demonstrating the preference-performance paradox documented by Nielsen and Levy [33].

**Design Implications.** Reading interfaces should support multiple modes aligned with goals: "quick scan" with maximal AI support versus "deep study" with strategic prompts. Eye-tracking can classify deep versus skim reading with 0.82 AUC [8], enabling real-time mode detection. Adaptive friction could adjust based on reader expertise, text difficulty (Coh-Metrix), and reading goal. Rather than hiding friction features, interfaces could present calibration data showing learning-effort trade-offs, enabling informed choice.

**Evaluation Implications.** Success metrics should expand beyond time-on-task and ease to include situation-model comprehension, delayed retention, metacognitive accuracy, and transfer. The field should move from non-inferiority tests to superiority tests with theoretically motivated measures, reporting effect sizes for meta-analysis.

**AI Augmentation Implications.** Current trends emphasize invisible, effortless AI [1, 2, 16, 34]. But if friction is productive, AI might be most valuable when it doesn't immediately answer questions, that is, when it guides sense-making rather than replacing it. This suggests AI-as-coach rather than AI-as-oracle, where systems scaffold active processing. Generate-first, reveal-second embodies this, as AI support is available but deliberately delayed.

## 6 Limitations and Future Work

Productive friction is not universal. For readers with cognitive or attentional differences, reducing friction may be essential for access. Future work should examine how neurodivergent readers respond, identify individual differences moderating effectiveness, and design adaptive systems calibrating friction to reader characteristics.

Our evaluation focuses on scientific papers, raising generalization questions about narrative texts, news, and documentation. Does friction help across text types, expertise levels, and reading goals? Longitudinal studies are needed to assess long-term skill transfer: do readers internalize strategies? How long do retention benefits persist beyond seven days?

Design parameters such as prompt frequency, delay duration, and any similarity-based heuristics are provisional and would require piloting. Optimal friction likely varies by reader, requiring adaptive adjustment. Our work focuses on individual reading, but much consequential reading occurs collaboratively. How does productive friction interact with collaborative sensemaking? These questions deserve further exploration.

## 7 Conclusion

As AI makes reading increasingly effortless, we face a choice: optimize for speed or for learning. This paper argues these goals are often in tension. Our theoretical synthesis reveals that current interfaces likely optimize textbase retrieval at the expense of situation model construction. We propose three mechanisms for productive friction (active recall prompts, elaborative interrogation, and generate-first interactions), each grounded in learning science with documented effect sizes. We outline an evaluation methodology measuring what matters: situation-model comprehension, delayed retention, and metacognitive calibration.

The broader implication is that success metrics should expand beyond time-on-task and ease. Speed is valuable, but not the primary goal when reading to learn. By treating the speed-depth trade-off as an explicit design variable rather than always optimizing for speed, we can create tools supporting the full range of reading goals.

## References


[1] Adobe Research. 2024. How Adobe Research is helping unlock the intelligence inside trillions of PDFs.

[2] Apple Inc. 2025. Apple Intelligence - AI for the rest of us.

[3] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Transactions on Computer-Human Interaction* 30, 5, Article 74 (2023), 38 pages. doi:10.1145/3589955

[4] Robert A. Bjork. 1994. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing*, Janet Metcalfe and Arthur Shimamura (Eds.). MIT Press, Cambridge, MA, 185–205.

[5] Robert A. Bjork and Elizabeth L. Bjork. 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 ed.). Worth Publishers, 56–64.

[6] Garvin Brod. 2021. Predicting as a learning strategy. *Psychonomic Bulletin & Review* 28 (2021), 1839–1847. doi:10.3758/s13423-021-01904-1

[7] Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and Instruction* 8, 4 (1991), 293–332.

[8] Yue Chen, Yifan Li, Xin Zhang, et al. 2023. Characteristics of deep and skim reading on smartphones vs. desktop: A comparative study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM.

[9] Michelene T. H. Chi. 2000. Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In *Advances in instructional psychology*, Robert Glaser (Ed.). Vol. 5. Lawrence Erlbaum Associates, Mahwah, NJ.

[10] Michelene T. H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13, 2 (1989), 145–182.

[11] Pablo Delgado, Cristina Vargas, Rakefet Ackerman, and Ladislao Salmerón. 2018. Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review* 25 (2018), 23–38. doi:10.1016/j.edurev.2018.09.003

[12] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest* 14, 1 (2013), 4–58.

[13] Jeffrey S. Fisher and Gabriel A. Radvansky. 2018. Patterns of forgetting. *Journal of Memory and Language* 102 (2018), 130–141. doi:10.1016/j.jml.2018.05.008

[14] Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S. Weld. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, 476–490. doi:10.1145/3581641.3584034

[15] Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes* 25, 2-3 (1998), 285–307.

[16] Google LLC. 2024. Summarize documents, text, and more with generative AI and LLMs. https://ai.google/.

[17] Ziwei Gu, Ian Arawjo, Kenneth Li, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. An AI-Resilient Text Rendering Technique for Reading and Skimming Documents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–22. doi:10.1145/3613904.3642908

[18] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, P. A. Hancock and N. Meshkati (Eds.). North-Holland, Amsterdam, 139–183.

[19] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, 1–18. doi:10.1145/3411764.3445648

[20] Anne Sophie Jensen and Carsten Elbro. 2022. Clozing in on reading comprehension: a deep cloze test of global inference making. *Reading and Writing* 35 (2022), 2303–2322. doi:10.1007/s11145-021-10230-w

[21] Manu Kapur. 2012. Productive failure in learning the concept of variance. *Instructional Science* 40, 4 (2012), 651–672.

[22] Manu Kapur. 2014. Productive failure in learning math. *Cognitive Science* 38, 5 (2014), 1008–1022.

[23] Jeffrey D. Karpicke and Janell R. Blunt. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 331, 6018 (2011), 772–775.

[24] Walter Kintsch. 1994. Text comprehension, memory, and learning. *American Psychologist* 49, 4 (1994), 294–303.

[25] Walter Kintsch. 1998. *Comprehension: A paradigm for cognition.* Cambridge University Press, Cambridge, UK.

[26] Kenneth R. Koedinger and Vincent Aleven. 2007. Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review* 19, 3 (2007), 239–264.

[27] Nate Kornell, Matthew J. Hays, and Robert A. Bjork. 2009. Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35, 4 (2009), 989–998.

[28] Suzanne M. Mannes and Walter Kintsch. 1987. Knowledge organization and text organization. *Cognition and Instruction* 4, 2 (1987), 91–115.

[29] Danielle S. McNamara. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers* 36, 2 (2004), 222–233.

[30] Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press, Cambridge, UK.

[31] Danielle S. McNamara and Joe Magliano. 2009. Toward a comprehensive model of comprehension. In *Psychology of Learning and Motivation*. Vol. 51. Academic Press, 297–384.

[32] Danielle S. McNamara, Tenaha O'Reilly, Rachel M. Best, and Yasuhiro Ozuru. 2006. Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research* 34, 2 (2006), 147–171.

[33] Jakob Nielsen and Jonathan Levy. 1994. Measuring usability: preference vs. performance. *Commun. ACM* 37, 4 (1994), 66–75.

[34] OpenAI. 2025. Introducing Deep Research. https://openai.com/.

[35] Steven C. Pan and Shana K. Carpenter. 2023. The effect of prequestions on learning: A multilevel meta-analysis. *Educational Psychology Review* 36, 1, Article 5 (2023). doi:10.1007/s10648-023-09814-5

[36] Franz Schmalhofer and Danilo Glavanov. 1986. Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language* 25, 3 (1986), 279–294.

[37] Keith W. Thiede, Mary C. Anderson, and David Therriault. 2003. Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology* 95, 1 (Mar 2003), 66–73. doi:10.1037/0022-0663.95.1.66

[38] Vera E. Woloshyn, Allan Paivio, and Michael Pressley. 1994. Use of elaborative interrogation to help students acquire information consistent with prior knowledge and information inconsistent with prior knowledge. *Journal of Educational Psychology* 86, 1 (1994), 79–89.

[39] Harle Xie, Zhongmin Zhou, and Qianling Liu. 2018. Review of the cognitive effects of disfluent typography on functional reading. *The Design Journal* 21, 4 (2018), 549–570. doi:10.1080/14606925.2018.1478237

Received 12 February 2026