Enhancing Multilingual Mathematical Document Retrieval: A Hindi-to-English Translation and ColBERT-based Approach

Ayush Sur[1,*,†], Venkatesh Shukla[1,†], Aditya Rai[1,†] and Lucky Garg[1,†]

[1] 1National Institute of Technology, Kurukshetra

Abstract

Mathematical Information Retrieval (MIR) systems tend to be monolingual, limiting their capacity to access multilingual scientific information. Cross-Lingual Mathematical Information Retrieval (CLMIR) strives to fill this gap by enabling mathematical knowledge retrieval across languages. This paper presents a CLMIR system for Hindi-English scientific and scholarly articles that uses IndicTrans2 for machine translation, ColBERTv2 with FAISS indexing for dense retrieval, and a cross-encoder (MiniLM-L-6-v2) for re-ranking the retrieved documents. In the FIRE CLMIR 2025 shared task, the submission applying a re-ranking procedure with a ColBERT + MiniLM model attained a Precision at 10 (P@10) of 0.08, a Mean Average Precision (MAP) of 0.1484, and a normalized Discounted Cumulative Gain (nDCG) of 0.2202 over the official test collection. The proposed system can process both mathematical and text expressions, with a future possibility of LaTeX-aware filtering and tagging. By providing improved access to mathematical resources for English- and Hindi-speaking researchers, this work leads to greater inclusiveness and fairness in technical knowledge exchange.

Keywords

Mathematical Information Retrieval, Cross-lingual Retrieval, Hindi–english, ColBERT, Dense Retrieval, Reranking, Machine translation,

1. Introduction

Mathematical Information Retrieval (MIR) focuses on finding relevant mathematical expressions, equations, and documents in response to user queries. While effective in monolingual settings, most MIR systems unable to steer the retrieval on multilingual contexts, limiting access to relevant sources and valuable content. Cross-Lingual Mathematical Information Retrieval (CLMIR) addresses these inherent computational limitations with enabling search and retrieval across languages, an important capability for communities where technical resources are fragmented across linguistic boundaries. In the context of Hindi–English retrieval, this challenge is particularly significant, as hindi-speaking scholars and researchers struggle to access high-quality English-language materials, while mathematical knowledge written in Hindi remains underutilized by the wider research community. The difficulty lies in accurately translating both text and mathematical expressions, aligning cross-lingual concepts, and compensating for the scarcity of bilingual datasets. The CLMIR 2025 shared task establishes a benchmark for this challenge by providing a Hindi corpus of 40,000 mathematical entries from the Math Stack Exchange dataset (ARQMath-1), along with English text–formula queries. This setting demands robust and precise matching strategies that extend beyond simple textual alignment to capture both semantic context and mathematical structure. This paper proposes a retrieval framework that integrates translation, dense retrieval, and re-ranking mechanisms to improve access to mathematical content for both Hindi- and English-speaking researchers, thereby promoting inclusivity in technical knowledge sharing. The experimental evaluation demonstrates the potential of extending the framework with features such as LaTeX-aware search and tag-based filtering, which are also showcased in this work. The manuscript

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

ayushsur26@gmail.com (A. Sur); venkateshshukla18012005@gmail.com (V. Shukla); adityarai735@gmail.com (A. Rai); luckygarg452@gmail.com (L. Garg)

is organized as follows: Introduction section presents fundamentals of the proposed work. Related work section summarizes a brief literature of relevant strategy of MIR. Next, details of the proposed framework and the experimental setup, along with the evaluation metrics, are presented. Towards the end, the proposed work is concluded and directions for future research are suggested.

2. Related Work

Research in cross-language mathematical information retrieval remains limited due to the combined challenge of translating natural language alongside mathematical expressions, the lack of large multilingual datasets, and the limitations of general-purpose machine translation for domain-specific terminology. Furthermore, effective retrieval requires structure-aware representation and matching techniques, yet the niche nature of the field has drawn relatively little research attention. To situate the work, prior studies in Mathematical Information Retrieval and CLMIR are first reviewed.

2.1. Mathematical Information Retrieval

MIR techniques can be broadly classified into structural search and data-driven search.

2.1.1. Structural Search

Structural search methods operate on the syntactical and hierarchical representation of mathematical formulas, often modeled as Operator Trees or Symbol Layout Trees (SLTs). Retrieval is then performed using pattern matching, tree indexing, or approximate matching algorithms. MathWebSearch addresses formula retrieval from a semantic perspective by converting expressions into canonical tree forms, enabling subtree and approximate matching regardless of superficial notation differences. Tangent-CFTED [1] represents formulas as SLTs and introduces the Compressed Formula Tree Edit Distance (CFTED) metric, which measures structural similarity while accounting for substitutions, deletions, and reordered symbols. This allows efficient approximate matching by compressing redundant substructures. While these approaches excel at capturing formula structure, they often ignore surrounding text, omitting valuable contextual information.

2.1.2. Data Driven Search

Data-driven approaches use machine learning and embedding models to jointly encode mathematical expressions and their textual context into a continuous semantic space, enabling retrieval via vector similarity. Approach0 [2] employs transformer-based models with contrastive learning to generate dense embeddings that integrate both syntactic and semantic features, achieving state-of-the-art results on MIR benchmarks. Another work [3] evaluates the dense retrieval at token-and passage-levels, showing that combining both granularity levels improves the retrieval of math content by balancing fine-grained symbol encoding with broader contextual understanding. The proposal CLMIR strategy in this paper adopts the data-driven paradigm presented in [3] to jointly model formulas and text, ensuring richer semantic retrieval across languages.

2.2. Cross Lingual Mathematical Information Retrieval

CLMIR enables queries in one language to retrieve math content in another, requiring simultaneous handling of linguistic differences and the semantic–structural complexity of formulas. This often involves machine translation, multilingual embeddings, or hybrid methods combining formula matching with semantic alignment. CrossMath [4] introduces a CLMIR benchmark with manually translated queries in Croatian, Czech, Persian, and Spanish. The system uses mBART [5] and NLLB [6] translation models with a formula masking strategy to preserve mathematical notation during translation. On ARQMath datasets, CrossMath [4] matches monolingual performance, demonstrating its effectiveness in bridging language barriers for math retrieval.

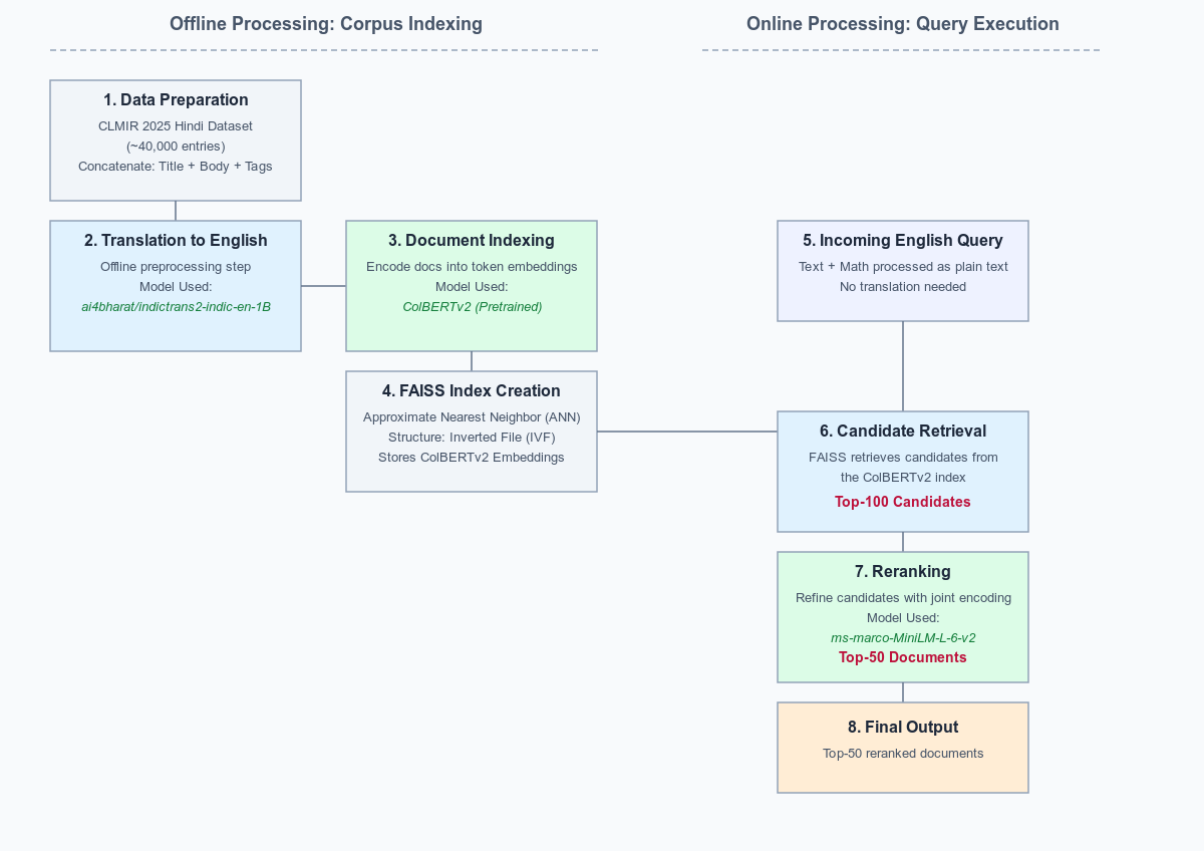

Figure 1: Overview of the proposed CLMIR pipeline, illustrating offline corpus indexing and online query execution.

2.3. Late-Interaction based Dense Retrieval

Recent advances in dense retrieval have explored late-interaction architectures, which allow token-level matching between queries and documents after independent encoding. ColBERT [7] introduced this paradigm by storing contextualized token embeddings for each document and computing maximum-similarity scores at retrieval time, striking a balance between expressiveness and efficiency. ColBERTv2 [8] improves on the original by introducing more compact representations, improved scoring functions, and optimized indexing, enabling faster retrieval with reduced memory requirements. These characteristics make ColBERTv2 particularly suitable for CLMIR tasks, where fine-grained alignment between translated text and query tokens, including partial matches within mathematical contexts, is essential for high recall and precision.

3. Proposed Methodology

This section describes the implemented pipeline for cross-lingual mathematical information retrieval from the Hindi corpus provided in the CLMIR 2025 shared task. Figure 1 illustrates the proposed pipeline. The aim is to retrieve Hindi-origin mathematical content in response to English text–formula queries by translating, indexing, and ranking documents in a unified English representation space.

3.1. Data Preparation

The CLMIR 2025 Shared Task dataset [9] is adapted for the experimental analysis, consisting approximately 40,000 Hindi-language scientific entries derived from the Math Stack Exchange corpus(ARQMath-1). Each entry consists of Title (a short summary of the problem or topic), Body (descriptive text and embedded mathematical expressions in LaTeX format) and Tags (categorical labels denoting subject

domains). For retrieval purposes, each document was converted into a single composite chunk by concatenating its title, body, and tags. This representation allows the retrieval model to exploit multiple aspects of document content within a unified indexing framework.

3.2. Translation and Document Indexing with ColBERTv2

To enable cross-lingual retrieval between Hindi documents and English queries, the entire Hindi corpus was translated into English offline (i.e., in a single preprocessing step). The translation was performed using the ai4bharat/indictrans2-indic-en-1B [10] model, a high-capacity neural machine translation system optimized for Indic languages. Translating the corpus in advance eliminates the need for on-the-fly translation during query processing and ensures that both queries and documents reside in the same linguistic space. Further, in the proposed strategy, dense retrieval was performed using the ColBERTv2 architecture, a late-interaction retrieval model that enables fine-grained token-level matching between queries and documents. The publicly available pretrained ColBERTv2 model provided by the original authors was used without additional fine-tuning. Each composite document (title + body + tags) was encoded into contextualized token embeddings, which were stored in an index for efficient matching.

3.3. Candidate Retrieval with FAISS

Approximate nearest neighbor (ANN) search was implemented using FAISS (Facebook AI Similarity Search) [11]. The FAISS index was configured with its default inverted file (IVF) structure, balancing retrieval speed and accuracy for large-scale search. For each incoming query, FAISS retrieves the Top-100 candidate documents from the ColBERTv2 index based on vector similarity.

3.4. Reranking with Cross Encoder

The initial candidate set is refined using a cross-encoder model, cross-encoder/ms-marco-MiniLM-L-6-v2 [12], which jointly encodes the query and each candidate document to produce a more precise relevance score. This reranking step captures deeper semantic relationships and improves the quality of the final results. After reranking, the Top-50 documents are returned as the system's output for each query.

3.5. Query Processing

Since the translated corpus is entirely in English, incoming English queries are used directly without additional translation. Queries may contain both textual and mathematical elements, which are processed as plain text by the current pipeline.

4. Performance Evaluation and Analysis

To assess the effectiveness of the system, multiple retrieval and reranking configurations were evaluated using three widely adopted information retrieval metrics:

- Precision at 10 (P@10): Measures the fraction of relevant documents in the Top-10 results, indicating immediate usefulness to the user.
- Mean Average Precision (MAP): Averages the precision over all relevant documents for each query, reflecting the overall ranking quality.
- Normalized Discounted Cumulative Gain (nDCG): Accounts for the position of relevant documents in the ranked list, rewarding higher placement of relevant items.

4.1. Experimental Setup

In this view, since the official test set was limited, a custom evaluation corpus consisting of 500 documents and 20 representative user queries was created to validate the effectiveness of the proposed retrieval strategy. Each query–document relevance judgment was manually curated to reflect realistic retrieval scenarios. The proposed pipeline is configured and tested for five configurations for the employed models:

- ColBERTv2 only
- ColBERTv2 + cross-encoder (MiniLM-L-6-v2)
- all-mpnet-base-v2 [13] only
- all-mpnet-base-v2 + cross-encoder (MiniLM-L-6-v2)
- MathBERT (tb17/MathBERT) + cross-encoder (MiniLM-L-6-v2)

4.2. Proposed Extension

The proposed pipeline focuses on CLIR strategy, through translation and dense re-ranking, in this view, there are several promising directions to further enhance performance in CLMIR tasks. The following ideas are under consideration for future development and have not yet been implemented:

- LaTeX-Aware Retrieval: Many queries in mathematical domains rely heavily on the accurate interpretation of mathematical expressions. Future work involves extracting LaTeX segments from documents and representing them separately in a dense vector space using models such as Tangent-CFTED. This would allow parallel retrieval from two spaces textual and mathematical combining results before re-ranking.
- Query-Time Dual-Space Search: If a query contains mathematical expressions, the same LaTeX extraction and encoding process will be applied to the query. Retrieval will then be performed on both the formula vector space and the text vector space, with results merged before final scoring
- Fine-Tuning MathBERT for ColBERT-Based Retrieval: Future work includes fine-tuning MathBERT [14] embeddings for use within the ColBERT late-interaction framework. This would combine MathBERT's domain knowledge of mathematical text with ColBERT's token-level matching capabilities.

4.3. Evaluation

The performance statistics are listed in Table 1 and Table 2, clearly asserting the addition of a re-ranker (MiniLM-L-6-v2) substantially improves performance across all baselines, increasing both ranking precision and relevance placement. In particular:

- ColBERT + MiniLM-L-6-v2 and all-mpnet-base-v2 + MiniLM-L-6-v2 emerged as close competitors, achieving strong P@10 and nDCG scores.
- While all-mpnet-base-v2 + reranker reached the highest nDCG (0.77), ColBERT + reranker was selected as the final configuration because the main CLMIR corpus is heavily mathematics-oriented. ColBERT's late-interaction design better captures fine-grained token-level alignments, making it more suitable for formula-rich content compared to MPNet's sentence-level embeddings.
- Fine-Tuning MathBERT for ColBERT-Based Retrieval: Future work includes fine-tuning MathBERT [14] embeddings for use within the ColBERT late-interaction framework. This would combine MathBERT's domain knowledge of mathematical text with ColBERT's token-level matching capabilities.

Overall, the experiments confirm that reranking is essential for high-quality CLMIR retrieval, and ColBERT remains a robust backbone when handling mixed text–formula queries.

Table 1
Evaluation Benchmarks

Model Configuration	MAP	P@10	nDCG
ColBERTv2	0.2607	0.4000	0.4879
ColBERTv2 + miniLM-L-6-v2	0.4760	0.6476	0.6637
all-mpnet-base-v2	0.1658	0.0667	0.4604
all-mpnet-base-v2 + miniLM-L-6-v2	0.5681	0.6381	0.7698
tb17/MathBERT + miniLM-L-6-v2	0.1965	0.3667	0.3735

Table 2
Official Track Results

Metrics	Score
P@10	0.08
MAP	0.1484
nDCG	0.2202

4.4. Overall Analysis

Table 1 presents the scores assigned to the proposed method by the CLMIR 2025 organizers on the official test set. The results reflect the performance of the implemented pipeline described in Section 3. The proposed method achieves strong precision in the Top-10 retrieved results, indicating that the combination of ColBERT-based dense retrieval and cross-encoder reranking is effective for high-ranking positions. The MAP score shows that relevance is maintained throughout the ranked list, not only in the top results. While these results are promising, there is room for improvement, particularly in handling formula-rich queries where the exact mathematical structure is crucial for relevance. The future work described in Section 8 aims to address these areas by incorporating LaTeX-aware retrieval and tag-based filtering, which are expected to boost both MAP and nDCG scores by improving matching accuracy and domain specificity.

5. Conclusion

This work presents a cross-lingual mathematical information retrieval system designed for the CLMIR 2025 shared task. The proposed approach translates a Hindi mathematical corpus into English using the ai4bharat/indictrans2-indic-en-1B model, indexes the translated corpus with a ColBERTv2-based dense retrieval framework, and refines results through cross-encoder reranking. The official evaluation from the organizers demonstrates that this pipeline achieves competitive performance on P@10, MAP, and nDCG metrics, confirming its effectiveness in retrieving relevant mathematical content across languages. While the current system performs well for text–formula queries, there is scope for further enhancement. Future work will explore LaTeX-aware retrieval, dual-space search for parallel text and formula matching, and tag-based filtering for domain-specific refinement. These improvements aim to boost retrieval precision and ranking quality, especially for formula-intensive queries, ultimately contributing to more inclusive access to mathematical knowledge for Hindi- and English-speaking communities alike.

Acknowledgement

The authors express their gratitude to the organizers of the CLMIR 2025 shared task for providing the dataset, evaluation platform, and benchmark metrics that enabled this research. Appreciation is also extended to the AI4Bharat team for releasing the ai4bharat/indictrans2-indic-en-1B translation model, which served as a core component of the proposed pipeline.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Schubotz, A. Youssef, et al., A large corpus for mathematical formula search and recognition, in: Proceedings of the ACM Symposium on Document Engineering (DocEng '16), ACM, 2016. doi:10.1145/2960811.2960819, tangent-CFTED.

[2] J. Mansouri, N. Goharian, Approach0: Mathematical information retrieval via dense embeddings, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), ACM, 2022.

[3] K. L. Clarkson, Algorithms for Closest-Point Problems, Ph.D. thesis, Stanford University, Palo Alto, CA, 1985. UMI Order Number: AAT 8506171.

[4] S. Gore, A. Stupariu, et al., Crossmath: A benchmark for cross-lingual mathematical information retrieval, in: Proceedings of the European Conference on Information Retrieval (ECIR '24), Springer, 2024.

[5] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics (2020).

[6] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al., No language left behind: Scaling human-centered machine translation (2022).

[7] O. Khattab, M. Zaharia, Colbert: Efficient and effective passage search via contextualized late interaction over bert, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), ACM, 2020, pp. 39–48. doi:10.1145/3397271.3401075.

[8] K. Santhanam, O. Khattab, et al., Colbertv2: Effective and efficient retrieval via lightweight late interaction, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), ACM, 2022, pp. 332–341. doi:10.1145/3477495.3531856.

[9] Clmir 2025 shared task dataset, 2025. Shared task dataset.

[10] AI4Bharat, Indictrans2: ai4bharat/indictrans2-indic-en-1b, 2023. URL: https://aikosh.indiaai.gov.in/home/models/details/aibharat_indictrans2_en_indic_1_1b.html.

[11] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data (2019). doi:10.1109/TBDATA.2019.2921572, fAISS library.

[12] Sentence Transformers, cross-encoder/ms-marco-minilm-l-6-v2, 2020. URL: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2.

[13] Sentence Transformers, all-mpnet-base-v2, 2021. URL: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

[14] Y. Peng, J. Lu, J. Yu, et al., Mathbert: A pre-trained language model for mathematical text, arXiv preprint arXiv:2105.00377 (2021).