

# THaMES Framework Impact on Factual Consistency in 7B and 70B Language Models

Assignee Research

June 8, 2026

## Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: To what extent does the THaMES framework improve factual consistency scores for 7B versus 70B LLMs across domain-specific QA benchmarks. 5 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: To what extent does the THaMES framework improve factual consistency scores for 7B versus 70B LLMs across domain-specific QA benchmarks?.

## 2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

## 3 Results

12 papers retrieved. 5 claims extracted; 0 independently verified. Quality review score: 3.8/10.

## 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 framework is divided into three main components: Testset generation from a user-provided corpus, Baseline metric	×	0.05
THaMES is designed to be capable of processing a variety of corpora, including political news articles, academic papers,	×	0.08
THaMES framework is compatible with various file formats including PDF, TXT, and CSV.	×	0.03
The QA-set generation process in THaMES includes seven steps: Knowledge Base Processing, Ground-Truth Weighted Sampling,	×	0.08
The VectorStoreIndex module provided by LlamaIndex is utilized to build a knowledge base out of the raw corpus in THaMES	×	0.04

## References

- <http://arxiv.org/abs/2409.11353v3>
- <http://arxiv.org/abs/2605.04999v2>
- <http://arxiv.org/abs/2509.12382v1>