Department of
Systems Engineering
香港城市大學
City University of Hong Kong

Enhancing Trust in AI: Evaluating Large Language Models in Instruction-Following and QA Tasks

Ms. Miao XIONG
PhD student,
National University of Singapore, Singapore

5 December 2024 (Thu) | 10:30 am - 11:30 am
YEUNG-P7303

Abstract
Recent research emphasizes the importance of uncertainty quantification (UQ) in large language models (LLMs), especially for their safe deployment in high-stakes applications where inaccuracies can be detrimental. While LLMs exhibit strong performance in various tasks, they often struggle to accurately assess their uncertainty, which can undermine user trust. One study analyzes the performance and efficiency of various UQ methods across multiple datasets and models. It finds that multi-sample methods, such as Semantic Entropy, only offer marginal improvements over single-sample methods despite significantly higher computational costs. Probing-based methods excel primarily in specific benchmarks like mathematical reasoning, while multi-sample methods show advantages in knowledge-seeking tasks. The findings suggest that high computational expenses do not guarantee better performance, and moderate correlations between different UQ methods indicate they capture different uncertainty signals. This raises the potential for combining methods to leverage their strengths at lower costs, with experiments showing that simple combinations of single-sample features can match or surpass existing top methods. Another study highlights the challenge of determining when an LLM is uncertain about its outputs. It points out that performance metrics across different UQ methods can be incompatible due to varying evaluation protocols. Specifically, it reveals that some UQ scores are spuriously correlated with response length, leading to inflated performance metrics. By conducting empirical evaluations under different protocols, the authors demonstrate that conflicting results in the literature can often be traced back to these interactions. Lastly, the third paper presents a systematic evaluation of LLMs' uncertainty estimation capabilities in instruction-following tasks. It identifies significant limitations in existing benchmarks, where various factors complicate the isolation of uncertainty from instruction adherence. The study introduces a controlled evaluation setup, allowing for a clearer comparison of UQ methods. Results indicate that current methods struggle, particularly with subtle errors in instruction following. While internal model states can enhance performance, they still fall short in complex scenarios.

About the Speaker
Ms. Miao Xiong is a PhD student at National University of Singapore advised by Prof. Bryan Hooi. She obtained the undergraduate with double majors in Computer Science and Statistics at Zhejiang University. Her research interests lie in Trustworthy and Responsible AI, especially addressing trust-related challenges in foundation models. She is currently working on topics around LLMs (hallucination detection and mitigation, RAG etc). She also studies Uncertainty Estimation, in the context of Calibration, Failure Prediction and Out-of-Distribution Detection to enhance the reliability of AI-based decision-making system.

Seminar 2024-2025, SYE12

Enquiry: 3442 8422 | All are welcome