Language Models Solving Competition-Level Software Engineering Problems: Techniques and Benchmark Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What techniques enable language models to solve competition-level software engineering problems v14. 13 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Kodezi Chronos: A Debugging-First Language Model for Repository-Scale Code Understanding. Research question: What techniques enable language models to solve competition-level software engineering problems v14.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 13 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
Chronos-1 achieves 65.3% debugging success compared to 31.2% for other solutions.	×	0.08
Chronos-1 operates with continuous memory updates, specialized debugging knowledge, and efficient graph-guided retrieval	×	0.08
Chronos-1 maintains persistent debug memory (PDM) across millions of debugging sessions.	×	0.10
Chronos-1 is trained specifically on 15M+ debugging scenarios.	×	0.06
Chronos-1 implements a specialized 7-layer debugging architecture with automatic test validation, iterative refinement,	×	0.05
Chronos-1 achieves 4-5x better debugging performance than state-of-the-art alternatives.	×	0.08
Claude 4.5 Opus achieves 74.40% on SWE-bench Verified.	✓	0.24
Gemini 3 Pro reaches 76.2% on SWE-bench Verified.	✓	0.18
Gemini 3 Pro achieves 91.9% GPQA Diamond, 95-100% AIME 2025, 2,439 Elo on LiveCodeBench Pro.	×	0.05
Claude 4.1 Opus and Claude 4.5 Sonnet achieve 72.5% and 72.7% respectively on SWE-bench Full (code generation).	×	0.11
Claude 4.1 Opus and Claude 4.5 Sonnet achieve 67.60% and 70.60% respectively on SWE Bench Bash Only.	×	0.08
Traditional LLM Planning has a 23% success rate for implementing a state machine.	×	0.03
AGR-Enhanced Debugging has an 87% success rate for implementing a state machine.	×	0.04

References

- http://arxiv.org/abs/cs/0101002v1
- http://arxiv.org/abs/2507.12482v4
- http://arxiv.org/abs/2406.04710v2