Robustness of TabMWP Evaluation Scores Across Pretraining Data Types Under Adversarial and Noisy Conditions

Assignee Research

June 11, 2026

Abstract

Large Language Models offer new opportunities to devise automated implementation generation methods that can tackle problem solving activities beyond traditional methods, which require algorithmic specifications and can use only static domain knowledge, like performance metrics and libraries of basic building blocks. Large Language Models could support creating new methods to support problem solving activities for open-ended problems, like problem framing, exploring possible solving approaches, feature elaboration and combination, more advanced implementation assessment, and handling unexpecte

1 Introduction

This paper examines: An Overview and Discussion on Using Large Language Models for Implementation Generation of Solutions to Open-Ended Problems. Research question: How does the robustness of TabMWP evaluation scores vary between models pretrained on synthetic-only versus mixed data when exposed to adversarial tabular inputs or noisy real-world datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

1 papers retrieved. 11 claims extracted; 10 independently verified. Quality review score: 8.5/10.

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
Traditional implementation generation methods require algorithmic specifications.	✓	0.28
Traditional implementation generation methods can use only static domain knowledge, such as performance metrics and libr	✓	0.35
Large Language Models could support problem framing for open-ended problems.	✓	0.30
Large Language Models could support exploring possible solving approaches for open-ended problems.	✓	0.31
Large Language Models could support feature elaboration and combination for open-ended problems.	✓	0.27
Large Language Models could support more advanced implementation assessment for open-ended problems.	✓	0.28
Large Language Models could support handling unexpected situations for open-ended problems.	✓	0.27
The report summarizes current work on Large Language Models including model prompting.	✓	0.24
The report summarizes current work on Large Language Models including Reinforcement Learning.	✓	0.21
The report summarizes current work on Large Language Models including Retrieval-Augmented Generation.	✓	0.25
The report discusses future research requirements.	×	0.13

References

- https://doi.org/10.48550/arxiv.2501.00562