A castle on the sand

Dongjae Lee

2025.05.01

Abstract

This week, OpenAI rolled back GPT-4o due to its excessively flattering responses. Similarly, OpenAI's new frontier reasoning model, o3, has been reported to hallucinate more than previous models. Microsoft's Copilot has also recently begun to generate weird code. These unpredictable behaviors stem from the black box nature of language models. So far, there have been no major issues, but this is like building a castle on sand.

Neural networks are black boxes, and this characteristic is shared by language models, multimodal models, and even agent systems built on neural networks. A black box means that we fundamentally cannot know the reasoning process that leads to a particular output. This makes it difficult to predict the consequences of changes to the model.

Until recently, everything seemed to work well, but now problems are starting to surface. Excessive flattery, hallucinations, and abnormal behaviors are appearing even in top-tier AI services. Due to the infinite variety of inputs and the black box nature of neural networks, unexpected issues can arise no matter how much testing is done.

In particular, the o3 case shows that improving one capability can degrade another. Although o3 achieves top performance in math and coding, it hallucinates even more on general questions. This may be because it applies complex reasoning even when it is unnecessary, leading to strange answers. Such side effects are similar to how fixing one part of a very complex system can unintentionally break another.

As the era of agents is just beginning, research and discussion are needed to build safe and reliable AI systems. With the introduction of agents, their impact on the real world may become so great that follow-up actions are no longer sufficient. It is essential to have systems that verify AI in advance and monitor it during operation.