SOVEREIGN: How do different prompt engineering strategies affect the accuracy of long-context retrieval in multimodal lan

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

This article surveys and organizes research works in a new paradigm in natural language processing, which we dub "prompt-based learning." Unlike traditional supervised learning, which trains a model to take in an input x and predict an output y as P (y|x), prompt-based learning is based on language models that model the probability of text directly. To use these models to perform prediction tasks, the original input x is modified using a template into a textual string prompt x' that has some unfilled slots, and then the language model is used to probabilistically fill the unfilled informatio

1 Introduction

Analysis of: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. Research goal: How do different prompt engineering strategies affect the accuracy of long-context retrieval in multimodal language models?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 6.2/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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.

References

- https://doi.org/10.1145/3560815
- https://doi.org/10.48550/arxiv.2303.12712
- https://doi.org/10.1007/s12599-023-00834-7