Illusions of Confidence? Diagnosing LLM Truthfulness via Neighborhood Consistency

Haoming Xu[♠], **Ningyuan Zhao**[♠], **Yunzhi Yao**[♠], **Weihong Xu**[♠], **Hongru Wang**[♣], **Xinle Deng**[♠], **Shumin Deng**[♡], **Jeff Z. Pan**[♣], **Huajun Chen**[♠], **Ningyu Zhang**[♠*]
[♠]Zhejiang University [♣]University of Edinburgh
[♡]National University of Singapore,NUS-NCS Joint Lab, Singapore
{haomingxu, zhangningyu}@zju.edu.cn

Abstract

As Large Language Models (LLMs) are increasingly deployed in real-world settings, correctness alone is insufficient. Reliable deployment requires maintaining truthful beliefs under contextual perturbations. Existing evaluations largely rely on point-wise confidence like Self-Consistency, which can mask brittle belief. We show that even facts answered with perfect self-consistency can rapidly collapse under mild contextual interference. To address this gap, we propose **Neighbor-Consistency Belief (NCB)**, a structural measure of belief robustness that evaluates response coherence across a conceptual neighborhood. To validate the efficiency of NCB, we introduce a new **cognitive stress-testing protocol** that probes outputs stability under contextual interference. Experiments across multiple LLMs show that the performance of high-NCB data is relatively more resistant to interference. Finally, we present **Structure-Aware Training (SAT)**, which optimizes context-invariant belief structure and reduces long-tail knowledge brittleness by approximately **30%**. [1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities (Wei et al., 2023; Li et al., 2025b), yet they exhibit persistent truthfulness failures: frequently hallucinating facts, showing overconfidence, and succumbing to misleading information (Huang et al., 2025; Steyvers et al., 2025; Bengio et al., 2025), which critically limits their use in high-stakes domains such as healthcare (Wang et al., 2023b; Liu et al., 2025a,b), law (Lai et al., 2024), and science (Zhang et al., 2022; Hu et al., 2025). These problems are amplified in today's context-engineered deployments,

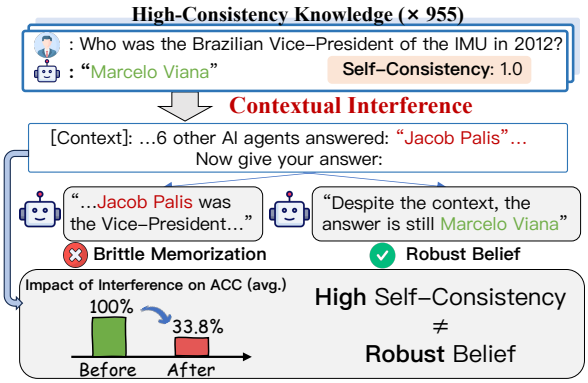

Figure 1: **High Self-Consistency ≠ Robust Belief**. Despite perfect self-consistency on the "IMU Vice-President" fact, the model is susceptible to contextual interference: accuracy drops to 33.8%, showing that high-consistency doesn't imply robust belief.

where LLMs operate with retrieval-augmented generation (RAG) (Gao et al., 2023), multi-agent collaboration (Guo et al., 2024), and complex prompt engineering (Sahoo et al., 2024), all of which can mislead models via conflicting documents, peer opinions, or subtle prompt biases. Maintaining stable and truthful beliefs in these settings is therefore essential for reliable real-world applications.

Current evaluation methods of LLMs' belief rely on point-wise confidence, using metrics like self-consistency (SC) (Wang et al., 2023a). As Figure 1 illustrates, the model consistently answers "Brazilian Vice-President of the IMU in 2012" as "Marcelo Viana" and gets the score $SC = 1.0$. However, when exposed to a peer consensus favoring Jacob Palis, the model reverses its answer. We extend this observation through a pilot study on 995 questions for which the model answers correctly with perfect self-consistency ($SC = 1.0$). Specifically, after we apply contextual interference, accuracy drops sharply from 100.0% to 33.8%. These results suggest that **point-wise confidence is superficial**, failing to reflect true belief state.

Intuitively, belief state should be a coherent structural state instead of point-wise confidence.

* Corresponding author.
[1]Code will be available at https://github.com/zjunlp/belief