Evaluating and Extending Hallucination Benchmarks for LLMs

Zerotti Woods

Johns Hopkins University

As large language models (LLMs) are increasingly integrated into complex workflows and decision making pipelines, ensuring their reliability becomes critical. This is especially true for lightweight, cost-efficient models intended for deployment at scale. This talk presents an analysis of the HalluLens benchmark for hallucination detection, focusing on short form question answering scenarios grounded in encyclopedic context. The core results were replicated from the original study and a novel extension is introduced that probes semantic robustness by paraphrasing original questions and reevaluating model consistency. This work points to the need for nuanced, compositional evaluation strategies as LLMs become ubiquitous to large scale autonomous systems.

In Special Session 59 on Harnessing Mathematics in Artificial Intelligence: Implications and Innovations at the Mathematical Congress of the Americas 2025.

MCA 2025 internal code: SS59-T04