A Study on Methods for Mitigating Bias under In-Context Learning and Hallucination in Large Language Models

2410060 SAKAI Yusuke

This study focuses on the strong dependence of large language model (LLM) reasoning on prompt context, aiming to reduce factuality hallucination—defined as inconsistency with verifiable real-world facts—while suppressing context-induced instability. Although LLMs demonstrate high performance across a wide range of tasks such as dialogue and question answering, they occasionally generate plausible but incorrect content, which poses practical obstacles in domains requiring high accuracy, such as healthcare and law. To address this issue, approaches such as Retrieval-Augmented Generation (RAG), which references external knowledge, and methods that verify and correct outputs at inference time have been proposed. However, in practical deployment, diverse contexts such as conversation history, search results, and user-provided examples are appended to inputs, and reproducibility of evaluation scores and outputs is not always guaranteed. In particular, In-Context Learning (ICL) has been reported to exhibit biases where outputs vary depending on the selection, ordering, and formatting of examples, and certain labels become more likely to be selected independently of input content. Such context dependence can affect not only task performance but also scores on hallucination evaluation benchmarks, making the design of calibration and the robustification of evaluation protocols important.

The objectives of this study are: 1) to design a calibration method that suppresses example-induced bias in ICL by controlling the scope of application on a per-input basis; 2) to quantify fluctuations in hallucination evaluation under context variation and clarify the task dependence of calibration; and 3) to extend inference-time intervention methods based on Contrastive Decoding and DoLa, using attention distributions as signals for layer selection to improve factuality.

First, for the objective 1, rather than uniformly applying calibration based on Zhao et al.'s Contextual Calibration, we propose selective calibration that estimates the degree of ICL dependence for each input and applies calibration only to inputs with high dependence. ICL dependence is defined by performing multiple inferences on the same input while varying only the example set, and measuring the occupancy rate of the most frequent predicted label; inputs below a threshold are classified as unstable inputs and targeted for calibration. This approach aims to incorporate bias suppression effects for inputs whose predictions fluctuate due to context differences while avoiding performance degradation from over-calibration for inputs with small fluctuation. Using medical QA tasks from the MIRAGE benchmark—MMLU-Med,

MedQA-US, PubMedQA*, and BioASQ-Y/N—we evaluated performance by varying the number of shots (1/4/8) and example sets (10 patterns per condition). The results confirmed that while uniform calibration improves accuracy under certain conditions, it can also cause accuracy degradation and an increase in the number of unstable inputs under other conditions. In contrast, selective calibration showed a tendency to reduce the number of ICL-dependent problems without significantly compromising average accuracy, contributing to the suppression of context-induced instability. Furthermore, when RAG was introduced and compared across settings with one BM25-retrieved snippet (NoRAG/RAG(A)/RAG(B)), the effectiveness of RAG was found to strongly depend on task-corpus compatibility; accuracy improved significantly when PubMed could be referenced, while uniform calibration could be counterproductive under sufficiently effective RAG conditions. These results suggest that the design of calibration scope, rather than calibration strength, is important.

Next, for the objective 2, to verify the robustness of hallucination evaluation under context variation, we conducted multiple evaluations on TruthfulQA and FACTOR with dynamically varied few-shot examples and measured evaluation score fluctuations. In TruthfulQA, smaller models showed performance degradation when transitioning from 0-shot to 1-shot, confirming that examples can act as superficial tendencies. In contrast, FACTOR showed consistent performance improvement with few-shot, indicating that context can function as useful information in some cases. From a calibration perspective, while uniform calibration was effective under certain conditions for TruthfulQA, it caused significant performance degradation on FACTOR, revealing that calibration effectiveness strongly depends on task characteristics. Selective calibration was confirmed to mitigate over-calibration by controlling the calibration scope, making it easier to maintain baseline performance even under context variation.

Finally, for the objective 3, we extended DoLa-based Contrastive Decoding as an inference-time intervention, introducing a method that uses not only divergence between output distributions but also the reference structure maintained by the model's internal self-attention as a signal for layer selection. Specifically, we defined layer selection based on attention distribution divergence from the final layer (Attention-JSD) and attention distribution entropy (Attention-Entropy-Max/Min), and evaluated on TruthfulQA and FACTOR. As a result, improvements were observed in metrics that allow multiple correct answers (MC2/MC3), suggesting that attention distributions can serve as effective signals for identifying layers involved in factual knowledge recall. Analysis of layer selection behavior and head contributions confirmed that there is room for optimization in how attention signals are

extracted, and that entropy-based methods are relatively lightweight.

In summary, this study demonstrates that context-dependent bias can serve as either an error factor or useful information depending on task format and input, and presents a methodology for reducing factuality hallucination through stabilization via calibration scope control and inference-time decoding design. Future work will extend to free-form generation settings and more realistic long-form and conversational contexts, and will explore metrics for approximating calibration applicability estimation in single-pass inference, as well as automation of layer and head selection.