# AIR-BENCH: Automated Heterogeneous Information Retrieval Benchmark

**Anonymous ACL submission**

## Abstract

Evaluation plays a crucial role in the advancement of information retrieval (IR) models. However, current benchmarks, which are based on predefined domains and human-labeled data, face limitations in addressing evaluation needs for emerging domains both cost-effectively and efficiently. To address this challenge, we propose the **A**utomated Heterogeneous **I**nformation **R**etrieval **Bench**mark (**AIR-BENCH**). AIR-BENCH is distinguished by three key features: 1) Automated. The testing data in AIR-BENCH is automatically generated by large language models (LLMs) without human intervention. 2) Heterogeneous. The testing data in AIR-BENCH is generated with respect to diverse tasks, domains and languages. 3) Dynamic. The domains and languages covered by AIR-BENCH are constantly augmented to provide an increasingly comprehensive evaluation benchmark for community developers. We develop a reliable and robust data generation pipeline to automatically create diverse and high-quality evaluation datasets based on real-world corpora. Our findings demonstrate that the generated testing data in AIR-BENCH aligns well with human-labeled testing data, making AIR-BENCH a dependable benchmark for evaluating IR models. The resources in AIR-BENCH will be made publicly available.

## 1 Introduction

As information retrieval (IR) models grow in complexity and capability, the need for sophisticated evaluation techniques becomes increasingly critical. In recent years, a series of milestone works have significantly advanced the field by introducing comprehensive evaluation datasets and benchmarks. Early contributions to IR evaluation include MS MARCO (Bajaj et al., 2016) and Natural Questions (Kwiatkowski et al., 2019), both designed for open-domain question answering (QA) tasks in English. Recognizing the importance of multilingual information retrieval, researchers developed Mr.TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023). More recently, the focus has shifted towards creating general-domain, zero-shot IR benchmarks. BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) represent this trend by aggregating multiple existing datasets from diverse tasks and domains.

Despite their contributions, existing benchmarks are constrained to pre-defined domains and rely heavily on human-labeled data, making it challenging to efficiently address evaluation needs in emerging domains. With the emergence of powerful large language models (LLMs), several studies have explored their application for retrieval evaluation in retrieval-augmented generation (RAG) systems (Es et al., 2024; Saad-Falcon et al., 2024; Salemi and Zamani, 2024), presenting a promising solution to this challenge. However, a comprehensive IR benchmark that addresses this limitation remains insufficiently developed.

In this work, we present the **A**utomated Heterogeneous **I**nformation **R**etrieval **Bench**mark (**AIR-BENCH**), which is characterized by three features:

1. **Automated:** We develop a comprehensive data generation pipeline to automatically produce diverse and high-quality testing data with large language models (LLMs). Therefore, it is able to instantly support the evaluation of new domains both cost-effectively and efficiently. Besides, the new testing data is almost impossible to be covered by the training sets of any existing retrievers.

2. **Heterogeneous:** AIR-BENCH is designed to be a heterogeneous IR benchmark including diverse tasks, domains and languages. It currently covers 2 tasks, 9 domains, and 13 languages, including a total of 69 datasets. This extensive coverage enables thorough evaluation across diverse scenarios, potentially ac-

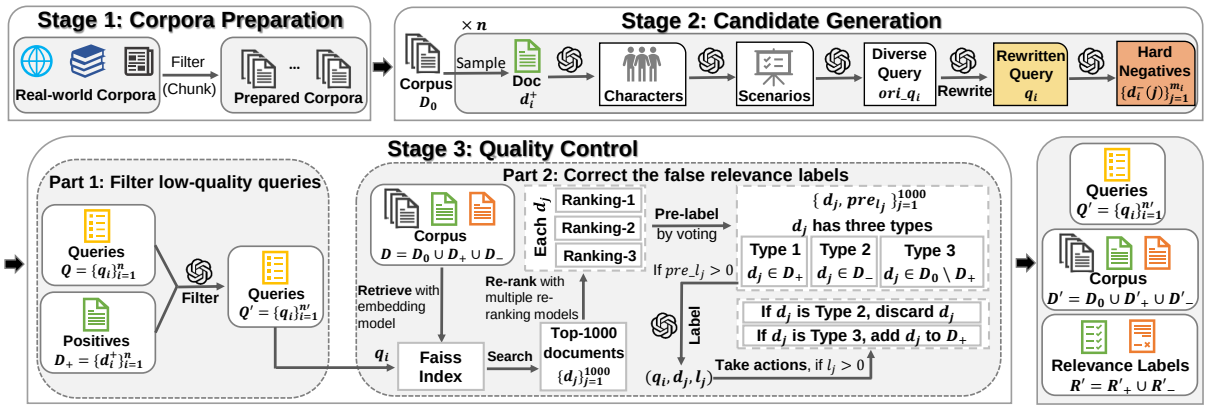



Figure 1: The three-stage data generation pipeline of AIR-BENCH.

celerating advancements in IR technology for both established and emerging domains.

3. **Dynamic:** The tasks, domains and languages covered by AIR-BENCH are planed to be augmented on regular basis. There are currently two distinct versions, 24.04 and 24.05, with more anticipated in the future. We hope AIR-BENCH is able to provide an increasingly comprehensive evaluation benchmark for community developers.

These features form the foundation of our proposed benchmark and directly address the limitation in existing benchmarks for information retrieval systems. To further elucidate the impact and scope of our work, we summarize our main contributions as follows: 1) We introduce AIR-BENCH, a new information retrieval benchmark highlighted by new features: automated, heterogeneous and dynamic. 2) We demonstrate that our data generation pipeline is able to produce diverse and high-quality testing data highly consistent with human-labeled testing data, making AIR-BENCH a dependable benchmark for evaluating IR models. 3) Additionally, we develop and release software tools enabling community developers to evaluate any IR model using AIR-BENCH. To foster collaboration and progress in the field, we establish and maintain a public leaderboard[1] to track and compare model performance across the community. These contributions collectively advance the field of information retrieval by providing a versatile, dynamic, and comprehensive evaluation framework.

[1] *Anonymous-Link-1*

## 2 Benchmark Construction

The entire data generation pipeline of AIR-BENCH consists of three stages: 1) Corpora preparation, 2) Candidate generation, and 3) Quality control.

### 2.1 Preliminary

AIR-BENCH focuses on the evaluation of information retrieval. The information retrieval task can be formulated as: Given a query $q$, retrieve a ranked list of $n$ most relevant documents $\mathcal{L} = [d_1, d_2, \cdots, d_n]$ from the corpus $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$.

To clarify the subsequent explanation, Table 1 lists the symbols that appear in this section along with their corresponding meanings for reference.

| Symbol | Meaning | Symbol | Meaning |
|---|---|---|---|
| $q$ | query | $\mathcal{Q}$ | queries set |
| $d$ | document | $d^+/d^-$ | positive/negative document |
| $l$ | relevance label | $n, m$ | number |
| $\mathcal{D}$ | documents set | $\mathcal{D}_+/\mathcal{D}_-$ | positive/negative documents set |
| $\mathcal{R}$ | relevance labels set | $\mathcal{R}_+/\mathcal{R}_-$ | positive/negative relevance labels set |
| $\mathcal{L}$ | documents list | $\mathcal{M}$ | re-ranking model |

Table 1: Corresponding meanings for the symbols appearing in this section.

### 2.2 Corpora Preparation

As shown in Figure 1, the first stage involves preparing diverse corpora. Specifically, given a task, we collect real-world datasets from diverse domains and languages, and apply distinct pre-processing strategies to the raw datasets based on the task requirements (See Appendix A.1 for more details).

The corpus prepared in this stage is denoted as $\mathcal{D}_0 = \{d_i\}_{i=1}^{n_0}$, including $n_0$ documents.

## 2.3 Candidate Generation

The candidate data for a retrieval dataset consists of three components: corpus, queries and qrels. After preparing the corpus in the initial stage, the candidate generation stage produces the remaining two components of the dataset: queries and qrels.

Based on the corpus, the candidate generation process is iteratively executed in a loop. As shown in Figure 1, the generation process can be summarized as the following steps: 1) Sample one document from the raw corpus as the positive document $d_i^+$. 2) Prompt LLM to generate the characters who might find the document useful. 3) Prompt LLM to generate the scenarios in which the character might find the document useful. 4) Prompt LLM to generate the query $ori_q_i$ based on the specific character and scenario. To diversify the generated queries, we consider the following attributes when designing the prompt: query length, query type, information-based type, and expression style. 5) Prompt LLM to rewrite the generated query for multiple times to try to avoid the duplicated tokens as in the raw corpus, and finally get query $q_i$. 6) Prompt LLM to generate some hard negative documents $\{d_i^-(j)\}_{j=1}^{m_i}$ based on the generated query $q_i$ and the positive document $d_i^+$. 7) Repeat Step 1-6. Considering both simplicity and the absence of examples in a new domain, the above prompting strategies are all zero-shot. For more details, please refer to Appendix A.2.

After repeating $n$ times of the above loop, we get the queries set $\mathcal{Q}$, the positive documents set $\mathcal{D}_+$, the hard negative documents set $\mathcal{D}_-$, the corpus $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_+ \cup \mathcal{D}_-$, the positive relevance labels set $\mathcal{R}_+$, and the negative relevance labels set $\mathcal{R}_-$.

## 2.4 Quality Control

In this stage, we design comprehensive quality control strategies to enhance the quality of the generated dataset. As shown in Figure 1, the quality control process can be summarized as two parts.

**Filter low-quality queries**. Since all of the queries in the candidate data are generated by LLM, there are potential low-quality queries. To improve the quality of generated queries, we utilize LLM to access the relevance between the query $q_i$ and the positive document $d_i^+$. If the LLM prediction is negative, indicating that $q_i$ is a low-quality query, we discard $q_i$ from $\mathcal{Q}$ and remove the relevance labels $\{(q_i, *, *)\}$ from $\mathcal{R}_+$ and $\mathcal{R}_-$. For details on how we utilize LLM to label the relevance, please refer to Appendix A.3.

| | $l_j$ is pos | $l_j$ is neg |
|---|---|---|
| Type 1: $d_j \in \mathcal{D}_+$ | - | * |
| Type 2: $d_j \in \mathcal{D}_-$ | discard $d_j$ from $\mathcal{D}_-$, remove $(q_i, d_j, 0)$ from $\mathcal{R}_-$ | - |
| Type 3: $d_j \in \mathcal{D}_0 \setminus \mathcal{D}_+$ | add $d_j$ to $\mathcal{D}_+$, add $(q_i, d_j, 1)$ to $\mathcal{R}_+$ | - |

Table 2: Specifications of different quality control strategies based on the type of document $d_j$ and the relevance label $l_j$ of $(q_i, d_j)$. Type 1 means that $d_j$ is the original positive document, Type 2 means that $d_j$ is the generated hard negative document, and Type 3 means that $(q_i, d_j)$ does not have a relevance label in the second stage. "-": Skip. "*": If the type of $d_j$ is Type 1, $l_j$ must be positive since we have filtered low-quality queries.

**Correct the false relevance labels**. The false relevant labels comprise two types of documents: the first type includes the generated hard negative documents, and the second type consists of relevant documents that were overlooked in the corpus. Given a query $q_i$, we design a three-step pipeline to correct the false relevance labels. 1) *Recall with embedding model*. Use the embedding model to search top-1000 relevant documents $\mathcal{L}_{recall} = [d_1, \cdots, d_{1000}]$ from the corpus for $q_i$. 2) *Pre-label with re-ranking models*. Use multiple re-ranking models to re-rank $\mathcal{L}_{recall}$. We pre-label each document $d_j$ according to their ranking $r_j(\mathcal{M})$ in the re-ranked top-1000 relevant documents $\mathcal{L}_{rerank}(\mathcal{M})$ from the re-ranking model $\mathcal{M}$. Specifically, if $r_j(\mathcal{M})$ is higher than the predetermined threshold, the label $l_j(\mathcal{M})$ for $d_j$ from $\mathcal{M}$ is positive. If more than half of re-ranking models label $d_j$ as positive, we pre-label $d_j$ as positive, otherwise we pre-label $d_j$ as negative. After this step, each document $d_j$ in $\mathcal{L}_{recall}$ has a preliminary label $pre_l_j$. 3) *Label with LLM*. In this step, we also utilize LLM to access the relevance between $q_i$ and the documents $\{d_j\}_{j=1}^{m_i}$ that are pre-labeled as positive in the last step. The prediction from LLM is denoted as $l_j$. As shown in Table 2, we categorize $d_j$ into three types, and take different actions by the type of $d_j$ and $l_j$. For details on how we select the embedding model and multiple re-ranking models, and set the predetermined threshold for pre-labeling, please refer to Appendix A.3.

After executing the above quality control process for each query, we get the new queries set $\mathcal{Q}'$, the new positive documents set $\mathcal{D}'_+$, the new hard negative documents set $\mathcal{D}'_-$, the new corpus $\mathcal{D}' = \mathcal{D}_0 \cup \mathcal{D}'_+ \cup \mathcal{D}'_-$, and the new relevance labels set $\mathcal{R}' = \mathcal{R}'_+ \cup \mathcal{R}'_-$, which form the final dataset.

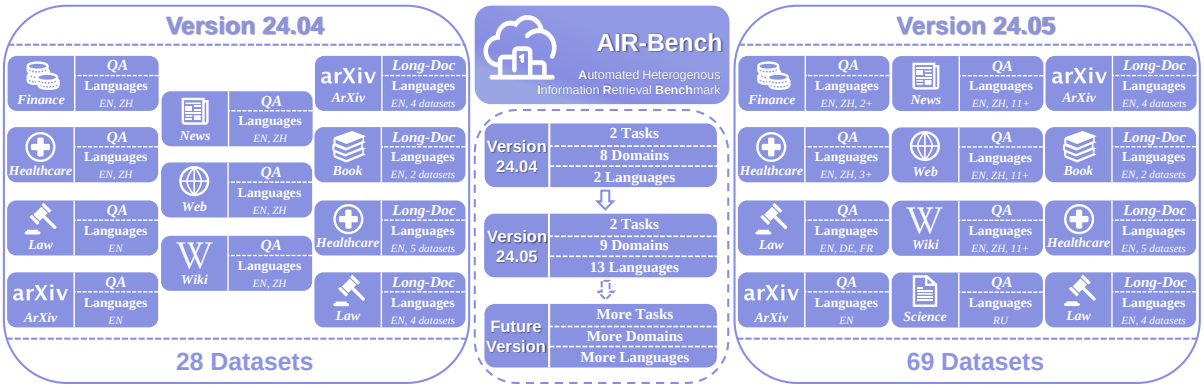



Figure 2: An overview of the diverse tasks, domains, languages, and datasets in AIR-BENCH 24.04 and 24.05.

## 3 Overview of AIR-BENCH

In this section, we provide an overview of AIR-BENCH from three perspectives: tasks, datasets and software.

**Tasks**. AIR-BENCH currently covers two retrieval tasks to meet the evaluation needs in different scenarios: 1) **QA**. This task focuses on the classic question answering scenarios (Voorhees et al., 1999), where the corpus consists of a large collection of documents. Following BEIR (Thakur et al., 2021), we utilize nDCG@10 as the main metric for the QA task. 2) **Long-Doc**. This task is closely related with today's LLM and RAG applications (Lewis et al., 2020), where the corpus consists of chunks from a lengthy document. Given that the proportion of positive documents precedes the ranking of positive documents in the RAG scenario, we utilize Recall@10 as the main metric for the Long-Doc task. AIR-BENCH will be extended to cover more retrieval tasks in the future.

**Datasets**. As shown in Figure 2, AIR-BENCH currently has two distinct versions, 24.04 and 24.05, where the latest version 24.05 consists of a total of 69 datasets, covering 9 domains and 13 languages on two retrieval tasks. We hope to incorporate more domains and languages in the future version to provide an increasingly comprehensive evaluation benchmark for community developers. More details are available in Appendix B.1.

**Software**. We develop the AIR-BENCH software[2] to facilitate the evaluation of any information retrieval methods. Besides, we maintain a Hugging Face leaderboard[3] with all datasets and models. For more details, please refer to Appendix C.

[2] *Anonymous-Link-2*
[3] *Anonymous-Link-1*

| Task → | QA | | Long-Doc | |
|---|---|---|---|---|
| Split → | dev | test | dev | test |
| # of datasets → | 54 | 53 | 4 | 11 |
| *Query Type* | | | | |
| HOW | 16.4% | 17.6% | 17.0% | 19.7% |
| WHAT | 34.1% | 30.9% | 28.5% | 33.1% |
| WHEN | 4.8% | 5.9% | 1.1% | 1.2% |
| WHERE | 3.0% | 3.2% | 0.9% | 0.8% |
| WHICH | 4.7% | 5.3% | 4.4% | 4.0% |
| WHO | 7.3% | 7.6% | 8.7% | 4.0% |
| WHY | 3.2% | 3.2% | 6.4% | 3.8% |
| YES/NO | 4.2% | 4.1% | 5.5% | 6.9% |
| CLAIM | 22.2% | 22.1% | 27.5% | 26.3% |
| OTHERS | 0.1% | 0.1% | 0% | 0.2% |

Table 3: The type distribution of queries in each split for each task in AIR-BENCH 24.05.

## 4 Diversity Analysis

To analyze the query type diversity of AIR-BENCH, we utilize GPT-4o[4] (Achiam et al., 2023) as labeler to label the type of the generated queries. Specifically, given a query, we prompt GPT-4o to select the most suitable type for the query from the optional types.

The statistics are grouped by tasks and splits in Table 3. Based on the results, we can make the following observations. Firstly, both the QA and Long-Doc tasks have the highest frequency of WHAT queries, followed by CLAIM queries as the second most common, and HOW queries as the third. Additionally, the QA task exhibits a more balanced distribution of the other query types, whereas the Long-Doc task shows a lower frequency of WHEN queries and WHERE queries. Lastly, a small number of queries are classified as OTHERS, reflecting the diverse types of queries present in AIR-BENCH to some extent. Further diversity analysis of AIR-BENCH is presented in Appendix B.3.

[4] gpt-4o-2024-08-06

| Model | Size | R-MSMARCO | | G-MSMARCO | | | |
|---|---|---|---|---|---|---|---|
| | | | | w/ quality control | | w/o quality control | |
| | | nDCG@10 | Rank | nDCG@10 | Rank | nDCG@10 | Rank |
| repllama-v1-7b-lora-passage (Ma et al., 2023) | 6.74B | 48.000 | 1 | 59.625 | 1 | 33.434 | 2 |
| e5-large-v2 (Wang et al., 2022b) | 335M | 45.232 | 2 | 55.260 | 4 | 32.581 | 5 |
| multilingual-e5-large (Wang et al., 2024) | 560M | 45.119 | 3 | 54.431 | 5 | 32.099 | 6 |
| multilingual-e5-base (Wang et al., 2024) | 278M | 44.130 | 4 | 52.581 | 8 | 30.870 | 8 |
| bge-large-en-v1.5 (Wang et al., 2024) | 335M | 44.122 | 5 | 55.513 | 3 | 33.119 | 4 |
| e5-mistral-7b-instruct (Wang et al., 2023) | 7.11B | 43.787 | 6 | 59.015 | 2 | 36.186 | 1 |
| e5-small-v2 (Wang et al., 2022b) | 33.4M | 43.104 | 7 | 51.456 | 10 | 30.471 | 10 |
| e5-base-v2 (Wang et al., 2022b) | 109M | 43.056 | 8 | 51.438 | 11 | 30.411 | 11 |
| bge-small-en-v1.5 (Xiao et al., 2024) | 33.4M | 42.553 | 9 | 51.528 | 9 | 30.155 | 13 |
| bge-base-en-v1.5 (Xiao et al., 2024) | 109M | 42.388 | 10 | 54.292 | 7 | 32.067 | 7 |
| multilingual-e5-small (Wang et al., 2024) | 118M | 42.253 | 11 | 47.989 | 15 | 28.579 | 16 |
| simlm-base-msmarco-finetuned (Wang et al., 2022a) | 110M | 41.675 | 12 | 48.102 | 14 | 30.548 | 9 |
| jina-embeddings-v2-base-en (Günther et al., 2023) | 137M | 39.887 | 13 | 51.112 | 12 | 29.560 | 14 |
| jina-embeddings-v3 (Sturua et al., 2024) | 572M | 39.787 | 14 | 51.098 | 13 | 30.297 | 12 |
| bge-m3 (Chen et al., 2024b) | 568M | 39.565 | 15 | 54.404 | 6 | 33.286 | 3 |
| contriever-msmarco (Izacard et al., 2022) | 109M | 36.570 | 16 | 47.127 | 16 | 29.231 | 15 |
| msmarco-roberta-base-ance-firstp (Xiong et al., 2021) | 125M | 33.637 | 17 | 42.107 | 17 | 24.798 | 17 |
| BM25 (Robertson and Zaragoza, 2009) | - | 26.211 | 18 | 34.155 | 18 | 22.582 | 18 |
| **Spearman Rank Correlation Coefficient (P-value)** | | - | | 0.8204 (3e-5) | | 0.7028 (1e-3) | |

Table 4: The consistency between the testing data generated by the pipeline of AIR-BENCH and the human-labeled testing data. We use the MS MARCO passage ranking dataset (Bajaj et al., 2016) to evaluate the consistency. For the public link of the models appearing in the table, please refer to Table 15.

## 5 Experiment

In this section, we aim to address the following research questions:

**RQ1:** How well does the LLM-generated testing data in AIR-BENCH align with the human-labeled testing data?

**RQ2:** What additional evaluation functionalities dose AIR-BENCH offer compared to MTEB/BEIR?

**RQ3:** How effectively can AIR-BENCH distinguish the capabilities of distinct IR models?

### 5.1 Consistency Analysis (RQ1)

Thomas et al. (2024) have demonstrated that LLMs like OpenAI's GPT-4 are as accurate as human labelers when generating high-quality golden labels for search system. Based on this conclusion, we attempt to examine how well the LLM-generated testing data aligns with human-labeled testing data.

**Setup**. We utilize MS MARCO passage ranking dataset (Bajaj et al., 2016) to access the consistency between the LLM-generated testing data in AIR-BENCH and human-labeled testing data. Specifically, we use the positive passages in the raw MS MARCO dev split as the candidate positives ($d_i^+$ in Stage 2, refer to Section 2.3), and finally generate a new MS MARCO passage ranking dataset. The raw MS MARCO passage ranking dataset (dev split) is denoted as *R-MSMARCO*, and the new generated MS MARCO passage ranking dataset is denoted as *G-MSMARCO*. Table 5 shows the comparison of R-MSMARCO and G-MSMARCO.

| | #corpus | #queries | #positives |
|---|---|---|---|
| R-MSMARCO | 8,841,823 | 6,980 | 7,437 |
| G-MSMARCO | 8,872,840 | 6,319 | 31,447 |
| w/o quality control | 8,878,865 | 7,429 | 7,429 |

Table 5: Comparison of R-MSMARCO and G-MSMARCO. R-MSMARCO is the raw MS MARCO passage ranking dataset (Bajaj et al., 2016), and G-MSMARCO is the generated MS MARCO passage ranking dataset in AIR-BENCH. **#corpus** represents the number of documents in the corpus, **#queries** represents the number of queries, and **#positives** represents the number of positive relevance labels. Since there are some generated hard negative documents in the corpus of G-MSMARCO, it is slightly larger than the corpus of R-MSMARCO.

To examine how well G-MSMARCO aligns with R-MSMARCO, we evaluate 18 IR models on R-MSMARCO and G-MSMARCO using nDCG@10, and compute the Spearman rank correlation coefficient (Spearman, 1961) between their rankings on

| Model | Size | MTEB (English) | | | | AIR-BENCH 24.05 (English, test) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Overall 56 datasets | | Retrieval (BEIR) 15 datasets | | QA 7 datasets | | Long-Doc 11 datasets | |
| | | Avg. | Rank | nDCG@10 | Rank | nDCG@10 | Rank | Recall@10 | Rank |
| *LLM-based Embedding Models* | | | | | | | | | |
| NV-Embed-v2 | 7.85B | 72.31 | 1 | 62.65 | 1 | 53.35 | 3 | 73.45 | 1 |
| bge-en-icl | 7.11B | 71.24 | 2 | 61.67 | 2 | 53.60 | 2 | 72.62 | 3 |
| bge-en-icl-e5data | 7.11B | 64.67 | 10 | 59.59 | 6 | 54.46 | 1 | 73.43 | 2 |
| SFR-Embedding-2_R | 7.11B | 70.31 | 3 | 60.18 | 5 | 50.80 | 7 | 65.83 | 9 |
| gte-Qwen2-7B-instruct | 7.61B | 70.24 | 4 | 60.25 | 3 | 51.87 | 5 | 63.97 | 11 |
| NV-Embed-v1 | 7.85B | 69.32 | 5 | 59.36 | 7 | 50.97 | 6 | 72.08 | 4 |
| Linq-Embed-Mistral | 7.11B | 68.17 | 6 | 60.19 | 4 | 49.76 | 9 | 70.02 | 5 |
| SFR-Embedding-Mistral | 7.11B | 67.56 | 7 | 59.00 | 8 | 52.78 | 4 | 68.10 | 6 |
| e5-mistral-7b-instruct | 7.11B | 66.63 | 8 | 56.89 | 10 | 49.88 | 8 | 66.91 | 7 |
| *Large-size Embedding Models* | | | | | | | | | |
| gte-large-en-v1.5 | 434M | 65.39 | 9 | 57.91 | 9 | 46.251 | 11 | 60.71 | 14 |
| multilingual-e5-large-instruct | 560M | 64.41 | 11 | 52.47 | 12 | 45.39 | 12 | 63.96 | 12 |
| bge-large-en-v1.5 | 335M | 64.23 | 12 | 54.29 | 11 | 44.91 | 13 | 61.86 | 13 |
| e5-large-v2 | 335M | 62.20 | 13 | 50.56 | 14 | 46.253 | 10 | 66.16 | 8 |
| multilingual-e5-large | 560M | 60.89 | 14 | 51.40 | 13 | 43.91 | 14 | 65.71 | 10 |
| *Lexical Method* | | | | | | | | | |
| BM25 | - | - | - | 40.76 | 15 | 39.16 | 15 | 53.09 | 15 |

Table 6: Comparison of the performance of 15 IR models on AIR-BENCH and MTEB/BEIR. The results on MTEB/BEIR are directly taken from the MTEB leaderboard. For detailed information of the models appearing in the table, please refer to Table 15. The detailed results for each dataset in AIR-BENCH are available in Appendix F.

R-MSMARCO and G-MSMARCO as the consistency metric.

**Main Results**. As shown in Table 4, the Spearman rank correlation coefficient is 0.8204 with a p-value of 3e-5, indicating that the LLM-generated testing data aligns well with the human-labeled testing data. Overall, each model achieves higher nDCG@10 on G-MSMARCO than on R-MSMARCO. This can be largely attributed to more comprehensive quality control strategy of AIR-BENCH, which results in more positives for each query (see Table 5).

**Ablation of Quality Control**. To demonstrate the necessity of the quality control stage in the data generation pipeline of AIR-BENCH, we also evaluate the consistency between R-MSMARCO and G-MSMARCO generated *without quality control*. As shown in Table 4, the correlation coefficient shows a significant degradation (0.8204 $\rightarrow$ 0.7028). Besides, the nDCG@10 of each model on G-MSMARCO without quality control also has a huge drop, due to some low-quality queries and very limited positives (see Table 5, there are 1,110 low-quality queries and only 7,429 positives). Therefore, quality control stage is necessary to ensure the data generation pipeline a reliable data generation pipeline.

**Robustness of Consistency**. To investigate the robustness of consistency, we simulate 30 generation processes by randomly sampling 2,000 generated queries from G-MSMARCO on each occasion. After each sampling, we access the consistency between the sampled G-MSMARCO and R-MSMARCO. As illustrated in Figure 3, the LLM-generated testing data exhibits stable and strong consistency with the human-labeled testing data, highlighting the robustness of this consistency.

### 5.2 Comparison with MTEB/BEIR (RQ2)

To investigate what additional evaluation functionalities AIR-BENCH can offer compared to MTEB (Muennighoff et al., 2023) and BEIR (Thakur et al., 2021), we compare the performance of 15 IR models on AIR-BENCH and MTEB/BEIR.

**Setup**. In addition to 14 large-size and LLM-based embedding models exhibiting superior performances on MTEB/BEIR, we also evaluate the performance of lexical method BM25 (Robertson and Zaragoza, 2009).

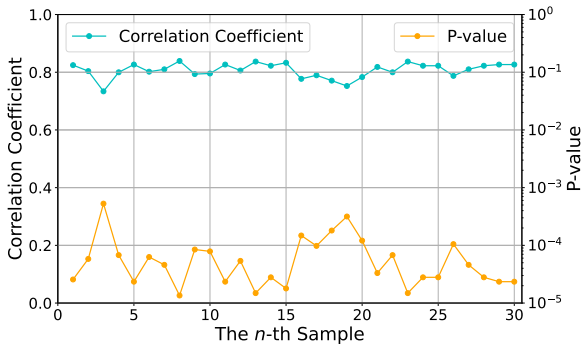



Figure 3: Robustness analysis of the consistency between the LLM-generated testing data and the human-labeled testing data. The mean correlation coefficient is 0.8067 with a mean p-value of 8e-5 across 30 simulated generation processes.

**Main Results**. As presented in Table 6, we can make the following observations based on the comparison results. 1) LLM-based embedding models generally outperform large-size embedding models on both AIR-BENCH and MTEB/BEIR, largely due to the superior generalization ability of LLMs. Besides, BM25 performs worse than all embedding models on both AIR-BENCH and BEIR. 2) The QA task and the Long-Doc task in AIR-BENCH exhibit a level of heterogeneity. The Spearman rank correlation coefficient between the rankings of the nine LLM-based embedding models across the two tasks is only 0.6, with a p-value of 0.0876. Moreover, as a large-size embedding model, `e5-large-v2` even outperforms some LLM-based embedding models on the Long-Doc task. 3) By comparing the results on AIR-BENCH and MTEB/BEIR, we observe that better performance on MTEB/BEIR may not indicate better performance on AIR-BENCH. For example, according to Li et al. (2024), `bge-en-icl` utilizes more in-domain training data in MTEB/BEIR than `bge-en-icl-e5data` and achieves more superior performance on MTEB/BEIR. However, compared to `bge-en-icl-e5data`, `bge-en-icl` shows performance degradation on AIR-BENCH, including both the QA task (54.46 $\rightarrow$ 53.60) and the Long-Doc task (73.43 $\rightarrow$ 72.62). This suggests that increased in-domain training data in MTEB/BEIR may introduce a risk of over-fitting, leading to a decline in the generalization ability of embedding models.

In conclusion, as a new benchmark, AIR-BENCH can offer additional evaluation functionalities for community developers compared to MTEB/BEIR.

| Dataset (↓) | mContriever | mContriever-finetuned |
|---|---|---|
| | nDCG@10 | nDCG@10 / Training Data |
| finance_en | 39.452 | 41.281 (↑ 1.829)<br>FiQA (Maia et al., 2018) |
| healthcare_zh | 14.557 | 17.351 (↑ 2.794)<br>cMedQAv2 (Zhang et al., 2018) |
| law_de | 5.614 | 6.687 (↑ 1.073)<br>Hoppe et al. (2021) |
| law_fr | 3.102 | 4.325 (↑ 1.223)<br>BSARD (Louis and Spanakis, 2022) |
| web_hi | 19.067 | 30.103 (↑ 11.036)<br>mMARCO (Bonifacio et al., 2021) |
| wiki_ar | 38.159 | 43.470 (↑ 5.311)<br>MIRACL (Zhang et al., 2023) |

Table 7: AIR-BENCH can showcase models' performance enhancement in specific domains. The training process takes 100 steps for cMedQAv2, and 50 steps for the other datasets.

### 5.3 Distinguishing Models (RQ3)

To examine how effectively AIR-BENCH can distinguish the capabilities of distinct IR models, we evaluate the performance of a single model before and after fine-tuning to illustrate that AIR-BENCH can reflect the performance enhancement of IR models in specific domains.

**Setup**. We fine-tune mContriever[5] (Izacard et al., 2021), using domain-specific training datasets and compare the model's performance on the corresponding datasets in AIR-BENCH before and after fine-tuning. Specifically, we fine-tune[6] mContriever with FlagEmbedding tool[7] to enhance its domain-specific capabilities. The domain-specific training data used for fine-tuning is independent of the corresponding testing data in AIR-BENCH.

**Main Results**. Table 7 presents the detailed information about each domain-specific training dataset and compares the model's performance on the corresponding dataset in AIR-BENCH before and after fine-tuning. For example, after fine-tuning with the Hindi training data from mMARCO (Bonifacio et al., 2021), the performance of mContriever on the web_hi dataset in AIR-BENCH improves from 19.067 to 30.103. This trend is also observed in other domains, such as finance, healthcare, law and wiki. Therefore, AIR-BENCH effectively reflects the performance enhancement of IR mod-

[5] https://huggingface.co/facebook/mcontriever-msmarco

[6] The learning rate is $2 \times 10^{-4}$, the warmup ratio is 0.1, and the weight decay is 0.01. The training process takes around a hundred steps with a total batch size of 64 on 8 A800 GPUs.

[7] https://github.com/FlagOpen/FlagEmbedding

els in specific domains following fine-tuning with domain-specific training datasets.

We also evaluate a diverse set of IR models on AIR-BENCH to further demonstrate its capability of distinguishing different models across multiple dimensions, including model type, domain, and language. Refer to Appendix F.1 for the details.

## 6 Related Work

The related works are reviewed from two aspects: evaluation datasets for IR, and synthetic data generation for IR.

### 6.1 Evaluation Datasets for IR

Evaluation datasets are critically important for the development of IR models.

In recent years, a series of milestone works have been introduced to the community. As the earlier contributions, MS MARCO (Bajaj et al., 2016) includes Bing search questions paired with human-labeled relevant passages from Web documents. Natural Questions (NQ) (Kwiatkowski et al., 2019) consists of Google search queries with human-labeled relevant Wikipedia pages. Both MS MARCO and NQ are designed for open-domain question answering tasks in English. Recent works like Mr.TyDi (Zhang et al., 2021), while MIRACL (Zhang et al., 2023) focus on multilingual retrieval in non-English languages. Mr.TyDi covers 11 languages and MIRACL encompasses an extended 18 languages. BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) are introduced to benchmark IR models in a general-domain zero-shot setting, including multiple existing datasets from diverse tasks and domains.

However, all of these benchmarks, which rely on pre-defined domains and human-labeled data, face limitations in addressing evaluation needs for emerging domains both cost-effectively and efficiently. Recently, several studies have explored the application of large language models for retrieval evaluation in retrieval-augmented generation (RAG) systems (Es et al., 2024; Saad-Falcon et al., 2024; Salemi and Zamani, 2024), offering a promising solution to this challenge. Nonetheless, a comprehensive IR benchmark that addresses this limitation remains insufficiently developed.

### 6.2 Synthetic Data Generation for IR

The tasks and domains in IR applications are often diverse and dynamic, meaning that the the training and evaluation data are frequently unavailable for new tasks and domains. As a result, it becomes challenging to fine-tune and evaluate IR models in these contexts.

Several recent works (Bonifacio et al., 2022; Dai et al., 2023; Jeronymo et al., 2023; Khramtsova et al., 2024; Thakur et al., 2024) have focused on addressing the scarcity of domain-specific training data by prompting LLMs to generate synthetic training data. Wang et al. (2023) and Chen et al. (2024a) employ LLMs to generate synthetic task and training data. Lee et al. (2024b) further refines the synthetic training data by using LLMs to select more relevant positives and negatives.

However, there is currently limited research addressing the scarcity of domain-specific evaluation datasets. Thomas et al. (2024) have demonstrated that powerful LLMs can generate high-quality golden labels for search system with accuracy comparable to human labelers, laying a solid foundation for our work. Our experiment results also demonstrate that the LLM-generated testing data aligns well with the human-labeled testing data. To our knowledge, AIR-BENCH is the first comprehensive IR benchmark that utilizes the LLM-generated datasets to perform evaluation.

## 7 Conclusion

In this paper, we introduce a new IR benchmark AIR-BENCH, which is highlighted for three main features: 1) Automated, 2) Heterogeneous, and 3) Dynamic. We demonstrate that the generated testing data in AIR-BENCH is highly consistent with the human-labeled testing data, which makes AIR-BENCH a dependable benchmark for evaluating IR models. Additionally, we demonstrate that AIR-BENCH can offer additional evaluation functionalities compared to MTEB/BEIR. Last but not least, we demonstrate that AIR-BENCH can effectively distinguish the capabilities of distinct IR models from multiple dimensions.

AIR-BENCH currently covers 2 tasks, 9 domains and 13 languages, including a total of 69 datasets. In the future, AIR-BENCH will be extended to cover more tasks, domains and languages to provide an increasingly comprehensive evaluation benchmark for community developers. We welcome datasets contributions to AIR-BENCH[8] as well as the model submissions to our leaderboard[9].

[8] *Anonymous-Link-2*
[9] *Anonymous-Link-1*

## Limitations

While AIR-BENCH aims to be a comprehensive IR benchmark, incorporating new features to address limitations in existing benchmarks, it still has several inherent constraints: 1) Dependence on real-world corpora. The generation process for each dataset in AIR-BENCH begins with corpora preparation. Consequently, the availability of real-world corpora is essential for generating an evaluation dataset. Future extensions of AIR-BENCH will be significantly influenced by this factor. 2) Reliance on capabilities of LLM. We design complex prompts for the LLM to diversify the generated queries. The quality of these queries is largely dependent on the LLM's capabilities, including its instruction-following, domain-specific, and language-specific capabilities. 3) Potential biases from quality control models. In addition to the LLM, we employ several existing IR models during the quality control stage. This reliance may introduce potential biases in the final evaluation dataset.

## Ethics Consideration

Since AIR-BENCH is built on testing data generated by LLM, it may inherit potential biases, toxicity, and other issues present in the LLM used during the generation process. Additionally, considering that the corpora utilized in the generation process are derived from the real-world sources, they may contain sensitive content. Therefore, the testing data in AIR-BENCH may only be used for evaluation purposes.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2387–2392, New York, NY, USA. Association for Computing Machinery.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset. corr abs/2108.13897 (2021). *arXiv preprint arXiv:2108.13897*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2024a. Little giants: Synthesizing high-quality embedding data at scale. *arXiv preprint arXiv:2410.18634*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. 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 and virtual meeting. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Tobias Daudert and Sina Ahmadi. 2019. CoFiF: A corpus of financial reports in French language. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 21–26, Macao, China.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim

Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Peter Henderson*, Mark S. Krass*, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *arXiv preprint*.

Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck. 2021. Towards intelligent legal advisors for document retrieval and question-answering in german legal documents. In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32. IEEE.

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*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, and Guido Zuccon. 2024. Leveraging llms for unsupervised dense retriever ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1307–1317, New York, NY, USA. Association for Computing Machinery.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy-yong Sohn, and Chanyeol Choi. 2024. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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:2405.17428*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024b. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

David Lewis. 1997. Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C52G6M.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

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*.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset. *Preprint*, arXiv:2305.01526.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Jerry Liu. 2022. LlamaIndex.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in french. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, page 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv:2310.08319*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

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

Inc. NetEase Youdao. 2023. Bcembedding: Bilingual and crosslingual embedding for rag. https://github.com/netease-youdao/BCEmbedding.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *Preprint*, arXiv:2306.02069.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2395–2400, New York, NY, USA. Association for Computing Machinery.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Charles Spearman. 1961. The proof and measurement of association between two things.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Nandan Thakur, Jianmo Ni, Gustavo Hernandez Abrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7699–7724, Mexico City, Mexico. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. 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)*.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940, New York, NY, USA. Association for Computing Machinery.

Fabián Villena. 2019. Multilingual medical corpora.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Simlm: Pre-training with representation bottleneck for dense passage retrieval. *ArXiv*, abs/2207.02578.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. 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. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, 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*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

# A Details on Benchmark Construction

In this section, we provide more details on the construction of datasets in AIR-BENCH. Considering the powerful capability of GPT-4 (Achiam et al., 2023), we use *gpt-4-1106-preview*[10] as the LLM through the generation pipeline.

## A.1 Corpora Preparation

AIR-BENCH currently covers two different tasks: QA and Long-Doc. For QA task, we directly use the real-world dataset as the corpus, such as Wikipedia, mC4(Raffel et al., 2020), CC-News(Hamborg et al., 2017), etc. We filter out text that is either too short or too long and make a straightforward attempt to remove any information that names or uniquely identifies individuals, as well as any offensive content. For Long-Doc task, we first select one long document for each dataset, such as book, ArXiv paper, legal document, etc., and remove table of contents and references. Then we use the node parser[11] tool from LlamaIndex(Liu, 2022) to split the long document into fixed-size chunks[12] as the corpus. All corpora used in AIR-BENCH are available in Appendix B.1.

## A.2 Candidate Generation

### A.2.1 Query Generation

To diversify the generated queries, we consider the following attributes when designing the prompt.

**Query Length**. This refers to the length of the query. We consider four different categories based on word count: *less than 5 words*, *less than 10 words*, *10 to 20 words*, and *at least 20 words*. The ratio of the number of queries in these categories is 1:4:2:1.

**Query Type**. This refers to the type of the query. We consider three different types: *question*, *problem*, and *claim*. Based on our observation, the "problem" type is usually more difficult than the "question" type. The ratio of the number of queries in these three types is 3:1:1. For Long-Doc task, considering that the chunks in the corpus are derived from the same long document, the topics of these chunks are highly related. Therefore, we only utilized two types for Long-Doc task: question and claim. For the "claim" type, we observe that when the claim is too short, it will become too ambiguous

[10] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[11] SimpleNodeParser: https://github.com/run-llama/llama_index

[12] chunk_size=200, chunk_overlap=50

to be a high-quality query. Therefore, the query length for the "claim" type is only sampled from "between 10 and 20 words" and "at least 20 words".

**Information-based Type**. This refers to the type of the information used when formulating queries. We consider two different types: *queries based on the overall information in the document*, and *queries based on the partial information beyond the main topic of the document*. The ratio of the number of queries in these two types is 1:1.

**Expression Style**. This refers to the style of query formulation. The three attributes mentioned above are used in Step 4. In Step 5, we consider different types of expression styles, allowing the LLM to rewrite the queries using various styles, thereby enhancing the diversity of query formulations. There are seven different styles in total: *concise*, *casual*, *informal*, *formal*, *professional*, *complicated*, and *academic*. During the rewriting process, the sampling probabilities for these styles are in the ratio of 5:3:3:1:1:1:1.

### A.2.2 Hard Negative Generation

To improve the difficulty of the generated datasets, we prompt LLM to generate 3-7 hard negative documents based on the rewritten query and the original positive document. For Long-Doc task, considering the chunks are extracted from the same long document and some of them have been hard enough, we do not generate additional hard negatives. The statistics of the number of hard negatives in each dataset are available in Appendix B.1.

## A.3 Quality Control

We present more details on how we use LLM as labeler to label the relevance, select the embedding model and multiple re-ranking models, and set the predetermined threshold for pre-labeling.

**Use LLM as labeler**. Thomas et al. (2024) demonstrated that LLMs like OpenAI's GPT-4 are as accurate as human labelers when generating high-quality golden labels for search system. Zhuang et al. (2024) showed that incorporating fine-grained relevance labels into the prompt for LLM rerankers significantly improves their performance on zero-shot reranking. In our paper, we use GPT-4 as labeler with a 4-level relevance generation strategy. The prompt we used is shown in Table 8.

**Embedding Model**. Considering that the corpora in AIR-BENCH are multilingual, we use bge-m3[13] as the embedding model to recall the top-1000 relevant documents.

**Multiple Re-ranking Models**. For the datasets in English and Chinese, we use the following three re-ranking models: bge-reranker-large[14], bce-reranker-base_v1[15], mmarco-mMiniLMv2-L12-H384-v1[16]. For the datasets in the other languages, we use the following three re-ranking models: bge-reranker-v2-m3[17], mmarco-mMiniLMv2-L12-H384-v1, bge-reranker-v2-gemma[18].

**Predetermined Threshold**. For the hard negative documents, we set the threshold to 20. For the other documents, we set the threshold to 10.

| |
|---|
| For the following query and document, judge whether the document is relevant to the query.<br><br>Query:<br>'''<br>{query}<br>'''<br><br>Document:<br>'''<br>{doc}<br>'''<br><br>Your output must be one of the following:<br>- 0: The document is not relevant to the query.<br>- 1: The document is superficially relevant but actually not relevant to the query.<br>- 2: The document is somewhat relevant to the query.<br>- 3: The document is relevant to the query.<br><br>Do not explain your answer in the output. Your output must be a single number. |

Table 8: Prompt used for LLM to label the relevance. {query} and {doc} are placeholders of query and document, respectively.

## A.4 Queries Split

After the quality control stage, we split the generated queries into different sets. For QA task, we split the queries in each dataset into dev set and test set in a 1:4 ratio. For Long-Doc task, we select one dataset as the dev set for each domain, and remain other datasets as the test set. Refer to Appendix B.1 for more details.

[13] https://huggingface.co/BAAI/bge-m3
[14] https://huggingface.co/BAAI/bge-reranker-large
[15] https://huggingface.co/maidalun1020/bce-reranker-base_v1
[16] https://huggingface.co/nreimers/mmarco-mMiniLMv2-L12-H384-v1
[17] https://huggingface.co/BAAI/bge-reranker-v2-m3
[18] https://huggingface.co/BAAI/bge-reranker-v2-gemma

# B AIR-BENCH Datasets

## B.1 Specifications

The available versions of AIR-BENCH are listed in Table 9.

| Version | Release Date | #domains | #languages | #datasets | Statistics |
|---|---|---|---|---|---|
| 24.04 | May 21, 2024 | 8 | 2 | 28 | Table 16 |
| 24.05 | Oct 17, 2024 | 9 | 13 | 69 | Table 17-21 |

Table 9: Available versions of AIR-BENCH.

For each dataset, we use the same format as BEIR, i.e. corpus, queries and qrels[19], which are all available in the Hugging Face Hub[20] of AIR-BENCH. To avoid the possible data leakage, we keep the qrels in test splits private. For the qrels in dev splits, we make them public to enable the developers to perform evaluation by themselves.

As the initial version, AIR-BENCH 24.04 only covered 2 languages, English and Chinese. Additionally, each dataset in AIR-BENCH 24.04 only contains the test set, which means that the developers could not know the evaluation results until they submit their model's search results to the leaderboard. As for the latest version AIR-BENCH 24.05, we have covered 13 languages, and included dev set and test set. The golden labels of dev set are made public, and the golden labels of test set remain private. Furthermore, the corpus size of some datasets in AIR-BENCH 24.04 is too large (such as 6.7M for wiki_en dataset and 2.4M for finance_zh dataset in QA task), which makes the download of datasets and the evaluation of models relatively inefficient. Therefore, in AIR-BENCH 24.05, we trimmed the large corpora to maintain a corpus size of around 1M for each dataset.

For the detailed statistics of all datasets in AIR-BENCH 24.04 and 24.05, please refer to Table 16 and Table 17-21, respectively. Note that we use the tokenizer[21] of OpenAI's GPT-4o[22] to count the token number for every language.

## B.2 Licenses

In Table 16-21, we also list the licenses of the source corpora used for the dataset generation in AIR-BENCH. All generated testing data in AIR-BENCH is licensed under CC BY-NC-SA-4.0[23]. The testing data in AIR-BENCH may only be used for evaluation purposes.

| | QA | | Long-Doc | |
|---|---|---|---|---|
| split → | dev | test | dev | test |
| # of datasets → | 54 | 53 | 4 | 11 |
| *Query Style* | | | | |
| FORMAL | 31.3% | 35.3% | 17.7% | 17.9% |
| INFORMAL | 44.3% | 44.7% | 28.3% | 27.4% |
| PROFESSIONAL | 7.0% | 6.8% | 8.2% | 10.3% |
| CASUAL | 0.8% | 0.7% | 0.5% | 0.6% |
| COMPLICATED | 0.5% | 0.3% | 0.7% | 1.4% |
| CONCISE | 8.3% | 5.4% | 12.3% | 12.5% |
| ACADEMIC | 7.8% | 6.8% | 32.2% | 29.8% |
| OTHERS | < 0.1% | < 0.1% | 0.1% | 0.1% |

Table 10: The style distribution of queries in each split for each task in AIR-BENCH 24.05.

## B.3 Additional Diversity Analysis

We provide more analysis of diversity to better characterize AIR-BENCH.

### B.3.1 Query Diversity

We also analyze the **style diversity** of the generated queries in AIR-BENCH. We still utilize GPT-4o[24] as labeler to label the style of the queries in AIR-BENCH. The optional query styles include: FORMAL, INFORMAL, PROFESSIONAL, CASUAL, COMPLICATED, CONCISE, ACADEMIC, and OTHERS. The statistics are grouped by tasks and splits in Table 10.

We can make the following observations according to the results. First of all, since the optional styles given to GPT-4o are not mutually exclusive, the ratio of the number of different styles is not consistent with the ratio we set in the generation stage (Step 5 of the Candidate Generation stage). Secondly, the QA task tends to have more INFORMAL queries, and Long-Doc task tends to have more ACADEMIC queries, which may be due to the fact that the long documents in the Long-Doc task are more academic related, such as ArXiv papers, books, etc. Finally, PROFESSIONAL queries and COMPLICATED queries account for a certain portion, which means that some queries in AIR-BENCH are probably challenging for IR models.

[19] *qrels* are the relevance labels for queries. The relevance label is 1 for the positive document, and 0 for the negative document.

[20] *Anonymous-Link-3*

[21] https://github.com/openai/tiktoken

[22] https://openai.com/index/hello-gpt-4o/

[23] https://creativecommons.org/licenses/by-nc-sa/4.0

[24] gpt-4o-2024-08-06: https://platform.openai.com/docs/models/gpt-4o

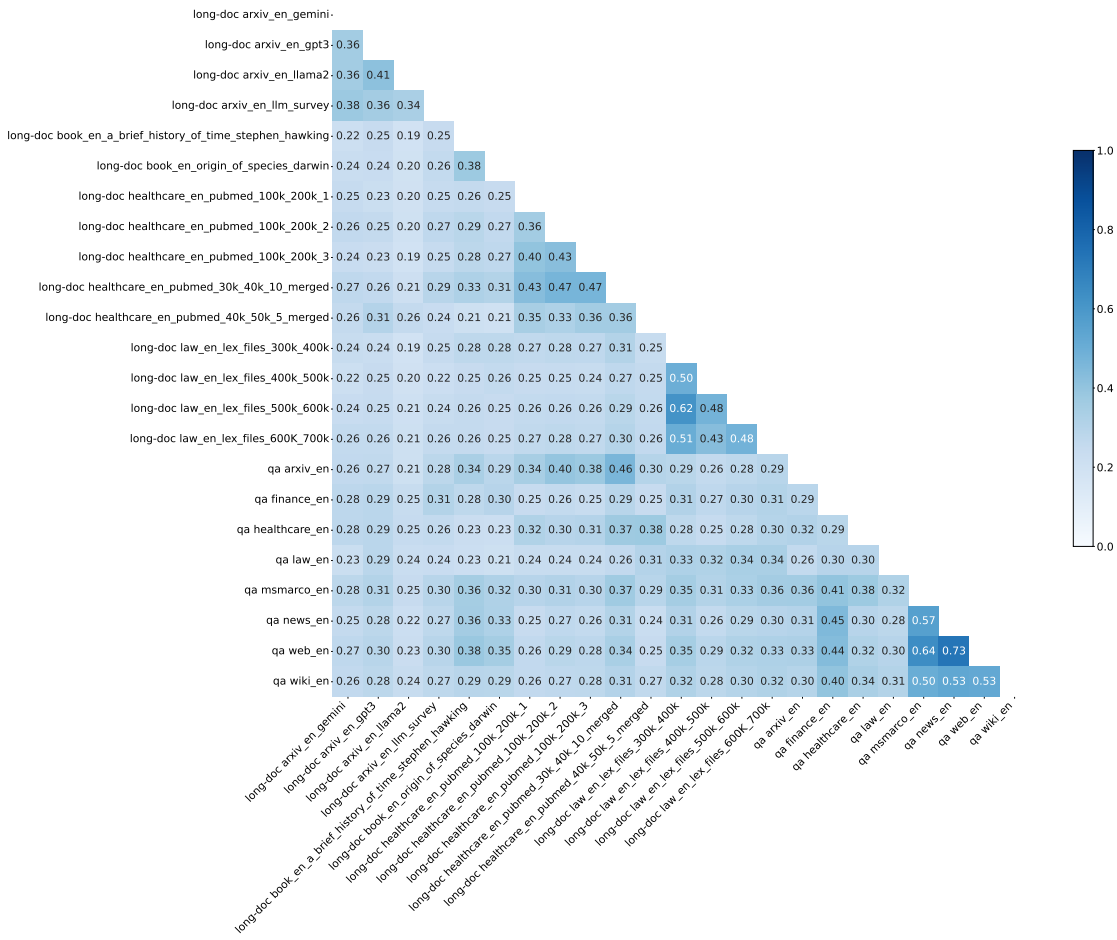



Figure 4: Pairwise weighted Jaccard similarity scores between AIR-BENCH English datasets. We use the tokenizer of GPT-4o to tokenize the corpus of each dataset.

### B.3.2 Corpus Diversity

Following the work of BEIR (Thakur et al., 2021), we compute the pairwise weighted Jaccard similaity scores between the datasets in AIR-BENCH. Considering that there are 69 datasets in total, we only present the results of datasets in English here. As shown in Figure 4, we can observe that the corpora from different domains have a low weighted Jaccard similarity word overlap, indicating that AIR-BENCH is a challenging benchmark where the IR methods must generalize well to diverse out-of-distribution domains.

## C AIR-BENCH Software

The AIR-BENCH software[25] makes it convenient for the evaluation of any information retrieval methods. With the provided Python framework, in order to evaluate a retrieval method, users only need to implement a `Retriever` that takes the queries and the corpus as the inputs, and returns the top-$k$ relevant documents for each query as the outputs. If the users want to evaluate the performance of retrieval-then-reranking method, they only need to additionally implement a `Reranker`, which takes the queries, the corpus, and the top-$k$ search results from `Retriever` as the inputs, and returns the re-ranked top-$k'$ ($k' \leq k$) relevant documents as the outputs.

We also maintain a Hugging Face leaderboard[26] with all datasets and models. To make the leaderboard more readable, we classify the submissions into three categories: 1) **Retrieval Only**. It means that this submission only uses a specific retrieval method to generate the top-$k$ search results. 2) **Reranking Only**. It means that this submission uses BM25 as the retrieval method and then uses a specific reranking method to re-rank the search results from BM25 to generate the re-ranked top-$k$ search results. 3) **Retrieval+Reranking**. It means that this submission first uses a specific retrieval method to generate the top-$k$ search results, and then uses a specific reranking method to re-rank

[25] *Anonymous-Link-2*

[26] *Anonymous-Link-1*

to get the final search results. It should be noted that our leaderboard only maintain the evaluation results for the test splits, and the evaluations results for the dev splits will be available on the MTEB leaderboard[27].

To facilitate the evaluation of existing IR models, we also develop the evaluation scripts based on two mainstream architectures: HuggingFace Transformers[28] (Wolf et al., 2020) and Sentence Transformers[29] (Reimers and Gurevych, 2019). These scripts are all available in our repository[30].

# D Examples

We list some examples of the generated testing data in Table 12-14.

# E Evaluation Details

## E.1 Models

For detailed information of the models appearing in this paper, please refer to Table 15. For the BM25 method, we employ the implementation from Pyserini[31] (Lin et al., 2021). For the evaluation of BM25-based re-ranking models, we evaluate the performance by re-ranking the top-100 search results from BM25 with the re-ranking models.

The models used in this paper are all publicly available (see Table 15 for the public link). We confirm that we did not violate the license of any model used in our paper.

## E.2 Parameters

When performing evaluation, we set the max length of both query and passage to 512 tokens. If the embedding models need task specific instruction, such as `e5-mistral-7b-instruct` (Wang et al., 2023), `SFR-Embedding-Mistral`, etc., we use the same instruction for all datasets: "Given a question, retrieve passages that answer the question", which is denoted as Instr-1. Considering that the queries in AIR-BENCH include both questions and claims, we also evaluate the performance of `e5-mistral-7b-instruct` with a more reasonable but more complex instruction: "Given a question or claim, retrieve passages that answer the question or support the claim", which is denoted as Instr-2. However, as shown in Table 11, the performance of e5-mistral-7b-instruct using Instr-2 is slightly worse than that using Instr-1, which may indicate that current models are not yet able to adapt well to more complex instruction.

For the total computational budget, we did not perform detailed statistics. However, based on our estimates, all evaluations in this paper required approximately 2000 GPU hours using 24 A800 (80GB) GPUs.

# F More Experiment Results

## F.1 Distinguishing Models

We evaluate a diverse set of IR models on AIR-BENCH to demonstrate its capability of distinguishing different models from multiple dimensions: model type, domain, language.

**Model Type**. As shown in Figure 5a and Figure 5b, we can observe the following three points on both QA task and Long-Doc task, regardless of whether the datasets are only in English or multilingual: 1) BM25 performs worse than all embedding models. 2) BM25 + bge-reranker-v2-m3 achieves more excellent performance than all of the embedding models. 3) The performance of embedding models from the same series scales with model size.

**Domain**. We evaluate three kinds of embedding models with the same model size (*large-size*), and compare their performances in each domain on AIR-BENCH. As shown in Figure 5c and Figure 5d, regardless of whether the task is QA or Long-Doc and whether the datasets are only in English or multilingual, no model is able to achieve the best performance on all domains. For more results, please refer to Appendix F.

**Language**. We evaluate three kinds of embedding models with the same model size (*large-size*) on the multilingual datasets of AIR-BENCH, and compare their performance on the datasets of each language. As shown in Figure 5e, we also observe that no model is able to achieve the best performance on all languages. For more results, please refer to Appendix F.

Apart from the results of large-size embedding models in Figure 5, we also perform investigation with base-size embedding models and LLM-based embedding models. The additional results are shown in Figure 6.

[27] https://huggingface.co/spaces/mteb/leaderboard
[28] https://github.com/huggingface/transformers
[29] https://github.com/UKPLab/sentence-transformers
[30] *Anonymous-Link-2*
[31] https://github.com/castorini/pyserini

| # of datasets → | **QA (English, test)** 7 datasets | **QA (Multilingual, test)** 53 datasets | **Long-Doc (English, test)** 11 datasets |
|---|---|---|---|
| e5-mistral-7b-instruct (Instr-1) | 49.880 | 48.077 | 66.908 |
| e5-mistral-7b-instruct (Instr-2) | 49.252 | 47.772 | 66.766 |

Table 11: Comparison of performances when using different evaluation parameters on AIR-BENCH. The metric for QA task is nDCG@10, and the metric for Long-Doc task is Recall@10.

### F.2 Detailed Evaluation Results

In this section, we provide the detailed evaluation results of each model on AIR-BENCH 24.05. Table 22 presents the detailed evaluation results of English IR models on AIR-BENCH 24.05. Table 23 presents the detailed evaluation results of multilingual IR models on AIR-BENCH. For detailed information of the models appearing in these tables, please refer to Table 15.

| |
|---|
| **Domain:** news; **Language:** English<br>**Original Positive:**<br>"It's hard to think of a part of the world that hasn't been touched by robotic advances this year.\n In 2016, strides were taken in the areas of robotic home delivery, cooking, tough terrain navigation and even attempts to conquer the beautiful game of football.\n Here are the top five robots of the year.\n While we're not quite at the singularity yet, more sophisticated automation is an inevitability of the future.\n The strides in Artificial Intelligence (AI) over the past decade have been huge, so expect to see a lot more in this area in the coming years.\n We just hope the tech guys making super AI fit it with an "off" switch so it can be unplugged when it wants to, you know, take over the world and destroy everything." |
| **Character:** Robotics Engineer |
| **Scenario:** Preparing a presentation on the yearly advancements in robotics technology. |
| **Original Query:** Which industries implemented robotic home delivery?<br>**Rewritten Query:** In which sectors has the implementation of autonomous delivery robots for residential services been observed? |
| **Hard Negative 1:**<br>"Autonomous technologies have been expanding rapidly across various industries, with drones making headway in aerial inspections and surveillance. Companies are investing in autonomous flight for package delivery, but primarily in commercial settings. The convenience and efficiency improvements in logistics are undeniable, but residential use isn't widespread yet."<br>**Hard Negative 2:**<br>"Residential sectors are increasingly relying on technology, with smart homes integrating systems for automated cleaning, energy management and advanced security. These innovations in domestic tech have redefined the way we live, promising a future where household chores are managed seamlessly through digital interfaces and remote controls."<br>**Hard Negative 3:**<br>"Recent developments in the robotics industry have witnessed significant progress in various sectors, such as industrial manufacturing, precision agriculture, and automated warehousing solutions. These robots have revolutionized production efficiency, crop management, and inventory control, enhancing economic output."<br>**Hard Negative 4:**<br>"In recent years, residential areas have seen an uptick in smart home innovations that include automated climate control, security systems with facial recognition, and voice-activated appliances. The integration of AI in household management has significantly enhanced the convenience and efficiency of daily living."<br>**Hard Negative 5:**<br>"Experts predict an expansion in the use of unmanned vehicles for military logistics and combat support missions. The autonomous systems being developed are designed for supply transport, surveillance, and even tactical offense, set to revolutionize battlefield strategies in the near future." |

Table 12: Random sampled examples for the generated testing data. Domain: news, Language: English.

**Domain:** healthcare; **Language:** English
**Original Positive:**
"Only two patients, 5 and 12 years old, with primary gastric NHL were found. Upper gastroduodenal endoscopy detected an ulcer in the lesser curvature of the body of the stomach, in both cases. Endoscopy revealed a moderate chronic gastritis in the antrum of both patients that was H. pylori associated in one of them who also suffered from chronic gastritis. Biopsy specimens demonstrated infiltration by Burkitt lymphoma (BL). The two patients received chemotherapy for 6 months. Additionally, one of the two patients received a triple therapy regimen with bismuth, amoxicillin, and metronidazole for H. pylori. Fifteen and six years later they are in complete remission, free of symptoms."

**Character:** College student

**Scenario:** Creating a presentation on the clinical manifestations and treatment outcomes of primary gastric non-Hodgkin's lymphoma in pediatric patients.

**Original Query:** How long did the pediatric patients receive chemotherapy for primary gastric NHL?
**Rewritten Query:** How long were the kids treated with chemo for their stomach lymphoma?

**Hard Negative 1:**
"In a recent clinical review, five pediatric cases of gastrointestinal complaints were assessed. The patients, ranging in age from 3 to 14 years, presented with various symptoms including abdominal pain, vomiting, and weight loss. In-depth medical evaluations, including blood tests, abdominal ultrasonography, and, for three patients, an upper gastroduodenal endoscopy, were conducted. The endoscopic examination in these three patients showed mild inflammation in the stomach lining and superficial gastric erosions in the antrum and the lesser curvature. None of the patients had a history of gastric malignancies, and there were no indications of Non-Hodgkin Lymphoma (NHL) or any other types of cancer. Helicobacter pylori infection was not detected in any of the cases. The patients' symptoms were managed with dietary modifications and antacid medications. Symptom relief was noted in all cases, and follow-up visits over the course of six months revealed significant improvement and no further gastrointestinal issues. The clinical team concluded that the symptoms were likely due to functional dyspepsia and emphasized the importance of considering less severe diagnoses when pediatric patients present with gastrointestinal symptoms."
**Hard Negative 2:**
"Two young individuals, aged 6 and 11, presented with abdominal discomfort and were subsequently screened for gastrointestinal disorders. Initial evaluation through pediatric upper gastrointestinal series indicated irregularities in the stomach lining, prompting further investigation. Comprehensive upper gastrointestinal endoscopies were performed, illuminating significant gastroesophageal reflux disease (GERD) in both patients, characterized by distinctive erosions in the esophagus and transient lower esophageal sphincter relaxations. GERD was particularly pronounced along the greater curvature of the stomach. Their evaluations also included biopsies of the gastric tissue, which fortunately ruled out malignancy, including lymphomas and other gastric cancers. To manage the GERD, both patients were placed on a rigorous treatment regimen including lifestyle modifications and proton-pump inhibitors (PPIs). Each was monitored regularly via follow-up endoscopies which demonstrated gradual improvements in esophageal tissue integrity. Concurrently, both were tested for H. pylori, with one testing positive. The H. pylori-positive patient underwent an eradication protocol with a combination therapy of clarithromycin, amoxicillin, and a PPI, resulting in successful elimination of the infection. Years later, through diligent management and follow-up, both individuals have achieved excellent control over their symptoms and maintain a good quality of life."
**Hard Negative 3:**
"Numerous pediatric cases have been reviewed to understand the duration and efficacy of chemotherapy in treating various forms of juvenile cancer. One study outlines the treatment plan for a pair of siblings, aged 7 and 14, diagnosed with acute lymphoblastic leukemia (ALL). The treatment protocol involved a comprehensive induction regimen followed by a consolidation phase. During the induction phase, which lasted for about a month, the patients were administered a combination of vincristine, prednisone, asparaginase, and an anthracycline. The consolidation phase incorporated methotrexate and 6-mercaptopurine and extended over several months. Intrathecal chemotherapy was included to prevent CNS disease. Maintenance therapy was subsequently initiated, which is scheduled to continue for a period of three years, with regular follow-ups to monitor remission status. It was observed that the older child had to face additional challenges due to the emergence of several therapy-related side effects. Despite the intensive treatment, both patients are currently responding positively with substantial remission observed in follow-up examinations. The study emphasizes the importance of a tailored approach to pediatric chemotherapy, taking into account not only the type of cancer but also individual patient factors and potential long-term outcomes."

Table 13: Random sampled examples for the generated testing data. Domain: healthcare, Language: English.

**Domain:** wiki; **Language:** English
**Original Positive:**
"Caffeine/ergotamine (trade name Cafergot) is the proprietary name of a medication consisting of ergotamine tartrate and caffeine. This combination is used for the treatment of headaches, such as migraine headache.\n\n Use\n\n Correct timing of use is important. Cafergot is an abortive headache treatment, which prevents the development of the headache, rather than a treatment for an established headache. The medication should be administered at the first sign of headache.\n\n There exist some limitations as to the maximum number of tablets that can be taken per day per week. Different sources of drug information may carry different information, and patients are encouraged to ask their pharmacist or prescriber about such details.\n\n Cafergot is currently available as a generic drug (ergotamine tartrate/caffeine)\n\n Mechanism of action\n\n According to a topic review on UpToDate, ërgotamine and dihydroergotamine (DHE 45) bind to 5HT 1b/d receptors, just as triptans do.T̈his along with binding to other serotonergic and dopaminergic receptors is their presumed mechanism of action in treating migraine.\n\n Adverse effects\n\n Because the vasoconstrictive effects of ergotamine and caffeine are not selective for the brain, adverse effects due to systemic vasoconstriction can occur. Cold feet or hands, angina pectoris, myocardial infarction, or dizziness are some examples. \n\n It has also been shown to be associated with mitral valve stenosis.\n\n References \n\n Antimigraine drugs\n Combination drugs"

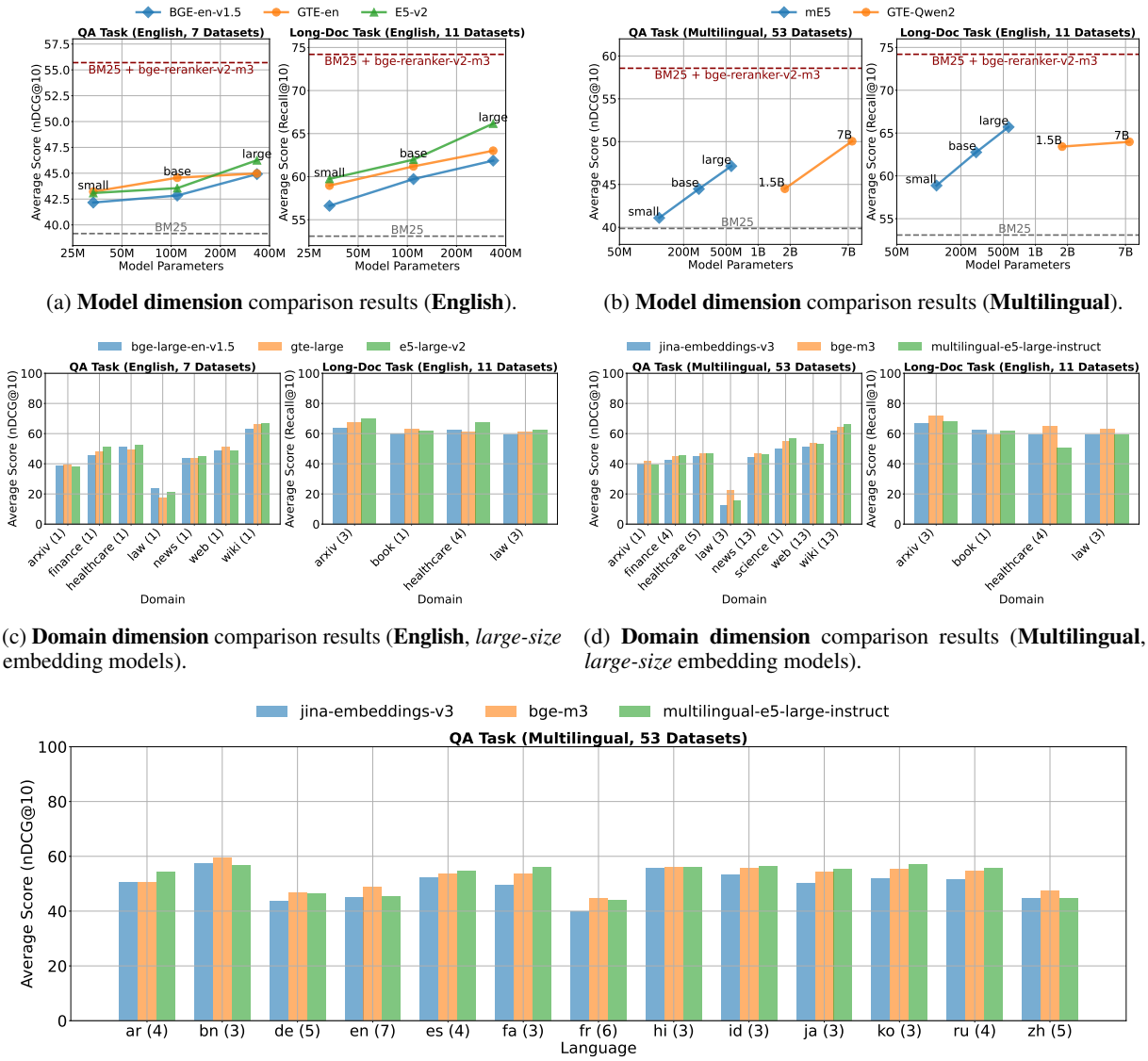
**Character:** Pharmacist

**Scenario:** Advising a patient on the proper usage of Cafergot, including timing and dosage limits.

**Original Query:** What is the optimal timing for administering Cafergot to treat migraine headaches?
**Rewritten Query:** At which temporal juncture is it considered most optimal to commence administration of Cafergot for the alleviation of cephalalgic discomfort characteristic of a migraine?

**Hard Negative 1:**
"The importance of adherence to a prescribed treatment regimen cannot be overstated, especially when managing chronic conditions such as hypertension and diabetes. Medications for these diseases, while different in function and timing from migraine treatments like Cafergot, require consistent and timely dosing to maintain health and prevent complications. For example, antihypertensive drugs must be taken daily to effectively control blood pressure and reduce the risk of heart attack and stroke. Similarly, diabetic patients must monitor their blood sugar levels regularly and administer insulin or oral hypoglycemic agents as directed to avoid hyperglycemic or hypoglycemic episodes. Although the precise timing may differ from abortive headache therapies, the principle of timing in medication administration is universally critical. Patients are advised to follow the specific instructions provided by their healthcare provider or pharmacist to achieve the best outcomes from their medication regimen. Furthermore, lifestyle modifications, such as diet and exercise, also play a vital role in the management of these conditions and should be initiated in conjunction with pharmacotherapy for an integrated approach to treatment."
**Hard Negative 2:**
"Caffeine and its Role in Pain Relief: An Overview\n\n Caffeine, a central nervous system stimulant, has been widely recognized for its ability to increase alertness and alleviate fatigue. Commonly found in various beverages such as coffee, tea, and energy drinks, caffeine is also included in certain pain relief medications. Its application in pain management is based on its pharmacological properties that enhance the efficacy of other analgesic compounds.\n\n Although not a primary treatment for migraine pain, caffeine is sometimes combined with analgesics like acetaminophen or aspirin to increase their effectiveness. The precise timing for administration of such combination therapies is generally flexible and tailored to individual patient needs. Unlike migraine-specific treatments, these over-the-counter remedies aim to reduce the severity of pain after onset of symptoms.\n\n Research into caffeine's role in pain relief extends beyond headaches to muscle soreness and other types of pain. While it possesses some anti-inflammatory properties, the exact mechanism through which caffeine exerts its effect on pain pathways is still being investigated. However, it is thought to involve adenosine receptor antagonism.\n\n Knowing the right amount of caffeine consumption for pain relief is crucial since excessive intake can cause side effects such as jitteriness, insomnia, and an increased heart rate. As with any medication or supplement, users should consult healthcare professionals to determine the appropriate dosage for their condition."

Table 14: Random sampled examples for the generated testing data. Domain: wiki, Language: English.




(a) **Model dimension** comparison results (**English**).

(b) **Model dimension** comparison results (**Multilingual**).

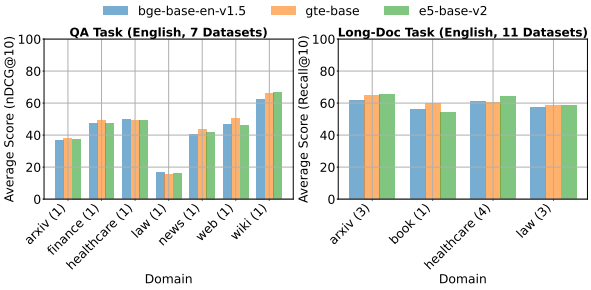
(c) **Domain dimension** comparison results (**English**, *large-size* embedding models).

(d) **Domain dimension** comparison results (**Multilingual**, *large-size* embedding models).

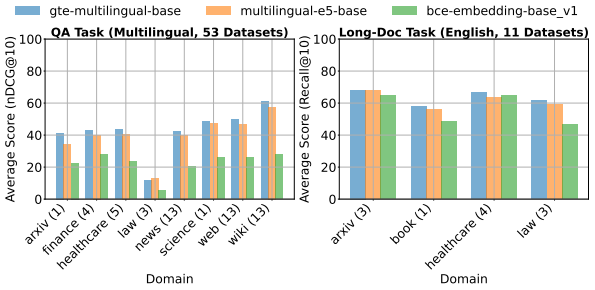



(e) **Language dimension** comparison results (**Multilingual**, *large-size* embedding models).

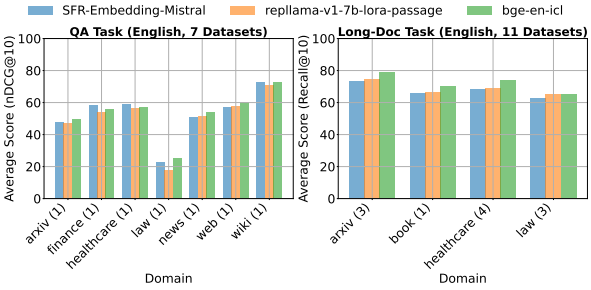
Figure 5: AIR-BENCH can distinguish models in different dimensions, including model dimension, domain dimension, and language dimension. For detailed information of the models appearing in this figure, please refer to Table 15. The detailed metric value and additional results on other model size are all available in Appendix F.



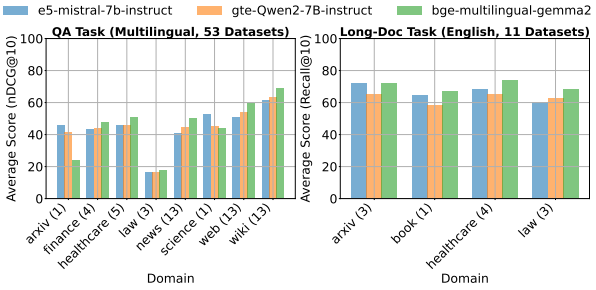

(a) **Domain dimension** comparison results (**English**, *base-size* embedding models).



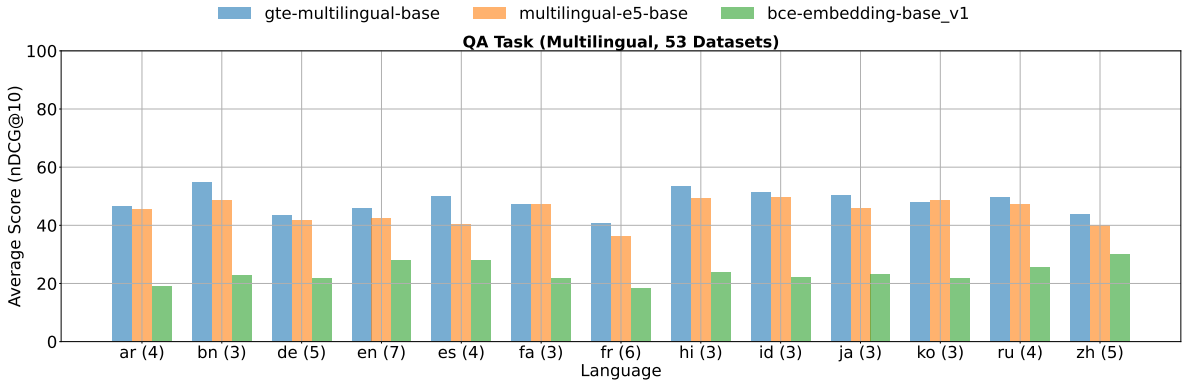

(b) **Domain dimension** comparison results (**Multilingual**, *base-size* embedding models).



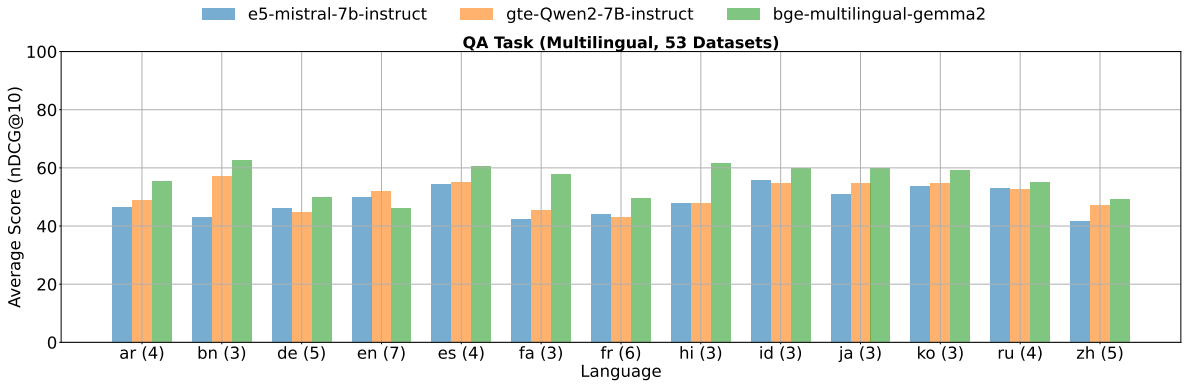

(c) **Domain dimension** comparison results (**English**, *LLM-based* embedding models).




(d) **Domain dimension** comparison results (**Multilingual**, *LLM-based* embedding models).




(e) **Language dimension** comparison results (**Multilingual**, *base-size* embedding models).




(f) **Language dimension** comparison results in **multilingual** datasets (*LLM-based* embedding models).

Figure 6: Additional results indicating that AIR-BENCH can distinguish models in different dimensions. For detailed information of the models appearing in this figure, please refer to Table 15.

| Model | Size | Model Link |
|---|---|---|
| *Lexical Method* | | |
| BM25 (Robertson and Zaragoza, 2009) | - | https://github.com/castorini/pyserini |
| *English Embedding Models* | | |
| bge-small-en-v1.5 (Xiao et al., 2024) | 33.4M | https://huggingface.co/BAAI/bge-small-en-v1.5 |
| bge-base-en-v1.5 (Xiao et al., 2024) | 109M | https://huggingface.co/BAAI/bge-base-en-v1.5 |
| bge-large-en-v1.5 (Xiao et al., 2024) | 335M | https://huggingface.co/BAAI/bge-large-en-v1.5 |
| bge-en-icl (Li et al., 2024) | 7.11B | https://huggingface.co/BAAI/bge-en-icl |
| bge-en-icl-e5data (Li et al., 2024) | 7.11B | https://huggingface.co/BAAI/bge-en-icl-e5data |
| e5-small-v2 (Wang et al., 2022b) | 33.4M | https://huggingface.co/intfloat/e5-small-v2 |
| e5-base-v2 (Wang et al., 2022b) | 109M | https://huggingface.co/intfloat/e5-base-v2 |
| e5-large-v2 (Wang et al., 2022b) | 335M | https://huggingface.co/intfloat/e5-large-v2 |
| gte-small (Li et al., 2023b) | 33.4M | https://huggingface.co/thenlper/gte-small |
| gte-base (Li et al., 2023b) | 109M | https://huggingface.co/thenlper/gte-base |
| gte-large (Li et al., 2023b) | 335M | https://huggingface.co/thenlper/gte-large |
| gte-large-en-v1.5 (Li et al., 2023b) | 434M | https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5 |
| jina-embeddings-v2-base-en (Günther et al., 2023) | 137M | https://huggingface.co/jinaai/jina-embeddings-v2-base-en |
| jina-embeddings-v3 (Sturua et al., 2024) | 572M | https://huggingface.co/jinaai/jina-embeddings-v3 |
| repllama-v1-7b-lora-passage (Ma et al., 2023) | 6.74B | https://huggingface.co/castorini/repllama-v1-7b-lora-passage |
| SFR-Embedding-Mistral | 7.11B | https://huggingface.co/Salesforce/SFR-Embedding-Mistral |
| SFR-Embedding-2_R | 7.11B | https://huggingface.co/Salesforce/SFR-Embedding-2_R |
| NV-Embed-v1 (Lee et al., 2024a) | 7.85B | https://huggingface.co/nvidia/NV-Embed-v1 |
| NV-Embed-v2 (Lee et al., 2024a) | 7.85B | https://huggingface.co/nvidia/NV-Embed-v2 |
| Linq-Embed-Mistral (Kim et al., 2024) | 7.11B | https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral |
| simlm-base-msmarco-finetuned (Wang et al., 2022a) | 110M | https://huggingface.co/intfloat/simlm-base-msmarco-finetuned |
| msmarco-roberta-base-ance-firstp (Xiong et al., 2021) | 125M | https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firstp |
| contriever-msmarco (Izacard et al., 2022) | 109M | https://huggingface.co/facebook/contriever-msmarco |
| *Multilingual Embedding Models* | | |
| bge-m3 (Chen et al., 2024b) | 568M | https://huggingface.co/BAAI/bge-m3 |
| bge-multilingual-gemma2 (Li et al., 2024) | 9.24B | https://huggingface.co/BAAI/bge-multilingual-gemma2 |
| e5-mistral-7b-instruct (Wang et al., 2023) | 7.11B | https://huggingface.co/intfloat/e5-mistral-7b-instruct |
| multilingual-e5-small (Wang et al., 2024) | 118M | https://huggingface.co/intfloat/multilingual-e5-small |
| multilingual-e5-base (Wang et al., 2024) | 278M | https://huggingface.co/intfloat/multilingual-e5-base |
| multilingual-e5-large (Wang et al., 2024) | 560M | https://huggingface.co/intfloat/multilingual-e5-large |
| multilingual-e5-large-instruct (Wang et al., 2024) | 560M | https://huggingface.co/intfloat/multilingual-e5-large-instruct |
| gte-multilingual-base (Zhang et al., 2024) | 305M | https://huggingface.co/Alibaba-NLP/gte-multilingual-base |
| bce-embedding-base_v1 (NetEase Youdao, 2023) | 278M | https://huggingface.co/maidalun1020/bce-embedding-base_v1 |
| gte-Qwen2-1.5B-instruct (Li et al., 2023b) | 1.78B | https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct |
| gte-Qwen2-7B-instruct (Li et al., 2023b) | 7.61B | https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct |
| *Re-ranking Models* | | |
| bge-reranker-large (Xiao et al., 2024) | 560M | https://huggingface.co/BAAI/bge-reranker-large |
| bge-reranker-v2-m3 | 568M | https://huggingface.co/BAAI/bge-reranker-v2-m3 |
| bge-reranker-v2-gemma | 2.51B | https://huggingface.co/BAAI/bge-reranker-v2-gemma |
| bce-reranker-base_v1 (NetEase Youdao, 2023) | 278M | https://huggingface.co/maidalun1020/bce-reranker-base_v1 |
| mmarco-mMiniLMv2-L12-H384-v1 | 118M | https://huggingface.co/nreimers/mmarco-mMiniLMv2-L12-H384-v1 |

Table 15: Detailed information on all of the models appearing in our paper.

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| qa | arxiv | English | en | default | long-summarization (Cohan et al., 2018) | Apache 2.0 | 222,877 | 334 | test | 1,731 | 19 | 5,340 | 6,288 |
| | finance | English | en | default | Reuters-21578 (Lewis, 1997) | CC BY 4.0 | 26,226 | 202 | test | 1,585 | 17 | 3,357 | 5,595 |
| | | Chinese | zh | default | Duxiaoman-DI/FinCorpus | Apache 2.0 | 2,398,095 | 1,616 | test | 1,805 | 29 | 7,836 | 7,211 |
| | healthcare | English | en | default | PubMedQA (Jin et al., 2019) | MIT | 847,395 | 103 | test | 1,707 | 19 | 5,052 | 7,023 |
| | | Chinese | zh | default | Huatuo-26M (Li et al., 2023a) | Apache 2.0 | 360,218 | 751 | test | 1,874 | 31 | 10,029 | 7,336 |
| | law | English | en | default | Pile of Law (Henderson* et al., 2022) | CC BY-NC-SA 4.0 | 141,678 | 1,509 | test | 1,801 | 19 | 5,372 | 6,574 |
| | news | English | en | default | CC-News (Hamborg et al., 2017) | Unknown | 574,417 | 531 | test | 1,614 | 16 | 5,798 | 6,784 |
| | | Chinese | zh | default | intfloat/multilingual_cc_news | Unknown | 935,162 | 1,263 | test | 1,697 | 31 | 7,381 | 6,618 |
| | web | English | en | default | mC4 (Raffel et al., 2020) | ODC-BY | 2,459,587 | 840 | test | 1,707 | 16 | 5,543 | 7,439 |
| | | Chinese | zh | default | mC4 (Raffel et al., 2020) | ODC-BY | 956,699 | 1,208 | test | 1,683 | 29 | 6,250 | 6,721 |
| | wiki | English | en | default | Wikipedia 20240101 | CC BY-SA 3.0, GFDL | 6,738,498 | 667 | test | 1,727 | 17 | 4,260 | 7,882 |
| | | Chinese | zh | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,161,226 | 557 | test | 1,679 | 30 | 4,745 | 6,963 |
| | web (msmarco) | English | en | default | MS MARCO (Bajaj et al., 2016) | MIT | 8,872,840 | 81 | test | 6,319 | 16 | 31,447 | 26,828 |
| long-doc | arxiv | English | en | gemini | Paper of Gemini | CC BY 4.0 | 276 | 136 | test | 249 | 18 | 249 | 0 |
| | | | | gpt3 | Paper of GPT-3 | arXiv.org perpetual, non-exclusive license 1.0 | 515 | 137 | test | 337 | 16 | 496 | 0 |
| | | | | llama2 | Paper of Llama 2 | arXiv.org perpetual, non-exclusive license 1.0 | 566 | 136 | test | 326 | 18 | 635 | 0 |
| | | | | llm-survey | Survey of LLM | arXiv.org perpetual, non-exclusive license 1.0 | 1,144 | 135 | test | 357 | 17 | 924 | 0 |
| | book | English | en | a-brief-history-of-time_stephen-hawking | *A Brief History of Time* | Unknown | 778 | 127 | test | 370 | 16 | 876 | 0 |
| | | | | origin-of-species_darwin | *On the Origin of Species* | Unknown | 1,758 | 126 | test | 366 | 16 | 1,145 | 0 |
| | healthcare | English | en | pubmed_100K-200K_1 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 899 | 133 | test | 372 | 20 | 1,008 | 0 |
| | | | | pubmed_100K-200K_2 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 872 | 136 | test | 355 | 18 | 980 | 0 |
| | | | | pubmed_100K-200K_3 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 873 | 133 | test | 357 | 19 | 978 | 0 |
| | | | | pubmed_30K-40K_10-merged | long-summarization (Cohan et al., 2018) | Apache 2.0 | 2,154 | 133 | test | 368 | 18 | 1,485 | 0 |
| | | | | pubmed_40K-50K_5-merged | long-summarization (Cohan et al., 2018) | Apache 2.0 | 1,731 | 136 | test | 336 | 21 | 1,046 | 0 |
| | law | English | en | lex_files_300K-400K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 2,797 | 137 | test | 339 | 15 | 1,307 | 0 |
| | | | | lex_files_400K-500K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 3,320 | 137 | test | 333 | 17 | 1,427 | 0 |
| | | | | lex_files_500K-600K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 4,087 | 136 | test | 346 | 17 | 1,324 | 0 |
| | | | | lex_files_600K-700K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 5,049 | 138 | test | 338 | 18 | 1,442 | 0 |

Table 16: Statistics of all datasets in AIR-BENCH 24.04.

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| qa | arxiv | English | en | default | long-summarization (Cohan et al., 2018) | Apache 2.0 | 222,877 | 334 | dev | 346 | 19 | 1,091 | 1,230 |
| | | | | | | | | | test | 1,385 | 19 | 4,249 | 5,058 |
| | finance | English | en | default | Reuters-21578 (Lewis, 1997) | CC BY 4.0 | 26,226 | 202 | dev | 317 | 17 | 627 | 1,122 |
| | | | | | | | | | test | 1,268 | 17 | 2,730 | 4,473 |
| | | Arabic | ar | default | asas-ai/financial_news | Apache 2.0 | 11,235 | 397 | dev | 293 | 49 | 635 | 727 |
| | | | | | | | | | test | 1,175 | 46 | 2,796 | 2,959 |
| | | French | fr | default | CoFiF (Daudert and Ahmadi, 2019) | CC BY-NC 4.0 | 1,006,801 | 92 | dev | 310 | 21 | 1,841 | 1,071 |
| | | | | | | | | | test | 1,243 | 20 | 7,206 | 4,362 |
| | | Chinese | zh | default | Duxiaoman-DI/FinCorpus | Apache 2.0 | 1,014,974 | 1,613 | dev | 361 | 29 | 1,516 | 1,471 |
| | | | | | | | | | test | 1,444 | 29 | 6,320 | 5,740 |
| | healthcare | English | en | default | PubMedQA (Jin et al., 2019) | MIT | 847,395 | 103 | dev | 341 | 20 | 1,008 | 1,382 |
| | | | | | | | | | test | 1,366 | 19 | 4,044 | 5,641 |
| | | German | de | default | MLSUM (Scialom et al., 2020) | MIT | 27,934 | 909 | dev | 360 | 21 | 1,102 | 1,137 |
| | | | | | | | | | test | 1,441 | 20 | 4,667 | 4,306 |
| | | Spanish | es | default | Multilingual Medical Corpora (Villena, 2019) | Unknown | 1,006,093 | 60 | dev | 300 | 21 | 1,210 | 930 |
| | | | | | | | | | test | 1,201 | 22 | 4,695 | 3,809 |
| | | French | fr | default | Multilingual Medical Corpora (Villena, 2019) | Unknown | 972,938 | 202 | dev | 331 | 23 | 1,885 | 1,261 |
| | | | | | | | | | test | 1,326 | 24 | 7,460 | 5,119 |
| | | Chinese | zh | default | Huatuo-26M (Li et al., 2023a) | Apache 2.0 | 360,218 | 751 | dev | 374 | 31 | 2,030 | 1,490 |
| | | | | | | | | | test | 1,500 | 31 | 7,999 | 5,846 |
| | law | English | en | default | Pile of Law (Henderson* et al., 2022) | CC BY-NC-SA 4.0 | 141,678 | 1,509 | dev | 360 | 20 | 1,080 | 1,341 |
| | | | | | | | | | test | 1,441 | 19 | 4,292 | 5,233 |
| | | German | de | default | MultiLegalPile (Niklaus et al., 2023) | CC BY-NC-SA 4.0 | 752,913 | 3,361 | dev | 345 | 24 | 1,373 | 1,099 |
| | | | | | | | | | test | 1,382 | 25 | 5,500 | 4,622 |
| | | French | fr | default | MultiLegalPile (Niklaus et al., 2023) | CC BY-NC-SA 4.0 | 649,017 | 2,540 | dev | 348 | 23 | 1,371 | 1,260 |
| | | | | | | | | | test | 1,394 | 22 | 5,535 | 4,968 |
| | science | Russian | ru | default | mlsa-iai-msu-lab/ru_sci_bench | MIT | 200,532 | 347 | dev | 345 | 34 | 1,577 | 1,160 |
| | | | | | | | | | test | 1,382 | 33 | 6,018 | 4,655 |
| | web (msmarco) | English | en | default | MS MARCO (Bajaj et al., 2016) | MIT | 8,872,840 | 81 | dev | 6,319 | 16 | 31,447 | 26,828 |

Table 17: Statistics of all datasets in AIR-BENCH 24.05 (Part 1).

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| qa | news | English | en | default | CC-News (Hamborg et al., 2017) | Unknown | 574,417 | 531 | dev | 322 | 16 | 1,206 | 1,375 |
| | | | | | | | | | test | 1,292 | 16 | 4,592 | 5,409 |
| | | Arabic | ar | default | intfloat/multilingual_cc_news | Unknown | 1,006,308 | 992 | dev | 349 | 42 | 1,810 | 1,307 |
| | | | | | | | | | test | 1,396 | 43 | 7,169 | 5,266 |
| | | Bengali | bn | default | intfloat/multilingual_cc_news | Unknown | 20,681 | 912 | dev | 289 | 84 | 562 | 741 |
| | | | | | | | | | test | 1,159 | 78 | 2,269 | 3,013 |
| | | German | de | default | intfloat/multilingual_cc_news | Unknown | 1,006,876 | 659 | dev | 336 | 23 | 1,448 | 1,234 |
| | | | | | | | | | test | 1,348 | 23 | 5,990 | 5,176 |
| | | Spanish | es | default | intfloat/multilingual_cc_news | Unknown | 1,007,104 | 615 | dev | 337 | 23 | 1,541 | 1,240 |
| | | | | | | | | | test | 1,351 | 23 | 6,246 | 5,257 |
| | | Persian | fa | default | intfloat/multilingual_cc_news | Unknown | 1,002,797 | 1,351 | dev | 346 | 50 | 1,952 | 1,341 |
| | | | | | | | | | test | 1,386 | 48 | 7,885 | 5,353 |
| | | French | fr | default | intfloat/multilingual_cc_news | Unknown | 1,007,592 | 641 | dev | 345 | 23 | 1,548 | 1,424 |
| | | | | | | | | | test | 1,383 | 22 | 6,224 | 5,594 |
| | | Hindi | hi | default | intfloat/multilingual_cc_news | Unknown | 1,006,218 | 1,465 | dev | 349 | 66 | 1,716 | 1,264 |
| | | | | | | | | | test | 1,398 | 67 | 7,039 | 5,162 |
| | | Indonesian | id | default | intfloat/multilingual_cc_news | Unknown | 1,007,724 | 548 | dev | 338 | 24 | 1,799 | 1,397 |
| | | | | | | | | | test | 1,356 | 24 | 7,485 | 5,618 |
| | | Japanese | ja | default | intfloat/multilingual_cc_news | Unknown | 834,364 | 1,559 | dev | 344 | 35 | 1,817 | 1,334 |
| | | | | | | | | | test | 1,378 | 36 | 6,590 | 5,197 |
| | | Korean | ko | default | intfloat/multilingual_cc_news | Unknown | 1,006,798 | 1,072 | dev | 361 | 34 | 1,967 | 1,413 |
| | | | | | | | | | test | 1,447 | 36 | 7,908 | 5,665 |
| | | Russian | ru | default | intfloat/multilingual_cc_news | Unknown | 1,004,550 | 776 | dev | 337 | 34 | 1,676 | 1,301 |
| | | | | | | | | | test | 1,352 | 33 | 6,689 | 5,158 |
| | | Chinese | zh | default | intfloat/multilingual_cc_news | Unknown | 935,162 | 1,263 | dev | 339 | 32 | 1,477 | 1,354 |
| | | | | | | | | | test | 1,358 | 30 | 5,904 | 5,264 |

Table 18: Statistics of all datasets in AIR-BENCH 24.05 (Part 2).

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| qa | web | English | en | default | mC4 (Raffel et al., 2020) | ODC-BY | 1,012,910 | 838 | dev | 341 | 16 | 1,087 | 1,511 |
| | | | | | | | | | test | 1,366 | 16 | 4,456 | 5,928 |
| | | Arabic | ar | default | mC4 (Raffel et al., 2020) | ODC-BY | 165,902 | 1,686 | dev | 334 | 42 | 1,133 | 1,119 |
| | | | | | | | | | test | 1,338 | 42 | 4,782 | 4,717 |
| | | Bengali | bn | default | mC4 (Raffel et al., 2020) | ODC-BY | 45,375 | 2,161 | dev | 362 | 73 | 1,142 | 1,073 |
| | | | | | | | | | test | 1,451 | 77 | 4,759 | 4,449 |
| | | German | de | default | mC4 (Raffel et al., 2020) | ODC-BY | 441,182 | 1,025 | dev | 357 | 20 | 1,320 | 1,345 |
| | | | | | | | | | test | 1,432 | 20 | 5,539 | 5,253 |
| | | Spanish | es | default | mC4 (Raffel et al., 2020) | ODC-BY | 403,020 | 912 | dev | 341 | 23 | 1,317 | 1,281 |
| | | | | | | | | | test | 1,368 | 24 | 5,317 | 5,302 |
| | | Persian | fa | default | mC4 (Raffel et al., 2020) | ODC-BY | 181,463 | 2,114 | dev | 338 | 49 | 1,389 | 1,160 |
| | | | | | | | | | test | 1,354 | 47 | 5,532 | 4,839 |
| | | French | fr | default | mC4 (Raffel et al., 2020) | ODC-BY | 387,210 | 1,076 | dev | 364 | 20 | 1,444 | 1,451 |
| | | | | | | | | | test | 1,457 | 20 | 5,572 | 5,552 |
| | | Hindi | hi | default | mC4 (Raffel et al., 2020) | ODC-BY | 50,501 | 2,396 | dev | 355 | 68 | 1,180 | 1,180 |
| | | | | | | | | | test | 1,423 | 64 | 4,706 | 4,481 |
| | | Indonesian | id | default | mC4 (Raffel et al., 2020) | ODC-BY | 245,878 | 1,059 | dev | 339 | 23 | 1,395 | 1,295 |
| | | | | | | | | | test | 1,356 | 23 | 5,605 | 5,202 |
| | | Japanese | ja | default | mC4 (Raffel et al., 2020) | ODC-BY | 547,419 | 1,026 | dev | 323 | 35 | 1,106 | 1,253 |
| | | | | | | | | | test | 1,293 | 36 | 4,473 | 4,976 |
| | | Korean | ko | default | mC4 (Raffel et al., 2020) | ODC-BY | 250,605 | 1,137 | dev | 327 | 34 | 1,156 | 1,083 |
| | | | | | | | | | test | 1,309 | 36 | 4,239 | 4,457 |
| | | Russian | ru | default | mC4 (Raffel et al., 2020) | ODC-BY | 490,581 | 1,266 | dev | 324 | 32 | 1,330 | 1,277 |
| | | | | | | | | | test | 1,297 | 33 | 5,096 | 5,152 |
| | | Chinese | zh | default | mC4 (Raffel et al., 2020) | ODC-BY | 956,699 | 1,208 | dev | 336 | 30 | 1,230 | 1,366 |
| | | | | | | | | | test | 1,347 | 29 | 5,020 | 5,355 |

Table 19: Statistics of all datasets in AIR-BENCH 24.05 (Part 3).

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| | wiki | English | en | default | Wikipedia 20240101 | CC BY-SA 3.0, GFDL | 1,012,092 | 665 | dev | 345 | 18 | 863 | 1,576 |
| | | | | | | | | | test | 1,382 | 17 | 3,397 | 6,306 |
| | | Arabic | ar | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,232 | 787 | dev | 338 | 40 | 1,112 | 1,438 |
| | | | | | | | | | test | 1,355 | 38 | 4,467 | 5,778 |
| | | Bengali | bn | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 152,064 | 2,129 | dev | 364 | 69 | 1,016 | 1,542 |
| | | | | | | | | | test | 1,456 | 71 | 4,203 | 5,841 |
| | | German | de | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,186 | 891 | dev | 350 | 23 | 817 | 1,481 |
| | | | | | | | | | test | 1,404 | 22 | 3,411 | 5,881 |
| | | Spanish | es | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,147 | 801 | dev | 345 | 22 | 879 | 1,451 |
| | | | | | | | | | test | 1,380 | 23 | 3,531 | 5,767 |
| | | Persian | fa | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 999,223 | 627 | dev | 332 | 41 | 1,179 | 1,444 |
| | | | | | | | | | test | 1,328 | 45 | 4,538 | 5,581 |
| | | French | fr | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,270 | 779 | dev | 356 | 20 | 768 | 1,496 |
| | | | | | | | | | test | 1,424 | 21 | 3,429 | 5,989 |
| | | Hindi | hi | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 162,188 | 997 | dev | 340 | 59 | 995 | 1,324 |
| | | | | | | | | | test | 1,360 | 59 | 3,911 | 5,225 |
| | | Indonesian | id | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 687,513 | 451 | dev | 343 | 22 | 1,003 | 1,365 |
| | | | | | | | | | test | 1,373 | 21 | 4,089 | 5,486 |
| | | Japanese | ja | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,365 | 1,470 | dev | 358 | 30 | 1,099 | 1,537 |
| | | | | | | | | | test | 1,432 | 32 | 4,303 | 6,101 |
| | | Korean | ko | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 665,227 | 775 | dev | 346 | 39 | 1,109 | 1,423 |
| | | | | | | | | | test | 1,384 | 35 | 4,604 | 5,810 |
| | | Russian | ru | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,008,405 | 1,211 | dev | 365 | 30 | 1,154 | 1,505 |
| | | | | | | | | | test | 1,463 | 30 | 4,516 | 6,283 |
| | | Chinese | zh | default | Wikipedia 20240401 | CC BY-SA 3.0, GFDL | 1,011,604 | 557 | dev | 335 | 30 | 952 | 1,301 |
| | | | | | | | | | test | 1,344 | 30 | 3,793 | 5,662 |

Table 20: Statistics of all datasets in AIR-BENCH 24.05 (Part 4).

| Task | Domain | Language | | Dataset Name | Source of Corpus | | #corpus | Avg #token of corpus | Split | # of queries | Avg #token of queries | # of positives | # of hard negatives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Name | ISO Code | | Link & Citation | License | | | | | | | |
| long-doc | arxiv | English | en | gemini | Paper of Gemini | CC BY 4.0 | 276 | 136 | test | 249 | 18 | 249 | 0 |
| | | | | gpt3 | Paper of GPT-3 | arXiv.org perpetual, non-exclusive license 1.0 | 515 | 137 | test | 337 | 16 | 496 | 0 |
| | | | | llama2 | Paper of Llama 2 | arXiv.org perpetual, non-exclusive license 1.0 | 566 | 136 | dev | 326 | 18 | 635 | 0 |
| | | | | llm-survey | Survey of LLM | arXiv.org perpetual, non-exclusive license 1.0 | 1,144 | 135 | test | 357 | 17 | 924 | 0 |
| | book | English | en | a-brief-history-of-time_stephen-hawking | *A Brief History of Time* | Unknown | 778 | 127 | dev | 370 | 16 | 876 | 0 |
| | | | | origin-of-species_darwin | *On the Origin of Species* | Unknown | 1,758 | 126 | test | 366 | 16 | 1,145 | 0 |
| | healthcare | English | en | pubmed_100K-200K_1 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 899 | 133 | test | 372 | 20 | 1,008 | 0 |
| | | | | pubmed_100K-200K_2 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 872 | 136 | test | 355 | 18 | 980 | 0 |
| | | | | pubmed_100K-200K_3 | long-summarization (Cohan et al., 2018) | Apache 2.0 | 873 | 133 | dev | 357 | 19 | 978 | 0 |
| | | | | pubmed_30K-40K_10-merged | long-summarization (Cohan et al., 2018) | Apache 2.0 | 2,154 | 133 | test | 368 | 18 | 1,485 | 0 |
| | | | | pubmed_40K-50K_5-merged | long-summarization (Cohan et al., 2018) | Apache 2.0 | 1,731 | 136 | test | 336 | 21 | 1,046 | 0 |
| | law | English | en | lex_files_300K-400K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 2,797 | 137 | dev | 339 | 15 | 1,307 | 0 |
| | | | | lex_files_400K-500K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 3,320 | 137 | test | 333 | 17 | 1,427 | 0 |
| | | | | lex_files_500K-600K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 4,087 | 136 | test | 346 | 17 | 1,324 | 0 |
| | | | | lex_files_600K-700K | LexFiles (Chalkidis et al., 2023) | CC BY-NC-SA 4.0 | 5,049 | 138 | test | 338 | 18 | 1,442 | 0 |

Table 21: Statistics of all datasets in AIR-BENCH 24.05 (Part 5).

| **Dataset** (↓) | bge-en-icl (zero-shot) | bge-en-icl-e5data (zero-shot) | bge-*-en-v1.5 | | | e5-*-v2 | | | gte-* | | | NV-Embed-* | | gte-large-en-v1.5 | SFR-Embedding-Mistral | SFR-Embedding-2_R | repllama-v1-7b-lora-passage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | small | base | large | small | base | large | small | base | large | v1 | v2 | | | | |
| **QA Task (English, 7 Datasets)** | | | | | | | | | | | | | | | | | |
| arxiv_en | 50.43 | 50.43 | 35.56 | 36.44 | 38.68 | 35.44 | 37.56 | 38.31 | 36.25 | 38.09 | 39.20 | 47.39 | 49.88 | 40.00 | 48.10 | 43.58 | 47.20 |
| finance_en | 55.48 | 55.48 | 44.68 | 47.35 | 45.59 | 49.44 | 47.33 | 51.34 | 47.60 | 49.08 | 47.91 | 51.21 | 54.09 | 54.33 | 58.34 | 56.02 | 54.14 |
| healthcare_en | 57.20 | 57.20 | 47.23 | 49.50 | 50.93 | 46.48 | 48.93 | 52.45 | 47.48 | 49.20 | 49.02 | 58.98 | 59.85 | 48.34 | 59.09 | 55.83 | 56.25 |
| law_en | 28.92 | 28.92 | 17.12 | 16.71 | 23.53 | 15.67 | 16.34 | 21.28 | 15.89 | 15.63 | 17.57 | 20.77 | 25.43 | 18.73 | 22.52 | 21.22 | 17.98 |
| news_en | 54.92 | 54.92 | 41.38 | 40.59 | 43.70 | 43.37 | 42.00 | 45.26 | 40.70 | 43.64 | 43.70 | 50.93 | 52.35 | 45.74 | 51.20 | 50.57 | 51.58 |
| web_en | 60.80 | 60.80 | 47.04 | 46.77 | 48.92 | 45.27 | 46.12 | 48.60 | 49.55 | 50.41 | 51.47 | 56.46 | 58.76 | 50.02 | 57.44 | 55.57 | 58.06 |
| wiki_en | 73.44 | 73.44 | 62.15 | 62.53 | 63.05 | 65.94 | 66.61 | 66.54 | 65.07 | 65.92 | 65.96 | 71.08 | 73.06 | 66.59 | 72.78 | 72.77 | 70.84 |
| **Long-Doc Task (English, 11 Datasets)** | | | | | | | | | | | | | | | | | |
| arxiv_en_gemini | 87.55 | 87.55 | 69.08 | 69.88 | 75.50 | 75.50 | 72.29 | 76.31 | 72.69 | 71.89 | 74.70 | 83.94 | 85.54 | 72.29 | 79.92 | 78.72 | 82.73 |
| arxiv_en_gpt3 | 81.65 | 81.65 | 65.23 | 64.91 | 64.37 | 65.83 | 69.44 | 75.07 | 66.79 | 67.41 | 70.62 | 83.31 | 83.28 | 68.30 | 76.88 | 72.28 | 76.81 |
| arxiv_en_llm_survey | 71.05 | 71.05 | 46.25 | 49.80 | 51.32 | 52.71 | 54.30 | 58.07 | 49.54 | 55.85 | 56.58 | 67.62 | 69.33 | 55.94 | 63.51 | 60.15 | 64.71 |
| book_en_origin_of_species_darwin | 71.07 | 71.07 | 52.50 | 56.09 | 60.22 | 53.03 | 54.07 | 61.51 | 58.37 | 59.76 | 63.32 | 68.21 | 70.77 | 51.35 | 66.18 | 62.49 | 66.64 |
| healthcare_en_pubmed_100k_200k_1 | 71.23 | 71.23 | 57.24 | 60.50 | 61.68 | 58.76 | 62.33 | 64.62 | 56.94 | 57.59 | 58.76 | 68.51 | 69.17 | 57.78 | 63.77 | 65.13 | 68.17 |
| healthcare_en_pubmed_100k_200k_2 | 75.59 | 75.59 | 57.23 | 59.39 | 62.54 | 62.14 | 66.00 | 69.10 | 58.30 | 61.70 | 64.49 | 75.53 | 77.96 | 63.91 | 73.15 | 70.78 | 69.78 |
| healthcare_en_pubmed_30k_40k_10_merged | 79.00 | 79.00 | 65.76 | 67.33 | 69.36 | 68.32 | 69.74 | 74.16 | 67.03 | 68.50 | 68.34 | 78.45 | 79.98 | 64.79 | 74.45 | 71.97 | 77.24 |
| healthcare_en_pubmed_40k_50k_5_merged | 65.57 | 65.57 | 52.03 | 57.15 | 57.28 | 57.94 | 57.83 | 61.35 | 50.79 | 54.53 | 53.57 | 64.89 | 64.57 | 55.14 | 61.95 | 61.19 | 60.55 |
| law_en_lex_files_400k_500k | 67.73 | 67.73 | 54.46 | 58.07 | 60.25 | 53.20 | 58.06 | 62.17 | 55.56 | 59.03 | 62.66 | 66.54 | 66.06 | 58.80 | 61.76 | 60.35 | 64.77 |
| law_en_lex_files_500k_600k | 68.52 | 68.52 | 53.03 | 60.23 | 61.68 | 57.98 | 60.08 | 65.89 | 58.75 | 61.63 | 63.54 | 68.14 | 71.56 | 63.36 | 66.31 | 63.16 | 67.55 |
| law_en_lex_files_600k_700k | 68.73 | 68.73 | 49.98 | 53.58 | 56.30 | 51.79 | 57.61 | 59.56 | 53.84 | 55.29 | 56.53 | 67.76 | 69.68 | 56.18 | 61.18 | 57.90 | 63.97 |

Table 22: Detailed evaluation results of English IR models on QA (English, test) datasets and Long-Doc (English, test) datasets of AIR-BENCH 24.05.

| **Dataset** (↓) | BM25 | BM25 + bge-reranker-v2-m3 | multilingual-e5-* small | multilingual-e5-* base | multilingual-e5-* large | gte-Qwen2-*-instruct 1.5B | gte-Qwen2-*-instruct 7B | bge-m3 | multilingual-e5-large-instruct | e5-mistral-7b-instruct | bge-multilingual-gemma2 | gte-multilingual-base | bce-embedding-base_v1 | jina-embeddings-v3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **QA Task (Multilingual, 53 Datasets)** | | | | | | | | | | | | | | |
| arxiv_en | 33.18 | 50.30 | 32.98 | 34.17 | 37.84 | 42.15 | 41.33 | 41.64 | 39.52 | 46.06 | 24.00 | 41.28 | 22.60 | 39.65 |
| finance_ar | 35.78 | 51.78 | 36.17 | 43.82 | 45.34 | 44.12 | 43.55 | 45.76 | 48.95 | 44.59 | 50.25 | 45.59 | 25.00 | 46.32 |
| finance_en | 45.13 | 58.04 | 47.32 | 50.29 | 49.05 | 55.21 | 59.23 | 52.92 | 52.79 | 55.90 | 50.08 | 53.24 | 41.67 | 51.70 |
| finance_fr | 27.63 | 52.08 | 25.90 | 33.83 | 36.41 | 36.52 | 39.57 | 41.44 | 42.73 | 38.98 | 51.10 | 35.47 | 19.27 | 37.14 |
| finance_zh | 22.43 | 42.35 | 30.46 | 32.07 | 34.71 | 34.28 | 34.61 | 40.23 | 37.72 | 33.10 | 39.23 | 36.84 | 25.72 | 33.96 |
| healthcare_de | 50.02 | 63.43 | 43.90 | 47.34 | 46.14 | 46.34 | 53.91 | 49.00 | 52.06 | 53.12 | 55.40 | 50.68 | 25.55 | 49.86 |
| healthcare_en | 34.84 | 53.76 | 44.21 | 49.16 | 50.63 | 52.11 | 54.46 | 49.05 | 54.02 | 56.24 | 47.48 | 47.48 | 29.89 | 49.42 |
| healthcare_es | 31.25 | 50.85 | 45.67 | 50.30 | 54.91 | 49.49 | 53.78 | 53.05 | 51.74 | 47.67 | 63.13 | 46.35 | 29.90 | 52.75 |
| healthcare_fr | 28.02 | 50.99 | 19.75 | 28.53 | 32.40 | 33.86 | 30.29 | 39.29 | 36.64 | 37.28 | 45.13 | 34.92 | 6.39 | 32.68 |
| healthcare_zh | 18.10 | 43.58 | 28.97 | 28.08 | 33.62 | 39.13 | 38.66 | 42.31 | 39.76 | 36.05 | 42.35 | 37.94 | 25.84 | 38.91 |
| law_de | 12.33 | 22.95 | 11.93 | 13.35 | 13.56 | 12.81 | 13.18 | 20.11 | 15.59 | 14.77 | 15.75 | 12.65 | 5.72 | 11.71 |
| law_en | 19.50 | 34.17 | 14.61 | 15.76 | 19.71 | 20.19 | 22.75 | 26.95 | 16.90 | 19.61 | 22.60 | 11.44 | 8.67 | 16.78 |
| law_fr | 13.16 | 23.19 | 7.34 | 10.30 | 9.94 | 12.72 | 13.15 | 20.20 | 15.12 | 14.38 | 14.29 | 11.68 | 2.64 | 9.76 |
| news_ar | 26.54 | 50.17 | 32.16 | 37.49 | 40.64 | 35.93 | 37.63 | 44.93 | 48.20 | 38.95 | 48.41 | 39.13 | 13.43 | 44.04 |
| news_bn | 29.33 | 41.60 | 44.97 | 46.48 | 52.17 | 20.27 | 61.31 | 59.03 | 49.31 | 25.50 | 58.77 | 56.00 | 17.90 | 53.73 |
| news_de | 38.52 | 55.11 | 39.06 | 43.70 | 43.34 | 43.08 | 44.89 | 47.87 | 47.84 | 46.48 | 52.05 | 43.93 | 21.15 | 46.39 |
| news_en | 39.72 | 57.63 | 38.70 | 43.05 | 43.48 | 47.44 | 52.74 | 47.34 | 44.27 | 47.89 | 50.29 | 47.55 | 30.74 | 45.61 |
| news_es | 33.09 | 54.65 | 36.14 | 38.88 | 40.41 | 39.90 | 45.21 | 44.70 | 45.99 | 45.34 | 49.90 | 40.47 | 19.76 | 42.94 |
| news_fa | 24.95 | 52.02 | 33.07 | 36.70 | 40.03 | 26.40 | 30.09 | 43.81 | 45.59 | 29.72 | 43.40 | 39.05 | 15.79 | 37.90 |
| news_fr | 41.20 | 60.79 | 28.56 | 40.51 | 36.59 | 45.60 | 49.76 | 49.52 | 50.59 | 49.61 | 56.80 | 45.86 | 22.75 | 46.56 |
| news_hi | 31.93 | 54.95 | 32.96 | 32.85 | 39.12 | 23.39 | 30.28 | 42.12 | 39.66 | 29.82 | 44.89 | 36.64 | 14.02 | 40.02 |
| news_id | 42.82 | 66.52 | 35.87 | 41.26 | 41.03 | 34.82 | 46.44 | 47.45 | 48.59 | 45.93 | 50.65 | 41.27 | 19.20 | 44.86 |
| news_ja | 38.12 | 57.95 | 37.42 | 39.06 | 45.24 | 41.95 | 44.13 | 47.09 | 47.60 | 43.47 | 51.51 | 42.62 | 21.44 | 41.96 |
| news_ko | 34.79 | 59.22 | 40.05 | 43.16 | 47.79 | 44.55 | 47.19 | 48.13 | 50.52 | 46.47 | 51.64 | 40.39 | 20.70 | 45.18 |
| news_ru | 31.67 | 55.72 | 37.90 | 42.06 | 43.24 | 43.09 | 46.55 | 48.31 | 48.81 | 46.59 | 51.48 | 42.93 | 20.56 | 46.65 |
| news_zh | 15.22 | 30.61 | 27.34 | 36.24 | 39.67 | 36.43 | 43.17 | 41.00 | 35.46 | 35.98 | 43.42 | 36.20 | 27.56 | 40.56 |
| science_ru | 39.78 | 62.84 | 43.70 | 47.01 | 51.87 | 54.04 | 45.21 | 55.18 | 56.86 | 53.07 | 44.13 | 48.69 | 26.06 | 50.24 |
| web_ar | 39.15 | 60.85 | 41.15 | 46.78 | 47.74 | 48.85 | 55.56 | 52.53 | 56.40 | 49.56 | 59.97 | 47.12 | 18.89 | 53.40 |
| web_bn | 47.47 | 68.73 | 44.65 | 46.71 | 51.10 | 38.37 | 51.45 | 55.53 | 56.17 | 46.83 | 59.68 | 50.89 | 25.03 | 55.55 |
| web_de | 46.14 | 61.30 | 45.06 | 45.90 | 46.89 | 47.73 | 48.62 | 51.89 | 50.87 | 50.88 | 57.72 | 47.22 | 26.31 | 48.06 |
| web_en | 41.46 | 60.51 | 38.71 | 43.24 | 42.81 | 52.68 | 58.99 | 53.88 | 41.58 | 52.08 | 56.48 | 52.05 | 30.55 | 47.38 |
| web_es | 42.52 | 60.89 | 42.57 | 46.04 | 46.44 | 50.69 | 54.11 | 51.78 | 52.24 | 54.45 | 58.20 | 49.56 | 26.77 | 49.42 |
| web_fa | 42.61 | 64.98 | 45.91 | 48.44 | 50.45 | 41.71 | 49.55 | 55.81 | 58.68 | 45.86 | 62.43 | 49.70 | 21.83 | 52.84 |
| web_fr | 46.62 | 63.48 | 30.61 | 43.13 | 39.56 | 51.60 | 55.16 | 51.46 | 50.20 | 54.52 | 59.54 | 50.34 | 26.94 | 48.80 |
| web_hi | 50.70 | 71.06 | 50.53 | 51.50 | 56.44 | 40.53 | 53.06 | 57.06 | 56.32 | 49.43 | 64.50 | 56.30 | 25.22 | 58.79 |
| web_id | 48.80 | 67.23 | 39.52 | 46.37 | 44.80 | 48.32 | 55.51 | 53.14 | 54.49 | 55.17 | 60.00 | 50.50 | 21.02 | 52.76 |
| web_ja | 47.41 | 64.53 | 45.49 | 47.36 | 52.21 | 52.21 | 57.27 | 54.75 | 54.89 | 51.80 | 60.26 | 51.18 | 27.65 | 50.10 |
| web_ko | 44.73 | 61.51 | 45.07 | 46.53 | 53.59 | 52.48 | 57.54 | 55.28 | 55.81 | 54.22 | 59.64 | 47.41 | 23.53 | 51.87 |
| web_ru | 42.92 | 63.59 | 42.85 | 47.59 | 48.51 | 52.35 | 55.88 | 54.53 | 54.97 | 53.85 | 60.12 | 49.77 | 28.24 | 50.51 |
| web_zh | 33.69 | 52.96 | 42.14 | 44.27 | 48.17 | 47.48 | 51.66 | 50.20 | 47.06 | 45.68 | 53.04 | 46.75 | 35.66 | 47.66 |
| wiki_ar | 43.66 | 63.82 | 50.61 | 54.35 | 60.65 | 47.74 | 59.44 | 59.65 | 63.21 | 52.98 | 63.42 | 54.40 | 19.38 | 57.89 |
| wiki_bn | 55.80 | 72.97 | 53.57 | 53.13 | 60.33 | 51.35 | 58.17 | 64.33 | 64.45 | 56.84 | 69.48 | 58.12 | 25.81 | 62.81 |
| wiki_de | 61.20 | 73.32 | 56.58 | 57.89 | 59.70 | 56.30 | 63.97 | 64.68 | 65.81 | 65.40 | 67.91 | 62.83 | 30.17 | 62.08 |
| wiki_en | 60.27 | 75.46 | 61.89 | 62.78 | 63.85 | 66.45 | 73.59 | 69.70 | 68.62 | 71.38 | 72.80 | 69.12 | 30.97 | 64.96 |
| wiki_es | 57.24 | 73.70 | 59.53 | 59.41 | 61.61 | 60.94 | 67.62 | 65.40 | 68.10 | 69.49 | 71.79 | 63.42 | 34.99 | 63.65 |
| wiki_fa | 48.02 | 67.43 | 54.07 | 56.47 | 59.69 | 44.29 | 57.05 | 61.15 | 64.20 | 51.77 | 67.57 | 53.24 | 27.63 | 57.75 |
| wiki_fr | 62.71 | 76.51 | 50.94 | 59.04 | 60.71 | 61.90 | 70.32 | 66.04 | 69.72 | 69.29 | 71.28 | 66.69 | 33.14 | 64.67 |
| wiki_hi | 57.81 | 74.76 | 62.73 | 63.59 | 68.59 | 51.57 | 60.54 | 69.02 | 71.81 | 63.93 | 75.39 | 67.62 | 32.02 | 68.74 |
| wiki_id | 58.14 | 75.16 | 59.00 | 60.95 | 61.82 | 54.47 | 61.81 | 66.30 | 66.36 | 66.23 | 68.91 | 62.79 | 25.92 | 62.75 |
| wiki_ja | 56.43 | 72.90 | 54.32 | 51.31 | 61.07 | 55.97 | 62.88 | 60.86 | 64.12 | 57.72 | 68.29 | 57.62 | 20.26 | 58.26 |
| wiki_ko | 43.93 | 67.17 | 55.75 | 56.26 | 62.64 | 54.89 | 59.17 | 62.36 | 64.79 | 60.30 | 66.78 | 55.63 | 20.96 | 58.28 |
| wiki_ru | 53.99 | 68.60 | 53.80 | 52.96 | 58.16 | 53.45 | 62.95 | 60.18 | 62.57 | 58.70 | 64.15 | 57.03 | 28.08 | 59.41 |
| wiki_zh | 40.24 | 63.51 | 53.63 | 59.44 | 61.83 | 58.33 | 67.50 | 63.52 | 62.82 | 57.19 | 68.64 | 61.86 | 35.46 | 62.70 |
| **Long-Doc Task (English, 11 Datasets)** | | | | | | | | | | | | | | |
| arxiv_en_gemini | 63.85 | 82.33 | 75.10 | 74.70 | 76.71 | 75.10 | 73.09 | 82.33 | 76.71 | 77.51 | 81.53 | 74.70 | 71.89 | 72.69 |
| arxiv_en_gpt3 | 56.13 | 74.56 | 67.21 | 70.23 | 73.71 | 73.39 | 71.61 | 71.93 | 69.12 | 76.85 | 75.12 | 72.13 | 69.98 | 71.39 |
| arxiv_en_llm_survey | 47.76 | 68.77 | 54.11 | 58.05 | 60.29 | 53.63 | 50.33 | 61.25 | 58.87 | 62.28 | 59.65 | 57.89 | 52.85 | 55.96 |
| book_en_origin_of_species_darwin | 42.07 | 65.42 | 50.12 | 55.93 | 59.39 | 63.02 | 58.39 | 59.41 | 61.94 | 64.50 | 67.08 | 57.78 | 48.53 | 62.20 |
| healthcare_en_pubmed_100k_200k_1 | 60.21 | 78.17 | 58.56 | 63.94 | 63.47 | 60.54 | 62.06 | 65.64 | 62.97 | 64.40 | 71.48 | 64.40 | 48.08 | 58.33 |
| healthcare_en_pubmed_100k_200k_2 | 61.78 | 82.29 | 59.41 | 61.79 | 63.94 | 66.71 | 69.05 | 67.31 | 64.42 | 71.40 | 79.21 | 68.32 | 50.88 | 57.63 |
| healthcare_en_pubmed_30k_40k_10_merged | 65.45 | 84.12 | 66.44 | 70.90 | 70.36 | 70.47 | 70.75 | 70.98 | 72.13 | 74.65 | 79.78 | 73.03 | 58.36 | 67.06 |
| healthcare_en_pubmed_40k_50k_5_merged | 53.90 | 72.44 | 55.84 | 57.31 | 60.13 | 56.65 | 59.22 | 56.45 | 59.07 | 62.91 | 66.72 | 60.10 | 45.09 | 53.67 |
| law_en_lex_files_400k_500k | 42.75 | 68.51 | 51.85 | 58.70 | 64.56 | 58.70 | 61.84 | 63.59 | 59.82 | 58.61 | 66.14 | 59.56 | 47.08 | 60.64 |
| law_en_lex_files_500k_600k | 42.99 | 67.93 | 56.03 | 61.41 | 67.72 | 60.78 | 64.33 | 64.12 | 60.90 | 62.73 | 69.83 | 64.56 | 49.92 | 60.82 |
| law_en_lex_files_600k_700k | 47.12 | 71.70 | 52.92 | 57.33 | 62.50 | 58.61 | 63.04 | 60.60 | 57.67 | 60.15 | 69.03 | 60.17 | 42.74 | 56.06 |

Table 23: Detailed evaluation results of multilingual IR models on QA (Multilingual, test) datasets and Long-Doc (English, test) datasets of AIR-BENCH 24.05.