I became aware of how brittle our faith in large language models (LLMs) can be after reading about AI hallucinations on OpenAI’s blog and in recent research papers.

When artificial intelligence (AI) systems produce erroneous or fake information that sounds accurate, hallucinations result. The OpenAI blog likened it to a student who answers with assurance even when unsure, not out of malice, but because it feels worse to remain silent than to be incorrect.

I discovered from the surveys conducted by Huang et al. (2023) and Tonmoy et al. (2024) that model behavior and data limitations both contribute to hallucinations. Instead of predicting actual text, LLMs are designed to predict likely text. This small distinction results in answers that are fluent but inaccurate. I found it fascinating how contemporary mitigation strategies, such as self-verification and retrieval-augmented generation (RAG), attempt to “anchor” the model in actual data, essentially endowing it with a conscience.

This has a lot to do with my research interest in developing trustworthy AI for practical applications. An agricultural or medical assistant, for instance, cannot afford to “make things up” regarding a product or illness. Designing systems that verify before responding requires an understanding of hallucinations.

This idea strongly connects to my ongoing IPD project, where my team is developing a machine learning model to predict soil nutrient deficiencies and detect pests using leaf images. The goal is to guide farmers on the precise amount of fertilizer required when low nutrient levels are detected. In such a context, even a minor hallucination, like an incorrect diagnosis or fertilizer recommendation, could be fatal to a farmer’s crops. Therefore, ensuring that our model minimizes hallucinations is not just a research goal but a matter of real-world impact and agricultural safety.

The confidence–accuracy gap most surprised me: AI can appear to be absolutely certain while actually being entirely incorrect. That makes us reevaluate “intelligence”; perhaps humility with verification, rather than confidence, is what true intelligence is. Even though they can be funny at times, hallucinations serve as a reminder that AI still requires human oversight, data, and truth.