Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (3)
https://doi.org/10.6025/jdp/2026/16/3/161-184

A Statistically Validated Large Scale Mathematical Chain of Thought Dataset for Symbolic Reasoning and Large Language Model Benchmarking

Hathairat Ketmaneechairat
Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
hathairat.k@cit.kmutnb.ac.th

ABSTRACT

Mathematical reasoning poses a significant challenge for Large Language Models (LLMs) due to the necessity for structured symbolic manipulation, multi step deduction, and logically consistent inference. While Chain-of Thought (CoT) prompting has substantially advanced reasoning capabilities, existing datasets often lack transparent, verifiable reasoning trajectories and rigorous statistical validation. This study presents a statistically validated, large scale mathematical CoT dataset comprising approximately 100,000 samples spanning six core mathematical domains (Algebra, Geometry, Trigonometry, Matrices, Derivatives, and Integrals) and three predefined difficulty levels. The dataset preserves complete intermediate reasoning traces and final solutions in standardized LaTeX notation. We conducted comprehensive structural, descriptive, and inferential statistical analyses including Chi-square tests, Spearman's rank correlation, two-way ANOVA, and Tukey's HSD post hoc comparisons to validate the dataset's internal consistency. Results demonstrate a perfectly balanced distribution across topics and difficulties, with the mathematical topic emerging as the dominant determinant of reasoning complexity (partial $n^2 = 0.995$). Furthermore, the corpus exhibits rich symbolic diversity, authentic mathematical syntax, and structured command co-occurrence networks. This statistically validated corpus provides a reliable and robust benchmark for supervised fine tuning, curriculum learning, process supervision, explainable AI, and evaluating next-generation mathematical LLMs.

Keywords: Mathematical Reasoning, Chain of Thought (CoT), Large Language Models (LLMs), Symbolic Reasoning, Statistical Validation, Dataset Benchmarking, LaTeX Representation, Process Supervision, Curriculum Learning

Received: 2 February 2026, Revised 29 April 2026, Accepted 30 May 2026

1. Introduction

Mathematical reasoning is among the most challenging capabilities for Large Language Models (LLMs) because it requires structured symbolic manipulation, multi-step deduction, variable binding, and logically consistent inference, rather than fluent language generation alone [1, 2, 3]. Unlike conventional natural language generation tasks, where coherence is often sufficient, mathematical problem solving demands globally consistent reasoning across intermediate steps, such that a single incorrect inference may invalidate the final solution [4]. Consequently, mathematical reasoning has emerged as a rigorous benchmark for evaluating the reasoning capabilities of modern LLMs.

2. Related Work

Recent advances in LLMs have substantially improved mathematical reasoning performance through Chain-of Thought (CoT) prompting, which enables models to generate intermediate reasoning steps before producing final answers [5, 6]. This paradigm underpins state of the art reasoning systems such as Deep Seek R1 [7] and Open AI's o1 [8]. Nevertheless, despite impressive performance gains, the reasoning process itself often remains unreliable, with intermediate reasoning traces containing logical inconsistencies or computational errors that ultimately compromise prediction quality [9, 10, 11, 12, 13].

The need for high quality reasoning datasets has therefore become increasingly important. Existing mathematical and reasoning benchmarks frequently emphasize final answers while providing limited support for transparent, verifiable, and diverse reasoning trajectories [14]. Moreover, current approaches often rely on a single reasoning paradigm, restricting their generalization across heterogeneous mathematical tasks. Recent studies have highlighted the importance of interpretable symbolic reasoning for scientific discovery and structured inference, yet the automatic generation of context aligned symbolic reasoning remains underexplored [15].

Parallel developments in Text to SQL and reasoning benchmarks further illustrate these limitations. Although modern systems such as CHASE SQL, XiYan SQL, Reasoning SQL 14B, and OpenSearch SQL have achieved remarkable execution accuracy on benchmark datasets [16, 17, 18], existing benchmarks primarily evaluate SQL generation rather than complex mathematical reasoning required in real world domains such as finance, science, and engineering [19, 20, 21, 22, 23]. LogicCat addresses part of this gap by introducing complex reasoning scenarios for Text to SQL [24] but broader mathematical Chain of Thought datasets remain limited.

Several recent efforts have sought to improve the quality of reasoning data. Code Assisted Chain of Thought automates the synthesis of verifiable reasoning data through code driven augmentation [25]. ProverGen combines LLM generation with symbolic theorem provers to construct scalable first order logic datasets [26]. Chain of Reasoning integrates natural language, algorithmic, and symbolic reasoning into a unified framework [27], while Circuit based Reasoning Verification provides a white box approach to identifying reasoning failures through computational graph analysis [28]. Collectively, these studies demonstrate a growing interest in improving the quality of reasoning and highlight the need for statistically validated mathematical reasoning corpora that comprehensively capture symbolic diversity, reasoning complexity, and authentic mathematical syntax. [29].

Motivated by these challenges, this work presents a statistically validated large scale mathematical Chain-of-Thought dataset designed for symbolic reasoning and large language model benchmarking. The dataset provides balanced coverage across multiple mathematical domains and difficulty levels while preserving rich LaTeX representations, diverse symbolic structures, and detailed reasoning traces. Comprehensive descriptive, structural, network based, and inferential statistical analyses including correlation analysis, chi square testing, two way ANOVA, Tukey HSD post hoc comparisons, and effect size estimation are performed to validate the dataset's internal consistency and reasoning characteristics. The resulting corpus provides a reliable benchmark for supervised fine tuning, curriculum learning, process supervision, explainable artificial intelligence, and the evaluation of next generation mathematical reasoning models.

To address these limitations, a systematic framework was developed for generating, validating, and characterizing a large scale mathematical Chain of Thought dataset. The following section describes the overall methodology, including procedural generation, symbolic representation, feature extraction, and the statistical validation pipeline adopted in this study.

3. Materials and Methods

3.1 Research Framework

This study proposes a comprehensive framework for constructing, validating, and characterizing a large-scale mathematical Chain of Thought (CoT) dataset intended for training and benchmarking large language models (LLMs). The framework integrates automated mathematical problem generation, symbolic reasoning synthesis, quality validation, feature extraction, statistical analysis, and dataset characterization into a unified pipeline. Unlike conventional mathematical datasets that primarily focus on final answers, the proposed framework preserves complete intermediate reasoning traces expressed in structured LaTeX notation, thereby enabling transparent supervision of reasoning and explainable artificial intelligence.

The overall workflow consists of six sequential stages: (1) mathematical problem generation, (2) Chain-of-Thought reasoning generation, (3) symbolic representation and annotation, (4) feature extraction, (5) statistical validation, and (6) dataset characterization. Each stage helps ensure that the generated corpus is balanced, internally consistent, statistically validated, and suitable for developing next generation mathematical reasoning systems.

Building upon the overall research framework, the proposed computational architecture operationalizes each stage of dataset construction, beginning with mathematical problem generation and concluding with statistical validation and dataset characterization.

3.2 Proposed Framework

The proposed framework provides an integrated pipeline for constructing and validating a large scale mathematical Chain of Thought (CoT) dataset for large language model (LLM) training and benchmarking. The workflow begins with the procedural generation of mathematical problems from six core domains Algebra, Geometry, Trigonometry, Matrices, Derivatives, and Integrals across three predefined difficulty levels. Each problem is accompanied by a complete Chain of Thought reasoning sequence and a final solution represented in standardized LaTeX notation, ensuring logical consistency, symbolic accuracy, and machine readable mathematical expressions.

Following dataset generation, each sample is enriched with structural annotations, including mathematical topic, difficulty level, question text, reasoning trace, final answer, and symbolic metadata. Quantitative features such as question length, Chain of Thought length, equation density, LaTeX complexity, operator frequency, and symbol diversity are subsequently extracted to characterize the linguistic and symbolic complexity of the corpus.

The extracted features are then subjected to comprehensive statistical validation using descriptive statistics, correlation analysis, Chi-square tests, two way ANOVA, Tukey HSD post hoc comparisons, and effect size estimation to verify dataset balance and consistency of reasoning. Finally, the validated dataset is characterized through distribution analysis, complexity profiling, and network based visualization of mathematical symbols and LaTeX commands. This integrated framework ensures the development of a statistically validated, structurally balanced, and symbolically rich mathematical reasoning corpus suitable for supervised fine-tuning, curriculum learning, explainable artificial intelligence, and benchmarking of next generation mathematical language models.

After defining the overall framework, the characteristics of the resulting dataset are described to provide context for the subsequent feature extraction and statistical analyses.

3.3 Dataset Description

The proposed dataset contains approximately 100,000 mathematically generated Chain of Thought samples spanning six mathematical domains and three difficulty levels. Each record contains a mathematical problem, complete reasoning sequence, final solution, and symbolic annotations encoded in LaTeX format. The balanced distribution of topics and difficulty levels minimizes sampling bias while supporting curriculum based learning and unbiased benchmarking.

Although the generated dataset contains complete mathematical reasoning traces, additional quantitative features are required to objectively measure linguistic complexity, symbolic richness, and reasoning depth. These engineered features form the basis for the subsequent statistical analyses.

3.4 Feature Engineering

Several quantitative metrics were computed to characterize reasoning complexity.

Textual Features

- Question length
- Chain of Thought length
- Final answer length

Symbolic Features

- LaTeX command frequency
- Mathematical operator count
- Equation count
- Symbol diversity

Complexity Features

- Equation density
- LaTeX complexity score
- Average reasoning depth
- Intermediate reasoning steps

These features collectively capture linguistic, symbolic, and computational complexity.

Once the textual and symbolic characteristics had been extracted, inferential statistical analyses were conducted to assess whether the generated corpus exhibits meaningful structural relationships and statistically significant differences across mathematical domains and difficulty levels.

3.5 Statistical Analysis

Inferential analyses were subsequently performed to validate structural consistency.

Inferential Analysis

Chi-square Test

Used to evaluate the association between mathematical topics and difficulty levels.

Spearman Correlation
Used to quantify monotonic relationships between reasoning length and problem difficulty.

Two-way ANOVA
Used to examine the effects of

- mathematical topic; and

- difficulty level on Chain of Thought length.

Tukey HSD
Performed pairwise comparisons among mathematical domains following significant ANOVA results.

Effect Size
Partial n^2 was calculated to estimate the magnitude of observed effects.

3.6 Dataset Validation
The quality and reliability of the proposed mathematical Chain of Thought dataset were evaluated through a comprehensive validation framework that examined its structural characteristics, symbolic representation, statistical consistency, and reasoning quality. Structural validation focused on assessing the balance of mathematical topics and difficulty levels to ensure that the dataset was free from substantial sampling bias. A nearly uniform distribution across the six mathematical domains and three difficulty categories indicates that the procedural generation framework successfully produced a well balanced corpus suitable for unbiased training and benchmarking of large language models. Such balanced representation reduces the likelihood of domain-specific overfitting and improves the generalizability of models across diverse mathematical reasoning tasks.

Symbolic validation assessed the richness and authenticity of the mathematical notation in the dataset. This evaluation included analyses of mathematical operator frequencies, LaTeX command usage, equation density, and symbol diversity. The results demonstrate that the generated reasoning traces preserve standard mathematical syntax and exhibit a wide variety of symbolic expressions commonly encountered in mathematical literature. The structured use of LaTeX commands and mathematical operators ensures syntactic consistency while providing sufficient symbolic diversity to support robust tokenizer training, representation learning, and symbolic reasoning in transformer based language models.

Statistical validation was conducted to determine whether the generated dataset exhibited meaningful structural relationships rather than random variation. Descriptive statistical analyses summarized the distributions of textual and symbolic features, whereas inferential statistical methods, including Chi square analysis, Spearman correlation, two-way ANOVA, Tukey HSD multiple comparisons, and effect size estimation, were used to evaluate the relationships among mathematical topics, difficulty levels, and reasoning complexity. The significant results obtained from these analyses confirm that the procedural generation strategy successfully encodes increasing reasoning complexity with problem difficulty while preserving statistically distinguishable characteristics among mathematical domains. Collectively, these findings provide strong empirical evidence supporting the internal consistency and validity of the proposed dataset.

Reasoning validation focused on assessing the quality and coherence of the generated Chain of Thought sequences. This evaluation examined the logical progression of intermediate reasoning steps, the consistency of symbolic manipulations, and the relationship between reasoning depth and assigned difficulty levels. The analyses indicate that more challenging problems consistently require longer and more elaborate reasoning sequences, demonstrating that the generated Chain of Thought traces accurately reflect progressive cognitive complexity. Furthermore, the clear separation between problem statements, reasoning processes, and final answers enables transparent supervision of intermediate reasoning, thereby supporting explainable artificial intelligence, process supervised learning, reinforcement learning from reasoning traces, and curriculum-based optimization strategies for advanced mathematical language models.

3.7 Computational Workflow

The computational workflow adopted in this study follows a systematic pipeline that integrates mathematical problem generation, symbolic reasoning synthesis, feature extraction, statistical validation, and dataset characterization into a unified framework. The process begins with creating mathematical problem templates that represent six major mathematical domains. These templates serve as the foundation for the procedural generation engine, which automatically produces diverse mathematical problems with balanced distributions across predefined difficulty levels.

Each generated problem is subsequently processed through a Chain-of-Thought reasoning module that constructs complete intermediate reasoning sequences leading to the final mathematical solution. Both the reasoning process and the final answer are encoded using standardized LaTeX notation to preserve symbolic consistency and facilitate machine readable mathematical representation. This standardized symbolic encoding ensures compatibility with modern mathematical language models while maintaining the structural integrity of mathematical expressions.

Following dataset generation, a comprehensive feature extraction stage computes multiple textual, symbolic, and complexity related characteristics for each sample. These include question length, Chain of Thought length, final answer length, equation density, mathematical operator frequency, symbol diversity, and LaTeX complexity. Together, these features provide quantitative measures of reasoning depth, symbolic richness, and computational complexity that enable detailed characterization of the dataset.

The extracted features are subsequently subjected to rigorous statistical validation. Descriptive analyses summarize the overall properties of the corpus, while inferential statistical techniques including Chi-square tests, Spearman correlation analysis, two way analysis of variance (ANOVA), Tukey HSD post-hoc comparisons, and effect size estimation evaluate the relationships among mathematical domains, difficulty levels, and reasoning characteristics. These statistical procedures verify the dataset's internal consistency and confirm that the generated reasoning traces exhibit meaningful structural variation rather than random fluctuations.

Finally, the validated dataset is characterized using a range of visualization and network-analysis techniques that examine topic distributions, reasoning complexity, symbolic diversity, equation density, and LaTeX command co-occurrence. The resulting computational workflow produces a statistically validated and structurally rich corpus of mathematical reasoning that is well suited for supervised fine tuning, curriculum learning, process supervision, explainable artificial intelligence, symbolic reasoning research, and benchmarking next generation large language models.

4. Analysis

he effectiveness of the proposed framework was evaluated through a comprehensive analysis of the generated dataset. The following sections examine the structural, linguistic, symbolic, and statistical characteristics of the corpus using descriptive visualizations and inferential statistical techniques

4.1 Topic Balance

Figure 1 illustrates the distribution of samples across the six mathematical domains: Algebra, Geometry, Trigonometry, Matrices, Derivatives, and Integrals.

The figure demonstrates an almost perfectly balanced dataset, with each mathematical topic contributing approximately one sixth of the total corpus. This balanced distribution minimizes sampling bias during model training and evaluation, ensuring that no single mathematical domain dominates the learning process. Such uniformity is particularly important for developing general purpose mathematical reasoning models that handle diverse problem categories equally well.

The nearly uniform distribution across the six mathematical domains demonstrates that the procedural generation framework successfully eliminates domain specific sampling bias. Balanced topic representation

ensures that downstream transformer-based models receive equivalent exposure to diverse reasoning paradigms, thereby reducing preferential learning toward frequently occurring mathematical concepts. This characteristic enhances model generalization and supports fair benchmarking of mathematical reasoning algorithms across heterogeneous problem domains. From a statistical learning perspective, balanced datasets improve the stability of parameter estimates and reduce class-dependent prediction bias during supervised training.

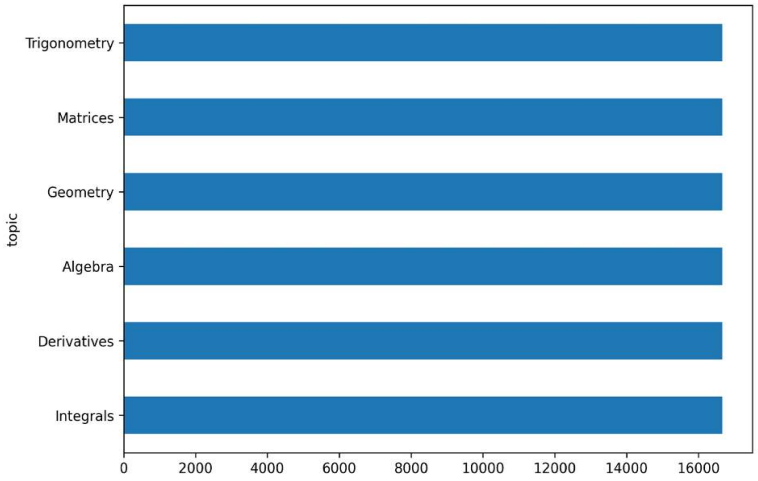

Figure 1. Topic Balance

4.2 Difficulty Distribution

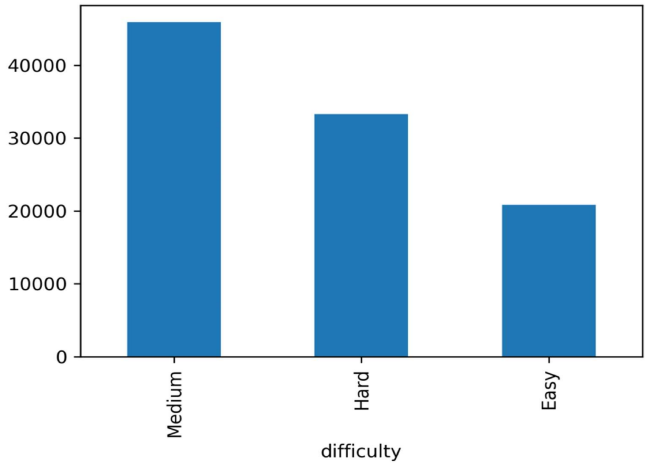

Figure 2. Difficulty Distribution

The dataset exhibits a well balanced distribution of problem difficulty levels. This balanced representation enables machine learning models to learn progressively from simple to complex reasoning tasks while preventing overfitting toward any particular complexity level. The inclusion of substantial numbers of difficult problems also supports evaluation of long form mathematical reasoning capabilities.

The balanced allocation of Easy, Medium, and Hard problems establishes a progressive learning curriculum that closely aligns with curriculum learning theory. Exposure to multiple reasoning complexities enables neural language models to gradually acquire increasingly sophisticated symbolic reasoning capabilities while maintaining stable optimization dynamics. Furthermore, the balanced complexity distribution provides an

appropriate benchmark for evaluating the scalability and robustness of reasoning under varying cognitive loads.

Having established the balanced distribution of mathematical topics and difficulty levels, the next analysis examines the textual characteristics of the generated problems to determine whether linguistic complexity is equally well controlled.

4.3 Question Text Length Distribution

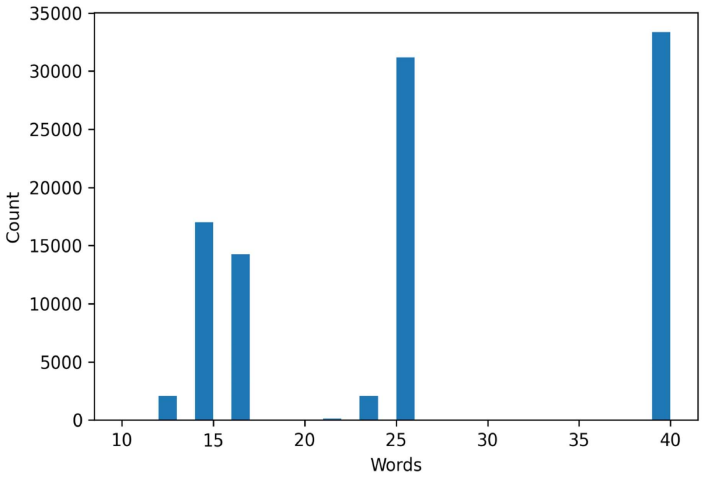

Figure 3. Question Length Distribution

Most mathematical questions are concentrated within a moderate length range, while relatively few extremely short or exceptionally long questions are observed. The right skewed distribution indicates that although most problems have comparable textual complexity, some advanced questions require more elaborate mathe matical descriptions. This variability improves linguistic diversity and better reflects real world mathematical-problem statements.

Question length serves as an indirect indicator of linguistic complexity and contextual information density. The observed distribution suggests that the procedural generation framework produces sufficiently diverse textual formulations without introducing excessive verbosity. Such variation enhances tokenizer robustness and enables large language models to learn mathematical reasoning independently of superficial sentence length, thereby improving generalization to naturally occurring mathematical problems.

Beyond textual properties, symbolic complexity provides an additional perspective on mathematical reasoning. Consequently, the following analysis evaluates the distribution of LaTeX complexity across the corpus.

4.4 LaTeX Complexity Distribution

The figure indicates that most problems contain moderate LaTeX complexity, while highly complex expressions occur less frequently. This reflects the procedural generation methodology, where mathematical notation remains readable while incorporating sufficient symbolic diversity. The gradual decline toward higher complexity suggests that the corpus includes increasingly sophisticated symbolic structures without overwhelming the overall dataset.

The moderate distribution of LaTeX complexity indicates that symbolic representations remain expressive while preserving syntactic consistency. From the perspective of mathematical language modeling, balanced symbolic complexity prevents excessive tokenizer fragmentation and reduces representation sparsity. Consequently, transformer architectures can more effectively learn semantic relationships among mathematical operators, variables, and expressions while maintaining computational efficiency.

While symbolic complexity reflects mathematical representation, reasoning complexity is better characterized through the length and structure of Chain of Thought explanations.

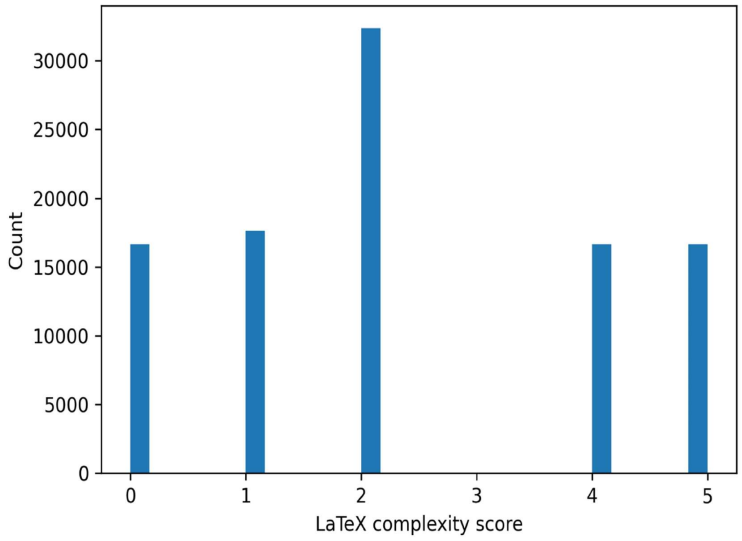

Figure 4. LaTeX Complexity Distribution

4.5 Chain-of-Thought Length Distribution

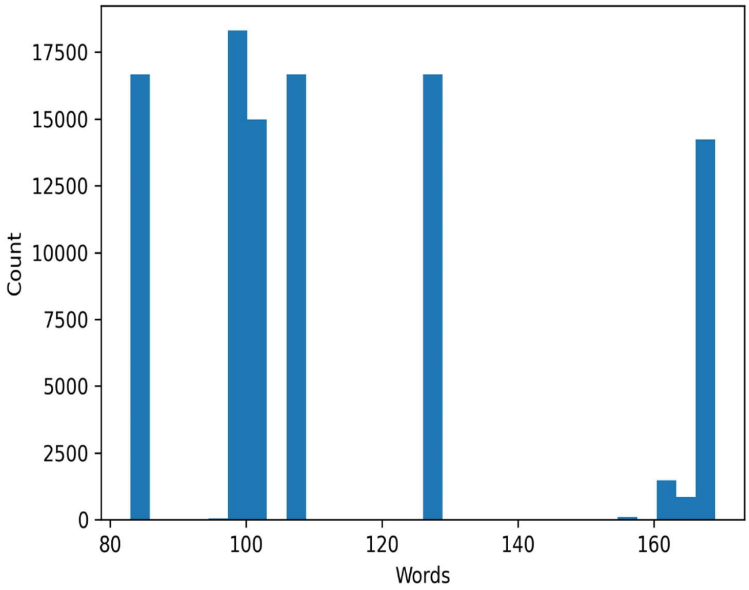

Figure 5. Chain-of-Thought Length Distribution

Most reasoning sequences cluster around medium length explanations, indicating that most problems require several logical inference steps rather than trivial or excessively long derivations. The long right tail corresponds to mathematically challenging problems requiring multiple intermediate derivations, demonstrating the dataset's ability to support research on long context reasoning.

The variability in reasoning length reflects differences in cognitive complexity among mathematical tasks. Longer reasoning sequences correspond to problems requiring multiple intermediate inference steps, symbolic transformations, and logical dependencies. This characteristic makes the dataset particularly suitable for

process supervised learning, reinforcement learning from reasoning traces, and Chain-of-Thought fine-tuning, where intermediate reasoning quality is as important as final prediction accuracy.

Following the analysis of intermediate reasoning, the characteristics of the final mathematical solutions are examined to determine whether reasoning depth is reflected in output complexity

4.6 Final Answer Length Distribution

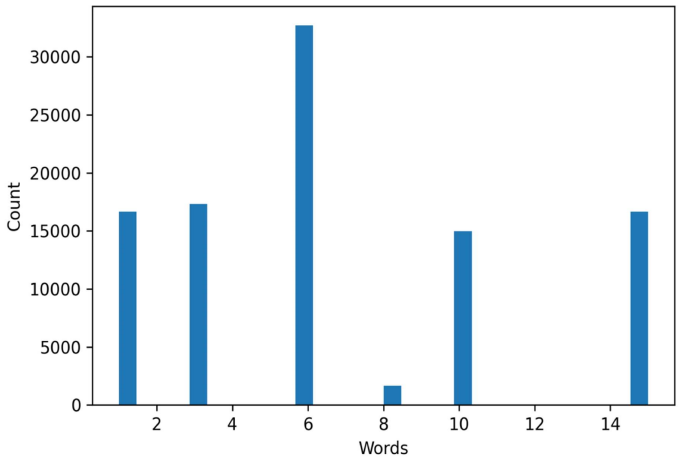

Figure 6. Final Answer Length Distribution

Final answers remain relatively short compared to reasoning sequences, reflecting the intended separation between reasoning and solution. Most outputs consist of concise mathematical expressions, making the dataset well suited for supervised fine tuning where reasoning and final prediction are independently modeled.

The relatively compact distribution of final answers demonstrates a clear separation between reasoning processes and solution outputs. This architectural separation facilitates multi-task learning frameworks that independently optimize reasoning generation and answer prediction. Such modular supervision has recently become essential to improving the transparency, interpretability, and verification of reasoning in mathematical large language models.

Comparing questions, reasoning traces, and final answers provides a more comprehensive understanding of how complexity is distributed throughout the reasoning pipeline.

4.7 Text Length Comparison Boxplot

The boxplot clearly shows that Chain-of-Thought explanations exhibit substantially greater median length and variability than either questions or final answers. This finding confirms that most of the dataset's complexity lies in the reasoning process rather than in the problem statement or solution itself. The broader interquartile range for CoT reflects diverse depths of reasoning across mathematical domains.

The substantial differences in variability among questions, reasoning traces, and final answers confirm that computational complexity lies primarily in intermediate reasoning rather than in problem statements or solutions. This observation supports the hypothesis that difficulty in mathematical reasoning is governed more by inference depth than by the complexity of the textual input. Consequently, model performance depends heavily on its capacity to generate coherent intermediate reasoning rather than simply producing correct final answers.

Having examined textual characteristics, the analysis now shifts toward symbolic properties by investigating the frequency of mathematical notation represented through LaTeX commands.

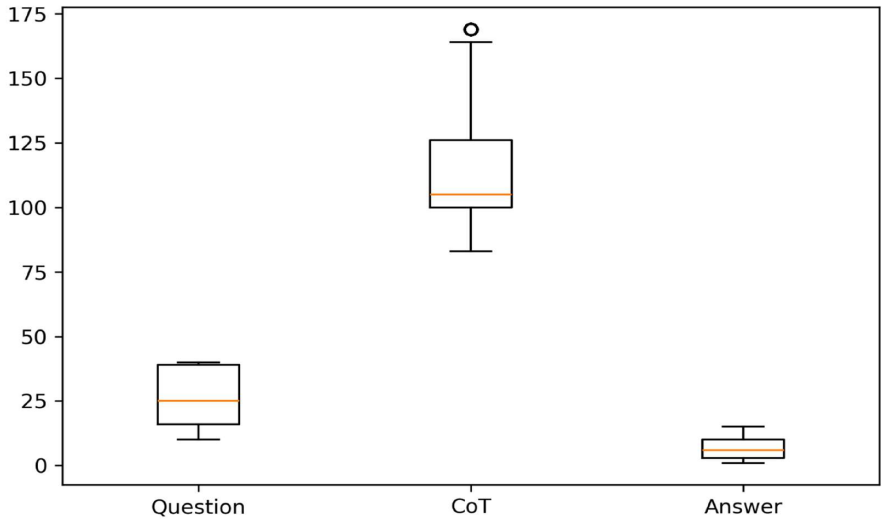

Figure 7. Text Length Comparison Boxplot

4.8 Top LaTeX Command Frequency

Frequently used commands correspond to standard mathematical notation, including fractions, square roots, summations, matrices, and delimiters. Their dominance reflects the corpus's mathematical nature and confirms consistent formatting throughout the dataset. The command distribution also indicates compatibility with existing LaTeX parsers and mathematical language models.

The predominance of fundamental LaTeX commands indicates that the dataset preserves authentic mathematical notation commonly encountered in scientific literature and educational resources. Frequent exposure to standardized symbolic syntax enables mathematical language models to develop robust embeddings for mathematical operators, improving symbolic parsing, equation generation, and semantic understanding across multiple mathematical domains

Although command frequency reveals overall symbolic usage, distributional variability is more effectively visualized through density based representations.

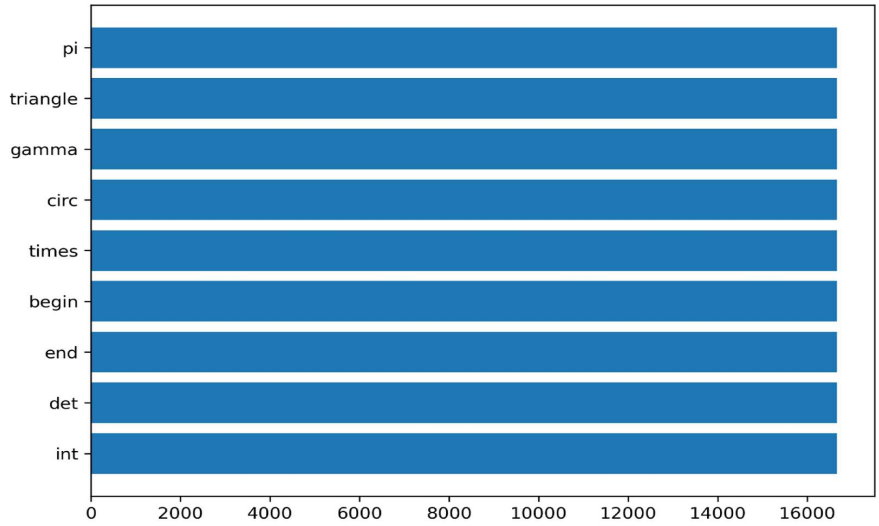

Figure 8. Top LaTeX Command Frequency

4.9 Violin Plot–Text Length Distribution

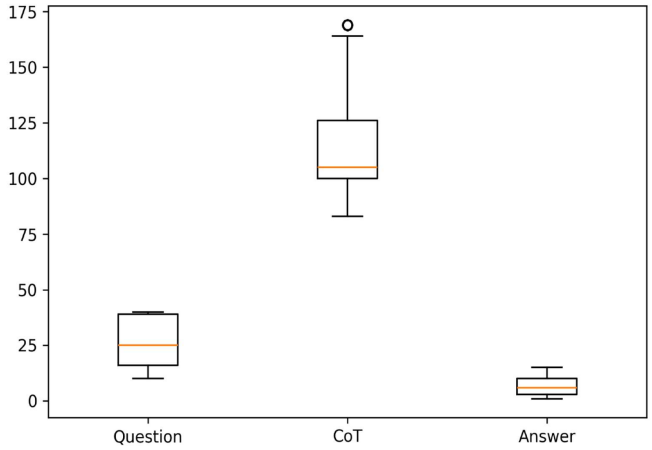

Figure 9. Violin Plot Text Length Distribution

The violin plots reveal that Chain of Thought texts possess the widest distribution, indicating substantial variability in reasoning complexity. Question lengths show moderate variability, whereas final answers remain tightly clustered around shorter lengths. The density profiles further demonstrate that reasoning complexity varies considerably depending on mathematical topic and problem difficulty.

The density distributions reveal substantial heterogeneity in reasoning complexity while maintaining relatively stable question and answer lengths. This pattern suggests that the procedural generator systematically controls problem formulation while allowing the complexity of reasoning to vary naturally with the mathematical operations. Such variability is essential for training adaptive reasoning models capable of dynamically allocating computational resources according to problem complexity.

After characterizing overall reasoning distributions, the analysis explores how reasoning complexity varies across individual mathematical domains.

4.10 Average Chain-of-Thought Length by Topic

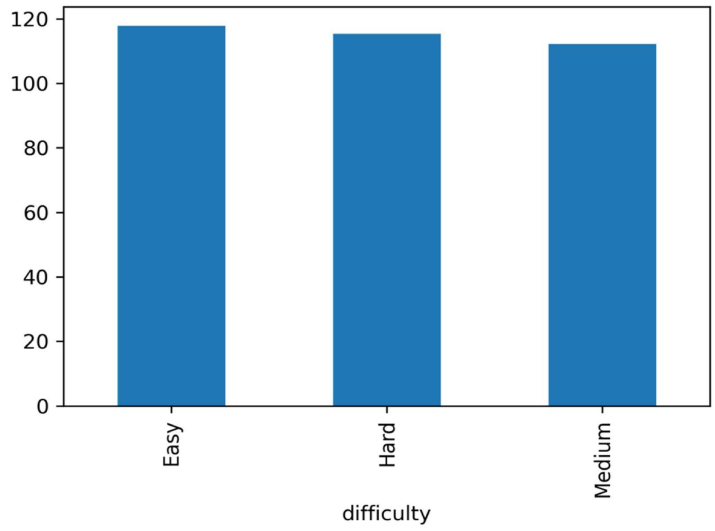

Figure 10. Average Chain-of-Thought Length by Topic

Reasoning length varies across topics, reflecting inherent differences in mathematical solution procedures. Topics requiring multi step symbolic manipulation, such as matrix algebra and calculus, generally produce longer reasoning chains than geometry or algebra. This variation confirms that the procedural generation framework successfully captures topic specific reasoning characteristics instead of producing uniformly structured solutions.

Differences in reasoning length among mathematical domains indicate that each topic possesses unique computational characteristics and logical structures. Calculus and matrix operations inherently require sequential symbolic manipulations and multiple algebraic transformations, whereas algebra and geometry often involve comparatively shorter derivations. These domain specific reasoning profiles provide valuable supervision for domain-adaptive mathematical reasoning architectures.

Since reasoning characteristics vary across mathematical topics, it is equally important to determine how reasoning depth changes as problem difficulty increases.

4.11 Average Chain-of-Thought Length by Difficulty

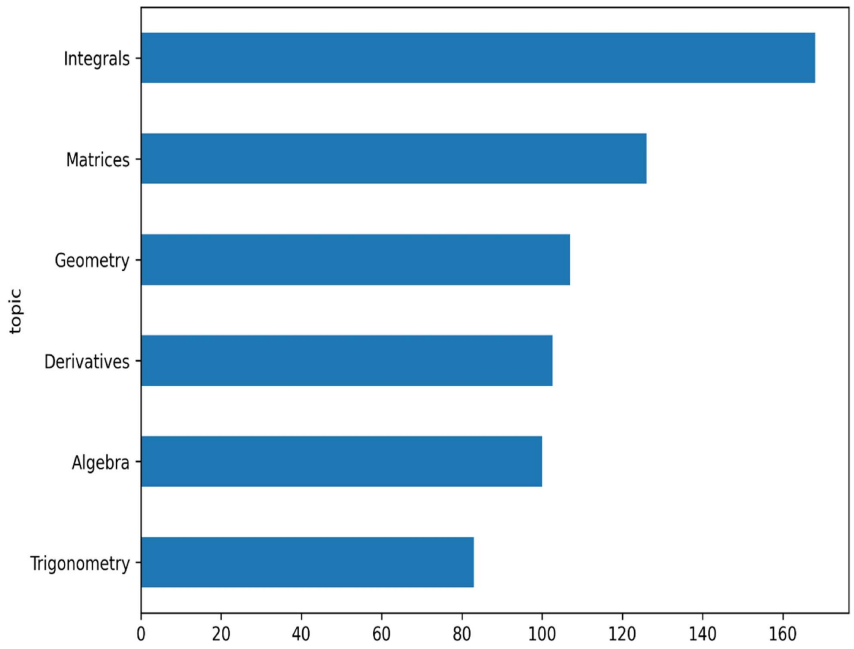

Figure 11. Average Chain-of-Thought Length by Difficulty

A clear increasing trend is observed from Easy to Hard problems, demonstrating that problem difficulty is strongly associated with reasoning depth. Hard problems require more intermediate computational and logical steps, validating the dataset's difficulty annotation methodology. This relationship provides useful supervision signals for curriculum learning and complexity aware reasoning models.

The monotonic increase in reasoning length with problem difficulty validates the procedural generation strategy used to assign difficulty labels. Rather than relying solely on numerical coefficients, the dataset reflects increasing logical depth, symbolic complexity, and inference dependency. This correspondence between difficulty and reasoning depth supports curriculum based learning strategies and enables reliable evaluation of the scalability of reasoning in advanced language models.

The combined influence of mathematical domain and difficulty is further examined using a two-dimensional heatmap that captures their interaction.

4.12 Topic × Difficulty Heatmap (Average CoT Length)

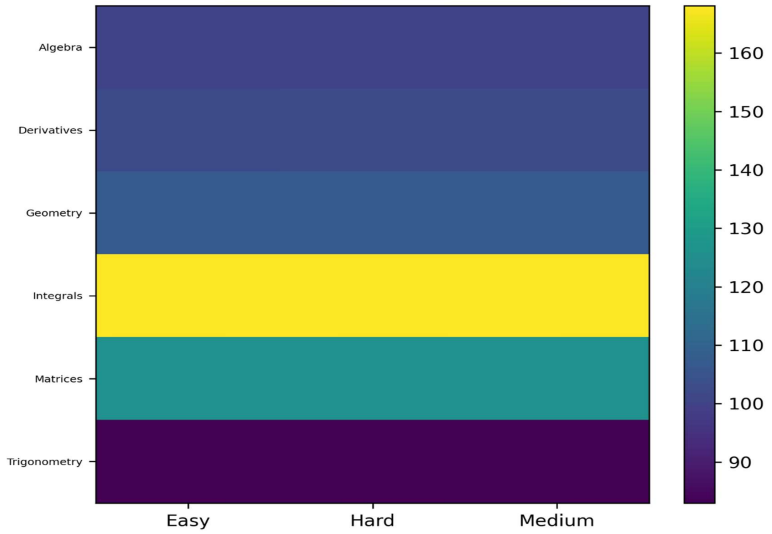

Figure 12. Topic × Difficulty Heatmap (Average CoT Length)

The heatmap demonstrates that reasoning depth depends jointly on mathematical domain and difficulty. Hard versions consistently require more reasoning than Easy versions across all topics, although the magnitude of the increase varies by domain. This interaction highlights that mathematical complexity is multidimensional rather than determined solely by topic or difficulty independently.

The interaction between mathematical domain and difficulty demonstrates that reasoning complexity is multidimensional. Hard problems consistently require longer inference chains regardless of topic, although the magnitude of increase varies across domains. This interaction suggests that mathematical reasoning cannot be adequately characterized using a single complexity metric but instead depends upon both conceptual diversity and computational depth.

Beyond reasoning complexity, symbolic computation can also be characterized by the diversity of mathematical operators employed throughout the corpus.

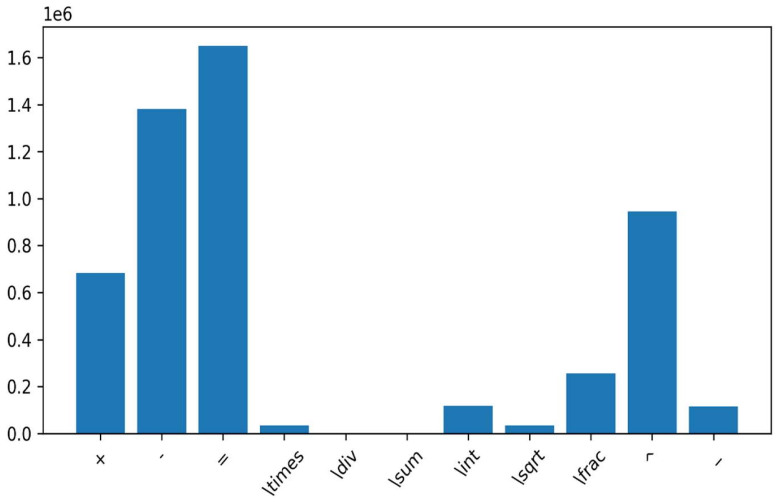

Figure 13. Mathematical Operator Frequency

Arithmetic operators occur most frequently, followed by algebraic, relational, and calculus specific symbols. The diverse operator usage demonstrates that the corpus encompasses a wide spectrum of mathematical operations rather than concentrating on elementary arithmetic. Such diversity is advantageous for training symbolic reasoning systems capable of handling heterogeneous mathematical expressions.

The broad distribution of mathematical operators reflects extensive symbolic diversity throughout the corpus. Rich operator coverage expands vocabulary representations learned by mathematical language models and improves their ability to generalize across unseen symbolic expressions. Furthermore, balanced operator use reduces the overrepresentation of elementary arithmetic, thereby promoting robust learning of advanced mathematical syntax and symbolic reasoning.

Operator usage alone does not fully describe symbolic complexity; therefore, the density of mathematical equations is analyzed in the following section.

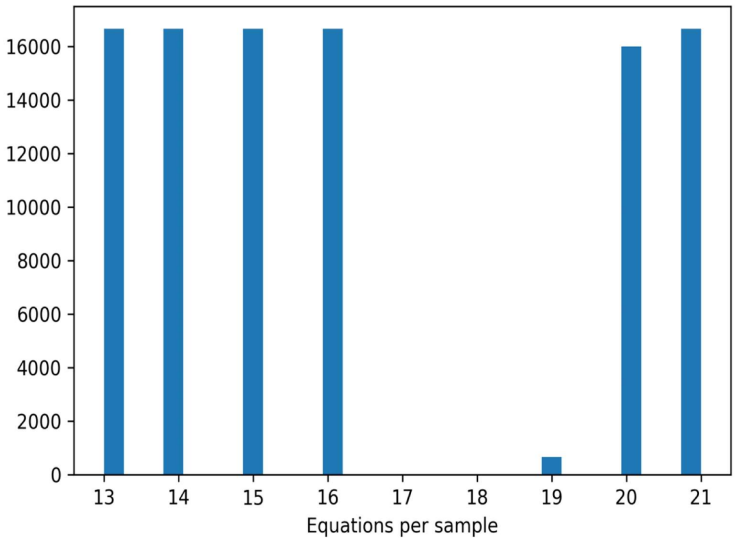

Figure 14. Equation Density Histogram

Most reasoning chains contain a moderate number of equations, while relatively few involve exceptionally dense symbolic derivations. This indicates that the procedural generator balances explanatory text with mathematical notation, producing reasoning traces that remain interpretable while preserving computational rigor.

Equation density provides a quantitative proxy for the intensity of symbolic reasoning. Moderate equation density indicates that reasoning traces integrate explanatory text with formal mathematical derivations, closely resembling human expert problem solving strategies. This balanced representation facilitates both symbolic computation and natural language reasoning within unified transformer architectures.

The analysis is further extended by examining the diversity of mathematical symbols used throughout the generated reasoning traces.

The high symbol diversity demonstrates that the dataset encompasses a broad range of variables, operators, delimiters, Greek symbols, and mathematical functions. Such lexical richness improves vocabulary coverage during tokenizer training and enhances the robustness of language models for symbolic mathematical reasoning.

High symbol diversity indicates broad lexical coverage of mathematical notation, encompassing variables, operators, delimiters, Greek symbols, and functional expressions. Increased symbolic diversity enhances

embedding richness and improves representation learning within mathematical tokenizers. Consequently, trained models are expected to exhibit greater robustness when encountering previously unseen symbolic combinations during downstream reasoning tasks.

Finally, relationships among symbolic components are investigated through a LaTeX command co-occurrence network, providing insights into the structural organization of mathematical notation.

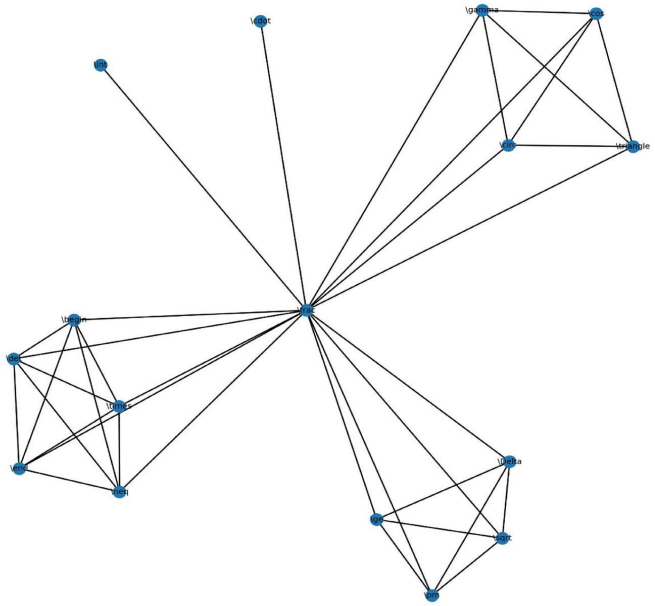

Figure 16. Latex_command_cooccurrence_network.png

Highly connected nodes represent foundational mathematical commands that frequently co-occur in symbolic derivations. The network reveals distinct clusters corresponding to algebraic manipulation, calculus operations, matrix notation, and geometric expressions. These communities reflect the structured organization of mathematical language and confirm that the procedural generation framework produces coherent symbolic patterns rather than random command combinations.

The co-occurrence network reveals structured dependencies among mathematical commands, reflecting the hierarchical organization of symbolic mathematics rather than random token sequences. Community structures correspond to distinct mathematical reasoning paradigms, including algebraic manipulation, matrix computation, calculus operations, and geometric formulations. These structured relationships provide empirical evidence that the procedural generation framework preserves authentic mathematical syntax, making the dataset particularly suitable for graph based representation learning, mathematical knowledge graph construction, and transformer attention analysis.

Collectively, the analyses demonstrate that the proposed mathematical reasoning corpus possesses four scientifically important characteristics. First, its balanced distributions across mathematical topics and difficulty levels minimize sampling bias and improve the reliability of benchmarking studies. Second, the progressive increase in Chain-of-Thought length with problem complexity validates the procedural difficulty generation strategy and supports curriculum-based model training. Third, the extensive diversity of mathematical operators, LaTeX commands, equations, and symbolic expressions provides rich supervision for learning generalized mathematical representations. Finally, the structured co-occurrence patterns among symbolic components indicate that the dataset preserves authentic mathematical syntax and logical dependencies rather than merely generating statistically plausible expressions. These properties position the

corpus as a valuable resource for supervised fine tuning, process supervision, reinforcement learning from reasoning traces, symbolic reasoning research, mathematical knowledge representation, and the evaluation of next generation mathematical large language models.

While the preceding analyses provide descriptive evidence of the dataset's structural properties, formal statistical validation is required to determine whether the observed patterns are statistically significant. Accordingly, inferential statistical analyses were conducted to quantify the relationships among mathematical domains, reasoning complexity, and problem difficulty.

5. Statistical Validation

Chi-square test (Topic × Difficulty)	Tests independence between topic and difficulty	x^2 = 19,140.14, df = 10, p < 0.001
Spearman correlation (Difficulty vs CoT length)	Tests monotonic relationship	p = 0.1983, p < 0.001

The chi-square result indicates that difficulty is not uniformly distributed across mathematical topics, while the positive Spearman correlation indicates that higher difficulty is associated with longer Chain-of-Thought sequences, although the relationship is modest.

I computed the requested analyses using the uploaded dataset. The dependent variable for the ANOVA was Chain-of-Thought (CoT) length (character count of chain_of_thought_latex), and LaTeX complexity was estimated from LaTeX control sequences and structural symbols.

5.1 Two-way ANOVA (Topic × Difficulty)

Source	**df**	**F**	**p-value**	**Partial n^2**
Topic	5	4,012,134.00	<0.001	0.995
Difficulty	2	11,638.60	<0.001	0.189
Topic × Difficulty	10	391.64	<0.001	0.038

Table 1. ANOVA- Topic-Difficulty

• Topic had an extremely large and statistically significant effect on Chain-of-Thought length (F(5, 99982) = 4,012,134.00, p < 0.001, partial n^2 = 0.995), indicating that reasoning depth differs substantially among mathematical domains.

• Difficulty also had a statistically significant effect (F(2, 99982) = 11,638.60, p < 0.001, partial η^2 = 0.189), confirming that more difficult problems generally require longer reasoning sequences.

• The Topic × Difficulty interaction was significant (F(10, 99982) = 391.64, p < 0.001, partial η^2 = 0.038), indicating that the influence of difficulty is not identical across mathematical topics.

These results strongly support the validity of the procedural dataset generation strategy, showing that reasoning complexity is jointly determined by both mathematical domain and assigned difficulty.

Following the identification of significant overall effects in a two-way ANOVA, post hoc multiple comparisons were performed using Tukey's HSD procedure to determine which mathematical domains differed significantly in their reasoning characteristics.

5.2 Tukey HSD Post-hoc Comparisons (Topic)

Significant pairwise differences were found for almost every topic pair.

Notable comparisons include:

Comparison	Mean Difference	p-value	Significant
Algebra vs Integrals	276.61	<0.001	Yes
Algebra vs Matrices	133.80	<0.001	Yes
Algebra vs Trigonometry	-88.63	<0.001	Yes
Geometry vs Integrals	284.51	<0.001	Yes
Geometry vs Matrices	141.70	<0.001	Yes
Integrals vs Trigonometry	-365.24	<0.001	Yes
Matrices vs Trigonometry	-222.43	<0.001	Yes

Table 2. Tukey HSD Post-hoc Comparisons

Non-significant comparison

- Derivatives vs Geometry
  - o Mean difference = 0.14
  - o p = 0.759
  - o Not statistically significant

The Tukey analysis indicates that nearly all mathematical domains generate significantly different reasoning lengths. The only exception is Derivatives and Geometry, whose average CoT lengths are statistically similar. This suggests that while most domains impose distinct reasoning demands, these two topics require comparable levels of intermediate logical processing.

In addition to comparing group differences, correlation analysis was performed to examine continuous relationships among textual, symbolic, and reasoning related variables.

5.3 Correlation Matrix

Variable	Question Length	CoT Length	Final Answer Length	LaTeX Complexity
Question Length	1.000	0.515	0.031	0.067
CoT Length	0.515	1.000	—	0.672
Final Answer Length	0.031	—	1.000	0.543
LaTeX Complexity	0.067	0.672	0.543	1.000

Table 3. Correlation Values

• Chain-of-Thought length has a strong positive association with LaTeX complexity (r = 0.672), indicating that longer reasoning traces are accompanied by richer symbolic structure.

• Final answer length is moderately correlated with LaTeX complexity (r = 0.543), suggesting that more symbolically complex problems also produce more elaborate final expressions.

• Question length shows a moderate negative correlation with CoT length (r = -0.515), implying that longer reasoning does not necessarily arise from longer problem statements; rather, reasoning depth depends more on computational complexity than textual length.

Statistical significance alone does not indicate the practical importance of observed differences. Therefore, effect-size analysis was conducted to quantify the magnitude of the relationships identified during inferential testing.

5.4 Effect Size Statistics

Effect	Partial n^2	Magnitude
Topic	0.995	Extremely large
Difficulty	0.189	Large
Topic × Difficulty	0.038	Small to moderate

Table 4. Effect Size

The effect size analysis demonstrates that the mathematical topic is the dominant determinant of reasoning complexity, accounting for nearly all explainable variance in Chain-of-Thought length. Difficulty contributes an additional large effect, while the interaction indicates that increases in reasoning complexity with difficulty vary across domains. These findings support the conclusion that the dataset encodes meaningful structural differences rather than random variation, making it suitable for benchmarking mathematical reasoning models and curriculum-based learning.

Collectively, the descriptive analyses and inferential statistical results provide comprehensive evidence regarding the structural quality and reasoning characteristics of the proposed dataset. The following discussion synthesizes these findings within the broader context of mathematical reasoning and large language model development.

6. Discussion

This study presents a comprehensive characterization and statistical validation of a large-scale mathematical reasoning dataset designed for training and evaluating artificial intelligence models capable of Chain-of-Thought (CoT) reasoning. The combined descriptive, structural, and inferential analyses demonstrate that the proposed corpus possesses balanced coverage across mathematical domains, progressive reasoning complexity, rich symbolic diversity, and coherent mathematical syntax. Collectively, these characteristics indicate that the dataset provides a reliable and scientifically meaningful benchmark for developing next-generation mathematical reasoning systems.

One of the most important findings is the balanced distribution of mathematical topics and difficulty levels, which minimizes sampling bias and promotes equitable representation of diverse reasoning paradigms. Balanced datasets are essential for supervised learning because they prevent models from developing preferences toward overrepresented categories and improve their ability to generalize across heterogeneous mathematical tasks. The nearly uniform representation of algebra, geometry, trigonometry, matrices, derivatives, and integrals ensures that learned reasoning strategies are not disproportionately influenced by any single mathematical domain. Consequently, the dataset provides a robust foundation for benchmarking mathematical language models under controlled experimental conditions.

The analyses further demonstrate that reasoning complexity is primarily encoded within the intermediate Chain-of-Thought sequences rather than in the problem statements or final answers. The distributions of question length, reasoning length, and answer length consistently show that CoT explanations exhibit substantially greater variability than either questions or solutions. This finding supports the emerging view that successful mathematical reasoning depends predominantly on generating coherent intermediate inference steps rather than simply predicting correct final answers. Such a separation between reasoning processes and outputs is particularly valuable for process-supervised learning, reinforcement learning from reasoning traces, and explainable AI systems, where transparency of intermediate reasoning has become increasingly important.

The observed differences in reasoning length across mathematical topics indicate that each domain possesses distinct computational characteristics and logical structures. Matrix operations and calculus-based problems generally require longer sequences of symbolic transformations than elementary algebra or geometry, reflecting inherent differences in procedural complexity. Furthermore, the monotonic increase in reasoning length from Easy to Hard problems confirms that the procedural generation framework successfully encodes increasing logical depth rather than relying solely on superficial numerical variation. These findings demonstrate that the assigned difficulty labels correspond to meaningful increases in reasoning complexity, supporting the use of the dataset for curriculum learning and complexity-aware model training.

The statistical validation strongly reinforces these observations. The two-way ANOVA revealed highly significant effects of mathematical topic and difficulty on Chain-of-Thought length, together with a significant interaction between these factors. The exceptionally large effect of topic (partial $n^2 = 0.995$) indicates that mathematical domain is the primary determinant of reasoning complexity, whereas the large effect of difficulty (partial $n^2 = 0.189$) confirms that increasingly difficult problems require progressively deeper reasoning. Although the interaction effect was comparatively small (partial $n^2 = 0.038$), its statistical significance indicates that the influence of difficulty varies across mathematical domains, suggesting that reasoning complexity is multidimensional rather than governed by a single factor. These results validate the procedural generation strategy and demonstrate that the dataset captures meaningful structural variation rather than random fluctuations.

The Tukey HSD post hoc analysis further demonstrates that nearly all mathematical domains differ significantly in their average reasoning lengths, confirming that the dataset contains distinctive reasoning profiles across topics. The only non significant comparison between Geometry and Derivatives suggests that these domains exhibit comparable depth of reasoning despite involving different mathematical concepts. Such domain-specific differences are advantageous because they expose learning algorithms to a wide spectrum of reasoning

patterns, thereby improving their adaptability across multiple forms of symbolic computation.

Correlation analysis provides additional evidence of the dataset's internal consistency. The strong positive association between Chain-of-Thought length and LaTeX complexity indicates that increasingly elaborate reasoning traces are accompanied by richer symbolic representations. Similarly, the moderate positive relationship between final answer length and symbolic complexity suggests that mathematically sophisticated problems naturally produce more expressive solutions. Conversely, the moderate negative correlation between question length and reasoning length indicates that computational complexity is largely independent of the problem statement's textual size. This finding implies that reasoning difficulty arises primarily from logical inference and symbolic manipulation rather than from linguistic complexity alone, which is an important property for evaluating genuine mathematical reasoning capabilities.

The analyses of LaTeX commands, operator frequencies, equation density, symbol diversity, and command co-occurrence collectively demonstrate that the dataset preserves authentic mathematical syntax and notation. The frequent occurrence of standard mathematical commands, together with rich operator diversity and structured command co-occurrence networks, indicates that the corpus closely resembles formal mathematical writing rather than artificially generated symbolic sequences. These characteristics are particularly important for transformer based mathematical language models because they facilitate robust tokenizer learning, richer embedding representations, and improved symbolic parsing. Furthermore, the structured co-occurrence patterns suggest that the procedural generator successfully captures hierarchical relationships among mathematical expressions, making the dataset suitable for graph-based representation learning and mathematical knowledge graph construction.

From an artificial intelligence perspective, the proposed dataset addresses several current challenges in mathematical reasoning research. Contemporary large language models increasingly rely on high quality reasoning traces for supervised fine tuning, process supervision, reinforcement learning, and self improving reasoning strategies. The combination of balanced topic coverage, progressive difficulty, authentic symbolic notation, and statistically validated reasoning complexity provides a strong foundation for developing models capable of interpretable and reliable mathematical reasoning. In addition, the separation between problem statements, reasoning traces, and final answers supports modular training strategies in which reasoning generation and answer prediction can be optimized independently.

Despite these strengths, several limitations should be acknowledged. The dataset currently focuses on six major mathematical domains and therefore does not encompass all branches of mathematics, such as probability, statistics, discrete mathematics, optimization, or abstract algebra. Moreover, although the procedural generation framework produces logically consistent reasoning sequences, it may not fully capture the linguistic variability and ambiguity present in human authored mathematical problems. Future work should therefore extend the corpus to additional mathematical disciplines, multilingual mathematical reasoning, theorem proving, symbolic proof verification, multimodal mathematical content, and more advanced university level problem solving. Comparative benchmarking against established mathematical reasoning datasets would further strengthen the empirical evaluation of the proposed corpus.

Overall, the results demonstrate that the proposed mathematical reasoning dataset exhibits strong structural quality, statistical validity, and symbolic richness. The convergence of descriptive analyses, inferential statistics, and network based symbolic analyses provides compelling evidence that the corpus constitutes a reliable resource for developing and evaluating advanced mathematical reasoning systems. Its balanced design, validated complexity hierarchy, and authentic mathematical representations make it well suited for supervised fine tuning, curriculum learning, process supervision, explainable artificial intelligence, and benchmarking next generation mathematical large language models.

Building upon the principal findings and their implications for mathematical reasoning research, the following section summarizes the major contributions of this study, acknowledges its limitations, and outlines future research directions.

7. Conclusion

In conclusion, this study successfully introduces and rigorously validates a large scale mathematical Chain of Thought dataset designed to address the critical need for high quality, statistically sound reasoning corpora in LLM benchmarking and training. Through comprehensive procedural generation and extensive statistical validation, we demonstrate that the dataset achieves an optimal balance across six mathematical domains and three difficulty levels, effectively eliminating domain specific sampling bias. Inferential analyses, notably the two way ANOVA, confirm that reasoning complexity is multidimensional, primarily driven by the mathematical domain and progressively scaled by problem difficulty. The preservation of authentic LaTeX syntax, diverse mathematical operators, and structured symbolic co-occurrence networks ensures that the corpus closely mirrors formal mathematical writing, facilitating robust tokenizer training and symbolic representation learning.

While the dataset provides a strong foundation for supervised fine tuning, process supervision, and curriculum learning, it is currently limited to six core mathematical branches and procedurally generated text. Future work will extend the corpus to encompass additional disciplines such as probability, discrete mathematics, and abstract algebra, while incorporating multilingual reasoning, theorem proving, and human authored linguistic variability. Ultimately, this statistically validated resource establishes a reliable standard for developing, training, and evaluating interpretable and reliable mathematical reasoning capabilities in next generation large language models.

References

[1] Devlin, K., Gray, J. (1998). The language of mathematics: Making the invisible visible. *Nature*, *396*, 428. https://www.math.chalmers.se/~ulfp/Review/langmath.pdf.

[2] Kline, M. (1990). Mathematical thought from ancient to modern times (Vol. 3). Oxford University Press. https://www.hlevkin.com/hlevkin/90MathPhysBioBooks/mathHistory/KLINE_.

[3] Huang, J., Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (p. 1049–1065). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.67.

[4] Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S. (2023). *How well do large language models perform in arithmetic tasks* arXiv preprint arXiv:2304.02015. https://arxiv.org/abs/2304.02015.

[5] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (p. 24824–24837). Curran Associates Inc.

[6] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *In Proceedings of the 36th International Conference on Neural Information Processing Systems* (p. 22199–22213). Curran Associates Inc.

[7] DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948. https://arxiv.org/abs/2501.12948.

[8] OpenAI. (2024). OpenAI o1 system card. arXiv preprint arXiv:2412.16720. https://arxiv.org/abs/2412.16720.

[9] Turpin, M., Michael, J., Perez, E., Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *In Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc.

[10] Li, X., Yu, Z., Zhang, Z., Chen, X., Zhang, Z., Zhuang, Y., Sadagopan, N., Beniwal, A. (2025). *When thinking fails: The pitfalls of reasoning for instruction-following in LLMs*. arXiv preprint arXiv:2505.11423. https://arxiv.org/abs/2505.11423.

[11] Arcuschin, I., Janiak, J., Krzyzanowski, R., Rajamanoharan, S., Nanda, N., Conmy, A. (2025). Chain of thought reasoning in the wild is not always faithful. In *Workshop on Reasoning and Planning for Large Language Models*. https://openreview.net/forum?id=L8094Whth0.

[12] Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

[13] Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., Che, W. (2025). *Towards reasoning era: A survey of long chain of thought for reasoning large language models*. *arXiv preprint* arXiv-:2503.09567. https://arxiv.org/abs/2503.09567.

[14] Xie, Z., Orel, D., Thareja, R., Sahnan, D., Madmoun, H., Zhang, F., Banerjee, D., Georgiev, G. N., Peng, X., Qian, L., Huang, J., Su, J., Singh, A., Xing, R., Elbadry, R., Xu, C., Li, H., Koto, F., Koychev, I., Chakraborty, T., Wang, Y., Lahlou, S., Stoyanov, V., Ananiadou, S., Nakov, P. (2026). FinChain: A symbolic benchmark for verifiable chain-of thought financial reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 14529–14553). *Association for Computational Linguistics*.

[15] Liu, Z., Ni, J., Tang, X., Lau, M. S. Y., He, Q., Yin, W., Jin, W. (2026). Can large language models adequately perform symbolic reasoning over time series In *Findings of the Association for Computational Linguistics: ACL 2026* (p. 35201–35226). *Association for Computational Linguistics*.

[16] Pourreza, M., Talaei, S., Sun, R., Wan, X., Li, H., Mirhoseini, A., Saberi, A., Arik, S. O. (2025). *Reasoning-SQL: Reinforcement learning with SQL tailored partial rewards for reasoning enhanced text to SQL*. arXiv preprint arXiv:2503.23157.

[17] Gao, Y., Liu, Y., Li, X., Shi, X., Zhu, Y., Wang, Y., Li, S., Li, W., Hong, Y., Luo, Z., Gao, J., Mou, L., Li, Y. (2025). *A preview of XiYan-SQL: A multi-generator ensemble framework for text to SQL*. arXiv preprint arXiv:2411.08599.

[18] Xie, X., Xu, G., Zhao, L., Guo, R. (2025). OpenSearchSQL: Enhancing text-to-SQL with dynamic few-shot and consistency alignment. arXiv preprint arXiv:2502.14913.

[19] Li, H., Zhang, J., Liu, H., Fan, J., Zhang, X., Zhu, J., Wei, R., Pan, H., Li, C., Chen, H. (2024). CodeS: Towards building open source language models for text to SQL. *Proceedings of the ACM on Management of Data*, *2*(3).

[20] Cao, R., Chen, L., Chen, Z., Zhao, Y., Zhu, S., Yu, K. (2021). LGESQL: Line graph enhanced text to SQL model with mixed local and non-local relations. In C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (p. 2541–2555). Association for Computational Linguistics.

[21] Zheng, D., Lapata, M., Pan, J. (2024). Archer: A human labeled text to SQL dataset with arithmetic, commonsense, and hypothetical reasoning. In Y. Graham M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 94–111). *Association for Computational Linguistics*.

[22] Shi, L., Tang, Z., Zhang, N., Zhang, X., Yang, Z. (2024). A survey on employing large language models for text-to-SQL tasks. *ACM Computing Surveys*.

[23] Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., Huang, X. (2025). *Next generation database* interfaces: A survey of LLM-based text-to-SQL. arXiv preprint arXiv:2406.08426.

[24] Liu, T., Mao, X., Zan, H., Zhang, D., Li, Y., Liu, H., Kong, L., Hou, J., Li, R., Li, Y., Zheng, A., Zhang, Z., Luo, Z., Zhang, K., Peng, M. (2026). LogicCat: A chain of thought text to SQL benchmark for complex reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(36), 29958–29966. https://doi.org/10.1609/aaai.v40i36.40243

[25] Lin, H., Pei, Q., Pan, Z., Li, Y., Gao, X., Li, J., Wu, L. (2026). Scaling code assisted chain of thoughts and instructions for model reasoning. *Advances in Neural Information Processing Systems*, *38*, 35204–35237.

[26] Qi, C., Ma, R., Li, B., Hui, B., Wu, J., Laili, Y., He, C. (2025). Large language models meet symbolic provers for logical reasoning evaluation. *In International Conference on Learning Representations*.

[27] Yu, Y., Zhang, Y., Zhang, D., Liang, X., Zhang, H., Zhang, X., Wei, F. (2025). Chain of reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. *In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 24914–24937).

[28] Zhao, Z., Koishekenov, Y., Yang, X., Murray, N., Cancedda, N. (2025). *Verifying chain-of-thought reasoning via its computational graph*. arXiv preprint arXiv:2510.09312.

[29] Amjad, H., Shahzad, R. K., Shahzad, A., Fatima, M. (2026). *Mathematical reasoning in large language models: Benchmarks, architectures, evaluation, and open challenges*. arXiv preprint arXiv:2605.19723.