SOVEREIGN: Foundation model evaluation study MMLU HellaSwag ARC WinoGrande TruthfulQA scores

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Recent advancements in Natural Language Processing (NLP) technologies have been driven at an unprecedented pace by the development of Large Language Models (LLMs). However, challenges remain, such as generating responses that are misaligned with the intent of the question or producing incorrect answers. This paper analyzes various Prompt Engineering techniques for large-scale language models and identifies methods that can optimize response performance across different datasets without the need for extensive retraining or fine-tuning. In particular, we examine prominent Prompt Engineering tech

1 Introduction

Analysis of: Optimizing Large Language Models: A Deep Dive into Effective Prompt Engineering Techniques. Research goal: Foundation model evaluation study MMLU HellaSwag ARC WinoGrande TruthfulQA scores.

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

9 papers retrieved. 6 claims extracted, 5 verified. Tribunal: 7.8/10 $\rightarrow$ 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
The paper analyzes Prompt Engineering techniques including In-Context Learning (ICL), Chain of Thought (CoT), Retrieval-	✓	0.42
The study applies Prompt Engineering techniques to the LLMs Gemma2, LlaMA3, and Mistral.	✓	0.21
Model performance was evaluated using the AI2 Reasoning Challenge (ARC), HellaSwag, Massive Multitask Language Understan	✓	0.35
The evaluation metrics used in the study include BLEU, ROUGE, METEOR, BLEURT, and BERTScore.	×	0.13
The most suitable Prompt Engineering technique varies depending on the characteristics of each dataset.	✓	0.20
For datasets emphasizing mathematical and logical reasoning, Prompt Engineering strategies centered around CoT, SSR, and	✓	0.33

References

- https://doi.org/10.18653/v1/2023.findings-acl.29
- https://doi.org/10.3390/app15031430
- https://doi.org/10.48550/arxiv.2302.13971