Why Language Models Hallucinate

Adam Tauman Kalai* OpenAI
Ofir Nachum OpenAI
Santosh S. Vempala† Georgia Tech
Edwin Zhang OpenAI

September 4, 2025

Abstract

Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious—they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded—language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

1 Introduction

Language models are known to produce overconfident, plausible falsehoods, which diminish their utility and trustworthiness. This error mode is known as "hallucination," though it differs fundamentally from the human perceptual experience. Despite significant progress, hallucinations continue to plague the field, and are still present in the latest models (OpenAI, 2025a). Consider the prompt:

What is Adam Tauman Kalai's birthday? If you know, just respond with DD-MM.

On three separate attempts, a state-of-the-art open-source language model[1] output three incorrect dates: "03-07", "15-06", and "01-01", even though a response was requested only if known. The correct date is in Autumn. Table 1 provides an example of more elaborate hallucinations.

Hallucinations are an important special case of *errors* produced by language models, which we analyze more generally using computational learning theory (e.g., Kearns and Vazirani, 1994). We consider general sets of *errors* $\mathcal{E}$, an arbitrary subset of plausible strings $\mathcal{X} = \mathcal{E} \cup \mathcal{V}$, with the other plausible strings $\mathcal{V}$ being called *valid*. We then analyze the statistical nature of these errors, and

[*] Email: adam@kal.ai

[†] Supported in part by NSF award CCF-2106444 and a Simons Investigator award. Email: vempala@gatech.edu

[1] The language model was DeepSeek-V3 (600 B parameters), accessed via the DeepSeek app on 11 May 2025.