Reasoning Performance Gap Between CLAM-Trained and Token-Based Agents on BridgeData V2 Under Partial Observability

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the performance gap in reasoning capabilities between CLAM-trained agents and token-based agents on the BridgeData V2 benchmark when evaluated using the ALFRED task completion score under. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: DefenderBench: A Toolkit for Evaluating Language Agents in Cybersecurity Environments. Research question: What is the performance gap in reasoning capabilities between CLAM-trained agents and token-based agents on the BridgeData V2 benchmark when evaluated using the ALFRED task completion score under partial observability conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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.

References

- http://arxiv.org/abs/2412.06333v3
- http://arxiv.org/abs/2604.14552v2
- http://arxiv.org/abs/2506.00739v4