DeepSeek-V3 and GPT-4 Precision and Recall in Code Smell Detection Against Human Annotations

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 18 peer-reviewed papers addressing the following research question: What are the precision and recall metrics for DeepSeek-V3 in detecting specific code smell categories compared to human-annotated ground truth. Determining which Large Language Model (LLM) is superior for code smell detection is a complex challenge. This study aims to establish a systematic methodology and evaluation matrix to address this question. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking LLM for Code Smells Detection: OpenAI GPT-4.0 vs DeepSeek-V3. Research question: What are the precision and recall metrics for DeepSeek-V3 in detecting specific code smell categories compared to human-annotated ground truth?.

2 Methodology

Systematic literature search across multiple databases yielded 18 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

18 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 4.7/10.

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
GPT-4.0 achieved a precision of 0.79 in code smell detection	×	0.09
GPT-4.0 achieved a recall of 0.41 in code smell detection	×	0.10
GPT-4.0 achieved an F1-score of 0.54 in code smell detection	×	0.11
DeepSeek achieved a precision of 0.42 in code smell detection	×	0.09
DeepSeek achieved a recall of 0.31 in code smell detection	×	0.09
DeepSeek achieved an F1-score of 0.35 in code smell detection	×	0.11
GPT-4.0 has superior precision compared to DeepSeek-V3 in code smell detection	✓	0.16
DeepSeek-V3 produces a significantly higher number of false positives than GPT-4.0	×	0.07
GPT-4.0 has a lower recall rate than DeepSeek-V3 in code smell detection	×	0.13

References

- http://arxiv.org/abs/2504.16027v1
- https://arxiv.org/abs/2504.16027
- https://www.semanticscholar.org/paper/e9d301dc0fc09141f20f46e309d884de2a58ff6d