Robust LaTeX-Aware Representation Learning for Semantic Retrieval of Formula-Dense STEM Manuscripts

Anonymous ACL submission

Abstract

General-purpose language models possess inherent limitations in capturing the structural semantics of formula-dense STEM academic manuscripts. To overcome this, this study proposes a LaTeX-aware embedding framework that directly utilizes raw source code. This methodology combines a two-stage LaTeX preprocessing that preserves key mathematical tokens with SciBERT, integrates LoRA-based domain-adaptive fine-tuning to precisely project manuscripts into a high-dimensional semantic space, and employs a hierarchical retrieval strategy for robustness. Experiments using a large-scale arXiv dataset show that the proposed framework outperformed all baselines, including standard domain-specific models. In particular, we demonstrated that the performance gap with standard models increases significantly for documents with higher formula density, proving the robustness and effectiveness of this approach in learning STEM document representations.

1 Introduction

The exponential growth of STEM (Science, Technology, Engineering, and Mathematics) literature has intensified the demand for robust semantic retrieval systems for highly specialized queries Kim et al. (2025). In this domain, expert knowledge is densely encapsulated not only in natural language but also in complex formulas and symbolic expressions. Despite recent advances in general-purpose Large Language Models (LLMs), capturing the semantics of formula-dense manuscripts remains challenging Peng et al. (2021); Singh et al. (2022). A crucial bottleneck in semantic retrieval tasks stems from the failure to preserve the structural integrity of the original text. Although most STEM manuscripts are written in LaTeX and published as PDFs, conventional PDF-to-text conversion often misparses mathematical structures or renders symbols as noise Bast and Korzen (2017). As demonstrated by Beltagy et al. (2019), such structural inaccuracies directly lead to a substantial loss of semantic fidelity, severely degrading the final document embedding quality and subsequent retrieval performance.

To fundamentally overcome these limitations, we focus on handling the raw LaTeX source code of STEM literature. The LaTeX source eliminates the risks of PDF parsing and provides an explicit, structurally rich representation of mathematical content. In this paper, we propose a LaTeX-aware semantic retrieval framework optimized for formula-reliant environments. We conduct representation learning based on SciBERT Beltagy et al. (2019) and integrate a parameter-efficient domain-adaptive fine-tuning strategy leveraging LoRA Hu et al. (2021). Our system preserves the fidelity of complex formulas while dynamically adapting to emerging research trends and sub-fields with minimal computational overhead.

The main contributions of this work are threefold. (1) We propose a LaTeX-aware representation learning framework that directly utilizes raw LaTeX sources for semantic retrieval of STEM manuscripts, preserving mathematically meaningful tokens while removing syntactic noise. (2) We incorporate LoRA-based domain-adaptive fine-tuning into SciBERT, enabling efficient adaptation to evolving mathematical sub-fields with minimal additional parameters. (3) Hierarchical semantic retrieval strategy that narrows the search space through coarse-to-fine domain selection, improves retrieval robustness without exhaustive corpus-wide search. Experiments on a large-scale Mathematics corpus further show that the proposed framework is effective for formula-dense manuscripts.

2 Related Work

Recent advances in LLMs and text embeddings have substantially improved automated scholarly

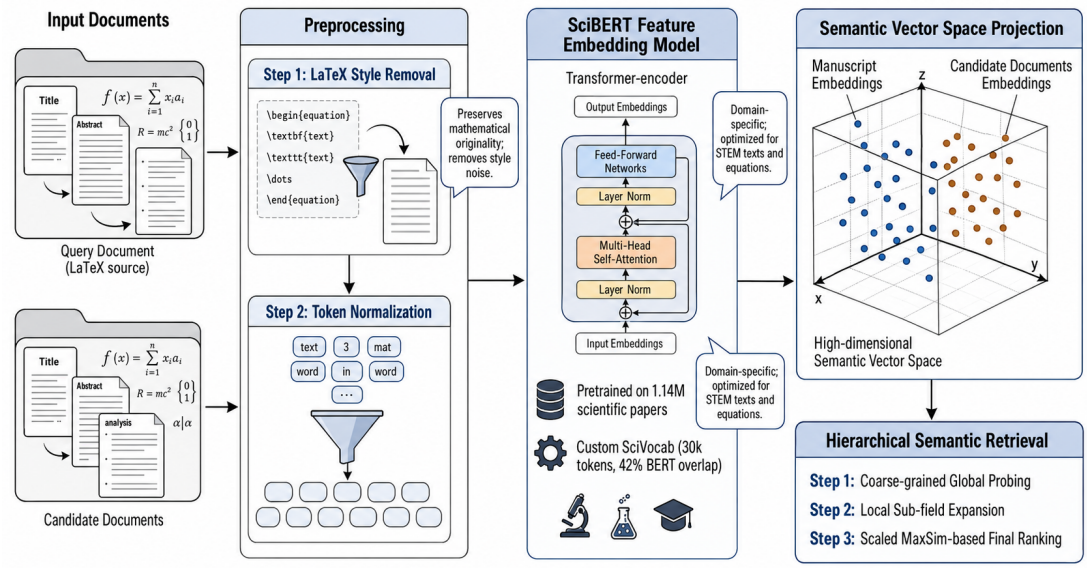


Figure 1: Feature embedding pipeline with LLMs, including how LaTeX source based query document and candidate documents are preprocessed, encoded by SciBERT, and mapped into a semantic vector space for similarity calculation of semantic relevance.

document processing and semantic matching Liu et al. (2014); Lin et al. (2023); Aksoy et al. (2025); Fu et al. (2025). Early embedding methods such as Word2Vec Mikolov et al. (2013) and GloVe Pennington et al. (2014) often suffered from vocabulary mismatches in technical documents, especially when mathematical notation was involved, and thus relied on external metadata or structural constraints Dumais and Nielsen (1992); Mimno and McCallum (2007); Rodriguez and Bollen (2008); Kobren et al. (2019).

Although modern LLMs provide stronger representations for scientific literature Karan et al. (2025); Zhu et al. (2025), their reliability remains limited in formula-intensive STEM domains Stelmakh et al. (2023). In semantic retrieval, embedding quality directly affects the ability to capture contextual relevance between a query and manuscripts Karpukhin et al. (2020); Humeau et al. (2019); Muennighoff et al. (2023); Reimers and Gurevych (2019); Thakur et al. (2021). Domain-adapted models such as SciBERT improve representations by incorporating scientific vocabulary Beltagy et al. (2019), while MathBERT highlights the importance of modeling mathematical structure Peng et al. (2021). However, when using these models, directly processing raw LaTeX introduces synthetic noise from formatting commands and vocabulary mismatches for mathematical tokens Drozdov et al. (2022); Weng et al. (2023). To address these limitations, we propose a LaTeX-aware representation learning pipeline that preserves structural and mathematical semantics in STEM manuscripts. Additionally, an in-depth discussion of related work associated with this study is provided in Section A.

3 Methodology

Figure 1 illustrates the overall architecture of our proposed framework. The core novelty lies in explicitly capturing the structural characteristics of formula-dense STEM literature by integrating four key components. Together, the unified framework enables highly accurate manuscript retrieval even within complex technical contexts.

3.1 LaTeX-Aware Representation Learning for STEM Manuscripts

To overcome the limitations of conventional PDF text extraction, which often fails to parse complex mathematical symbols and renders them as noise, we directly utilize raw LaTeX sources for effective STEM literature embedding. Based on Habib and Afzal (2019), we extract the title, abstract, and introduction as the core textual features to optimize the balance between computational efficiency and semantic richness.

To eliminate syntactic noise while preserving the essential meaning of mathematical expressions, our framework introduces a rigorous two-stage preprocessing pipeline: (1) **Style Removal** strips away semantic-agnostic formatting commands, such as (`\documentclass`), (`\usepackage`), tables, fig-

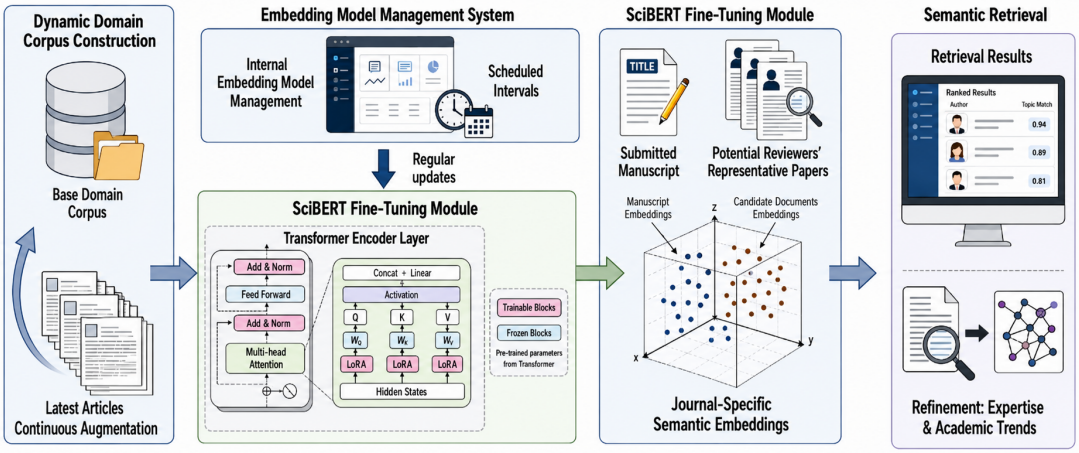


Figure 2: Journal-adaptive fine-tuning workflow for dynamic semantic retrieval. With journal-specific corpus and fine-tuned SciBERT based framework enables the retrieval system to adapt to emerging expertise and evolving academic trends with minimal computational overhead.

ures, and cross-reference tags (`\label`, `\ref`), allowing the model to focus purely on the semantic essence. (2) **Token Normalization** filters out comments, citation metadata, and special characters (e.g., $, %, &, _, {) that cause embedding interference, while cleaning up whitespaces within equations to increase embedding density. To prevent the loss of mathematically crucial commands, we construct a custom LaTeX Math Dictionary based on Mathmode Voß (2014) to explicitly preserve key formula tokens.

Since general-purpose LLMs are less effective at capturing fine-grained STEM semantics of LaTeX due to vocabulary limitations (Hallee et al., 2024), we adopt SciBERT as our representation learning model. SciBERT's pretraining dataset contains LaTeX source, and its specialized vocabulary (*SciVocab*) precisely captures the rich contextual nuances of technical jargon and mathematical formulas. As illustrated in Figure 1, the preprocessed LaTeX texts of the query and candidate documents are projected into high-dimensional semantic representations through SciBERT. Detailed discussions regarding model selection are provided in Sections B.2.1 and B.2.2.

3.2 Domain-Adaptive Fine-Tuning for Efficient Manuscript Representation

Although SciBERT provides effective embedding capabilities across broad STEM domains, it faces limitations in adapting to rapidly evolving trends within emerging sub-fields Hallee et al. (2025). While standard NLP models like SciBERT infer the semantics of out-of-vocabulary (OOV) words through tokenization, this flexibility does not extend to LaTeX formulas, which are structurally rigid and highly sensitive to token fragmentation. Consequently, when a general encoder encounters unseen LaTeX sequences or domain-specific expressions, the lack of optimized representations leads to a severe degradation in semantic fidelity.

However, retraining the entire model for every new sub-field is computationally prohibitive. Therefore, we adopt a domain-adaptive fine-tuning strategy based on Low-Rank Adaptation (LoRA) Hu et al. (2021). By keeping the pre-trained weights of SciBERT frozen and inserting low-rank matrices into selected layers, this approach updates only 0.2% of the total parameters, achieving high domain adaptability with minimal computational overhead. The overall workflow of the fine-tuning mechanism is illustrated in Figure 2. Automated by an Embedding Model Management System, this pipeline accumulates newly submitted manuscripts and incoming queries, periodically activating the fine-tuning process. Consequently, our framework moves beyond static knowledge, providing a sustainable semantic retrieval environment that dynamically adapts to the evolving academic ecosystem.

3.3 Hierarchical Semantic Retrieval for Robust Manuscript Representation

Computing semantic similarity across an entire large-scale STEM corpus can introduce substantial out-of-domain noise, which degrades retrieval robustness. To overcome these inefficiencies and maintain robust manuscript representation, we propose a three-stage hierarchical semantic retrieval

Category (Notation)	Entry	Category (Notation)	Entry	Category (Notation)	Entry
Commutative Algebra (AC)	4,651	Algebraic Geometry (AG)	16,029	Analysis of PDEs (AP)	14,098
Algebraic Topology (AT)	6,957	Classical Analysis and ODEs (CA)	7,309	Combinatorics (CO)	21,722
Category Theory (CT)	4,674	Complex Variables (CV)	6,752	Differential Geometry (DG)	13,055
Dynamical Systems (DS)	14,550	Functional Analysis (FA)	11,427	Group Theory (GR)	8,332
Geometric Topology (GT)	7,480	Mathematical Geometry (MG)	5,253	Numerical Analysis (NA)	10,666
Number Theory (NT)	10,049	Operator Algebras (OA)	4,683	Optimization and Control (OC)	18,004
Probability (PR)	18,005	Quantum Algebra (QA)	9,042	Rings and Algebras (RA)	6,766
Representation Theory (RT)	11,332	Symplectic Geometry (SG)	4,348	Spectral Theory (SP)	5,149
Statistics Theory (ST)	8,837	Group1 (G1)	4,829	Group2 (G2)	5,478

Table 1: The category notation used in this study and the corresponding number of papers available on arXiv.

strategy that improves retrieval robustness through a structured search strategy. Furthermore, we employ Scaled MaxSim (S) as our core similarity metric (a detailed formulation of S is provided in Section 4.2).

Step 1: Coarse-grained Global Probing This stage identifies the primary research domains relevant to the query. For each academic category C_j ($j = 1, \ldots, n_{\text{cat}}$), we construct a sampled document set $\mathcal{D}^{(j)}_{\text{global}}$ by randomly selecting N_{global} manuscripts. We then evaluate the query document d_q against these samples to retrieve the top-K most similar manuscripts based on scaled MaxSim S. Finally, we derive the three most frequent candidate categories, denoted as $\mathcal{C}_{\text{top3}}$:

$$\begin{aligned}
\mathcal{D}_{\text{global}} &= \bigcup_{j=1}^{n_{\text{cat}}} \mathcal{D}^{(j)}_{\text{global}}, \\
\mathcal{R}_{\text{global}} &= \left\{ (d_i, S(d_q, d_i)) \;\middle|\; d_i \in \mathcal{D}_{\text{global}} \right\}, \\
\mathcal{C}_{\text{top3}} &= \underset{C_j \in \mathcal{C}}{\text{Top3}} \sum_{(d_i, \cdot) \in \text{Top-}K(\mathcal{R}_{\text{global}})} \mathbf{1}(d_i \in C_j).
\end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function, Top-$K(\mathcal{R}_{\text{global}})$ denotes the top-K ranked document-score pairs sorted by similarity score, and Top3 identifies the three most frequent categories appearing among the retrieved documents.

Step 2: Local Sub-field Expansion To prevent the risk of false negatives (FN) and maintain high recall within the selected domains, we further expand the search space using top-ranked categories in $\mathcal{C}_{\text{top3}}$. Specifically, for each category $C_k \in \mathcal{C}_{\text{top3}}$, we randomly select N_{expand} documents and construct an expanded candidate set $\mathcal{D}_{\text{expand}}$ as follows:

$$\mathcal{D}_{\text{expand}} = \bigcup_{C_k \in \mathcal{C}_{\text{top3}}} \mathcal{D}^{(k)}_{\text{expand}} .$$

We then compute similarity scores between the query document d_q and all documents in the expanded candidate pool:

Step 3: Scaled MaxSim-based Final Ranking Finally, we merge the document-score pairs from the global and expanded retrieval stages and rerank them based on similarity scores to obtain the final top-K results, denoted as $\mathcal{R}_{\text{final}}$:

$$\begin{aligned}
\mathcal{R}_{\text{expand}} &= \{(d_i, S(d_q, d_i)) \mid d_i \in \mathcal{D}_{\text{expand}}\}, \\
\mathcal{R}_{\text{final}} &= \text{Top-}K \left(\mathcal{R}_{\text{global}} \cup \mathcal{R}_{\text{expand}} \right)
\end{aligned}$$

By focusing computations on a bounded subset of approximately $N_{\text{total}} = (n_{\text{cat}} \times N_{\text{global}}) + (3 \times N_{\text{expand}})$ data points rather than performing an exhaustive corpus-wide search, this hierarchical strategy significantly reduces computational overhead. Concurrently, while preserving semantic coverage across neighboring research domains, it ensures robust performance for interdisciplinary literature and emerging sub-fields.

4 Experimental Results

4.1 Datasets

We constructed a semantic retrieval evaluation database by collecting approximately 2.4 million STEM preprints from arXiv arXiv (2025), a large-scale open-access repository. Our experiments focus specifically on the Mathematics corpus, which provides a highly challenging testbed for domain-specific representation learning due to its dense integration of natural language and complex mathematical expressions Wang et al. (2025). The Mathematics category in arXiv consists of 32 sub-fields, which we utilize as target domain labels for adaptation and evaluation. To ensure statistical reliability, we filtered out categories with insufficient samples and consolidated closely related adjacent sub-fields. Detailed statistics of the final dataset distribution and category abbreviations are described in Table 1

Category	CO ($K = 50$)				CO ($K = 100$)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
w/ Gensim	0.458	0.468	0.486	0.454	0.470	0.474	0.477	0.468
w/ GoogleNews	0.247	0.273	0.294	0.240	0.225	0.312	0.276	0.221
w/ BERT	0.578	0.599	0.628	0.571	0.579	0.626	0.602	0.576
w/ MathBERT	0.552	0.541	0.583	0.572	0.623	0.586	0.546	0.538
w/ SBERT	0.338	0.307	0.355	0.344	0.380	0.353	0.334	0.306
w/ SciBERT	**0.672**	**0.706**	**0.733**	**0.666**	**0.644**	**0.710**	**0.705**	**0.640**
Category	CV ($K = 50$)				CV ($K = 100$)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
w/ Gensim	0.315	0.345	0.387	0.306	0.302	0.396	0.355	0.298
w/ GoogleNews	0.235	0.260	0.306	0.230	0.204	0.274	0.272	0.201
w/ BERT	0.425	0.452	0.486	0.418	0.372	0.455	0.449	0.369
w/ MathBERT	0.375	0.350	0.394	0.395	0.420	0.391	0.370	0.347
w/ SBERT	0.355	0.330	0.381	0.388	0.428	0.385	0.350	0.327
w/ SciBERT	**0.470**	**0.511**	**0.571**	**0.462**	**0.443**	**0.522**	**0.521**	**0.438**
Category	OC ($K = 50$)				OC ($K = 100$)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
w/ Gensim	0.675	0.696	0.730	0.668	0.642	0.672	0.700	0.639
w/ GoogleNews	0.460	0.482	0.526	0.452	0.458	0.497	0.497	0.453
w/ BERT	0.812	0.843	0.885	0.809	0.764	0.795	0.841	0.761
w/ MathBERT	0.680	0.642	0.705	0.680	0.743	0.704	0.673	0.639
w/ SBERT	0.652	0.599	0.671	0.639	0.700	0.667	0.648	0.596
w/ SciBERT	**0.840**	**0.859**	**0.892**	**0.837**	**0.790**	**0.814**	**0.855**	**0.788**

Table 2: Comparison of overall retrieval performance across embedding models. The best and the second-best results are highlighted in **bold**, and underlined, respectively.

and a comprehensive description of the arXiv repository is provided in Section B.1.

4.2 Performance Evaluations

Evaluation Procedure and Ground Truth For each academic category, a representative manuscript was randomly selected as a query manuscript, while the remaining documents formed the candidate retrieval corpus. The objective of our framework is to accurately retrieve manuscripts semantically related to the query. The ground truth is defined based on category membership: a candidate document is labeled as *Relevant* (Label 1) if it shares the same sub-field arXiv category as the query, and *Non-relevant* (Label 0) otherwise.

Similarity Scoring Function Accurately quantifying the similarity between formula-dense manuscripts in the latent space is a key determinant of retrieval performance. Common metrics such as Cosine or Jaccard similarity alone struggle to capture the complex interplay between mathematical structures and natural language. To address this limitation, we adopt Scaled MaxSim Choi et al. (2023), which fuses both metrics to capture both structural and semantic properties concurrently.

Specifically, for a given query d_q, we compute its Cosine and Jaccard similarity scores against all candidate manuscripts, constructing two comprehensive score vectors, $\mathbf{c}_q$ and $\mathbf{j}_q$. To address the scale mismatch between Cosine and Jaccard metrics, these vectors are normalized using the Softmax function (σ). The final blended similarity score vector is formulated as:

$$\text{Scaled MaxSim}(\mathbf{c}_q, \mathbf{j}_q) = \frac{1}{2}\big(\sigma(\mathbf{c}_q) + \sigma(\mathbf{j}_q)\big) \qquad (1)$$

As shown in Table 4, preliminary evaluations confirm that our proposed fusion metric is uniquely suited for LaTeX formula-dense contexts, consistently yielding higher similarity scores for intra-category documents and successfully elevating their ranking density. Detailed configurations on similarity functions are provided in Section B.3.1.

Evaluation Metrics To comprehensively assess retrieval quality, we employ standard information retrieval metrics: Precision, Mean Average Precision (MAP), Normalized Discounted Cumula-

Category	Precision@100	AP@100	nDCG@100	Pairwise F1@100
AC	0.170 → 0.380 (↑)	0.312 → 0.357 (↑)	0.471 → 0.504 (↑)	0.373 → 0.414 (↑)
AG	0.500 → 0.670 (↑)	0.707 → 0.727 (↑)	0.631 → 0.699 (↑)	0.667 → 0.818 (↑)
AP	0.670 → 0.680 (↑)	0.734 → 0.764 (↑)	0.703 → 0.717 (↑)	0.676 → 0.737 (↑)
CA	0.590 → 0.740 (↑)	0.756 → 0.785 (↑)	0.648 → 0.766 (↑)	0.646 → 0.657 (↑)
CO	0.940 → 0.950 (↑)	0.985 → 0.961 (↓)	0.954 → 0.956 (↑)	0.707 → 0.768 (↑)
CT	0.900 → 0.920 (↑)	0.926 → 0.932 (↑)	0.912 → 0.926 (↑)	0.879 → 0.901 (↑)
CV	0.790 → 0.870 (↑)	0.952 → 0.955 (↑)	0.838 → 0.898 (↑)	0.859 → 0.899 (↑)
DS	0.530 → 0.420 (↓)	0.688 → 0.699 (↑)	0.602 → 0.644 (↑)	0.686 → 0.768 (↑)
DG	0.460 → 0.460 (→)	0.559 → 0.570 (↑)	0.525 → 0.492 (↓)	0.459 → 0.570 (↑)

Table 3: Performance changes in evaluation metrics at $K = 100$ before and after fine-tuning. (Before → After)

Function	P@50	P@100	D@50	D@100	M@50	M@100
Cosine	0.612	0.584	0.707	0.655	0.536	0.457
Jaccard	0.506	0.449	0.575	0.518	0.354	0.284
MaxSim	0.636	0.606	0.728	0.687	0.551	0.490

Table 4: Retrieval performance comparison of similarity measures by function. MaxSim denotes scaled MaxSim fromEquation (1). P@K, D@K, and M@K denote Precision@K, nDCG@K, and MAP@K. A detailed explanation of each metric is in Section B.3.2.

tive Gain (nDCG), and Pairwise-F1 Ebrahimi et al. (2025). Formal mathematical definitions for each metric are detailed in Section B.3.2.

Implementation Details All experiments in Section 4.3 were conducted under an identical hierarchical semantic retrieval protocol for all embedding baselines. Specifically, for Step 1, the global sampling size per category was set to $N_{\text{global}} = 500$, with $p = 100$ for top-3 category identification. For Step 2, the expanded search space per category was configured to $N_{\text{expand}} = 2000$.

4.3 Main Results

Overall Performance of LaTeX-Aware Representation Learning Our first evaluation validates the semantic retrieval performance of the proposed LaTeX-aware preprocessing and hierarchical framework across various standard representation methods. We compare our framework against traditional embeddings (Word2Vec Rehurek (2025)), general LLMs (BERT Devlin et al. (2019), SBERT Reimers and Gurevych (2019)), and a math-specific encoder (MathBERT Peng et al. (2021)). Based on a random sample of eight manuscripts per category, the results in Section 3.3 demonstrate the overall retrieval performance. The results show that our proposed framework achieves optimal performance across most metrics when paired with SciBERT, a domain-specific encoder, substantially outperforming all other baselines.

For instance, in the OC category, the SciBERT-infused pipeline scored 0.8594 in nDCG@50, markedly surpassing BERT (0.8430) and Gensim (0.6959). This gap underscores that simple distributional semantics or general-purpose embeddings cannot sufficiently resolve technical document matching, highlighting the necessity of domain-specific representations. Crucially, our pipeline also outperforms MathBERT. In contrast, augmenting a broader scientific encoder like SciBERT with our structurally adaptive LaTeX pipeline avoids localized overfitting while precisely capturing and generalizing the underlying semantics of formula-dense manuscripts. Detailed comparisons are provided in Section D.2.

Effectiveness of Domain-Adaptive Fine-Tuning We conducted domain-adaptive fine-tuning using a LoRA strategy, inserting rank-8 adapters into the Query (Q) and Value (V) projection layers of the frozen SciBERT backbone. The adapters were trained via Masked Language Modeling (MLM) using the AdamW optimizer for 3 epochs and deployed as modular plug-ins. To simulate a real-world production environment, we treated the entire Mathematics corpus except evaluation set as a cumulative journal publication history for fine-tuning. Evaluation was performed using expert-validated representative test papers per category, focusing strictly on @100 metrics to isolate the fine-tuning effects.

The impact of fine-tuning on retrieval performance yields two key insights. First, as shown in Section 4.2, the most pronounced improvements occurred in theory-oriented mathematics fields (e.g., AG, AP, CA, CT, DG). This indicates that domain-specific terminology and structural patterns are ef-

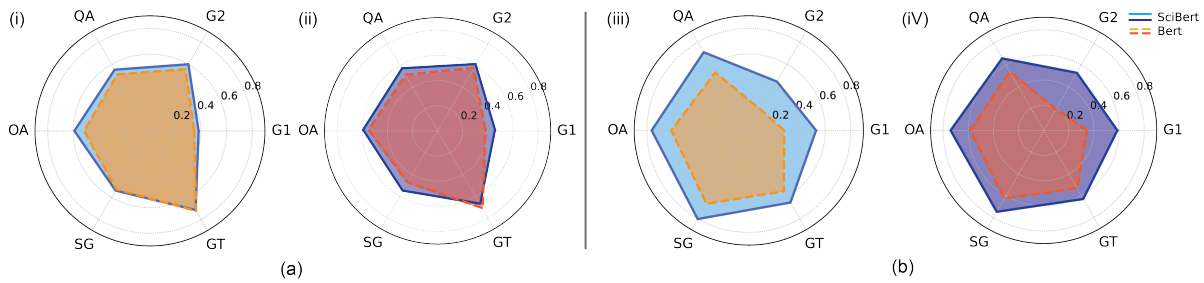


Figure 3: Performance gap between SciBERT and BERT under different formula density conditions. (a) The left radar plot reports AP@100 and the right radar plot presents nDCG@100 for the low formula-density group. (b) The left radar plot reports AP@100 and the right radar plot presents nDCG@100 for the high formula-density group.

fectively optimized, providing consistent performance gains even for low-resource minor fields. Second, Pairwise-F1 exhibited the most substantial improvement across all metrics. This underscores that the fine-tuning process enhances the clustering structure within the latent embedding space, precisely tightening the semantic proximity between manuscripts belonging to the same sub-field. Detailed results for additional sub-fields are available in Section D.4.

Robustness on Formula-Dense STEM Manuscripts To identify the core drivers of the performance gap between SciBERT and BERT, we investigate the impact of formula density on embedding quality and ranking accuracy. We divided the test papers into high-density and low-density formula groups based on the ratio of mathematical tokens (LaTeX commands retaining mathematical semantics) to overall text. Specifically, we identified manuscripts whose formula token proportions fell within the top 25% or bottom 25% relative to the category average. Following expert validation, a single representative manuscript per density group and category was selected. Given the single-query setup, we report Average Precision (AP) instead of MAP.

As illustrated in Figure 3, SciBERT outperforms BERT across both groups, but the performance gap widens distinctively depending on formula density. In the low-density group, SciBERT yields modest gains over BERT across most metrics, with a few exceptions like the GT category. In contrast, within the high-density group, SciBERT markedly widens the margin, achieving substantially higher performance. Notably, while the gap in basic Precision metrics remains relatively small, SciBERT consistently dominates BERT in ranking quality metrics such as AP, nDCG, and Pairwise-F1.

This discrepancy indicates that SciBERT excels not merely at broad document retrieval, but at precisely modeling relative ordering and semantic consistency to generate a top-heavy, tightly clustered ranking. Ultimately, these findings demonstrate the intrinsic robustness of our proposed framework for semantic retrieval in formula-dense STEM literature where complex LaTeX expressions are heavily integrated. Detailed empirical values are provided in Section D.3.

4.4 Ablation Studies

Effect of LaTeX-Aware Preprocessing To systematically analyze how preprocessing levels affect retrieval performance, we compare three configurations using raw LaTeX sources: (1) *Minimal processing* (comment removal only), (2) *Standard processing* (our pipeline; removing syntactic noise while preserving semantic symbols and formulas), and (3) *Aggressive processing* (stripping most LaTeX expressions to focus strictly on natural language). As illustrated in Figure 4-(b), Standard processing consistently achieves the highest performance across all metrics. Minimal processing retains excessive markup artifacts and formatting tokens, significantly degrading embedding quality. Conversely, Aggressive processing over-cleans structural information, destroying critical formula-based semantics and domain-specific symbols. These performance drops are especially pronounced in formula-dense categories (e.g., RT), proving that LaTeX formulas function as core semantic units rather than mere formatting elements. Thus, Standard processing serves as the optimal strategy to balance noise reduction and information preservation.

Effect of Hierarchical Semantic Retrieval We next evaluate the impact of each stage in our hierarchical pipeline. Specifically, we validate our

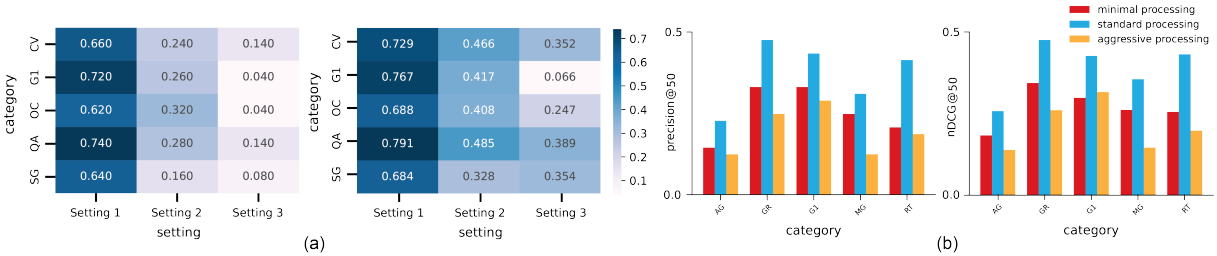


Figure 4: (a) Performance comparison across different content composition settings. The left heatmap reports Precision@50, and the right heatmap presents nDCG@50. (b) Performance comparison across different LaTeX processing levels. The left bar chart shows Precision@50, and the right bar chart illustrates nDCG@50.

Phase	P@50	P@100	D@50	D@100	M@50	M@100
Global	0.365	0.305	0.428	0.358	0.588	0.498
Final	0.607	0.570	0.705	0.655	0.650	0.640

Table 5: Comparison of retrieval performance of each step in hierarchical retrieval strategy. P@K, D@K, and M@K denote Precision@K, nDCG@K, and MAP@K. Detailed explanation on each metric is in Section B.3.2.

phased filtering strategy by comparing the performance of the initial domain probing phase (*Global*; Step 1) against the final expanded and re-ranked results (*Final*; Step 3), as summarized in Table 5. The Final stage substantially outperforms the Global stage across all metrics. For instance, P@50 improves from 0.3652 at the Global stage to 0.6065 at the Final stage, with consistent gains in ranking metrics like AP and Pairwise-F1. This leap verifies that our hierarchical approach simultaneously secures computational efficiency and semantic accuracy without requiring an exhaustive corpus-wide search. Identifying the query's global position via Global Probing and subsequently narrowing down the search via Final Retrieval provides substantial, practical performance gains in large-scale academic setups.

Ablation of Contextual Components for Manuscript Embedding Embedding an entire manuscript is computationally prohibitive and risks introducing noise from granular methodological details. Therefore, selecting contextual components that best represent the core semantics of a paper is essential. While the Title and Abstract provide condensed summaries of key contributions, the Introduction offers broader research context and background, playing a critical role in semantic similarity estimation. We evaluate three component combinations: *Setting 1* (Title + Abstract + Introduction), *Setting 2* (Title + Abstract), and *Setting 3* (Title only). As detailed in Section D.5, *Setting 1* achieves the most stable and superior performance across all categories. Notably, as shown in Figure 4-(a), including the Introduction heavily influences performance, as its motivational and background information serves as a key proxy for estimating inter-paper similarity. Conversely, *Settings 2* and *3* exhibit distinct performance drops that worsen in more complex domains. This confirms that *Setting 1* offers the most effective and optimized combination for constructing STEM manuscript representations.

5 Conclusion

We presented a LaTeX-aware representation learning and retrieval framework for semantic retrieval over STEM manuscripts. Our framework directly leverages raw LaTeX sources through a two-stage preprocessing strategy guided by a custom mathematical vocabulary. Combined with the domain-specific SciBERT encoder, it captures both natural-language context and mathematical semantics in a unified embedding space. We also incorporated a hierarchical retrieval strategy to reduce computational overhead while preserving semantic coverage across relevant research domains. Experiments on a large-scale arXiv dataset show that our framework outperforms domain-specialized baselines across multiple retrieval metrics. The gains are particularly pronounced for formula-dense manuscripts, demonstrating the robustness of preserving mathematical structure during representation learning. LoRA-based domain-adaptive fine-tuning further confirms the flexibility of our framework for efficient adaptation. Overall, our findings suggest that preserving structural and mathematical information in source-level representations is a promising direction for LaTeX-aware scientific document understanding.

Limitations

A limitation of this study is that the framework has been evaluated primarily within the mathematics domain, and its generalizability to other notation-heavy STEM fields, such as physics or computer graphics, requires further validation. We also acknowledge that arXiv category labels may not fully capture fine-grained semantic relatedness as judged by human experts. Nevertheless, they provide a practical and reproducible proxy for constructing a large-scale benchmark, which would be difficult to achieve through manual annotation at the same scale. Because arXiv category information is used in both candidate pruning and relevance construction, our results are best interpreted as category-consistent retrieval performance rather than fully expert-annotated semantic matching. Finally, while our LaTeX-aware preprocessing preserves formula-related tokens, it does not explicitly model deeper mathematical structures such as operator trees or symbol graphs. Future work will extend the evaluation to stronger retrieval baselines, significance testing, and detailed runtime analysis.

References

Meltem Aksoy, Seda Yanik, and Mehmet Fatih Amasyali. 2025. A data-driven optimization approach for automated reviewer assignment using natural language processing. *Intelligent Systems with Applications*, page 200587.

arXiv. 2025. arxiv submission statistics. https://arxiv.org/stats/monthly_submissions. Accessed: April 2025.

Hannah Bast and Claudius Korzen. 2017. A benchmark and evaluation for text extraction from pdf. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)*, pages 1–10. IEEE.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint* arXiv:1903.10676.

Jiaqi Cao, Jiarui Wang, Rubin Wei, Qipeng Guo, Kai Chen, Bowen Zhou, and Zhouhan Lin. 2025. Memory decoder: A pretrained, plug-and-play memory for large language models. *arXiv preprint arXiv:2508.09874*.

Dong-Hoon Choi, Jin Woo Hyun, and Young Rock Kim. 2023. An algorithm for peer reviewer recommendation based on scholarly activity assessment. *IEEE Access*, 11:39609–39620.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*.

S. T. Dumais and J. Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244.

Sajad Ebrahimi, Sara Salamat, Negar Arabzadeh, Mahdi Bashari, and Ebrahim Bagheri. 2025. exharmony: Authorship and citations for benchmarking the reviewer assignment problem. In *European Conference on Information Retrieval*, pages 1–16. Springer.

Yongfan Fu, Jian Luo, Guofang Nan, and Dahui Li. 2025. Peer review expert group recommendation: A multi-subject coverage-based approach. *Expert Systems with Applications*, 264:125971.

Raja Habib and Muhammad Tanvir Afzal. 2019. Sections-based bibliographic coupling for research paper recommendation. *Scientometrics*, 119(2):643–656.

Logan Hallee, Rohan Kapur, Arjun Patel, Jason P Gleghorn, and Bohdan Khomtchouk. 2024. Contrastive learning and mixture of experts enables precise vector embeddings. *arXiv preprint* arXiv:2401.15713.

Logan Hallee, Rohan Kapur, Arjun Patel, Jason P Gleghorn, and Bohdan B Khomtchouk. 2025. Contrastive learning and mixture of experts enables precise vector embeddings in biological databases. *Scientific Reports*, 15(1):14953.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

V. M. Karan, S. McQuistin, R. Yanagida, C. Perkins, G. Tyson, I. Castro, P. Healey, and M. Purver. 2025. A dataset for expert reviewer recommendation with large language models as zero-shot rankers. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 11422–11427.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the*

2020 conference on empirical methods in natural language processing (EMNLP), pages 6769–6781.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*.

A. Kobren, B. Saha, and A. McCallum. 2019. Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1247–1257.

Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information fusion*, 98:101830.

Weiming Liu, Xiaolin Zheng, Chaochao Chen, Jiahe Xu, Xinting Liao, Fan Wang, Yanchao Tan, and Yew-Soon Ong. 2024. Reducing item discrepancy via differentially private robust embedding alignment for privacy-preserving cross domain recommendation. In *Forty-first International Conference on Machine Learning*.

Xiang Liu, Torsten Suel, and Nasir Memon. 2014. A robust model for paper reviewer assignment. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 25–32.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4969–4983.

Jiafei Lyu, Chenjia Bai, Jingwen Yang, Zongqing Lu, and Xiu Li. 2024. Cross-domain policy adaptation by capturing representation mismatch. *arXiv preprint arXiv:2405.15369*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781.

D. Mimno and A. McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 500–509.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. Mathbert: A pre-trained model for mathematical formula understanding. *Preprint*, arXiv:2105.00377.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Radim Rehurek. 2025. Gensim.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

M. A. Rodriguez and J. Bollen. 2008. An algorithm to determine peer-reviewers. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 319–328.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

I. Stelmakh, J. Wieting, G. Neubig, and N. B. Shah. 2023. A gold standard dataset for the reviewer assignment problem. ArXiv preprint arXiv:2303.16750.

Petar Stojanov, Zijian Li, Mingming Gong, Ruichu Cai, Jaime Carbonell, and Kun Zhang. 2021. Domain adaptation with invariant representation learning: What transformations to learn? *Advances in Neural Information Processing Systems*, 34:24791–24803.

Chenxi Sun, Hongyan Li, Yaliang Li, and Shenda Hong. 2023. Test: Text prototype aligned embedding to activate llm's ability for time series. *arXiv preprint arXiv:2308.08241*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Herbert Voß. 2014. Tex mathmode.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Yi-Di Wang, Shu Yan, Cheng-Xing Jia, Xu-Hui Liu, Xin-Wei Chen, Jia-Cheng Xu, Ziniu Li, and 1 others. 2025. A survey on large language models for mathematical reasoning. *arXiv preprint* arXiv:2506.08446.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2023. Mastering symbolic operations: Augmenting language models with compiled neural networks. *arXiv preprint arXiv:2304.01665*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

Appendix

A Related work

A.1 NLP-Based Manuscript Embedding

Recent advances in large language models (LLMs) and text embedding techniques have substantially improved automated scholarly document processing and semantic matching tasks Liu et al. (2014); Lin et al. (2023); Aksoy et al. (2025); Fu et al. (2025). However, early embedding methods such as Word2Vec Mikolov et al. (2013) and GloVe Pennington et al. (2014) often suffered from vocabulary mismatches caused by complex mathematical notation, and therefore frequently relied on external metadata or structural constraints to compensate for these limitations Dumais and Nielsen (1992); Mimno and McCallum (2007); Rodriguez and Bollen (2008); Kobren et al. (2019). Although recent corpus-scale LLMs have significantly enhanced the representation capability for scientific literature Karan et al. (2025); Zhu et al. (2025), general-purpose LLMs still exhibit limited reliability in highly technical and formula-intensive domains Stelmakh et al. (2023). Consequently, recent studies have increasingly focused on optimizing representation learning for scholarly documents through domain adaptation and fine-tuning strategies, aiming to overcome the limitations of general-purpose models and establish more robust scientific document embedding frameworks Stojanov et al. (2021); Sun et al. (2023); Muennighoff et al. (2024); Liu et al. (2024); Lyu et al. (2024); Cao et al. (2025).

A.2 Semantic Retrieval with Manuscript Embeddings

In information retrieval systems, particularly semantic retrieval, embedding techniques serve as a core mechanism for measuring the semantic relevance between a query and target scholarly manuscripts Karpukhin et al. (2020); Humeau et al. (2019). According to Muennighoff et al. (2023), the reliability of such retrieval systems largely depends on the quality of the underlying embeddings. High-quality document embeddings must capture contextual similarity beyond simple keyword matching by accurately reflecting semantic relationships in vector space Reimers and Gurevych (2019). In contrast, low-quality embeddings can produce false negatives by failing to retrieve relevant documents, thereby degrading both recall and precision Thakur et al. (2021).

These limitations become particularly evident in STEM literature, where documents are densely populated with mathematical expressions and domain-specific terminology. To address this challenge, SciBERT Beltagy et al. (2019), trained on scientific corpora, improves domain adaptability by incorporating STEM-specific vocabulary into its representations. Furthermore, studies such as MathBERT Peng et al. (2021) introduced formula-aware embedding techniques that model mathematical expressions as hierarchical operational structures rather than simple text streams, demonstrating improved semantic retrieval performance in STEM domains.

A.3 LaTeX-Aware Scientific Representation Learning

In STEM manuscripts, core ideas are often conveyed through complex mathematical expressions and symbolic notation rather than natural language alone. Since mathematical formulas function not merely as auxiliary elements but as independent semantic units that define the logical structure of a paper Lo et al. (2020), effective representation learning over raw LaTeX source code is essential for accurately capturing document semantics. Early approaches that ignored formulas or treated them as plain text strings resulted in substantial degradation in embedding quality Peng et al. (2021). Moreover, attempts to directly process raw LaTeX using general-purpose NLP models face two major limitations. **Synthetic Noise**: LaTeX-specific formatting commands such as \begin{equation} and \label are intermixed with semantically meaningful mathematical symbols, introducing noise that disrupts the model's attention mechanism Drozdov et al. (2022). **Vocabulary Mismatch**: Tokenizers of general-purpose language models often fail to properly recognize mathematical reserved tokens such as \alpha and \nabla, fragmenting them into meaningless character-level units and thereby distorting their original semantics Weng et al. (2023). To address these limitations, we propose a LaTeX-aware representation learning pipeline that explicitly preserves the structural information of STEM manuscripts while accurately capturing mathematical context.

B Technical appendices and supplementary material

All datasets, pretrained models, and external artifacts used in this work were employed in accordance with their intended research use and associated licenses. We do not redistribute original manuscripts or restricted resources. Any released artifacts derived from the data are intended solely for research purposes and remain consistent with the original access conditions. The data used in this work consists of publicly available scholarly manuscripts collected from academic repositories. We did not intentionally collect personally identifying information (PII) beyond publicly available author metadata contained in the original manuscripts. Since our framework focuses on semantic retrieval of scientific documents, the dataset is not expected to contain offensive or harmful content at a meaningful scale. We do not redistribute raw manuscripts, and any released artifacts are limited to derived representations, preprocessing scripts, or evaluation resources intended solely for research purposes.

B.1 Categories within Mathematics in arXiv

B.1.1 arXiv Mathematics

The arXiv Mathematics repository contains scholarly publications across a wide range of subfields, broadly reflecting research trends in mathematics. Table 1 lists the mathematics categories used in this study, along with the number of publications in each category at the time of data collection. The Description column provides a brief overview of the academic scope of each field. For clarity, we refer to each category using the notations defined in Table 6. We use publicly available scholarly documents and comply with the corresponding licenses and terms of use. Our released artifacts, if any, will include explicit license information and usage restrictions.

B.1.2 Example of test papers used in Experiments

The LaTeX source files collected from arXiv are preprocessed and structurally refined for use as experimental data in this study. In particular, for embedding-based experiments, each paper is used as a test input to compare ranking results across models. Table 7 provides example test papers used in Experiment 1, where the listed arXiv IDs denote the unique identifiers assigned to each publication. Using raw source–based text preserves the structural characteristics of STEM documents, making it well suited for evaluating embedding quality.

For experiments other than those in Experiment 1, the full list of test papers can be provided upon request. Given the large number of papers used in our experiments, including the entire list in the main paper is impractical; instead, we maintain the data for reproducibility and release it as needed.

Table 6: Mathematics categories used in this study, their notations, and the number of corresponding papers available on arXiv. Paper counts reflect the state at the time of data collection.

Category	Notation	Entry	Description
Commutative Algebra	AC	4,651	Commutative rings, modules, ideals, and their applications.
Algebraic Geometry	AG	16,029	Algebraic varieties, schemes, and related geometric structures.
Analysis of PDEs	AP	14,098	Partial differential equations and related analysis.
Algebraic Topology	AT	6,957	Homotopy theory, homology, cohomology, and related topics.
Classical Analysis and ODEs	CA	7,309	Classical analysis, special functions, and ordinary differential equations.
Combinatorics	CO	21,722	Discrete mathematics, combinatorial structures, and counting methods.
Category Theory	CT	4,674	Categories, functors, and categorical methods.
Complex Variables	CV	6,752	Complex analysis and several complex variables.
Differential Geometry	DG	13,055	Geometry of manifolds and differential geometric structures.
Dynamical Systems	DS	14,550	Dynamical systems, ergodic theory, and chaos.
Functional Analysis	FA	11,427	Functional analysis and operator theory.
General Mathematics	GM	261	Mathematical material of general interest, topics not covered elsewhere.
General Topology	GN	1978	Continuum theory, point-set topology, spaces with algebraic structure, foundations, dimension theory, local and global properties.
Group Theory	GR	8,332	Theory of groups and group-related structures.
Geometric Topology	GT	7,480	Topology of manifolds and low-dimensional topology.
History and Overview	HO	780	Biographies, philosophy of mathematics, mathematics education, recreational mathematics, communication of mathematics, ethics in mathematics.
Information Theory	IT	0	math.IT is an alias for cs.IT. Covers theoretical and experimental aspects of information theory and coding.
K-Theory and Homology	KT	3500	Algebraic and topological K-theory, relations with topology, commutative algebra, and operator algebras.
Logic	LO	3788	Logic, set theory, point-set topology, formal mathematics.
Matric Geometry	MG	5,2353	Euclidean, hyperbolic, discrete, convex, coarse geometry, comparisons in Riemannian geometry, symmetric spaces.
Numerical Analysis	NA	10,666	Numerical methods and computational mathematics.
Number Theory	NT	10,049	Properties of integers and arithmetic structures.
Operator Algebras	OA	4,683	Operator algebras and noncommutative structures.
Optimization and Control	OC	18,004	Optimization theory and control systems.
Probability	PR	18,005	Probability theory and stochastic processes.
Quantum Algebra	QA	9,042	Algebraic structures related to quantum theory.
Rings and Algebras	RA	6,766	Non-commutative rings and algebras, non-associative algebras, universal algebra and lattice theory, linear algebra, semigroups.
Representation Theory	RT	11,332	Representation theory of groups and algebras.
Symplectic Geometry	SG	4,348	Symplectic and contact geometry.
Spectral Theory	SP	5,149	Spectral theory of operators.
Statistics Theory	ST	8,837	Statistical theory and inference.
General Mathematics, History and Overview, Logic	G1	4,829	Foundational and general-interest aspects of mathematics and its communication.
General Topology, K-Theory and Homology	G2	5,478	Topology and its algebraic, foundational, and structural extensions.

Table 7: List of test papers used in experiments for comparison of embedding models in Section 3.3

arXiv id	subject	title
2502.02473	NA	Parareal Algorithms for Stochastic Maxwell Equations Driven by Multiplicative Noise
2502.01938	NA	A Kolmogorov High Order Deep Neural Network for High Frequency Partial Differential Equations in High Dimensions
2502.00953	NA	Partitioned solution strategies for coupled BEM-FEM acoustic fluid-structure interaction problems
2502.00533	NA	Anderson acceleration of a Picard solver for the Oldroyd-B model of viscoelastic fluids
2502.0015	NA	Optimal sensor placement under model uncertainty in the weak-constraint 4D-Var framework
2501.18305	NA	A hybrid two-level weighted Schwartz method for time-harmonic Maxwell equations
2501.18013	NA	A comprehensive numerical investigation of a coupled mathematical model of neuronal excitability
2501.17956	NA	The Numerical Approximation of Caputo Fractional Derivative of Higher Orders Using A Shifted Gegenbauer Pseudospectral Method: Two-Point Boundary Value Problems of the Bagley Torvik Type Case Study
2205.03906	CT	Dynamic Operads, Dynamic Categories: From Deep Learning to Prediction Markets
2205.0364	CT	Weakening and Iterating Laws using String Diagrams
2204.09584	CT	An orthogonal approach to algebraic weak factorisation systems
2204.04274	CT	Rewriting for Symmetric Monoidal Categories with Commutative (Co)Monoid Structure
2204.03389	CT	Koszul duality for categories with a fixed object set
2203.16345	CT	An Algebraic Framework for Structured Epidemic Modeling
2203.11845	CT	Dualities in the complicial model of $infty$-categories
2203.00072	CT	Parametrized and equivariant higher algebra
2303.0827	AT	Meta-Diagrams for 2-Parameter Persistence
2303.0826	AT	On the Picard group of the stable module category for infinite groups
2303.01501	AT	Stability and Machine Learning Applications of Persistent Homology Using the Delaunay-Rips Complex
2302.12659	AT	On the motivic Segal conjecture
2302.05591	AT	2-local unstable homotopy groups of indecomposable $mathbfA_3^2$ -complexes
2301.07404	AT	Large simplicial complexes: Universality, Randomness, and Ampleness
2301.06824	AT	Equivariant Intrinsic Formality
2301.06776	AT	On the Homotopy Groups of the Suspended Quaternionic Projective Plane and Applications
2502.04118	ST	Maximum Likelihood Estimation of the Parameters of Matrix Variate Symmetric Laplace Distribution
2502.02726	ST	Multimarginal Schrödinger Barycenter
2502.01168	ST	On the Private Estimation of Smooth Transport Maps
2501.15725	ST	Eigenvector fluctuations and limit results for random graphs with infinite rank kernels

Continued on next page

arXiv id	subject	title
2501.12747	ST	Singular leaning coefficients and efficiency in learning theory
2501.10538	ST	Universality of Benign Overfitting in Binary Linear Classification
2501.09483	ST	Semiparametrics via parametrics and contiguity
2501.09336	ST	Estimating shared subspace with AJIVE: the power and limitation of multiple data matrices
2502.03288	OC	Optimal Mediated Graphs: The role of Combinatorics in Conic Optimization
2502.02783	OC	Runway capacity expansion planning for public airports under demand uncertainty
2502.01819	OC	Score as Action: Fine-Tuning Diffusion Generative Models by Continuous-time Reinforcement Learning
2502.01076	OC	qNBO: quasi-Newton Meets Bilevel Optimization
2502.00470	OC	Distributed Primal-Dual Algorithms: Unification, Connections, and Insights
2501.19315	OC	Differentially Private Linear Programming: Reduced Sub-Optimality and Guaranteed Constraint Satisfaction
2501.19000	OC	Bilevel optimization for the deployment of refuelling stations for electric vehicles on road networks
2501.18773	OC	Beyond Short Steps in Frank-Wolfe Algorithms
2405.17292	AG	Deformations with fibre constancy
2405.16936	AG	Construction of birational trilinear volumes via tensor rank criteria
2405.15195	AG	Dynamical degrees of automorphisms of K3 surfaces with Picard number 2
2405.14167	AG	Sesquilinear pairings on elliptic curves
2405.14084	AG	Algebraic Independence of Special Points on Shimura Varieties
2405.1262	AG	On the smoothability problem with rational coefficients
2405.12189	AG	The sign of scalar curvature on Kähler blowups
2405.10797	AG	K-stability of special Gushel-Mukai manifolds
2501.07522	GR	The Lodha–Moore groups and their n-adic generalizations are not SCY
2501.07438	GR	Bounded cohomology and scl of verbal wreath products
2501.00689	GR	Finite groups with exactly two nonlinear irreducible p-Brauer characters
2412.1441	GR	The proper geometric dimension of the mapping class group of an orientable surface with punctures
2412.14345	GR	Coxeter-type quotients of surface braid groups
2412.14081	GR	Subgroups of CAT(0) groups, exotic finiteness properties and non-QI-embeddings into linear groups
2412.12195	GR	Rewriting in Artin groups without A_3 or B_3 subdiagrams
2411.19837	GR	The connectivity of the normalising and permuting graph of a finite soluble group
2302.1293	CV	A Riemann–Hilbert approach to computing the inverse spectral map for measures supported on disjoint intervals
2302.1147	CV	Surjective morphisms from affine space to its Zariski open subsets
2302.1108	CV	A relative Nadel-type vanishing theorem
2302.07531	CV	On the Theory of Bergman Spaces on Homogeneous Siegel Domains
2302.0435	CV	One-parameter families of conformal mappings of the half-plane onto polygonal domains with several slits

Continued on next page

arXiv id	subject	title
2302.00889	CV	On a Class of Non-Univalent functions Associated with a Parabolic Region
2301.118	CV	Commuting Toeplitz operators and moment maps on Cartan domains of type III
2301.09526	CV	On the analytic version of the Mitiagin-DeLeeuw-Mirkhil non-inequality on bi-disc
2409.01796	CO	An exploration of the balance game
2409.00846	CO	Undecidability of Translational Tiling of the 4-dimensional Space with a Set of 4 Polyhypercubes
2409.00697	CO	Grundy Packing Coloring of Graphs
2409.00682	CO	Higher Order Bipartiteness vs Bi-Partitioning in Simplicial Complexes
2408.17283	CO	Non-Promise Version of Unique Sink Orientations
2408.17232	CO	Bubbles in Linear Chord Diagrams: Bridges and Crystallized Diagrams
2408.17212	CO	On log-concavity of the number of orbits in commuting tuples of permutations
2408.17191	CO	Tournament transitivity of graphs

B.2 Feature embedding with LLMs

B.2.1 SciBERT

SciBERT is trained on a large corpus of approximately 1.14 million scientific papers collected from Semantic Scholar, totaling about 3.1 billion tokens which are comparable in scale to the original BERT training data. A key innovation of SciBERT is its custom WordPiece vocabulary (SciVocab), consisting of approximately 30,000 tokens. Notably, SciVocab overlaps with the standard BERT vocabulary by only 42%, indicating strong specialization for STEM-domain linguistic patterns.

As a result, SciBERT captures rich contextual nuances of mathematical formulas and technical terminology commonly found in scholarly literature. Empirical studies have demonstrated that SciBERT consistently outperforms general BERT models on a wide range of scientific NLP downstream tasks, with particularly strong embedding quality in domains such as biomedicine and computer science.

In contrast, standard BERT models pretrained on general purpose corpora (e.g., Wikipedia and BookCorpus) exhibit limited understanding of science specific vocabularies and mathematical expressions. This limitation arises from their generic vocabularies, which inadequately cover complex STEM terminology, leading to suboptimal representations of technical concepts.

Therefore, to ensure maximal compatibility with the technical nature of our dataset, we adopt SciBERT, which is pretrained and optimized for STEM literature, as our feature embedding model.

B.2.2 Using SciBERT as an Embedding Model

The journal database consists of recently published articles from the target journal and is continuously updated as new manuscripts appear. At scheduled intervals, the embedding model management system fine-tunes SciBERT using this evolving corpus, producing a domain adapted embedding model that more precisely reflects the journal's specialized research areas. The fine-tuned model then replaces the generic SciBERT in the feature embedding module, mapping both submitted manuscripts and candidate document corpus into a journal specific semantic vector space.

Although the general SciBERT model is trained across broad scientific disciplines, it may not fully capture journal specific subfield vocabularies, frequently used equation forms, or notational conventions. Fine-tuning integrates this journal specific knowledge into the embedding space, enabling more accurate semantic similarity estimation between submitted manuscripts and publications. Moreover, sensitivity to recent research trends is substantially improved: as the journal database is continuously refreshed, the fine-tuned model rapidly incorporates emerging topics, newly coined terminology, and evolving methodological patterns. This allows semantic relevance assessment to reflect up to date research directions that static pretrained models may miss.

Fine-tuning also improves the granularity of similarity evaluation. Even within the same category, papers may address distinct subtopics; however, such documents often cluster closely in embedding space simply because they share a category label, hindering precise differentiation. By capturing fine-grained semantic nuances within categories, our approach enables granular topic identification at a more detailed level. This leads to more accurate similarity computation in the relevance comparison and yields tangible improvements in relevant ranking and final ranking quality.

B.3 Metrics

B.3.1 Similarity Measure

The selection of an appropriate similarity metric is a critical design decision in semantic retrieval systems, as an ill-suited measure can misalign the high-dimensional representations of semantically related documents, thereby significantly degrading retrieval performance. To identify the most robust metric for formula-dense STEM manuscripts, we conducted a comparative analysis of various similarity measures. The empirical results, summarized in Table Table 8, illustrate the retrieval effectiveness across different metrics evaluated on a consistent benchmarking dataset.

Cosine Similarity: This metric primarily captures the semantic proximity between documents by measuring the angular distance between their latent vector representations. While effective at identifying thematic overlap, it often fails to account for the exact keyword-level matches or the structural recurrence of specific mathematical tokens that are vital in STEM contexts. Jaccard Similarity: In contrast, Jaccard similarity focuses on set-level intersectional characteristics, reflecting the degree of overlap in specific terminology and formulaic symbols. Although it provides a high degree of pre-

cision for keyword-heavy documents, it lacks the ability to capture deeper contextual nuances and synonymic relations inherent in natural language.

To leverage the complementary advantages of both approaches, we use Scaled MaxSim, a hybrid similarity measure designed to integrate global semantic context with local symbolic fidelity. As defined in Equation (1), Scaled MaxSim applies a min-max normalization and a softmax-based weighting to both cosine and Jaccard scores. This integration ensures that the system maintains high sensitivity to technical sub-fields while simultaneously capturing the broader research context. Our results demonstrate that Scaled MaxSim consistently yields superior discrimination in ranking relevant documents compared to single-metric baselines.

To evaluate the effectiveness of our retrieval system, we conducted a comparative analysis using three different similarity measurement approaches: Cosine similarity, Jaccard similarity, and scaled Max-similarity, while maintaining the same data pool and embedding space. Our empirical results demonstrate significant performance variances depending on the choice of similarity metric, even when utilizing identical representations.

As shown in Table 4, Table 8, and Table 9 Max-similarity consistently outperforms both Cosine and Jaccard similarities across all key information retrieval metrics, including Precision, MAP, and NDCG. For instance, Max-similarity achieved a Precision@50 of 0.6360 and NDCG@100 of 0.6874, whereas Cosine similarity and Jaccard similarity yielded lower scores (Precision@50: 0.6120 and 0.5060; NDCG@100: 0.6558 and 0.5186, respectively).

The relatively poor performance of Cosine and Jaccard similarities suggests that these measures fail to assign sufficiently high similarity scores to items within the same category in our high-dimensional embedding space. This indicates a "representation-metric mismatch" where standard distance measures do not fully capture the nuanced semantic overlaps present in scientific documents. Specifically, Cosine similarity tends to be overly sensitive to the global orientation of vectors, while Jaccard similarity—primarily designed for sets—often ignores the rich semantic information embedded in dense vectors.

In contrast, scaled Max-similarity effectively identifies the most relevant local semantic features, leading to higher discriminative power in categorizing similar documents. By selecting the maximum similarity across multiple views or components, the model can more accurately measure the degree of relatedness between complex academic texts. Consequently, based on its superior empirical performance and its ability to better align with the underlying semantic structure of our data, we adopted Max-similarity as our primary metric for the retrieval tasks.

B.3.2 Evaluation Metrics

Below we provide detailed definitions of the evaluation metrics used in our experiments.

Precision@K measures the proportion of relevant documents among the top-K recommended candidates:

$$\text{Precision@}K = \frac{|\text{Relevant documents in top } K|}{K}. \tag{2}$$

MAP@K (Mean Average Precision at K) is defined as the mean of AP@K over all test manuscripts within a category. AP@K for a single manuscript is computed as the average precision at the ranks where relevant documents appear:

$$\text{AP@}K = \frac{\sum_{i=1}^{K} P(i) \cdot \text{rel}(i)}{\text{Number of relevant documents}}, \tag{3}$$

where $P(i)$ denotes the precision at position i, and $\text{rel}(i) = 1$ if the document at rank i is relevant, and 0 otherwise.

MAP@K is then given by

$$\text{MAP@}K = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP@}K(q), \tag{4}$$

where Q denotes the total number of test manuscripts.

nDCG@K (Normalized Discounted Cumulative Gain) evaluates ranking quality by assigning higher weights to top-ranked items. For fair comparison with AP, we adopt binary relevance (0/1):

$$\begin{aligned} \text{nDCG@}K &= \frac{\text{DCG@}K}{\text{IDCG@}K}, \\ \text{DCG@}K &= \sum_{i=1}^{K} \frac{2^{\text{rel}(i)} - 1}{\log_2(i+1)}. \end{aligned} \tag{5}$$

Here, IDCG@K denotes the DCG of the ideal ranking for the given manuscript.

Pairwise-F1@K evaluates how well relevant documents are ranked ahead of non-relevant documents

Table 8: Overall Performance Metrics by Similarity Method1

Category	Method	Ap@100	Ap@50	nDCG@100	nDCG@50	Prec@100	Prec@50
AC	cos	0.1477	0.1702	0.3838	0.3863	0.3400	0.3000
	jac	0.0685	0.0908	0.2338	0.2657	0.2200	0.2600
	max	0.5177	0.6222	0.7089	0.7887	0.6700	0.7600
AP	cos	0.3433	0.3610	0.5916	0.5874	0.5600	0.5200
	jac	0.3586	0.5054	0.5567	0.6687	0.5200	0.6600
	max	0.4920	0.5533	0.7026	0.7432	0.6700	0.7000
AG	cos	0.2218	0.3194	0.4491	0.5422	0.3800	0.4600
	jac	0.0979	0.1342	0.2885	0.3312	0.2600	0.3000
	max	0.3533	0.4953	0.5717	0.6803	0.5000	0.6000
AT	cos	0.0381	0.0438	0.1368	0.1408	0.0600	0.0600
	jac	0.3855	0.5484	0.5891	0.7303	0.5300	0.6400
	max	0.6295	0.7473	0.7818	0.8710	0.7100	0.8800
CA	cos	0.4627	0.4729	0.6389	0.6212	0.5800	0.5600
	jac	0.5601	0.6496	0.7323	0.8123	0.7100	0.7600
	max	0.5510	0.5681	0.7161	0.7483	0.6900	0.7000
CO	cos	0.6494	0.6781	0.7788	0.8130	0.6900	0.6800
	jac	0.7203	0.7744	0.8415	0.8931	0.8400	0.8000
	max	0.9328	0.9490	0.9654	0.9782	0.9600	0.9999
CV	cos	0.5739	0.6120	0.7073	0.7408	0.6100	0.5800
	jac	0.6095	0.7088	0.7432	0.8258	0.6900	0.7800
	max	0.8655	0.9168	0.9272	0.9620	0.9100	0.9800
DG	cos	0.7400	0.6700	0.6360	0.5325	0.7883	0.7171
	jac	0.5000	0.4500	0.3141	0.2572	0.5109	0.4704
	max	0.4400	0.4500	0.2632	0.2395	0.5024	0.4879
DS	cos	0.2800	0.1900	0.2171	0.1189	0.4070	0.2882
	jac	0.3800	0.3100	0.2099	0.1465	0.4170	0.3493
	max	0.6000	0.5300	0.4747	0.3644	0.6702	0.5890
FA	cos	0.8000	0.5900	0.7280	0.4976	0.8434	0.6669
	jac	0.5000	0.5200	0.3053	0.2968	0.5401	0.5416
	max	0.8400	0.8000	0.7448	0.6847	0.8600	0.8221
GR	cos	0.2600	0.2200	0.1260	0.0849	0.3359	0.2761
	jac	0.6600	0.6400	0.4957	0.4501	0.6637	0.6472
	max	0.9200	0.8200	0.8641	0.7504	0.9323	0.8525
GT	cos	0.6200	0.5700	0.5076	0.4060	0.6920	0.6254
	jac	0.8200	0.7700	0.6909	0.6306	0.8366	0.7919
	max	0.9200	0.9100	0.8861	0.8556	0.9378	0.9234
MG	cos	0.4200	0.3900	0.3209	0.2346	0.5278	0.4639
	jac	0.5800	0.5100	0.4171	0.3380	0.5989	0.5413
	max	0.7400	0.6900	0.5843	0.5232	0.7715	0.7219
NA	cos	0.8400	0.8800	0.7686	0.7860	0.8725	0.8914
	jac	0.6600	0.6500	0.4476	0.4400	0.6387	0.6408
	max	0.9400	0.9300	0.9299	0.8926	0.9569	0.9424
NT	cos	0.5400	0.5600	0.4167	0.3750	0.6245	0.6092
	jac	0.5800	0.5900	0.3687	0.3606	0.6007	0.6003
	max	0.9000	0.8500	0.8458	0.7688	0.9187	0.8727
OA	cos	0.3400	0.2500	0.1879	0.1172	0.4127	0.3157
	jac	0.7200	0.6700	0.5441	0.4891	0.7066	0.6741
	max	0.8600	0.8100	0.7476	0.6897	0.8681	0.8265
OC	cos	0.8000	0.7600	0.6521	0.6095	0.8121	0.7775
	jac	0.8600	0.9000	0.7895	0.8092	0.8848	0.9057
	max	0.7600	0.7400	0.6621	0.6103	0.8061	0.7752
PR	cos	0.8400	0.7200	0.7567	0.6048	0.8646	0.7621
	jac	0.6000	0.4900	0.3633	0.2795	0.5838	0.5042
	max	0.6800	0.6400	0.5693	0.4971	0.7411	0.6908
QA	cos	0.0400	0.0200	0.0400	0.0200	0.1265	0.0779
	jac	0.3000	0.2900	0.1182	0.0989	0.3137	0.2994
	max	0.6400	0.6000	0.4067	0.3762	0.6516	0.6163
RA	cos	0.3800	0.3100	0.1910	0.1375	0.4360	0.3612
	jac	0.3200	0.3000	0.1189	0.1016	0.3189	0.3030
	max	0.5600	0.6200	0.4046	0.4030	0.6265	0.6461
RT	cos	0.4400	0.4800	0.2447	0.2436	0.4924	0.5027
	jac	0.5800	0.4700	0.3590	0.2740	0.5677	0.4885
	max	0.8200	0.7500	0.7424	0.6299	0.8557	0.7869
SG	cos	0.6800	0.5500	0.5904	0.4214	0.7481	0.6220
	jac	0.4200	0.3800	0.2240	0.1826	0.4330	0.3983
	max	0.6800	0.6100	0.5476	0.4501	0.7327	0.6596

Table 9: Overall Performance Metrics by Similarity Method2

Category	Method	Ap@100	Ap@50	nDCG@100	nDCG@50	Prec@100	Prec@50
SP	cos	0.3400	0.3100	0.2553	0.1736	0.4575	0.3895
	jac	0.6400	0.5900	0.4337	0.3742	0.6608	0.6121
	max	0.8800	0.8200	0.8534	0.7484	0.9116	0.8536
ST	cos	0.8000	0.6300	0.6910	0.5081	0.8306	0.6888
	jac	0.6000	0.5700	0.3888	0.3548	0.5800	0.5654
	max	0.9400	0.9200	0.8930	0.8682	0.9478	0.9301
G1	cos	0.3200	0.2600	0.1881	0.1243	0.4091	0.3297
	jac	0.3600	0.2700	0.2126	0.1326	0.4427	0.3409
	max	0.8000	0.6500	0.7244	0.5436	0.8423	0.7118
G2	cos	0.4400	0.2700	0.2572	0.1446	0.5035	0.3480
	jac	0.4400	0.2600	0.3308	0.1769	0.5411	0.3630
	max	0.6600	0.5700	0.5490	0.4219	0.7255	0.6317
Average	cos	0.5015	0.4463	0.3893	0.3091	0.5723	0.5030
	jac	0.5704	0.5193	0.3954	0.3421	0.5818	0.5384
	max	0.7822	0.7270	0.6889	0.6079	0.8163	0.7613

within the top-K list. Among all (relevant, non-relevant) pairs in the top-K, correctly ordered pairs are counted as correct, from which pairwise precision and recall are computed:

$$\begin{aligned} P_K^{\text{pw}} &= \frac{C_K}{A_K}, \\ R_K^{\text{pw}} &= \frac{C_K}{B_K}, \\ F_K^{\text{pw}} &= \frac{2P_K^{\text{pw}} R_K^{\text{pw}}}{P_K^{\text{pw}} + R_K^{\text{pw}}}, \end{aligned} \tag{6}$$

where C_K is the number of correctly ordered pairs, A_K is the number of comparable pairs produced by the model, and B_K is the total number of ground truth pairs.

C Statistics for experiments

Table 10: Statistical significance of SciBERT over BERT across retrieval metrics. The reported p-values are computed under the one-sided hypothesis that SciBERT achieves higher performance than BERT.

Metric	p**-value**
Precision@50	0.028^*
Precision@100	0.021^*
AP@50	0.045^*
AP@100	0.031^*
nDCG@50	0.033^*
nDCG@100	0.024^*
Pairwise-F1@50	0.082
Pairwise-F1@100	0.021^*

* denotes statistical significance at $p < 0.05$.

To further examine whether the performance gains of SciBERT over BERT are statistically reliable, we conduct a one-sided significance test for each retrieval metric, where the alternative hypothesis assumes that SciBERT achieves higher performance than BERT. As shown in Table 10, SciBERT shows statistically significant improvements over BERT for most metrics, including Precision@50 ($p = 0.028$), Precision@100 ($p = 0.021$), AP@50 ($p = 0.045$), AP@100 ($p = 0.031$), nDCG@50 ($p = 0.033$), nDCG@100 ($p = 0.024$), and Pairwise-F1@100 ($p = 0.021$). These results indicate that the observed improvements are unlikely to be due to random variation alone. Pairwise-F1@50 also shows a positive trend in favor of SciBERT, although the difference is not statistically significant at the conventional $p < 0.05$ threshold ($p = 0.082$).

Table 11: Robust analysis for precision and nDCG across all mathematical categories. Values denote mean and standard deviation.

SciBERT				
Category	**P@50**	**P@100**	**nDCG@50**	**nDCG@100**
AC	0.440 ± 0.012	0.369 ± 0.012	0.631 ± 0.075	0.646 ± 0.086
AG	0.398 ± 0.026	0.390 ± 0.034	0.436 ± 0.032	0.468 ± 0.045
AP	0.593 ± 0.027	0.565 ± 0.020	0.624 ± 0.028	0.623 ± 0.036
AT	0.590 ± 0.026	0.560 ± 0.027	0.639 ± 0.033	0.634 ± 0.038
CA	0.515 ± 0.023	0.475 ± 0.021	0.551 ± 0.022	0.573 ± 0.042
CO	0.672 ± 0.037	0.644 ± 0.033	0.707 ± 0.045	0.710 ± 0.055
CT	0.732 ± 0.011	0.705 ± 0.012	0.745 ± 0.032	0.718 ± 0.038
CV	0.470 ± 0.020	0.443 ± 0.029	0.511 ± 0.023	0.522 ± 0.035
DG	0.675 ± 0.021	0.618 ± 0.022	0.709 ± 0.017	0.665 ± 0.017
DS	0.515 ± 0.006	0.496 ± 0.013	0.566 ± 0.004	0.601 ± 0.011
FA	0.540 ± 0.026	0.496 ± 0.031	0.591 ± 0.042	0.606 ± 0.044
G1	0.562 ± 0.076	0.486 ± 0.077	0.619 ± 0.080	0.613 ± 0.086
G2	0.425 ± 0.077	0.388 ± 0.074	0.468 ± 0.080	0.500 ± 0.095
GR	0.675 ± 0.014	0.650 ± 0.013	0.713 ± 0.011	0.687 ± 0.013
GT	0.470 ± 0.052	0.468 ± 0.049	0.507 ± 0.053	0.515 ± 0.055
MG	0.615 ± 0.010	0.588 ± 0.010	0.649 ± 0.016	0.628 ± 0.021
NA	0.780 ± 0.039	0.754 ± 0.041	0.801 ± 0.043	0.793 ± 0.044
NT	0.623 ± 0.045	0.589 ± 0.036	0.669 ± 0.044	0.635 ± 0.040
OA	0.708 ± 0.035	0.650 ± 0.043	0.736 ± 0.034	0.702 ± 0.034
OC	0.840 ± 0.003	0.790 ± 0.008	0.859 ± 0.004	0.814 ± 0.007
PR	0.587 ± 0.033	0.569 ± 0.020	0.629 ± 0.023	0.631 ± 0.011
QA	0.647 ± 0.044	0.598 ± 0.046	0.695 ± 0.048	0.638 ± 0.046
RA	0.535 ± 0.037	0.501 ± 0.042	0.573 ± 0.038	0.546 ± 0.044
RT	0.698 ± 0.029	0.614 ± 0.021	0.727 ± 0.034	0.680 ± 0.031
SG	0.728 ± 0.058	0.685 ± 0.059	0.761 ± 0.061	0.716 ± 0.055
SP	0.555 ± 0.020	0.529 ± 0.021	0.609 ± 0.022	0.616 ± 0.038
ST	0.788 ± 0.007	0.768 ± 0.016	0.816 ± 0.006	0.795 ± 0.014
BERT				
Category	**P@50**	**P@100**	**nDCG@50**	**nDCG@100**
AC	0.375 ± 0.022	0.359 ± 0.013	0.520 ± 0.029	0.547 ± 0.020
AG	0.398 ± 0.030	0.362 ± 0.034	0.429 ± 0.031	0.441 ± 0.043
AP	0.510 ± 0.046	0.511 ± 0.062	0.541 ± 0.047	0.572 ± 0.070
AT	0.505 ± 0.056	0.455 ± 0.053	0.540 ± 0.063	0.528 ± 0.062
CA	0.528 ± 0.011	0.472 ± 0.023	0.585 ± 0.016	0.587 ± 0.022
CO	0.577 ± 0.048	0.579 ± 0.048	0.599 ± 0.053	0.626 ± 0.053
CT	0.635 ± 0.020	0.585 ± 0.011	0.664 ± 0.007	0.617 ± 0.007
CV	0.425 ± 0.016	0.372 ± 0.015	0.452 ± 0.021	0.455 ± 0.034
DG	0.638 ± 0.033	0.604 ± 0.025	0.670 ± 0.028	0.650 ± 0.019
DS	0.412 ± 0.058	0.361 ± 0.034	0.468 ± 0.058	0.493 ± 0.051
FA	0.427 ± 0.035	0.411 ± 0.033	0.449 ± 0.043	0.475 ± 0.048
G1	0.383 ± 0.049	0.315 ± 0.040	0.481 ± 0.066	0.529 ± 0.072
G2	0.290 ± 0.029	0.256 ± 0.030	0.322 ± 0.032	0.298 ± 0.042
GR	0.665 ± 0.022	0.654 ± 0.033	0.713 ± 0.029	0.716 ± 0.047
GT	0.312 ± 0.057	0.323 ± 0.068	0.343 ± 0.068	0.373 ± 0.089
MG	0.630 ± 0.043	0.581 ± 0.047	0.657 ± 0.042	0.610 ± 0.042
NA	0.767 ± 0.034	0.729 ± 0.044	0.796 ± 0.041	0.773 ± 0.055
NT	0.627 ± 0.040	0.593 ± 0.027	0.657 ± 0.036	0.619 ± 0.032
OA	0.720 ± 0.065	0.672 ± 0.055	0.756 ± 0.069	0.721 ± 0.065
OC	0.812 ± 0.008	0.764 ± 0.004	0.843 ± 0.009	0.795 ± 0.001
PR	0.568 ± 0.035	0.519 ± 0.034	0.616 ± 0.040	0.622 ± 0.044
QA	0.547 ± 0.021	0.522 ± 0.039	0.596 ± 0.031	0.597 ± 0.039
RA	0.443 ± 0.092	0.399 ± 0.095	0.477 ± 0.096	0.447 ± 0.099
RT	0.507 ± 0.020	0.487 ± 0.014	0.556 ± 0.015	0.544 ± 0.006
SG	0.700 ± 0.016	0.664 ± 0.010	0.730 ± 0.013	0.695 ± 0.008
SP	0.598 ± 0.049	0.560 ± 0.042	0.637 ± 0.052	0.626 ± 0.051
ST	0.718 ± 0.026	0.672 ± 0.023	0.759 ± 0.021	0.717 ± 0.019

Table 12: Robust analysis for MAP and Pairwise-F1 across all mathematical categories. Values denote mean and standard deviation.

SciBERT				
Category	**MAP@50**	**MAP@100**	**Pairwise-F1@50**	**Pairwise-F1@100**
AC	0.673 ± 0.085	0.615 ± 0.076	0.429 ± 0.013	0.362 ± 0.013
AG	0.492 ± 0.040	0.446 ± 0.035	0.388 ± 0.024	0.385 ± 0.033
AP	0.665 ± 0.029	0.624 ± 0.027	0.584 ± 0.026	0.561 ± 0.019
AT	0.714 ± 0.044	0.652 ± 0.036	0.582 ± 0.026	0.556 ± 0.027
CA	0.616 ± 0.022	0.563 ± 0.019	0.505 ± 0.022	0.470 ± 0.021
CO	0.733 ± 0.043	0.705 ± 0.041	0.666 ± 0.036	0.640 ± 0.033
CT	0.761 ± 0.037	0.737 ± 0.027	0.605 ± 0.067	0.703 ± 0.011
CV	0.571 ± 0.029	0.521 ± 0.027	0.462 ± 0.053	0.438 ± 0.029
DG	0.758 ± 0.011	0.708 ± 0.016	0.668 ± 0.022	0.614 ± 0.023
DS	0.658 ± 0.005	0.587 ± 0.005	0.505 ± 0.006	0.491 ± 0.013
FA	0.659 ± 0.055	0.597 ± 0.039	0.531 ± 0.026	0.491 ± 0.031
G1	0.720 ± 0.083	0.639 ± 0.076	0.554 ± 0.074	0.481 ± 0.077
G2	0.539 ± 0.086	0.480 ± 0.079	0.416 ± 0.075	0.383 ± 0.073
GR	0.760 ± 0.007	0.717 ± 0.010	0.668 ± 0.014	0.646 ± 0.014
GT	0.553 ± 0.051	0.512 ± 0.051	0.462 ± 0.052	0.463 ± 0.049
MG	0.697 ± 0.017	0.656 ± 0.012	0.610 ± 0.009	0.585 ± 0.009
NA	0.833 ± 0.049	0.804 ± 0.046	0.664 ± 0.089	0.752 ± 0.040
NT	0.735 ± 0.047	0.676 ± 0.045	0.615 ± 0.045	0.585 ± 0.036
OA	0.771 ± 0.033	0.733 ± 0.035	0.702 ± 0.036	0.646 ± 0.043
OC	0.892 ± 0.006	0.855 ± 0.004	0.837 ± 0.003	0.788 ± 0.008
PR	0.706 ± 0.007	0.642 ± 0.015	0.579 ± 0.033	0.564 ± 0.020
QA	0.769 ± 0.053	0.702 ± 0.050	0.640 ± 0.044	0.593 ± 0.046
RA	0.639 ± 0.041	0.588 ± 0.034	0.528 ± 0.037	0.497 ± 0.042
RT	0.758 ± 0.037	0.712 ± 0.031	0.691 ± 0.028	0.610 ± 0.020
SG	0.805 ± 0.067	0.757 ± 0.064	0.722 ± 0.058	0.682 ± 0.059
SP	0.702 ± 0.037	0.626 ± 0.032	0.546 ± 0.020	0.524 ± 0.020
ST	0.855 ± 0.010	0.821 ± 0.008	0.658 ± 0.007	0.765 ± 0.016
BERT				
Category	**MAP@50**	**MAP@100**	**Pairwise-F1@50**	**Pairwise-F1@100**
AC	0.547 ± 0.030	0.521 ± 0.039	0.367 ± 0.021	0.355 ± 0.013
AG	0.468 ± 0.033	0.432 ± 0.033	0.390 ± 0.029	0.359 ± 0.033
AP	0.586 ± 0.047	0.548 ± 0.051	0.500 ± 0.045	0.506 ± 0.061
AT	0.594 ± 0.079	0.544 ± 0.066	0.497 ± 0.055	0.451 ± 0.053
CA	0.677 ± 0.024	0.601 ± 0.018	0.518 ± 0.012	0.467 ± 0.023
CO	0.628 ± 0.060	0.602 ± 0.055	0.571 ± 0.047	0.576 ± 0.048
CT	0.696 ± 0.019	0.657 ± 0.005	0.630 ± 0.022	0.582 ± 0.011
CV	0.486 ± 0.031	0.449 ± 0.026	0.418 ± 0.015	0.369 ± 0.015
DG	0.722 ± 0.024	0.676 ± 0.025	0.630 ± 0.033	0.600 ± 0.026
DS	0.576 ± 0.052	0.496 ± 0.053	0.403 ± 0.057	0.356 ± 0.033
FA	0.489 ± 0.052	0.451 ± 0.043	0.418 ± 0.035	0.407 ± 0.032
G1	0.567 ± 0.076	0.485 ± 0.062	0.372 ± 0.048	0.309 ± 0.040
G2	0.373 ± 0.033	0.328 ± 0.031	0.286 ± 0.028	0.254 ± 0.030
GR	0.789 ± 0.042	0.733 ± 0.037	0.658 ± 0.084	0.650 ± 0.032
GT	0.401 ± 0.091	0.360 ± 0.078	0.306 ± 0.054	0.319 ± 0.067
MG	0.694 ± 0.040	0.654 ± 0.042	0.625 ± 0.042	0.578 ± 0.047
NA	0.839 ± 0.055	0.800 ± 0.046	0.762 ± 0.033	0.726 ± 0.044
NT	0.697 ± 0.034	0.659 ± 0.031	0.622 ± 0.040	0.590 ± 0.027
OA	0.815 ± 0.077	0.765 ± 0.071	0.714 ± 0.065	0.669 ± 0.054
OC	0.885 ± 0.010	0.841 ± 0.006	0.809 ± 0.008	0.761 ± 0.004
PR	0.687 ± 0.049	0.625 ± 0.041	0.559 ± 0.034	0.514 ± 0.034
QA	0.671 ± 0.057	0.613 ± 0.037	0.538 ± 0.020	0.518 ± 0.038
RA	0.522 ± 0.101	0.479 ± 0.097	0.436 ± 0.091	0.395 ± 0.094
RT	0.617 ± 0.011	0.556 ± 0.016	0.500 ± 0.021	0.484 ± 0.014
SG	0.766 ± 0.008	0.727 ± 0.010	0.694 ± 0.016	0.660 ± 0.010
SP	0.693 ± 0.051	0.638 ± 0.048	0.589 ± 0.049	0.556 ± 0.042
ST	0.822 ± 0.014	0.763 ± 0.018	0.712 ± 0.027	0.669 ± 0.023

D Overall Experimental Results

All experiments were performed on an Intel(R) Xeon(R) Gold 6348 CPU and 3 NVIDIA RTX PRO 4000 Blackwell GPUs (24GB VRAM each).

Human Participant Instructions. This study did not involve human-subject experiments, crowdworkers, or annotators for constructing relevance labels. The retrieval ground truth was automatically derived from arXiv category labels. Expert validation was used only to confirm the representativeness of selected evaluation manuscripts, and no additional participant instruction, risk disclosure, or annotation guideline was used.

D.1 Implementation and Package Details

All experiments were implemented in Python using PyTorch 2.3.0+cu118 and HuggingFace Transformers 4.45.2. Documents were tokenized using the corresponding BERT and SciBERT tokenizers with a maximum sequence length of 512 tokens, and document representations were obtained from 768-dimensional embeddings. For domain-adaptive fine-tuning, we applied LoRA adapters to the query and value projection layers of SciBERT with rank $r = 8$, and trained the adapters using masked language modeling with the AdamW optimizer for 3 epochs. Traditional Word2Vec-based baselines were implemented using Gensim 4.3.3. The LaTeX preprocessing pipeline was implemented with custom regular-expression-based scripts and a manually constructed mathematical-token dictionary based on **?**. All hyperparameters used in the hierarchical retrieval stage were fixed across models, with $N_{\text{global}} = 500$, $p = 100$, and $N_{\text{expand}} = 2000$.

D.2 Comparison of performance by embedding models in document retrieval algorithms

The complete results of Experiment 1 are reported in Table 13. Across most categories, SciBERT clearly outperforms all other compared models on all evaluation metrics. In a few categories, such as CA and QA, BERT slightly exceeds SciBERT; however, these differences are marginal and do not alter the overall trend.

We also observe an interesting pattern in which the classical embedding model Gensim achieves relatively strong performance in specific categories such as AC and NT. This can be attributed to the lower frequency of equations in these fields, where simple distributional representations can still capture semantic similarity to some extent. Nevertheless, even in these categories, SciBERT remains competitive and maintains stable advantages across most metrics.

Overall, except for a small number of outliers, these results strongly indicate that SciBERT provides the most robust and consistently high quality embeddings across the entire dataset.

Table 13: Comparison of embedding performance of classical models and LLMs

category	AC (K=50)				AC (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.3600	0.4311	0.4586	0.3546	0.2950	0.4289	0.4157	0.2917
GoogleNews	0.1725	0.2299	0.2580	0.1633	0.1588	0.3017	0.2179	0.1540
BERT	0.3750	0.5203	0.5465	0.3673	0.3588	0.5474	0.5211	0.3548
MathBERT	0.4050	0.4824	0.5224	0.3929	0.3600	0.5316	0.4717	0.3535
SBERT	0.4750	0.4362	0.5196	**0.5740**	0.5921	0.5093	0.4643	0.4306
SciBERT	**0.4400**	**0.6314**	**0.6728**	0.4286	**0.3688**	**0.6458**	**0.6151**	**0.3624**
category	AP (K=50)				AP (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.5400	0.5848	**0.7220**	0.5332	0.4950	0.5603	0.6017	0.4912
GoogleNews	0.3600	0.4102	0.4158	0.3520	0.3563	0.4343	0.4015	0.3523
BERT	0.5100	0.5413	0.5862	0.5000	0.5112	0.5722	0.5481	0.5063
MathBERT	0.4600	0.4979	0.5470	0.4515	0.4575	0.5098	0.5075	0.4533
SBERT	0.6550	0.6100	0.7064	0.6845	0.8411	0.7246	0.5230	0.6061
SciBERT	**0.5925**	**0.6241**	0.6652	**0.5842**	**0.5650**	**0.6226**	**0.6244**	**0.5606**
category	CA (K=50)				CA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.3100	0.3500	0.4080	0.2959	0.3025	0.4575	0.3645	0.2955
GoogleNews	0.3025	0.3393	0.3979	0.2934	0.2975	0.3858	0.3538	0.2929
BERT	**0.5275**	**0.5848**	**0.6767**	**0.5179**	0.4725	**0.5868**	**0.6009**	0.4672
MathBERT	0.4875	0.5192	0.5659	0.4796	0.4537	0.5126	0.5190	0.4495
SBERT	0.3450	0.3037	0.3873	0.4139	0.4392	0.3883	0.3367	0.2992
SciBERT	0.5150	0.5512	0.6161	0.5051	**0.4750**	0.5725	0.5633	**0.4697**
category	DG (K=50)				DG (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.5675	0.6127	0.6668	0.5587	0.5050	0.6135	0.6171	0.5000
GoogleNews	0.4150	0.4377	0.4897	0.4056	0.4113	0.4830	0.4593	0.4066
BERT	0.6375	0.6705	0.7218	0.6301	0.6038	0.6499	0.6759	0.5997
MathBERT	0.5475	0.5900	0.6456	0.5383	0.5287	0.5965	0.5915	0.5240
SBERT	0.6475	0.5762	0.6871	0.6366	0.7441	0.6887	0.6403	0.5720
SciBERT	**0.6750**	**0.7086**	**0.7584**	**0.6684**	**0.6175**	**0.6645**	**0.7077**	**0.6136**
category	DS (K=50)				DS (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.3925	0.4135	0.4532	0.3801	0.3750	0.5051	0.4263	0.3687
BERT	0.4125	0.4682	0.5755	0.4031	0.3613	0.4932	0.4962	0.3561
MathBERT	0.4425	0.4704	0.5109	0.4337	0.3950	0.4731	0.4731	0.3902
SBERT	0.2775	0.2450	0.2914	0.2900	0.3135	0.2907	0.2730	0.2424
SciBERT	**0.5150**	**0.5665**	**0.6577**	**0.5051**	**0.4963**	**0.6014**	**0.5871**	**0.4912**
category	FA (K=50)				FA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.2575	0.2705	0.2898	0.2551	0.2475	0.2575	0.2707	0.2462
GoogleNews	0.1875	0.2006	0.2214	0.1811	0.1775	0.2512	0.2073	0.1742
BERT	0.4275	0.4493	0.4890	0.4184	0.4113	0.4748	0.4515	0.4066
MathBERT	0.4275	0.4535	0.4922	0.4184	0.4225	0.4758	0.4586	0.4179
SBERT	0.3975	0.3538	0.4153	0.3846	0.4469	0.4195	0.3929	0.3510
SciBERT	**0.5400**	**0.5914**	**0.6588**	**0.5306**	**0.4963**	**0.6065**	**0.5968**	**0.4912**
category	G1 (K=50)				G1 (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	**0.5975**	**0.6323**	0.6884	**0.5918**	**0.5112**	**0.6212**	0.6316	**0.5076**
GoogleNews	0.1400	0.1499	0.1665	0.1352	0.1412	0.1957	0.1516	0.1389
BERT	0.3825	0.4807	0.5674	0.3724	0.3150	0.5294	0.4845	0.3093
MathBERT	0.2150	0.2345	0.2625	0.2117	0.2087	0.2294	0.2406	0.2071
SBERT	0.2225	0.1937	0.2438	0.2770	0.2695	0.2477	0.2168	0.1907
SciBERT	0.5625	0.6186	**0.7198**	0.5536	0.4863	0.6128	**0.6387**	0.4811
category	G2 (K=50)				G2 (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

GoogleNews	0.1250	0.1546	0.1832	0.1199	0.1300	0.1876	0.1553	0.1275
BERT	0.2900	0.3221	0.3733	0.2857	0.2562	0.2979	0.3279	0.2538
MathBERT	0.3750	0.3965	0.4356	0.3673	0.3762	0.4038	0.4058	0.3725
SBERT	0.1900	0.1900	0.2145	0.2275	0.2562	0.2232	0.1862	0.1881
SciBERT	**0.4250**	**0.4684**	**0.5390**	**0.4158**	**0.3875**	**0.5000**	**0.4796**	**0.3826**
category	GT (K=50)				GT (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.2600	0.2700	0.2785	0.2577	0.2412	0.2649	0.2681	0.2399
GoogleNews	0.2075	0.2540	0.3136	0.1990	0.2087	0.2985	0.2695	0.2045
BERT	0.3125	0.3429	0.4014	0.3061	0.3225	0.3726	0.3596	0.3194
MathBERT	0.2650	0.2886	0.3339	0.2577	0.2863	0.3386	0.3062	0.2828
SBERT	0.2750	0.2637	0.2885	0.2896	0.3044	0.2899	0.2704	0.2614
SciBERT	**0.4700**	**0.5068**	**0.5534**	**0.4617**	**0.4675**	**0.5148**	**0.5119**	**0.4634**
category	MG (K=50)				MG (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.1975	0.2052	0.2130	0.1964	0.1787	0.1878	0.2015	0.1780
GoogleNews	0.1850	0.1847	0.1930	0.1811	0.1862	0.1881	0.1882	0.1843
BERT	**0.6300**	**0.6571**	0.6943	**0.6250**	0.5813	0.6101	0.6544	0.5783
MathBERT	0.4900	0.5146	0.5428	0.4821	0.4675	0.5173	0.5130	0.4634
SBERT	0.5800	0.5562	0.6226	0.6290	0.6943	0.6360	0.5714	0.5518
SciBERT	0.6150	0.6485	**0.6970**	0.6097	**0.5875**	**0.6279**	**0.6556**	**0.5846**
category	NA (K=50)				NA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.7000	0.7075	0.7297	0.6938	0.6688	0.6858	0.7102	0.6655
GoogleNews	0.6133	0.6346	0.6696	0.6054	0.5811	0.6232	0.6370	0.5768
BERT	0.7666	0.7962	**0.8390**	**0.7619**	0.7288	0.7727	0.7995	0.7261
MathBERT	0.6000	0.6360	0.7006	0.5918	0.5938	0.6349	0.6530	0.5896
SBERT	0.5375	0.5338	0.5704	0.5974	0.6218	0.5841	0.5306	0.5303
SciBERT	**0.7800**	**0.8013**	0.8334	0.6643	**0.7544**	**0.7927**	**0.8044**	**0.7519**
category	NT (K=50)				NT (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	**0.7825**	**0.8064**	**0.8310**	**0.7781**	**0.7800**	**0.7954**	**0.8079**	**0.7778**
GoogleNews	0.5150	0.5392	0.5700	0.5051	0.5375	0.5718	0.5527	0.5328
BERT	0.6275	0.6573	0.6971	0.6224	0.5925	0.6194	0.6588	0.5896
MathBERT	0.5400	0.5688	0.6136	0.5332	0.5262	0.5565	0.5768	0.5227
SBERT	0.5425	0.5325	0.5679	0.5673	0.6102	0.5773	0.5383	0.5303
SciBERT	0.6225	0.6694	0.7352	0.6148	0.5888	0.6350	0.6762	0.5846
category	OA (K=50)				OA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.2675	0.3124	0.3591	0.2602	0.2275	0.3198	0.3112	0.2235
BERT	**0.7200**	**0.7560**	**0.8153**	**0.7143**	**0.6725**	**0.7212**	**0.7646**	**0.6692**
MathBERT	0.4300	0.4642	0.5197	0.4235	0.4300	0.4700	0.4761	0.4268
SBERT	0.7300	0.6787	0.7663	0.7155	0.8280	0.7746	0.7245	0.6755
SciBERT	0.7075	0.7359	0.7713	0.7015	0.6500	0.7016	0.7328	0.6465
category	PR (K=50)				PR (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.2050	0.2501	0.3149	0.1939	0.2125	0.3266	0.2659	0.2071
GoogleNews	0.4175	0.4484	0.5019	0.4082	0.4225	0.4733	0.4693	0.4179
BERT	0.5675	0.6158	0.6868	0.5587	0.5188	0.6225	0.6254	0.5139
MathBERT	0.5075	0.5307	0.5668	0.4974	0.4738	0.5541	0.5352	0.4684
SBERT	0.3525	0.3150	0.3894	0.3789	0.4440	0.3947	0.3469	0.3119
SciBERT	**0.5875**	**0.6291**	**0.7059**	**0.5791**	**0.5687**	**0.6314**	**0.6418**	**0.5644**
category	QA (K=50)				QA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.4100	0.4555	0.5243	0.3980	0.4113	0.5178	0.4726	0.4053
BERT	0.5475	0.5961	0.6709	0.5383	0.5225	0.5969	0.6131	0.5177
MathBERT	0.4100	0.4382	0.4791	0.4005	0.3900	0.4458	0.4462	0.3851
SBERT	0.4475	0.4625	0.5012	0.5315	0.5999	0.5279	0.4388	0.4583
SciBERT	**0.6475**	**0.6948**	**0.7686**	**0.6403**	**0.5975**	**0.6384**	**0.7019**	**0.5934**

category	RA (K=50)				RA (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.1350	0.1762	0.2184	0.1301	0.1363	0.2017	0.1819	0.1338
BERT	0.4425	0.4769	0.5218	0.4362	0.3987	0.4472	0.4793	0.3952
MathBERT	0.3250	0.3508	0.3875	0.3189	0.3125	0.3509	0.3539	0.3093
SBERT	0.5625	0.5475	0.6064	0.6113	0.6710	0.6222	0.5536	0.5429
SciBERT	**0.5350**	**0.5733**	**0.6392**	**0.5281**	**0.5012**	**0.5462**	**0.5876**	**0.4975**

category	RT (K=50)				RT (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.2600	0.2841	0.3314	0.2526	0.2575	0.3253	0.2881	0.2538
BERT	0.5075	0.5557	0.6170	0.5000	0.4875	0.5437	0.5561	0.4836
MathBERT	0.3725	0.3923	0.4227	0.3648	0.3688	0.4041	0.4013	0.3649
SBERT	0.6300	0.5825	0.6762	0.6675	0.7361	0.6780	0.6224	0.5783
SciBERT	**0.6975**	**0.7269**	**0.7580**	**0.6913**	**0.6138**	**0.6797**	**0.7120**	**0.6098**

category	SG (K=50)				SG (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0550	0.0651	0.0864	0.0536	0.0462	0.0702	0.0718	0.0455
GoogleNews	0.5000	0.5118	0.5321	0.4949	0.4750	0.4905	0.5134	0.4722
BERT	0.7000	0.7304	0.7656	0.6939	0.6638	0.6947	0.7269	0.6604
MathBERT	0.5625	0.6030	0.6657	0.5561	0.5250	0.5732	0.6119	0.5215
SBERT	0.6500	0.5938	0.6770	0.6250	0.7168	0.6772	0.6454	0.5909
SciBERT	**0.7275**	**0.7613**	**0.8048**	**0.7219**	**0.6850**	**0.7155**	**0.7570**	**0.6818**

category	SP (K=50)				SP (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.4725	0.4940	0.5235	0.4694	0.4288	0.4523	0.4900	0.4268
GoogleNews	0.2275	0.2516	0.2940	0.2219	0.2263	0.2940	0.2582	0.2235
BERT	**0.5975**	**0.6367**	0.6929	**0.5893**	**0.5600**	**0.6264**	**0.6377**	**0.5556**
MathBERT	0.4250	0.4640	0.5158	0.4158	0.4250	0.4961	0.4743	0.4205
SBERT	0.6575	0.6262	0.6849	0.6682	0.7149	0.6843	0.6505	0.6225
SciBERT	0.5550	0.6090	**0.7025**	0.5459	0.5288	0.6161	0.6261	0.5240

category	OC (K=50)				OC (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.6750	0.6959	0.7296	0.6684	0.6425	0.6722	0.7002	0.6389
GoogleNews	0.4600	0.4822	0.5258	0.4515	0.4575	0.4973	0.4971	0.4533
BERT	0.8125	0.8430	0.8852	0.8087	0.7637	0.7948	0.8406	0.7614
MathBERT	0.6800	0.7048	0.7429	0.6735	0.6425	0.6804	0.7036	0.6389
SBERT	0.6525	0.5988	0.6710	0.6391	0.7000	0.6672	0.6480	0.5960
SciBERT	**0.8400**	**0.8594**	**0.8922**	**0.8367**	**0.7900**	**0.8135**	**0.8551**	**0.7879**

category	AG (K=50)				AG (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.1150	0.1279	0.1481	0.1122	0.1138	0.1319	0.1302	0.1124
GoogleNews	0.2400	0.2471	0.2546	0.2321	0.2562	0.2953	0.2557	0.2525
BERT	**0.3975**	0.4288	0.4683	**0.3903**	0.3625	0.4410	0.4320	0.3586
MathBERT	0.2650	0.3012	0.3686	0.2577	0.2612	0.3131	0.3199	0.2576
SBERT	0.3125	0.2875	0.3352	0.3651	0.3630	0.3364	0.3036	0.2828
SciBERT	**0.3975**	**0.4359**	**0.4922**	0.3878	**0.3900**	**0.4681**	**0.4461**	**0.3851**

category	GR (K=50)				GR (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.3000	0.3246	0.3634	0.2959	0.2637	0.3146	0.3281	0.2614
GoogleNews	0.4050	0.4393	0.4940	0.3980	0.3950	0.4370	0.4552	0.3914
BERT	0.6650	0.7128	**0.7894**	0.6582	**0.6537**	**0.7155**	**0.7326**	**0.6503**
MathBERT	0.3900	0.4258	0.4819	0.3827	0.3762	0.4206	0.4392	0.3725
SBERT	0.6400	0.5925	0.6681	0.6328	0.7038	0.6609	0.6352	0.5896
SciBERT	**0.6750**	**0.7129**	0.7604	**0.6684**	0.6500	0.6868	0.7170	0.6465

category	CV (K=50)				CV (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.3150	0.3452	0.3873	0.3061	0.3025	0.3962	0.3549	0.2980
GoogleNews	0.2350	0.2601	0.3061	0.2296	0.2037	0.2742	0.2716	0.2008
BERT	0.4250	0.4519	0.4864	0.4184	0.3725	0.4552	0.4486	0.3687

MathBERT	0.3750	0.3944	0.4198	0.3699	0.3500	0.3954	0.3906	0.3472
SBERT	0.3550	0.3300	0.3807	0.3883	0.4282	0.3849	0.3495	0.3270
SciBERT	**0.4700**	**0.5106**	**0.5714**	**0.4617**	**0.4425**	**0.5222**	**0.5206**	**0.4381**
category	CO (K=50)				CO (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.4575	0.4676	0.4855	0.4541	0.4700	0.4735	0.4768	0.4684
GoogleNews	0.2475	0.2726	0.2944	0.2398	0.2250	0.3116	0.2757	0.2210
BERT	0.5775	0.5987	0.6279	0.5714	0.5787	0.6258	0.6023	0.5758
MathBERT	0.5525	0.5825	0.6232	0.5459	0.5413	0.5718	0.5863	0.5379
SBERT	0.3375	0.3075	0.3549	0.3442	0.3801	0.3528	0.3342	0.3056
SciBERT	**0.6725**	**0.7065**	**0.7326**	**0.6658**	**0.6438**	**0.7098**	**0.7049**	**0.6402**
category	CT (K=50)				CT (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.0575	0.0580	0.0560	0.0561	0.0587	0.0585	0.0562	0.0581
BERT	0.6350	0.6635	0.6960	**0.6301**	0.5850	0.6173	0.6566	0.5821
MathBERT	0.5250	0.5967	0.6431	0.5179	0.5162	0.6008	0.5868	0.5126
SBERT	0.5500	0.5100	0.5862	0.5816	0.6300	0.5868	0.5434	0.5063
SciBERT	**0.7325**	**0.7448**	**0.7609**	0.6046	**0.7050**	**0.7185**	**0.7371**	**0.7033**
category	AT (K=50)				AT (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
GoogleNews	0.3775	0.4086	0.4567	0.3699	0.3287	0.4349	0.4181	0.3245
BERT	0.5050	0.5400	0.5936	0.4974	0.4550	0.5283	0.5440	0.4508
MathBERT	0.3250	0.3552	0.4030	0.3163	0.3050	0.3886	0.3650	0.3005
SBERT	0.3250	0.2825	0.3435	0.3270	0.3727	0.3496	0.3214	0.2803
SciBERT	**0.5900**	**0.6386**	**0.7141**	**0.5816**	**0.5600**	**0.6344**	**0.6523**	**0.5556**
category	ST (K=50)				ST (K=100)			
metrics / model	Precision	nDCG	MAP	Pairwise F1	Precision	nDCG	MAP	Pairwise F1
Gensim	0.1300	0.1519	0.1532	0.1276	0.1175	0.1620	0.1408	0.1162
GoogleNews	0.6400	0.6447	0.6489	0.6352	0.6300	0.6353	0.6424	0.6275
BERT	0.7175	0.7587	0.8224	**0.7117**	0.6725	0.7165	0.7631	0.6692
MathBERT	0.5550	0.5942	0.6526	0.5485	0.5413	0.5727	0.6007	0.5379
SBERT	0.7575	0.7212	0.7931	0.7899	0.8540	0.8051	0.7526	0.7184
SciBERT	**0.7875**	**0.8159**	**0.8549**	0.6582	**0.7675**	**0.7954**	**0.8214**	**0.7652**

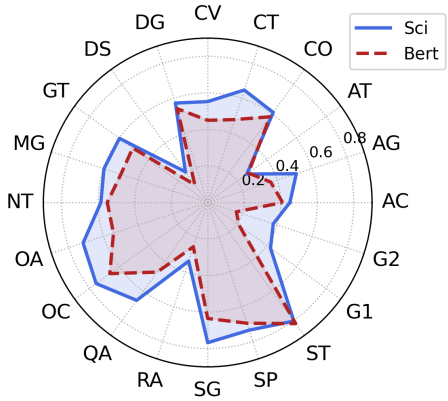
D.3 Equation density and feature embedding performances

This section presents the results of the experiments conducted in Experiment 2, shown in Figure 5-Figure 8. We analyze embedding performance differences by dividing documents into a high equation density group and a low equation density group based on the relative proportion of equations within each document. Categories in which equations are intrinsically rare are excluded for experimental validity, and experiments are conducted on test papers selected from the remaining 20 categories.

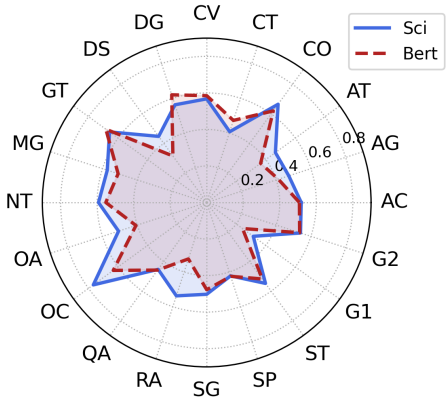
Selected categories: AC, AG, AT, CO, CT, CV, DG, DS, GT, MG, NT, OA, OC, QA, RA, SG, SP, ST, G1, G2

Experimental results show that SciBERT outperforms BERT in most categories, with performance gaps becoming more pronounced in formula dense fields. Given that mathematical expressions convey critical semantic information in scientific literature, this suggests that SciBERT's domain specific representations effectively capture formula-driven semantics.

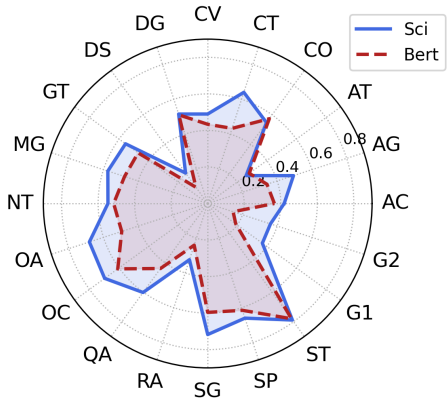
In contrast, in categories with intrinsically low formula usage, such as CT and DG, BERT occasionally exceeds SciBERT on certain metrics. These observations indicate that SciBERT's advantages become more evident as equation driven semantics play a larger role in the document content.



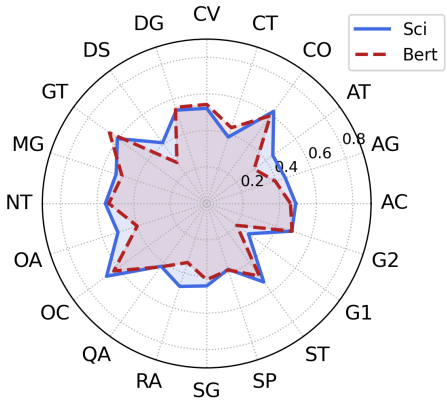
(a) Precision@50 with high equation density data



(b) Precision@50 with low equation density data

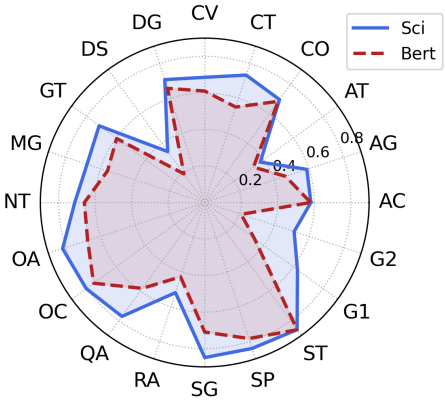


(c) Precision@100 with high equation density data

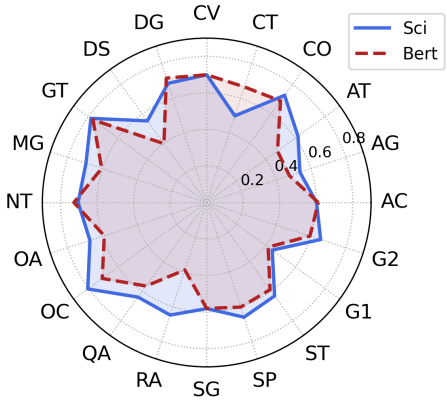


(d) Precision@100 with low equation density data

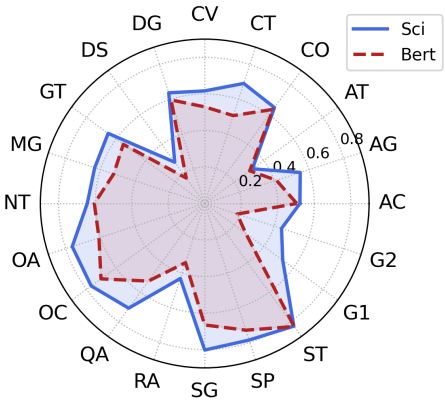
Figure 5: Performance gap between SciBERT and BERT in Precision



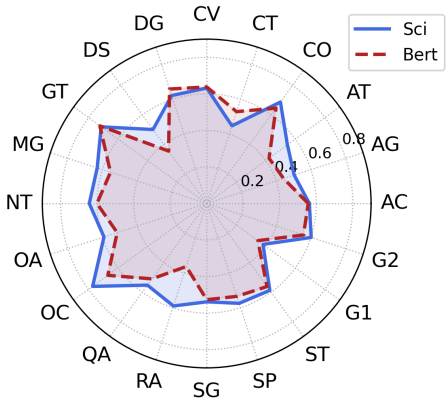
(a) MAP@50 with high equation density data



(b) MAP@50 with low equation density data

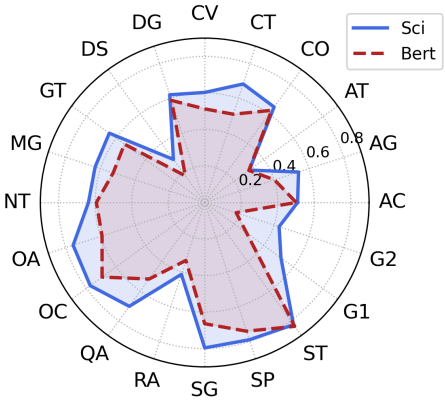


(c) MAP@100 with high equation density data

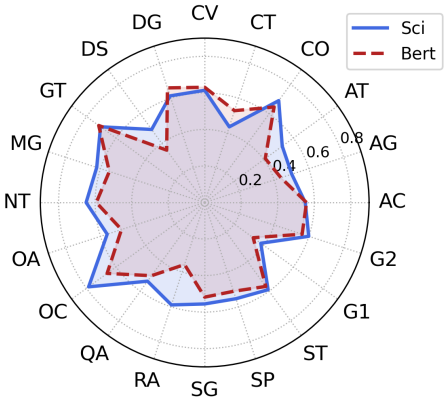


(d) MAP@100 with low equation density data

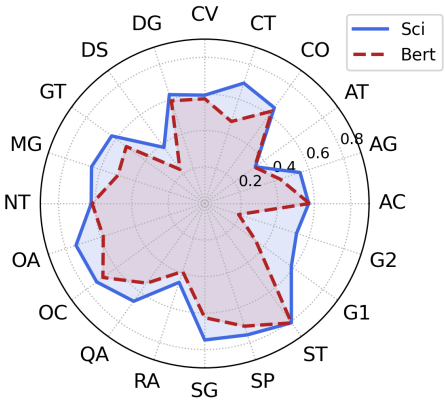
Figure 6: Performance gap between SciBERT and BERT in MAP



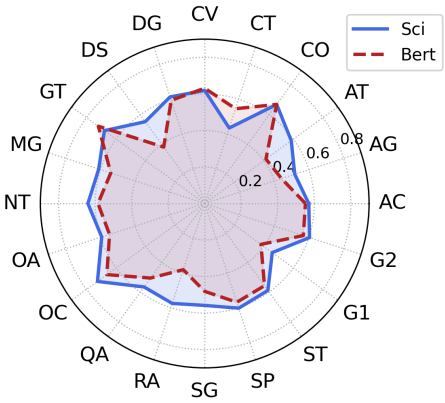
(a) nDCG@50 with high equation density data



(b) nDCG@50 with low equation density data

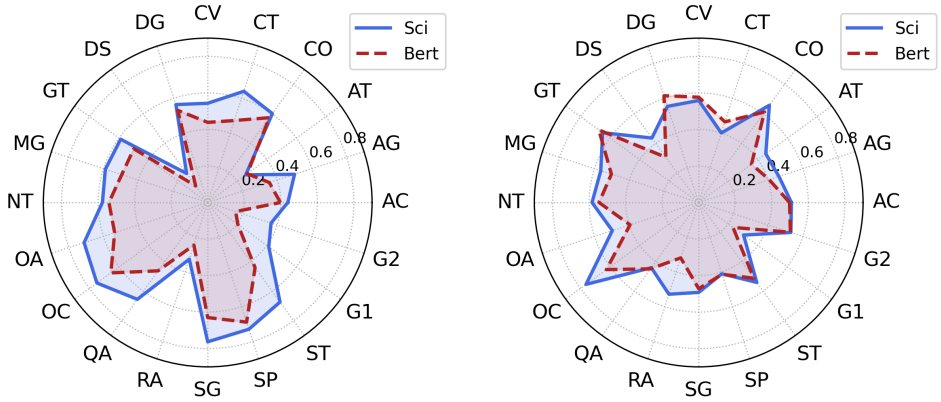


(c) nDCG@100 with high equation density data

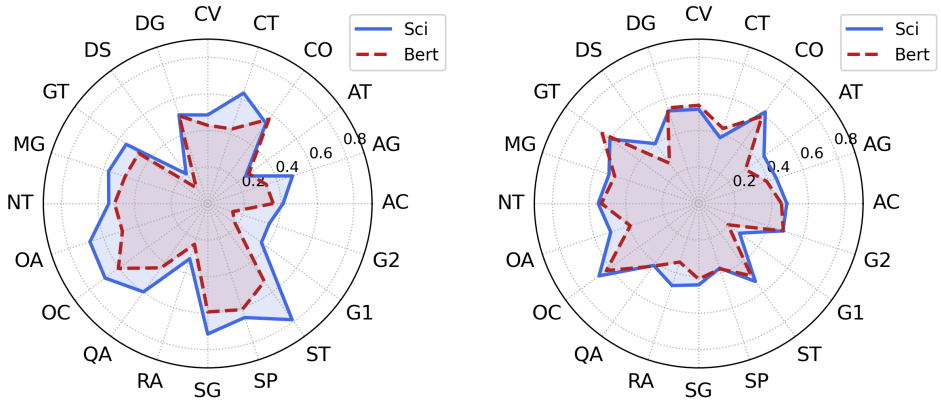


(d) nDCG@100 with low equation density data

Figure 7: Performance gap between SciBERT and BERT in nDCG



(a) Pairwise-F1@50 with high equation density data (b) Pairwise-F1@50 with low equation density data



(c) Pairwise-F1@100 with high equation density data (d) Pairwise-F1@100 with low equation density data

Figure 8: Performance gap between SciBERT and BERT in Pairwise-F1

D.4 The effect of domain-adaptive journal-specific finetuning

To build a general purpose semantic document retrieval system, we perform category wise domain fine-tuning using category specific data, with full results reported in Table 14. As Experiment 1 establishes the superiority of SciBERT based models, our analysis focuses on @100 metrics. Moreover, since fine-tuning is conducted on category specific data in isolation, performance is reported using AP rather than MAP.

We observe stable performance improvements after fine-tuning in most categories. However, in a few fields, such as FA and GR, gains are marginal or slightly negative. Notably, these categories already exhibit strong baseline performance and possess mathematical or structural characteristics that distinguish them from other fields. We therefore interpret these minor degradations not as general model failures, but as a recalibration of domain specific representations that occurs as SciBERT adapts to balanced representations across broader STEM domains during fine-tuning.

D.5 Experimental results of contents composition

We use papers from the arXiv repository as experimental data. Since embedding the full text of each paper is inefficient, we select only key components for representation learning. Accordingly, we conduct an ablation study to evaluate which manuscript components provide the most informative signals for document retrieval task. The full results are shown in Figure 9-Figure 16.

The results show that Setting1, which includes the title, abstract, and introduction, consistently achieves substantially higher performance than other settings across most categories. A particularly striking case is observed in the NT category. As shown in Figure 11 and Figure 13, when using Setting3 (title only), the relevance check becomes effectively non-informative for NT. In contrast, Setting1 enables NT to achieve performance comparable to other categories, clearly demonstrating that the selection of textual context used for embedding has a decisive impact on retrieval quality. This further suggests that richer contextual information is essential for domains involving complex conceptual and technical contexts.

Category	Precision@100	AP@100	nDCG@100	Pairwise F1@100
AC	0.170 → 0.380 (↑)	0.312 → 0.357 (↑)	0.471 → 0.504 (↑)	0.374 → 0.415 (↑)
AG	0.500 → 0.670 (↑)	0.707 → 0.727 (↑)	0.631 → 0.699 (↑)	0.667 → 0.818 (↑)
AP	0.670 → 0.680 (↑)	0.734 → 0.764 (↑)	0.703 → 0.717 (↑)	0.676 → 0.737 (↑)
CA	0.590 → 0.740 (↑)	0.756 → 0.785 (↑)	0.648 → 0.766 (↑)	0.646 → 0.657 (↑)
CO	0.940 → 0.950 (↑)	0.985 → 0.961 (↓)	0.954 → 0.956 (↑)	0.707 → 0.768 (↑)
CT	0.900 → 0.920 (↑)	0.926 → 0.932 (↑)	0.912 → 0.926 (↑)	0.879 → 0.901 (↑)
CV	0.790 → 0.870 (↑)	0.952 → 0.955 (↑)	0.838 → 0.898 (↑)	0.859 → 0.899 (↑)
DG	0.460 → 0.460 (→)	0.559 → 0.570 (↑)	0.525 → 0.492 (↓)	0.459 → 0.570 (↑)
G1	0.650 → 0.660 (↑)	0.836 → 0.792 (↓)	0.712 → 0.710 (↓)	0.687 → 0.667 (↓)
G2	0.590 → 0.610 (↑)	0.750 → 0.804 (↑)	0.663 → 0.675 (↑)	0.586 → 0.606 (↑)

Table 14: Full results of finetuning vs. without finetuning

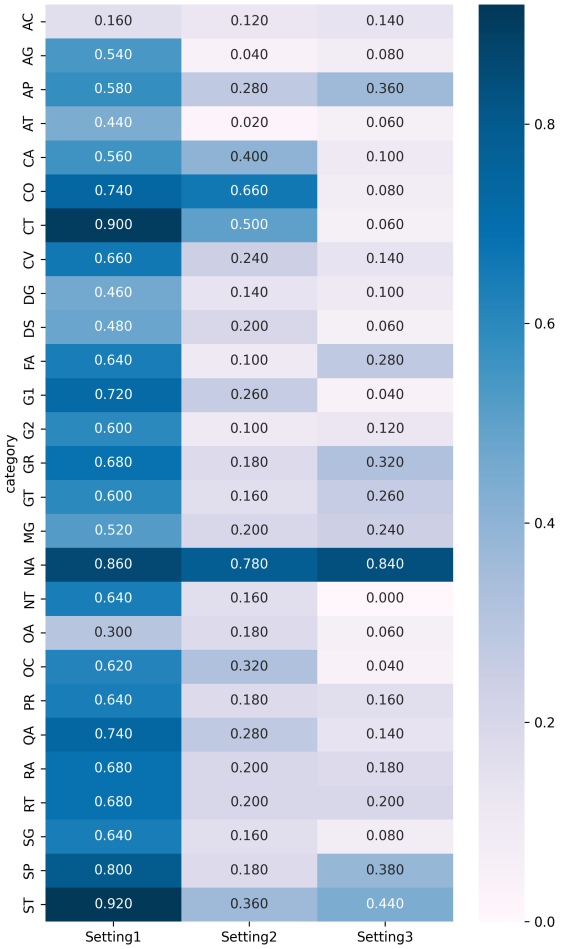


Figure 9: Heatmap of Precision@50 across categories and settings

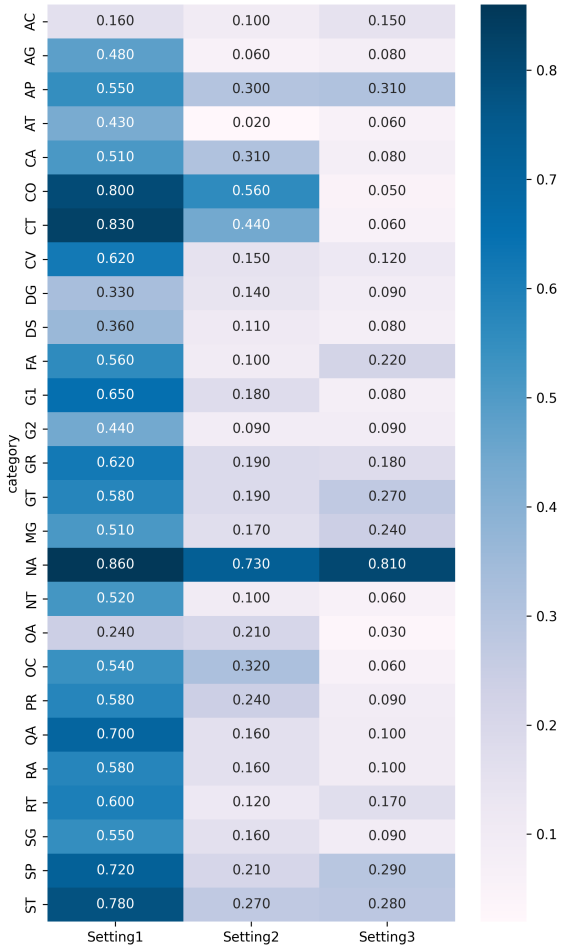


Figure 10: Heatmap of Precision@100 across categories and settings

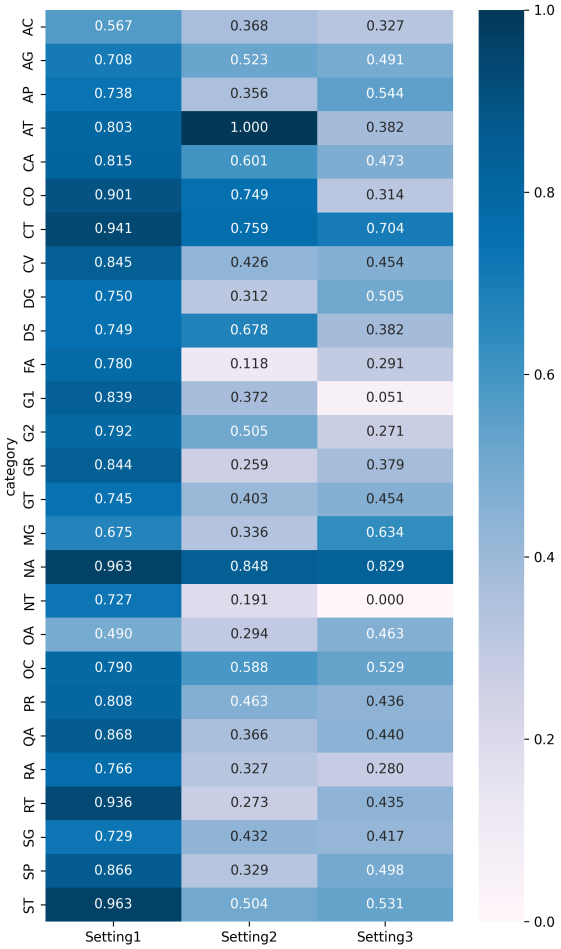


Figure 11: Heatmap of MAP@50 across categories and settings

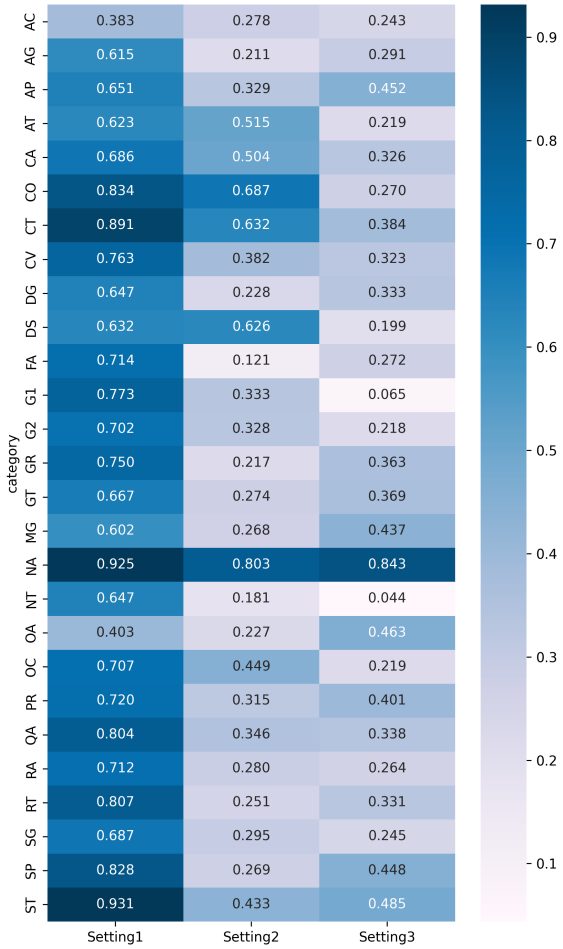


Figure 12: Heatmap of MAP@100 across categories and settings

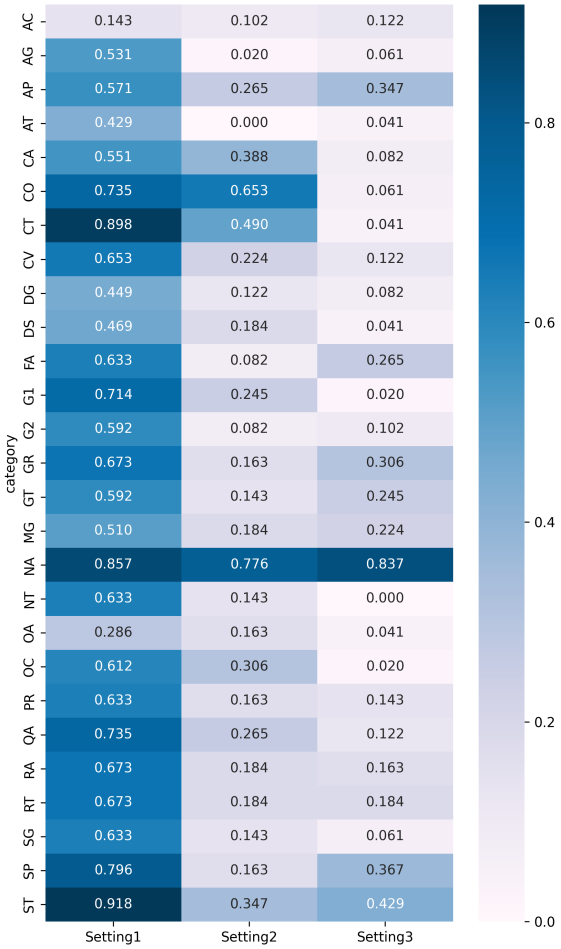


Figure 13: Heatmap of Pairwise-F1@50 across categories and settings

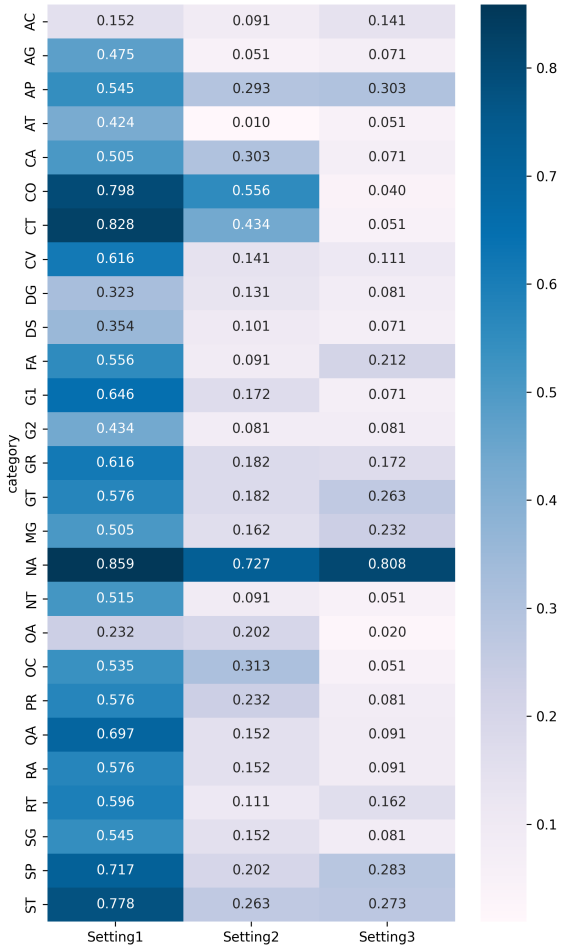


Figure 14: Heatmap of Pairwise-F1@100 across categories and settings

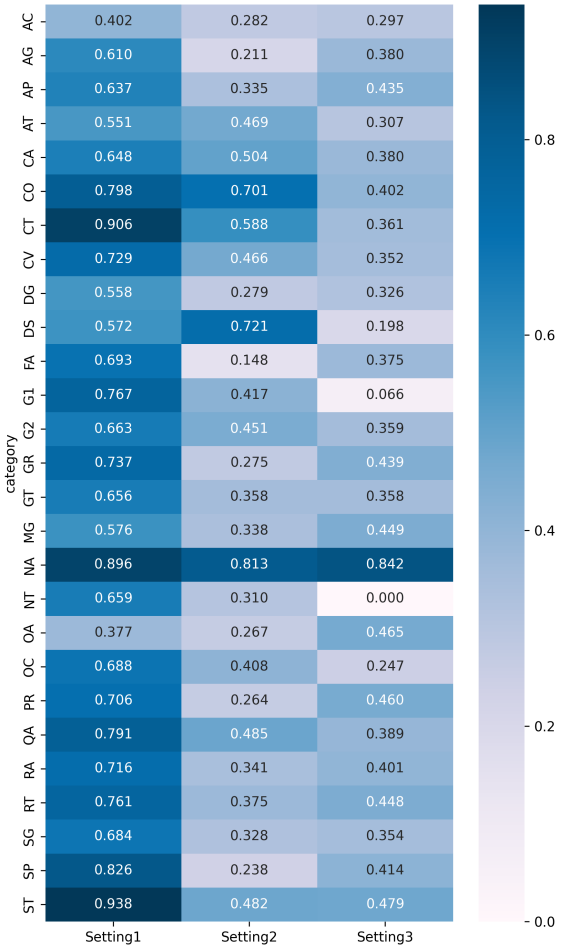


Figure 15: Heatmap of nDCG@50 across categories and settings

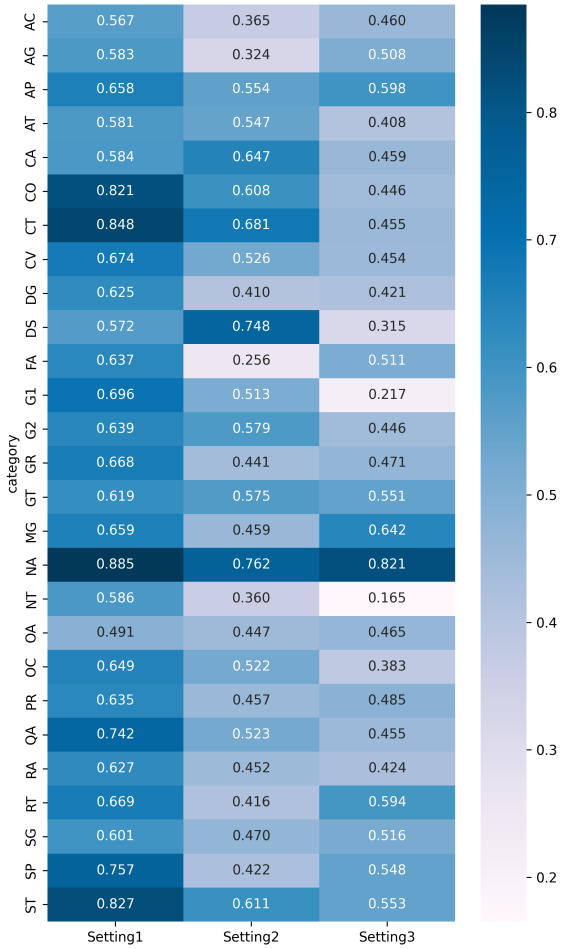


Figure 16: Heatmap of nDCG@100 across categories and settings

D.6 Experimental results by LaTeXprocessing level

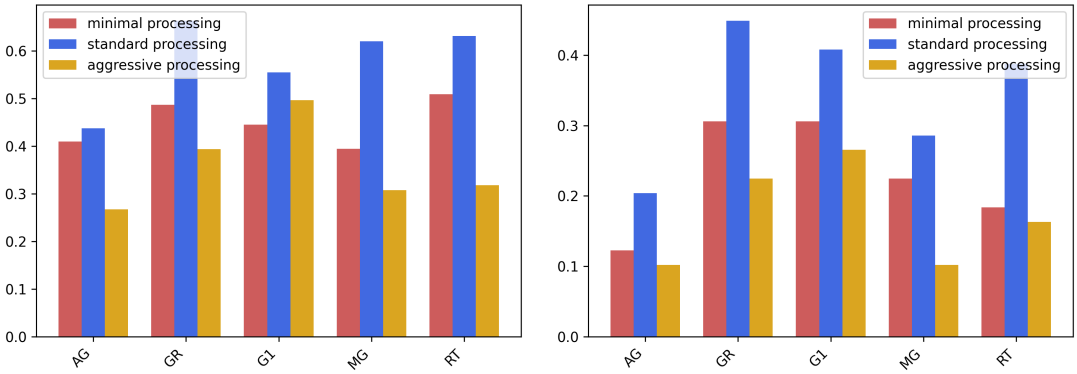
Since we operate on raw LaTeX source files, preprocessing requires strategies distinct from standard natural language processing pipelines. STEM manuscripts are typically written in LaTeX, which faithfully reflects domain-specific structure, but raw LaTeX sources also contain numerous syntactic artifacts and formatting commands unrelated to semantic content. Similar to removing stopwords in NLP, LaTeX data must be cleaned to discard elements irrelevant to meaning.

Accordingly, we define three preprocessing levels (minimal, standard, and aggressive) based on processing intensity. These strategies are evaluated for their ability to provide stable and consistent performance across all categories, rather than being tailored to specific fields. In this section, we report average metric performance across all categories in Figure 18.

Our results show that standard processing consistently outperforms both minimal and aggressive processing across all metrics. This indicates that the proposed standard preprocessing effectively balances structural noise removal and semantic preservation in LaTeX documents, making it the most suitable strategy for feature embedding.

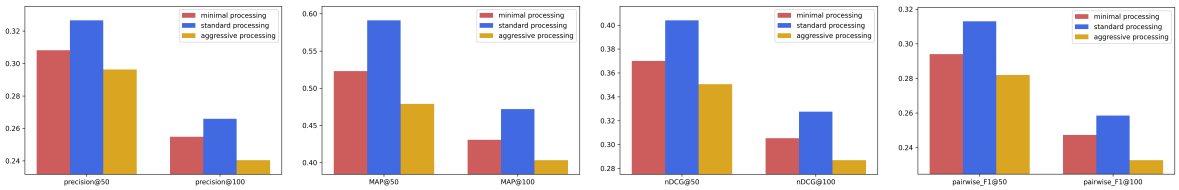
E Use of AI Assistants

AI-based writing assistants were used only for limited language refinement and proofreading during manuscript preparation. All core research ideas, experimental design, implementation, analyses, and conclusions were developed and verified by the authors.



(a) MAP@50 (b) Pairwise-F1@50

Figure 17: Performance across categories by processing levels



(a) Precision (b) MAP (c) nDCG (d) Pairwise-F1

Figure 18: Average performance across processing levels

	Pairwise-F1 @ 50			Pairwise-F1 @ 100		
Measure	cosine	jaccard	MaxSim	cosine	jaccard	MaxSim
OC	0.7168	0.6531	0.8367	0.6856	0.7184	0.7879
AG	0.2934	0.4541	0.3878	0.2980	0.4432	0.3851
GR	0.5536	0.4107	0.6684	0.5051	0.3472	0.6465
CV	0.0918	0.6556	0.4617	0.0884	0.6402	0.4381
CO	0.5638	0.6148	0.6658	0.5088	0.5808	0.6402

Table 15: Pairwise-F1 comparison across similarity measures