SOVEREIGN: How does the predictive expert caching strategy in ExpertFlow affect multi-object hallucination rates (measure

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Since the introduction of ChatGPT, large language models (LLMs) have demonstrated significant utility in various tasks, such as answering questions through retrieval-augmented generation. Context can be retrieved using a vectorized database, serving as a foundation for LLMs to generate responses. However, hallucinations in responses can undermine the reliability of LLMs in practical applications, and they are not easily detectable in the absence of ground truth, particularly in question-and-answer scenarios. This paper proposes a framework that integrates multiple small language models to veri

1 Introduction

Analysis of: Hallucination Detection with Small Language Models. Research goal: How does the predictive expert caching strategy in ExpertFlow affect multi-object hallucination rates (measured by mPOPE or entity-level F1) compared to dense models of equivalent FLOPs in MoE-VLMs across varying numbers of active experts?.

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

12 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 3.3/10 → REJECT (revision_round=0). 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.

5 Extracted Claims

Claim	Verified	Confidence
The proposed framework for detecting hallucinations using multiple SLMs improves verification accuracy while mitigating	×	0.09
A dataset was prepared to verify the proposed framework, demonstrating an improvement of 10% over the baseline.	×	0.06
The proposed method for computing hallucination scores using multiple SLMs is superior to both P(yes) and ChatGPT approa	×	0.07
Utilizing multiple SLMs improves performance in detecting correct responses from partial responses.	×	0.14
The 'max' method achieves the highest F1 score of 0.99 in hallucination detection among different mean calculation appro	×	0.11
The 'harmonic' method achieves an F1 score of 0.81, which is the highest among the methods tested in Fig. 5(b).	×	0.08

References

- http://arxiv.org/abs/2505.20225v1
- http://arxiv.org/abs/2506.22486v1
- http://arxiv.org/abs/2407.06192v2