Quando a IA Lembra de Você Errado

Uma IA pode errar uma resposta. Uma memória persistente pode ficar reproduzindo o erro.

Rafael M. Ehlers · rafaelmehlers.com.br · Julho de 2026

> **Da Capacidade à Continuidade** — o primeiro ensaio de um programa de pesquisa em cinco partes. Quando uma IA transforma uma observação temporária numa crença duradoura, o erro pode moldar as interações que vêm depois. Uma introdução à alucinação biográfica.

A primeira alucinação

Quando as pessoas começaram a usar grandes modelos de linguagem, a falha mais visível era fácil de reconhecer.

O modelo inventava algo.

Ele inventava uma citação, confundia eventos históricos, atribuía uma frase à pessoa errada ou descrevia com confiança um lugar que não existia. Esses erros ficaram conhecidos como *alucinações*. Eram perceptíveis porque apareciam diretamente na resposta.

Num sistema sem memória além da sessão, o erro costumava ficar confinado àquela troca. O usuário podia contestá-lo, ignorá-lo ou começar de novo. A conversa seguinte não precisava herdar o engano.

À medida que esses erros foram sendo mais bem compreendidos, outro tipo de falha começou a chamar atenção: erros que não terminam com a resposta, porque o sistema os preserva como memória.

A segunda alucinação

Imagine dizer a uma IA:

> “Parei de tomar café esta semana para ver se durmo melhor.”

Um mês depois, ela lembra:

> “Você evita cafeína.”

Nada de dramático aconteceu. O sistema não inventou o café, e a nova afirmação ainda soa plausível. Ele apenas transformou um experimento temporário numa preferência duradoura.

Essa pequena mudança pode viajar. Meses depois, o sistema pode filtrar recomendações, interpretar de forma diferente uma queixa de cansaço, ou justificar seu conselho referindo-se ao que acredita que você prefere. Cada nova interação pode fazer a interpretação original parecer mais estabelecida do que era.

O erro aconteceu uma única vez, mas a memória continua a reproduzi-lo, e isso é um tipo diferente de problema.

Uma resposta desaparece. Uma memória se acumula.

Num sistema limitado à sessão, uma alucinação costuma ser local à resposta.

Num sistema persistente, um erro pode virar infraestrutura.

Uma vez que uma memória entra na imagem que o agente tem de uma pessoa, respostas posteriores podem depender dela. O erro deixa de ser apenas algo que o sistema disse. Ele passa a fazer parte daquilo que o sistema usa para decidir o que dizer em seguida.

Uma resposta é um evento; uma memória é um estado.

Corrigir uma resposta costuma ser local. Corrigir um estado persistente pode exigir reparar tudo o que depende dele. Mesmo quando o usuário fornece o fato correto, o sistema pode precisar revisar resumos antigos, inferências posteriores e padrões construídos a partir do erro original.

A personalização torna os erros mais convincentes

Quanto mais um agente parece lembrar de nós, mais confiável ele pode parecer.

Essa impressão pode tornar erros sutis mais difíceis de perceber. Uma resposta personalizada chega com a autoridade da continuidade:

> “Você sempre preferiu…”

> “Isso vem acontecendo há meses…”

Frases assim fazem mais do que recuperar informação. Elas colocam o presente dentro de uma história sobre o passado. Quando essa história é precisa, a continuidade pode ser útil. Quando não é, a fluência pode fazer uma inferência frágil soar como história consolidada.

O perigo vai além de errar um fato: com a repetição, o erro adquire uma biografia.

Memória é interpretação

Memória é mais do que armazenamento.

Todo sistema de memória seleciona, comprime e interpreta. Ele precisa decidir o que é relevante, o que pertence a um momento específico, quão certa é uma afirmação, qual contexto lhe dá sentido e por quanto tempo ela deve permanecer ativa.

Essas decisões são inevitáveis. Uma transcrição completa não é o mesmo que uma memória útil, e um resumo conciso não consegue preservar cada ressalva. O problema começa quando a compressão esconde a diferença entre o que a pessoa disse e o que o sistema concluiu.

No momento em que uma IA transforma uma observação temporária numa característica permanente, ela começa a construir uma biografia em vez de apenas preservar evidências.

O exemplo do café é simples, mas o mesmo padrão pode tocar partes mais consequentes de uma vida: saúde, trabalho, relações, planos, medos ou crenças. Um detalhe não precisa ser fabricado para se tornar enganoso. Basta ser desligado de seu tempo, contexto ou incerteza.

Esquecer às vezes é mais seguro

Um assistente esquecido pergunta de novo; um assistente equivocado pode seguir em frente com confiança. Um expõe uma lacuna, enquanto o outro a esconde atrás da personalização. É por isso que lembrar da coisa errada muitas vezes é pior do que não lembrar de nada.

Não porque esquecer seja inofensivo ou sempre preferível. Pedir repetidamente a mesma informação pode ser frustrante, ineficiente ou insensível. Mas a incerteza deixa espaço para correção. Uma memória falsa pode fechar esse espaço antes que o usuário perceba que ele era necessário.

Uma nova responsabilidade

Uma vez que um agente carrega a história de alguém ao longo do tempo, a exatidão factual sozinha já não basta.

Ele precisa distinguir:

- fatos de inferências;
- informação atual de informação obsoleta;
- as palavras do usuário das conclusões do sistema;
- estados temporários de características duradouras;
- informação sobre o usuário de informação sobre terceiros;
- afirmações diretas de interpretações geradas pelo assistente.

A memória persistente cria novas obrigações. Um sistema precisa preservar não só o conteúdo, mas também a proveniência, o contexto, a incerteza e a mudança. Ele deve ser capaz de segurar

uma afirmação com leveza quando a evidência é temporária, e revisá-la quando as circunstâncias da pessoa seguem em frente.

Isso não é simplesmente uma exigência de mais memória. Mais memória pode preservar mais erros. O desafio mais profundo é continuidade sem distorção.

A próxima pergunta

A primeira geração de IA perguntou:

> As máquinas conseguem responder às nossas perguntas?

A próxima geração talvez pergunte:

> As máquinas conseguem lembrar de nós sem reescrever silenciosamente quem somos?

Talvez essa falha seja mais do que apenas outra variedade de alucinação, e mereça um nome próprio.

O primeiro paper deste programa propõe um.

Alucinação Biográfica.

Referências

1. Park, J. S., et al. (2023). "Generative Agents: Interactive Simulacra of Human Behavior." *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. doi.org/10.1145/3586183.3606763
2. Packer, C., et al. (2023). "MemGPT: Towards LLMs as Operating Systems." *arXiv:2310.08560*. arxiv.org/abs/2310.08560
3. Wu, D., et al. (2025). "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory." *International Conference on Learning Representations*. arxiv.org/abs/2410.10813
4. Chen, D., et al. (2025). "HaluMem: Evaluating Hallucinations in Memory Systems of Agents." *arXiv:2511.03506*. arxiv.org/abs/2511.03506
5. Uddin, M. N., et al. (2026). "From Recall to Forgetting: Benchmarking Long-Term Memory for Personalized Agents." *arXiv:2604.20006*. arxiv.org/abs/2604.20006