Abstract

Large Language Models (LLMs) exhibit a persistent tendency to generate fluent yet factually incorrect outputs, commonly referred to as hallucinations. This failure mode remains prevalent despite advances in model scaling, instruction tuning, and alignment, posing significant risks in high-stakes applications. Recent work has shown that hallucinations can be mitigated at inference time by intervening directly in a model's internal representations, without retraining or external knowledge sources. In particular, representation engineering (RepE) demonstrates that behavioral properties such as truthfulness correspond to approximately linear directions in activation space, and that injecting these directions during inference can bias model outputs toward factual correctness.

However, existing inference-time steering methods apply interventions with a fixed strength across all tokens and decoding steps, implicitly assuming that hallucination risk is uniform throughout generation or decision-making. This assumption ignores substantial variation in model uncertainty across tokens and answer options. While token-level uncertainty estimation has been extensively studied as a diagnostic tool for calibration and hallucination detection, its use as a continuous control signal for inference-time intervention remains poorly understood.

This thesis investigates whether model-internal uncertainty can be used to adaptively modulate inference-time representation steering. We introduce *Uncertainty-Aware Scaling* (UAS), a framework that dynamically adjusts the magnitude of a fixed representation-level steering direction based on token-level uncertainty signals computed during inference. UAS is evaluated using multiple uncertainty estimators, including predictive entropy, Mahalanobis distance-based uncertainty, and recurrent attention-based uncertainty, and is applied selectively to the final transformer layers where interventions most strongly influence output decisions.

Experiments are conducted on TruthfulQA and CREAK, using both multiple-choice and generative evaluation protocols across several open-source LLMs. Results show that static representation steering consistently improves factual decision-making across models, confirming its robustness as an inference-time hallucination mitigation technique. In contrast, uncertainty-aware scaling exhibits mixed and strongly model-dependent behavior. For some configurations, uncertainty modulation provides modest gains, while in many cases it attenuates effective steering or degrades performance.

Detailed analysis reveals that these outcomes are driven by weak alignment between token-level uncertainty signals and incorrect content at the layers where steering is applied. As a result, uncertainty-aware modulation often operates in a noise-dominated regime unless the base steering signal is sufficiently strong. In free-form generation, uncertainty-aware scaling consistently degrades truthfulness and stability, indicating that token-level uncertainty is ill-suited as a control signal for generative decoding.

Overall, this work clarifies both the potential and the limitations of uncertainty-driven inference-time control. While uncertainty-aware modulation can refine representation steering under specific conditions, static representation engineering remains the most robust and broadly effective

inference-time intervention for hallucination mitigation in large language models.