Training Strategies for Language Model Generalization in Mathematical Reasoning

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What training strategies improve language model generalization to novel mathematical reasoning problems v9. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Unraveling Arithmetic in Large Language Models: The Role of Algebraic Structures. Research question: What training strategies improve language model generalization to novel mathematical reasoning problems v9.

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 3.1/10.

3 Results

15 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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 experiments use addition problems testing commutativity and identity for operators '+' and '$\oplus$', and operations invol	×	0.04
The experiments set n = 7, 11, and 13 for Zn.	×	0.01
The number of input elements M is set to 6.	×	0.05
The language model used is GPT-2.	×	0.05
The GPT-2 model weights were reinitialized before training to remove pre-existing knowledge.	×	0.03
The tokenizer was customized so that each element (e.g., z10) is represented as a single token.	×	0.03
Each training or testing set with scale K contains 10K instances.	×	0.02
For scale K, 4K instances involve operators with commutativity and identity ('+', '$\oplus$'), with K instances for each of the	×	0.05
For scale K, 6K instances involve operators without commutativity and identity (", ", "), with 2K instances for ea	×	0.03
The testing set scale K is fixed at 1000 throughout the experiments.	×	0.01
The training set scale K ranges from 100 to 30,000.	×	0.02
In the Z7 case with training set K=3000, the model achieved 100% accuracy on the training set.	×	0.03
In the Z7 case with training set K=3000, the model did not achieve high accuracy on the testing set for the commutativit	×	0.02
Zn is defined as the set of integers modulo n under addition modulo n.	×	0.01
In Z5, the elements are {0, 1, 2, 3, 4} and the identity element is 0.	×	0.04
The dataset consists of addition problems with M terms expressed as zi1 + zi2 + $\cdot$ $\cdot$ $\cdot$ + ziM = z(i1+i2+$\cdot$$\cdot$$\cdot$+iM mod n).	×	0.01

References

- http://arxiv.org/abs/2509.25160v1

- http://arxiv.org/abs/2411.16260v3
- http://arxiv.org/abs/2510.17496v2