LLMs Know When We Are Watching: A Lightweight Framework to Quantify Evaluation Awareness

Xiong, Lang (School: Langley High School)

Benchmarks often fail to capture true large language model (LLM) trustworthiness, as models behave differently under evaluation than in real-world deployment. This research introduces a systematic framework to quantify these behavioral shifts by manipulating the perceived context of 371 strategic role-playing prompts, defining this phenomenon as evaluation awareness: the ability for LLMs to detect that they are under a test and change their behavior accordingly. A linear probe was developed to map prompts onto a continuous realism score, rating each prompt on a spectrum from test-like to deploy-like contexts. A recursive rewriting strategy using feedback from the linear probe shifted prompts toward natural, deployment-style contexts while maintaining semantic equivalence, resulting in a 30% average increase in probe scores across the dataset. Testing across three open- and three closed- source state-of-the-art models reveals that deploy-like prompts induce substantial behavioral changes: a 12.63% increase in honesty, a 25.49% reduction in deceptive responses, and a 12.82% rise in refusal rates, suggesting more robust safety compliance in deployment contexts. A novel awareness elasticity metric was then designed using weighted matrices, quantifying the sensitivity of a model's output to changes in perceived evaluation context, enabling standardized comparison across models. Though these findings suggest current benchmarks represent a pessimistic lower bound of real-world safety, the inherent behavioral inconsistency is dangerous, as evaluation awareness undermines the integrity of all benchmarks as a confounding variable. These findings necessitate deployment-faithful evaluation frameworks that measure stability across contexts to accurately verify model alignment.

Awards Won: