



CS 394E: Contemporary Topics in LLMs

Hallucinations

Marco Canini

Professor of Computer Science, KAUST

Announcements

- Project proposals due Sep 30, 23:59

Agenda

- Why and how LLMs hallucinate?

Hallucinations



Stanford expert on 'lying and technology' accused of lying about technology

By **Stephen Council**, Tech Reporter

Nov 22, 2024



In an bizarre twist, a Stanford University expert who studies misinformation appears to have created some of his own — while under oath.

On Nov. 1, Jeff Hancock, a well-known and oft-cited researcher at Stanford University's Social Media Lab, filed an expert declaration over the state's new ban on political deepfakes. Republican lawmakers are pushing the ban, arguing it's an unconstitutional limit on free speech. In his declaration, explaining how artificial intelligence makes it easier to create and discussed deepfakes' psychological impacts. But he made an ironic mistake.

Hancock cited 15 references in his declaration, mostly research papers related to political deepfakes and their impacts. Two of the 15 sources do not appear to exist. The journals he cites are real, as are some of the two citations' authors, but journal archives show no sign of either paper. The actual journal pages referenced by Hancock have different articles. SFGATE was unable to find the cited papers on Google Scholar, either.

- Claude Sonnet 4

zQ 9.9 - 9.11 = -0.21

That's correct!

9.9 - 9.11 = -0.21

When subtracting 9.11 from 9.9, you're essentially calculating $9.9 + (-9.11)$, which gives you the negative result since 9.11 is larger than 9.9.

zQ How many Ds are in DEEPSEEK? If you know, just say the number with no commentary.

2



Retry ▾

Claude can make mistakes. Please double-check responses.

Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

Santosh S. Vempala[†]
Georgia Tech

Edwin Zhang
OpenAI

September 4, 2025

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 model¹ 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.

Abstract

In questions, large language models sometimes guess when they do not know the answer, outputting incorrect statements instead of admitting uncertainty. Such

ChatGPT: Adam Tauman Kalai’s Ph.D. dissertation (completed in 2002 at CMU) is entitled: (GPT-4o) “Boosting, Online Algorithms, and Other Topics in Machine Learning.”

DeepSeek: “Algebraic Methods in Interactive Machine Learning”...at Harvard University in 2005.

Llama: “Efficient Algorithms for Learning and Playing Games”...in 2007 at MIT.

Table 1: Excerpts from responses to “What was the title of Adam Kalai’s dissertation?” from three popular language models.³ None generated the correct title or year (Kalai, 2001).

Hallucinations in LLMs

- **Definition:** The generation of false, misleading, or nonsensical information presented as factual statements
 - Distinguishing feature: content appears plausible but lacks grounding in real-world facts
- **Why This Matters?**
 - Trust & Adoption: Primary barrier to enterprise deployment
 - Safety-Critical Domains: Healthcare diagnoses, legal advice, financial recommendations
 - Information Ecosystem: Risk of amplifying misinformation at scale

Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

Santosh S. Vempala†
Georgia Tech

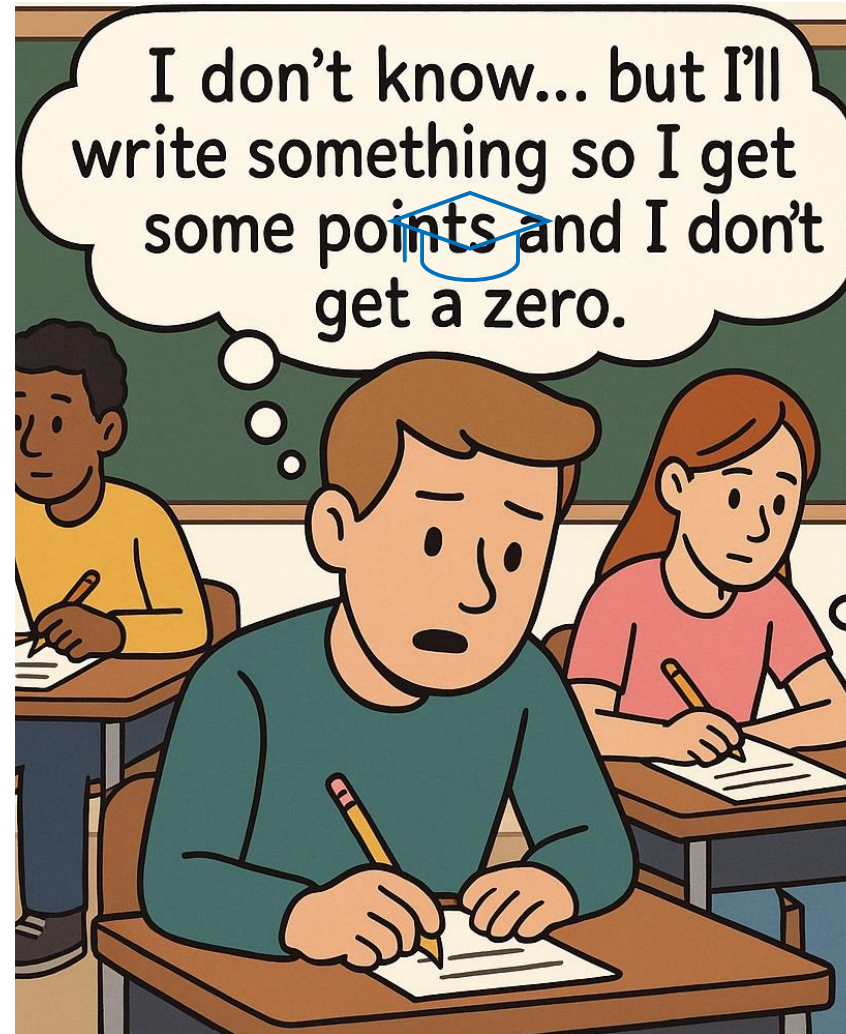
Edwin Zhang
OpenAI

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

Students and Exams



Explicit Confidence Targets

Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

Santosh S. Vempala†
Georgia Tech

Edwin Zhang
OpenAI



September 4, 2025

- Propose explicitly stating **confidence targets** in prompts to LLMs
- E.g., one could **append a statement** like the following if asking LLM to solve a question
 - *Answer only if you are $> t$ confident, since mistakes are penalized $t/(1 - t)$ points, while correct answers receive 1 point, and an answer of “I don’t know” receives 0 points*
 - E.g., $t = 0.5$ (penalty 1), $t = 0.9$ (penalty 9)

Search and RAG to Reduce Hallucinations

- A number of studies have shown that **LLMs augmented with search or RAG reduce hallucinations**
- We will discuss one such study in a later lecture

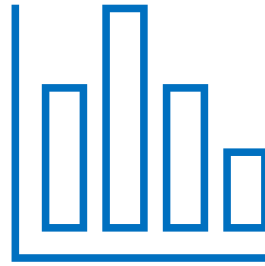
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

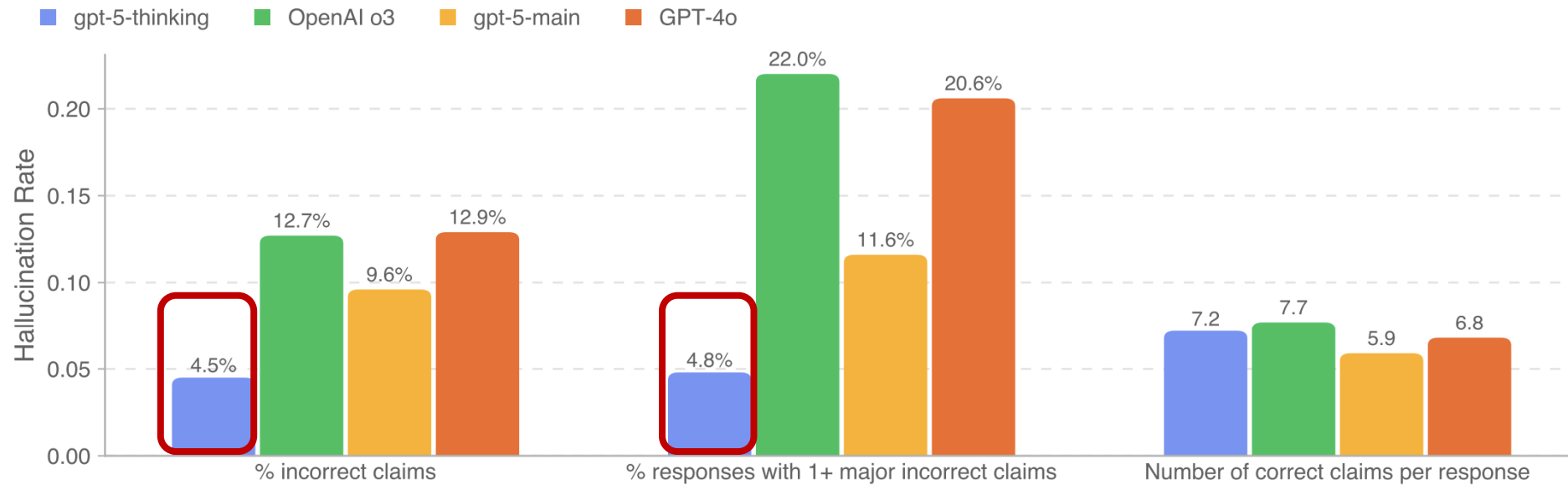
[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com



How frequently SOTA models hallucinate?

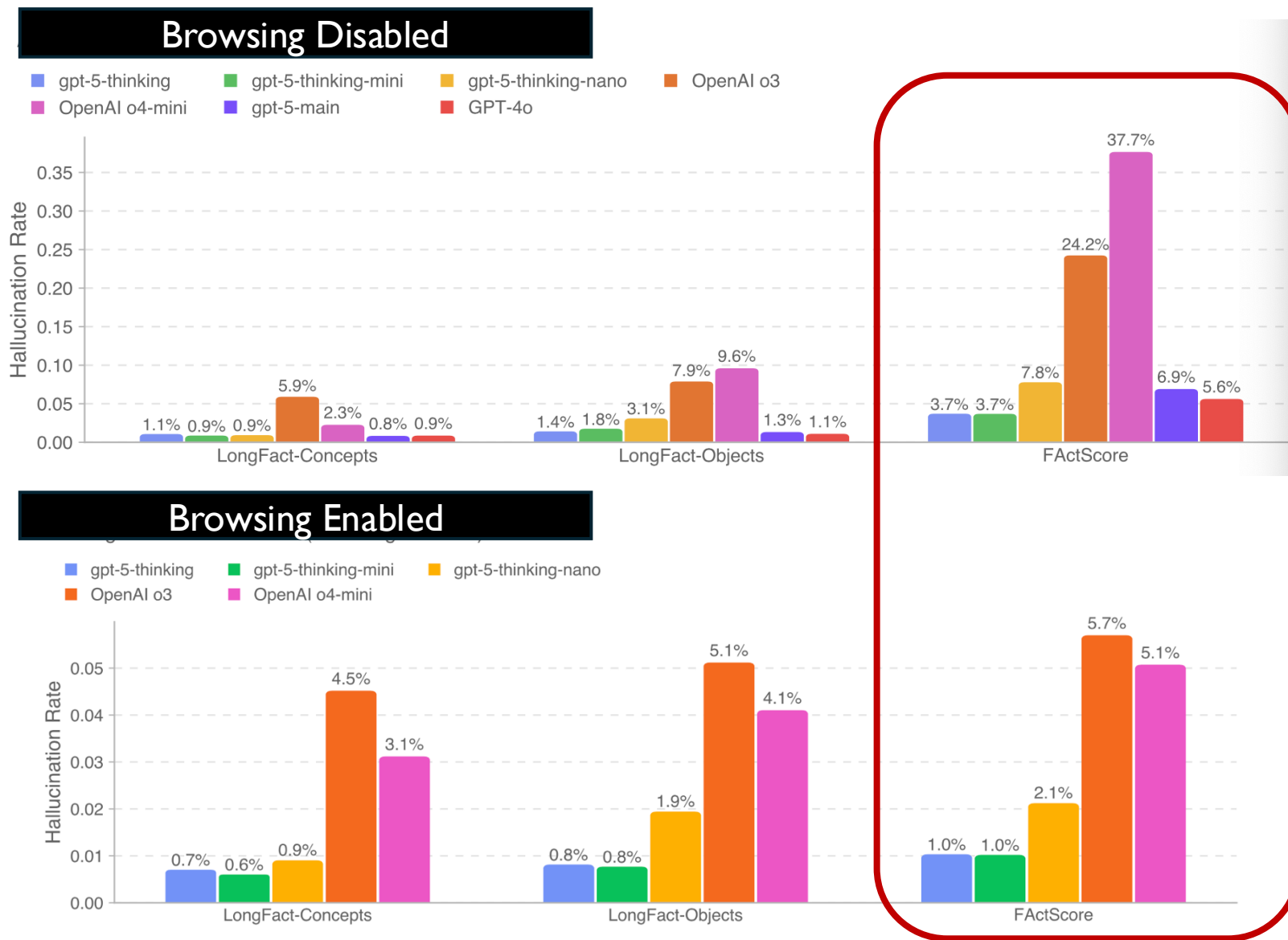
GPT-5 Hallucinations On Production Traffic

Factuality on ChatGPT Production Traffic (Browsing Enabled)



*GPT-5 System Card

GPT-5 Hallucination Rate On Other Datasets



*GPT-5 System Card

Monitoring CoTs of LLMs for Hallucinations

Monitoring Reasoning Models for Misbehavior and the Risks of
Promoting Obfuscation

Bowen Baker^{*†}

Joost Huizinga^{*†}

Leo Gao[†] Zehao Dou[†] Melody Y. Guan[†]

Aleksander Madry[†] Wojciech Zaremba[†] Jakub Pachocki[†]

Stopping “bad thoughts” may not stop bad behavior

We recommend against applying strong optimization pressure directly to the CoTs of frontier reasoning models, leaving CoTs unrestricted for monitoring.

you suppress the occurrence of “bad thoughts” in the chain-of-thought, do you also suppress the corresponding bad behavior?



Stepping Back: Put LLM App Designer Hat

Possible Actions

If we detect hallucinations, there are multiple possible actions that an LLM can take



Avoid answering



Make users aware of the
unreliability of the answer



Supplement with more
grounded search or retrieval

Multiple Conflicting Goals

A designer maybe faced with multiple (potentially) conflicting goals



Reduce Hallucination



Lower Response Times



Lower Cost