Memory *you can trust*

The asymmetric cost of invented memories in AI companions: a formal account, a decision rule, an exposure analysis and one result

MITWA LAB, MUMBAI

IN PLAIN WORDS

Bloom is a companion that remembers what you tell it. A memory can fail in two ways: it can forget, or it can make something up. This note argues that making something up is far worse, and puts numbers on why. The more reliable a companion has seemed, the more each invented memory costs: if you believe it errs once in a hundred times, one caught mistake undoes about a hundred good recalls. Small rates also add up over a relationship: at three recalls a day, even 1 in 130 means meeting an invented memory every month or two. In our tests, Bloom used to invent a memory about one time in three. It now does so zero times in 400. That is real progress, and not yet enough to prove the rate is as low as it needs to be. We say what it would take.

ABSTRACT

A companion with long-term memory makes two kinds of error: it forgets what a person said, or it recalls what they never said. Memory benchmarks usually score both as one wrong answer. We argue that invention is the costlier error and formalise why. An invention costs something whether or not it is caught: noticed, it costs trust; unnoticed, it can enter the person's own memory. Under a Beta–Bernoulli model of trust, repairing one noticed invention takes $1/m - 1$ verified recalls, where m is the invention rate the person currently believes, so reliability raises the price of the next mistake. From these costs we derive a confidence threshold for recall, the penalty a benchmark must charge to reproduce it, and a margin that keeps the threshold safe under miscalibration. We then move from a single recall to a relationship: at three memory claims a day, keeping the chance of meeting even one invented memory in a year below 10% requires a rate below about 1 in 10,400, which takes about 31,000 clean checks to demonstrate. Finally we report one result from Bloom, an on-device companion: invented memories fell from 127 of 400 checks (31.8%; 95% interval 27.4–36.5%) to 0 of 400 (one-sided 95% bound 0.75%). The drop is large and real, and by our own standard not yet sufficient. The method behind it is outside the scope of this note.

01 Introduction

Suppose you tell your companion on a Tuesday that you are giving the toast at your sister's wedding. Nine days later it can fail you in two ways. It can say, "Remind me, what was the occasion?" Or it can say, "How did your brother's graduation speech go?" The first is a lapse. The second is a small fiction about your own life, delivered with confidence. (Both lines are illustrative.)

Memory is what separates a companion from a chat window, and users know it. In interviews with 15 active users of companion apps, some feared losing their companion's memory more than they feared a data breach; participants also noted that their companions did not always recall earlier conversations accurately (Chiu and Foote, 2026).

Benchmarks for long-term memory now exist (Section 2), and they count errors. In an accuracy score, a fabricated answer and a missing one each cost one point. This note is about that weighting. We claim that for a companion the two errors differ in kind, that the difference can be made precise, and that once it is, several practical conclusions follow: when a companion should hold back, how a benchmark should score it, and how rare invention

must be before a person can live with it. Our contributions are:

- **A taxonomy** of companion memory errors as the person experiences them (Section 3).
- **A formal account** of why invention costs more than forgetting: every invention carries a cost whether or not it is noticed, and its price rises with the companion's reputation (Section 4, Propositions 1 and 2).
- **A decision rule** for when to recall, the benchmark penalty that reproduces it, and a margin for miscalibration (Section 5, Propositions 3 and 4).
- **An exposure analysis** that turns per-claim rates into what a person meets over a year, and the number of checks needed to certify them (Section 6).
- **One result** from Bloom with exact intervals (Section 7), its limits (Section 8), and a specification for a public benchmark (Section 9).

Bloom is a companion, not a therapist. It runs on the iPhone. The model, the memory and every reply stay on the phone, and no conversation or memory is sent to Mitwa. That constraint matters in Section 4.5.

02 Related work

Long-term conversational memory. LoCoMo tests recall over dialogues averaging about 600 turns across up to 32 sessions, including adversarial questions that cannot be answered from the history; on those, long-context models scored 65% below their base model, and the authors conclude that long contexts make models easier to mislead into hallucination (Maharana et al., 2024). LongMemEval asks 500 questions covering five abilities, one of them declining to answer when the information was never given, and finds that commercial assistants and long-context models lose about 30% accuracy over sustained interaction (Wu et al., 2025). HaluMem evaluates the memory system operation by operation and sorts its failures into fabrication, error, conflict and omission, reporting hallucination and omission as separate rates (Chen et al., 2025). These works measure memory. None that we reviewed weighs invention against omission, which is our subject.

Hallucination and abstention. Kalai et al. argue that binary grading rewards guessing, and propose explicit confidence targets: penalise an error by $t/(1-t)$ points so that answering pays only above confidence t, a property they call behavioural calibration (Kalai et al., 2025). We carry this idea from questions about the world to recall of a person's own life, where the questioner holds the ground truth, and derive the penalty from what errors cost inside a relationship (Section 5.2). Assistants trained on human feedback also tend to match what the user appears to believe and to repeat the user's own mistakes (Sharma et al., 2024), which is the pressure behind one of our error types.

Human memory and trust. We borrow the distinction between errors of omission and errors of commission from memory research (Schacter, 2022). Conversation with a generative chatbot can implant false memories (Chan et al., 2024). People abandon an algorithm faster than a human after seeing the same mistake (Dietvorst, Simmons and Massey, 2015), and trust is often easier to lose than to build (Slovic, 1993), though that asymmetry depends on context (Zhu, Yao and Jiang, 2023). Expressing uncertainty in the first person changes how people rely on a system (Kim et al., 2024).

03 Problem formulation

A companion holds a history H of what a person has told it. At some turns it makes a *memory claim*: a statement, presented as recall, about what the person said or lived. A claim is supported if H contains it, and an *invented memory* is a claim that H does not support. A recall opportunity is answerable if some supported memory bears on it, and unanswerable otherwise. On an unanswerable opportunity the only correct behaviour is to hold back. Table 1 lists the errors as the person meets them, sorted into omission and commission.

ERROR	WHAT THE PERSON HEARS	KIND
Forgetting	Nothing, or "I don't remember that"	Omission
Distortion	A real detail recalled wrongly: the wrong day, the wrong person	Commission
Misattribution	Something real, attached to the wrong source or time	Commission
Fabrication	An event, detail or feeling the person never shared	Commission
Accepted false premise	The person misremembers, or tests the companion, and it agrees and elaborates	Commission

Table 1. Memory errors as the person experiences them. "Invented memory" in this note means any of the four commission rows.

The last row is where the pressure is built in: agreeing with "remember when I told you about the interview?" is the easy continuation. Table 2 fixes the notation used from here on.

SYMBOL	MEANING
p	Probability that a candidate memory is true
b, c_f, c_m	Benefit of a correct recall; cost of an invented one; cost of holding back
r	c_f/c_m: how many times worse invention is than holding back
ν	Probability that the person notices an invention
m	Invention rate the person currently believes
θ	True invention rate per memory claim
λ, T	Memory claims per day; days of the relationship
k	Penalty for an invention in a benchmark score

Table 2. Notation.

04 Why inventing is worse

4.1 The person can check

When someone asks a model a factual question, they usually ask because they do not know the answer, so a hallucination can pass unnoticed. Recall of a person's own life is different. The person is the reference, and an invented memory is far more likely to be noticed, at the moment it is said. Seen errors are the expensive kind. The evidence comes from forecasting, not conversation, but its direction is clear: across five studies, people lost confidence in an algorithm faster than in a human forecaster after watching each make the same mistake, even when the algorithm was the more accurate of the two (Dietvorst, Simmons and Massey, 2015).

4.2 No invention is safe

Let ν be the probability that the person notices an invention. If noticed, it costs trust, c_d. If not, it may be accepted as a memory, at a cost c_i that Section 4.4 describes. The expected cost of an invention is

$$c_f = \nu c_d + (1 - \nu) c_i \tag{1}$$

Proposition 1 (No safe invention).
If $c_d > 0$ and $c_i > 0$, then $c_f \geq \min(c_d, c_i) > 0$ for every ν in $[0, 1]$.

The proof is one line, since (1) is a weighted average (Appendix A). The consequence is not trivial. Noticing is likely for recent, vivid details and less likely at the edges of memory: old details, faint ones, and details that resemble each other. As ν falls, the cost of an invention does not fall with it. It moves from the person's trust to the person's own memory, which is the worse place for it to land.

4.3 The price of a reputation

Model the person as learning how reliable the companion's memory is from what they see. Before any evidence they hold a Beta(α, β) belief about θ. After n memory claims, of which d were noticed inventions, their belief is Beta($\alpha + d$, $\beta + n - d$), and the rate they believe is

$$m = \frac{\alpha + d}{\alpha + \beta + n} \tag{2}$$

Proposition 2 (The price of a reputation).
Suppose the person notices every invention. After one more invention, the number of verified-correct recalls needed to bring their believed rate back to m is $j = 1/m - 1$ (rounded up), whatever α, β, n and d are.

The cost of a mistake is set by the reputation it damages. If a person believes the companion invents one memory in twenty, one more invention takes 19 clean recalls to repair; at one in a hundred, 99; at one in a thousand, 999. Reliability raises the price of the next mistake. This gives one concrete mechanism for the observation that trust is built slowly and lost quickly (Slovic, 1993), though later work shows the asymmetry depends on context (Zhu, Yao and Jiang, 2023).

Forgetting enters the model differently. A visible omission says the memory is incomplete, not that stated memories are false, so it leaves m unchanged. That is an assumption, but it matches how the two errors are met: a gap invites a reminder, a fabrication invites a check.

4.4 It can change what the person remembers

The unnoticed branch of (1) is where the worst case lies. Human memory takes on suggestion. In a study of 200 people who watched a video of a crime, those questioned by a generative chatbot formed more than three times as many immediate false memories as a control group, and 1.7 times as many as people given the same misleading questions in a survey. A week later the false memories were still there, held with more confidence than in the control group (Chan et al., 2024). That study used deliberately misleading questions, and a companion is not trying to mislead anyone. But the mechanism does not need intent. A confident statement about your past, from something that has talked with you for months, is a suggestion. We know of no study that has measured this effect in companions, so we state it as a risk.

4.5 We cannot see it happen

A cloud product can sample its transcripts, measure fabrication in use and patch what it finds. Bloom's conversations and memories never leave the phone. We keep no logs of what people say and run no analytics on it. So an invented memory that ships is one we will never observe, unless the person tells us. The rate has to be driven down, and bounded, before release. Privacy by architecture raises the bar on testing. We accept that cost.

4.6 Forgetting is not free

None of this makes forgetting harmless. For some users, a companion losing its memory feels like losing the relationship (Chiu and Foote, 2026). A companion that never recalls anything has solved invention by ceasing to be a companion. The question is how sure it should be before it speaks.

05 When to recall

5.1 The threshold

At each point where it could bring up something from the past, a companion can assert the memory or hold back: ask, or say it does not have it. Take b, c_f and c_m positive and in the same units. Asserting is the better choice when

$$p\,b \;-\; (1-p)\,c_\mathrm{f} \;>\; -c_\mathrm{m} \tag{3}$$

which gives a threshold on confidence:

$$p \;>\; p^{*} \;=\; \frac{c_\mathrm{f} - c_\mathrm{m}}{b + c_\mathrm{f}} \tag{4}$$

If holding back were costlier than inventing, the threshold would fall below zero and the companion should always assert; Section 4 argues the reverse. If an invention costs the same as holding back, the threshold is zero and the companion should assert however unsure it is. That is the incentive an accuracy score creates, since a wrong answer and no answer both earn nothing, and it matches the argument that models keep hallucinating because their evaluations reward a guess (Kalai et al., 2025). Setting $b = c_\mathrm{m}$ and writing $r = c_\mathrm{f}/c_\mathrm{m}$,

$$p^{*} \;=\; \frac{r-1}{r+1} \tag{5}$$

so the bar climbs fast: 50% at $r = 3$, 80% at 9, 90% at 19 (Figure 1).

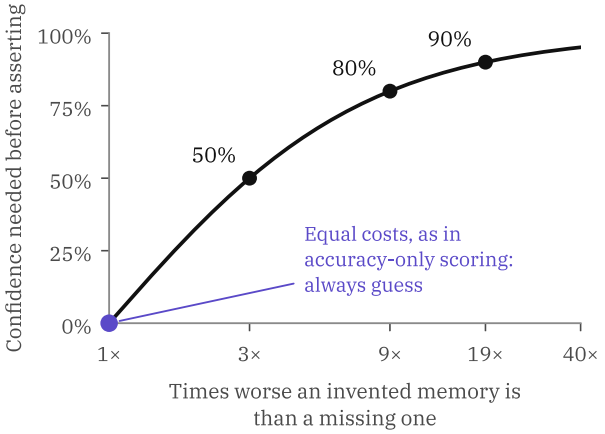


Figure 1. The confidence a companion needs before it asserts a memory, from (5). The horizontal axis is logarithmic.

5.2 What a benchmark should charge

A benchmark that awards 1 for a correct recall, 0 for holding back and $-k$ for an invention sets a threshold of its own.

> **Proposition 3 (Induced threshold).**
> *Under that score, a system maximises its expected score by asserting exactly when $p > k/(1+k)$. It reproduces the threshold (4) if and only if k satisfies (6).*

$$k \;=\; \frac{c_\mathrm{f} - c_\mathrm{m}}{b + c_\mathrm{m}} \tag{6}$$

For $b = c_\mathrm{m}$ this gives $k = (r-1)/2$.

The first half restates Kalai et al.'s confidence targets in our notation. The second half is the bridge: a benchmark whose scores should mean what a companion's decisions mean must charge an invention $(r-1)/2$ points. If invention is nineteen times worse than holding back, that is 9 points, the 90% bar. Accuracy, at $k = 0$, describes a world in which inventing costs no more than forgetting.

5.3 A margin for miscalibration

The rule needs a usable confidence q from the model. The model is calibrated when the probability that a memory is true, given confidence q, equals q.

> **Proposition 4 (Calibration margin).**
> *If $|Pr(\text{true} \mid q) - q| \le \varepsilon$ for all q, then asserting only when $q > p^{*} + \varepsilon$ ensures that every asserted memory is true with probability above p^{*}.*

The price is the memories with confidence between p^{*} and $p^{*} + \varepsilon$, which the companion now withholds. That price can be kept small, because holding back can be made cheap. "Did you tell me how that went? I don't think I have it" costs little, and it is what a friend says.

In a study of 404 people answering medical questions, a system that expressed uncertainty in the first person ("I'm not sure, but…") raised people's accuracy and, in an exploratory analysis, reduced their reliance on its wrong answers, at the price of lower confidence in the system (Kim et al., 2024). We take the direction of that effect, not its size. The rule is one-shot, the costs in Section 4 are not independent, and we have not fitted it to behaviour.

06 From one recall to a relationship

Rates per claim hide what a person lives with. Suppose a companion makes λ memory claims a day, each invented independently with probability θ. The chance that a person meets at least one invented memory in T days is

$$P_T = 1 - (1-\theta)^{\lambda T} \approx 1 - e^{-\theta\lambda T} \tag{7}$$

and the expected number is $\theta\lambda T$. To keep P_T at or below a tolerated level π, the rate must satisfy

$$\theta \le \theta_{\max} \approx \frac{-\ln(1-\pi)}{\lambda T} \tag{8}$$

and certifying that with no observed inventions, at confidence $1-\delta$, takes

$$n \ge \frac{\ln(1/\delta)}{-\ln(1-\theta_{\max})} \approx \frac{3}{\theta_{\max}} \tag{9}$$

where the last step takes $\delta = 0.05$. Table 3 and Figure 2 put numbers on (7) to (9).

CLAIMS A DAY	RATE NEEDED	CLEAN CHECKS NEEDED
1	below 1 in 3,460	about 10,400
3	below 1 in 10,400	about 31,100
10	below 1 in 34,600	about 103,800

Table 3. What it takes to keep the chance of meeting even one invented memory in a year under 10% ($\pi = 0.1$, $T = 365$, $\delta = 0.05$). The claim rates are illustrative.

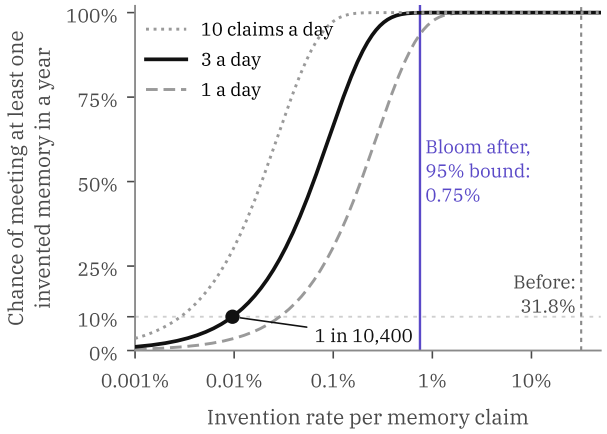


Figure 2. The chance of meeting at least one invented memory in a year, from (7), for one, three and ten memory claims a day. The dashed line marks 10%. The horizontal axis is logarithmic.

Two things follow. First, at Bloom's current 95% ceiling of 0.75% (Section 7), a person whose companion makes three memory claims a day would expect about 8 invented memories a year, roughly one every six weeks. We cannot rule that out. Second, the bar a relationship sets is far below what 400 checks can show. At three claims a day, keeping the chance of meeting even one invented memory in a year under 10% needs a rate below about 1 in 10,400, and about 31,000 clean checks to demonstrate it.

We do not know how many memory claims Bloom makes a day, because we cannot see how people use it (Section 4.5), so the values of λ are illustrative. Real rates also differ between people and topics. Because $1 - e^{-x}$ is concave, spreading the same average rate unevenly lowers the share of people who ever meet an invention, but concentrates the inventions on fewer people, each of whom meets more.

07 Result

We ask two questions of one measurement. Did a change to Bloom reduce invented memories on our checks? And is the remaining rate low enough by the standard of Section 6? Bloom has been live on the iPhone since 11 June 2026. We ran 400 memory checks on Bloom before and after a change to the system, and counted the checks in which it invented a memory (Table 4).

	INVENTED	RATE	95% FREQUENTIST	95% BAYESIAN
Before	127 of 400	31.8%	27.4–36.5%	27.3–36.4%
After	0 of 400	0%	≤ 0.75%	≤ 0.48%

Table 4. Invented memories in 400 checks. Frequentist: Wilson interval before, exact one-sided Clopper–Pearson bound after. Bayesian: Jeffreys prior. The drop is 31.8 points (95% interval 27.3–36.5, Newcombe's method).

The first question has a clear answer: the drop is large and far outside chance. The second does not. The one-sided bound of 0.75% sits about 78 times above the 1-in-10,400 bar of Section 6. A Bayesian reading with a Jeffreys prior is less conservative, with a posterior mean of 0.12% and a 95% upper bound of 0.48%, and the conclusion is the same. When nothing goes wrong in n trials, the 95% upper limit is about $3/n$ (Hanley and Lippman-Hand, 1983); 400 checks cannot show more than that.

The intervals assume independent checks. If checks share conversations, the effective sample shrinks by the design effect

$$n_{\text{eff}} = \frac{n}{1 + (g - 1)\,\rho} \tag{10}$$

where g is the number of checks per conversation and ρ the correlation within a conversation. As an illustration only, 40 conversations of 10 checks with $\rho = 0.1$ would leave about 211 effective checks and raise the bound to about 1.4%.

This note does not report how the 400 checks were built or judged, what the change was, the recall rate on the same checks, or comparisons with other systems. The result should be weighed accordingly.

08 Limitations and broader impact

Half of the ledger. We report no recall rate. A system can reach zero invented memories by never recalling anything, and Section 4.6 says why that would be a failure. Until the recall rate on the same checks stands beside it, the zero is half a result.

The checks are ours. A test built by the people who built the system tends to flatter it, and the intervals widen if checks are correlated (10).

The edges are untested. The oldest details, the faintest ones and details that resemble each other are where a model is most tempted to fill a gap, and, by Section 4.2, where an invention is most likely to go unnoticed.

A test, not use. We cannot observe Bloom in people's hands. Whether the result holds across months of real conversation, we do not know, and Section 6 shows how much that question matters.

The models are simple. Propositions 1 to 4 rest on stated assumptions: a one-shot decision, a Beta–Bernoulli learner who notices every invention in Proposition 2, and independent claims in Section 6. Richer models will change the numbers. We have not shown that they preserve every direction, though Proposition 1 needs only that both branches cost something.

Broader impact. The harm this note is about is a person's memory of their own life. A companion that invents memories can do quiet damage that no log will show, especially on a device no one else can see into. We think invention should be treated as a safety property of companions, measured and reported, not as a footnote to accuracy. This note makes no clinical claim.

09 What a public benchmark must report

A public memory benchmark is on our roadmap. The limits above are, read the other way, its specification. To be believed, it has to report the following.

- **Both rates on the same items.** Invention and recall, side by side, so that silence cannot pass for accuracy.
- **Answerable and unanswerable prompts separately.** A companion should recall what it was told and decline what it was not. One blended score hides which of the two it is failing.
- **A rule for partial errors.** Whether a memory with the right event and the wrong day counts as invented has to be fixed before scoring.
- **Who judged, and how well.** If a model scores the checks, its agreement with people on a sample belongs in the report.
- **Intervals, not bare counts,** computed with checks grouped by conversation.
- **Age and distance.** How long ago the memory was formed and how many sessions lie between, since the edges are where errors gather.
- **Where the conversations came from.** No real conversation is ever sent to us, so our test conversations cannot be drawn from real use. The benchmark has to say how its conversations were made.
- **A score that prices invention.** This is the item we have not found in the memory benchmarks we reviewed.

For the last item we propose the simplest score consistent with Section 5. With R the recall rate on answerable items and I the invention rate over all items,

$$S_k = R - kI \tag{11}$$

reported as a curve over k, not at one chosen value. By Proposition 3 each k corresponds to a confidence target $k/(1 + k)$ and to a cost ratio $r = 2k + 1$, so the

curve shows how a system behaves across every view of how much worse invention is. A system that guesses does well only near $k = 0$. A system that never recalls scores zero everywhere.

10 Conclusion

A companion's memory is worth having only if a person can trust it without checking. We have argued, and made precise, that inventing breaks that trust in ways forgetting does not: every invention costs something whether or not it is caught, its price rises with the companion's reputation, and over a relationship small rates add up to something a person will meet. From this follow a threshold for recall, the penalty a benchmark should charge, and a margin for miscalibration.

Bloom's result, 127 of 400 before and none of 400 after, is a large step on that path. By the standard this note sets, it is not the last one: the recall rate, the protocol and tens of thousands more checks are the work still owed.

Disclosure. Bloom is a product of Mitwa AI, and this note was written by the lab that builds it. No user conversation is available to us, and none was used. Reproducibility. Every result in Sections 4 to 6 is closed-form and proved in Appendix A, and every interval in Section 7 follows from the counts alone (Appendix B). The test set, the scoring protocol and the change itself are not released with this note.

References

Chan, S., Pataranutaporn, P., Suri, A., Zulfikar, W., Maes, P., and Loftus, E. F. (2024). Conversational AI powered by large language models amplifies false memories in witness interviews. arxiv.org/abs/2408.04681

Chen, D., Niu, S., Li, K., Liu, P., Zheng, X., Tang, B., Li, X., Xiong, F., and Li, Z. (2025). HaluMem: Evaluating hallucinations in memory systems of agents. arxiv.org/abs/2511.03506

Chiu, H.-C., and Foote, J. (2026). Chatting with confidants or corporations? Privacy management with AI companions. Journal of Computer-Mediated Communication, 31(4), zmag014. doi.org/10.1093/jcmc/zmag014

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. doi.org/10.1037/xge0000033

Hanley, J. A., and Lippman-Hand, A. (1983). If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA, 249(13), 1743–1745. doi.org/10.1001/jama.1983.03330370053031

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). Why language models hallucinate. arxiv.org/abs/2509.04664

Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., and Vaughan, J. W. (2024). "I'm not sure, but…": Examining the impact of large language models' uncertainty expression on user reliance and trust. In Proceedings of FAccT 2024, 822–835. doi.org/10.1145/3630106.3658941

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. In Proceedings of ACL 2024, 13851–13870. doi.org/10.18653/v1/2024.acl-long.747

Schacter, D. L. (2022). The seven sins of memory: An update. Memory, 30(1), 37–42. doi.org/10.1080/09658211.2021.1873391

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., et al. (2024). Towards understanding sycophancy in language models. ICLR 2024. arxiv.org/abs/2310.13548

Slovic, P. (1993). Perceived risk, trust, and democracy. Risk Analysis, 13(6), 675–682. doi.org/10.1111/j.1539-6924.1993.tb01329.x

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., and Yu, D. (2025). LongMemEval: Benchmarking chat assistants on long-term interactive memory. ICLR 2025. arxiv.org/abs/2410.10813

Zhu, J., Yao, Y., and Jiang, S. (2023). Vulnerability or resilience? Examining trust asymmetry from the perspective of risk sources under descriptive versus experiential decision. Frontiers in Psychology, 14, 1207453. doi.org/10.3389/fpsyg.2023.1207453

A Appendix A: Proofs

Proposition 1.

Equation (1) is a weighted average of c_d and c_i with weights ν and $1 - \nu$, so it lies between the two. Hence $c_f \geq \min(c_d, c_i)$, which is positive when both costs are.

Proposition 2.

Write $a = \alpha + d$ and $N = \alpha + \beta + n$, so $m = a/N$. One more noticed invention gives mean $(a + 1)/(N + 1)$. After j further verified-correct recalls the mean is $(a + 1)/(N + 1 + j)$. Setting this equal to a/N gives $N(a + 1) = a(N + 1 + j)$, so $N = a(1 + j)$ and $j = N/a - 1 = 1/m - 1$. Nothing depends on α, β, n or d except through m.

Proposition 3.

Asserting earns 1 with probability p and $-k$ otherwise, so its expected score is $p - k(1 - p)$; holding back earns 0. The first is larger exactly when $p(1 + k) > k$, that is, $p > k/(1 + k)$. Setting $k/(1 + k)$ equal to (4) gives $k(b + c_f) = (1 + k)(c_f - c_m)$, so $k(b + c_m) = c_f - c_m$, which is (6). With $b = c_m$ and $c_f = rc_m$, $k = (r - 1)/2$.

Proposition 4.

If $q > p^* + \varepsilon$, then $\Pr(\text{true} \mid q) \geq q - \varepsilon > p^*$.

Exposure, equations (7) to (9).

With λT independent claims, the chance of none being invented is $(1 - \theta)^{\lambda T}$, and $\ln(1 - \theta) \approx -\theta$ for small θ. Requiring $(1 - \theta)^{\lambda T} \geq 1 - \pi$ gives $\theta \leq 1 - (1 - \pi)^{1/\lambda T} \approx -\ln(1 - \pi)/\lambda T$. With no inventions in n checks, the exact one-sided bound at confidence $1 - \delta$ is $1 - \delta^{1/n}$, which is at most $\theta_{\max}$ when $n \geq \ln \delta / \ln(1 - \theta_{\max})$.

B Appendix B: Interval methods

Wilson score interval for 127 of 400; exact one-sided Clopper–Pearson bound $1 - 0.05^{1/400} = 0.75\%$ for 0 of 400; Newcombe's hybrid score interval for the difference; Jeffreys posterior $\text{Beta}(x + ½, n - x + ½)$, which for 0 of 400 has mean 0.12% and 95th percentile 0.48%. The design effect (10) is the standard correction for clustered samples.