SOVEREIGN: To what extent does the accuracy of multi-step retrieval pipelines for multi-hop QA degrade under noisy or adv

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

This paper proposes a framework for quantitatively evaluating interactive LLMs such as ChatGPT using publicly available data sets. We carry out an extensive technical evaluation of ChatGPT using 23 data sets covering 8 different common NLP application tasks. We evaluate the multitask, multilingual and multi-modal aspects of ChatGPT based on these data sets and a newly designed multimodal dataset. We find that ChatGPT outperforms LLMs with zero-shot learning on most tasks and even outperforms fine-tuned models on some tasks. We find that it is better at understanding non-Latin script languages

1 Introduction

Analysis of: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. Research goal: To what extent does the accuracy of multi-step retrieval pipelines for multi-hop QA degrade under noisy or adversarial intermediate contexts, and does this degradation scale with the number of hops (2 vs 5) across GPT-4 and open-source 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

9 papers retrieved. 10 claims extracted, 10 verified. Tribunal: 9.2/10 → 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
This paper proposes a framework for quantitatively evaluating interactive LLMs such as ChatGPT using publicly available	✓	0.33
We carry out an extensive technical evaluation of ChatGPT using 23 data sets	✓	0.30
We evaluate the multitask, multilingual and multi-modal aspects of ChatGPT based on these data sets and a newly designed	✓	0.36
ChatGPT outperforms LLMs with zero-shot learning on most tasks and even outperforms fine-tuned models on some tasks	✓	0.29
It is better at understanding non-Latin script languages than generating them	✓	0.22
ChatG-PT is 63.41% accurate on average in 10 different reasoning categories under logical reasoning, non-textual reasoni	✓	0.31
It is able to generate multimodal content from textual prompts, via an intermediate code generation step	✓	0.26
ChatGPT suffers from hallucination problems like other LLMs and it generates more extrinsic hallucinations from its para	✓	0.33
The interactive feature of ChatGPT enables human collaboration with the underlying LLM to improve its performance	✓	0.26
8% ROUGE-1 on summarization and 2% ChrF++ on machine translation, in a multi-turn 'prompt engineering' fashion	✓	0.25

References

- https://doi.org/10.48550/arxiv.2302.04023
- https://doi.org/10.48550/arxiv.2402.07927
- https://doi.org/10.1109/access.2021.3140175