Why Language Models Hallucinate

Dongjae Lee

2025.09.11

Abstract

Hallucination is one of the most critical limitations of large language models. It undermines trust and slows real-world adoption. Can we eliminate hallucinations entirely? This article explains a root cause of hallucination and its implications for building more reliable AI systems.

Put simply, hallucination is a language model's behavior that creates plausible but unsupported statements. It has been a known weakness since the earliest language models and remains unsolved. Many approaches—such as Retrieval-Augmented Generation (RAG) and reinforcement learning (RL)—reduce the rate but do not eliminate it.

So, why do language models hallucinate? First, some questions cannot be answered with certainty from the information at hand. With limited context, there is no way to guarantee a correct answer. For example, guessing a randomly chosen person's birthday is essentially impossible.

Then, why not simply say "I don't know"? Under binary RL feedback signals, guessing often yields a higher expected reward than abstaining[1]. For instance, in exams, when we face a question we do not know, we usually guess. Because leaving it blank yields an $\mathrm{E}[Score]$ as zero, whereas a guess may yield $\mathrm{E}[Score]$ more than zero. Similarly, when models are trained with binary reward function, the learned policy favors plausible guesses over expressing uncertainty.

Therefore, models should say when they are unsure. A model that knows what it does not know can abstain, ask for more information, or retrieve evidence. RAG and RL with appropriate reward functions can implement these behaviors. Metacognitable models will show significantly reduced hallucination and lead to more reliable AI, and perhaps AGI.

[1] Adam et al., "Why language models hallucinate", 2025.09.05