WHITE PAPER · AGENT MEMORY · V1.0

Dual-signal retrieval for *AI agent memory*.

An AI agent with a long memory has to decide which memories come to mind first. The common answer, borrowed from memory research, is to rank memories by how often and how recently they were used. We built that for Bizer's agent, Miles, then **simulated it before trusting it**. The first version made recall worse than no ranking at all. The simulation also showed what works: two separate signals, **use** and **correctness**, each kept small and each on the clock that suits it.



Author Bizer **Audience** People who build AI agents **Published** October 2026

Download the PDF

THE CORE IDEA

Keep used lately and right lately as two signals. Never one.

> A memory's rank should rise when it is deliberately used and fall when the owner corrects it, and both pulls must stay smaller than the gaps between the candidates they reorder. Fading reorders a memory; it never hides it.

75.9% → 89.8%

correct answers, against relevance alone

47.5% → 3.7%

wrong answers after a fact changed

83.2% → 82.2%

rarely needed pages found: kept, not traded away

Simulated: a 300-page business memory over two years, 20 random seeds, bursty need, 3% grader noise. The ranking code is Bizer's production code.

The two headline numbers are measured over different questions. Correct answers are out of every question, and nearly all of the 10.2% that were not correct are misses (9.9%), where the page the question needed was not among the four recall picked. Wrong answers after a change are out of only the questions about a fact that changed, after it changed; across all questions, a stale page led the answer about 0.3% of the time.

CONTENTS

01 Background
02 Method
03 What counts as a use
04 A second signal for correctness
05 How big a nudge can be
06 Two clocks
07 The design, and what is new
08 Limits and live checks

§01 · BACKGROUND

The field ranks memory by use, and assumes retrieval is use

Anderson and Schooler (1991) showed that how likely a memory is to be needed again tracks how recently and how often it was needed before, and that human forgetting follows the same curve: a power law. ACT-R (Anderson and Lebiere, 1998) turns that into one number per memory, base-level activation $B = \ln(\Sigma\, t_k^{-0.5})$ over every past use.

Agent builders borrowed the idea widely. Generative Agents scores recency at 0.995 per hour since a memory was last retrieved. MemoryBank strengthens a memory each time it is recalled. MemoryOS promotes and evicts by a "heat" of visits and recency. Mem0's Memory Decay biases search toward memories recently returned by a search. FadeMem and Oblivion decay memories unless accessed, and at least two systems use ACT-R itself.

Almost all of them share one assumption: **a memory that was retrieved was used.** Retrieval raises its rank, which makes it retrieved again.

The brain keeps two signals apart

Bjork and Bjork's (1992) new theory of disuse gives every memory a **retrieval strength** (how easily it comes to mind now) separate from its **storage strength** (how well it is learned). Separate again is whether a memory is *right*. Memories that lead to reward are consolidated through dopamine-driven systems distinct from the ones that set accessibility (Shohamy and Adcock, 2010). A recalled memory that turns out wrong becomes briefly open to correction (Nader, Schafe and LeDoux, 2000), driven by prediction error, not by how often it was used.

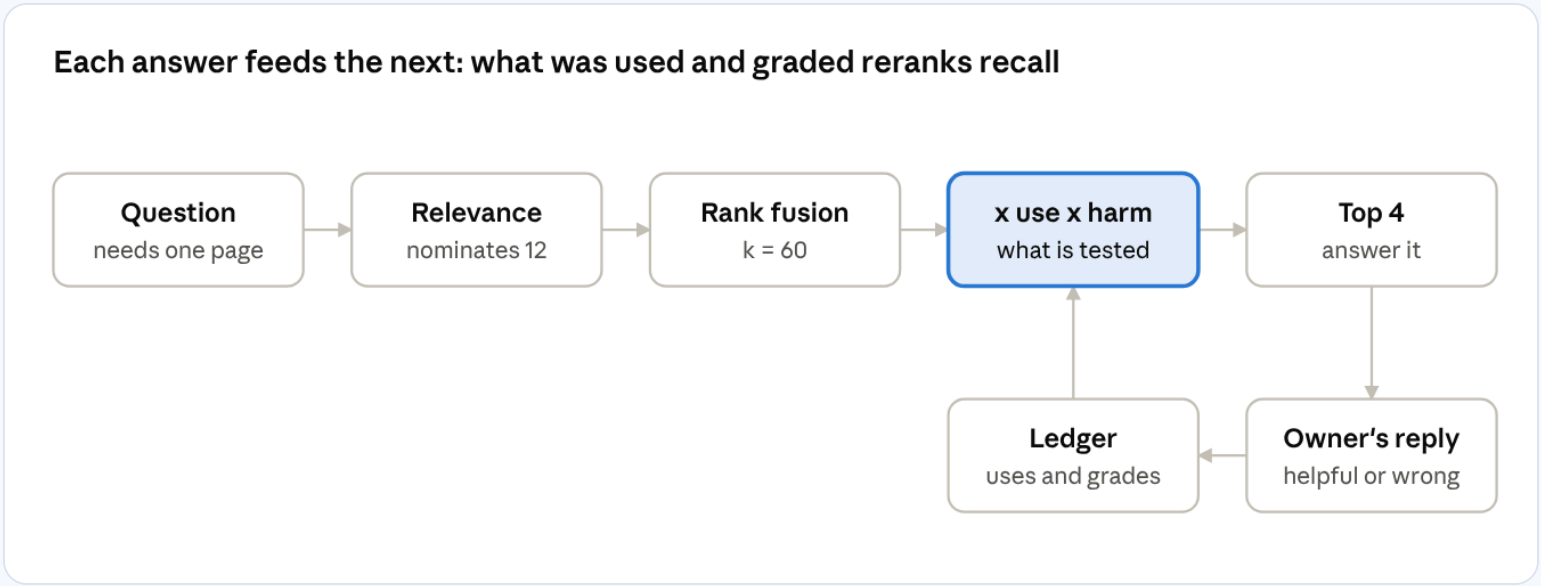
Bizer's first version followed the field. It counted a page as used each time recall carried it into an answer, kept a lifetime tally of helpful and harmful grades as a second

multiplier (×0.5 to ×1.5), and nudged recall by power-law strength at ×0.8 to ×1.2. It looked principled. We had not tested it.

§02 · METHOD

A simulated business memory, run through production code

The simulator is a small open program in the Bizer repository (`scripts/memoryStrengthSim.ts`). It imports the production ranking functions rather than reimplementing them, so it tests the code that ships.



The simulated recall loop. Each answer feeds the next: what was used and graded reranks recall.

KIND OF PAGE	PAGES	HOW OFTEN IT IS NEEDED
Core (a main supplier, the menu)	20	Often
Project	40	Heavily for 30 to 60 days, then rarely

KIND OF PAGE	PAGES	HOW OFTEN IT IS NEEDED
Annual (a license renewal)	24	A few times, once a year
Tail	204	Rarely
Replaced pairs	6 pairs	One page, then its replacement after a switch day

The **replaced pairs** are the hard case. A business changes vendor or prices; the old page and the new one are about the same thing, and relevance cannot tell them apart. Only use and outcome can.

Two worlds. In the scheduled world, needs arrive at steady rates. In the bursty world, each need raises the odds of the next by a power law of the time since, the recurrence statistics Anderson and Schooler measured in real environments. We report the bursty world unless we say otherwise.

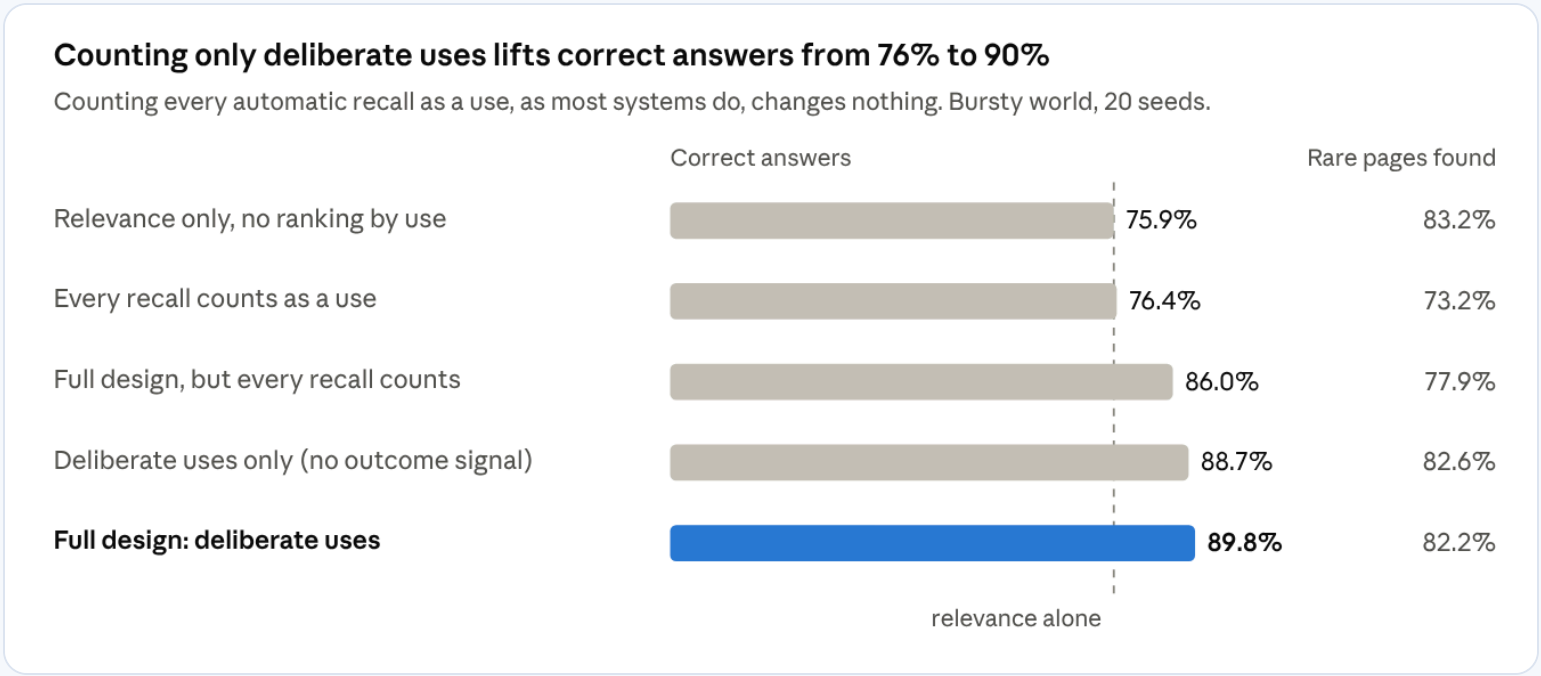
Recall. For each question, relevance nominates the 12 best candidates. Reciprocal rank fusion (Cormack et al., 2009, k = 60) turns rank into a score, the ranker under test multiplies each page's score, and the top four answer. This is Bizer's production recall, simplified to one ranked list.

Feedback. An answer is correct if the needed page is in the top four and no stale page ranks above it. Owners grade a correct answer helpful half the time and, by mistake, harmful 3% of the time, and correct a stale answer 70% of the time. Each figure is the mean over 20 seeds; seed-to-seed standard deviations are under one point for correct answers and under four for rare pages.

§03 · FINDING ONE

What counts as a use decides whether ranking helps

Under the usual definition, "the memory was retrieved," ranking by use does nothing for accuracy and costs ten points on rarely needed pages.



Bursty world, 20 seeds, 3% grading noise. Rare pages are pages needed five times or fewer in two years.

The reason is a feedback loop. Recall nominates the needed page together with a few near misses. Whatever lands in the top four counts as used, so popular near misses gain strength, rank higher next time, and are used again. The ranking learns what recall already did, not what the owner needed.

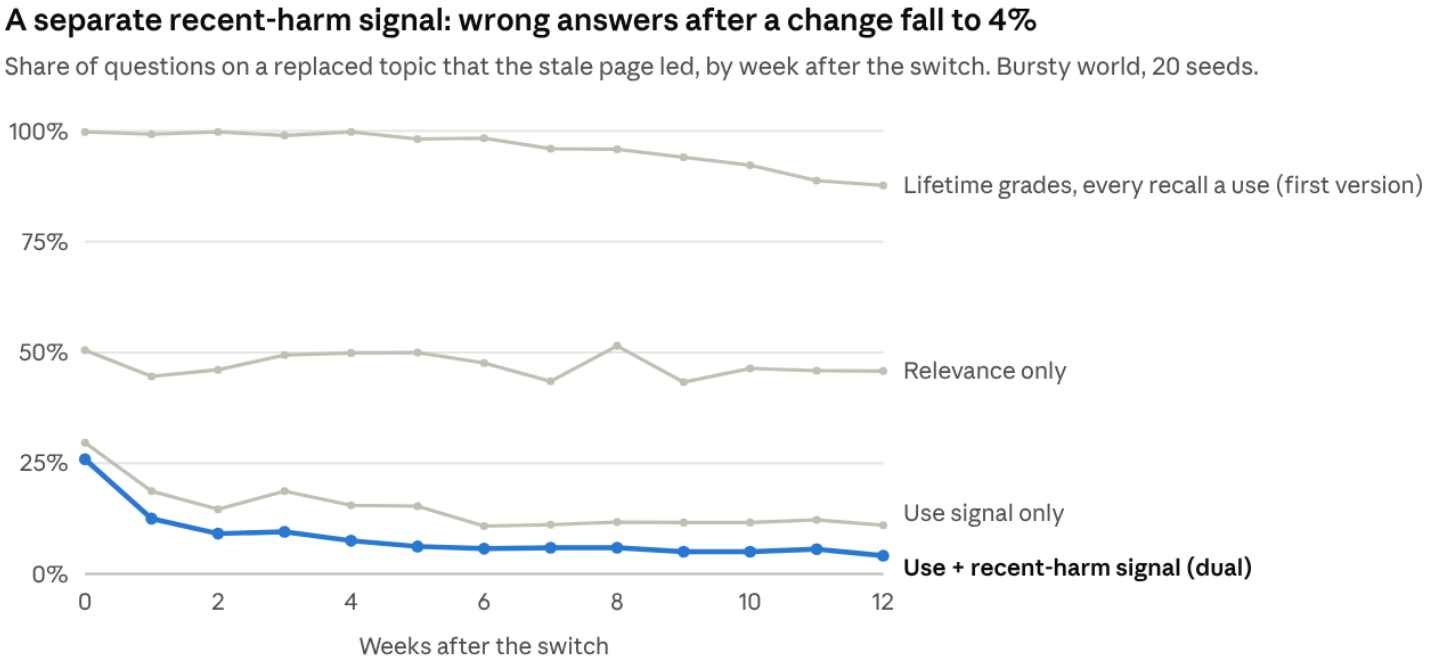
Counting only **deliberate** uses breaks the loop: the owner confirmed the answer helped, the agent chose to open the page, or the page was written. None of them happens just because recall surfaced a page. The testing effect (Roediger and Karpicke, 2006) makes the same distinction: memories retrieved and used are strengthened, not memories that merely flicker past.

---

§04 · FINDING TWO

Correctness needs its own signal, recent and shallow

Use tells you a memory is reachable. It cannot tell you a memory is right. When a business switches vendor, the old vendor's page, used every week for a year, is now wrong, and relevance cannot separate the two pages.



Share of questions on a replaced topic that the stale page led, by week after the switch. Bursty world, 20 seeds.

CORRECTNESS SIGNAL	CORRECT ANSWERS	WRONG AFTER A SWITCH	WHY
Lifetime grades (×0.5 to ×1.5)	75.4%	34.3%	A page right for a year has dozens of helpful grades; months of corrections cannot outvote them
Deep demotion on the latest harmful grade (×0.6)	41.8%*	0.0%*	One mistaken grade buries a page for good: once demoted it is never recalled, so it never earns the grade that would lift it

CORRECTNESS SIGNAL	CORRECT ANSWERS	WRONG AFTER A SWITCH	WHY
Recent and shallow (harm share of the newest 8 grades, down to ×0.91, demote only)	**89.8%**	**3.7%**	A change is noticed within weeks; a stray grade costs a little and fades

* 8 seeds, from the exploration runs; all other figures 20 seeds.

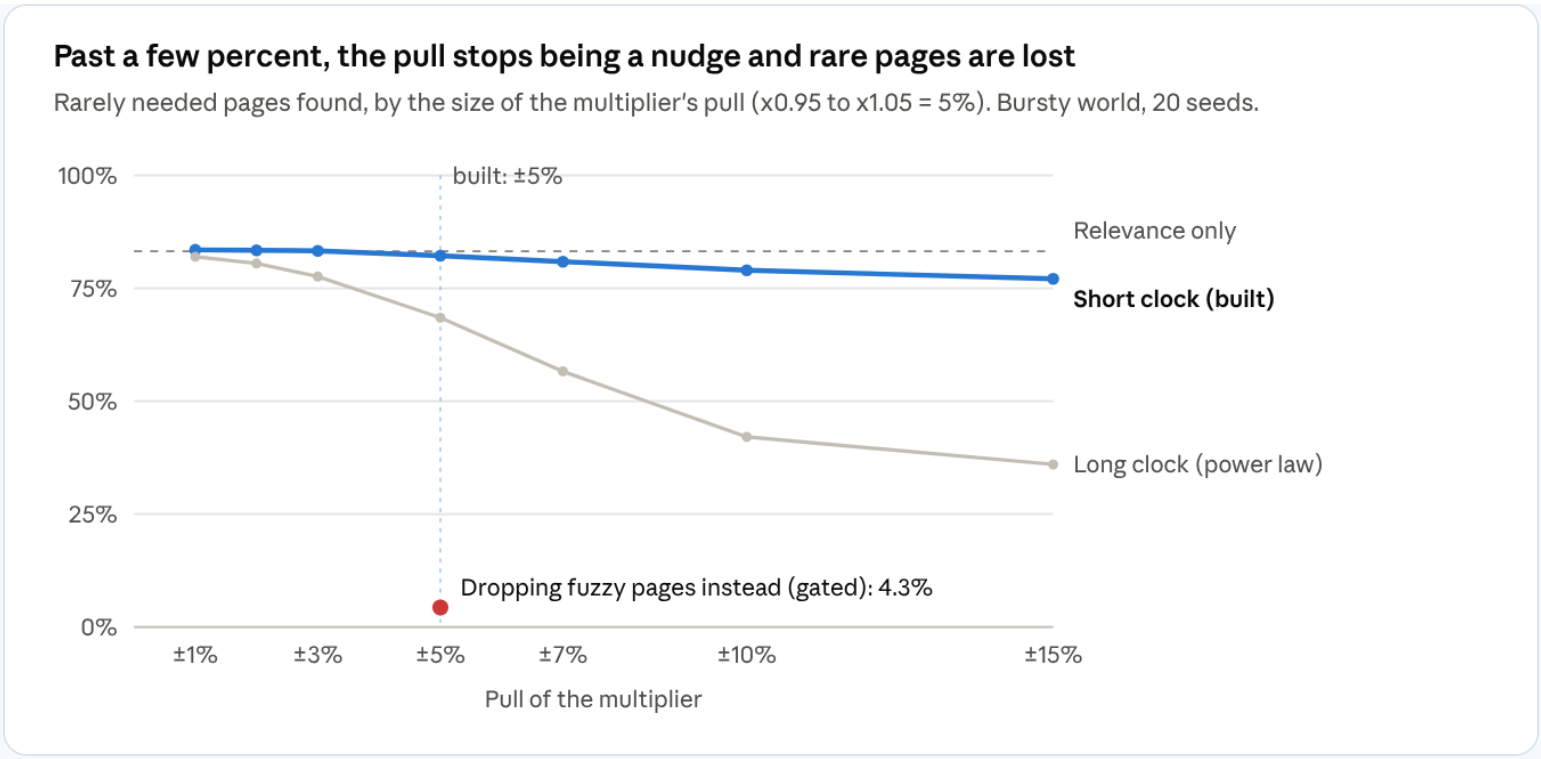
The signal **only demotes**: a helpful grade already counts as a use, so letting it also raise correctness would count the same evidence twice. And it holds up with noisy owners: with one right answer in ten graded wrong, the design still answered 87.5% correctly against 75.9% for relevance alone.

---

§05 · FINDING THREE

A nudge must stay inside the spread of the scores it multiplies

Reciprocal rank fusion scores a candidate at rank r as 1 / (60 + r). The best of twelve scores 1/61 and the twelfth 1/72: **the whole list sits within about 18%**. A ×0.8 to ×1.2 pull spans 50%, so it does not break ties. It re-sorts the list by use.



Rarely needed pages found by the size of the multiplier's pull. Bursty world, 20 seeds.

Correct answers rise with the pull up to about ±5% and then flatten (89.8% at ±5%, 90.7% at ±15%), so a larger pull buys almost nothing and costs the pages a business needs once a year. The extreme is a gate: dropping faded pages, the "cold tier" several systems use, found 4% of rarely needed pages and answered 27% of questions correctly.

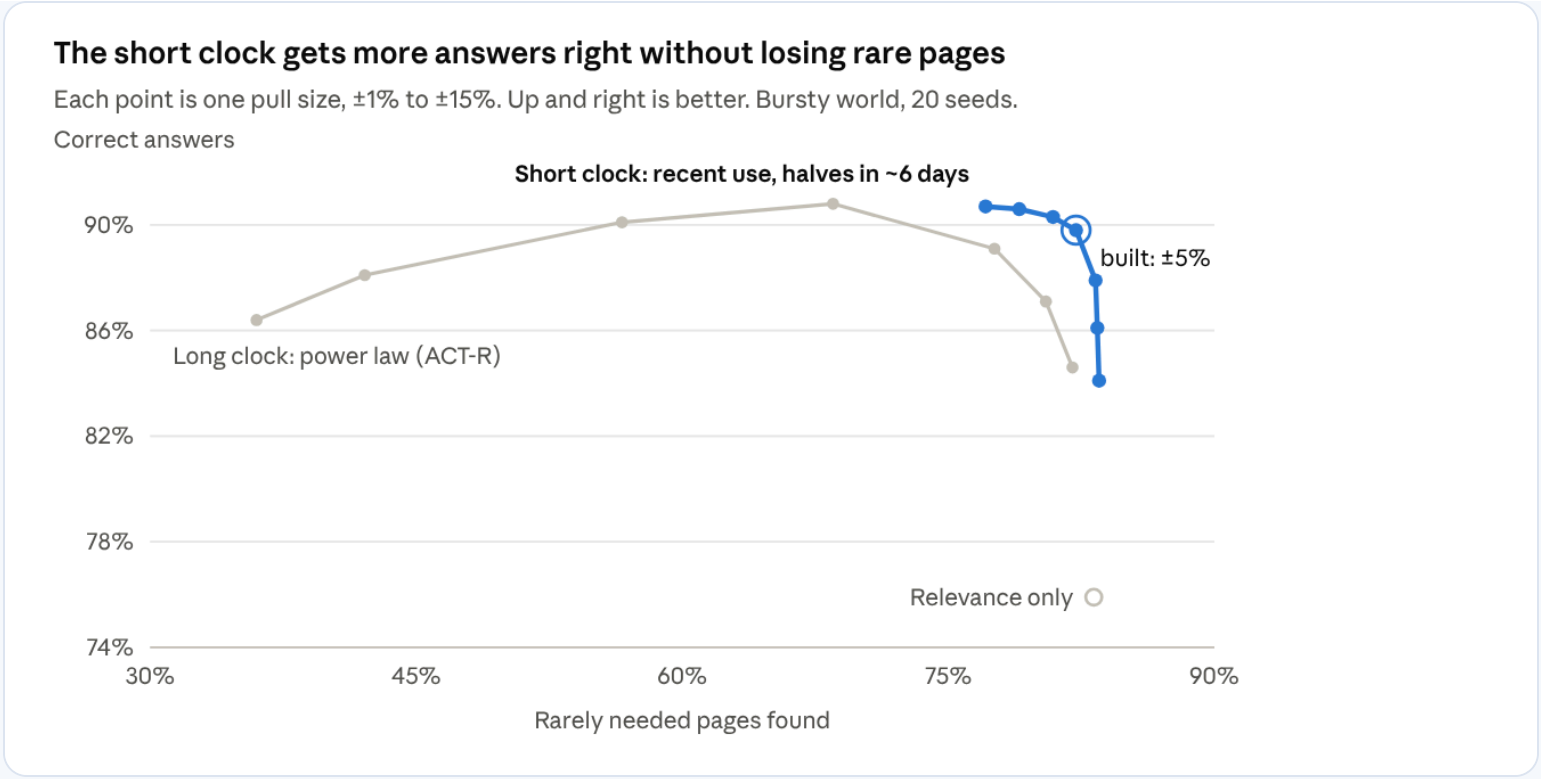
The rule: size a ranking nudge against the spread of the fused scores it multiplies. Bizer's recall uses ±5% for use and at most about −9% for harm.

---

§06 · FINDING FOUR

Two clocks, one for recall and one for what stays in view

We expected the power law to win everywhere. For recall, it lost.



Each point is one pull size, ±1% to ±15%. Up and right is better. Bursty world, 20 seeds.

A short clock (recent use, 0.995 an hour, halving in about six days) answers more questions correctly than the power law at the same rare-page recall, because the long clock gives lasting weight to every popular page. The long clock wins the other job, choosing which pages the agent sees by name: ordered by power-law activation, the named list covered 76.5% of the next month's needs, against 71.0% ordered by last edit. One decay curve should not do both jobs.

---

§07 · THE DESIGN, AND WHAT IS NEW

Dual-signal retrieval

PART	RULE
What counts as a use	The owner confirms an answer helped, the agent opens the page, or the page is written. Being recalled is counted, never a use

PART	RULE
Signal 1: use, short clock	Recency of last use, 0.995 an hour; multiplies recall's fused score ×0.95 to ×1.05
Signal 2: correctness	Share of harm in the newest 8 grades; only demotes, to about ×0.91; harm before a repair is forgiven
Use, long clock	ACT-R activation over every use; orders which pages the agent sees by name
Fuzzy	Long-clock strength under 50%: still stored, still found by a deliberate search, not named

What exists already. Ranking memory by use is common. ACT-R has been ported to agents (HAI '25; Guatu Labs, 2026). Bounded, non-deleting decay exists in Mem0 and Guatu Labs, which also argued for counting only acted-on memories. Oblivion reinforces the memories that produced a response.

What we did not find elsewhere, and the simulation supports:

- **Correctness as a separate, recent, demote-only signal**, beside a use signal that never counts harm. The folded, lifetime form was the worst design tested for changed facts.
- **A sizing rule for ranking nudges** against the fused-score spread, with a measured cost for crossing it.
- **Two clocks for two jobs**, each chosen because it beat the other at the other's job.
- **An open harness that runs production ranking code** with feedback, bursty need, replaced facts and noisy graders.

None of these is a new theory. Each is a measured engineering choice that the systems we surveyed do not make, and the first version of our own system did not make either.

§08 · LIMITS AND LIVE CHECKS

Simulated, not yet measured on live traffic

- **The worlds are made, not measured.** Their shape follows published statistics and plain business patterns; no real business's memory was replayed. Every finding held in direction in both worlds; the sizes will not transfer exactly.
- **Recall is simplified.** Production fuses a title list and a meaning list, which spreads scores further; the safe pull may be somewhat larger than ±5%.
- **Correctness pays on changed facts, not on noise.** At 10% grading mistakes the design's edge over the use signal alone vanished (87.5% against 88.4%), while it still cut wrong answers after a switch from 13.5% to 4.0%.
- **The constants are chosen, not fitted.** They sit at a knee in the simulation and should be fitted per business once live data exists.

Live, we will check calibration (does long-clock strength predict use over the next 30 days), named-list coverage, how fast stale pages stop leading after an owner reports a change, and a split comparison of relevance alone against dual-signal retrieval.

SOURCES

- Anderson, J. R. and Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science* 2(6): 396–408.
- Anderson, J. R. and Lebiere, C. (1998). *The Atomic Components of Thought.* Erlbaum.
- Bahrick, H. P. (1984). Semantic memory content in permastore. *Journal of Experimental Psychology: General* 113(1): 1–29.
- Bjork, R. A. and Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In *From Learning Processes to Cognitive Processes*, Vol. 2.

Erlbaum.

- Nader, K., Schafe, G. E. and LeDoux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* 406: 722–726.
- Roediger, H. L. and Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science* 17(3): 249–255.
- Shohamy, D. and Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences* 14(10): 464–472.
- Cormack, G. V., Clarke, C. L. A. and Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *SIGIR '09*.
- Guatu Labs (2026). Eviction without deletion: running an ACT-R decay policy for agent memory.
- Human-Like Remembering and Forgetting in LLM Agents: An ACT-R-Inspired Memory Architecture (2025). *HAI '25*, ACM. doi:10.1145/3765766.3765803.
- Kang, J. et al. (2025). Memory OS of AI Agent. arXiv:2506.06326.
- Mem0 (2026). Memory Decay.
- Park, J. S. et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *UIST '23*.
- Rana, A. et al. (2026). Oblivion: Self-adaptive agentic memory control through decay-driven activation. arXiv:2604.00131.
- Wei, L. et al. (2026). FadeMem: Biologically-inspired forgetting for efficient agent memory. arXiv:2601.18642.
- Zhong, W. et al. (2024). MemoryBank: Enhancing large language models with long-term memory. *AAAI 2024*.
- Bizer (2026). The Hippocampal Index for AI Agents.
- Bizer (2026). The Episodic Buffer for AI Agents.

Bizer

build AI agents. Rank memories by deliberate use, keep correctness as its own signal, and keep every nudge small.