LLM Microscope: What Model Internals Reveal About Answer Correctness and Context Utilization

Jiarui Liu*, **Jivitesh Jain***, **Mona Diab**, **Nishant Subramani**
Carnege Mellon University
{jiarui15, jivitesj, mdiab, nishant2}@andrew.cmu.edu

Abstract

Although large language models (LLMs) have tremendous utility, trustworthiness is still a chief concern: models often generate incorrect information with high confidence. While contextual information can help guide generation, identifying when a query would benefit from retrieved context and assessing the effectiveness of that context remains challenging. In this work, we operationalize interpretability methods to ascertain whether we can predict the correctness of model outputs from the model's activations alone. We also explore whether model internals contain signals about the efficacy of external context. We consider correct, incorrect, and irrelevant context and introduce metrics to distinguish amongst them. Experiments on six different models reveal that a simple classifier trained on intermediate layer activations of the first output token can predict output correctness with nearly 80% accuracy, enabling early auditing. Our model-internals-based metric significantly outperforms prompting baselines at distinguishing between correct and incorrect context, guarding against inaccuracies introduced by polluted context. These findings offer a lens to better understand the underlying decision-making processes of LLMs.