# Novelty Assessment Report

**Paper**: MathFimer: Enhancing Mathematical Reasoning by Expanding Reasoning Steps through Fill-in-the-Middle Task

**PDF URL**: https://openreview.net/pdf?id=14i2wzPPfn

**Venue**: ICLR 2026 Conference Submission

**Year**: 2026

**Report Generated**: 2026-01-05

## Abstract

Mathematical reasoning represents a critical frontier in advancing large language models (LLMs). While step-by-step approaches have emerged as the dominant paradigm for mathematical problem-solving in LLMs, the quality of reasoning steps in training data fundamentally constrains model performance. Recent studies has demonstrated that more detailed intermediate steps can enhance model performance, yet existing methods for step expansion either require more powerful external models or incur substantial computational costs. In this paper, we introduce MathFimer, a novel framework for mathematical reasoning step expansion inspired by the "Fill-in-the-middle" from code completion. By decomposing solution chains into prefix-suffix pairs and training models to reconstruct missing intermediate steps, we develop a specialized model, MathFimer-7B, on our carefully curated NuminaMath-Fim dataset. We then apply these models to enhance existing mathematical reasoning datasets by inserting detailed intermediate steps into their solution chains, creating MathFimer-expanded versions. Through comprehensive experiments on multiple mathematical reasoning datasets, including MathInstruct and MetaMathQA, we demonstrate that models trained on MathFimer-expanded data consistently outperform their counterparts trained on original data across various benchmarks such as GSM8K and MATH. Our approach offers a practical, scalable solution for enhancing mathematical reasoning capabilities in LLMs without relying on more powerful external models or expensive inference procedures.

**Disclaimer**

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

## Core Task Landscape

This paper addresses: **Expanding Mathematical Reasoning Steps through Fill-in-the-Middle Task**

A total of **18 papers** were analyzed and organized into a taxonomy with **14 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Step Expansion and Intermediate Reasoning Generation**
- **Reasoning Correction and Verification Mechanisms**
- **Fill-in-the-Middle Training Paradigms and Architectures**
- **Backward Reasoning and Inverse Problem Formulation**
- **Tool Use and External Knowledge Grounding**
- **Context Utilization and Long-Context Modeling**
- **Formal Verification and Proof Generation**
- **Tokenization and Byte-Level Modeling**
- **Educational Perspectives and Pedagogical Applications**

### Complete Taxonomy Tree

- Expanding Mathematical Reasoning Steps through Fill-in-the-Middle Task Survey Taxonomy
- Step Expansion and Intermediate Reasoning Generation
  - Fill-in-the-Middle Based Step Expansion ★ (2 papers)
  - [0] MathFimer: Enhancing Mathematical Reasoning by Expanding Reasoning Steps through Fill-in-the-Middle Task (Anon et al., 2026) View paper
  - [6] ClozeMath: Improving Mathematical Reasoning in Language Models by Learning to Fill Equations (Pham Quang Hieu, 2025) View paper
  - Thought Leap Detection and Bridging (1 papers)
  - [7] Mind the Gap: Bridging Thought Leap for Improved Chain-of-Thought Tuning (Yan, 2025) View paper
  - Enriched Instruction Tuning for Multi-Step Reasoning (1 papers)
  - [1] System-2 mathematical reasoning via enriched instruction tuning (Cai, 2024) View paper
- Reasoning Correction and Verification Mechanisms
  - Search-Based Reasoning Correction (1 papers)
  - [2] Search-Based Correction of Reasoning Chains for Language Models (M Kim, 2025) View paper
  - Pseudo-Dual Task Reexamination (1 papers)
  - [8] Solving math word problems with reexamination (Bin Yi, 2023) View paper
- Fill-in-the-Middle Training Paradigms and Architectures
  - Planning-Aware Infilling with Horizon Prediction (2 papers)
  - [13] Planning-Aware Code Infilling via Horizon-Length Prediction (Yifeng Ding, 2025) View paper
  - [14] Horizon-Length Prediction: Advancing Fill-in-the-Middle Capabilities for Code Generation with Lookahead Planning (Ding YiFeng, 2024) View paper
  - Masked Diffusion Models for Infilling (2 papers)
  - [12] No Compute Left Behind: Rethinking Reasoning and Sampling with Masked Diffusion Models (Singhal Raghav, 2025) View paper

    - [15] Scaling Diffusion Language Models via Adaptation from Autoregressive Models (Gong, 2024) View paper
    - Self-Supervised Skip-Tree Training for Formal Mathematics (1 papers)
    - [11] Mathematical Reasoning via Self-supervised Skip-tree Training (Markus N. Rabe, 2021) View paper
- Backward Reasoning and Inverse Problem Formulation (1 papers)
    - [17] Fill in the Blank: Exploring and Enhancing LLM Capabilities for Backward Reasoning in Math Word Problems (Deb Aniruddha, 2023) View paper
- Tool Use and External Knowledge Grounding (1 papers)
    - [5] Efficient tool use with chain-of-abstraction reasoning (Gao, 2025) View paper
- Context Utilization and Long-Context Modeling (1 papers)
    - [3] Make your llm fully utilize the context (Shengnan An, 2024) View paper
- Formal Verification and Proof Generation (1 papers)
    - [4] From Informal to Formal--Incorporating and Evaluating LLMs on Natural Language Requirements to Verifiable Formal Proofs (Jialun Cao, 2025) View paper
- Tokenization and Byte-Level Modeling (1 papers)
    - [9] Exact byte-level probabilities from tokenized language models for fim-tasks and model ensembles (Phan, 2024) View paper
- Educational Perspectives and Pedagogical Applications (3 papers)
    - [10] Seeing the continuity behind â□□double discontinuityâ□□: Investigating Hong Kong prospective mathematics teachers' secondaryâ□□tertiary transition (Biyao Liang, 2023) View paper
    - [16] Open Middle Math (Robert Kaplinsky, 2023) View paper
    - [18] The Science of Proof: Mathematical Reasoning and Its Limitations (Faris, n.d.) View paper

### Narrative

Core task: expanding mathematical reasoning steps through fill-in-the-middle task. The field addresses how language models can generate, verify, and refine intermediate reasoning steps in mathematical problem-solving. The taxonomy reveals several complementary directions: some branches focus on step expansion and intermediate reasoning generation, exploring how models can produce missing derivations between problem statements and solutions; others examine reasoning correction and verification mechanisms that detect and fix errors in multi-step chains. Fill-in-the-middle training paradigms investigate architectural choices for infilling tasks, while backward reasoning explores inverse problem formulation. Additional branches address tool use and external knowledge grounding (such as Chain-of-Abstraction[5]), context utilization for long-form reasoning (Fully Utilize Context[3]), formal verification and proof generation (Informal to Formal[4]), and even tokenization strategies at the byte level. Educational perspectives consider pedagogical applications, reflecting the dual role of these methods in both advancing AI capabilities and supporting human learning.

Particularly active lines of work contrast autoregressive step expansion with explicit infilling objectives, and explore whether models benefit more from forward generation or backward reasoning (Backward Reasoning[17]). Search-based correction methods (Search-Based Correction[2]) and system-level reasoning frameworks (System-2 Mathematical Reasoning[1]) highlight ongoing debates about how to balance generation fluency with verification rigor. MathFimer[0] sits within the fill-in-the-middle based step expansion cluster, closely aligned with ClozeMath[6], which similarly frames intermediate step generation as a cloze-style infilling problem. Compared to approaches that emphasize post-hoc verification or tool-augmented reasoning, MathFimer[0] and its neighbors prioritize training models to natively predict missing derivation steps, leveraging the fill-in-the-middle objective to encourage coherent bridging between given context and conclusions. This positions the work as part of a growing effort to make reasoning expansion an intrinsic model capability rather than an external search or correction process.

## Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. ClozeMath: Improving Mathematical Reasoning in Language Models by Learning to Fill Equations

**Authors**: Pham Quang Hieu, Nguyen Thuy Duong, Pham, Tung, Luu Anh Tuan, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

The capabilities of large language models (LLMs) have been enhanced by training on data that reflects human thought processes, such as the Chain-of-Thought format. However, evidence suggests that the conventional scheme of next-word prediction may not fully capture how humans learn to think. Inspired by how humans generalize mathematical reasoning, we propose a new approach named ClozeMath to fine-tune LLMs for mathematical reasoning. Our ClozeMath involves a text-infilling task that predicts ma...

#### Relationship Analysis

Both papers belong to the Fill-in-the-Middle Based Step Expansion category, using fill-in-the-middle tasks to enhance mathematical reasoning by predicting missing intermediate steps. While MathFimer focuses on decomposing existing solutions into prefix-suffix pairs and training specialized models to insert detailed intermediate steps between consecutive reasoning steps, ClozeMath specifically targets mathematical equations for masking and prediction, combining text-infilling with language modeling objectives using a PrefixLM architecture. The key distinction is that MathFimer expands general reasoning steps in existing solutions, whereas ClozeMath selectively masks and predicts equations while preserving textual rationales, aiming to align more closely with human learning patterns through cloze-style exercises.

## Contributions Analysis

**Overall novelty summary.** The paper introduces MathFimer, a framework that applies fill-in-the-middle training to expand mathematical reasoning steps, producing a specialized 7B model and an enhanced dataset. Within the taxonomy, it resides in the 'Fill-in-the-Middle Based Step Expansion' leaf alongside one sibling paper (ClozeMath). This leaf is part of a broader 'Step Expansion and Intermediate Reasoning Generation' branch containing three leaves total, indicating a moderately active but not overcrowded research direction focused on generating missing intermediate steps.

The taxonomy reveals neighboring branches addressing reasoning correction and verification mechanisms, fill-in-the-middle training paradigms across domains, and backward reasoning approaches. MathFimer's leaf sits adjacent to 'Thought Leap Detection and Bridging' and 'Enriched Instruction Tuning for Multi-Step Reasoning', which tackle similar step-completion goals through different mechanisms (detecting omissions versus human-AI feedback synergy). The broader taxonomy includes 18 papers across diverse directions, suggesting the field balances step expansion with verification, tool use, and formal proof generation, positioning MathFimer within a specific niche of proactive infilling-based expansion.

Among 30 candidates examined, none clearly refute any of the three contributions. The MathFimer framework (10 candidates examined, 0 refutable), the NuminaMath-FIM dataset and model (10 candidates, 0 refutable), and the empirical performance demonstrations (10 candidates, 0 refutable) all appear novel within this limited search scope. The single sibling paper in the same taxonomy leaf suggests the specific application of fill-in-the-middle to mathematical step expansion remains relatively underexplored, though the broader step expansion category contains multiple alternative approaches that address overlapping goals through different technical means.

Based on the top-30 semantic matches and taxonomy structure, the work appears to occupy a distinct position within mathematical reasoning step expansion. The limited search scope and sparse sibling count suggest novelty, though the taxonomy shows active neighboring research in related verification and training paradigm directions. The analysis covers semantic proximity and structural taxonomy placement but does not exhaustively survey all mathematical reasoning literature or adjacent code-completion domains that inspired the approach.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

## Contribution 1: MathFimer framework for mathematical reasoning step expansion

**Description**: The authors introduce MathFimer, a framework that adapts the fill-in-the-middle paradigm from code reasoning to mathematical problem-solving. By decomposing solution chains into prefix-suffix pairs and training models to reconstruct missing intermediate steps, the framework enables targeted expansion of reasoning steps without generating entirely new solution chains.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Codegemma: Open code models based on gemma

**URL**: View paper

**Brief Assessment**

CodeGemma[40] focuses on code completion using fill-in-the-middle techniques for programming tasks, not mathematical reasoning step expansion. The domains and applications are fundamentally different.

### 2. Constrained Decoding for Fill-in-the-Middle Code Language Models via Efficient Left and Right Quotienting of Context-Sensitive Grammars

**URL**: View paper

**Brief Assessment**

Constrained Decoding[41] focuses on fill-in-the-middle for code generation with syntax constraints, not mathematical reasoning step expansion. The candidate addresses code completion in programming languages, while the original paper targets mathematical problem-solving with step-by-step reasoning enhancement.

### 3. Mind the Gap: Bridging Thought Leap for Improved Chain-of-Thought Tuning

**URL**: View paper

**Brief Assessment**

Mind the Gap[7] focuses on detecting and bridging 'thought leaps' (missing steps between existing steps) in complete reasoning chains, while MathFimer uses fill-in-the-middle training to reconstruct missing intermediate steps. Though both address step completeness, they employ fundamentally different technical approaches and problem formulations.

### 4. Efficient tool use with chain-of-abstraction reasoning

**URL**: View paper

**Brief Assessment**

Chain-of-Abstraction[5] focuses on decoupling general reasoning from domain-specific knowledge using abstract placeholders and external tools for multi-step reasoning across mathematical and Wikipedia QA domains. This differs from MathFimer's fill-in-the-middle paradigm that specifically targets inserting missing intermediate steps within existing mathematical solution chains without generating entirely new reasoning paths.

### 5. Make your llm fully utilize the context

**URL**: View paper

**Brief Assessment**

Fully Utilize Context[3] focuses on training LLMs to utilize information across long contexts (4k-32k tokens) through information-intensive training, not on mathematical reasoning step expansion. The candidate addresses the lost-in-the-middle problem in long-context understanding, while the original paper introduces a fill-in-the-middle approach specifically for expanding mathematical reasoning steps.

### 6. Mathematics and Plausible Reasoning: Logic, Symbolic and mathematical

**URL**: View paper

**Brief Assessment**

Plausible Reasoning[39] appears to be a classical mathematics text focused on mathematical reasoning principles and derivation methods. The sparse context provided shows discussion of sequences and Bernoulli's method, but contains no evidence of fill-in-the-middle techniques, LLM training frameworks, or automated step expansion systems that would challenge MathFimer's novelty in adapting FIM paradigms to mathematical reasoning.

### 7. Seeing the continuity behind â□□double discontinuityâ□□: Investigating Hong Kong prospective mathematics teachers' secondaryâ□□tertiary transition

**URL**: View paper

**Brief Assessment**

Double Discontinuity[10] focuses on Hong Kong prospective mathematics teachers' secondary-tertiary transition in education, not on fill-in-the-middle techniques for mathematical reasoning or LLM training frameworks. The candidate paper addresses pedagogical continuity in teacher education rather than computational methods for expanding reasoning steps.

### 8. From Informal to Formal--Incorporating and Evaluating LLMs on Natural Language Requirements to Verifiable Formal Proofs

**URL**: View paper

**Brief Assessment**

Informal to Formal[4] focuses on formal verification tasks (converting natural language requirements to verifiable formal proofs in languages like Coq, Lean4, Dafny), not mathematical reasoning step expansion. The candidate addresses a completely different domain and problem space than MathFimer's fill-in-the-middle approach for mathematical problem-solving.

### 9. CombiBench: Benchmarking LLM Capability for Combinatorial Mathematics

**URL**: View paper

**Brief Assessment**

CombiBench[42] focuses on benchmarking LLMs for combinatorial mathematics using formal verification in Lean~4, not on fill-in-the-middle techniques for expanding reasoning steps in natural language mathematical solutions.

### 10. Beyond the last answer: Your reasoning trace uncovers more than you think

**URL**: View paper

**Brief Assessment**

Beyond Last Answer[43] focuses on analyzing intermediate reasoning steps (subthoughts) to aggregate multiple answer candidates via mode selection, rather than expanding or inserting missing steps into existing solution chains as MathFimer does. The candidate examines answer consistency across reasoning checkpoints, while the original contribution specifically addresses step-level expansion using fill-in-the-middle techniques.

## Contribution 2: NuminaMath-FIM dataset and MathFimer-7B model

**Description**: The authors construct NuminaMath-FIM by decomposing NuminaMath-CoT solutions into prefix-suffix pairs with missing intermediate steps, resulting in 2.5M training samples. They train MathFimer-7B on this dataset using Qwen2.5-Math-7B as the base model, creating a specialized model for step expansion that can be applied to enhance existing mathematical reasoning datasets.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Mathscale: Scaling instruction tuning for mathematical reasoning

**URL**: View paper

**Brief Assessment**

MathScale[32] focuses on generating diverse mathematical reasoning data through concept extraction and graph-based composition, not on fill-in-the-middle training for step expansion. The datasets and methodologies serve fundamentally different purposes.

### 2. Analysing mathematical reasoning abilities of neural models

**URL**: View paper

**Brief Assessment**

Analysing Neural Reasoning[35] focuses on evaluating mathematical reasoning abilities of neural models using a procedurally generated dataset with diverse question types, but does not address training datasets for step generation or fill-in-the-middle tasks for expanding reasoning steps.

### 3. AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset

**URL**: View paper

**Brief Assessment**

AIMO-2 Solution[34] focuses on creating a large-scale dataset of 540k math problems with 3.2m solutions and integrating code execution, not on fill-in-the-middle training for step expansion. The technical approaches are fundamentally different.

### 4. Lila: A unified benchmark for mathematical reasoning

**URL**: View paper

**Brief Assessment**

Lila[37] focuses on creating a unified benchmark for mathematical reasoning evaluation across diverse tasks, not on training step-expansion models. The paper constructs a benchmark by extending existing datasets with Python programs as solutions, rather than developing specialized models for generating intermediate reasoning steps through fill-in-the-middle tasks.

### 5. A survey of deep learning for mathematical reasoning

**URL**: View paper

**Brief Assessment**

Deep Learning Survey[31] is a comprehensive survey paper that reviews existing work in mathematical reasoning but does not present novel datasets or models. It discusses various pre-training corpora and methods but does not describe the specific FIM-based approach or the NuminaMath-FIM dataset construction methodology proposed in the original paper.

### 6. Ovm, outcome-supervised value models for planning in mathematical reasoning

**URL**: View paper

**Brief Assessment**

OVM[30] focuses on outcome-supervised value models for planning in mathematical reasoning, not on training datasets for step generation. The candidate does not address fill-in-the-middle task construction or step expansion datasets.

### 7. Processbench: Identifying process errors in mathematical reasoning

**URL**: View paper

**Brief Assessment**

ProcessBench[29] focuses on identifying process errors in mathematical reasoning through human expert annotation of erroneous steps, not on training datasets for step generation models or fill-in-the-middle tasks for expanding reasoning steps.

### 8. Measuring multimodal mathematical reasoning with math-vision dataset

**URL**: View paper

**Brief Assessment**

Math-Vision Dataset[38] focuses on evaluating multimodal mathematical reasoning through visual contexts from real math competitions, not on training datasets for step generation models or fill-in-the-middle tasks for reasoning step expansion.

### 9. Deepseekmath: Pushing the limits of mathematical reasoning in open language models

**URL**: View paper

**Brief Assessment**

DeepSeekMath[36] focuses on pre-training language models on large-scale mathematical web data (120B tokens from Common Crawl) and reinforcement learning techniques (GRPO), not on fill-in-the-middle training for step expansion. The candidate does not address the specific methodology of decomposing solutions into prefix-suffix pairs for training step-expansion models.

### 10. A survey on large language models for mathematical reasoning

**URL**: View paper

**Brief Assessment**

LLM Reasoning Survey[33] is a survey paper that reviews existing methods for mathematical reasoning in LLMs. It does not present a specific dataset construction method or model training approach comparable to NuminaMath-FIM/MathFimer-7B.

## Contribution 3: Empirical demonstration of consistent performance improvements

**Description**: The authors conduct comprehensive experiments showing that models trained on MathFimer-expanded data consistently outperform those trained on original data across various benchmarks including GSM8K and MATH. The improvements are observed across both general-purpose and math-specialized models, demonstrating the practical effectiveness and scalability of the approach.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. SBSC: Step-By-Step Coding for Improving Mathematical Olympiad Performance

**URL**: View paper

**Brief Assessment**

SBSC[24] focuses on step-by-step coding for mathematical problem-solving through program generation, not on step granularity effects in natural language reasoning chains. The candidate's multi-turn program generation approach is fundamentally different from the original paper's fill-in-the-middle framework for expanding natural language reasoning steps.

### 2. Masked Thought: Simply Masking Partial Reasoning Steps Can Improve Mathematical Reasoning Learning of Language Models

**URL**: View paper

**Brief Assessment**

Masked Thought[28] focuses on masking tokens within reasoning chains during training, while the original paper expands reasoning steps through fill-in-the-middle tasks. These are fundamentally different approaches to improving mathematical reasoning.

### 3. Coarse-to-fine process reward modeling for mathematical reasoning

**URL**: View paper

**Brief Assessment**

Coarse-to-fine Rewards[25] focuses on process reward modeling with step granularity through merging adjacent steps, not on expanding reasoning steps through fill-in-the-middle tasks as in the original paper. The candidate addresses redundancy reduction while the original addresses step expansion.

### 4. Evaluating mathematical reasoning beyond accuracy

**URL**: View paper

**Brief Assessment**

Beyond Accuracy[21] focuses on evaluating the quality of reasoning steps (validity and redundancy) rather than demonstrating performance improvements from step expansion methods. The candidate paper's evaluation methodology is orthogonal to the original paper's contribution about training on expanded data.

### 5. Cumulative reasoning with large language models

**URL**: View paper

**Brief Assessment**

Cumulative Reasoning[19] focuses on logical inference and mathematical problem-solving through a proposer-verifier-reporter framework with DAG construction, not on the effects of step granularity in training data. The candidate does not address training data expansion or step-level granularity effects that are central to the original paper's contribution.

### 6. Mathprompter: Mathematical reasoning using large language models

**URL**: View paper

**Brief Assessment**

MathPrompter[20] focuses on zero-shot prompting techniques for mathematical reasoning and reports accuracy improvements on MultiArith dataset (78.7% → 92.5%). The original paper's contribution concerns step granularity effects through fill-in-the-middle training on multiple datasets (GSM8K, MATH, etc.), which is a fundamentally different approach from MathPrompter's prompting-based verification method.

### 7. We-math: Does your large multimodal model achieve human-like mathematical reasoning?

**URL**: View paper

**Brief Assessment**

We-math[27] focuses on evaluating visual mathematical reasoning through step-wise problem decomposition and knowledge concepts, not on training data expansion or step granularity effects that would refute the original paper's claims about MathFimer's performance improvements.

### 8. Knowledge-centered dual-process reasoning for math word problems with large language models

**URL**: View paper

**Brief Assessment**

Dual-process Reasoning[26] focuses on knowledge-centered reasoning for math word problems using dual-process theory, not on step granularity effects. The candidate evaluates performance improvements from knowledge invocation/verification/injection mechanisms, whereas the original paper examines improvements from step expansion via fill-in-the-middle tasks.

### 9. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models

**URL**: View paper

**Brief Assessment**

ConceptMath[22] focuses on concept-wise evaluation of mathematical reasoning across different granularities, not on step granularity effects or step expansion methods like MathFimer. The candidate does not address step-by-step solution quality or reasoning step expansion.

### 10. Step-kto: Optimizing mathematical reasoning through stepwise binary feedback

**URL**: View paper

**Brief Assessment**

Step-KTO[23] focuses on optimizing mathematical reasoning through stepwise binary feedback combining process-level and outcome-level signals, while the original paper demonstrates improvements from expanding reasoning steps via fill-in-the-middle task. These are complementary approaches to improving mathematical reasoning rather than competing claims about step granularity effects.

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References

- [0] MathFimer: Enhancing Mathematical Reasoning by Expanding Reasoning Steps through Fill-in-the-Middle Task View paper
- [1] System-2 mathematical reasoning via enriched instruction tuning View paper
- [2] Search-Based Correction of Reasoning Chains for Language Models View paper
- [3] Make your llm fully utilize the context View paper
- [4] From Informal to Formal--Incorporating and Evaluating LLMs on Natural Language Requirements to Verifiable Formal Proofs View paper
- [5] Efficient tool use with chain-of-abstraction reasoning View paper
- [6] ClozeMath: Improving Mathematical Reasoning in Language Models by Learning to Fill Equations View paper
- [7] Mind the Gap: Bridging Thought Leap for Improved Chain-of-Thought Tuning View paper
- [8] Solving math word problems with reexamination View paper
- [9] Exact byte-level probabilities from tokenized language models for fim-tasks and model ensembles View paper
- [10] Seeing the continuity behind â□□double discontinuityâ□□: Investigating Hong Kong prospective mathematics teachers' secondaryâ□□tertiary transition View paper
- [11] Mathematical Reasoning via Self-supervised Skip-tree Training View paper
- [12] No Compute Left Behind: Rethinking Reasoning and Sampling with Masked Diffusion Models View paper
- [13] Planning-Aware Code Infilling via Horizon-Length Prediction View paper
- [14] Horizon-Length Prediction: Advancing Fill-in-the-Middle Capabilities for Code Generation with Lookahead Planning View paper
- [15] Scaling Diffusion Language Models via Adaptation from Autoregressive Models View paper
- [16] Open Middle Math View paper
- [17] Fill in the Blank: Exploring and Enhancing LLM Capabilities for Backward Reasoning in Math Word Problems View paper
- [18] The Science of Proof: Mathematical Reasoning and Its Limitations View paper
- [19] Cumulative reasoning with large language models View paper
- [20] Mathprompter: Mathematical reasoning using large language models View paper
- [21] Evaluating mathematical reasoning beyond accuracy View paper
- [22] Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models View paper
- [23] Step-kto: Optimizing mathematical reasoning through stepwise binary feedback View paper
- [24] SBSC: Step-By-Step Coding for Improving Mathematical Olympiad Performance View paper
- [25] Coarse-to-fine process reward modeling for mathematical reasoning View paper
- [26] Knowledge-centered dual-process reasoning for math word problems with large language models View paper
- [27] We-math: Does your large multimodal model achieve human-like mathematical reasoning? View paper
- [28] Masked Thought: Simply Masking Partial Reasoning Steps Can Improve Mathematical Reasoning Learning of Language Models View paper
- [29] Processbench: Identifying process errors in mathematical reasoning View paper
- [30] Ovm, outcome-supervised value models for planning in mathematical reasoning View paper
- [31] A survey of deep learning for mathematical reasoning View paper
- [32] Mathscale: Scaling instruction tuning for mathematical reasoning View paper
- [33] A survey on large language models for mathematical reasoning View paper
- [34] AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset View paper
- [35] Analysing mathematical reasoning abilities of neural models View paper
- [36] Deepseekmath: Pushing the limits of mathematical reasoning in open language models View paper
- [37] Lila: A unified benchmark for mathematical reasoning View paper
- [38] Measuring multimodal mathematical reasoning with math-vision dataset View paper
- [39] Mathematics and Plausible Reasoning: Logic, Symbolic and mathematical View paper
- [40] Codegemma: Open code models based on gemma View paper
- [41] Constrained Decoding for Fill-in-the-Middle Code Language Models via Efficient Left and Right Quotienting of Context-Sensitive Grammars View paper
- [42] CombiBench: Benchmarking LLM Capability for Combinatorial Mathematics View paper
- [43] Beyond the last answer: Your reasoning trace uncovers more than you think View paper