Techniques for Solving Competition-Level Software Engineering Problems with Language Models

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What techniques enable language models to solve competition-level software engineering problems v17. 16 claims were extracted from source literature; 2 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: LLM-ProS: Analyzing Large Language Models' Performance in Competitive Problem Solving. Research question: What techniques enable language models to solve competition-level software engineering problems v17.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 16 claims extracted; 2 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
The study uses a curated dataset of 166 World Finals problems from 2011 to 2024.	✓	0.23
The study evaluates five models: GPT-4o, Mistral Large, Llama-3.1-405B, o1-mini, and o1-preview.	✓	0.25
In the benchmark results, o1-mini achieved 16 Accepted (AC) verdicts out of the total test cases.	×	0.05
In the benchmark results, o1-preview achieved 15 Accepted (AC) verdicts.	×	0.06
In the benchmark results, GPT-4o achieved 0 Accepted (AC) verdicts.	×	0.06
In the benchmark results, Mistral-Large achieved 0 Accepted (AC) verdicts.	×	0.06
In the benchmark results, Llama-3.1 achieved 0 Accepted (AC) verdicts.	×	0.04
o1-mini received 124 Wrong Answer (WA) verdicts in the evaluation.	×	0.05
o1-preview received 120 Wrong Answer (WA) verdicts in the evaluation.	×	0.04
GPT-4o received 41 Compile Error (CE) verdicts and 39 Runtime Error (RE) verdicts.	×	0.03
The problem 'Allied Chute Manufacturers' involves input where each test case starts with an integer n representing the n	×	0.02
The sample output for Case 1 of the 'Allied Chute Manufacturers' problem is 2.40.	×	0.03
The sample output for Case 2 of the 'Allied Chute Manufacturers' problem is 14.15.	×	0.03
o1-mini and o1-preview consistently outperform other evaluated LLMs in accuracy, verdict distribution, and resource effi	×	0.11
Models with specialized training for chain-of-thought reasoning exhibit greater robustness and adaptability to unseen pr	×	0.09
General-purpose models showed a significant performance drop on unseen data compared to specialized models.	×	0.07

References

- http://arxiv.org/abs/2502.04355v1
- http://arxiv.org/abs/2502.20868v1
- http://arxiv.org/abs/2305.06599v3