# MathCoder2 Pretraining Enhances Adversarial Robustness in Sub-3B Math Models

Assignee Research

June 6, 2026

**Abstract**

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does the MathCoder2 pretraining approach improve robustness against adversarial perturbations in competition-level math problems for models under 3B parameters. 17 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: Nemotron-CC-Math: A 133 Billion-Token-Scale High Quality Math Pretraining Dataset. Research question: Does the MathCoder2 pretraining approach improve robustness against adversarial perturbations in competition-level math problems for models under 3B parameters?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 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

13 papers retrieved. 17 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 |
|---|---|---|
| Nemotron-CC-Math contains 133.26 billion tokens. | ✓ | 0.15 |
| Nemotron-CC-Math consists of 101.15 million documents. | × | 0.08 |
| The Nemotron-CC-Math construction pipeline starts from 98 Common Crawl snapshots. | ✓ | 0.15 |
| The base Nemotron-T 8B model was pretrained on 9 trillion tokens. | × | 0.09 |
| In the ablation studies, the target math dataset constitutes 30% of the total data blend. | × | 0.04 |
| The 100B Token Ablation setting targets math datasets typically below 30 billion tokens. | × | 0.07 |
| Code generation quality is measured on MBPP, HumanEval, HumanEval+, and MBPP+. | × | 0.05 |
| Code task results are reported as avg@20, representing the average accuracy from generating 20 samples per prompt. | × | 0.05 |
| Code samples are generated using nucleus sampling with a temperature of 0.6 and a top-p value of 0.95. | × | 0.04 |
| Mathematical reasoning is evaluated on the GSM8K and MATH benchmarks. | × | 0.14 |
| Mathematical reasoning evaluations use greedy decoding and Math-Verify5 for symbolic matching. | × | 0.10 |
| Knowledge understanding is assessed using MMLU, MMLU-Pro, and MMLU-STEM datasets. | × | 0.02 |
| Knowledge understanding results are reported as exact match (EM) accuracy. | × | 0.02 |
| All model evaluations were run using lm-evaluation-harness. | × | 0.03 |
| Mathematics constitutes 60.28% of the content in the Nemotron-CC-Math dataset. | × | 0.11 |
| Physics constitutes 11.22% of the content in the Nemotron-CC-Math dataset. | × | 0.09 |
| Chemistry constitutes 1.71% of the content in the Nemotron-CC-Math dataset. | × | 0.10 |

## References


- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2410.08196v1
- http://arxiv.org/abs/2508.15096v1