THaMES Evaluation Pipelines and Multimodal Model Robustness to Factual Caption Errors

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of THaMES evaluation pipelines on the robustness of multimodal models against factually incorrect captions in the ScienceQA dataset. 5 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: THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models. Research question: What is the impact of THaMES evaluation pipelines on the robustness of multimodal models against factually incorrect captions in the ScienceQA dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 5 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
THaMES is divided into three main components: Testset generation from a user-provided corpus, Baseline metric evaluation	×	0.06
The QA-set generation process includes seven steps: Knowledge Base Processing, Ground-Truth Weighted Sampling, Baseline	×	0.09
THaMES is designed to be capable of processing a variety of corpora, including political news articles, academic papers,	×	0.09
THaMES is compatible with various file formats including PDF, TXT, and CSV.	×	0.03
The VectorStoreIndex module provided by LlamaIndex [Liu, 2022] is utilized to build a knowledge base out of the raw corp	×	0.03

References

- http://arxiv.org/abs/2409.11353v3
- http://arxiv.org/abs/2604.00086v1
- http://arxiv.org/abs/2306.09265v1