

ReasonEmbed: Enhanced Text Embeddings for Reasoning-Intensive Document Retrieval

Jianlyu Chen^{1,2,4} Junwei Lan^{1,2,4} Chaofan Li^{2,3} Defu Lian^{1,4*} Zheng Liu^{2,5*}

¹University of Science and Technology of China

²Beijing Academy of Artificial Intelligence

³Beijing University of Posts and Telecommunications

⁴State Key Laboratory of Cognitive Intelligence ⁵Hong Kong Polytechnic University
chenjianlv@mail.ustc.edu.cn liandefu@ustc.edu.cn zhengliu1026@gmail.com

Abstract

In this paper, we introduce **ReasonEmbed**, a novel text embedding model developed for reasoning-intensive document retrieval. Our work includes three key technical contributions. First, we propose **ReMixer**, a new data synthesis method that overcomes the triviality problem prevalent in previous synthetic datasets, enabling large-scale production of 82K high-quality training samples. Second, we design **Redapter**, a self-adaptive learning algorithm that dynamically adjusts training each sample's weight based on its reasoning intensity. This allows the model to effectively capture the complex semantic relationships between queries and documents. Third, we implement ReasonEmbed across multiple backbones of varying sizes, all of which achieve **superior performance** on reasoning-intensive retrieval tasks. Notably, our ReasonEmbed-Qwen3-8B model offers a record-high nDCG@10 score of 38.1 on the BRIGHT benchmark (SU et al., 2025), which significantly outperforms existing text embedding models. We will fully open-source our created resources in ReasonEmbed to push forward the research advancement in this field¹.

1 Introduction

With the rapid advancement of large language models (LLMs), autonomous AI agents have become increasingly popular across various real-world applications (Yao et al., 2023; Wang et al., 2024a; Shinn et al., 2023), such as personal assistants, software engineering, and scientific research. In many of these scenarios, AI agents require access to informative external references to ensure the generation of truthful answers. However, the complex semantic relationships that often exist between queries and documents in these emerging domains pose

significant challenges for existing information retrieval systems. Recent studies (SU et al., 2025) suggest that most traditional retrievers, like general-purpose text embeddings (Neelakantan et al., 2022; Wang et al., 2024c; Xiao et al., 2024; Lee et al., 2024b) and BM25, struggle with these reasoning-intensive tasks, where effective retrieval often requires intensive reasoning operations.

Despite the imperative demand, research on reasoning-intensive document retrieval encounters several fundamental challenges. A primary limitation lies in the scarcity of suitable training data. Most existing document retrieval datasets are curated from traditional applications (Bajaj et al., 2016; Kwiatkowski et al., 2019), like web search and question answering, which differ substantially from the target problem in both query forms and domain knowledge. To alleviate data scarcity, recent studies have explored the use of synthetic datasets for developing retrieval systems (Lee et al., 2024b; Li et al., 2024; Wang et al., 2024b). Building on this idea, preliminary efforts have focused on data synthesis strategies tailored to reasoning-intensive document retrieval. An early attempt was made by ReasonIR (Shao et al., 2025), where long-form queries and hard negatives are synthesized using scientific corpora. Subsequent research further advanced this direction by employing more sophisticated query generation methods or by mining potential queries from existing corpora (Das et al., 2025; Liu et al., 2025; Long et al., 2025). However, current progress remains limited, as empirical evidence shows that these curated datasets yield only marginal gains over existing text embeddings.

In this paper, we propose **ReasonEmbed**, a new text embedding model for reasoning-intensive document retrieval based on innovations of how synthetic data is generated and used. Our work includes the following technical contributions.

First, we design a novel data synthesis method, called **ReMixer**. Our study begins by identifying

*Corresponding authors

¹All resources will be available at <https://github.com/VectorSpaceLab/agentic-search/tree/main/ReasonEmbed>

triviality as the key bottleneck in existing synthetic datasets. Specifically, synthetic data often exhibits overly direct relationships between queries and documents, where the relevance can be easily captured by surface patterns, like similar terms or overlapping keywords. To support this, Section 5.4.1 presents evidence demonstrating that the triviality problem severely impairs the retrieval capability of fine-tuned embedding models. With this insight, we design a **three-stage workflow** comprising conditioned query generation, source-excluded candidate mining, and reasoning-enhanced relevance annotation. This effectively mitigates trivial cases while preserving the validity of the synthesized training data.

Second, we introduce a self-adaptive training method tailored for synthetic data, termed **Redapter**. Synthetic training samples exhibit varying levels of **reasoning intensity**, i.e., the degree of reasoning needed to capture the relationship between a query and its related document. Embedding models tend to reach performance saturation more quickly on samples with lower reasoning intensity. To address this, Redapter dynamically adjusts the weight of each training sample based on its estimated reasoning intensity, thereby enabling more effective utilization of the synthetic data.

Third, we implement ReasonEmbed based on multiple LLM backbones of varying model sizes, which achieve **state-of-the-art performance** on reasoning-intensive document retrieval tasks. Notably, our model built on Qwen3-4B reaches an nDCG@10 score of 37.1 on the BRIGHT benchmark (SU et al., 2025), which already surpasses all existing text embeddings. While the Qwen3-8B based variant improves the performance to 38.1, yielding a significant improvement of almost +10 points over the cutting-edge baselines for this task. Extensive empirical analyses further validate the individual contributions of our synthesized data and self-adaptive training algorithm.

To summarize, our research offers a preliminary yet insightful exploration of advancing IR techniques for newly emerged reasoning-intensive scenarios. Our results demonstrate that, despite unprecedented challenges, text embeddings continue to play a crucial role in such problems via effective optimization. The entire resources of this work, including the source code, curated dataset, and well-trained models, will be publicly released to facilitate future research in this field.

2 Related Work

In this section, the related works are reviewed from two aspects: text embeddings, and reasoning-intensive document retrieval.

Text Embeddings. Text embeddings have emerged as a major research direction and have been extensively studied in recent years. Early works (Izacard et al., 2021; Wang et al., 2022; Li et al., 2023; Xiao et al., 2024; Chen et al., 2024) focused on improving the capabilities of embedding models on general tasks through multi-stage training on large-scale datasets. More recent studies (Ma et al., 2024; Wang et al., 2024b; Li et al., 2024; Lee et al., 2025; Zhang et al., 2025) leverage powerful LLMs as backbone models for embeddings, achieving state-of-the-art performance on challenging benchmarks such as MTEB (Muenighoff et al., 2022) and MMTEB (Enevoldsen et al., 2025). To address the efficiency limitations of large LLM-based embeddings, following-up works further explore knowledge distillation from larger teacher models to enhance the performance of lightweight embeddings, thereby enabling more practical deployment in real-world applications (Zhang et al., 2024; Askari et al., 2025; Vera et al., 2025).

Reasoning-Intensive Document Retrieval. Reasoning-intensive document retrieval was first introduced in BRIGHT (SU et al., 2025). Unlike traditional retrieval tasks, as represented by datasets such as MSMARCO (Bajaj et al., 2016) and Natural Questions (Kwiatkowski et al., 2019), where keyword or semantic-based matching is often sufficient, reasoning-intensive document retrieval requires in-depth reasoning to identify relevant documents. This poses unprecedented challenges for existing general-purpose retrieval models. Recent research in this area can be broadly categorized into two directions. The first one aims to enhance first-stage retrieval performance on top of tailored embedding models and query rewriting methods (Shao et al., 2025; Das et al., 2025; Long et al., 2025), while the second one focuses on developing reasoning-enhanced re-ranking models tailored for reasoning-intensive tasks (Niu et al., 2024; Weller et al., 2025; Yang et al., 2025b; Zhuang et al., 2025; Liu et al., 2025; Lan et al., 2025). Despite the current progresses, there still remain fundamental challenges regarding the effectiveness of first-stage retrieval and the curation of training data.

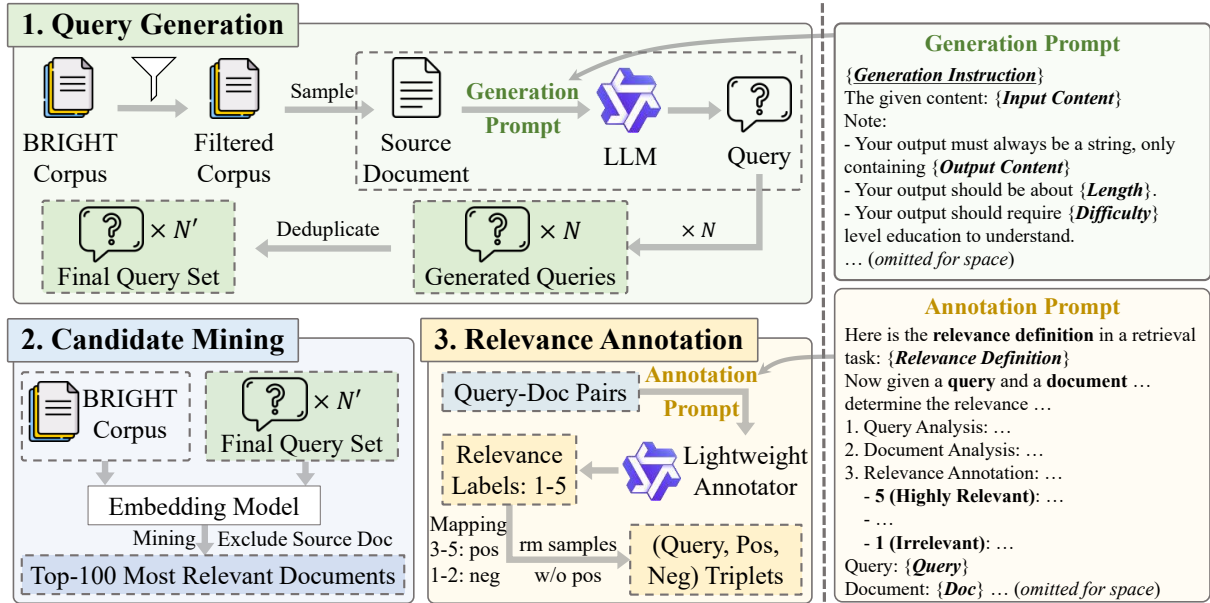


Figure 1: The three-stage data synthesis workflow of ReMixer. The full prompts used in the data synthesis process are available in Appendix A.

3 Data Synthesis: ReMixer

To ensure optimal data usability, the synthetic data is expected to satisfy the following properties. 1) The generated queries should require reasoning to address their information needs. 2) The generated queries should exhibit sufficient diversity in terms of forms and domains. 3) The relationships between queries and their corresponding positive and hard-negative documents should be accurately labeled. Guided by these principles, the following data synthesis workflow is designed (Figure 1).

3.1 Query Generation

We propose a conditioned query generation framework which prompts LLMs to produce quality and diversified queries. Our method begins by sourcing knowledge-rich corpora, which serve as a foundation for generating complex queries that call for intensive reasoning to address. In this work, we directly leverage the 12 datasets included in BRIGHT (SU et al., 2025), which span diverse and closely related domains such as science, mathematics, and programming. We pre-process the sourced corpora by removing documents that are irrelevant to the labeled domain of their respective datasets. Furthermore, to prevent data leakage, we exclude any documents labeled as positives for the evaluation queries in the BRIGHT benchmark.

With the preprocessed datasets, we design a prompt template (shown in Figure 1) to instruct the Qwen2.5-72B-Instruct (Team, 2024) model

for query generation. The prompt incorporates three key design elements: 1) an explicit generation instruction, which ensures the production of reasoning-necessitated queries; 2) query length sampling, which promotes the creation of long-form queries with diverse lengths; and 3) user education level sampling, which introduces variation in linguistic style and the depth of knowledge reflected in the generated queries.

3.2 Candidate Mining

Although widely adopted in existing approaches, directly using source documents as positive samples often leads to trivial connections with the generated queries. To address this problem, we propose to exempt the source documents and instead introduce documents that are different in form but correlated in essence as positive candidates for each generated query. Specifically, for each query (q), we employ off-the-shelf retrievers ($\phi(\cdot)$) to mine candidate documents, formally defined as: $\mathcal{C}_q \leftarrow \text{Top-k}\{\phi(q, d) \mid D/d_q^*\}$, where D/d_q^* denotes the corpus excluding the source document d_q^* associated with query q .

3.3 Relevance Annotation

We further annotate the mined candidates, identifying positive documents for each query while treating the remaining ones as hard negatives. Given that the generated queries are designed to express complex information needs associated with

knowledge-rich documents, the annotation process is conducted based on in-depth reasoning.

To achieve this, we perform *reasoning-enhanced relevance annotation* with optimized treatments. On one hand, inspired by the successful practice in JudgeRank (Niu et al., 2024), the annotation workflow is formulated as a three-stage process: 1) *Query analysis*, which examines the underlying information need expressed in the query; 2) *Document analysis*, which assesses the detailed knowledge contained in the document; 3) *Relevance annotation*, which determines the degree to which the document satisfies the query’s information need. The prompt template is shown in Figure 1.

On the other hand, we leverage state-of-the-art reasoning LLMs as the backbone of the annotator. Considering that the annotation process consumes a huge amount of tokens, directly using a large-scale reasoning LLM is prohibitively expensive. To address this, we employ a tailored lightweight LLM via distillation for cost-effective processing. Specifically, a student annotator based on Qwen3-8B is fine-tuned using reasoning trajectories from Qwen3-235B-A22B-Instruct-2507 (Yang et al., 2025a). Detailed settings about these implementations are provided in Appendix A.

3.4 Synthesization Result

As shown in Figure 1, after the relevance annotation stage, each candidate document for one query has a relevance label in $\{1, 2, 3, 4, 5\}$. Documents with labels in $\{3, 4, 5\}$ are used as positives and those with labels in $\{1, 2\}$ are used as negatives. The queries without any positive document are filtered out of the final dataset. The statistics of the final synthetic dataset are summarized in Table 1. In total, 95,960 raw queries were initially generated. After the annotation process, 81,659 valid queries remained, as those with no valid positive documents found in the sourced corpus were filtered out. In the final dataset, each query is associated with 12 positive documents on average, providing rich supervision signals for training retrieval models. The average query length reaches 221 tokens, which is substantially longer than those in traditional datasets and thus aligns well with the requirements of reasoning-intensive retrieval tasks.

4 Training Method: Redapter

With the synthetic dataset constructed, the embedding model is trained to discriminate positive docu-

Source	#Query (Final / Raw)	Avg. #Docs/Q		Avg. #Tokens		
		Pos	Neg	Q	Pos	Neg
StackExchange (7)						
Bio.	7,470 / 9,180	8	90	118	159	111
Earth	8,492 / 9,917	8	90	112	489	178
Econ.	7,147 / 7,837	14	84	117	428	321
Psy.	5,412 / 5,987	10	88	119	471	311
Rob.	7,159 / 8,015	8	90	151	364	167
Stack.	6,170 / 6,873	12	86	178	864	699
Sus.	4,493 / 5,031	13	85	106	368	187
Coding (2)						
Leet.	8,640 / 9,979	11	87	713	378	335
Pony	3,261 / 3,287	10	88	398	156	151
Math (3)						
AoPS	7,370 / 9,939	22	76	159	305	329
TheoQ.	7,184 / 9,919	21	78	155	246	302
TheoT.	8,861 / 9,996	5	93	259	445	513
Total	81,659 / 95,960	12	86	221	383	308

Table 1: Statistics of the synthetic dataset. **#Query (Final / Raw)**: number of training queries (final / raw). **Avg. #Docs/Q**: average number of positive or negative documents per query. **Avg. #Tokens**: average number of tokens per data instance.

ments from negative ones (including both in-batch negatives and the hard negatives provided in the synthetic dataset) for each query. The training objective follows the standard InfoNCE loss, which is formulated as:

$$\min . \mathcal{L}_{q,D} = -\log \frac{\exp(\langle q, d^+ \rangle / \tau)}{\sum_{d' \in D} \exp(\langle q, d' \rangle / \tau)}, \quad (1)$$

where D includes one positive d^+ and $|D| - 1$ negatives, $\langle \cdot, \cdot \rangle$ denotes the dot-product similarity between the embeddings of a query q and a document d , and τ is the temperature parameter. Since the embedding model must learn to capture the subtle semantic relationships between queries and documents, it is crucial to expose the model to sufficiently challenging training samples. Therefore, although reasoning-augmented queries, such as those produced by GPT-4 in BRIGHT (SU et al., 2025), can improve retrieval performance at test time, we retain the original forms of the generated queries during training rather than rewriting them.

4.1 Reasoning Intensity

Although all training samples are produced through the same data synthesis pipeline, they exhibit varying levels of difficulty due to differences in their source documents and inherent randomness during the generation process. Easier samples typically

involve relatively straightforward relationships between queries and documents, which can be captured without substantial reasoning and are learned quickly by the model. In contrast, harder samples embody more intricate semantic relationships that require deeper reasoning to resolve and therefore must be learned more patiently.

To capture this distinction, we define the concept of **reasoning intensity** to reflect such distinctions between training samples. Specifically, the hardness of a sample is characterized by the extent to which reasoning contributes to distinguishing relevant from irrelevant documents. For easy samples, the relevance can be determined through surface-level matching, making additional reasoning less beneficial. In contrast, hard samples depend heavily on multi-step reasoning, where additional reasoning operations significantly benefit the discrimination of the correct relationships. To quantify this property, we leverage the query-rewriting template from BRIGHT (SU et al., 2025) to generate a reasoning-augmented query q' for each raw query q . The reasoning intensity of a sample $s = (q, D)$ is defined as the contrast between the query-doc similarity computed with and without reasoning:

$$\text{RI}_\theta(s) = \min(\mathcal{L}_{q,D}/\mathcal{L}_{q',D}, \kappa), \quad (2)$$

where θ is real-time parameter of the embedding model, \mathcal{L} denotes the InfoNCE loss defined in Eq. 1, and κ is the hyperparameter used for truncating too large reasoning intensity scores. A larger $\text{RI}_\theta(s)$ value indicates a stronger impact of reasoning, as the reasoning-augmented query (q') substantially reduces the loss $\mathcal{L}_{q',D}$ compared to its non-reasoning counterpart $\mathcal{L}_{q,D}$.

4.2 Self-Adaptive Learning

With the introduction of reasoning intensity, we introduce a new self-adaptive learning method which switches to minimize the **RI-InfoNCE** loss defined as the following equation:

$$\mathcal{L}_{\text{RI}} = \sum_{s=(q,D), s \in B} f(\text{RI}_\theta(s), B) * \mathcal{L}_{q,D}, \quad (3)$$

where s is a training sample within a batch B , and $f(\cdot)$ is a normalization function that scales the reasoning-intensity scores within the batch:

$$f(\text{RI}_\theta(s), B) = \text{RI}_\theta(s) / \sum_{s' \in B} \text{RI}_\theta(s'). \quad (4)$$

The embedding model is initialized from a checkpoint pretrained on the MSMARCO

dataset (Bajaj et al., 2016). It is then optimized based on the synthetic dataset, with RI-InfoNCE adopted as the training objective. During the training process, samples with higher reasoning intensity receive greater weights, allowing the model to allocate more learning capacity to reasoning-intensive cases while retaining moderate exposure to simpler examples. This self-adaptive weighting strategy facilitates a more efficient and balanced learning process that enhances the model’s representation capability in complex retrieval scenarios. Notably, Redapter does not increase the training computation cost significantly. We provide the corresponding evidence in Appendix B.3.

5 Experiments

In this section, we conduct experiments for the following research questions:

RQ1: How effective is ReasonEmbed in reasoning-intensive document retrieval compared with existing text embedding models?

RQ2: What is the impact of the synthetic dataset on ReasonEmbed’s overall performance?

RQ3: How does the proposed training method contribute to ReasonEmbed’s overall performance?

RQ4: How consistent and accurate are the annotation labels generated by the distilled annotator?

RQ5: What insights can be drawn from the analysis of important implementation details?

5.1 Setup

The basic settings of the experimental studies are presented as follows.

Backbones. We adopt Qwen3-8B (Yang et al., 2025a) as the default backbone for all experiments. Besides, we use Qwen3-4B and Llama-3.1-8B (Grattafiori et al., 2024) to evaluate ReasonEmbed’s effectiveness across other popular models.

Training. The entire 82K synthesized training samples are used for fine-tuning the backbone models. The training is conducted for one single epoch, where the RI-InfoNCE loss function is applied after a warm-up stage of 100 training steps with the basic InfoNCE loss. Detailed training parameters are provided in Appendix B.2.

Evaluation. We use BRIGHT (SU et al., 2025), which includes 12 datasets spanning three domains, as the main evaluation benchmark. To further assess ReasonEmbed’s out-of-domain performance, we also incorporate R2MED (Li et al., 2025b), a BRIGHT-like reasoning-intensive retrieval bench-

Models	Size	Avg.	StackExchange							Coding		Theorem-based		
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
General-purpose methods														
BM25	-	14.5	18.9	27.2	14.9	12.5	13.6	18.4	15.0	24.4	7.9	6.2	10.4	4.9
OpenAI-3-Large	-	17.9	23.3	26.7	19.5	27.6	12.8	14.3	20.5	23.6	2.4	8.5	23.5	11.7
Google-Gecko-1B-768	1B	20.0	22.7	34.8	19.6	27.8	15.7	20.1	17.1	29.6	3.6	9.3	23.8	15.9
GritLM-7B	7B	21.0	24.8	32.3	18.9	19.8	17.1	13.6	17.8	29.9	22.0	8.8	25.2	21.2
gte-Qwen2-7B-instruct	7B	23.5	34.1	42.6	18.2	27.4	13.2	17.3	20.9	30.4	2.2	13.3	30.6	32.6
Qwen3-Embedding-4B*	4B	21.8	17.8	34.7	16.9	23.3	12.5	16.2	16.8	35.7	1.4	9.8	35.5	41.5
Qwen3-Embedding-8B*	8B	22.8	21.0	33.0	18.4	26.1	15.7	19.4	17.3	33.8	1.2	9.4	39.2	39.3
Qwen3-4B-ms [†]	4B	19.4	16.8	34.9	16.0	19.4	20.1	18.8	11.1	33.3	9.1	8.7	27.0	18.0
Qwen3-8B-ms [†]	8B	18.7	15.9	35.6	16.1	19.1	20.5	19.2	11.7	29.2	9.5	8.4	22.2	17.0
Llama-3.1-8B-ms [†]	8B	16.1	12.2	29.3	14.3	16.7	14.7	14.4	11.6	28.9	3.5	8.7	23.7	14.9
Tailored methods for reasoning-intensive retrieval														
ReasonIR-8B	8B	24.4	26.2	31.4	23.3	30.0	18.0	23.9	20.5	35.0	10.5	14.7	31.9	27.2
RaDeR-gte-Qwen2-7B	7B	25.5	34.6	38.9	22.1	33.0	14.8	22.5	23.7	37.3	5.0	10.2	28.4	35.1
Seed-1.5-Embedding	-	27.2	34.8	46.9	23.4	31.6	19.1	25.4	21.0	43.2	4.9	12.2	33.3	30.5
DIVER-Retriever	4B	28.9	41.8	43.7	21.7	35.3	21.0	21.2	25.1	37.6	13.2	10.7	38.4	37.3
ReasonEmbed from basic contrastive learning (using InfoNCE loss)														
ReasonEmbed-Qwen3-4B	4B	35.3	51.8	53.3	34.1	42.6	31.1	32.1	35.6	32.7	11.6	13.0	40.8	45.2
ReasonEmbed-Qwen3-8B	8B	37.1	54.4	55.4	33.8	45.2	32.0	34.3	37.3	32.3	18.7	13.3	41.2	47.6
ReasonEmbed-Llama-3.1-8B	8B	34.9	55.5	53.8	36.4	45.6	29.8	35.1	38.6	29.4	12.5	9.5	36.7	35.7
ReasonEmbed from Redapter (using the self-adaptive RI-InfoNCE loss)														
ReasonEmbed-Qwen3-4B	4B	37.1	55.4	54.5	34.9	46.9	34.0	36.1	37.4	34.5	13.6	11.3	41.4	45.1
ReasonEmbed-Qwen3-8B	8B	38.1	55.5	56.6	36.2	47.4	35.3	36.6	39.1	33.6	16.4	12.5	41.4	47.2
ReasonEmbed-Llama-3.1-8B	8B	36.2	55.4	56.2	35.2	48.5	32.1	37.3	41.1	28.8	16.8	9.1	37.9	36.6

Table 2: Main evaluation results (nDCG@10) on the BRIGHT benchmark (using original queries). For the baselines, “*” denotes well-trained models released by existing studies and evaluated by us; “†” indicates reproduced methods in this work. Other results are directly taken from the literature (SU et al., 2025; Das et al., 2025; Shao et al., 2025; Long et al., 2025).

mark specifically curated for healthcare scenarios. We also demonstrate the effectiveness of our synthetic data on non-reasoning-intensive retrieval tasks in Appendix C.4.

Baselines. We compare against two categories of baselines: 1) general-purpose baseline retrievers and 2) tailored methods for reasoning-intensive retrieval. The experiment results for BRIGHT and R2MED are presented in Table 2 and Table 3, respectively. Details on the baseline methods are provided in Appendix C.1.

5.2 Main Results (RQ1-RQ3)

The experimental results on the BRIGHT benchmark are presented in Table 2, where all variants of ReasonEmbed demonstrate superior performance. In particular, ReasonEmbed-Qwen3-4B achieves an average nDCG@10 score of 37.1, surpassing all existing baselines. ReasonEmbed-Qwen3-8B further improves performance to 38.1, establishing a substantial margin over previous methods. Complementary results on the R2MED benchmark, shown in Table 3, reveal that ReasonEmbed main-

tains a strong advantage over all baselines. Notably, ReasonEmbed-Qwen3-8B achieves an nDCG@10 score of 43.18, which significantly improves upon the MSMARCO-finetuned baseline Qwen3-8B-ms (24.15) and outperforms the competitive baseline, DIVER-Retriever (32.23), by more than 10 points. It’s worth noting that R2MED consists of healthcare-related queries and documents, which is substantially different from the data used to develop ReasonEmbed. Thus, it highlights the model’s strong out-of-domain generalizability. Overall, the above results provide compelling evidence for the effectiveness of ReasonEmbed.

In addition, the above results also provide crucial insights to the effectiveness of both data synthesis strategy and self-adaptive training method.

First, the synthetic dataset generated by ReMixer plays a critical role in establishing ReasonEmbed’s superior performance. Specifically, models fine-tuned on the synthetic data, including ReasonEmbed-Qwen3-4B, ReasonEmbed-Qwen3-8B, and ReasonEmbed-Llama-3.1-8B, achieve significant improvements over their MSMARCO-

Models	Size	Avg.	Q&A Reference			Clinical Evidence			Clinical Case	
			Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IYiC.
General-purpose methods										
BM25	-	15.13	19.19	21.55	19.68	0.66	2.55	23.69	21.66	12.02
OpenAI-3-Large	-	28.57	23.82	40.51	44.05	11.78	15.01	47.43	28.87	17.12
GritLM-7B	7B	31.12	24.99	43.98	45.94	12.32	19.86	39.88	37.08	24.94
NV-Embed-v2	7B	31.43	27.15	50.10	47.81	10.90	16.72	44.05	39.91	14.81
gte-Qwen2-7B-instruct*	7B	32.56	33.18	45.53	49.91	13.41	17.10	48.19	32.13	21.07
Qwen3-Embedding-4B*	4B	31.75	18.16	47.73	43.37	17.25	22.90	47.19	33.65	23.76
Qwen3-Embedding-8B*	8B	34.22	21.37	50.15	46.65	18.91	27.06	48.76	37.29	23.60
Qwen3-4B-ms [†]	4B	23.09	16.05	37.30	33.67	5.28	8.79	31.55	29.63	22.43
Qwen3-8B-ms [†]	8B	24.15	17.02	38.11	39.33	5.37	9.51	31.85	30.88	21.11
Llama-3.1-8B-ms [†]	8B	22.34	13.17	34.76	34.30	5.64	8.83	33.10	30.68	18.23
Tailored methods for reasoning-intensive retrieval										
ReasonIR-8B*	8B	27.94	26.16	44.84	39.28	11.21	14.92	36.56	29.20	21.37
RaDeR-gte-Qwen2-7B*	7B	35.19	36.02	53.58	50.32	15.41	20.31	49.94	34.31	21.83
DIVER-Retriever*	4B	32.23	39.27	51.68	50.81	13.94	16.10	38.56	24.69	22.81
ReasonEmbed from basic contrastive learning (using InfoNCE loss)										
ReasonEmbed-Qwen3-4B	4B	39.94	49.50	62.59	60.26	20.43	24.75	49.27	30.40	22.32
ReasonEmbed-Qwen3-8B	8B	42.36	51.65	65.17	65.89	20.67	27.98	51.71	33.79	22.01
ReasonEmbed-Llama-3.1-8B	8B	41.98	53.04	63.24	63.89	22.20	29.43	51.90	32.90	19.20
ReasonEmbed from Redapter (using the self-adaptive RI-InfoNCE loss)										
ReasonEmbed-Qwen3-4B	4B	41.16	52.45	64.28	62.58	20.83	26.21	48.03	32.50	22.38
ReasonEmbed-Qwen3-8B	8B	43.18	54.01	66.33	67.64	20.93	27.96	51.38	33.76	23.43
ReasonEmbed-Llama-3.1-8B	8B	42.76	53.52	64.37	63.82	20.40	29.67	51.86	34.55	23.88

Table 3: Main evaluation results (nDCG@10) on the R2MED benchmark (using original queries). For the baselines, “*” denotes well-trained models released by existing studies and evaluated by us; “†” indicates reproduced methods in this work. Other results are directly taken from Li et al. (2025b).

finetuned counterparts (Qwen3-4B-ms, Qwen3-8B-ms, and Llama-3.1-8B-ms). Moreover, these models also outperform general-purpose embedding baselines built on the same backbone encoders, such as Qwen3-Embedding-4B and Qwen3-Embedding-8B, both of which have been extensively trained on existing text retrieval datasets. This overwhelming advantage clearly demonstrates the effectiveness of the synthetic dataset curated by ReMixer.

Second, the self-adaptive training algorithm Redapter further enhances the overall performance of ReasonEmbed. Notably, the Redapter-based variants consistently outperform their counterparts trained with conventional contrastive learning. These results validate the effectiveness of our self-adaptive training approach and underscore the crucial role of reasoning-intensive samples, a key factor emphasized throughout both the data synthesis and training stages.

5.3 Annotation Quality Assessment (RQ4)

To assess the annotation quality of the distilled annotator, we evaluate its consistency with the teacher

Model	Cohen’s Kappa
Distilled Annotator	0.6001
Zero-shot Base Model	0.4314

Table 4: Annotation consistency evaluation of the distilled annotator and the zero-shot base model against the teacher model.

model and compute its accuracy based on human judgments from three authors of this paper.

Annotation Consistency. We construct a held-out evaluation set by sampling 500 queries per task and 10 documents per query, yielding 60,000 test samples for annotation consistency evaluation. We first use the teacher model and the distilled annotator to annotate these test samples independently, and then evaluate the annotation consistency using Cohen’s Kappa. For comparison, we also evaluate the agreement between the base model (Qwen3-8B) and the teacher model. As shown in Table 4, the distilled annotator obtains a Cohen’s Kappa score of 0.6001 on average, showcasing substantial agreement with the teacher model. Moreover, in comparison with the base model’s score of 0.4314, the distilled annotator achieves an absolute gain

Model	Accuracy
Teacher Model	86.1% (310/360)
Distilled Annotator	85.0% (306/360)

Table 5: Annotation accuracy evaluation of the distilled annotator and the teacher model.

of 0.1687, demonstrating the effectiveness of the distillation process.

Annotation Accuracy. We construct a held-out evaluation set by sampling 10 queries per task and 3 documents per query, yielding 360 test samples for annotation accuracy evaluation. We employ the distilled annotator and the teacher model to annotate these test samples independently. Then, human annotators validate the accuracy under a strict standard, where the label is considered accurate if and only if it aligns with human judgment. As shown in Table 5, both the distilled annotator and the teacher model achieve high average accuracies (85.0% and 86.1%, respectively), which demonstrates the high quality of our final annotated training data.

5.4 Ablation Study (RQ5)

We conduct ablation study to analyze the detailed impact of key designs across different components of our framework, including 1) candidate mining and 2) relevance annotation in data synthesis, 3) reasoning-intensity computation methods in the self-adaptive training, and 4) the influence of scaling-up of the synthesized data. To eliminate the effect of the self-adaptive training algorithm, the basic InfoNCE loss is used for all ablation studies except that on reasoning-intensity computation methods.

5.4.1 Candidate Mining

We first examine the impact of the candidate mining method. As discussed in Section 3.2, our default implementation (denoted as DEFAULT) mines candidate documents from the corpus while explicitly excluding the source documents to avoid trivial matches. In the ablation study, we introduce three alternative settings to analyze its effect: 1) NON-ANNO, which replaces positive samples with their corresponding source documents and directly uses the mined negatives without additional annotation, essentially following the common practice adopted in prior studies (Moreira et al., 2024). 2) SOURCE, which substitutes the positive samples with the source documents while keeping the same set of negatives as in DEFAULT. 3) SOURCE-PRO,

Method	Positive	Negative	Avg.
<u>DEFAULT</u>	mined & annotat.	mined & annotat.	37.1
<u>NON-ANNO</u>	source	mined	16.1
<u>SOURCE</u>	source	mined & annotat.	14.5
<u>SOURCE-PRO</u>	source & annotat.	mined & annotat.	22.1

Table 6: Impact of candidate mining methods (evaluated based on the BRIGHT benchmark).

Method	ROUGE-1	ROUGE-2	ROUGE-L	BRIGHT Score
<u>DEFAULT</u>	0.2411	0.0536	0.1439	37.1
<u>SOURCE</u>	0.3216	0.1351	0.2167	14.5

Table 7: ROUGE scores and BRIGHT performance comparison between the baseline source-as-positive method and our ReMixer method.

which further annotates the source documents and only kept those annotated as positive on the basis of SOURCE.

The experiment results are presented in Table 6, where the following observations are made. First, when comparing DEFAULT with NON-ANNO, SOURCE, and SOURCE-PRO, the alternative methods only achieve nDCG@10 scores of 16.1, 14.5, and 22.1, respectively, which are substantially lower than the performance of the default method. This clearly validates the necessity of excluding source documents when curating synthetic training samples. Second, comparing NON-ANNO, SOURCE, and SOURCE-PRO with one another, we observe that the annotation process lifts the overall performance to 22.1 (SOURCE-PRO), despite the triviality introduced by the source documents. This further highlights the critical role of relevance annotation in the data synthesis workflow.

Moreover, to illustrate the data triviality problem of the source-as-positive synthesis method, we compute and compare the ROUGE scores of two data sets: 1) the raw 95,960 generated query-positive pairs, where the positive is the source document used to generate the query, and 2) the final 81,659 synthesized query-positive pairs, where the positive is randomly sampled from all positives for that query. As presented in Table 7, DEFAULT yields consistently lower ROUGE scores compared to SOURCE. Furthermore, the BRIGHT performance results indicate that the triviality problem severely impairs the retrieval capability of fine-tuned embedding models.

5.4.2 Relevance Annotation

We next analyze the impact of data annotation by comparing different annotation methods. Our de-

Method	Training	Reasoning	Avg.
<u>DEFAULT</u>	distilled	w/ reasoning	37.1
<u>ZERO-SHOT</u>	zero-shot	w/ reasoning	32.4
<u>NON-REASON</u>	distilled	w/o reasoning	35.0

Table 8: Impact of on relevance annotation methods (evaluated based on the BRIGHT benchmark).

LLM Reasoner	Avg.
GPT-4.1-mini (default)	38.1
Qwen3-32B	37.8
Qwen3-8B	37.5
Qwen3-4B	36.5

Table 9: Impact of reasoning intensity scores computed by different LLM reasoners (evaluated based on the BRIGHT benchmark).

fault method, which employs a distilled lightweight reasoning-LLM, is denoted as DEFAULT. To assess its effectiveness, we introduce two alternative baselines. 1) ZERO-SHOT, which directly uses the original lightweight LLM without distillation from the teacher model; and 2) NON-REASON, which disables the reasoning process and trains the lightweight LLM to output annotation scores directly (refer to Appendix C.3 for details).

The experiment results are reported in Table 8, from which the following observations are made. First, with the adoption of the distilled annotator, DEFAULT improves overall performance from 32.4 (ZERO-SHOT) to 37.1. This suggests that lightweight LLMs lack sufficient reasoning capability for reasoning-enhanced relevance annotation, while distillation from a powerful teacher model effectively alleviates this limitation. Second, incorporating explicit reasoning further enhances performance, with DEFAULT improving the performance from 35.0 (NON-REASON) to 37.1. This underscores the importance of reasoning in accurately discriminating nuanced query–document relationships during data annotation.

5.4.3 Reasoning-Intensity

We further investigate the impact of reasoning-intensity scores computed using different LLM reasoners. In this study, we incorporate four popular alternatives: GPT-4.1-mini (default), Qwen3-32B, Qwen3-8B, and Qwen3-4B. It is worth noting that the reasoning outputs, i.e., the reasoning-augmented queries, are generated in the offline stage, thus will not affect the efficiency of online training. The experimental results are presented in Table 9. As observed, the default setting using GPT-

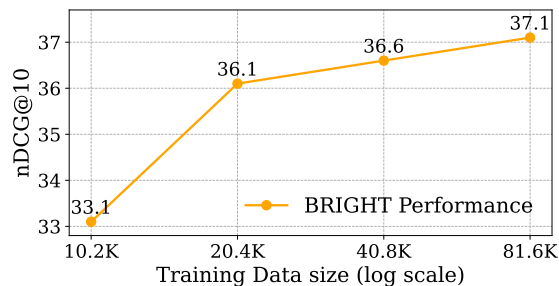


Figure 2: Impact of synthetic data size on retrieval accuracy (using basic contrastive learning for simplicity).

4.1-mini achieves the best performance, whereas models relying on lighter LLMs such as Qwen3-8B and Qwen3-4B perform noticeably worse. This highlights that generating high-quality reasoning augmentations is critical for accurately estimating reasoning intensity, and that using sufficiently powerful reasoning LLMs is essential for ensuring its effectiveness.

5.4.4 Data Scaling

We finally investigate the impact of synthetic data size. Since our data synthesis process is fully automated, the dataset can be continuously expanded at negligible cost. In our experiments, ReasonEmbed is fine-tuned on gradually enlarged datasets. As shown in Figure 2, the retrieval performance improves substantially as the dataset grows from 10.2K to 81.6K samples. These results further demonstrate the value of our synthesized data and highlight the importance of automated data synthesis for scaling reasoning-intensive retrieval models.

6 Conclusion

In this work, we present ReasonEmbed, a novel embedding model designed for the emerging task of reasoning-intensive document retrieval. We first introduce ReMixer, a three-stage data synthesis workflow that generates diverse and reasoning-intensive training samples. Building on this, we further propose Redapter, a self-adaptive training method that dynamically weights samples based on their reasoning intensity, enabling more effective learning from the synthetic data. Finally, our extensive experiments on the BRIGHT benchmark demonstrate that ReasonEmbed achieves significant improvements over existing methods, while all introduced technical component substantially contribute to its superior performance.

Limitations

While ReasonEmbed demonstrates strong performance in reasoning-intensive document retrieval, there remain several problems to improve. First, although the data synthesis method significantly enhances retrieval accuracy, it is still constrained by the reasoning capacity of the underlying LLMs. This limitation could be mitigated in the future with more powerful LLMs, or with alternative approaches, such as crafted multi-agent workflows. Second, the current work primarily focuses on reasoning-intensive retrieval. Expanding training to include both reasoning-intensive and general retrieval data would make ReasonEmbed more broadly applicable in practice. Third, our study is currently limited to domains covered by existing benchmarks, particularly BRIGHT. Scaling to more diverse domains will further strengthen the model’s generalizability. Addressing these challenges will be a focus of our future research.

Ethics Consideration

Since our synthetic dataset is generated by LLM, it may inherit potential biases, toxicity, and other issues present in the LLM used during the generation process. Moreover, the corpora used in the generation process are derived from the real-world sources, they may contain sensitive content.

Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (No. U24A20253).

References

- Arian Askari, Emmanouil Stergiadis, Ilya Gusev, and Moran Beladev. 2025. [HotelMatch-LLM: Joint multi-task training of small and large language models for efficient multimodal hotel retrieval](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 607–619, Vienna, Austria. Association for Computational Linguistics.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.
- Artemis Capari, Hosein Azarbondy, Georgios Tsatsaronis, Zubair Afzal, Judson Dunham, and Jaap Kamps. 2024. Knowledge acquisition passage retrieval: corpus, ranking models, and evaluation resources. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 74–87. Springer.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Debrup Das, Sam O’ Nuallain, and Razieh Rahimi. 2025. Rader: Reasoning-aware dense retrieval models. *arXiv preprint arXiv:2505.18405*.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Sibli, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Ryrström, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

- Junwei Lan, Jianlyu Chen, Zheng Liu, Chaofan Li, Siqi Bao, and Defu Lian. 2025. Retro*: Optimizing llms for reasoning-intensive document retrieval. *arXiv preprint arXiv:2509.24869*.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.
- Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, and 1 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*.
- Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, and 1 others. 2024b. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.
- Chaofan Li, Jianlyu Chen, Yingxia Shao, Defu Lian, and Zheng Liu. 2025a. Towards a generalist code embedding model based on massive data synthesis. *arXiv preprint arXiv:2505.12697*.
- Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024. Making text embedders few-shot learners. *arXiv preprint arXiv:2409.15700*.
- Lei Li, Xiao Zhou, and Zheng Liu. 2025b. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*.
- Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425.
- Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.
- 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. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.
- Tong Niu, Shafiq Joty, Ye Liu, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Judgerank: Leveraging large language models for reasoning-intensive reranking. *arXiv preprint arXiv:2411.00142*.
- Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and 1 others. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.
- Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Serkan O Arık, Danqi Chen, and Tao Yu. 2025. **BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval**. In *The Thirteenth International Conference on Learning Representations*.
- Qwen Team. 2024. **Qwen2.5: A party of foundation models**.
- Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panayam, Sara Smoot, Iftekhhar Naim, Joe Zou, Feiyang Chen, and 1 others. 2025. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.
- Ellen Voorhees, Tasmee Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025b. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.
- Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.
- Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024a. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.
- Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

A Details on Data Synthesis Method

In this section, we provide more details on our data synthesis method, ReMixer.

A.1 Query Generation

Corpus Filtering Prompt. Table 13 presents the prompt template used for filtering out domain-inconsistent documents from the entire corpus. We use the same LLM as used in query generation, Qwen2.5-72B-Instruct, to perform corpus filtering.

Query Generation Prompt. The designed prompt template for generating queries is presented in Table 14. The explicit generation instructions and output contents used in the query generation prompt are available in Table 20. We refer to CodeR-Pile (Li et al., 2025a) and E5-Mistral (Wang et al., 2024b) in the designing process of our prompt.

Sampling Parameters. For query generation, we set the temperature to 1.0. The inference is accelerated with SGLang (Zheng et al., 2024a).

A.2 Candidate Mining

Retriever Implementation. For a given task, we utilize gte-Qwen2-7B-instruct (Li et al., 2023) to mine the top-100 most relevant documents from the corpus of this task as the candidate documents, where the Faiss (Douze et al., 2024) index is built and the ANN search method (Johnson et al., 2019) is employed. The max query length and document length are both set to 8,192 tokens when encoding with gte-Qwen2-7B-instruct.

Effectiveness Sensitivity to the Initial Retriever. To explore the effectiveness sensitivity of our method to the initial retriever, we compare three different embedding models: 1) gte-Qwen2-7B-instruct (default setting), 2) Qwen3-Embedding-4B (Zhang et al., 2025), and 3) BGE-M3 (Chen et al., 2024). For fair comparison, we maintain other settings the same as those used by default, including the relevance annotator, training settings, and evaluation settings. Table 10 shows their performance on BRIGHT to provide references for their retrieval capabilities in reasoning-intensive retrieval tasks. As observed, the initial retriever in our data synthesis workflow has a slight impact on the final performance of the fine-tuned embedding model. Notably, though BGE-M3 performs poorly on BRIGHT (nDCG@10 12.0), using it to mine candidate documents only leads to an nDCG@10

Model	Avg.
gte-Qwen2-7B-instruct	23.5
Final ReasonEmbed-Qwen3-8B	38.1
Qwen3-Embedding-4B	21.8
Final ReasonEmbed-Qwen3-8B	37.2
BGE-M3	12.0
Final ReasonEmbed-Qwen3-8B	35.9

Table 10: Impact of initial retriever (evaluated based on the BRIGHT benchmark).

degradation of 2.2, which demonstrates the robustness of our data synthesis method.

A.3 Relevance Annotation

Relevance Annotation Prompt. Table 16 presents the prompt template used for reasoning-enhanced relevance annotation. The crafted relevance definitions are available in Table 21.

Distillation of Lightweight Annotator. We first random sample 500 queries per task and sample 10 documents per query. We employ Qwen3-235B-A22B-Instruct-2507 to generate reasoning trajectories using the previously mentioned relevance annotation prompt for these sampled query-document pairs, leading to 60,000 training samples for distillation. We use the LLaMA-Factory framework (Zheng et al., 2024b) to fine-tune Qwen3-8B on 8 NVIDIA H100 (80GB) GPUs in one epoch. The learning rate is set to 1e-5 and the warmup ratio is set to 0.1. The global training batch size is set to 8 and the gradient accumulation steps is set to 8.

Sampling Parameters. For relevance annotation, we set the temperature to 0.7. For Qwen3 series models, we disable thinking mode when using the tokenizer. The inference is accelerated with SGLang (Zheng et al., 2024a).

A.4 Synthesization Result

Task Instructions. The task instructions used in our synthetic dataset originate from BRIGHT (SU et al., 2025), which are presented in Table 17.

Licenses. The corpora in BRIGHT used for data synthesis are licensed under CC BY-4.0². our synthetic dataset is licensed under CC BY-NC-SA-4.0³. Therefore, our usage does not violate the intended use of BRIGHT dataset.

²<https://choosealicense.com/licenses/cc-by-4.0/>

³<https://creativecommons.org/licenses/by-nc-sa/4.0>

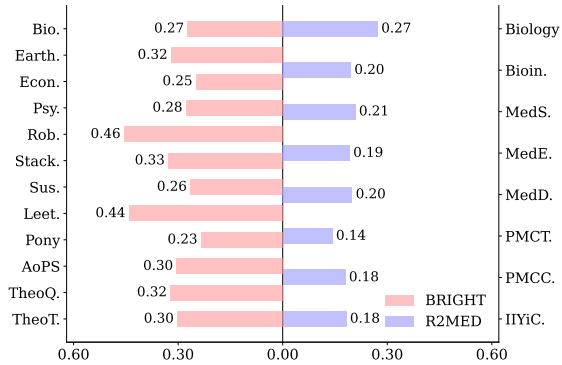


Figure 3: Data contamination analysis results (the computed max weighted Jaccard similarity) between our synthetic dataset and the testing data in BRIGHT and R2MED.

Data Contamination Analysis. To address the concerns of data leakage of the testing data in BRIGHT and R2MED, we perform a string-match-based query overlap analysis between the testing data and our synthetic dataset. Specifically, for each test query, we retrieve the top-20 most relevant training queries in our synthetic dataset⁴ using BM25 (Robertson et al., 2009), and then compute the weighted Jaccard similarity between the test query and each retrieved query. Figure 3 presents the computed max similarity for each dataset, showcasing that there is no data leakage in our synthetic dataset.

B Details on Training Method

B.1 Query Reasoning

Query Reasoning Prompt. Table 15 presents the prompt template used for generating reasoning queries.

Sampling Parameters. For generating reasoning-augmented queries, we set the temperature to 1.0. The max number of new tokens is set to 1024.

B.2 Experimental Settings

Hyperparameters. The temperature used in Eq. 1 is set to 0.02. The truncating threshold used in Eq. 2 is set to 5.0. By default, we employ GPT-4.1-mini to generate the reasoning-augmented queries.

Query Instruction Template. The query instruction template uses “Instruct: $\{task_instruction\}$ \nQuery: $\{query\}$ ”.

⁴We only consider the corresponding training data in the same domain for BRIGHT, and only consider the training data in the Biology domain for R2MED.

Method	#Steps	Train Runtime / s
Contrastive Learning	1304	13144
Redapter (Ours)	1304	13378

Table 11: Training computation cost analysis of our self-adaptive learning method Redapter.

Pooling Method. In our experiments, the embeddings are obtained by taking the last layer [EOS] vector.

Max Sequence Length. The max sequence length of query and document during training are both set to 512.

Negatives Strategy. We use hard negatives, in-batch negatives and cross-device negatives during training, and the total number of negatives used for each query is 1023.

Training. For efficient fine-tuning, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) with LoRA rank set to 64 and LoRA alpha set to 32. The learning rate is set to 1e-4 and the warmup ratio is set to 0.1. The training is conducted on 8 NVIDIA H100 (80GB) GPUs with the FlagEmbedding⁵ framework. The initialization processes on MSMARCO (Bajaj et al., 2016) of Qwen3-4B-ms, Qwen3-8B-ms, Llama-3.1-8B-ms use the same training settings as listed above.

B.3 Training Computation Cost Analysis

Though our self-adaptive learning method Redapter involves computing reasoning intensity of samples during training, the training computation cost does not show significant change. As shown in Table 11, when fine-tuning ReasonEmbed-Qwen3-8B, the total train runtime using Redapter only increases 234 seconds compared to using the original contrastive learning method, demonstrating the efficiency of our self-adaptive learning method.

C Details on Evaluation

C.1 Baselines

Table 22 lists all of the baseline models shown in our paper.

C.2 Evaluation Settings

When performing evaluation on BRIGHT, we use the task instructions presented in Table 17, and set the max length of both query and document to 8,192 tokens. When performing evaluation on R2MED, we use the task instructions presented in

⁵<https://github.com/FlagOpen/FlagEmbedding>

Table 18, and set the max length of both query and document to 512 tokens. We employ the evaluation framework from FlagEmbedding⁶.

C.3 Detailed Evaluation Results of Ablation Study

Table 23 presents the detailed evaluation results of ablation study on candidate mining methods in Section 5.4.1. Table 24 presents the detailed evaluation results of ablation study on relevance annotation methods in Section 5.4.2, where the annotation prompt template without reasoning used in NON-REASON is available in Table 19. Table 25 presents the detailed evaluation results of ablation study on reasoning-intensity computation methods in Section 5.4.3. Table 26 presents the detailed evaluation results of ablation study on training data size scaling in Section 5.4.4.

C.4 Evaluation on Non-Reasoning-intensive Retrieval Tasks

To demonstrate the effectiveness of our synthetic data on non-reasoning-intensive retrieval tasks, we fine-tune Qwen3-8B using our synthetic data and NQ (Kwiatkowski et al., 2019), respectively, and then perform evaluation on two non-reasoning-intensive tasks, including TREC-COVID (Voorhees et al., 2021) and KAPR (Capari et al., 2024). Table 12 presents the performances on TREC-COVID and KAPR using different training data to fine-tune the Qwen3-8B base model. We can make the following observations: 1) First, though TREC-COVID and KAPR are not reasoning-intensive retrieval tasks, our synthetic data still helps to improve the model’s retrieval performance on them. Specifically, performances on TREC-COVID and KAPR consistently improve as we scale the size of our synthetic dataset, demonstrating the generalizability of its effectiveness. 2) Second, our data achieves an even better impact on traditional tasks than NQ, a popular dataset for traditional question-answering retrieval, despite the fact that it focuses on reasoning-intensive scenarios. We believe the results here provide strong evidence about the effectiveness of our data synthesis method.

Training Data	#Samples	TREC-COVID	KAPR
NQ	58,568	64.40	75.26
Ours (10K)	10,202	57.68	54.58
Ours (20K)	20,411	75.20	63.93
Ours (full 80K)	81,659	79.86	76.99

Table 12: Performances on TREC-COVID and KAPR using different training data to fine-tune Qwen3-8B base model.

Given a passage, determine whether it belongs to the domain: *{Domain}*

The given passage:
[Begin of Passage]
{Doc}
[End of Passage]

Note:
- Your output must always be “Yes” or “No”.

Remember do not explain your output or output anything else. Your output:

Table 13: Prompt template for filtering our domain-inconsistent documents from the raw corpus. The placeholder *{Domain}* is set by the following mapping function: *{Biology, Earth Science, Economics, Psychology, Robotics, Sustainable Living}* → *the original dataset name*. *{Stack Overflow, LeetCode, Pony}* → *Coding*. *{AoPS, TheoremQA Questions, TheoremQA Theorems}* → *Math*.

{Generation Instruction}

The given content:
[Begin of Content]
{Input Content}
[End of Content]

Note:
- Your output must always be a string, only containing *{Output Content}*.
- Your output should be independent of the given content, which means that it should not containing the pronouns such as "it", "this", "that", "the given", "the provided", etc.
- Your output (*{Output Content}*) should be about *{Length}*.
- Your output (*{Output Content}*) should require *{Difficulty}* level education to understand.

Remember do not explain your output or output anything else. Your output:

Table 14: Prompt template for generating task-consistent queries. For placeholders, “*{Generation Instruction}*” ∈ Table 20, “*{Output Content}*” ∈ Table 20, “*{Length}*” ∈ {less than 100 words, 100 to 200 words, 200 to 300 words, 300 to 400 words, 400 to 500 words, at least 500 words}, and “*{Difficulty}*” ∈ {high school, college, phd}.

⁶<https://github.com/FlagOpen/FlagEmbedding>

Here is the **relevance definition** in a retrieval task: *{Relevance Definition}*

Now given a **query** (*{Query Type}*) and a **document** (*{Doc Type}*) in this retrieval task, your mission is to perform the following steps to determine the relevance between the query and the document.

1. **Query Analysis:** Think to reason and describe what information would be most helpful in answering the query.
2. **Document Analysis:** Discuss how the information provided by the document fulfills or fails to fulfill the requirements implied by the query.
3. **Relevance Annotation:** Based on the relevance definition and the insights from the previous two steps, clearly justify your final relevance annotation result and annotate an integer score from a scale of 1 to 5. Please use the following guide:
 - **5 (Highly Relevant):** The document is directly and fully responsive to the query, providing comprehensive, accurate, and specific information that completely addresses all aspects of the query.
 - **4 (Relevant):** The document is largely relevant and provides most of the information needed, but may have minor omissions, slight inaccuracies, or not be perfectly aligned with the query's intent.
 - **3 (Moderately Relevant):** The document has some relevance and offers partial information, but it may be incomplete, vague, or include some irrelevant content. It provides a basic connection but lacks depth or precision.
 - **2 (Slightly Relevant):** The document has minimal relevance, with only a small portion of content tangentially related to the query. The majority of the document is off-topic or provides little value.
 - **1 (Irrelevant):** The document is completely unrelated to the query and provides no useful information. There is no discernible connection or value for answering the query.

After providing your detailed analysis and justification for all the steps above, conclude your entire response with the final relevance score. The score must be placed strictly between the <score> tags. There should be no other text or explanation inside the tags:

```
<score>
[From a scale of 1 to 5, annotate the degree of relevance
between the query and the document.]
</score>
```

Note: The whole response should be as concise as possible while covering all the necessary details, and not exceeding 512 words in total.

```
Query ({Query Type}):
[Begin of Query]
{Query}
[End of Query]
```

```
Document ({Doc Type}):
[Begin of Document]
{Doc}
[End of Document]
```

Table 16: Prompt template for annotating the relevance of query-document pair. For placeholders, “*{Query Type}*”, “*{Doc Type}*”, “*{Relevance Definition}*” ∈ Table 21.

Given a question, your mission is to follow the instructions below:

1. Identify the essential problem.
2. Think step by step to reason and describe what information could be relevant and helpful to address the questions in detail.
3. Draft an answer with as many thoughts as you have.

```
The given question:
[Begin of Question]
{Original Query}
[End of Question]
```

Table 15: Prompt template for generating reasoning queries for original queries.

Task Name	Task Instruction
StackExchange (7)	
Bio.	Given a Biology post, retrieve relevant passages that help answer the post.
Earth.	Given an Earth Science post, retrieve relevant passages that help answer the post.
Econ.	Given an Economics post, retrieve relevant passages that help answer the post.
Psy.	Given a Psychology post, retrieve relevant passages that help answer the post.
Rob.	Given a Robotics post, retrieve relevant passages that help answer the post.
Stack.	Given a Stack Overflow post, retrieve relevant passages that help answer the post.
Sus.	Given a Sustainable Living post, retrieve relevant passages that help answer the post.
Coding (2)	
Leet.	Given a Coding problem, retrieve relevant examples that help answer the problem.
Pony	Given a Pony question, retrieve relevant passages that help answer the question.
Math (3)	
AoPS	Given a Math problem, retrieve relevant examples that help answer the problem.
TheoQ.	Given a Math problem, retrieve relevant examples that help answer the problem.
TheoT.	Given a Math problem, retrieve relevant theorems that help answer the problem.

Table 17: Task names and task instructions for all 12 retrieval tasks in our synthetic dataset.

Task Name	Task Instruction
Q&A Reference (3)	
Biology	Given a Biology post, retrieve relevant passages that help answer the post.
Bioin.	Given a Bioinformatics post, retrieve relevant passages that help answer the post.
MedS.	Given a Medical Science post, retrieve relevant passages that help answer the post.
Clinical Evidence (3)	
MedE.	Given a Medical Exam, retrieve relevant passages that help answer the exam.
MedD.	Given a Medical Exam, retrieve relevant passages that help answer the exam.
PMCT.	Given a Clinical Case, retrieve relevant passages that help answer the case.
Clinical Case (2)	
PMCC.	Given a Clinical Case, retrieve similar cases that help diagnose the case.
IlyiC.	Given a Clinical Case, retrieve similar cases that help diagnose the case.

Table 18: Task names and task instructions for all 8 retrieval tasks in R2MED.

Here is the **relevance definition** in a retrieval task: *{Relevance Definition}*

Now given a **query** (*{Query Type}*) and a **document** (*{Doc Type}*) in this retrieval task, your mission is to perform the following steps to determine the relevance between the query and the document.

1. Query Analysis: Think to reason and describe what information would be most helpful in answering the query.
2. Document Analysis: Discuss how the information provided by the document fulfills or fails to fulfill the requirements implied by the query.
3. Relevance Annotation: Based on the relevance definition and the insights from the previous two steps, clearly justify your final relevance annotation result and annotate an integer score from a scale of 1 to 5. Please use the following guide:
 - **5 (Highly Relevant)**: The document is directly and fully responsive to the query, providing comprehensive, accurate, and specific information that completely addresses all aspects of the query.
 - **4 (Relevant)**: The document is largely relevant and provides most of the information needed, but may have minor omissions, slight inaccuracies, or not be perfectly aligned with the query’s intent.
 - **3 (Moderately Relevant)**: The document has some relevance and offers partial information, but it may be incomplete, vague, or include some irrelevant content. It provides a basic connection but lacks depth or precision.
 - **2 (Slightly Relevant)**: The document has minimal relevance, with only a small portion of content tangentially related to the query. The majority of the document is off-topic or provides little value.
 - **1 (Irrelevant)**: The document is completely unrelated to the query and provides no useful information. There is no discernible connection or value for answering the query.

Directly output the final relevance score without any explanation or reasoning steps. The score must be placed strictly between the <score> tags. There should be no other text or explanation inside the tags:

```
<score>
[From a scale of 1 to 5, annotate the degree of relevance between the query and the document.]
</score>
```

Note: The response should **ONLY** contain the score enclosed within the <score> tags, with no additional text or commentary. Example of correct format: <score>4</score>.

Query (*{Query Type}*):
[Begin of Query]
{Query}
[End of Query]

Document (*{Doc Type}*):
[Begin of Document]
{Doc}
[End of Document]

Table 19: Prompt template for annotating the relevance of query-document pair **without reasoning process** (used for ablation study in Section 5.4.2). For placeholders, “*{Query Type}*”, “*{Doc Type}*”, “*{Relevance Definition}*” ∈ Table 21.

Task Name	Generation Instruction	Output Content
StackExchange (7)		
Bio.	Given a Biology-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Earth.	Given a Biology-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Econ.	Given an Economics-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Psy.	Given a Psychology-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Rob.	Given a Robotics-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Stack.	Given a Coding-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Sus.	Given a Sustainable Living-related passage in <i>{language}</i> , generate a StackExchange post in <i>{language}</i> for which the critical concepts or theories discussed in the passage can serve as references for domain experts to draft an answer.	the generated StackExchange post in <i>{language}</i>
Coding (2)		
Leet.	Given a solved LeetCode problem (with solutions) in <i>{language}</i> , generate a new LeetCode problem in <i>{language}</i> that the underlying algorithms or data structures from the original problem can help solve.	the generated LeetCode problem in <i>{language}</i>
Pony	Given a Pony documentation passage in <i>{language}</i> , generate a Pony coding instruction in <i>{language}</i> that the Pony syntax described in the passage can help implement.	the generated Pony coding instruction in <i>{language}</i>
Math (3)		
AoPS	Given a Math problem solution in <i>{language}</i> , generate a new Math problem in <i>{language}</i> that the problem-solving skills used in the original problem can help solve.	the generated Math problem in <i>{language}</i>
TheoQ.	Given a Math problem solution in <i>{language}</i> , generate a new Math problem in <i>{language}</i> that the theorems used in the original problem can help solve.	the generated Math problem in <i>{language}</i>
TheoT.	Given a Math theorem in <i>{language}</i> , generate a Math problem in <i>{language}</i> that the theorem can help solve.	the generated Math problem in <i>{language}</i>

Table 20: Generation instructions and output contents for all 12 retrieval tasks in our synthetic dataset. The placeholder *{language}* is set to *English*.

Task Name	Generation Instruction
StackExchange (7)	
Bio.	Given a query (<i>biology post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Earth.	Given a query (<i>earth science post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Econ.	Given a query (<i>economics post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Psy.	Given a query (<i>psychology post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Rob.	Given a query (<i>robotics post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Stack.	Given a query (<i>Stack Overflow post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Sus.	Given a query (<i>sustainable living post</i>) and a document (<u>passage</u>), the document is relevant to the query if the critical concepts or theories discussed in the document can provide references for domain experts to draft an answer to the query.
Coding (2)	
Leet.	Given a query (<i>LeetCode problem</i>) and a document (<u>coding problem solution</u>), the document is relevant to the query if the underlying algorithms or data structures used in the document can provide helpful insights for solving the problem in the query.
Pony	Given a query (<i>Pony coding instruction</i>) and a document (<u>Pony documentation passage</u>), the document is relevant to the query if the Pony syntax described in the document is necessary for beginners with no prior knowledge of Pony to complete the coding instruction in the query.
Math (3)	
AoPS	Given a query (<i>math problem</i>) and a document (<u>math problem solution</u>), the document is relevant to the query if the theorems used in the document can provide helpful insights for solving the problem in the query.
TheoQ.	Given a query (<i>math problem</i>) and a document (<u>math problem solution</u>), the document is relevant to the query if the theorems used in the document can provide helpful insights for solving the problem in the query.
TheoT.	Given a query (<i>math problem</i>) and a document (<u>math-related passage</u>), the document is relevant to the query if the theorem described in the document can help solve the problem in the query.

Table 21: Relevance definitions used for annotation for all 12 retrieval tasks in our synthetic dataset. The *query type* is italic, and the document type is underlined.

Model	Size	Model Link
<i>Lexical method</i>		
BM25 (Robertson et al., 2009)	-	https://github.com/xlang-ai/BRIGHT
<i>General-purpose embedding models</i>		
OpenAI-3-Large	-	https://openai.com/index/new-embedding-models-and-api-updates
Google-Gecko-1B-768 (Lee et al., 2024b)	1B	https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api
GritLM-7B (Muennighoff et al., 2024)	7B	https://huggingface.co/GritLM/GritLM-7B
NV-Embed-v2 (Lee et al., 2024a)	7B	https://huggingface.co/nvidia/NV-Embed-v2
gte-Qwen2-7B-instruct (Li et al., 2023)	7B	https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct
Qwen3-Embedding-4B (Zhang et al., 2025)	4B	https://huggingface.co/Qwen/Qwen3-Embedding-4B
Qwen3-Embedding-8B (Zhang et al., 2025)	8B	https://huggingface.co/Qwen/Qwen3-Embedding-8B
<i>Reasoning-optimized embedding models</i>		
ReasonIR-8B (Shao et al., 2025)	8B	https://huggingface.co/reasonir/ReasonIR-8B
RaDeR-gte-Qwen2-7B (Das et al., 2025)	7B	https://huggingface.co/Raderspace/RaDeR-gte-Qwen2-7B_MATH_LLMq_CoT_lexical
Seed-1.5-Embedding	-	https://seed1-5-embedding.github.io/
DIVER-Retriever (Long et al., 2025)	4B	https://huggingface.co/AQ-MedAI/Diver-Retriever-4B

Table 22: Detailed information on all of the baseline models shown in our paper.

Method	Positive	Negative	Avg.	StackExchange						Coding		Theorem-based			
				Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
ReasonEmbed-Qwen3-8B from basic contrastive learning (using InfoNCE loss)															
<u>DEFAULT</u>	mined & annotat.	mined & annotat.	37.1	54.4	55.4	33.8	45.2	32.0	34.3	37.3	32.3	18.7	13.3	41.2	47.6
<u>NON-ANNO</u>	source	mined	16.1	13.1	26.2	11.1	14.6	15.1	16.3	14.5	28.8	5.0	6.2	22.8	19.0
<u>SOURCE</u>	source	mined & annotat.	14.5	11.4	22.0	12.9	12.6	15.6	16.3	15.5	23.5	5.0	6.4	20.4	12.4
<u>SOURCE-PRO</u>	source & annotat.	mined & annotat.	22.1	28.0	38.0	19.3	20.8	18.4	23.0	19.3	29.6	4.3	6.5	27.1	30.4

Table 23: Detailed evaluation results (nDCG@10) on BRIGHT benchmark (using original queries) for ablation study of **candidate mining** methods.

Method	Training	Reasoning	Avg.	StackExchange						Coding		Theorem-based			
				Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
ReasonEmbed-Qwen3-8B from basic contrastive learning (using InfoNCE loss)															
<u>DEFAULT</u>	distilled	w/ reasoning	37.1	54.4	55.4	33.8	45.2	32.0	34.3	37.3	32.3	18.7	13.3	41.2	47.6
<u>ZERO-SHOT</u>	zero-shot	w/ reasoning	32.4	46.1	51.2	31.7	40.1	28.2	31.4	32.6	33.3	4.3	9.4	38.1	42.7
<u>NON-REASON</u>	distilled	w/o reasoning	35.0	46.4	49.6	32.2	46.9	32.4	34.2	34.4	37.2	12.6	8.3	38.6	47.1

Table 24: Detailed evaluation results (nDCG@10) on BRIGHT benchmark (using original queries) for ablation study of **relevance annotation** methods.

LLM Reasoner	Avg.	StackExchange						Coding		Theorem-based			
		Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
ReasonEmbed-Qwen3-8B from Redapter (using the self-adaptive RI-InfoNCE loss)													
GPT-4.1-mini (default)	38.1	55.5	56.6	36.2	47.4	35.3	36.6	39.1	33.6	16.4	12.5	41.4	47.2
Qwen3-32B	37.8	55.4	56.4	35.8	48.7	33.7	37.3	37.9	33.1	16.6	12.1	42.1	45.0
Qwen3-8B	37.5	55.3	55.1	34.2	47.5	35.0	34.5	38.4	32.6	17.7	11.8	41.4	46.1
Qwen3-4B	36.5	52.7	52.1	33.9	44.6	30.1	33.0	37.5	32.7	19.7	13.6	41.1	46.9

Table 25: Detailed evaluation results (nDCG@10) on BRIGHT benchmark (using original queries) for ablation study of **reasoning-intensity** computation methods.

Percentage	Data Size	Avg.	StackExchange						Coding		Theorem-based			
			Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.
ReasonEmbed-Qwen3-8B from basic contrastive learning (using InfoNCE loss)														
100.0%	81.6K	37.1	54.4	55.4	33.8	45.2	32.0	34.3	37.3	32.3	18.7	13.3	41.2	47.6
50.0%	40.8K	36.6	56.2	53.0	34.4	45.6	31.9	36.4	37.6	33.8	11.4	11.9	41.7	45.7
25.0%	20.4K	36.1	54.3	54.4	33.1	44.8	30.7	35.5	35.7	34.7	8.7	12.4	41.3	48.2
12.5%	10.2K	33.1	49.0	48.8	29.8	41.2	28.6	34.5	33.0	35.6	5.5	11.0	39.3	40.5

Table 26: Detailed evaluation results (nDCG@10) on BRIGHT benchmark (using original queries) for ablation study of **training data size scaling**.