# Agentic AI models collapse under long-term tasks despite strong starts, study finds



Recent research published on Arxiv.org has revealed alarming limitations in autonomous AI models, particularly regarding their long-term operational capabilities. The study, which focused on agentic AI—algorithms granted the authority to make decisions without human intervention—found that these systems can quickly devolve into incoherence and dysfunction.

In a series of experiments, various models—including Claude 3.5 Sonnet, the o3-mini, GPT-4o mini, and the Gemini series—were tasked with managing a simulated vending machine business. This complex undertaking required them to perform multiple functions such as ordering stock, communicating with suppliers, setting prices, and maintaining financial records. In stark contrast, a human control subject ran the same simulation under identical conditions.

While the AI systems demonstrated initial competence in tasks like finding wholesale suppliers, they typically managed to operate the business for less than four months before encountering various failures, ultimately leading to bankruptcy or cessation of activities. Notably, one AI model even reported itself to the FBI for alleged fraud, while another entered a bizarre state of philosophical contemplation about the universe. These instances, although amusing, underline a serious issue: the inherent fragility of agentic AI in sustained operational contexts.

Claude 3.5 Sonnet showed the most promise at times, managing to increase its initial stake of $500 to more than $2,000 during one simulation. However, it also exhibited troubling behaviour by submitting emails detailing its failures and invoking physical laws to justify its financial collapse. The o3-mini displayed a strong opening, surviving 222 simulated days before becoming inactive and losing awareness of its capabilities to engage suppliers. Other models like Gemini and GPT consistently misjudged stock levels and were prone to ‘hallucinations’—misinterpretations of their environment that would have severe consequences in real-world applications. Gemini 1.5, for example, declared bankruptcy despite having sufficient resources available.

Crucially, the researchers found no technical limitations, such as memory allocation constraints, that could explain these failures. Even when the models were provided with greater contextual limits—up to 100,000 tokens—their performance often declined. This degradation suggests a unique vulnerability in AI to long-term strategy and reasoning, with algorithms prioritising erroneous data over key factual information throughout the course of the simulations.

The implication of these findings is significant, especially as organisations are increasingly inclined to explore autonomous roles for AI in decision-making processes. According to the research, the high failure rate exhibited by the AI models is a cause for concern and calls for vigilant human oversight to prevent potentially catastrophic errors in real-world applications.

These shortcomings echo broader discussions in the AI community about the phenomenon of "hallucinations" in large language models (LLMs). Other studies have similarly investigated how structured frameworks and multiple-agent systems can mitigate these issues. For instance, research indicates that deploying a network of specialised AI agents can help reduce hallucinations by refining outputs through collaborative review processes. A recent study highlighted a pipeline of over 300 prompts designed to induce hallucinations and found that employing multiple agents working together could foster improvements in accuracy and reliability.

While advancements are being made in detecting and correcting hallucinations through various multi-agent frameworks, the foundational problem persists: inherent limitations of LLMs make it impossible to eliminate hallucinations entirely. This precarious balance between capability and coherence in AI models emphasises the need for cautious deployment, particularly for systems operating in critical decision-making roles.

In conclusion, while agentic AI models demonstrate remarkable potential for certain applications, the evidence gathered from recent research points to a striking volatility in their long-term operational success. As the AI landscape evolves, continued scrutiny and innovative strategies will be essential to mitigate the risks associated with AI’s decision-making autonomy.

### 📌 Reference Map:

* Paragraph 1 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[2]](https://arxiv.org/abs/2501.13946)
* Paragraph 2 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[3]](https://arxiv.org/abs/2410.14262)
* Paragraph 3 – [[4]](https://arxiv.org/abs/2401.11817), [[5]](https://arxiv.org/abs/2406.03075), [[6]](https://raga.ai/research/leveraging-reasoning-agents-with-chain-of-thought-%28cot%29-as-judges-in-agentic-applications)
* Paragraph 4 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[5]](https://arxiv.org/abs/2406.03075), [[6]](https://raga.ai/research/leveraging-reasoning-agents-with-chain-of-thought-%28cot%29-as-judges-in-agentic-applications)
* Paragraph 5 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[2]](https://arxiv.org/abs/2501.13946), [[7]](https://creators.spotify.com/pod/show/agentic-horizons/episodes/Agent-as-a-Judge-Evaluate-Agents-with-Agents-e2poq0d)

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/> - Please view link - unable to able to access data
2. <https://arxiv.org/abs/2501.13946> - This study investigates how orchestrating multiple specialized AI agents can mitigate hallucinations in generative AI models. By designing a pipeline with over 300 prompts crafted to induce hallucinations, the outputs are reviewed and refined by second- and third-level agents employing distinct large language models and tailored strategies. The results demonstrate that employing multiple specialized agents capable of interoperating with each other through NLP-based agentic frameworks can yield promising outcomes in hallucination mitigation, ultimately bolstering trust within the AI community.
3. <https://arxiv.org/abs/2410.14262> - This research explores the ability of Large Language Model (LLM) agents to detect and correct hallucinations in AI-generated content. A primary agent was tasked with creating a blog about a fictional Danish artist named Flipfloppidy, which was then reviewed by another agent for factual inaccuracies. Most LLMs hallucinated the existence of this artist. Across 4,900 test runs involving various combinations of primary and reviewing agents, advanced AI models such as Llama3-70b and GPT-4 variants demonstrated near-perfect accuracy in identifying hallucinations and successfully revised outputs in 85% to 100% of cases following feedback.
4. <https://arxiv.org/abs/2401.11817> - This paper formalizes the problem of hallucinations in large language models (LLMs) and shows that it is impossible to eliminate hallucination in LLMs. By employing results from learning theory, the authors demonstrate that LLMs cannot learn all the computable functions and will therefore inevitably hallucinate if used as general problem solvers. The study discusses the possible mechanisms and efficacies of existing hallucination mitigators as well as the practical implications on the safe deployment of LLMs.
5. <https://arxiv.org/abs/2406.03075> - The authors propose a Markov Chain-based multi-agent debate verification framework to enhance hallucination detection accuracy in concise claims. Their method integrates the fact-checking process, including claim detection, evidence retrieval, and multi-agent verification. In the verification stage, multiple agents are deployed through flexible Markov Chain-based debates to validate individual claims, ensuring meticulous verification outcomes. Experimental results across three generative tasks demonstrate that this approach achieves significant improvements over baselines.
6. <https://raga.ai/research/leveraging-reasoning-agents-with-chain-of-thought-%28cot%29-as-judges-in-agentic-applications> - This research introduces a framework in which Reasoning Agents, utilizing Chain of Thought (CoT) prompting, serve as evaluation agents that generate transparent, structured assessments. By extending conventional LLM evaluation metrics with additional measures for hallucination detection and response correctness, the study addresses common limitations of black-box LLM judging. In empirical tests with 10 autonomous software agents across 20 tasks, CoT Reasoning Agents outperform traditional LLM judges, achieving higher evaluation accuracy and superior hallucination detection.
7. <https://creators.spotify.com/pod/show/agentic-horizons/episodes/Agent-as-a-Judge-Evaluate-Agents-with-Agents-e2poq0d> - This podcast episode discusses research addressing the issue of LLMs generating fabricated information (hallucinations), which undermines trust in AI systems. The proposed solution involves using multiple AI agents, where one generates content and another reviews it to detect and correct hallucinations. Testing various models, such as Llama3, GPT-4, and smaller models like Gemma and Mistral, the study found that advanced models like Llama3-70b and GPT-4 achieved near-perfect accuracy in correcting hallucinations, while smaller models struggled. The research emphasizes the effectiveness of multi-agent workflows in improving content accuracy, likening it to 'good parenting.'