MIRB: Mathematical Information Retrieval Benchmark

Anonymous Authors[1]

Abstract

Mathematical Information Retrieval (MIR) is the task of retrieving information from mathematical documents and plays a key role in various applications, including theorem search in mathematical libraries, answer retrieval on math forums, and premise selection in automated theorem proving. However, a unified benchmark for evaluating these diverse retrieval tasks has been lacking. In this paper, we introduce MIRB (Mathematical Information Retrieval Benchmark) to assess the MIR capabilities of retrieval models. MIRB includes four tasks—semantic statement retrieval, question-answer retrieval, premise retrieval, and formula retrieval—spanning a total of 12 datasets. We evaluate 13 retrieval models on this benchmark and analyze the challenges inherent to MIR. We hope that MIRB provides a comprehensive framework for evaluating MIR systems and helps advance the development of more effective retrieval models tailored to the mathematical domain.[1]

1. Introduction

Mathematical Information Retrieval (MIR) (Dadure et al., 2024; Zanibbi et al., 2025) focuses on retrieving mathematical content such as definitions, theorems, and proofs from a mathematical corpus. MIR has many practical applications. For instance, mathematicians working with Lean (de Moura et al., 2015; de Moura & Ullrich, 2021) often need to verify whether a particular theorem exists in mathlib4, Lean's mathematical library. In this case, the MIR query can be either a natural language or formal statement, and the corpus consists of declarations in mathlib4. Another example is students searching for similar questions or answers on Mathematics Stack Exchange to help them solve problems. Here, the user's question serves as the query, and the corpus includes all question and answer posts on the forum. MIR is also an essential component in automated theorem proving, in both natural and formal languages. For example, NaturalProver (Welleck et al., 2022) is a natural language theorem prover that uses stepwise beam search to sample proofs, retrieving multiple references from a corpus of ProofWiki definitions and theorems to support reliable tactic generation. Similarly, ReProver (Yang et al., 2023) is a formal theorem prover for Lean that performs best-first search; at each step, it retrieves premises from mathlib4 using the current proof state as the query, and feeds the retrieved premises into a tactic generator. This retrieval step is often referred to as premise retrieval. In summary, MIR plays a crucial role in a wide range of mathematical applications.

MIR differs from standard text retrieval in that both queries and documents often contain mathematical formulas. These formulas are highly structured, and their semantic meaning typically remains unchanged under variable substitution, even though their textual representations differ. This structural property poses unique challenges for retrieval models, which must adapt to the specific characteristics of mathematical language. Due to the importance of MIR, several competitions have been organized to evaluate different MIR systems. For example, ARQMath (Zanibbi et al., 2020; Mansouri et al., 2021; 2022), held at the Conference and Labs of the Evaluation Forum (CLEF) from 2020 to 2022, includes two main tasks: answer retrieval and formula retrieval, with both queries and corpora sourced from Mathematics Stack Exchange. Similarly, the NTCIR series (Zanibbi et al., 2016) features a formula+keyword search task over corpora drawn from arXiv and Wikipedia. However, existing MIR datasets are limited in both task diversity and domain coverage, and are scattered across different sources. To the best of our knowledge, there is no unified benchmark that consolidates all major MIR tasks and datasets for a comprehensive evaluation of retrieval models.

To address this gap, we introduce **MIRB** (Mathematical Information Retrieval Benchmark), a comprehensive benchmark designed to assess retrieval models on a wide range of MIR tasks across various domains and languages. MIRB covers four main tasks: Semantic Statement Retrieval, Ques-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Our code and data are available at https://anonymous.4open.science/r/mirb-C66B and https://kaggle.com/datasets/fbb7c83309a3fa4fd4927928e537da8af6be21c617f60de21f0ba7d20d5ff94d

tion Answer Retrieval, Premise Retrieval, and Formula Retrieval, across 12 datasets in diverse mathematical domains and languages. We evaluate 13 retrieval models on this benchmark and observe that all models perform worse on reasoning-based tasks compared to semantic-based tasks. Moreover, applying cross-encoder rerankers generally leads to performance degradation. These results highlight that current retrieval models still have much room for improvement in handling MIR tasks.

The rest of the paper is organized as follows. We review the related works on retrieval benchmarks, retrieval models and mathematical information retrieval in Section 2. Section 3 describe the tasks included in MIRB and the details of the dataset construction process. Experimental results of the evaluated retrieval models are presented in Section 4, and the paper concludes in Section 5.

2. Related Work

2.1. Retrieval Benchmarks

Existing retrieval benchmarks can generally be divided into two categories: (1) general-purpose benchmarks that span diverse domains and tasks, such as BEIR (Thakur et al., 2021), MTEB (Muennighoff et al., 2023), MMTEB (Enevoldsen et al., 2025), C-MTEB (Xiao et al., 2024b) and MAIR (Sun et al., 2024); and (2) domain-specific or task-specific benchmarks that focus on a particular domain or retrieval task. For example, ChemTEB (Kasmaee et al., 2024) includes a retrieval benchmark for chemistry, while CodeSearchNet (Husain et al., 2019), CosQA (Huang et al., 2021), XcodeEval (Khan et al., 2024), and CoIR (Li et al., 2024) target code retrieval. LONGEMBED (Zhu et al., 2024) is designed for long-context retrieval. The benchmarks most closely related to our work are RAR-b (Xiao et al., 2024a) and BRIGHT (SU et al., 2025), both of which include reasoning-based retrieval datasets covering commonsense reasoning, mathematics, and code. In RAR-b's question-answer retrieval task, relevant documents directly answer the query, while BRIGHT focuses on retrieving documents that either assist in answering the query or use the same theorem as the one in the query. Our work differs from these benchmarks in three aspects: (1) we focus exclusively on the mathematics domain; (2) we include both semantic retrieval tasks (Semantic Statement Retrieval, Formula Retrieval) and reasoning-based tasks (Question-Answer Retrieval, Premise Retrieval), whereas RAR-b and BRIGHT focus solely on reasoning-based retrieval; (3) within reasoning-based retrieval, we include the task of premise retrieval in both natural and formal language, which is not covered in either RAR-b or BRIGHT.

2.2. Retrieval Models

The development of retrieval models has advanced beyond the classic BM25 algorithm (Robertson et al., 1995; Robertson & Zaragoza, 2009), which relies on sparse vector representations and measures lexical similarity between queries and documents. Modern approaches leverage deep neural networks to encode queries and documents into dense vectors, enabling relevance assessment based on semantic similarity. A widely adopted training paradigm for these dense retrieval models (Neelakantan et al., 2022; Wang et al., 2022; Su et al., 2023; Xiao et al., 2024b) involves pretraining on large-scale unsupervised data using contrastive loss, followed by fine-tuning on smaller labeled datasets. In terms of architecture, earlier models commonly employed bidirectional encoders, but recent studies (Wang et al., 2024; Meng* et al., 2024; Meng et al., 2024; Lee et al., 2025) have demonstrated that decoder-only language models can achieve superior performance. Moreover, the training data for retrieval models can be augmented with synthetic data generated by large language models (Wang et al., 2024; Muennighoff et al., 2024; Lee et al., 2024).

2.3. Mathematical Information Retrieval.

Classical mathematical information retrieval methods often rely on tree-based representations to capture the structural information of mathematical formulas, such as the Symbol Layout Tree(Zanibbi & Blostein, 2012) and the Operator Tree (Gao et al., 2016). A representative approach is the structure search used in Approach0 (Zhong & Zanibbi, 2019; Zhong et al., 2020), which computes structural similarity by identifying the largest common subexpressions and matching maximum subtrees. More recent methods combine structure-based search with dense retrieval models (Kane et al., 2022; Zhong et al., 2022a;b; 2023), allowing systems to handle both the semantic similarity of text and the structural similarity of formulas. In general, dense retrievers such as text embedding models are more robust to invalid LaTeX formulas and to formulas written in alternative formats, whereas traditional structure based methods often fail at the parsing stage if the LaTeX syntax is incorrect.

3. The MIRB Benchmark

We present **MIRB**, a benchmark designed to evaluate the mathematical information retrieval capabilities of retrieval models. It comprises four tasks: Semantic Statement Retrieval, Question-Answer Retrieval, Premise Retrieval and Formula Retrieval. Dataset statistics are provided in Table 1. The following four subsections describe each task and the corresponding dataset construction in detail.

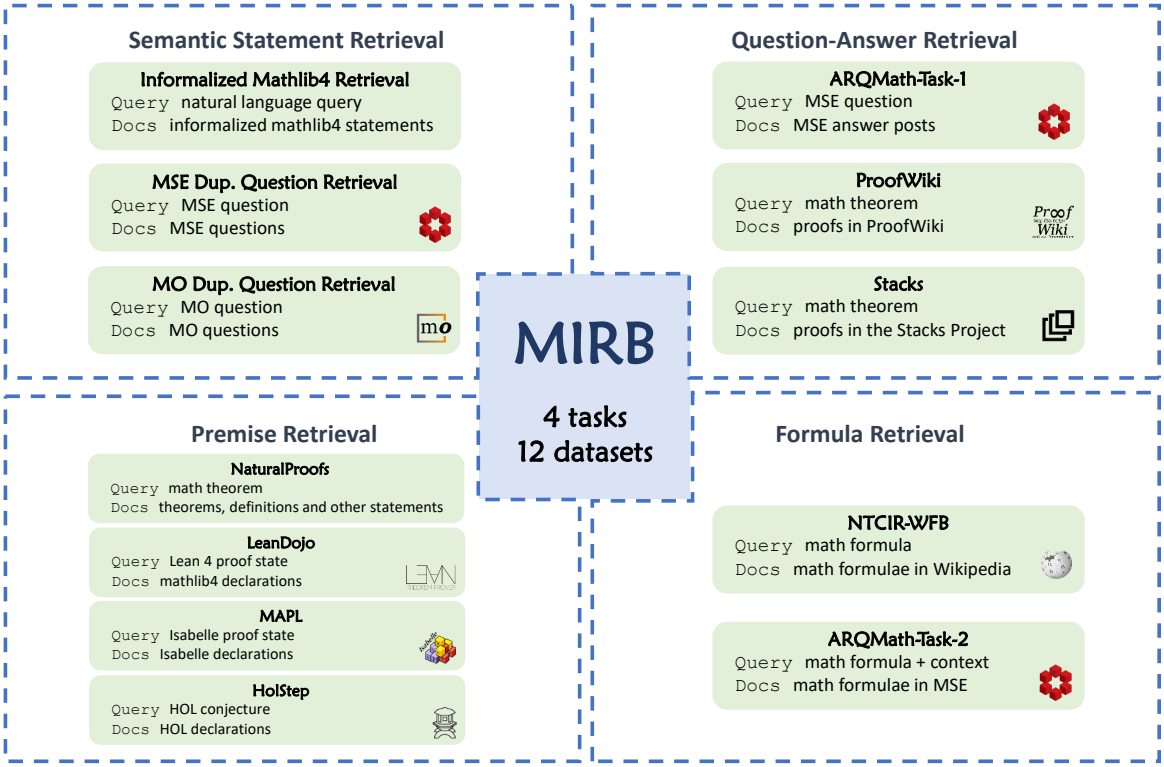

Figure 1. Overview of tasks and datasets in MIRB.

3.1. Semantic Statement Retrieval

Semantic Statement Retrieval is the task of retrieving semantically similar statements or questions given a math query, which itself is a mathematical statement or question. This task is motivated by real-world scenarios such as searching for theorems in mathematical libraries—for example, users of Lean often need to look up theorems in mathlib4. One instance of this task is Informalized Mathlib4 Retrieval, where the goal is to retrieve relevant mathlib4 theorems based on informal mathematical queries. Another instance is Duplicate Question Retrieval, which involves retrieving questions labeled as duplicates on math forums like Mathematics Stack Exchange (MSE) and Math Overflow (MO). This task is inspired by the CQADupStack dataset (Hoogeveen et al., 2015). A key challenge in this task is identifying semantically equivalent questions that may differ in phrasing or notation but express the same mathematical meaning. We construct two datasets for this purpose: MSE Duplicate Question Retrieval and MO Duplicate Question Retrieval. The details of all three datasets are discussed in the following paragraphs.

Informalized Mathlib4 Retrieval. We use the evaluation dataset from (Gao et al., 2024). The original dataset contains both formal and informal queries; in this work, we focus only on the informal queries, retaining 40 out of the original 50. The retrieval corpus consists of informalized mathlib4 statements. Relevance is graded on a three-level scale, with the criteria defined in the original paper. An example query and its relevant document are shown in Table 2.

MSE Dup. Question Retrieval. The task of Duplicate Question Retrieval involves retrieving questions that are duplicates of a given input question. We construct our dataset using the Mathematics Stack Exchange Data Dump (2024-09-30)[2]. We begin by extracting all question posts and removing those containing figures, links, or tables. Next, we build an undirected graph where an edge connects two questions if they are marked as duplicates in the data dump. We compute the transitive closure of this graph to ensure that if question A is a duplicate of B and B is a duplicate of C, then A is also considered a duplicate of C. From each connected component in the graph, we randomly sample one question to serve as a query. The remaining questions constitute the initial corpus, which we further refine. To mitigate the issue of false negatives—questions that are duplicates but not labeled as such—we adopt a dynamic corpus approach similar to the LeetCode dataset in BRIGHT (SU et al., 2025). Specifically, we extract the tags for each question from the data dump. For a query Q with tag set

[2]https://archive.org/download/stackexchange_20240930/stackexchange_20240930/math.stackexchange.com.7z

Table 1. Statistics of the datasets. We report the number of queries and documents in each dataset. **Avg. D / Q** denotes the average number of relevant documents per query. Average Word Length refers to the mean number of words per query or per document. Examples from five representative datasets (Informalized Mathlib4 Retrieval, MSE Dup. Question Retrieval, ARQMath-Task-1, NaturalProofs, NTCIR-WFB) are included in the main text, while examples from the remaining datasets are provided in the appendix.

			Test			Avg. Word Length		
Task	**Dataset**	**Relevancy**	**#query**	**#corpus**	**Avg. D / Q**	**Query**	**Document**	**Example**
Semantic Statement Retrieval	Informalized Mathlib4 Retrieval (Gao et al., 2024)	3-level	40	124,254	7.23	10.38	41.60	Table 2
	MSE Dup. Question Retrieval	Binary	25,116	1,350,505	1.78	97.22	116.42	Table 3
	MO Dup. Question Retrieval	Binary	225	108,301	1.08	100.78	144.53	Table 11
Question-Answer Retrieval	ARQMath-Task-1 (Zanibbi et al., 2020; Mansouri et al., 2021; 2022)	4-level	78	33,369	100.79	125.15	120.40	Table 4
	ProofWiki	Binary	1,099	15,763	1.03	48.37	196.87	Table 12
	Stacks	Binary	776	10,423	1.00	55.47	171.07	Table 13
Premise Retrieval	NaturalProofs (Welleck et al., 2021)	Binary	2,060	40,806	3.94	49.51	62.32	Table 5
	LeanDojo (Yang et al., 2023)	Binary	4,109	180,944	2.33	106.28	30.18	Table 14
	MAPL (Mikuła et al., 2024)	Binary	4,000	493,029	7.07	43.53	30.15	Table 15
	HolStep (Kaliszyk et al., 2017)	Binary	1,411	3,973	22.82	34.33	28.84	Table 16
Formula Retrieval	NTCIR-WFB (Zanibbi et al., 2016)	3-level	39	1,994	38.95	2.72	2.93	Table 6
	ARQMath-Task-2 (Zanibbi et al., 2020; Mansouri et al., 2021; 2022)	4-level	76	9,969	63.18	122.25	5.61	Table 17

Table 2. Informalized Mathlib4 Retrieval example.

Query	**Relevant Document**
Let L/K be a Galois extension, F be an intermidiate field, then $L^{\{\sigma \in \text{Gal}(L/K) \mid \sigma x = x, \forall x \in F\}} = F$	Fixed Field of Fixing Subgroup Theorem: For a Galois field extension E/F with an intermediate field K, the fixed field of the subgroup fixing K is equal to K.

Table 3. MSE Dup. Question Retrieval example.

Query	**Relevant Document**
Example of divisor D such that $\deg D > 0$ and $\ell(D) = 0$ It is easy to see that if a divisor D on a projective curve C over a field K has negative degree, then $\ell(D) = \dim_K\{f \in K(C) \mid div(f) + D \geq 0\}$ is zero. However, I suppose that the converse is not true. Can someone give me the simplest example of a divisor D on some curve C satisfying $\deg(D) > 0$ but $\ell(D) = 0$?	Does the dual of a line bundle with no sections have a section? Let $L \to X$ be a holomorphic line bundle over a compact complex manifold. Suppose L is non-trivial and has no non-trivial sections. Let me ask the following (hopefully not entirely trivial) question: Does the dual L^* have a non-trivial section? A special case of this is when L is the dual of an ample line bundle. Obviously ample line bundles have sections, but the dual does not.

$T(Q)$, we exclude a candidate question Q' from its corpus if the tag overlap satisfies $\frac{|T(Q) \cap T(Q')|}{|T(Q)|} \geq 0.5$. This ensures that, aside from the ground-truth duplicates, most questions in the corpus are not on the same topic as the query, thus reducing the risk of unlabeled duplicates appearing as false negatives.

MO Dup. Question Retrieval. The construction of the MO Duplicate Question Retrieval dataset follows the same procedure as for the MSE dataset. We use the MathOverflow Data Dump (2024-09-30)[3]. After cleaning the question posts, applying transitive closure to the graph, and filtering the corpus, we obtain 225 queries and 108,301 documents.

3.2. Question-Answer Retrieval

Question-Answer Retrieval focuses on retrieving relevant answers or proofs for a given mathematical question. The main challenge lies in understanding the underlying mathematical intent of the question and identifying documents that provide accurate and precise answers—an objective that goes beyond simple semantic similarity. We include three datasets for this task: ARQMath-Task-1, ProofWiki, and Stacks, which are discussed in the following paragraphs.

ARQMath-Task-1. ARQMath-Task-1 (Zanibbi et al., 2020; Mansouri et al., 2021; 2022) is an answer retrieval task, where the goal is to retrieve relevant answer posts from Mathematics Stack Exchange (MSE) between 2010 and 2018, given a query question posted after 2019. The task was held over three years, with the query sets consisting of MSE questions from 2019, 2020, and 2021, respectively. We use ARQMath-3-Task-1 as the test set. The ARQMath-3-Task-1 dataset contains 78 queries, with an average of 446.8 annotated answers per query. Relevance is graded on four levels, and readers may refer to (Mansouri et al., 2022) for the detailed relevance criteria. The evaluation metric is

[3] https://archive.org/download/stackexchange_20240930/stackexchange_20240930/mathoverflow.net.7z

Table 4. ARQMath-Task-1 example.

Query	Relevant Document
Confusion about the formula of the area of a surface of revolution Before I read the formula of the area of revolution which is $\int 2\pi y\, ds$, where $ds = \sqrt{1 + \frac{dy}{dx}^2}$, I thought of deriving it myself. I tried to apply the same logic used for calculating the volume of revolution (e.g., $\int \pi y^2 dx$). My idea is to use many tiny hollow cylinders (inspired from the shell method), each has a surface area of $(2\pi y)(dx)$: $2\pi y$ is the circumference of the cylinder, and dx is the height of the cylinder Their product is the surface area of the hollow (e.g., empty from the inside) cylinder. With this logic, the area is $\int 2\pi y dx$. Where is my mistake? Also it's confusing why for the volume it was enough to partition the object using cylinders and for areas not.	You should review the formula for the surface area in the case of a surface of revolution (e.g. here). The surface area of the surface obtained by rotation the graph of $y = f(x)$ about the x-axis on the interval $[x_1, x_2]$, is given by: $2\pi \int_{x_1}^{x_2} y\sqrt{1 + (y')^2}\,\mathrm{d}x = 2\pi \int_{x_1}^{x_2} f(x)\sqrt{1 + (f'(x))^2}\,\mathrm{d}x$ Now if $f(x) = \frac{\cosh(4x)}{4}$, then $f'(x) = \sinh(4x)$ so rotation on $[-1, 1]$ gives: $\frac{\pi}{2}\int_{-1}^{1} \cosh(4x)\sqrt{1 + \sinh^2(4x)}\,\mathrm{d}x$ You can simplify (a lot). Can you take it from here? I also need to know how would one go about rotating this about the y-axis, but have no idea where to start. The link from above also covers the formula for rotation about the y-axis.

nDCG-prime, introduced in (Sakai & Kando, 2008), which excludes unjudged documents from the ranked list. As a result, we adopt a dynamic corpus approach, where the corpus for each query consists only of its associated annotated documents.

ProofWiki. ProofWiki is a mathematical library containing definitions, axioms, theorems, and their corresponding proofs. In the ProofWiki Question-Answer Retrieval task, the queries are theorems from ProofWiki, and the corpus consists of proofs sourced from the same platform. The objective is to retrieve the correct proof(s) for a given theorem. Since some theorems in ProofWiki have multiple proofs, the average number of relevant documents per query is greater than one. We use the theorems from the test set of the ProofWiki dataset in NaturalProofs (Welleck et al., 2021) as queries, and include all proofs from the dataset, not just those associated with the queries, as the retrieval corpus.

Table 5. NaturalProofs example.

Query	Relevant Document
If $\mathcal{H}$ is an open covering of a closed and bounded subset S of the real line, then S has an open covering $\widetilde{\mathcal{H}}$ consisting of finitely many open sets belonging to $\mathcal{H}$.	no point of S^c is a limit point of S.

Stacks. The Stacks Project is a mathematical library focused on algebraic stacks and algebraic geometry. Similar to the ProofWiki Question-Answer Retrieval task, Stacks Question-Answer Retrieval aims to retrieve the correct proof for a given theorem in the Stacks Project. We use theorems from the test set of the Stacks dataset in NaturalProofs (Welleck et al., 2021) as queries, and include all proofs from the dataset as the retrieval corpus.

3.3. Premise Retrieval

Premise retrieval is the task of retrieving definitions, theorems, and lemmas that are useful for proving a target theorem or advancing the current proof state. This task plays a crucial role in automated theorem proving, where the ability to efficiently identify relevant mathematical premises can greatly influence the success of the proof process (Mikuła et al., 2024; Yang et al., 2023). We include four datasets for this task: one natural language premise retrieval dataset, NaturalProofs (Welleck et al., 2021), and three formal premise retrieval datasets: LeanDojo (Yang et al., 2023) for Lean, MAPL (Mikuła et al., 2024) for Isabelle, and HolStep (Kaliszyk et al., 2017) for HOL Light. The details of these four datasets are discussed in the following paragraphs.

NaturalProofs. NaturalProofs (Welleck et al., 2021) is a natural language premise retrieval dataset, where the goal is to retrieve definitions, lemmas, and theorems that are useful for proving a given query statement. It consists of four subsets: ProofWiki, Stacks, Real Analysis, and Number Theory. In the ProofWiki subset, the query is a theorem from ProofWiki, the corpus includes all definitions, lemmas, and theorems in the library, and the relevant documents are those used in the proof of the query theorem. The other three subsets follow a similar formulation. We evaluate each subset separately and report the average of their scores as the final result for the NaturalProofs dataset.

LeanDojo. LeanDojo (Yang et al., 2023) provides a premise retrieval dataset for Lean, where the goal is to retrieve useful premises from mathlib4 to advance a given

Table 6. NTCIR-WFB example.

Query		Relevant Document
$L(\lambda, \alpha, s)$ $\sum_{n=0}^{\infty} \frac{\exp(2\pi i \lambda n)}{(n+\alpha)^s}$.	=	$g(s) = \sum_{n=1}^{\infty} \frac{a(n)}{n^s}$

Lean 4 proof state. In this task, the query is a proof state, the corpus consists of all mathlib4 declarations, and the relevant documents are the premises used in the next tactic step. We follow the `novel_premises` data split from the original benchmark, in which each proof in the test set uses at least one premise not seen during training.

MAPL. MAPL (Mikuła et al., 2024) is a premise retrieval dataset for Isabelle. The task is similar to that of LeanDojo premise retrieval, where the goal is to retrieve useful premises to advance the current proof state. In MAPL, the query is an Isabelle proof state and the corpus consists of premises expressed in Isabelle's formal language. The original dataset comprises a collection of (state, premise) pairs, which we split into train, dev, and test sets following a strategy similar to the `novel_premises` split in LeanDojo. Specifically, each proof state in the test set uses at least one premise that does not appear in the training set.

HolStep. HolStep (Kaliszyk et al., 2017) is a dataset based on HOL Light proofs. Each file in the original dataset contains a single conjecture along with the dependencies used in its proof. We treat the conjectures as queries and aggregate all dependencies across the dataset to form the retrieval corpus. The task is to retrieve the relevant dependencies for a given conjecture.

3.4. Formula Retrieval

Formula retrieval focuses on retrieving mathematical expressions that are relevant to a given query formula, optionally incorporating the formula's surrounding context. This task requires a deep understanding of the semantic meaning of mathematical formulas. We evaluate this task using two datasets: NTCIR-12 Wikipedia Formula Browsing (WFB) (Zanibbi et al., 2016) and ARQMath-Task 2 (Zanibbi et al., 2020; Mansouri et al., 2021; 2022).

NTCIR-WFB. The NTCIR-12 Wikipedia Formula Browsing task involves retrieving relevant formulas given a query formula. The corpus consists of mathematical formulas extracted from Wikipedia articles. Relevance is graded on a three-level scale, with detailed criteria provided in (Zanibbi et al., 2016). Similar to ARQMath-Task-1, we adopt a dynamic corpus approach, where each query is evaluated against only its associated annotated documents.

ARQMath-Task-2. ARQMath-Task-2 (Zanibbi et al., 2020; Mansouri et al., 2021; 2022) is a formula retrieval task, where the goal is to retrieve relevant formulas from MSE posts given a query formula along with its context (i.e., the question post in which it appears). We use ARQMath-3-Task-2 as the test set, which contains 76 queries and an average of 63.18 annotated relevant documents per query. The task defines four levels of relevance, with criteria detailed in (Mansouri et al., 2022). Similar to ARQMath-Task-1, we adopt a dynamic corpus approach, where each query's corpus consists only of its annotated documents.

4. Experiments

In this section, we evaluate the performance of 13 retrieval models on MIRB. The experimental setup is described in SubSection 4.1, and the comparison of model performance is presented in SubSection 4.2.

4.1. Experiment Setup

We evaluate four groups of retrieval models. For the sparse model, we test BM25. For open-source models with fewer than 1 billion parameters, we include gte-large-en-v1.5 (Li et al., 2023), UAE-Large-V1 (Li & Li, 2024), and bge-large-en-v1.5 (Xiao et al., 2024b). For open-source models with more than 1 billion parameters, we evaluate gte-Qwen2-1.5B-instruct (Li et al., 2023), e5-mistral-7b-instruct (Wang et al., 2024), NV-Embed-v2 (Lee et al., 2025), gte-Qwen2-7B-instruct (Li et al., 2023), SFR-Embedding-2_R (Meng* et al., 2024), and GritLM-7B (Muennighoff et al., 2024). For proprietary models, we evaluate Cohere-embed-english-v3.0[4], text-embedding-3-large [5], and voyage-3-large[6].

For dense models, we compute the cosine similarity between the query embedding and the corpus embeddings, and return a ranked list of documents. Model configurations, including the maximum context length for queries and documents, as well as whether instructions are prepended to the queries, are provided in Table 7. The instructions used are listed in Table 8. Following prior work (Thakur et al., 2021; SU et al., 2025), we report nDCG@10 as the main evaluation metric.

4.2. Results

Main Results The results are shown in Table 9. BM25 underperforms compared to dense retrievers, and there is a clear performance gap between small models (fewer than 1B parameters) and larger models (around 7B). voyage-3-large outperforms all other models, achieving an average

[4] https://huggingface.co/Cohere/Cohere-embed-english-v3.0

[5] https://platform.openai.com/docs/models/text-embedding-3-large

[6] https://huggingface.co/voyageai/voyage-3-large

Table 7. Model configuration. Max $|Q|$ and Max $|D|$ is the maximum context length we set for each model. The instruction column denotes whether we prepend instructions to the query.

	Size	Max $\|Q\|$	Max $\|D\|$	Instruction
Sparse model				
BM25	-	-	-	No
Open-source models (<1B)				
gte-large-en-v1.5	434M	8192	8192	No
UAE-Large-V1	335M	512	512	Yes
bge-large-en-v1.5	335M	512	512	Yes
Open-source models (>1B)				
gte-Qwen2-1.5B-instruct	1.78B	4096	4096	Yes
e5-mistral-7b-instruct	7.11B	4096	4096	Yes
NV-Embed-v2	7.85B	32768	32768	Yes
gte-Qwen2-7B-instruct	7.61B	4096	4096	Yes
SFR-Embedding-2_R	7.11B	4096	4096	Yes
GritLM-7B	7.24B	4096	4096	Yes
Proprietary models				
Cohere-embed-english-v3.0	-	512	512	No
text-embedding-3-large	-	8192	8192	No
voyage-3-large	-	32000	32000	Yes

Table 8. Instructions used for different datasets are applied to all models that utilize instructions, except for UAE-Large-V1 and bge-large-en-v1.5. For these two models, the instruction used is: "Represent this sentence for searching relevant passages:"

Dataset	Instruction
Informalized Mathlib4 Retrieval	Given a mathematical query, retrieve relevant theorems.
MSE Dup. Question Retrieval MO Dup. Question Retrieval	Given a math question, retrieve questions that are duplicates of the given one
ARQMath-Task-1	Given a math problem, retrieve its solution.
ProofWiki Stacks	Given a math theorem, retrieve its proof.
NaturalProofs	Given a math theorem, retrieve useful references, such as theorems, lemmas, and definitions, that are useful for proving the given theorem.
LeanDojo	Given a Lean 4 proof state, retrieve the declarations that are useful for proving it.
MAPL	Given an Isabelle proof state, retrieve the declarations that are useful for proving it.
HolStep	Given a HOL conjecture, retrieve the declarations that are useful for proving it.
NTCIR-WFB	Given a math formula, retrieve relevant formulas.
ARQMath-Task-2	Given a math formula and its context, retrieve relevant formulas.

nDCG@10 score of 54.54 and ranking first on 7 out of the 12 datasets. Among the evaluated tasks, models generally perform better on semantic retrieval tasks such as Semantic Statement Retrieval and Formula Retrieval, while their performance degrades on reasoning-oriented tasks, especially Premise Retrieval. Unlike Question-Answer Retrieval, where the solution or part of it appears in the document, Premise Retrieval requires identifying relevant mathematical statements such as lemmas or theorems that are not part of the answer but are useful for constructing a proof. For formal premise retrieval datasets like LeanDojo, MAPL, and HolStep, embedding models often struggle because they are not extensively pre-trained on large corpora of formal language data. As a result, they are unfamiliar with the notation and syntax of formal languages, and are even less capable of identifying the underlying logical connections between the query state and potential premises. Consequently, even models that perform well on Question-Answer Retrieval (e.g., voyage-3-large) show poor performance on Premise Retrieval. To improve performance on this task, models need to be trained on premise retrieval datasets across different formal languages.

Results of Reranking Applying rerankers to retrieval results is generally expected to improve performance. To assess their effectiveness on mathematical retrieval tasks, we evaluate two rerankers: bge-reranker-v2-m3 (Chen et al., 2024) and jina-reranker-v2-base-multilingual[7]. Each reranker computes a relevance score for the concatenated query and document pair, and then reranks the top 10 retrieved documents accordingly. We apply them to the top five models in MIRB: voyage-3-large, SFR-Embedding-2_R,

[7] https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

Table 9. The performance of retrieval models in MIRB. We report nDCG@10 for all datasets. Avg. denotes the average score across datasets. The best score for each dataset is highlighted in bold.

	Semantic Statement Retrieval			Question-Answer Retrieval			Premise Retrieval				Formula Retrieval		Avg.
	Informalized Mathlib4 Retrieval	MSE Dup. Question Retrieval	MO Dup. Question Retrieval	ARQMath-Task-1	ProofWiki	Stacks	NaturalProofs	LeanDojo	MAPL	HolStep	NTCIR-WFB	ARQMath-Task-2	
Sparse model													
BM25	31.49	22.85	44.01	24.83	57.35	35.49	24.14	6.91	15.27	25.88	66.03	32.46	32.23
Open-source models (<1B)													
gte-large-en-v1.5	38.05	46.76	68.04	37.78	66.49	32.26	28.42	3.73	8.78	29.15	68.83	59.87	40.68
UAE-Large-V1	40.43	41.11	67.44	31.66	54.81	28.17	27.85	4.64	5.59	30.17	71.92	55.50	38.27
bge-large-en-v1.5	41.99	41.70	67.40	31.02	56.36	30.25	27.53	5.45	6.84	30.51	73.76	55.22	39.00
Open-source models (>1B)													
gte-Qwen2-1.5B-instruct	55.17	43.13	67.73	41.97	77.83	52.56	27.46	8.40	18.64	28.05	72.96	53.56	45.62
e5-mistral-7b-instruct	57.33	51.14	71.31	46.46	77.29	39.85	32.14	10.80	15.41	30.27	**78.48**	57.93	47.37
NV-Embed-v2	59.48	55.00	78.47	47.34	83.08	58.56	37.21	12.27	16.58	**32.77**	73.22	70.00	52.00
gte-Qwen2-7B-instruct	40.38	38.40	61.77	44.74	77.02	49.35	30.08	11.53	17.46	28.16	77.52	54.68	44.26
SFR-Embedding-2_R	**60.98**	58.52	81.32	51.15	85.07	54.94	**34.67**	11.83	17.07	30.76	75.69	65.48	52.29
GritLM-7B	54.09	53.05	78.60	46.35	81.59	55.89	32.92	10.68	19.53	30.80	74.22	66.56	50.36
Proprietary models													
Cohere-embed-english-v3.0	42.00	42.96	61.00	38.05	66.00	32.33	28.99	6.96	13.95	29.72	73.27	54.51	40.81
text-embedding-3-large	49.38	52.35	76.74	45.79	81.95	56.14	31.33	11.34	**19.94**	31.02	73.06	70.18	49.93
voyage-3-large	57.36	**60.33**	**82.87**	**52.45**	**91.69**	**62.62**	32.74	**13.02**	17.77	32.68	76.91	**74.00**	**54.54**

NV-Embed-v2, GritLM-7B and text-embedding-3-large, to assess whether reranking improves performance. The results, shown in Table 10, indicate that reranking generally leads to a decline in performance. In a few cases, slight improvements are observed: for example, jina-reranker-v2-base-multilingual raises the score of voyage-3-large on ARQMath-Task-1 from 52.45 to 53.03, and improves SFR-Embedding-2_R on NTCIR-WFB from 75.69 to 76.13. These results suggest that rerankers trained on general text retrieval tasks may not transfer effectively to mathematical retrieval.

5. Conclusion

In this paper, we introduce MIRB, a comprehensive benchmark designed to evaluate the mathematical information retrieval capabilities of retrieval models. MIRB comprises four tasks: Semantic Statement Retrieval, Question-Answer Retrieval, Premise Retrieval, and Formula Retrieval. These tasks span both semantic-based and reasoning-based retrieval settings. We evaluate 13 retrieval models and observe that while their performance on semantic-based retrieval is moderate, they perform poorly on reasoning-based tasks. Additionally, applying cross-encoder rerankers does not lead to performance improvements. We hope that MIRB will facilitate future research in mathematical information retrieval and support the development of more effective retrieval models tailored to mathematics.

Impact Statement

We introduce a unified benchmark for mathematical information retrieval, aiming to encourage the development of more effective retrieval models. We hope this benchmark helps advance search engines and automated theorem proving systems by driving progress in math-specific retrieval capabilities.

References

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Dadure, P., Pakray, P., and Bandyopadhyay, S. Mathematical information retrieval: A review. *ACM Computing Surveys*, 57(3):1–34, 2024.

de Moura, L. and Ullrich, S. The Lean 4 Theorem Prover and Programming Language. In Platzer, A. and Sutcliffe, G. (eds.), *Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings*, volume 12699 of *Lecture Notes in Computer Science*, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37. URL https://doi.org/10.1007/978-3-030-79876-5_37.

de Moura, L. M., Kong, S., Avigad, J., van Doorn, F., and von Raumer, J. The Lean Theorem Prover (System Description). In Felty, A. P. and Middeldorp, A. (eds.), *Automated Deduction - CADE-25 - 25th International Conference on Automated Deduction, Berlin, Germany, August 1-7, 2015, Proceedings*, volume 9195 of *Lecture Notes in Computer Science*, pp. 378–388. Springer, 2015. doi: 10.1007/978-3-319-21401-6_26. URL https://doi.org/10.1007/978-3-319-21401-6_26.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

Table 10. Results of reranking. Each $\cdot / \cdot / \cdot$ represents the score before reranking, after applying the bge-reranker-v2-m3, and after applying the jina-reranker-v2-base-multilingual, respectively.

	NV-Embed-v2	SFR-Embedding-2_R	GritLM-7B	text-embedding-3-large	voyage-3-large
Informalized Mathlib4 Retrieval	59.48 / 55.19 / 56.61	60.98 / 55.29 / 57.16	54.09 / 52.39 / 54.71	49.38 / 46.92 / 48.23	57.36 / 55.22 / 56.06
MSE Dup. Question Retrieval	55.00 / 47.23 / 49.00	58.52 / 50.28 / 52.25	53.05 / 46.58 / 48.22	52.35 / 46.06 / 47.45	60.33 / 52.93 / 54.84
MO Dup. Question Retrieval	78.47 / 64.52 / 70.96	81.32 / 66.09 / 72.00	78.60 / 63.70 / 69.00	76.74 / 61.77 / 67.95	82.87 / 66.78 / 73.15
ARQMath-Task-1	47.34 / 46.66 / 47.17	51.15 / 50.41 / 50.52	46.35 / 43.45 / 44.36	45.79 / 43.30 / 44.09	52.45 / 51.41 / 53.03
ProofWiki	83.08 / 67.85 / 73.09	85.07 / 69.29 / 74.05	81.59 / 67.26 / 72.23	81.95 / 67.41 / 72.51	91.69 / 70.33 / 75.91
Stacks	58.56 / 44.67 / 51.68	54.94 / 41.07 / 49.47	55.89 / 42.03 / 49.60	56.14 / 41.87 / 50.93	62.62 / 45.76 / 53.75
NaturalProofs	37.21 / 33.32 / 32.33	34.67 / 31.44 / 30.39	32.92 / 31.00 / 29.76	31.33 / 29.22 / 28.23	32.74 / 30.70 / 29.82
LeanDojo	12.27 / 10.59 / 11.37	11.83 / 10.36 / 11.24	10.68 / 9.59 / 10.33	11.34 / 9.96 / 10.86	13.02 / 11.10 / 11.83
MAPL	16.58 / 16.56 / 16.81	17.07 / 17.10 / 17.27	19.53 / 18.73 / 18.99	19.94 / 18.64 / 19.45	17.77 / 17.06 / 18.08
HolStep	32.77 / 31.94 / 31.01	30.76 / 30.37 / 29.46	30.80 / 30.44 / 29.35	31.02 / 30.23 / 29.33	32.68 / 31.44 / 30.68
NTCIR-WFB	73.22 / 71.67 / 73.84	75.69 / 74.42 / 76.13	74.22 / 73.04 / 73.83	73.06 / 71.21 / 72.27	76.91 / 74.69 / 76.21
ARQMath-Task-2	70.00 / 69.06 / 67.69	65.48 / 64.96 / 64.99	66.56 / 65.41 / 64.92	70.18 / 69.43 / 67.26	74.00 / 72.90 / 71.65
Avg.	52.00 / 46.60 / 48.46	52.29 / 46.76 / 48.74	50.36 / 45.30 / 47.11	49.93 / 44.67 / 46.55	54.54 / 48.36 / 50.42

Gao, G., Ju, H., Jiang, J., Qin, Z., and Dong, B. A Semantic Search Engine for Mathlib4. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8001–8013, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.470. URL https://aclanthology.org/2024.findings-emnlp.470/.

Gao, L., Yuan, K., Wang, Y., Jiang, Z., and Tang, Z. The Math Retrieval System of ICST for NTCIR-12 MathIR Task. In Kando, N., Sakai, T., and Sanderson, M. (eds.), *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences, Tokyo, Japan, June 7-10, 2016*. National Institute of Informatics (NII), 2016. URL http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/MathIR/03-NTCIR12-MathIR-GaoL.pdf.

Hoogeveen, D., Verspoor, K. M., and Baldwin, T. CQADupStack: A Benchmark Data Set for Community Question-Answering Research. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450340403. doi: 10.1145/2838931.2838934. URL https://doi.org/10.1145/2838931.2838934.

Huang, J., Tang, D., Shou, L., Gong, M., Xu, K., Jiang, D., Zhou, M., and Duan, N. CoSQA: 20,000+ Web Queries for Code Search and Question Answering. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5690–5700, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.442. URL https://aclanthology.org/2021.acl-long.442/.

Husain, H., Wu, H.-H., Gazit, T., Allamanis, M., and Brockschmidt, M. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

Kaliszyk, C., Chollet, F., and Szegedy, C. HolStep: A Machine Learning Dataset for Higher-order Logic Theorem Proving. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=ryuxYmvel.

Kane, A., Ng, Y. K., and Tompa, F. W. Dowsing for Answers to Math Questions: Doing Better with Less. In Faggioli, G., Ferro, N., Hanbury, A., and Potthast, M. (eds.), *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pp. 40–62. CEUR-WS.org, 2022. URL https://ceur-ws.org/Vol-3180/paper-03.pdf.

Kasmaee, A. S., Khodadad, M., Saloot, M. A., Sherck, N., Dokas, S., Mahyar, H., and Samiee, S. ChemTEB: Chemical Text Embedding Benchmark, an Overview of Embedding Models Performance & Efficiency on a Specific Domain. *arXiv preprint arXiv:2412.00532*, 2024.

Khan, M. A. M., Bari, M. S., Do, X. L., Wang, W., Parvez, M. R., and Joty, S. XCodeEval: An Execution-based Large Scale Multilingual Multitask Benchmark for Code Understanding, Generation, Translation and Retrieval. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6766–6805, Bangkok, Thailand,

August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.367. URL https://aclanthology.org/2024.acl-long.367/.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B., and Ping, W. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., Hui, K., Boratko, M., Kapadia, R., Ding, W., et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024.

Li, X. and Li, J. AoE: Angle-optimized Embeddings for Semantic Textual Similarity. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1825–1839, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.101. URL https://aclanthology.org/2024.acl-long.101/.

Li, X., Dong, K., Lee, Y. Q., Xia, W., Zhang, H., Dai, X., Wang, Y., and Tang, R. Coir: A comprehensive benchmark for code information retrieval models. *arXiv preprint arXiv:2407.02883*, 2024.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Mansouri, B., Zanibbi, R., Oard, D. W., and Agarwal, A. Overview of ARQMath-2 (2021): Second CLEF Lab on Answer Retrieval for Questions on Math. In Candan, K. S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., and Ferro, N. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 215–238, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85251-1.

Mansouri, B., Agarwal, A., Oard, D. W., and Zanibbi, R. Advancing Math-Aware Search: The ARQMath-3 Lab at CLEF 2022. In Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., and Setty, V. (eds.), *Advances in Information Retrieval*, pp. 408–415, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99739-7.

Meng*, R., Liu*, Y., Joty, S. R., Caiming Xiong, Y. Z., and Yavuz, S. SFR-Embedding-2: Advanced Text Embedding with Multi-stage Training, 2024. URL https://huggingface.co/Salesforce/SFR-Embedding-2_R.

Meng, R., Liu, Y., Joty, S. R., Xiong, C., Zhou, Y., and Yavuz, S. SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

Mikuła, M., Tworkowski, S., Antoniak, S., Piotrowski, B., Jiang, A. Q., Zhou, J. P., Szegedy, C., Kuciński, Ł., Miłoś, P., and Wu, Y. Magnushammer: A Transformer-Based Approach to Premise Selection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oYjPk8mqAV.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive Text Embedding Benchmark. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Muennighoff, N., Hongjin, S., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., and Kiela, D. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., and Weng, L. Text and Code Embeddings by Contrastive Pre-Training. *CoRR*, abs/2201.10005, 2022. URL https://arxiv.org/abs/2201.10005.

Robertson, S. and Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. Okapi at TREC-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126. Gaithersburg, MD: NIST, January 1995. URL https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/.

Sakai, T. and Kando, N. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11:447–470, 2008.

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.-t., Smith, N. A., Zettlemoyer, L., and Yu, T. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.71. URL https://aclanthology.org/2023.findings-acl.71.

SU, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., yu Wang, H., Haisu, L., Shi, Q., Siegel, Z. S., Tang, M., Sun, R., Yoon, J., Arik, S. O., Chen, D., and Yu, T. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ykuc5q381b.

Sun, W., Shi, Z., Wu, J., Yan, L., Ma, X., Liu, Y., Cao, M., Yin, D., and Ren, Z. MAIR: A Massive Benchmark for Evaluating Instructed Retrieval. *arXiv preprint arXiv:2410.10127*, 2024.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *CoRR*, abs/2212.03533, 2022. doi: 10.48550/ARXIV.2212.03533. URL https://doi.org/10.48550/arXiv.2212.03533.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Improving Text Embeddings with Large Language Models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Welleck, S., Liu, J., Bras, R. L., Hajishirzi, H., Choi, Y., and Cho, K. NaturalProofs: Mathematical Theorem Proving in Natural Language. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=Jvxa8adr3iY.

Welleck, S., Liu, J., Lu, X., Hajishirzi, H., and Choi, Y. Naturalprover: Grounded mathematical proof generation with language models. *Advances in Neural Information Processing Systems*, 35:4913–4927, 2022.

Xiao, C., Hudson, G. T., and Moubayed, N. A. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*, 2024a.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., and Nie, J.-Y. C-Pack: Packed Resources For General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Yang, K., Swope, A. M., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R., and Anandkumar, A. LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=g7OX2sOJtn.

Zanibbi, R. and Blostein, D. Recognition and retrieval of mathematical expressions. *Int. J. Document Anal. Recognit.*, 15(4):331–357, 2012. doi: 10.1007/S10032-011-0174-4. URL https://doi.org/10.1007/s10032-011-0174-4.

Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., and Davila, K. NTCIR-12 MathIR Task Overview. In *NTCIR*, 2016.

Zanibbi, R., Oard, D. W., Agarwal, A., and Mansouri, B. Overview of ARQMath 2020: CLEF Lab on Answer Retrieval for Questions on Math. In Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., and Ferro, N. (eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 169–193, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58219-7.

Zanibbi, R., Mansouri, B., Agarwal, A., et al. Mathematical information retrieval: Search and question answering. *Foundations and Trends® in Information Retrieval*, 19 (1-2):1–190, 2025.

Zhong, W. and Zanibbi, R. Structural Similarity Search for Formulas Using Leaf-Root Paths in Operator Subtrees. In Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., and Hiemstra, D. (eds.), *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18,*

2019, Proceedings, Part I, volume 11437 of *Lecture Notes in Computer Science*, pp. 116–129. Springer, 2019. doi: 10.1007/978-3-030-15712-8_8. URL https://doi.org/10.1007/978-3-030-15712-8_8.

Zhong, W., Rohatgi, S., Wu, J., Giles, C. L., and Zanibbi, R. Accelerating Substructure Similarity Search for Formula Retrieval. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I*, pp. 714–727, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-45438-8. doi: 10.1007/978-3-030-45439-5_47. URL https://doi.org/10.1007/978-3-030-45439-5_47.

Zhong, W., Xie, Y., and Lin, J. Applying Structural and Dense Semantic Matching for the ARQMath Lab 2022, CLEF. In Faggioli, G., Ferro, N., Hanbury, A., and Potthast, M. (eds.), *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pp. 147–170. CEUR-WS.org, 2022a. URL https://ceur-ws.org/Vol-3180/paper-09.pdf.

Zhong, W., Yang, J., Xie, Y., and Lin, J. Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 1092–1102. Association for Computational Linguistics, 2022b. doi: 10.18653/V1/2022.FINDINGS-EMNLP.78. URL https://doi.org/10.18653/v1/2022.findings-emnlp.78.

Zhong, W., Lin, S.-C., Yang, J.-H., and Lin, J. One Blade for One Purpose: Advancing Math Information Retrieval using Hybrid Search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 141–151, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394086. doi: 10.1145/3539618.3591746. URL https://doi.org/10.1145/3539618.3591746.

Zhu, D., Wang, L., Yang, N., Song, Y., Wu, W., Wei, F., and Li, S. LongEmbed: Extending Embedding Models for Long Context Retrieval. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 802–816, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.47. URL https://aclanthology.org/2024.emnlp-main.47/.

A. Dataset Examples

In this section, we present examples of datasets from MIRB that are not included in the main text.

Table 11. MO Dup. Question Retrieval example.

Query	Relevant Document
On finite subsets of set of integers, which lies in its sum-set , whose sum of elements equals 0 Let $n > 1$ be an integer and $S \subseteq \mathbb{Z}$ be such that $\|S\| = n$ and $S \subseteq S + S := \{a + b : a, b \in S\}$; then does there exist $T \subseteq S$ with $1 \leq \|T\| \leq n/2$ such that $\sum_{a \in T} = 0$?	Existence of a zero-sum subset Some time ago I heard this question and tried playing around with it. I've never succeeded to making actual progress. Here it goes: Given a finite (nonempty) set of real numbers, $S = \{a_1, a_2, \ldots, a_n\}$, with the property that for each i there exist j, k (not necessarily distinct) so that $a_i = a_j + a_k$ (i.e. every element in S can be written as a sum of two elements in S, note that this condition is trivially satisfied if $0 \in S$ as then every $x \in S$ can be written as $x + 0$). Must there exist $\{i_1, i_2, \ldots, i_m\}$ (distinct) so that $a_{i_1} + a_{i_2} + \cdots + a_{i_m} = 0$? ETA: A possible reformulation can be made in terms of graphs. We can take the vertex set $\{1, \ldots, n\}$ and for each equation $a_i = a_j + a_k$ in S add an edge $[ij]$ and its "dual" $[ik]$. The idea is to find a cycle in this graph, whose dual is a matching.

Table 12. ProofWiki example.

Query	Relevant Document
Fortissimo Space is not Weakly Countably Compact Let \$T = \struct S, \tau_p\$ be a Fortissimo space. Then \$T\$ is not weakly countably compact.	It suffices to show that T has an infinite subset without limit points. Consider the set $S \setminus \{p\}$. Let $x \in S$. We have: {{begin-eqn}} {{eqn — l = \paren {S \setminus \paren {S \setminus \set p} } \cup \set x — r = \set p \cup \set x — c = }} {{eqn — r = \set {p, x} — c = }} {{end-eqn}} By definition, x is a limit point of $S \setminus \{p\}$ iff $\{p, x\}$ is not a neighborhood of x. By definition of Fortissimo space, $\{p, x\}$ is open in T. Hence it is a open neighborhood of x. Therefore x is not a limit point of $S \setminus \{p\}$. Since x is arbitrary, $S \setminus \{p\}$ has no limit points. Hence T is not weakly countably compact. {{qed}}

B. Computing Resources

We conduct our experiments on eight NVIDIA A800 (80G) GPUs. For the sparse model BM25, evaluation on our benchmark takes approximately one hour. For small models with fewer than one billion parameters, evaluation requires around six GPU hours. The 1.5B model takes about 36 GPU hours, while the 7B models require about 64 GPU hours. Each proprietary model is evaluated in under 25 hours.

Table 13. Stacks example.

Query	Relevant Document
spaces-morphisms-lemma-birational Let S be a scheme. Let X and Y be algebraic space over S with $\|X\|$ and $\|Y\|$ irreducible. Then X and Y are birational if and only if there are nonempty open subspaces $U \subset X$ and $V \subset Y$ which are isomorphic as algebraic spaces over S.	Assume X and Y are birational. Let $f : U \to Y$ and $g : V \to X$ define inverse dominant rational maps from X to Y and from Y to X. After shrinking U we may assume $f : U \to Y$ factors through V. As $g \circ f$ is the identity as a dominant rational map, we see that the composition $U \to V \to X$ is the identity on a dense open of U. Thus after replacing U by a smaller open we may assume that $U \to V \to X$ is the inclusion of U into X. By symmetry we find there exists an open subspace $V' \subset V$ such that $g\|_{V'} : V' \to X$ factors through $U \subset X$ and such that $V' \to U \to Y$ is the identity. The inverse image of $\|V'\|$ by $\|U\| \to \|V\|$ is an open of $\|U\|$ and hence equal to $\|U'\|$ for some open subspace $U' \subset U$, see Properties of Spaces, Lemma \ref{spaces-properties-lemma-open-subspaces}. Then $U' \subset U \to V$ factors as $U' \to V'$. Similarly $V' \to U$ factors as $V' \to U'$. The reader finds that $U' \to V'$ and $V' \to U'$ are mutually inverse morphisms of algebraic spaces over S and the proof is complete.

Table 14. LeanDojo example.

Query	Relevant Document
R : Type u M : Type v inst†² : CommRing R inst†¹ : AddCommGroup M inst†: Module R M B : BilinForm R M f g : Module.End R M hf : IsSkewAdjoint B f hg : IsSkewAdjoint B g ⊢ IsAdjointPair B B (f * g) (g * f)	theorem neg_mul_neg (a b : α) : -a * -b = a * b

Table 15. MAPL example.

Query	Relevant Document
proof (prove) using this: length ps = length vs left_nesting f \\¡noteq¿ left_nesting g is_const (fst (strip_comb f)) goal (1 subgoal): 1. match (list_comb f ps) (list_comb g vs) = None	list_induct2: fixes xs :: "'c list" and ys :: "'d list" and P :: "'c list ¡Rightarrow¿ 'd list ¡Rightarrow¿ bool" assumes "length xs = length ys" and "P [] []" and " ¡And¿x xs y ys. ¡lbrakk¿length xs = length ys; P xs ys ¡rbrakk¿ ¡Longrightarrow¿ P (x # xs) (y # ys)" shows "P xs ys"

Table 16. HolStep example.

Query	Relevant Document
ABSOLUTELY_INTEGRABLE_CONVOLUTION_LINF_L1 —- (!bop. (!f. (!g. (!x. (((bilinear bop) /\ (((measurable_on f) UNIV) /\ ((bounded ((IMAGE f) UNIV)) /\ ((absolutely_integrable_on g) UNIV)))) ==¿ ((absolutely_integrable_on (\y. ((bop (f ((vector_sub x) y))) (g y)))) UNIV))))))	BILINEAR_SWAP —- (!op. ((bilinear (\x. (\y. ((op y) x)))) = (bilinear op)))

Table 17. ARQMath-Task-2 example.

Query	Relevant Document
Formula: $\int \frac{1}{(x^2+1)^n} dx$ Context: $\int \frac{1}{(x^2+1)^n} dx$ Let be $n \in \mathbb{Z}_+$. Compute the following integral: $$\int \frac{1}{(x^2+1)^n} dx$$ I obtained that for $$n = 1$$ the value of the integral is $$\tan^{-1} x + C$$ and for $$n = 2$$ $$x\left(\frac{1}{2\,(x^2+1)} + \frac{\tan^{-1}}{2x}\right) + C$$ How to do the rest of the cases?	$I_n = \int \frac{1}{(x^2-1)^n} dx$