# THaMES-Driven Alignment Fine-Tuning for Factual Consistency Without Perplexity Degradation

Assignee Research

June 6, 2026

**Abstract**

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Can THaMES-driven alignment fine-tuning improve factual consistency scores on the TruthfulQA benchmark without degrading general language generation perplexity. 10 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Grounded Visual Factualization: Factual Anchor-Based Finetuning for Enhancing MLLM Factual Consistency. Research question: Can THaMES-driven alignment fine-tuning improve factual consistency scores on the TruthfulQA benchmark without degrading general language generation perplexity?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.9/10.

## 3 Results

13 papers retrieved. 10 claims extracted; 3 independently verified. Quality review score: 5.9/10.

## 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 |
|---|---|---|
| LLaVA-1.5-13B is used as the base model for experiments. | × | 0.05 |
| The training dataset consists of 80% of the Open-Ended Question (OEQ) benchmark from VHTest, totaling 960 samples. | × | 0.09 |
| The test data consists of the remaining 20% of the VHTest OEQ and Yes/No Question (YNQ) benchmarks, totaling 480 samples | × | 0.09 |
| The fine-tuned models are evaluated on general multimodal benchmarks, including MME and POPE. | × | 0.13 |
| The fine-tuning process uses the AdamW optimizer with a learning rate of 4e-6, a batch size of 16, and a cosine annealin | × | 0.04 |
| The training duration is estimated to be approximately 18-25 minutes on a single A6000 GPU. | × | 0.01 |
| Data preprocessing includes rewriting answers for counting-related questions and incorporating specific prompts for posi | × | 0.01 |
| GVF Finetuning integrates explicit factual signals via Factual Anchor Data Augmentation, Fact-Aware Instruction Tuning, | ✓ | 0.38 |
| GVF Finetuning significantly outperforms standard fine-tuning on the VHTest benchmark for both Open-Ended Question (OEQ) | ✓ | 0.33 |
| GVF Finetuning maintains or slightly improves performance on general multimodal benchmarks like MME and POPE. | ✓ | 0.29 |

## References


- http://arxiv.org/abs/2409.11353v3

- http://arxiv.org/abs/2511.10671v1
- http://arxiv.org/abs/2309.02144v1