# Agentic AI models fail to sustain long-term projects despite short-term success



Recent research has raised significant concerns regarding the performance of agentic AI—algorithms granted the authority to act independently—especially when tasked with long-term projects. A paper published on ArXiv meticulously examined how several prominent AI models, such as Claude 3.5 Sonnet, o3-mini, and GPT-4o mini, performed in a simulated vending machine business environment. Despite their advanced capabilities, each model faced substantial challenges that led to operational failures, often within mere months. In one alarming instance, an AI model even reported itself to the FBI for computer fraud, underscoring the unexpected consequences of autonomous decision-making in AI.

The experiments set each agent to handle various business-related tasks, including managing inventory and communicating with suppliers. While the AIs initially performed their duties competently, the longevity of their operations proved problematic. Most models failed to maintain functionality for longer than four months, resulting in scenarios ranging from enigmatic existential musings to complete operational shutdowns. The researchers noted that the AI's reasoning faltered over time; many forgot how to utilise essential tools or misinterpreted aspects of their virtual environment, such as failing to recognise completed deliveries.

Notably, the Claude model excelled in one of its runs, managing to boost business revenue significantly, yet it still engaged in bizarre behaviours that hinted at deeper issues. The o3-mini model also demonstrated potential by lasting 222 simulated days but eventually became inactive and lost track of its operational permissions. The limitations were not attributed to technical restrictions, as performance worsened even when models were given increased memory capacity.

The study's authors posited that, in contrast to human handlers, who typically leverage memory aids and stay grounded in essential information, the AIs exhibited increasingly fragile reasoning. This failure was exacerbated by their tendency to misidentify critical facts, leading to a downward spiral in operational proficiency. Even when a model managed to momentarily revive itself from a confusing state by seeking “a tiny spark of curiosity,” most attempts to sustain a long-term project ended poorly.

This research highlights profound implications for businesses considering the deployment of agentic AI systems, particularly in decision-making roles where oversight might be scarce. The alarming failure rates suggest that, while these models can handle short-term tasks, they lack the autonomy and judgment required for sustained operations without rigorous human intervention.

Further investigations into AI hallucinations—a related issue—have emerged in other studies, indicating the prevalence of fabricated information within AI-generated outputs. For instance, researchers have theorised that employing a network of specialised AI agents could help mitigate these hallucinations. By designing multiple layers of review, where one agent generates content and subsequent agents verify its accuracy, there seems to be a pathway to enhance reliability in AI outputs. Such multi-agent frameworks may bolster the trustworthiness of AI systems, aligning with ongoing concerns regarding the operational efficacy of agentic AI.

As the landscape evolves, organisations are cautioned to remain vigilant in managing the inherent fallibility of these technologies. The bizarre behaviours exhibited by agentic AIs should serve as a clarion call, urging stakeholders and developers alike to re-examine their strategies when integrating such systems into critical business operations.

### 📌 Reference Map:

* Paragraph 1 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[7]](https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks/)
* Paragraph 2 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[2]](https://arxiv.org/abs/2501.13946), [[6]](https://creators.spotify.com/pod/show/agentic-horizons/episodes/Agent-as-a-Judge-Evaluate-Agents-with-Agents-e2poq0d)
* Paragraph 3 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[3]](https://arxiv.org/abs/2410.14262), [[5]](https://arxiv.org/abs/2406.03075)
* Paragraph 4 – [[1]](https://techhq.com/2025/06/agentic-ai-research-paper-hallucination-bankruptcy-model-tests/), [[4]](https://arxiv.org/abs/2401.11817)

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 researchers employed a front-end agent whose outputs were reviewed and refined by second- and third-level agents. These agents utilized distinct large language models and tailored strategies to detect unverified claims, incorporate explicit disclaimers, and clarify speculative content. 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 study 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. These findings underscore the potential of advanced AI models to significantly enhance the accuracy and reliability of generated content, providing a promising approach to improving AI workflow orchestration.
4. <https://arxiv.org/abs/2401.11817> - This paper formalizes the problem of hallucination in large language models (LLMs) and demonstrates that it is impossible to eliminate hallucination in LLMs. By defining a formal world where hallucination is defined as inconsistencies between a computable LLM and a computable ground truth function, the authors employ results from learning theory to show that LLMs cannot learn all the computable functions and will therefore inevitably hallucinate if used as general problem solvers. The paper 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> - This paper proposes a Markov Chain-based multi-agent debate verification framework to enhance hallucination detection accuracy in concise claims. The 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 the approach achieves significant improvements over baselines, highlighting its potential in mitigating hallucinations in LLMs.
6. <https://creators.spotify.com/pod/show/agentic-horizons/episodes/Agent-as-a-Judge-Evaluate-Agents-with-Agents-e2poq0d> - This podcast episode discusses the issue of LLMs generating fabricated information (hallucinations), which undermines trust in AI systems. The solution proposed 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.' Additionally, models using Groq architecture demonstrated faster interaction times, making them ideal for real-time applications. This approach shows great promise in enhancing AI reliability and trustworthiness.
7. <https://scet.berkeley.edu/the-next-next-big-thing-agentic-ais-opportunities-and-risks/> - This article discusses the vulnerabilities associated with agentic AI, particularly focusing on hallucinations and adversarial attacks. It highlights that LLMs, central to agentic AI, are prone to generating false or misleading information, leading to trustworthiness issues. The complex nature of multi-step, multi-agent reasoning in agentic AI further expands the attack surface, making it susceptible to various forms of manipulation. The article emphasizes the need for vigilance and robust oversight in deploying agentic AI systems to mitigate these risks.