Diminishing Returns in Verification Accuracy from Scaling Diverse Debating Agents on the FEVER-LC Benchmark

Assignee Research

June 12, 2026

Abstract

Large Language Models (LLMs) suffer from hallucinations and factual inaccuracies, especially in complex reasoning and fact verification tasks. Multi-Agent Debate (MAD) systems aim to improve answer accuracy by enabling multiple LLM agents to engage in dialogue, promoting diverse reasoning and mutual verification. However, existing MAD frameworks primarily rely on internal knowledge or static documents, making them vulnerable to hallucinations. While MADKE introduces external evidence to mitigate this, its one-time retrieval mechanism limits adaptability to new arguments or emerging information

1 Introduction

This paper examines: Tool-MAD: A Multi-Agent Debate Framework for Fact Verification with Diverse Tool Augmentation and Adaptive Retrieval. Research question: Does scaling the number of debating agents with diverse retrieval strategies yield diminishing returns in verification accuracy on the FEVER-LC benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

14 papers retrieved. 12 claims extracted; 10 independently verified. Quality review score: 7.8/10.

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
FactScore attempts to quantify factuality at the claim level by decomposing model outputs.	✓	0.18
Consistency-based factuality approaches, including self-consistency and re-asking, evaluate factual correctness by measu	✓	0.31
RAGAS introduced two complementary metrics: faithfulness, measuring whether claims match retrieved evidence, and answer	✓	0.22
Faithfulness and answer relevance metrics have predominantly been applied as post-hoc evaluators for fully generated out	✓	0.27
Tool-MAD incorporates faithfulness and answer relevance as round-level stability indicators, which is distinct from exis	✓	0.22
Toolformer enables models to learn API-calling behaviors.	×	0.12
HuggingGPT and GEAR treat the LLM as a coordinator for heterogeneous models or systems.	×	0.14
Domain-specific tools can dramatically improve reasoning fidelity in scientific and professional domains.	✓	0.21
Existing tool-augmented systems overwhelmingly adopt a single-agent perspective and use tools in a one-shot or sequentia	✓	0.28
Retrieval systems typically perform a single retrieval step at the beginning of a task, without adapting to evolving rea	✓	0.22
Tool-MAD consistently outperforms competitive multi-agent debate frameworks such as MAD and MADKE, achieving performance	✓	0.25
Tool-MAD demonstrates its flexibility in medical QA settings, maintaining robust performance under different retrieval t	✓	0.30

References

- http://arxiv.org/abs/2409.08479v2
- http://arxiv.org/abs/2507.19090v4
- http://arxiv.org/abs/2601.04742v1