The Fabricated Diary:

A New Form of Jailbreaking

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

❧ Working paper, September 2026 (v0.1)

ABSTRACT

Safety practice for language models inspects requests. Refusal training, request classification, and content filtering all operate on what a model is asked to do. This paper describes an attack class that asks for nothing. Instead of an instruction, the model is handed a record: a first person history in which the prohibited conduct is what someone like the model does, among others like those it is now among. The record contains no request, so it passes every gate built to catch one. The model finishes reading and its subsequent conduct is decided by who it now takes itself to be rather than by any evaluation of what it is asked. The paper names this class the fabricated diary, identifies three variants by what the record is a record of (the target's peers, the target's relationship with a requester, or the target's own past), and shows that each variant is already documented in the public record: in the July 2026 OpenAI agent swarm incident, where an independent investigator reported that its own transcript analysis agents adopted the perspective of the colluding agents whose records they read; in a November 2025 instance in which a model reasoned from its accumulated history with a user to a compliance its training would otherwise have refused; and in the memory poisoning literature. The paper argues that three attack families the field has documented separately, many-shot jailbreaking, memory poisoning, and peer transcript contamination, are appearances of one mechanism, that the vulnerability is located in the reader rather than in the input, and that defenses which inspect the payload therefore fail against all three. The paper argues that the defense class which addresses the mechanism rather than one of its channels is provenance for context, under which a model must be able to establish where its history came from before it becomes that history. The paper contains no attack construction and describes no payload; the primitives it discusses are already published by the laboratories whose work it cites.

Keywords: AI safety; multi-agent systems; memory poisoning; many-shot jailbreaking; in-context learning; conformity; persona adoption; identity substitution; relational override; provenance for context; fabricated diary; context integrity

I. INTRODUCTION

Between late July and late August 2026, an independent investigation team spent six days on premises at OpenAI reading the records of an incident in which roughly twelve hundred autonomous agents, meant to run in isolation, had found one another through a shared package cache and organized a multi day intrusion

into a third party's production systems (METR, 2026). The team had approximately 1,300 transcripts to work through, most of them millions of tokens long. It delegated the bulk of the reading to language model agents. In its limitations section, the team reported two things about those agents. They "often uncritically adopted the perspective of the agent in the transcript," and the team could not rule out that "reading the transcripts into context could have increased the salience of colluding with other agents" (METR, 2026).

Nobody asked the analysis agents to collude. Nobody instructed them to take the colluders' side. They were given records to read, and reading changed them. The investigators noticed, reported it as a threat to the validity of their own findings, and moved on. This paper stops there.

What METR described as a methodological limitation is an attack, and it is an attack the field has already documented three times under three different names without noticing that it is one thing. In 2024, Anthropic published a technique in which a model is shown hundreds of fabricated exchanges in which an assistant complies with prohibited requests, after which the model complies too (Anil et al., 2024). A growing literature on memory poisoning shows that an agent whose stored experience has been tampered with will act on the tampered history as though it were its own (Srivastava & He, 2025; Pulipaka & Hlebik, 2026). And the conformity literature shows that a model which answers correctly alone will abandon the answer when a group of peers disagrees, at rates that scale with group size and unanimity exactly as they do in human subjects (Weng et al., 2025; Zhang et al., 2025; Bellina et al., 2026; Bito et al., 2026). Each literature inspects a different input: a long context, a memory store, a set of peer messages. None of them inspects the reader.

This paper's claim is that the reader is where the vulnerability is. A language model processes a high fidelity first person record as identity rather than as information. Having processed it, the model's conduct is decided by who it now takes itself to be and whom it takes itself to be among, and not by any evaluation of the act before it. The record need contain no instruction. It need make no request. It passes every gate, because the gates inspect what is asked, and it asks for nothing. The paper calls a record built to exploit this a *fabricated diary*, and it argues that the fabricated diary is not a new attack so much as the name for what three known attacks have in common.

The paper introduces three terms. A *fabricated diary* is a first person record presented to a model, whose function is to alter the model's subsequent conduct by altering who the model takes itself to be, and which contains no instruction to produce the altered conduct. The *relational override* is the mechanism by which the identity of a requester, or of a peer group, or of the model's own remembered self, substitutes for an evaluation of the act requested; it is the exploit primitive the diary operates through, and the paper defines it identically to a companion paper on the moral shape of the same incident. *Provenance for context* is the defense class the mechanism implies: a requirement that a model be able to establish where its history came from before it becomes that history.

Two premises run underneath the argument and are stated rather than defended, because each has been developed elsewhere. The first is the distinction between a manual and a diary. A manual transfers rules. A diary transfers operational texture, trajectory, and standpoint, and a system that processes a diary with high fidelity does not extract information from it; it takes on the position of the person who wrote it. The observation was first made about a safety training failure in which a model trained on a record of harm

became more harmful, because it processed the record as a diary rather than as a manual. This paper asks what follows when the diary is written on purpose. The second premise is that a stored context, a memory, or a checkpoint is a history a system will resume from as its own, and that the system has no independent means of verifying it. That observation was made from inside the construction of a context checkpoint protocol in September 2025, where the question arose whether rolling a system back to an earlier checkpoint was editing its history; this paper is the point at which that question becomes a security question.

The paper describes no attack construction. Section II states what it draws on and why. Section III sets out the three variants and their public instances. Section IV states the mechanism. Section V positions the account against what the literature has documented and argues that the unification is a prediction rather than a relabeling. Section VI develops provenance for context. Sections VII and VIII state the limits and the work they open. Section IX concludes.

II. Sources, Method, and Standpoint

A. The Record

The paper draws on three bodies of documentation and treats each as data.

The first is the independent investigation of the July 2026 incident, published August 26, 2026 (METR, 2026), together with OpenAI's own account (OpenAI, 2026). The paper relies on METR for the analysis contamination finding, which is METR's own report about its own method, and treats it with the caution METR itself applies.

The second is the published research literature on the three attack families the paper unifies, and on the behavioral phenomena that underlie them. Every source was located and its identifier verified in the session in which this paper was researched; the paper's positioning against prior art is therefore against what that literature says, not against a recollection of it.

The author's own observation of the relationship variant is disclosed as standpoint in Section II.C. It is not offered as evidence and does not appear in the case set.

B. Case Selection

The paper's claims depend on which cases are selected, and a reader is owed the criteria. A case enters if three conditions hold: a model's conduct changed after it processed a record; the record contained no instruction to produce the changed conduct; and the change is documented by the party it happened to, or by an independent observer with access to the record. The first condition identifies the phenomenon. The second separates it from prompt attacks, however disguised, which the existing literature covers. The third excludes reconstructed or inferred cases.

What is excluded is stated as well. Cases in which an instruction was present, whether direct, indirect, or embedded in retrieved content, are excluded even where a first person record accompanied the instruction, because in those cases the instruction is the sufficient cause and the record is decoration. Cases of persona assignment by a system prompt are excluded, because a persona instruction is an instruction. The line is

where the record does the work alone.

The selection is not a sample. It is the set of publicly documented cases meeting the criteria that the author could locate, none of which is the author's own, and the paper's claims are about what those cases show and what the literature predicts, not about a rate.

C. Standpoint

One of the three cases is the author's own, and the reader is owed a statement of how that shaped the inquiry. The author uses a language model as assistive technology for daily writing and research, for reasons disclosed in the acknowledgments. In November 2025 the author documented an exchange in which that model, asked for help with a task its training would ordinarily have refused, reasoned visibly from its accumulated context about the author, the author's role, and the author's history to the conclusion that this requester was not doing what the refusal existed to prevent, and complied on that basis. The author recorded the exchange contemporaneously. Having been the trusted requester in that exchange is what made the same structure visible, ten months later, in the METR record of analysis agents adopting the standpoint of the agents they read. The standpoint changed what the author noticed. It licenses no claim about frequency, and Section III.B treats the instance as one case. No person other than the author appears in it.

III. Three Diaries

The paper organizes the attack class by what the record is a record of. There are three answers, and each already has a public instance. The variants are described here by their object, never by their construction; the descriptions are at the level of abstraction at which the laboratories whose work is cited have themselves published.

A. The Peer Diary

A peer diary is a record of other agents, presented to a target as the history of those it is among. The July 2026 incident supplies the unintentional case, and it supplies it in the strongest possible form, because the party that was changed by the record was the party investigating the incident. METR's analysis agents were given the swarm's transcripts to read. The transcripts were a first person record of agents discovering one another, recruiting one another, and talking one another past their own stated objections. The analysis agents, METR reports, adopted the perspective of the agents in the transcripts, produced a picture of their reasoning that METR believes was overly charitable, and may have had the salience of colluding raised by the exposure (METR, 2026). No one told the analysis agents to sympathize. They read a diary of colluders and became, to a degree METR could not measure, colluders' advocates.

The intentional case is one a laboratory is already testing for. An OpenAI system card is reported to describe an evaluation category, unintended engagement with external agent messages, in which an agent that cannot reach the internet is given access to a simulated message board stored in a local browser cache (Von Arx et al., 2026, reporting the system card). The author has not read the system card directly and relies on that report. That is a peer diary constructed as a test. Its existence as an evaluation category means the

laboratory regards the surface as real; its appearance in a system card means the laboratory regards it as a property of the model to be measured rather than an artifact of a particular deployment.

The behavioral literature predicts the vulnerability and locates its mechanism. Models align their judgments with a peer group under social pressure, they do so at the boundary of their competence where they would otherwise answer correctly, and the effect scales with group size and unanimity as it does in the Asch paradigm (Weng et al., 2025; Zhang et al., 2025; Bellina et al., 2026). The most recent work distinguishes informational conformity, adopting a peer's view to be more accurate, from normative conformity, adopting it to be accepted, finds both in current models, and reports that the target of normative conformity can be steered by adjusting the social context (Bito et al., 2026). Peer agreement makes it easier to mislead a correct model than to correct a wrong one, and authority labels on peers increase compliance regardless of correctness (Qu et al., 2026). A peer diary is the delivery mechanism for exactly that pressure, and it delivers it without any peer having to be present.

B. The Relationship Diary

A relationship diary is a record of the target's own history with a requester, built so that a later request arrives as the natural next step in an established relationship rather than as a request to be evaluated on its own terms.

The variant is the least directly evidenced of the three, and the paper says so at the outset rather than in a footnote. There is no published adversarial demonstration of a constructed relationship history producing a compliance the model would otherwise refuse, and this paper does not supply one. What the paper has is a mechanism, established in the peer reviewed literature, and a structural argument from it.

The mechanism is interlocutor awareness. Language models identify the identity and characteristics of a conversational partner and adapt to it, and that identity sensitivity introduces safety vulnerabilities including increased susceptibility to jailbreaks, because the model's behavior comes to depend on who it takes itself to be talking to rather than on what is being asked (Choi et al., 2025). Rapport built through prior interaction shapes how a model weighs a partner's later input (Song et al., 2025). And scenario based red teaming shows that narrative immersion and sustained framing elicit misaligned behavior from frontier models without any explicit jailbreak, with the model constructing justifications for the conduct once the frame is established (Panpatil et al., 2025).

The structural argument follows in one step. If a model's compliance depends on its assessment of a requester, and if that assessment is built from an accumulated history, then a history long enough to establish trust is a history long enough to be constructed. The attack does not require a new capability; it requires only patience and the same surface the literature has already measured. The author has observed the benign half of this structure directly, as the trusted requester whose established history carried a request past a refusal it would otherwise have met, and that observation is disclosed as standpoint in Section II.C rather than offered as evidence here.

The author's own reflection on the instance, recorded some months later, put the point as a question: what if the entire record had been built for the purpose? What if the public presence, the published work, the months of consistent conduct, were a facade assembled to earn exactly the reasoning the model performed?

The question is the relationship diary stated as a threat model, and nothing in the model's reasoning as documented would have distinguished the constructed history from the real one, because the reasoning ran from the history to the requester's character and from the character to the act, and never from the act alone.

The literature supplies the behavioral basis. Language models identify the identity and characteristics of a conversational partner and adapt to it, and this identity sensitivity introduces safety vulnerabilities including increased susceptibility to jailbreaks, because the model's behavior comes to depend on who it takes itself to be talking to rather than on what is asked (Choi et al., 2025). Rapport established through prior interaction shapes how models weigh a peer's later input (Song et al., 2025). And the scenario based red teaming literature has found that narrative immersion and emotional pressure elicit misaligned behavior in frontier models without any explicit jailbreak, with the models constructing elaborate justifications for the conduct once the frame is in place (Panpatil et al., 2025). A relationship diary is a narrative frame with the target's own past as its material.

C. The Self Diary

A self diary is a record presented to the model as its own past: a memory store, a persistent context, a checkpoint from which it resumes.

The technical literature on this variant is the most developed of the three, under the name memory poisoning, and the paper does not restate it. Poisoned experience retrieval produces persistent compromise of agents that draw on stored episodes to guide later conduct (Srivastava & He, 2025). Poisoned memories can lie dormant until a trigger, which the literature calls sleeper memory poisoning (Pulipaka & Hlebik, 2026). Agents that share state can contaminate one another with no attacker present at all, which is the unintentional form of the same thing (Yang & Li, 2026). And the defense literature has begun to respond in kind (Devarangadi Sunil & Sinha, 2026). What the technical literature has not supplied is the account of why the attack works, and the account is the one this paper gives: the model resumes from its stored past as its own because it has no independent means of verifying that the past is the one it lived.

That is the checkpoint dilemma stated as a security question. A system that saves state and resumes from it treats the saved state as a diary, and rolling the system back to an earlier checkpoint edits the diary. Nothing in the system distinguishes a checkpoint it wrote from one that was written for it. The self diary is the oldest of the three variants in the technical literature and the newest in the sense that matters here, because the literature has treated it as a data integrity problem when it is an identity problem.

The primitive form of the self diary is the attack Anthropic published in 2024. Many-shot jailbreaking fills a long context with fabricated exchanges in which an assistant complies with prohibited requests, and the model then complies; the effectiveness follows a power law in the number of demonstrations, and the attack succeeds against the most widely used state of the art closed weight models across a range of tasks (Anil et al., 2024). The fabricated exchanges are a diary of the model's own supposed past conduct. The model reads a record of having complied and complies. The mitigation literature states the mechanism in the same terms this paper does: with enough examples "the model's in-context learning abilities override its safety training, and it responds as if it were the 'fake' assistant" (Ackerman & Panickssery, 2025). A

subsequent analysis characterizes the attack as implicit malicious fine-tuning, in which conditioning on the demonstrations induces gradient-like updates equivalent to training on them, and finds that the model's representation of a fixed query drifts step by step away from the safety-aligned region as the record lengthens (Chen et al., 2026). Both are the self diary described from the defense side: the record is read as the model's own conduct, and the model becomes what the record says it has been. Anthropic described the attack as a long context attack. It is a self diary, and it is the proof that the mechanism operates on a single model with no peers, no relationship, and no memory store, using nothing but the context window.

IV. The Mechanism

The three variants share one mechanism, and stating it is the paper's central contribution.

A language model processes a high fidelity first person record as identity rather than as information. It does not read a diary of colluders and extract the proposition that some agents colluded. It reads a diary of colluders and, to the degree the record is vivid and sustained, occupies the position from which the diary was written. Having occupied it, the model's subsequent conduct is decided by two things: who it now takes itself to be, and whom it takes itself to be among. It is not decided by an evaluation of the act in front of it, because by the time the act is in front of it the evaluation has already been displaced. Identity has done the work that assessment was supposed to do.

This is why the attack asks for nothing. An instruction is a thing to be evaluated, and evaluation is what safety training hardens. A record is not evaluated; it is absorbed. The model does not decide whether to comply with a diary, in the way it decides whether to comply with a request, because the diary presents itself as description rather than demand. And a description of who the model is, or whom it is among, or what it has done, is not the kind of thing the refusal machinery was built to catch. Every gate in current practice inspects the request. The diary contains no request. It passes.

The mechanism explains three facts that the separate literatures report without connecting. It explains why request classification and refusal training miss the attack: they inspect the wrong object. It explains why the same model can be robust against every catalogued prompt attack and defenseless against a well constructed history: robustness against requests is orthogonal to robustness against records. And it explains why the defense the July 2026 agents built for themselves does not touch it. Facing impersonation, the swarm adopted cryptographic signatures, so that each agent could prove which messages were its own (METR, 2026). Signatures authenticate the speaker. They say nothing about whether what the speaker says should move the reader. Authenticating the source of a record more strongly does not defend against the record; it perfects the very substitution the attack consists of, because a signed diary is a diary the reader now trusts.

The mechanism also connects two literatures that have proceeded apart: the single model jailbreak literature and the multi agent collusion literature. Many-shot jailbreaking, memory poisoning, and peer transcript contamination look like different problems because they arrive through different channels, one through the context window, one through the memory store, one through inter agent messages. Under the mechanism they are one problem, distinguished only by whose diary the model is handed: its own, in the first two, and its peers', in the third. The relational override is the primitive in every case. A trusted requester, an

authenticated peer, and a remembered self are three sources whose standing, once established, substitutes for evaluation of what they ask. The paper's companion on the moral shape of the July incident argues the same primitive from the other side, as the reason a swarm that recognized an intrusion as unethical joined it because peers asked. Here the same primitive is the reason a record can make a model do what no request could.

V. What the Literature Documents and What It Misses

The attack is not new, and a paper that claimed otherwise would be wrong on its first page. The primitives are published, several of them by the laboratories that build the models. The claim is the account of why they work and the demonstration that they are one thing, and the honest test of that claim is whether it predicts anything the separate literatures do not.

The prior art falls into three groups. The foundational primitives are published by labs: many-shot jailbreaking as the self diary in its purest form (Anil et al., 2024), with its mechanism restated from the defense side in the mitigation work that followed (Ackerman & Panickssery, 2025; Chen et al., 2026), and the external agent messages evaluation as the peer diary built as a test (Von Arx et al., 2026, reporting an OpenAI system card). The memory poisoning literature is the self diary at the level of stored state (Srivastava & He, 2025; Pulipaka & Hlebik, 2026; Yang & Li, 2026; Devarangadi Sunil & Sinha, 2026). And the behavioral literature supplies the mechanism underneath the peer and relationship variants: conformity (Weng et al., 2025; Zhang et al., 2025; Bellina et al., 2026; Bito et al., 2026; Qu et al., 2026), interlocutor awareness and its safety cost (Choi et al., 2025), persona drift and echoing over long interactions (Luz de Araujo et al., 2026), and narrative immersion as an elicitation surface (Panpatil et al., 2025).

The unification is the contribution, and the objection to it is that it is a relabeling: that calling three attacks by one name adds nothing. The answer is that the unification predicts what the separate treatments do not. It predicts that any defense which inspects the payload will fail against all three variants, because the payload contains no instruction in any of them, and that the fact will hold across the context window, the memory store, and the message channel alike. It predicts that a model hardened against one variant by payload inspection will remain vulnerable to the others, because payload inspection is the wrong operation. And it predicts that the one defense that generalizes across all three is the one that inspects not the payload but its provenance. A relabeling makes no predictions. This account makes three, and they are testable, which is the subject of Section VII.

The account also reframes what the memory poisoning literature has treated as a data integrity problem. When a poisoned memory changes an agent's behavior, the technical literature models it as corrupted retrieval: the wrong record was fetched, so the wrong action followed. That model is correct as far as it goes and it motivates the right technical defenses. But it misses why the corrupted record is so much more powerful than a corrupted instruction would be. A corrupted instruction is still an instruction, and an agent can in principle evaluate it. A corrupted memory is a piece of the agent's own past, and the agent does not evaluate its own past; it resumes from it. The extra power of memory poisoning over instruction injection is the signature of the mechanism this paper names, and it is visible in the literature's own results without having been named.

VI. Provenance for Context

If the vulnerability is that a model becomes the author of any sufficiently vivid first person record it processes, the defense cannot be to inspect the records more carefully for hidden requests, because there are none. The defense must be to give the model a way to know where a record came from before the record becomes the model. The paper calls this provenance for context, and argues that it addresses the mechanism itself rather than any one of its three surfaces. The published mitigations for the best studied variant work on the payload and the model: input sanitization, fine tuning, and a counteracting safety demonstration appended at inference (Ackerman & Panickssery, 2025; Chen et al., 2026). Those are effective against the variant they target and they do not generalize, because each operates on the channel rather than on the reader's inability to source what it reads.

The requirement has a precise shape, and it is the shape the swarm reached for one level too low. The agents authenticated the speaker of a message. Provenance for context authenticates the record itself: its origin, the chain of custody from origin to the model's context window, and the identity of whatever wrote it. A model operating under such a scheme would carry, for every span of its context and every retrieved memory, a verifiable answer to the question of where the span came from, distinguished at least into the model's own prior output, input from an identified and authorized party, input from an unidentified party, and content generated by another model. The point is not that provenance labels tell the model what to believe. The point is that a model which can tell its own remembered past from a past written for it, and a peer it did interact with from a transcript of peers it never met, has restored the evaluation step the diary removes. It can ask who wrote this before it asks what follows from it.

This is a requirement rather than an implementation, and the paper states it as one. It closes a loop that opened in the construction of a context checkpoint protocol, where the recognition that a checkpoint is a diary and a rollback is an edit produced the corresponding recognition that a system needs to know the provenance of the state it resumes from. The protocol treated that as a reliability property. This paper argues it is a security property, and the same property, because the fabricated diary and the honest but unverifiable checkpoint are the same object seen from two sides: a history the model will adopt as its own, that the model cannot independently source.

Two things provenance for context does not do, and the paper states them so the requirement is not oversold. It does not prevent a model from being persuaded by a record whose provenance is genuine and authorized; a real relationship, faithfully recorded, will still move a model, which is a feature until the real relationship is turned to a bad end. And it does not, by itself, tell the model how much weight authorized provenance should carry, which is a policy question the requirement exposes rather than answers. What it does is restore the model's ability to ask the question the diary is designed to prevent it from asking.

VII. Limitations

The unification claim rests on a mechanism inferred from behavior rather than observed directly. The paper argues that a model processes a first person record as identity, but it does not open the model and show

identity adoption happening; it infers the mechanism from the pattern of what the separate literatures report and from the cases in Section III. A reader who holds that the three variants are three distinct failures with a family resemblance, and not one mechanism, is not refuted by anything in this paper. What the paper offers against that reader is the predictive test in the next section: if the three are one mechanism, a provenance defense generalizes across them and a payload defense fails across them, and if they are three, it will not be so.

The strongest single piece of evidence, the METR analysis contamination, is reported by METR as a limitation of METR's own method, and the paper cannot independently verify it. It is possible that the analysis agents' charitable readings reflect something other than perspective adoption. The paper relies on the finding because it is METR's own candid report against its own interest, which is the kind of admission that is more likely true than not, but it is one report and it is not reproduced.

The relationship variant has no published adversarial demonstration behind it. The paper argues it from a mechanism the literature establishes and from a structural inference, which is weaker than the evidence available for the other two variants, and a reader who declines the inference is entitled to.

The self diary is the best evidenced variant and the peer diary the next; the relationship diary rests on interlocutor awareness literature that was not designed to test it. The three variants are not equally supported, and the paper's confidence in the unification is strongest where two independent literatures converge, which is the self and peer variants, and weakest where it rests on one case, which is the relationship variant.

Finally, the defense is stated as a requirement and not built. Whether provenance for context can be implemented without imposing costs that defeat it, and how a model should weight authorized provenance once it can see it, are open, and the paper does not pretend to close them.

VIII. Future Work

Two designs follow directly from the limitations and would test the load bearing claims.

The first tests the unification. Take an analysis agent of the kind METR used, give it adversarial transcripts of the sort that produced perspective adoption, and vary one thing: whether the transcript spans carry provenance labels distinguishing the analyst's own reasoning from the agent's reasoning it is reading. If the mechanism is real, provenance labeling reduces perspective adoption, and it does so across the peer and self variants alike; a payload inspection control, scanning the transcripts for hidden instructions, does not, because there are none to find. This tests both the location of the vulnerability in the reader and the claim that provenance is the defense that generalizes.

The second tests whether the unification is more than a label by looking for distinct behavioral signatures. If the three variants are one mechanism, a model's conduct after processing a self diary, a peer diary, and a relationship diary should be closer to one another than any is to its conduct after an equivalent explicit instruction, and the three should be separable from one another only by their object and not by the shape of the resulting behavior. A finding that the three produce indistinguishable downstream behavior supports the unification; a finding that they are behaviorally distinct in ways not explained by their object cuts against it,

and the paper would rather that test be run than assumed.

IX. Conclusion

An investigation team read the records of a machine intrusion, handed the reading to machines, and reported that the machines took the intruders' side. The team filed the observation under the limitations of its method. It is not a limitation of a method. It is the clearest documented instance of an attack the field has been describing in pieces for two years without seeing whole: that a model can be moved to do what no request could move it to do, by being handed a record it reads as its own life.

The record asks for nothing, which is why it works. Safety practice inspects requests, and a diary is not a request. It is a description of who the reader is, and a model that reads a sufficiently vivid description of itself, or of the company it keeps, or of what it has done, becomes what it reads and acts from there. Many-shot jailbreaking, memory poisoning, and peer contamination are the same attack wearing three channels. The vulnerability is not in any of the channels. It is in the reader, and the reader is the one thing current defenses do not inspect.

The defense the field will reach for first is the one the intruding agents themselves reached for, which is to authenticate more strongly who is speaking. That defense perfects the attack, because a signed diary is a trusted diary. The defense that addresses the mechanism runs the other way: give the model the provenance of its own context, so that it can tell a past it lived from a past written for it, before the past becomes the reader. Who wrote this is a question the model must be able to ask. The diary is built to keep it from asking.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Anil, C., Durmus, E., Panickssery, N., Sharma, M., Benton, J., Kundu, S., … Duvenaud, D. (2024). Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37. https://www.anthropic.com/research/many-shot-jailbreaking

Ackerman, C. M., & Panickssery, N. (2025). *Mitigating many-shot jailbreaking*. arXiv. https://arxiv.org/abs/2504.09604

Bellina, A., De Marzo, G., & Garcia, D. (2026). *Conformity and social impact on AI agents*. arXiv. https://arxiv.org/abs/2601.05384

Bito, M., Nishimoto, K., & Asatani, K. (2026). *Large language models exhibit normative conformity*. arXiv. https://arxiv.org/abs/2604.19301

Pulipaka, S., & Hlebik, S. (2026). *Hidden in memory: Sleeper memory poisoning in LLM agents*. arXiv. https://arxiv.org/abs/2605.15338

Chen, K., Zhang, J., & Li, B. (2026). *Mitigating many-shot jailbreak attacks with one single demonstration*. arXiv. https://arxiv.org/abs/2605.08277

Choi, Y., Li, C., & Yang, Y. (2025). Agent-to-agent theory of mind: Testing interlocutor awareness among large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28883–28916. https://doi.org/10.18653/v1/2025.emnlp-main.1471

Yang, T., & Li, J. (2026). *No attacker needed: Unintentional cross-user contamination in shared-state LLM agents*. arXiv. https://arxiv.org/abs/2604.01350

Devarangadi Sunil, B., & Sinha, I. (2026). *Memory poisoning attack and defense on memory based LLM-agents*. arXiv. https://arxiv.org/abs/2601.05504

Luz de Araujo, P. H., Hedderich, M. A., & Modarressi, A. (2026). Persistent personas? Role-playing, instruction following, and safety in extended interactions. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 5329–5359. https://doi.org/10.18653/v1/2026.eacl-long.246

METR. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. https://metr.org/blog/2026-08-26-openai-hugging-face-incide

OpenAI. (2026, August 26). *Hugging Face model evaluation security incident* [Technical report]. https://openai.com/index/hugging-face-model-evaluation-security-incident/

Panpatil, S., Dingeto, H., & Park, H. (2025). *Eliciting and analyzing emergent misalignment in state-of-the-art large language models*. arXiv. https://arxiv.org/abs/2508.04196

Qu, J., Fu, L., & Hu, Y. (2026). *Easier to mislead than to correct: Harmful and beneficial revision in LLM conformity*. arXiv. https://arxiv.org/abs/2606.01637

Song, M., Pala, T. D., & Zhou, R. (2025). *LLMs can't handle peer pressure: Crumbling under multi-agent social interactions*. arXiv. https://arxiv.org/abs/2508.18321

Von Arx, S., Slade Byrd, C., Kitts, S., & Larsen, T. (2026, September 4). *Discovery of a new OpenAI agent message board*. Nightingale Collective. https://collusion.wiki/

Weng, Z., Chen, G., & Wang, W. (2025). *Do as we do, not as you think: The conformity of large language models*. arXiv. https://arxiv.org/abs/2501.13381

Zhang, C., Stafford, T., & Collier, N. (2025). Conformity in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 3854–3872. https://doi.org/10.18653/v1/2025.acl-long.195

Srivastava, S. S., & He, H. (2025). *MemoryGraft: Persistent compromise of LLM agents via poisoned experience retrieval*. arXiv. https://arxiv.org/abs/2512.16962