SOVEREIGN: How do Llama-3-70B, Mistral-8x22B, and Qwen-2.5-72B compare on F1 score for multi-hop QA across HotpotQA, 2Wik

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Prompt engineering has emerged as an indispensable technique for extending the capabilities of large language models (LLMs) and vision-language models (VLMs). This approach leverages task-specific instructions, known as prompts, to enhance model efficacy without modifying the core model parameters. Rather than updating the model parameters, prompts allow seamless integration of pre-trained models into downstream tasks by eliciting desired model behaviors solely based on the given prompt. Prompts can be natural language instructions that provide context to guide the model or learned vector repr

1 Introduction

Analysis of: A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. Research goal: How do Llama-3-70B, Mistral-8x22B, and Qwen-2.5-72B compare on F1 score for multi-hop QA across HotpotQA, 2WikiMultihop, and MuSiQue when context window is fixed at 16k tokens?.

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

8 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 9.0/10 → APPROVE (revision_round=1). 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
Prompt engineering leverages task-specific instructions, known as prompts, to enhance model efficacy without modifying t	✓	0.34
Prompts can be natural language instructions that provide context to guide the model or learned vector representations t	✓	0.31
This survey paper provides a structured overview of recent advancements in prompt engineering, categorized by applicatio	✓	0.25
Prompt engineering has enabled success across various applications, from question-answering to commonsense reasoning	✓	0.25

References

- https://doi.org/10.18653/v1/d18-1259
- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.48550/arxiv.2402.07927