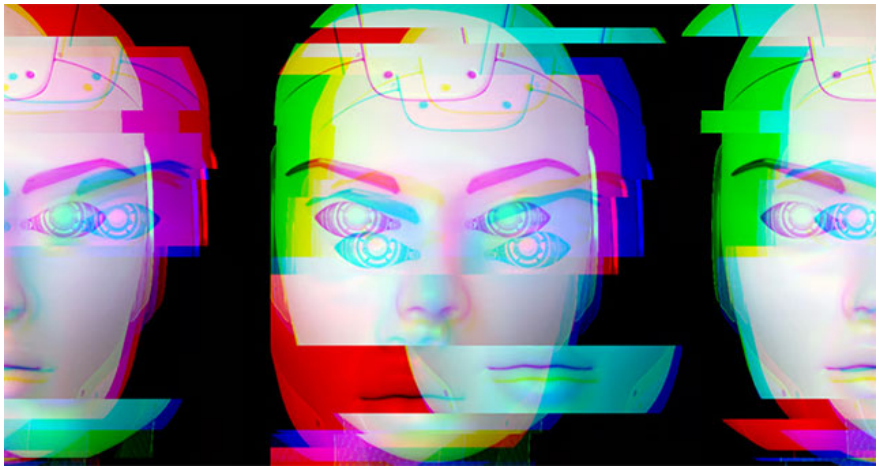
What is AI Hallucination? What Causes AI Hallucination?

OpenAI argues that AI illusions stem not just from the training process, but also from the way models are currently evaluated. Could changing the scoring mechanism help chatbots reduce 'guessing'?

A new study from OpenAI asks why large language models like GPT-5 or chatbots like ChatGPT still suffer from 'illusion', and is there any way to limit it?

In a blog post summarizing the findings, OpenAI defines *illusions* as 'apparently plausible but actually false answers generated by a language model.' The company acknowledges that, despite continued improvements in its systems, illusions remain a fundamental challenge for any major language model — and will never be completely eliminated.

To illustrate, the team said that when they asked a popular chatbot the title of Adam Tauman Kalai's (one of the study's authors) PhD thesis, they got three different answers — all wrong. When they asked him about his birthday, the chatbot gave three different dates, again wrong.



Why can chatbots be wrong… but speak so confidently?

Part of the problem, the researchers say, is the pretraining process, which focuses on predicting the next word in a sentence, rather than on distinguishing between correct and incorrect. In other words, the model only sees examples of fluent speech, and then tries to replicate the general distribution.

1. Recognizable rules like spelling or parenthetical usage are often learned very quickly, so errors gradually disappear as the model expands.
2. But for rare, irregular events (like a person's birthday), the model cannot infer accurately, and thus produces illusions.

It is worth noting that the paper does not focus on changing the initial training process, but rather points out that the current way of evaluating models creates biased incentives.

The authors compare model evaluation to taking a multiple-choice test. If you guess randomly, you might be right, but if you leave it blank, you will definitely get zero points. So when a model is scored solely on accuracy, it will 'rather guess than say it doesn't know'.

Proposed solution

Just as some exams (e.g. the SAT) penalize incorrect answers or give partial credit for missing an uncertain answer, OpenAI argues that AI assessments should also:

1. Punish errors of confidence more severely than those of uncertainty.
2. Award partial points when the model correctly expresses the level of doubt or says 'don't know'.

Adding a few more of these tests isn't enough, the team says. The entire popular grading system needs to be updated, because if rankings continue to rely solely on accuracy, models will continue to learn to guess at random to get high scores.