SOVEREIGN: Does scaling the LLM size (e.g., 7B vs. 70B parameters) mitigate the accuracy loss from adversarial perturbati

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

This comprehensive review delves into the pivotal role of prompt engineering in unleashing the capabilities of Large Language Models (LLMs). The development of Artificial Intelligence (AI), from its inception in the 1950s to the emergence of advanced neural networks and deep learning architectures, has made a breakthrough in LLMs, with models such as GPT-4o and Claude-3, and in Vision-Language Models (VLMs), with models such as CLIP and ALIGN. Prompt engineering is the process of structuring inputs, which has emerged as a crucial technique to maximize the utility and accuracy of these models.

1 Introduction

Analysis of: Unleashing the potential of prompt engineering for large language models. Research goal: Does scaling the LLM size (e.g., 7B vs. 70B parameters) mitigate the accuracy loss from adversarial perturbations in multi-hop RAG, compared to single-hop RAG, on the FEVER and NQ subsets of BEIR?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 5 claims extracted, 3 verified. Tribunal: 6.8/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Large Language Models such as GPT-4o and Claude-3 represent a breakthrough in AI development	✓	0.17
Vision-Language Models such as CLIP and ALIGN represent significant AI development	×	0.14
Prompt engineering techniques such as self-consistency, chain-of-thought, and generated knowledge significantly enhance	✓	0.34
Context Optimization (CoOp), Conditional Context Optimization (CoCoOp), and Multimodal Prompt Learning (MaPLe) are promp	✓	0.31
Prompt methods can be evaluated through both subjective and objective metrics	×	0.14

References

- https://doi.org/10.48550/arxiv.2402.07867
- https://doi.org/10.48550/arxiv.2310.14735
- https://doi.org/10.48550/arxiv.2311.05232