SOVEREIGN: What is the trade-off between answer accuracy (F1/EM) and inference time when applying Vendi-RAG's iterative d

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Abstract Large language model (LLM) systems, such as ChatGPT 1 or Gemini 2 , can show impressive reasoning and question-answering capabilities but often 'hallucinate' false outputs and unsubstantiated answers 3,4 . Answering unreliably or without the necessary information prevents adoption in diverse fields, with problems including fabrication of legal precedents 5 or untrue facts in news articles 6 and even posing a risk to human life in medical domains such as radiology 7 . Encouraging truthfulness through supervision or reinforcement has been only partially successful 8 . Researchers need a

1 Introduction

Analysis of: Detecting hallucinations in large language models using semantic entropy. Research goal: What is the trade-off between answer accuracy (F1/EM) and inference time when applying Vendi-RAG's iterative diversity-aware retrieval to multi-document QA on the TriviaQA dataset compared to standard relevance-only RAG?.

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

6 papers retrieved. 10 claims extracted, 8 verified. Tribunal: 8.0/10 → 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
Large language model systems like ChatGPT or Gemini can show hallucinations that prevent adoption in diverse fields	✓	0.16
Large language models can fabricate legal precedents or create untrue facts in news articles	✓	0.17
Hallucinations in large language models can pose a risk to human life in medical domains such as radiology	✓	0.22
Supervision or reinforcement for encouraging truthfulness has been only partially successful	×	0.15
Researchers need a general method for detecting hallucinations in LLMs that works with new and unseen questions	✓	0.32
The entropy-based uncertainty estimators can detect confabulations in LLMs	✓	0.18
	×	0.00
The method works across datasets and tasks without a priori knowledge of the task	✓	0.23
The method requires no task-specific data and robustly generalizes to new tasks not seen before	✓	0.25
The method helps users understand when LLMs must take extra care	✓	0.18

References

- https://doi.org/10.18653/v1/d18-1259
- https://doi.org/10.1038/s41586-024-07421-0

- https://doi.org/10.48550/arxiv.2312.10997