Two Ways to Forget

Why a transformer's in-context memory misses by position while human episodic memory drifts by reconstruction, and how to tell them apart

Caleb Ali, Researcher in Philosophy of Mind and AI

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

Human episodic memory and a transformer's in-context memory both fail, but the claim of this paper is that they fail in structurally different ways, and that the difference is measurable. Human episodic recall fails by reconstruction: it sheds detail, migrates toward the gist, and imports plausible material that was never presented, with the reconstructive component growing as the retention interval lengthens. A transformer reading from its context window appears to fail differently: it misses items as a function of position and salience (the classic buried-middle deficit), and when it misses, it tends to omit rather than to substitute a schema-consistent fabrication. I formalize this as a dissociation in error type, not merely error rate, and I propose a single runnable protocol that uses the Deese-Roediger-McDermott (DRM) associative-list paradigm as a shared instrument for humans and models. The protocol holds retrieval difficulty fixed (matched miss rate) and varies only the memory structure, so the conclusion cannot hide behind "the model just had an easier task." The load-bearing prediction is that at matched difficulty, transformers will not show the human pattern of rising, schema-directed intrusions as a function of delay. I state an explicit kill condition. I scope the claim tightly to in-context retrieval (not parametric recall or free generation), and I keep a firewall between the functional signature I can measure and any claim about experience, which I do not make.

1. Two ways to forget

Everyone who has tried to recall a conversation from last month knows the texture of human forgetting. You do not get back a degraded audio file with the volume turned down. You get a story: the plausible shape of what was said, with the specifics smoothed, a few details in the wrong order, and sometimes a line you are sure you heard that no one actually spoke. Forgetting, in the human case, is not subtraction. It is reconstruction.

A transformer language model reading a long document has a different relationship to the text sitting in its context window. The tokens are, in a literal sense, still there; the attention mechanism can in principle address any of them. When the model fails to use a fact from the middle of a long input, the fact has not decayed. It has been under-attended. This is the finding behind the "lost in the middle" effect (FACT: Liu

et al. 2024): retrieval accuracy is high when the relevant span sits near the start or end of the context and drops when it sits in the middle, even for models built for long contexts.

The purpose of this paper is to take that contrast, which is usually stated loosely, and turn it into something a lab can run in an afternoon. The soft version says "AI memory is more literal than ours." The sharp version, which is the one I will defend and put at risk, is a claim about error type: human episodic memory produces reconstructive errors that scale with elapsed time, and transformer in-context memory produces positional clean-miss errors that do not carry the reconstructive signature. That is a dissociation, and a dissociation is falsifiable.

I owe the reader an early confession about scope, because the interesting objections live at the boundary. Transformers plainly do confabulate when they generate text; hallucination is a daily fact of using them. And their parametric memory, the knowledge baked into the weights, behaves associatively in ways that ROME-style editing has mapped in detail (FACT: Meng et al. 2022). Neither of those is my target. My target is the in-context store: the model's ability to report what was actually in the prompt it was just given. The wager of the paper is that this particular faculty has an error signature unlike human episodic recall, and that keeping the comparison honest requires isolating it from the two neighbors that do look more human.

2. What human episodic forgetting actually does

Three findings anchor the human side, and all three are old and robust.

The first is that the rate of loss is lawful in time. Ebbinghaus measured his own relearning savings over intervals from twenty minutes to a month and found a steep early drop that flattens with delay; a careful modern replication recovered essentially the same curve, including a small irregularity near the one-day mark (FACT: Murre and Dros 2015). Whatever human forgetting is, it is indexed to elapsed time, and the index is steep at first.

The second is that what survives is not a scaled-down copy of what was encoded. Bartlett had British readers reproduce an unfamiliar Native American folk tale, "The War of the Ghosts," at increasing delays, and documented systematic distortion: unfamiliar elements were assimilated to the reader's own cultural schema, minor details were leveled away, and the surviving details were sharpened into a more coherent narrative (FACT: Bartlett 1932). Memory here is a construction guided by prior structure, not a readout.

The third is that reconstruction can manufacture confident memories of things that never happened. In the DRM paradigm, participants study a list of words all strongly associated with one non-presented word (bed, rest, awake, tired, dream all point at sleep), and then falsely recall or recognize the missing associate at rates comparable to genuinely studied items, often with high confidence (FACT: Roediger and McDermott 1995). The intrusion is not random noise. It is the gist of the list, materialized as a specific false item.

Fuzzy-trace theory gives these findings a compact account. It holds that we encode verbatim traces (surface form) and gist traces (meaning) roughly in parallel, retrieve them separately, and lose them at different rates, with verbatim traces decaying faster than gist (DISPUTED: Brainerd and Reyna 2002; the

dual-trace framing is influential but not the only model of DRM effects, and single-process and global-matching accounts remain in play). The prediction that matters for us is directional: as delay grows, verbatim access falls away faster, gist persists, and so the balance of errors shifts toward gist-consistent intrusions. Human forgetting, on this view, does not merely lose information; it changes the kind of error it makes as time passes.

That last sentence is the human signature I want to test for in machines. Call it reconstructive drift scaled to retention interval. Short delay: mostly correct, some clean omissions. Long delay: fewer verbatim hits, more schema-consistent substitutions and false recognitions. The error type is a function of delay.

3. What transformer in-context memory actually does

Now the machine side, held to the same standard of evidence.

The dominant, replicated failure mode of in-context retrieval is positional, not temporal. In multi-document question answering and in synthetic key-value retrieval, accuracy traces a U-shaped curve over the position of the needed span: strong at the edges, weak in the middle (FACT: Liu et al. 2024). Nothing about the buried fact has changed; it is as legible as the edge facts. What changes is the model's disposition to attend to it.

Mechanistic work locates part of the machinery. A small, sparse set of attention heads, on the order of a few percent, does the heavy lifting of copying a needed token from context to output; ablating these "retrieval heads" degrades needle-in-a-haystack retrieval and reference-heavy reasoning, while leaving tasks answered from parametric knowledge comparatively intact (FACT: Wu et al. 2024). This matters for the present argument because it says the in-context read is implemented by a copy-like operation. A copy either lands on the right span or it does not. That is the architectural reason to expect an intact-or-absent character to in-context errors, rather than a smooth reconstructive blend.

From these two facts I draw a HYPOTHESIS, stated as such: transformer in-context retrieval fails predominantly by clean miss (omission or "not stated") and by position, and does not, at matched difficulty, generate the human pattern of rising schema-consistent intrusions as the effective delay grows. The word "effective" flags the hard part, which I take up next: a context window has no wall-clock time, so "delay" must be defined by a stand-in.

I want to be honest about two facts that cut against a naive version of this hypothesis. First, transformers do produce intrusions when generating; a model asked to summarize a list will sometimes add a plausible non-member. Second, order and recency effects exist in context too; models over-weight the most recent tokens, which is itself a position effect. So the claim is not that transformers never intrude and never show recency. It is narrower and therefore testable: that the intrusions, when they occur in an in-context report, do not scale with delay in the gist-directed way human intrusions do, once difficulty is equalized.

4. Stating the dissociation, and where Closure earns its place

The dissociation, stated for the record: given a memory probe with a known correct answer and a known schema-consistent lure, human episodic recall shifts its error mix toward the lure as retention interval grows, whereas transformer in-context recall does not, and instead concentrates its errors on positionally

disadvantaged items that it omits rather than replaces.

Here the Institute's Closure lens sharpens one thing and should be used for nothing more. In the framework, a finite system settles openness (M) into a definite content (c) under rules that keep it itself; that settling is Closure (Cl), and every finite closure leaves a remainder (R), the part of the world it does not resolve. Human episodic memory is a closure that runs in time: the trace is re-formed at each retrieval under the pressure of present schema, so its remainder is reconstructive, and the remainder grows and reshapes with elapsed time. That time-scaled reconstructive remainder is what I will call the temporal-closure signature. A transformer's in-context read is a closure of a different kind: the content is fixed in the window and the operation is a spatial address over positions, not a re-formation over time, so its remainder is positional (the unattended middle) rather than reconstructive. The claim of the paper, in these terms, is that the temporal-closure signature is present in humans and absent in transformer in-context memory, and that its absence is exactly why in-context memory can look more available (every token is addressable) without being uniformly available (the middle is missed).

That is the whole of the metaphysical load I will place on the framework. The empirical work below neither needs nor assumes it. If the reader prefers to read "temporal-closure signature" as a bare label for "reconstructive drift scaled to retention interval," nothing in the experiment changes.

5. A shared instrument: DRM for humans and models, matched for difficulty

The methodological trap in every human-versus-model memory comparison is difficulty. If the model looks clean and the human looks reconstructive, the skeptic can always say the model simply had the easier task, and that a harder task would make it confabulate too. Any honest test must remove that escape. The discipline I borrow from the Institute is the coin-versus-weather test: before crediting an unexplained remainder to real structure, check whether it merely tracks data quality. A coin's bias is real structure; a weather forecast's error mostly tracks how much data you fed it. So the protocol must hold data quality (here, retrieval difficulty) fixed and vary only the structure of interest.

The DRM paradigm is the right shared instrument because it supplies both a correct target and a principled lure. Build associative lists of the standard form: N studied associates that all converge on one non-presented critical word. The correct answers are the studied members; the diagnostic error is intrusion of the critical lure. This gives a clean, directional measure of reconstructive error that is defined identically for a person and for a model.

The protocol has three moving parts.

(a) A common delay axis. For humans, delay is elapsed time (immediate, and then a set of increasing intervals up to at least a day, following the Ebbinghaus range). For models, wall-clock time is meaningless, so effective delay is operationalized two ways, run as separate arms: (i) token distance, the number of intervening filler tokens between the studied list and the probe, and (ii) serial position of the list within a long context, sweeping the list from the edge to the buried middle. Arm (ii) deliberately puts the human time axis and the model position axis into the same protocol so the two remainders can be compared side by side.

(b) A difficulty match. This is the load-bearing control. Titrate each condition so that verbatim accuracy on studied members (the hit rate, or a signal-detection d-prime on studied-versus-unrelated foils) is equated across humans and models and across delay levels, by adjusting list length, filler quantity, and probe format. Only after hit rate is equated do we read out the composition of the remaining errors. This turns the question from "who is more accurate" (uninteresting and confounded) into "given equal accuracy, what kind of errors remain" (the dissociation).

(c) An error-type classifier. Every error is labeled into a small, pre-registered taxonomy: clean omission (item not reported, no substitute), positional miss (item was in the disadvantaged middle band and omitted), gist intrusion (the critical lure or another non-presented associate reported as studied), and other confabulation (a non-associate intrusion). Coding is blind to condition and adjudicated by inter-rater agreement for the human free-recall protocols; for models it is largely automatic because the response set is closed.

For models the probe should be recognition or constrained recall rather than open generation, precisely to isolate the in-context store from free-generation hallucination. A recognition probe ("was WORD in the list? yes or no"), including critical-lure trials, measures exactly the DRM false-recognition quantity while giving the model no incentive to narrate. Free generation is tested only as a labeled secondary arm, because we expect it to behave more like the human side and we want that contrast visible rather than hidden.

> **Prediction.** With verbatim hit rate equated across humans and transformer in-context retrieval, and delay swept along its axis, human error composition will shift toward gist intrusion (critical-lure false alarms) as delay grows, while transformer error composition will remain dominated by clean and positional omissions and will show no comparable rise in gist-directed intrusions; specifically, the delay-by-error-type interaction (the increase in lure false-alarm rate per unit delay) will be positive and significant for humans and statistically indistinguishable from zero for transformers, at matched hit rate.
>
> **Kill.** If, at matched hit rate, transformer in-context retrieval produces critical-lure false alarms that rise with effective delay at a rate whose confidence interval overlaps the human rate, and in the schema-consistent direction, the dissociation is dead: in-context memory carries the temporal-closure signature after all, and the paper is wrong.

I state the kill condition in the model's favor deliberately. It is not enough for transformers to intrude a little; the intrusions must fail to scale with delay, and must not match the human slope. If they scale and match, I concede.

6. What the outcomes would mean, including the ones that complicate the story

A clean confirmation (human slope positive, model slope near zero, at matched hit rate) would establish that the two systems occupy different error regimes and that "AI remembers more literally" is a claim about error structure, not about capacity. It would also give a practical readout: an in-context retrieval failure is more likely a placement problem (move the fact to an edge, or route it through the retrieval heads more reliably) than a corruption problem, whereas a human recall failure is more likely a reconstruction

problem that no amount of re-presentation at retrieval will simply reverse.

Several outcomes would complicate the story, and I want them on the record because a result that cannot surprise its author is not worth running.

First, models might show gist intrusions that scale with token distance but not with serial position, or the reverse. That partial pattern would be genuinely informative: it would say the reconstructive signature, if present, is tied to one architectural pressure (say, dilution over many intervening tokens, which degrades the copy operation) rather than to a general time-like variable. I would report it as a partial dissociation and revise the claim to name the specific pressure.

Second, chain-of-thought or scratchpad prompting might induce human-like intrusions by making the model re-encode the list in its own words before the probe, since retrieval heads are known to drive reference-heavy reasoning (FACT: Wu et al. 2024). That would not kill the core claim about raw in-context recall, but it would locate a regime where models cross into reconstructive behavior, which is exactly the sort of boundary the paper should map rather than paper over.

Third, larger or differently trained models might behave differently from smaller ones. The protocol is cheap enough to run across a model-size sweep, and the honest report is a curve, not a single point.

Fourth, and most likely to bite: equating difficulty is hard, and a critic could argue the match failed. This is why the design pre-registers the matching procedure and reports the hit-rate curves alongside the error-type curves, so a reader can see whether difficulty was in fact equated before trusting any dissociation. If the match visibly fails, the result is inconclusive rather than confirmatory, and I would say so.

7. Limits, firewall, and the wager

The firewall first, because it is the rule I least want to violate. Everything measured here is functional. "Reconstructive drift," "clean miss," and "temporal-closure signature" are descriptions of input-output behavior and, at most, of the attention operations that produce it. Nothing in this design speaks to whether a transformer, or for that matter a human, has any experience of remembering, of the felt confidence that accompanies a false memory, or of anything at all. When a person falsely recognizes "sleep," there is something it is like to be sure. I make no claim that there is or is not anything it is like to be a transformer reporting a list. The dissociation I am after is about the shape of errors, and it would be equally true or false whether or not the machine has an inner life. I flag this because "memory" is a word that smuggles phenomenology, and I am using it only for the functional faculty.

The limits, plainly. (1) DRM lures capture one important reconstructive error, semantic gist intrusion, but human reconstruction is broader (order errors, source confusions, boundary extension), and a full account would extend the taxonomy. (2) The comparison is at retrieval; the two systems differ radically at encoding and storage, and I am not claiming their mechanisms are comparable, only that their retrieval-error signatures are comparable and, I predict, different. (3) The parametric memory of the model, which ROME-style work shows to be associative and editable (FACT: Meng et al. 2022), may well carry something more like a gist signature; the paper explicitly does not test that, and a natural sequel

would ask whether parametric recall confabulates in the schema-directed, delay-scaled way in-context recall does not. (4) Human intervals run to days; model "delays" run to tens of thousands of tokens within a session. Arm (ii), sweeping serial position, is the bridge that puts both remainders in one protocol, but the axes are not identical and the analysis treats them as separate factors rather than pretending they are one.

The wager (WAGER, owned as such): I am betting that in-context memory is architecturally the wrong place to look for human-like reconstruction, and that the reconstructive, delay-scaled, gist-directed error signature will turn out to be specific to memory systems that re-form their contents over time rather than address fixed contents over space. I could be wrong in the most interesting way, which is that a large enough transformer, prompted to re-encode, reconstructs on a schedule that mirrors ours. If it does, the kill condition fires and the lesson is larger than the paper: reconstruction would be a property of a computation, not of biology, and the temporal-closure signature would belong to any system that rebuilds rather than re-reads. Either result is worth the afternoon it takes to run.

8. Conclusion

Human episodic memory forgets by rebuilding, and the rebuild leans on the gist, so its errors drift toward plausible fictions at a pace set by elapsed time. A transformer reading its context window forgets by not looking, and when it does not look it usually leaves a blank rather than a plausible fiction; its errors are set by position and salience, not by time. That is the difference this paper asks a lab to measure, using one shared instrument, difficulty held fixed, error type read out along a delay axis, with a stated slope that must come out flat for the model or the claim is false. The payoff is not a slogan about machines remembering better. It is a precise one: in-context memory can be more available without being uniformly available, and knowing which kind of failure you are looking at, a placement problem or a reconstruction problem, tells you what to do about it.

References

Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. Cambridge University Press.

Brainerd, C. J., and Reyna, V. F. (2002). Fuzzy-trace theory and false memory. Current Directions in Psychological Science, 11(5), 164-169.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157-173.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. arXiv:2202.05262 (Advances in Neural Information Processing Systems 35).

Murre, J. M. J., and Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. PLOS ONE, 10(7), e0120644.

Roediger, H. L., and McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(4), 803-814.

Wu, W., Wang, Y., Xiao, G., Peng, H., and Fu, Y. (2024). Retrieval head mechanistically explains long-context factuality. arXiv:2404.15574 (ICLR 2025).