TAE Token Misalignment Threshold Effects on Hallucination and Coherence in Vicuna-13B and Baichuan-2

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the TAE token misalignment threshold affect the trade-off between hallucination rates and response coherence in Vicuna-13B and Baichuan-2 when evaluated on the TruthfulQA benchmark. 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. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Hallucination Detection with Small Language Models. Research question: How does the TAE token misalignment threshold affect the trade-off between hallucination rates and response coherence in Vicuna-13B and Baichuan-2 when evaluated on the TruthfulQA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.0/10.

3 Results

15 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The proposed framework for detecting hallucinations using multiple SLMs demonstrates an improvement of 10% over the base	×	0.10
The transformer architecture marked a pivotal turning point in NLP by enabling models to effectively capture long-range	×	0.03
LLMs are known to produce hallucinations in their outputs, and there is no clear definition of hallucination.	×	0.05
Detecting hallucinations is not analogous to conventional LLM measurements such as ROUGE metric and BLEU score.	×	0.03
Utilizing multiple SLMs improves performance in detecting correct responses from partial responses.	×	0.13
The proposed method is superior to both P(yes) and ChatGPT approaches in hallucination detection.	×	0.12
In the F1 score comparison, the 'max' method achieved the highest score of 0.99.	×	0.02
In the F1 score comparison for partial responses, the 'harmonic' method achieved the highest score of 0.81.	×	0.04

References

- http://arxiv.org/abs/2604.17982v1
- http://arxiv.org/abs/2506.22486v1
- http://arxiv.org/abs/2506.09886v2