SOVEREIGN: Can Promptriever's prompting capability be extended to improve robustness against adversarial query perturbati

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: Can Promptriever's prompting capability be extended to improve robustness against adversarial query perturbations in multi-hop QA retrieval tasks, evaluated via accuracy drop under attack compared to non-promptable retrieval models?.

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

10 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.5/10 → 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
Prompt-based learning uses language models that model the probability of text directly, unlike traditional supervised le	✓	0.30
In prompt-based learning, the original input x is modified using a template into a textual string prompt x' with unfille	✓	0.37
Prompt-based learning allows the language model to perform few-shot or even zero-shot learning, adapting to new scenario	✓	0.33
The article introduces a unified set of mathematical notations that can cover a wide variety of existing work on prompt-	✓	0.28
The article organizes existing work along dimensions such as the choice of pre-trained language models, prompts, and tun	✓	0.31

References

- https://doi.org/10.18653/v1/d17-1215
- https://doi.org/10.1145/3560815
- https://doi.org/10.48550/arxiv.2311.05232