SOVEREIGN: Mistral evaluation benchmark results MMLU HumanEval GSM8K performance scores comparison

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: Mistral evaluation benchmark results MMLU HumanEval GSM8K performance scores comparison.

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

12 papers retrieved. 10 claims extracted, 10 verified. Tribunal: 8.8/10 $\rightarrow$ 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
Recent advancements in Natural Language Processing (NLP) technologies have been driven at an unprecedented pace by the d	✓	0.32
Challenges remain in generating responses that are misaligned with the intent of the question or producing incorrect ans	✓	0.26
This paper analyzes various Prompt Engineering techniques for large-scale language models and identifies methods that ca	✓	0.41
The paper examines prominent Prompt Engineering techniques including In-Context Learning (ICL), Chain of Thought (CoT),	✓	0.42
These techniques were applied to leading LLMs such as Gemma2, LlaMA3, and Mistral.	✓	0.17
The performance of these models was evaluated using the AI2 Reasoning Challenge (ARC), HellaSwag, Massive Multitask Lang	✓	0.40
The evaluation metrics included BLEU, ROUGE, METEOR, BLEURT, and BERTScore.	✓	0.16
The experimental results indicate that the most suitable Prompt Engineering technique can vary depending on the characte	✓	0.30
For datasets emphasizing mathematical and logical reasoning, Prompt Engineering strategies centered around CoT, SSR, and	✓	0.35
For datasets focusing on natural language understanding, ICL-centered strategies were found to be more effective.	✓	0.22

References

- https://doi.org/10.48550/arxiv.2312.11444

- https://doi.org/10.3390/app15031430
- https://doi.org/10.48550/arxiv.2403.01976