SOVEREIGN: Does GPT-4's multi-hop reasoning accuracy on HotpotQA degrade monotonically with increasing retrieval steps (2

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Few-shot prompting is a surprisingly powerful way to use Large Language Models (LLMs) to solve various tasks. However, this approach struggles as the task complexity increases or when the individual reasoning steps of the task themselves are hard to learn, especially when embedded in more complex tasks. To address this, we propose Decomposed Prompting, a new approach to solve complex tasks by decomposing them (via prompting) into simpler sub-tasks that can be delegated to a library of prompting-based LLMs dedicated to these sub-tasks. This modular structure allows each prompt to be optimized f

1 Introduction

Analysis of: Decomposed Prompting: A Modular Approach for Solving Complex Tasks. Research goal: Does GPT-4's multi-hop reasoning accuracy on HotpotQA degrade monotonically with increasing retrieval steps (2 vs 5) under controlled context length, and how does the accuracy-throughput tradeoff compare against a single-step retrieval with wider context window?.

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

7 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 8.2/10 → 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
Few-shot prompting struggles as task complexity increases or when individual reasoning steps are hard to learn	✓	0.27
Decomposed Prompting decomposes complex tasks into simpler sub-tasks that can be delegated to a library of prompting-bas	✓	0.31
Decomposed Prompting allows each prompt to be optimized for its specific sub-task	✓	0.28
Decomposed Prompting can outperform prior work on few-shot prompting using GPT3	✓	0.25
On symbolic reasoning tasks, sub-tasks that are hard for LLMs can be further decomposed into simpler solvable sub-tasks	✓	0.30
When complexity comes from input length, tasks can be recursively decomposed into the same task but with smaller inputs	✓	0.21
On long-context multi-hop QA tasks, sub-tasks can be more effectively taught via separate sub-task prompts	✓	0.27
On open-domain multi-hop QA, symbolic information retrieval can be incorporated within the decomposition framework	✓	0.23

References

- https://doi.org/10.4230/tgdk.1.1.7
- https://doi.org/10.48550/arxiv.2302.04023

- https://doi.org/10.48550/arxiv.2210.02406