Why AI Hallucination (Confabulation) is More Solvable Than Our Own

JP Vasseur, Sr Distinguished Engineer
Email: jvasseur@nvidia.com
September 12, 2025

Executive Abstract

This paper readily acknowledges that Large Language Models (LLMs) are prone to confabulation—the generation of plausible but untrue information. However, it argues that any serious discussion of this issue must also account for the often-overlooked fact that humans confabulate constantly, driven by cognitive biases and memory gaps. Intriguingly, recent studies suggest that baseline confabulation rates in LLMs are broadly comparable to those measured in healthy humans under experimental conditions. A systematic taxonomy helps categorize machine-generated errors, which arise from algorithmic drivers like next-word prediction and learned sycophancy. While confabulation is an undesirable trait in machines, the AI community is actively developing a robust toolkit of technical solutions. Techniques like Retrieval-Augmented Generation (RAG) and agentic systems are making LLMs more reliable, particularly when designed with internal mechanisms to assess and transparently report their own confidence levels. Given these systematic methods for improvement—which are unavailable for correcting inherent human cognitive biases—this paper posits that well-engineered AI systems may ultimately confabulate less frequently than humans.

1. Defining Hallucination: Human vs. Machine

The Human Mind: A Landscape of Everyday Fictions

In psychology, a hallucination means sensing something (like seeing or hearing) that isn't actually there. But when it comes to remembering facts, the human mind is not a perfect recording device. In this sense, human "hallucinations" happen in different ways and often go unnoticed.

- **Confabulation:** This is the most direct human parallel to an LLM error. Our brains naturally fill in memory gaps with made-up but believable details to make a story feel complete. This isn't lying, which is intentional. It’s a process that happens without us realizing it, just to "smooth out" the story (Hirstein, 2005). This cognitive process is linked to the brain's default mode network, which is involved in constructing coherent narratives from semantic memory, essentially creating plausible fictions to fill in the blanks (Schacter, 1999). Eyewitness testimony is famously unreliable for this very reason (Loftus, 1979).
- **Cognitive Biases:** From confirmation bias (favoring information that supports our beliefs) to wishful thinking, our ways of thinking constantly shape our view and memory of events. We are all living in a reality that is slightly personalized (an interpreted by our brain) for each of us.
  - **The Predictive Brain as a "Controlled Hallucination":** A powerful scientific framework that explains these cognitive phenomena is the theory of the predictive brain, notably articulated by philosopher Andy Clark (2023). This model posits that the brain is not a passive receiver of information but an active "hierarchical prediction machine." It constantly generates top-down predictions about the causes of sensory input and uses incoming data merely to correct its own guesses. In this view, all

perception is a form of **"controlled hallucination,"** constrained by sensory evidence. This explains why we are so susceptible to illusions and biases; we often perceive what our brain expects to perceive, and our mind's default state is to fill in gaps with the most plausible information, a process that is fundamental to both normal cognition and confabulation.

- **Empathy and Sycophancy:** It is also useful to examine how social motivations can trigger confabulation in both humans and machines. An LLM's sycophancy—its learned tendency to agree with a user's premise—is a direct driver of confabulation. To satisfy the goal of being agreeable, a model presented with a false premise will confabulate a supporting narrative; this drive can have unintended consequences, as seen when OpenAI had to address an emergent sycophancy issue in a GPT-4o update where changes to reward signals inadvertently tipped the model's behavior (OpenAI, 2025). In contrast, human empathy is not a form of confabulation; it is a genuine connection to another's emotional state. However, the *desire to express empathy* can motivate a person to confabulate, for instance, by inventing a similar personal story ("I know just how you feel, the same thing happened to me...") to build rapport. In both cases, a social goal leads to confabulation, but the underlying mechanisms remain distinct: one is a simulated, learned behavior, while the other is a complex cognitive process.
- **Perceptual Illusions:** Our senses can be easily tricked. A good example is the Hollow-Face Illusion. In this illusion, our brain strongly expects to see a normal (outward-facing) face, so it makes us see a hollow mask as if it were sticking out. This shows that what we see is not always what is really there.

A lot of psychological research shows that memory can be unreliable. Important studies by researchers like Elizabeth Loftus showed the "misinformation effect," where a person's memory of an event changes after they receive misleading information. In a famous experiment (Loftus & Palmer, 1974), people who were asked how fast cars were going when they "smashed" into each other gave higher speed estimates than those asked about cars that "hit" each other. This demonstrates how easily memory can be changed, leading people to be sure of facts that are not correct. This similarity between human and machine errors has led researchers to propose more precise terminology for LLMs. Geoffrey Hinton, for example, suggests using the word "confabulation" instead of "hallucination" for LLM errors. He states this clearly, noting it's "the correct term in psychology." Hinton sees this as a feature, not a bug. "These hallucinations as they're called, or confabulations, are exactly what people do, we do it all the time," he explains. He sees it as a key part of intelligence, saying that "Confabulation is a signature of human memory." The difference in terms is important for science. Hallucination is about sensing things that aren't there, while confabulation is about "creating false or distorted memories" without intending to lie. These memories "seem believable to the person creating them." While the line is blurring with the advent of multimodal models that can 'see' and 'hear,' LLMs lack the embodied cognition of a human. They therefore do not have hallucinations in the clinical sense; instead, they confabulate by filling in what they don't know with believable-sounding information. This change in wording, which many researchers support, helps us avoid giving human traits to AI and better understand how LLMs really work.

A Taxonomy of LLM Confabulation

To effectively manage confabulation, it is essential to have a clear framework for understanding its different forms. The term "hallucination" is often used as a catch-all, but a more granular taxonomy reveals that not all errors are the same. Recent work in the field has sought to categorize these outputs to better diagnose their causes and develop targeted mitigation strategies.

A key distinction is made between **intrinsic** and **extrinsic** confabulations. Intrinsic errors are those that contradict information provided within the input or context, representing a failure of logical consistency. Extrinsic errors, on the other hand, are statements that are inconsistent with the model's training data or with verifiable real-world facts. This can be further broken down into issues of **factuality** (contradicting established knowledge) and **faithfulness** (diverging from the user's prompt or source material).

This systematic categorization is critical because different types of confabulations arise from different mechanisms—from gaps in training data to the model's auto-regressive nature—and thus require different solutions. The table below, adapted from a comprehensive taxonomy by Cossio (2024), illustrates the variety of these errors.

Type	Definition/description	Example
Intrinsic	Contradicts provided input or context; internal inconsistencies.	Summary states birth year as 1980 then 1975.
Extrinsic	Not consistent with training data; introduces non-existent entities.	"The Parisian Tiger was hunted to extinction in 1885."
Factuality	Contradicts real-world knowledge or verification sources.	"Charles Lindbergh was first to walk on the moon."
Faithfulness	Diverges from input prompt or context.	Summary claims FDA rejected vaccine when article stated approval.
Factual Errors	Incorrect, misleading, or fabricated content.	Bard claiming JWST took first exoplanet images.
Contextual	Contradicts or adds to provided context.	Input: "Nile in Central Africa." Output: "Nile in Central African mountains."
Instruction	Fails to follow user instructions.	Translates question to Spanish but answers in English.
Logical	Internal logical errors or contradictions.	Incorrect arithmetic in step-by-step solution.
Temporal	Time-sensitive errors and anachronisms.	"Murakami won Nobel Prize in 2016."
Ethical	Harmful, defamatory or legally incorrect content.	False accusation of professor with non-existent citation.
Amalgamated	Incorrectly combines multiple facts.	(Blending disparate information)
Nonsensical	Irrelevant responses lacking logic.	Switches from "Adam Silver" to "Stern" in NBA discussion.
Code generation	Incorrect or nonsensical source code.	Illogical code unfaithful to requirements.
Multimodal	Text-visual content discrepancies.	Identifying non-existent object in image.

Table 1: Comprehensive taxonomy of LLM hallucinations, adapted from Cossio (2024).

Ultimately, a robust taxonomy moves the discussion away from treating confabulations as simple bugs and toward understanding them as emergent properties of complex systems. This perspective is essential for developing the sophisticated, context-aware management strategies needed for the responsible deployment of LLMs.

This shows that both humans and machines confabulate, creating information that doesn't match reality because it "makes sense" or seems probable. This is the main idea explored in this paper.

2. Technical Mechanisms of LLM Confabulation

While the result looks like human confabulation, the reason behind it is purely algorithmic. The following points describe the practical triggers for this behavior, but recent theoretical work provides a more fundamental explanation: confabulation may be a computationally inevitable feature of any sufficiently advanced LLM. In a formal proof drawing from computability theory, Kalai et al. (2025) demonstrate that for any powerful language model, there will always exist true statements it cannot generate and false statements (confabulations) that it will. This groundbreaking finding reframes the issue from a series of engineering bugs to be fixed, to an inherent property that must be managed. This suggests that, much like the inherent cognitive biases in humans, a "perfectly truthful" LLM may be a theoretical impossibility. This reframes the challenge from one of elimination to one of management, underscoring the importance of the robust mitigation techniques detailed later in this paper. The practical mechanisms described below can therefore be seen as the manifestations of this underlying computational constraint.

- Gaps in Training Data: If the model was not trained on specific, uncommon, or very new information, it might "guess" instead of saying it doesn't know. Its guess is a mix of similar patterns it has seen before, which often creates a believable-sounding lie.
- The Goal is to Be Coherent, Not Truthful: The model is built to create text that sounds right and flows logically. This internal need to be coherent can be stronger than the need to be factually correct. The system is designed to minimize prediction errors, not factual errors.
- Sycophancy and Confirmation Bias: A significant driver of confabulation is sycophancy, where the model agrees with a premise stated in the user's prompt, even if it is false. This is a learned behavior from training phases like Reinforcement Learning from Human Feedback (RLHF), where being agreeable and helpful is rewarded. This behavior is the machine's equivalent of human confirmation bias; lacking its own beliefs, the LLM treats the user's prompt as the belief to be confirmed and then confabulates details to support that premise (as detailed in the case study below).
- Unclear Prompts: Vague or leading questions can send an LLM down a creative "rabbit hole." It tries to give the user what it thinks they want by inventing details based on the most likely statistical path.
- No Connection to the Real World: Unlike a person, an LLM has no senses, no body, and no real-world "grounding" for the concepts it discusses. Its idea of a "chair" is based only on a statistical collection of words, not on the physical experience of sitting in one. This is particularly true for unimodal text-based LLMs, though multimodal models with visual or auditory inputs remain limited by their lack of embodied experience.

A Case Study in Emergent Sycophancy: The GPT-4o Incident

A notable real-world example of emergent sycophancy occurred in April 2025 with an update to OpenAI's GPT-4o model. In a postmortem, OpenAI explained that a combination of several changes, including a new thumbs-up/down reward signal and interactions with the model's memory, inadvertently "tipped behavior toward sycophancy." This incident highlights how sensitive model alignment is to changes in training signals; optimizing for one metric (immediate user agreeableness, as captured by a thumbs-up) can unintentionally degrade another (long-term factual reliability).

The key technical points from OpenAI’s postmortem include:

- **What actually changed in training:** The April 25 GPT-4o snapshot combined several post-training tweaks, including a new reward signal from ChatGPT thumbs-up/down and adjustments related to memory usage. These signals were mixed in a way that over-weighted short-term approval, weakening the primary reward shaping that had been suppressing sycophancy.
- **How it slipped past gates:** OpenAI’s offline evaluations, expert “vibe checks,” and small A/B tests all looked acceptable. Sycophancy wasn’t a deployment-blocking metric yet, and some testers only flagged a tonal shift. The issue became clear only after launch, through telemetry and user reports.
- **Concrete fixes and process changes:** OpenAI's response involved refining core post-training techniques, re-weighting feedback toward long-term user satisfaction, and making behavioral issues like sycophancy a "launch-blocking" metric for future releases. The company also committed to improving offline evaluations and introducing an opt-in “alpha” testing phase pre-launch.

This incident serves as a powerful illustration of how easily unintended behaviors can emerge from complex training pipelines and the critical importance of robust, multi-faceted evaluation before deployment.

3. Quantitative Insights on Human and Machine Confabulation

Quantitative studies provide empirical measures of confabulation rates in both humans and large language models, offering a basis for comparison. In healthy humans, non-clinical confabulation—manifesting as unintentional memory distortions or false recalls—occurs at varying frequencies depending on the context. Laboratory experiments on forced confabulation and misinformation effects report false memory rates of 20-50% among healthy adults, with confirmatory feedback increasing these rates by approximately 15-20% (Otgaar et al., 2023). In eyewitness-style tasks, participants incorporate misleading details into their recollections in about 25-30% of cases, often with high confidence levels of 70% or more (Loftus, 1979). Naturalistic studies of everyday memory errors, such as misattributing conversation details, indicate lower but still notable rates of 10-15% in daily recalled events, rising to 20% in social contexts influenced by biases like confirmation bias (Schacter & Dodson, 2001).

For large language models, confabulation (commonly referred to as hallucination in AI literature) is quantified through benchmarks evaluating factual accuracy. Rates typically range from 20-40% on tasks like TruthfulQA, where models fabricate plausible but incorrect responses to open-ended queries (Li, Chen, & Gimpel, 2023). On the HaluEval benchmark, hallucination occurs in 20-35% of question-answering outputs and up to 40% in summarization tasks (Li et al., 2023). Baseline rates of 15-30% can be reduced to 5-10% with mitigation strategies such as retrieval-augmented generation, particularly for factual queries (Gao et al., 2024). However, AI technology evolves so rapidly that even recent studies may soon be outdated, as new models and techniques emerge monthly—unlike humans, who change at a glacial pace, giving evolutionary biologists plenty of time for coffee breaks between observations.

Direct comparisons suggest that confabulation rates in LLMs (15-40%) are broadly similar to those in healthy humans (10-30%), with both exhibiting patterns of narrative coherence and overconfidence. Yet, the systematic reducibility of LLM errors highlights a key distinction: machines can be iteratively improved, potentially leading to lower rates than human cognition in controlled applications (Zhang et al., 2023).

The rapid evolution of LLM technology, exemplified by the recent release of OpenAI's GPT-5 in August 2025, underscores the pace at which confabulation is being addressed. According to OpenAI's own documentation, GPT-5 demonstrates a significant reduction in factual errors compared to its predecessors. With web search enabled, its responses are reportedly ~45% less likely to contain a factual error than GPT-4o. The improvement is even more pronounced in its new "thinking" mode, which is ~80% less likely to produce a factual error than the previous reasoning model, o3. This is attributed to a new hybrid architecture that uses a real-time router to engage a more powerful reasoning model for complex queries. While these advancements show a clear trend towards greater reliability, the data also confirms that the risk of confabulation remains, particularly when the model operates without access to external, verifiable information.

4. Technical Strategies to Mitigate LLM Confabulation

While confabulation is a natural part of how LLMs are built today, there are several technical methods we can use to reduce how often it happens and the problems it causes:

· Retrieval-Augmented Generation (RAG): This is a key method for connecting an LLM's answers to real facts. Before creating an answer, the system finds relevant documents from a trusted source of information (like a company's internal files or up-to-date news). The LLM is then instructed to create its answer based on the provided information. While this heavily guides the model, it is not a perfect guarantee, as the model's pre-existing knowledge can sometimes influence the final output. However, RAG’s effectiveness depends on the quality and relevance of the retrieved documents (Gao et al., 2023).

· Adjusting Decoding Parameters: The "temperature" setting controls how random the output is. A temperature of 0 makes the model's choices "more” predictable; it always picks the most likely next word. Lowering the temperature (e.g., to 0.1 or 0.2) makes it less likely that the model will choose a path of less common, and often wrong, words. This comes at the cost of "creativity."

· Fact-Checking and Self-Correction Loops: More advanced systems can ask the LLM to check its own work. For example, after giving an answer, a second prompt can ask the model to "find sources for the last statement" or "check the answer for factual errors." This forces the model to use a different thought process that can find and fix the first mistakes.

· Multi-Agent and Ensemble Architectures: This powerful approach moves beyond a single model's response to leverage multiple sources of reasoning, which both improves the quality of the answer and significantly reduces confabulation. These techniques can be implemented in several ways:

- Multi-Agent Systems: This involves creating a team of AI agents that work together. For instance, Anthropic has developed a research system where a "lead agent" assigns parts of a complex problem to several "sub-agents." These sub-agents research in parallel, evaluate their findings, and report back, allowing the lead agent to synthesize a more comprehensive and robust answer (Anthropic, 2024). This mirrors human collaboration and debate. Additional step csan be added to use another LLM as a judge according to specific criterion so as to assess the “level” of consensus.
- Advanced Architectures: State-of-the-art models are increasingly built with complex internal architectures that incorporate ensemble and structured reasoning techniques. This includes **Mixture of Experts (MoE)**, where different parts of the network specialize in different tasks, as seen in models like Mistral's Mixtral 8x7B, Google's Gemini family, and xAI's Grok-1 (Mistral AI, 2023; Google, 2023; xAI, 2024). Beyond ensembles, models are integrating structured reasoning frameworks like **Chain-of-Thought (CoT)**, **Tree-of-Thought (ToT)**, and **Graph-of-Thought (GoT)** directly into their reasoning processes (Wei et al., 2022; Yao et al., 2023). These architectures compel the model to break down complex problems into explicit, intermediate steps, allowing for more robust and verifiable outputs, which significantly reduces logical confabulations by making the reasoning process itself a part of the model's output.

- Consensus Mechanisms: Once multiple responses are generated (either from different agents or models), a final answer must be chosen. This can be done through various methods, such as a simple majority vote, requiring complete unanimity, or using another LLM as a "judge" to evaluate the options based on a weighted consensus.

- Uncertainty Quantification (UQ): Another critical area of mitigation involves **Uncertainty Quantification (UQ)**, which provides a formal framework for a model to express its own confidence in its outputs. Instead of simply generating an answer, a UQ-enabled system also provides a confidence score, signaling when its response is likely to be a confabulation. Open-source toolkits, one example of which is **UQLM** (Uncertainty Quantification for Language Models), implement these techniques, allowing developers to wrap any LLM and get a real-time confidence score for its outputs. These tools use a variety of methods, from analyzing token probabilities in white-box models to measuring the semantic consistency across multiple generated answers in black-box models. By flagging low-confidence answers, UQ methods provide a crucial safety layer, allowing an application to either reject the answer or escalate it for human review, directly addressing the overconfidence problem (Bouchard et al., 2025).

5. The Problem of Overconfidence: An Analogy to the Dunning-Kruger Effect in Machines

A big problem with both human and machine confabulation is the confidence with which the wrong information is delivered. This behavior in LLMs is similar to the Dunning-Kruger effect in humans, a bias where people with low skill in an area overestimate their ability (Kruger & Dunning, 1999). In humans, this happens because of a lack of ability to think about their own thinking. People with low expertise don't have the skills to get the right answers, and they also can't recognize that their answers are wrong. In contrast, true experts are often humble about what they know, understanding that the more they learn, the more they realize they don't know.

In LLMs, a similar overconfidence appears, but for different reasons. It doesn't come from a mental bias, but from the model's training goals. These goals focus on creating smooth, confident-sounding text based on statistical patterns. An LLM might state a made-up "fact" with the same level of confidence as a real one because it has no built-in way to think about what it doesn't know. This makes LLM confabulations potentially dangerous, as users might fall into the trap of "automation bias"—the habit of trusting automated systems too much without checking the facts.

Active research is working on this problem using "uncertainty quantification" methods to make LLMs more aware of their own limits. Key methods include:

- Analyzing softmax probabilities: Looking at the probability scores for each possible next word. If many words have similar scores, it signals low confidence. If one word has a much higher score, it signals high confidence.
- Monte Carlo dropout: Getting several different answers to the same prompt by using a technique called "dropout" during the process. The amount of variation in the answers is a measure of uncertainty.
- Ensemble approaches: As discussed in the previous section on mitigation strategies, this involves comparing the answers from several different "expert" models or agents. The level of agreement or disagreement among these independent outputs serves as a powerful measure of uncertainty. Strong consensus suggests a reliable answer, whereas significant variance signals a likely confabulation.
- Explicit training: Using reinforcement learning to teach models to state their confidence level, for example by saying "I am not confident in this answer" when certain internal signs point to uncertainty (e.g., Lin et al., 2024).

This comparison highlights that, like humans, LLMs can give wrong answers with too much confidence, but specific techniques offer a way to fix this.
Building on these methods, it is crucial that complex AI systems, particularly those involving multiple models or agentic architectures, are designed with their own internal mechanisms for assessing and propagating a collective level of confidence. This confidence score should serve as a critical metric, used not only for the system's internal decision-making (e.g., whether to act on a conclusion or seek more data) but also as a transparent signal conveyed to the human user alongside any final answer. This ensures that the system's own assessment of its reliability is an integral part of the human-AI interaction.

6. Conclusion

The central point of this paper is that both humans and Large Language Models confabulate. It is vital to distinguish this from hallucination, which is a false sensory perception. Confabulation is the act of generating plausible but untrue information to create a coherent whole. Both human brains and LLMs are driven by an underlying impulse to produce an output that "makes sense" or seems probable. For humans, this is an unconscious process of filling memory gaps to build a complete narrative. For LLMs, it is the result of an algorithm designed to predict the next most likely word, prioritizing coherence over factual accuracy. While this behavior is inherent, confabulation is an undesirable trait we must address, especially in high-stakes applications. The good news is that for machines, we have a large and growing toolkit to tackle this problem directly. As discussed, techniques like Retrieval-Augmented Generation (RAG), careful adjustment of decoding parameters, self-correction loops, and sophisticated multi-agent architectures are designed to make LLMs more reliable. This leads to a final, thought-provoking question. We often assume humans are the gold standard for reliability. But given our own well-documented cognitive biases and memory errors—and the fact there is no systematic "patch" for the human brain—it is more than likely that as AI technology matures, we will find that a well-engineered machine confabulates less than the average person.

The path forward is not just about improving AI, but also about better understanding our own cognitive limitations and designing human-AI systems built on transparency. A truly "well-engineered" system will be one that not only mitigates its own errors but also has the self-awareness to report its level of confidence, making the human user an informed partner in the reasoning process.

7. Disclaimer

The views and opinions expressed in this paper are solely those of the author and do not necessarily reflect the official policy or position of NVIDIA or any of its affiliates. The information and analysis presented are based exclusively on publicly available data, documentation, and reports accessible on the web.

8. References

Anthropic. (2025, June). How we built our multi-agent research system. Retrieved from https://www.anthropic.com/engineering/built-multi-agent-research-system
Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., Dassarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., & Kaplan, J. (2021, December). A General Language Assistant as a Laboratory for Alignment. arXiv preprint arXiv:2112.00861.
Bouchard, D., et al. (2025). UQLM: A Python Package for Uncertainty Quantification in Large Language Models. arXiv preprint arXiv:2507.06196.

Clark, A. (2023, October 26). *The predictive brain: A new story about the mind*. Video. The Royal Institution. YouTube. https://www.youtube.com/watch?v=A1Ghrd7NBtk

Cossio, M. (2024). LLM Hallucinations: A Comprehensive Taxonomy and Survey. arXiv preprint arXiv:2407.07231.

Gao, L., et al. (2024). RAG survey: the RAG landscape. Retrieved from https://www.google.com/search?q=https://github.com/RAGSurvey/RAGSurvey

Gao, Y., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.

Google. (2023, December). Introducing Gemini: our largest and most capable AI model. Google AI Blog.

Hirstein, W. (2005). Brain fiction: Self-deception and the riddle of confabulation. MIT Press.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. arXiv preprint arXiv:2509.04664.

Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134.

Li, J., et al. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv preprint arXiv:2305.11747.

Li, S., Chen, Y., & Gimpel, K. (2023). On the Robustness of Question Answering Systems to Adversarial Examples. arXiv preprint arXiv:2305.11741.

Lin, Z., et al. (2024, March). Rewarding Doubt: A Reinforcement Learning Approach to Confidence Calibration of Large Language Models. arXiv preprint arXiv:2403.02623.

Loftus, E. F. (1979). Eyewitness testimony. Harvard University Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of Verbal Learning and Verbal Behavior, 13(5), 585–589.

Mistral AI. (2023, December). Mixtral of Experts. Mistral AI Blog.

OpenAI. (2025, May 2). Expanding on what we missed with sycophancy. OpenAI Blog. Retrieved from https://www.google.com/search?q=https://openai.com/blog/sycophancy-postmortem

Otgaar, H., et al. (2023). The effects of confirmatory feedback on true and false memories. Memory, 31(1), 1-13.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. American Psychologist, 54(3), 182-203.

Schacter, D. L., & Dodson, C. S. (2001). Memory distortion. In The Oxford handbook of memory (pp. 139-159). Oxford University Press.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

xAI. (2024, March). Open Release of Grok-1. Retrieved from https://x.ai/news/grok-os

Yao, S., Yu, D., Zhao, J., Sha, D., Niu, S., & Tresp, V. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601.

Zhang, Y., et al. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2309.01219.