From Highlights to Interests: Recovering What Readers Care About

Ozzie Kirkby
Ladder Media Lab
San Francisco, CA, USA

Andy Matuschak
Ladder Media Lab
San Francisco, CA, USA

Abstract

AI reading tools generate summaries, flashcards, and study guides from what users read, but typically assume uniform interest. Digital highlights offer a low-cost personalization signal, yet every mark looks the same. The passage that shifted how you think and the passage that merely caught your eye leave identical traces. We asked 42 readers to highlight articles freely, then select which ideas they would actually want turned into spaced-repetition flashcards. Spatial overlap between highlights and interest anchors predicts selections at F1 = 0.62; frontier LLMs given the full article with highlights marked inline improve this modestly (F1 = 0.67). But the aggregate hides a sharp split. Broad readers who select most interests are easy to predict (F1 = 0.87). Selective readers, who highlight comparably but choose far fewer interests, drop to F1 = 0.47. Only 45% of their highlighted interests are ultimately selected for spaced-repetition flashcards. For these readers, highlights encode where attention went—not what mattered.

1 Introduction

AI systems are increasingly embedded in reading tools, turning what people read into summaries, flashcards, and recommendations. As these systems begin to decide what to store, revisit, or resurface in personal knowledge workflows, they face a personalization challenge: different readers care about different things in the same text. Digital highlights offer an appealing signal—cheap, natural, and already produced during reading—but every highlight looks the same. The passage that changed how you think and the passage that merely caught your eye leave identical marks.

Winchell et al. show that highlighting patterns predict quiz performance and reflect stable individual differences in what readers find important, even though the *amount* highlighted does not [6]. They use these patterns to predict what a reader would highlight in new material. Highlights also shape the reading process itself: constraining how much readers can highlight makes them more selective, encourages deeper processing, and improves comprehension [1]. Yet even so, highlights remain only "a peek into [cognitive] processes, but obviously not a complete record" [6].

AI reading tools increasingly use highlights in a different way: as *input* for downstream action [4]. The problem is no longer simply what highlights reveal about the reader, but whether an external system can infer from those traces what the reader actually cares about enough to act on. Interest is not monolithic [2], and a single highlight may flatten curiosity, confusion, surprise, and genuine importance into the same trace. As Marshall observed, personal annotations are inherently telegraphic, incomplete, and tacit: they "pose interpretive difficulties for anyone other than the original annotator" [3].

Table 1: Articles and example interests.

Article	# Interests	Example Interest
Greening the Solar System[2]	12	Silica aerogels could be manufactured on Mars from rock silica to create local greenhouse bubbles.
How the Bitcoin Protocol Actually Works[3]	10	Attackers need >50% hash power to reliably rewrite history; with less, catch-up probability is tiny.
Three Kinds of Tacit Knowledge[4]	13	Scientists couldn't replicate TEA lasers just from details in the papers: success required tiny details the authors didn't think to document.

We test whether highlights can recover what readers actually want to carry forward. Forty-two readers highlighted freely, then selected which ideas they wanted turned into spaced-repetition flashcards—a deliberately high bar. Prediction works well when attention and intent coincide, and fails when they come apart.

2 Study & Findings

Forty-two experienced spaced-repetition users[1] were each assigned one of three casual explainer articles (Table 1), based on their stated preferences in an entrance survey. They read in a web interface augmented with a digital highlighter and were asked to highlight whatever struck them. After reading, they were shown a fixed list of 10–13 interests for that article. Each interest was a specific claim or idea from the text, distilled in advance and manually anchored to the contiguous passage that best represented it. Readers then selected which ideas they would want turned into flashcards, knowing they would receive a deck based on their selections. Participants had to decide whether an idea was interesting enough to revisit over the coming weeks and months.

Readers differed substantially in what they found interesting. Pairwise Jaccard similarity—the fraction of interests two readers share out of all interests either selected—averaged just 45%. Among selective readers (excluding the 19% who picked everything), overlap dropped to 39%. Fewer than half of interests overlapped between any two readers. This disagreement means that surfacing the same takeaways for everyone will reliably miss the mark. Any useful

[1] Participants were recruited from r/anki and X (formerly Twitter), then filtered for self-reported spaced-repetition experience of at least three months. Fifty-one participants completed the study; nine were excluded by this experience filter.

Figure 1: Both predictors over-predict for selective readers. Rows are readers (sorted by number selected); columns are interests. Green = selected, grey = unselected. Blue circles: highlight-overlap heuristic; purple diamonds: GPT-5.2.

system has to recover individual interest from the reading trace itself.

2.1 Recovering Interest

We first evaluate a simple spatial-overlap heuristic. Each interest is associated with an anchor passage, and we predict that an interest was selected when that passage was highlighted by the reader. Across all readers, it achieves precision 0.68 and recall 0.61 (F1 = 0.62). A uniform baseline that surfaces every interest to every reader achieves perfect recall but precision of only 0.63, since the average reader selects 7 of ~11 available interests. The heuristic improves on this precision while cutting recall losses, filtering out irrelevant interests without sacrificing much of what readers want.

As a coarse filter, it works, but its limits are systematic. First, coverage: 83% of readers have at least one selected interest whose anchor was never highlighted. Second, and more critically, overlap recovers where attention landed, not what the reader intends to carry forward. As readers become more selective, these false positives accumulate: highlighted anchor passages increasingly fail to indicate selected interests.

The best LLM tested[5] improves aggregate performance to F1 = 0.67 (purple diamonds in Figure 1), recovering some interests the spatial heuristic misses. But the pattern of failure remains unchanged. Both predictors systematically over-predict for readers who select few interests. Inference raises the ceiling; it does not help the system distinguish what a reader noticed from what they actually intended to keep.

[5]We evaluated four frontier LLMs: GPT-5.2, Claude Opus 4.5, Claude Opus 4.6, and Gemini 3 Pro. Each was given the full article with the reader's highlights marked inline and the list of candidate interests (tested with and without reasoning). GPT-5.2 (medium reasoning) performed best overall.

Table 2: Prediction performance by reader type. Readers are grouped into three equal-sized groups based on how many interests they selected. % Intent describes, among the interests whose anchors intersected a highlight, the fraction the reader ultimately selected. LLM F1: Performance of GPT-5.2 (medium reasoning), the best-performing model tested.

Reader type	*n*	Avg sel.	% Intent	Heur. F1	LLM F1
Broad (9–12)	14	10.5	94%	0.818	0.869
Mid (7–8)	13	7.4	79%	0.639	0.695
Selective (1–6)	15	3.7	45%	0.407	0.474
Overall	42	7.1	—	0.616	0.674

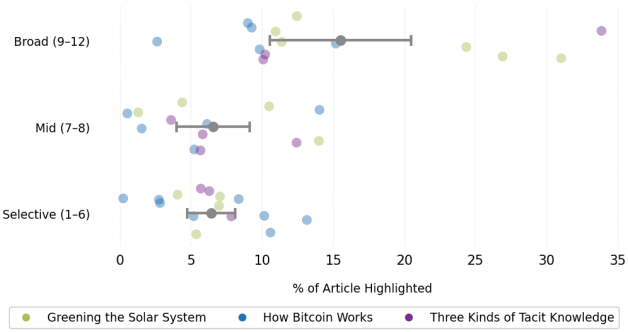

Figure 2: Percentage of article highlighted by reader type. Broad readers highlight roughly twice as much of their article as mid-range and selective readers. Individual readers are colored by article; black dots and error bars show group means and 95% confidence intervals.

2.2 Three Kinds of Readers

To unpack this variation, we grouped readers into terciles based on how many interests they ultimately selected, yielding three reader types: broad, mid-range, and selective (Table 2). These groups mark three qualitatively different regimes of reading behavior. What separates them is not how much readers highlight, but how often a highlighted passage corresponds to an interest they later choose.

Broad readers select nearly everything: on average 10.5 of ∼11 available interests. For them, attention and intent are nearly identical. 94% of highlighted interests became selections, and both the heuristic and the LLM performed well (F1 = 0.87). Their heuristic success is also partly mechanical: broad readers highlight roughly 15% of the article by word coverage, so a wider net catches more interest anchors by chance. Personalization adds little over simply surfacing everything; almost anything surfaced will be welcomed.

Mid-range readers select about 7 interests on average. Here the gap between attention and intent begins to open: only 79% of highlighted interests become selections. LLM inference helps, pushing F1 from 0.64 to 0.70 by filtering out some interests that caught the reader's eye but didn't ultimately matter enough to keep.

Selective readers behave differently. They make about as many highlights as mid-range readers (20 on average versus 24), and both groups cover about 7% of the article by word count—far less than broad readers, who highlight about 15% (Figure 2). But selective readers choose far fewer interests afterward, only about 3 on average. Just 45% of the interests linked to passages they highlighted were later selected. In other words, for every interest they both highlight and choose, there is roughly another they highlight but leave unselected. Even with the full article and highlighted passages in context, the LLM improves only marginally for these readers (F1 = 0.41 → 0.47). The gap is in the signal itself, not in the inference method.

For these readers, highlighting functions as exploration, not decision. It marks careful reading, orientation, and sense-making. Selection is a separate, later act with a much higher bar. Their highlights faithfully encode what caught their attention, but not what they intended to revisit. The aggregate F1 of 0.67, which suggests highlights "mostly work," is therefore misleading. Prediction is easy when nearly everything a reader notices also matters enough to act on. It fails precisely where personalization would matter most: for readers whose attention is wide but whose intent is narrow.

3 Discussion

Across readers, simple spatial overlap already removes substantial noise relative to surfacing all candidate interests (F1 = 0.62), and LLM inference improves this modestly (F1 = 0.67). For readers whose attention and intent largely coincide, this level of personalization is often sufficient.

The fracture appears when attention and intent diverge. Selective readers highlight about as much text as others by word coverage, but fewer than half of their highlighted interests become selections. Highlights therefore move us from no personalization to only coarse personalization. When attention and intent coincide, the system can seem prescient. But what makes a passage feel worth marking during reading—surprise, mismatch with expectations, sense-making [2]—is not always what makes it worth revisiting later. For selective readers, highlighting often reflects ongoing comprehension rather than a decision to preserve an idea. The same visible mark can signal curiosity, confusion, or genuine long-term importance. Once those motives are collapsed into a single trace, none of the methods explored can reliably tell them apart.

Many emerging AI reading tools treat highlights and other passive reading traces as signals of what should be summarized, recommended, or acted upon [4]. Our findings suggest a structural constraint: passive attention traces reliably encode where a reader looked, but not how much weight they assign to what they saw. Systems that infer intent from attention alone risk over-personalizing—amplifying content that was merely noticed rather than meaningfully valued. Qin et al. found the complementary failure: when readers know their highlights drive AI-generated suggestions, highlighting becomes a strategic act of feeding the system rather than a reading strategy, and the resulting engagement is superficial [4]. Together, these findings point to a structural tension in the signal itself: when readers highlight naturally, the trace is ambiguous; when they know highlights drive downstream artifacts, the behavior shifts. The gap between what readers highlight and what they actually care about is not easily closed from either direction.

One response is to extract richer signals from the same gesture, incorporating features such as highlight length, density, or thematic

structure. Joshi and Vogel showed that constraining the number of words a reader can highlight forces self-regulation and improves comprehension, without adding cognitive overhead [1]. However, our selective readers already highlight sparingly—about 7% of the article by word coverage—and their highlights remain difficult to interpret. Selectivity in highlighting may benefit the reader's own processing without necessarily producing a trace that is more legible to downstream systems. More broadly, LLM inference helps at the margin, improving F1 from 0.62 to 0.67, but does not close the gap when attention and intent come apart. Inference can refine estimates of attention, but it cannot reliably infer importance that the gesture itself did not encode.

An alternative is to ask readers for more explicit signals of intent. Prior work suggests that readers often want richer annotation mechanisms: Tashman and Edwards report requests for annotation layers, levels of importance, and multiple marker colors [5]. But the appeal of highlighting lies in its low cognitive overhead. It is a lightweight, in-the-moment reaction during reading. As Marshall observed, readers immersed in a text "seldom make more explicit than that which is required for the task at hand" [3]. Augmented reading systems therefore face a trade-off: accept coarse personalization from weak signals, or ask for richer signals at the cost of a more disruptive reading experience.

Our findings don't argue against highlight-based personalization. They clarify its scope. As AI systems increasingly transform what we read into summaries, recommendations, and actions, the risk is not a noisy prediction but a subtler drift: output that stays grounded in something the reader marked yet misrepresents what they found worth preserving. The question is not whether to personalize from reading traces, but how much weight to place on what was simply noticed.

References

[1] Nikhita Joshi and Daniel Vogel. 2024. Constrained Highlighting in a Document Reader can Improve Reading Comprehension. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 893, 10 pages. doi:10.1145/3613904.3642314

[2] Walter Kintsch. 1980. Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics* 9, 1 (1980), 87–98. doi:10.1016/0304-422X(80)90013-3

[3] Catherine C. Marshall. 1998. Toward an ecology of hypertext annotation. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '98)*. Association for Computing Machinery, New York, NY, USA, 40–49. doi:10.1145/276627.276632

[4] Peinuan Qin, Chi-Lan Yang, Nattapat Boonprakong, Jingzhu Chen, Yugin Tan, and Yi-Chieh Lee. 2026. AI Personalization Paradox: Personalized AI Increases Superficial Engagement in Reading while Undermines Autonomy and Ownership in Writing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 17 pages.

[5] Craig S. Tashman and W. Keith Edwards. 2011. Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 2927–2936. doi:10.1145/1978942.1979376

[6] Adam Winchell, Andrew Lan, and Michael Mozer. 2020. Highlights as an Early Predictor of Student Comprehension and Interests. *Cognitive Science* 44, 11 (2020), e12901. doi:10.1111/cogs.12901