MAIR: A Massive Benchmark for Evaluating Instructed Retrieval

Weiwei Sun[1] **Zhengliang Shi**[2] **Jiulong Wu**[3] **Lingyong Yan**[4]
Xinyu Ma[4] **Yiding Liu**[4] **Min Cao**[3] **Dawei Yin**[4*] **Zhaochun Ren**[5*]
[1]Carnegie Mellon University [2]Shandong University
[3]Soochow University [4]Baidu Inc. [5]Leiden University
{sunnweiwei,zhengliang.shii,lingyongy,xinyuma2016}@gmail.com
yindawei@acm.org, z.ren@liacs.leidenuniv.nl

Abstract

Recent information retrieval (IR) models are pre-trained and instruction-tuned on massive datasets and tasks, enabling them to perform well on a wide range of tasks and potentially generalize to unseen tasks with instructions. However, existing IR benchmarks focus on a limited scope of tasks, making them insufficient for evaluating the latest IR models. In this paper, we propose MAIR (Massive Instructed Retrieval Benchmark), a heterogeneous IR benchmark that includes 126 distinct IR tasks across 6 domains, collected from existing datasets. We benchmark state-of-the-art instruction-tuned text embedding models and re-ranking models. Our experiments reveal that instruction-tuned models generally achieve superior performance compared to non-instruction-tuned models on MAIR. Additionally, our results suggest that current instruction-tuned text embedding models and re-ranking models still lack effectiveness in specific long-tail tasks. MAIR is publicly available at https://github.com/sunnweiwei/Mair.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities in performing a wide range of natural language processing (NLP) tasks by being first pre-trained on large-scale corpora and then instruction-tuned on numerous downstream tasks (Chung et al., 2024; Wang et al., 2023b, 2022b; Wei et al., 2022). This advancement has garnered significant attention in other fields, including Information Retrieval (IR) (Gao and Callan, 2021; Neelakantan et al., 2022; Wang et al., 2022a; Izacard et al., 2021; Wang et al., 2023a; Asai et al., 2022; Su et al., 2022).

IR techniques aim to retrieve a set of relevant candidates from a large corpus based on semantic relevance or other query-specific criteria (Yates et al., 2021; Fan et al., 2022). These techniques are critical components of many AI applications, from web search (Zhu et al., 2023) to various retrieval-augmented tasks (Gao et al., 2023; Shi et al., 2024b). However, most traditional IR models, typically trained on a single task, often exhibit poor generalization to other IR tasks or domains (Thakur et al., 2021; Zhao et al., 2024). Inspired by the success of LLMs, recent research explores training models for the general-purpose IR (Sachan et al., 2022; Sun et al., 2023; Oh et al., 2024; Weller et al., 2024). These models, instruction-tuned on multiple retrieval tasks, show significant improvements in aligning with the user intent across different IR tasks.

To evaluate the generalization capabilities of newly emerged IR models, several benchmarks such as BEIR (Thakur et al., 2021), KILT (Petroni et al., 2020), and MTEB (Muennighoff et al., 2022) have been established recently, compiling a variety of IR tasks. However, as shown in Table 1, these benchmarks either (i) contain a relatively small number of tasks or (ii) feature tasks that are too similar, thus providing limited coverage of the broad IR landscape. In contrast, instruction-tuned LLMs have been evaluated on hundreds or thousands of diverse NLP tasks (Wang et al., 2022b; Chung et al., 2024).

To fill this gap, this paper introduces MAIR (Massive Instructed Retrieval Benchmark), a large-scale IR benchmark consisting of 126 diverse retrieval tasks with 805 distinct instructions to evaluate model generalization on unseen tasks. In MAIR, we collect tasks from existing IR datasets, such as those found in (i) SIGIR resource track papers, (ii) tasks in existing benchmarks, (iii) publicly accessible shared tasks in TREC (trec.nist.gov), and (iv) recent LLM benchmarks. Tasks in MAIR are elaborately selected to cover various information retrieval requirements in practice, including diverse types of queries and document types, as well as specific relevance criteria. As a result, a

*Corresponding authors

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14044–14067
November 12-16, 2024

	MAIR (this work)	BEIR Thakur et al. (2021)	KILT Petroni et al. (2020)	FollowIR Weller et al. (2024)	InstructIR Oh et al. (2024)
Number of tasks	126	18	11	3	1
Number of domains	6	4	1	1	1
Number of instructions	805	14	5	104	9,906
Number of test queries	10,038	47,233	50,736	104	9,906
Number of collections	426	18	1	3	1
Total number of docs	4,274,916	38,506,129	5,903,530	98,312	16,072

Table 1: Data statistics of MAIR and other relevant IR benchmarks.

total of 126 unique tasks are selected to constitute the benchmark MAIR.

Subsequently, we clean and merge the data of these tasks, and then sample them, resulting in a dataset that includes 10,038 queries and 4,274,916 documents. We perform the sampling mainly due to the data distribution imbalance. We validate that the sampled benchmark produces results highly correlated with full-scale testing, achieving a balance between evaluation accuracy and cost.

Finally, we manually annotate retrieval instructions based on the queries and corresponding documents for these IR tasks. Following previous work (Weller et al., 2024), these instructions specify the types of queries, documents, and relevance criteria. We ultimately annotate 805 distinct instructions, representing 805 different query-document-relevance combinations. Some tasks include query-level instructions, meaning each query has a different relevance criterion. These annotations make MAIR particularly challenging and suitable for evaluating instruction-tuned retrievers in terms of their ability to follow instructions in completing unseen tasks.

Based on MAIR, we benchmark various different types of retrieval models, including (i) sparse retriever, (ii) single-task text embedding models (Ni et al., 2021; Izacard et al., 2021), (iii) non-instruction-tuned multi-task text embedding models (Li et al., 2023b; Xiao et al., 2023; Wang et al., 2022a), (iv) instruction-tuned embedding models (Wang et al., 2023a; Lee et al., 2024), and (v) re-ranking models (Sun et al., 2023). We evaluate on both *no instruction* and *with instruction*, and found instruction-tuned embedding models show clear improvement when instructions are added. `GritLM-7B` achieves the best overall score, with an average nDCG@10 of 55.20. Furthermore, both `e5-mistral-7b-instruct` and `GritLM-7B` show notable improvement when instructions are added.

2 Data Construction

This section illustrates the process of data collection and benchmark construction. As aforementioned, the benchmark for instruction-tuned IR models should be capable of evaluating IR baselines on a variety of tasks across different domains and using as realistic instruction as possible. To this end, we construct our benchmark based on the following criteria: **(a) Task and domain variety**: To assess the generalization of different IR baselines on various tasks from different domains, we collect data from a large range of domains and tasks. Specifically, the data is collected from 126 distinct IR tasks across 6 domains. Each task is manually filtered to avoid duplication. **(b) Instruction diversity**: For each task, we annotate and review a large number of detailed instructions. These instructions can assist in thoroughly evaluating models' instruction-following capabilities when searching queries.

2.1 Data Collection

To build a comprehensive IR benchmark for instruction-following evaluation, we collect data from the following four well-known sources:

- **Existing IR Resources**: The specialized resource tracks in some information retrieval conferences (e.g. SIGIR). Specifically, we collect released papers from 2021-2024 SIGIR conferences and finally collect 10 tasks across 2 domains.
- **Other IR Benchmarks**: We leverage tasks from existing IR benchmarks such as BEIR (Thakur et al., 2021) and KILT (Petroni et al., 2020), as well as domain-specific benchmarks for evaluating Voyage embedding[1]. These benchmarks have gained attention in the IR community and provide diverse tasks

[1] `https://blog.voyageai.com/`

tailored to specific applications. Finally, We collect 55 tasks across 6 domains from the IR benchmarks.

- **TREC Tracks**: The *Text REtrieval Conference* (TREC) is a long-standing and well-established IR conference series organized by the National Institute of Standards and Technology (NIST). TREC provides unique and realistic use cases along with rigorously annotated data. Finally, We collect 34 tasks and across 5 domains from TREC tracks.

- **LLM Evaluation Datasets**: In addition to the above datasets, we also integrate several public LLM evaluation datasets released in the LLM era. Including those datasets can help extending our benchmark to diverse potential instruction-following scenarios. Finally, We collect 27 tasks across 4 domains from the LLM Evaluation Datasets.

Finally, we collect 126 tasks. There is a simple process for us to collect data: We first review the documentation of various conference tracks, merging tasks with identical corpus and settings to avoid duplication. Then, based on task requirements, we download publicly available datasets from official channels. For tasks with incomplete corpus, we use web crawling and other techniques to supplement the data as much as possible. Since the data sources are diverse and the original formats vary substantially, we perform necessary data cleaning operations such as deduplication, keyword extraction, and text normalization on the data. Thus, we unify these task data into a consistent format.

2.2 Data Sampling

After filtering out 126 tasks, we find the data distributions are quite imbalanced among different tasks. Besides, since the scales of some datasets are extremely large with amounts of redundant content, model evaluation of the whole dataset is inefficient and unnecessary. To alleviate the influence of task data imbalance and improve evaluation efficiency, we lightweight our benchmark while preserving its evaluative capability via effective data sampling.

Specifically, for each task, we reduce its sample size following the following two steps:

- **Query Sampling**: First, for all search queries in the task, we perform the K-means algorithm over query embeddings to cluster the queries. And we randomly sample one single query in each cluster to ensure query diversity while avoiding redundancy. And we set the cluster number as 100 for each task. In this way, we can finally sample 100 queries for each task.

- **Document Sampling**: Second, for each sampled query above, we use all originally annotated documents for evaluation (for example, each document is labeled related or not). Since there are some queries will few annotated documents, we randomly sample some unlabeled documents as negative documents for these queries.

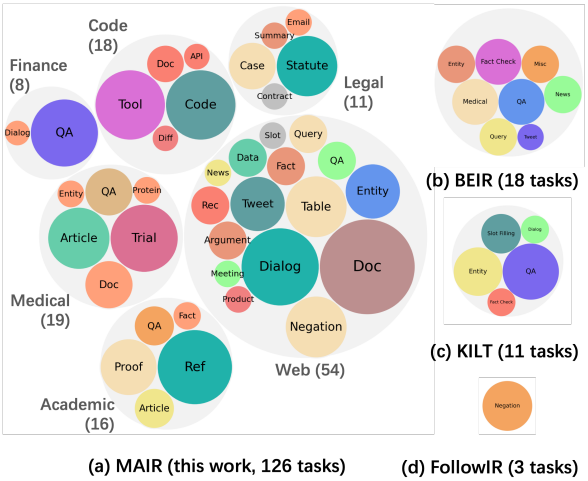

Figure 1: Compared to other datasets, MAIR covers a more diverse types of task. Bubble size represents the number of tasks of each type.

It is noteworthy that some tasks may be originally evaluated on too small size of corpus making it difficult to sample enough documents in the above step 2. To address this issue, we manually combine the documents together if a small task is similar to a larger task. As a result, both tasks can share the same large-scale corpus.

The resulting benchmark comprises 10,038 queries and 4,274,916 documents (about 2 billion tokens based on OpenAI cl32k tokenizer), providing a computationally efficient yet representative testbed for IR models.

2.3 Instruction Annotation

After collecting the tasks above, we manually write the retrieval instructions for each task. All the instructions are written and reviewed by the experts in information retrieval. Following Asai et al. (2022), the basic format of the instruction describes

the query, the passages, and the relevance criterion. For example, instruction of `ClinicalTrials` is "*Given a patient descriptions, retrieve clinical trials that suitable for that patient. The patient description (query) is a questionnaire that is filled by the patient or their clinician. Trials are relevant if the patient met the inclusion criteria and did not meet any exclusion criteria; Trials are partially relevant if the patient met the inclusion criteria but was excluded by one or more exclusion criteria.*",.

For tasks where different search queries require unique instructions, we provide query-level annotations. The following are some examples of instructions besides the basic format: For example, task `Genomics-AdHoc_2007` (Hersh et al., 2007) aims to retrieve passages that contain a specific type of biomedical entity eg. antibodies, proteins, and strains. We annotate the entity type and its definition in instructions for each query.

After the above steps, we obtain the final datasets, which consist of 10,038 queries from 126 IR tasks, with 805 instructions annotated for each task. There are in total 426 document collections and 4,274,916 documents across various domains, such as news articles, scientific papers, web pages, and code repositories. These datasets contain well-designed instructions and various document collections. They can serve as a comprehensive benchmark to evaluate the ability of retrieval systems in understanding natural language instructions and retrieving relevant information from different sources. See

3 Dataset Analysis

The Table 3 and 4 lists the full list of the tasks in MAIR and their input / output, and task type and domain. Tasks in MAIR mainly come from six domains: (i) Web refers to retrieving information from the general web. 47 IR tasks are included in the web domain. (ii) Academic domain focus on retrieving academic literature or retrieval for academic applications. (iii) The code domain consists of 18 tasks, ranging from code retrieval to tool retrieval, and to code agent. (iv) Legal domain focuses on legal-related tasks, such as case retrieval and statute retrieval. (v) Finance domain consists of 8 tasks that concern IR application in Finance task. (vi) Medical domain mainly consists of the shared tasks in TREC-CDS and GENOMICS, which wide range of retrieval and matching tasks in biomedical.

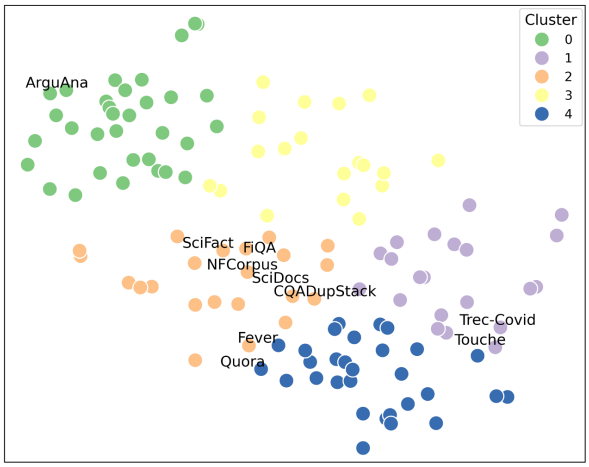

Figure 2: Visualization of the correlation among 126 tasks in MAIR, with annotations for tasks from BEIR. MAIR includes more diverse tasks. Task similarity is determined based on the performance correlation of all baseline models. We employ KMeans for clustering and t-SNE for visualization.

Figure 1 highlights the domain and task categories of MAIR, and compares them with previous datasets. MAIR covers more diverse domains and types of retrieval tasks. Next, we analyze the correlation between different tasks to further study the task diversity of MAIR, and validate the effectiveness of the data sampling approaches.

Task Correlation To measure the task diversity of MAIR, we calculate the similarity of two tasks using the Pearson correlation coefficient of the different models' performance on two tasks. Based on the results of all baseline models on the MAIR, we calculated the correlation between all tasks, and got a correlation matrix in $\mathbf{M} = \mathbb{R}^{126\times126}$, $\mathbf{M}[i, j]$ denote correlation between task i and j. We plot this matrix in Figure 7. To better visualize the matrix, we use t-SNE to visualize the task correlation matrix in 2D.

Figure 2 presents the t-SNE visualization of the correlation matrix, where each task is represented as a point, and the distance between two tasks reflects their correlation. We have annotated the tasks from BEIR and observed that most of these tasks cluster together in the orange group, indicating that the performance of different models on these tasks is highly correlated. In contrast, tasks in MAIR show greater diversity, covering a larger area and thus providing a more comprehensive evaluation of the models. The specific pairwise correlations are displayed in the heatmap in Figure 7, revealing that many tasks exhibit negative correlations. This

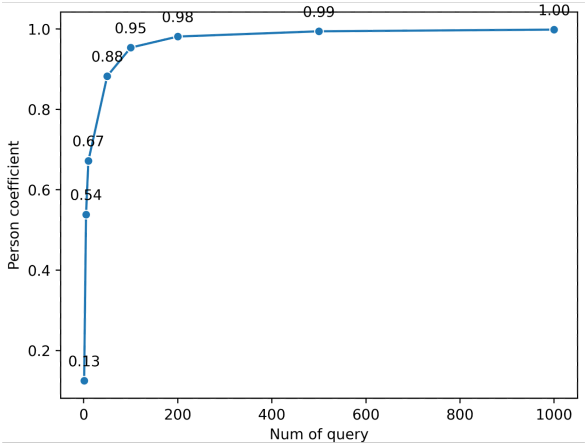

Figure 3: The performance correlation of baseline models with different sampled numbers of queries. Sampling 100 queries achieves a good trade-off between correlation and cost.

suggests that these tasks have significantly different definitions, leading models that perform well on one task to perform poorly on another.

Sampling Effectiveness In our data construction process, we sample the data to obtain a lightweight test set. However, the sampled data may lead to bias in the evaluation results. To measure the effectiveness of our data sampling process, we build a test set with different maximum test sizes cut off, ranging from [1, 5, 10, 50, 100, 200, 500, 10000]. Then, we run the baseline models on these test sets, can compute the task correlation between the full test set (i.e., unsampled test set), and the sampled test set. A high correlation means that evaluation results on the sampled test set are very similar to the evaluation on the full set. Figure 3 illustrates the correlation of different test sets cut off. We can see that retaining more instances leads to a higher correlation, i.e., the evaluation is more robust. Notably, a cut-off of 100 instances achieves over 0.95 Pearson correlation coefficient between the evaluation results, indicating that our sampling method could achieve good reliability while using minimal cost.

4 Experimental Setup

4.1 Models

Sparse Retrieval These measure relevance by computing term overlap. We benchmark BM25, implemented in BM25S (Lù, 2024).

Single-Task Text Embedding These models are trained on a single IR dataset. We benchmark `gtr-t5-base`, `gtr-t5-large`, and `contriever-msmarco`, all of which are trained on MS MARCO (Ni et al., 2021; Izacard et al., 2021).

Non-Instruction-Tuned Multi-Task Text Embedding These models are typically trained on various annotated IR training datasets combined with massive weekly supervised data. We benchmark (i) the GTE series (Li et al., 2023b): `gte-base-en-v1.5`, `gte-large-en-v1.5`; (ii) the BGE series (Xiao et al., 2023): `bge-base-en-v1.5`, `bge-large-en-v1.5`; (iii) the E5 series (Wang et al., 2022a): `e5-small-v2`, `e5-base-v2`, `e5-large-v2`; and (iv) `all-MiniLM-L6-v2` from Sentence Transformer. We also evaluated `text-embedding-3-small` by the OpenAI API[2].

Instruction-Tuned Text Embedding These models are fine-tuned on instruction datasets, where the model input is a query paired with an instruction describing the retrieval task. We benchmark (i) `e5-mistral-instruct`, which optimizes instruction-following ability using LLM-generated data (Wang et al., 2023a); (ii) `NV-Embed-v1`, which utilizes bidirectional attention with an additional latent attention layer to enhance text embedding model (Lee et al., 2024); (iii) `GritLM-7B` (Muennighoff et al., 2024), a unified text embedding and generation model; and (iv) `gte-Qwen2-1.5B-instruct` (Li et al., 2023b), general text embedding based on Qwen2-1.5B.

Cross-Encoder Re-Rankers These models measure the relevance of paired queries and passages using bidirectional or unidirectional Transformers. We benchmark (i) `monoT5-Base`, a T5 encoder model trained on MS MARCO, (ii) `mxbai-rerank-large-v1` and `jina-reranker-v2-base`, multi-task reranker developed by Mxbai and Jina.ai, respectively, and (iii) `bge-reranker-v2-m3` and `bge-reranker-v2-gemma`, trained on massive ranking data with XLM-R and Gemma-2B as their respective backbones.

LLM-based Re-Rankers These models prompt general-purpose LLMs to perform re-ranking in a zero-shot setting. We benchmark RankGPT with the `gpt-3.5-turbo` (Sun et al., 2023).

[2] https://platform.openai.com/docs/guides/embeddings

Model	Avg		Web		Academic		Legal		Medical		Finance		Code	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	40.75	-	40.21	-	38.50	-	47.60	-	46.01	-	51.22	-	32.72	-
`contriever-msmarco`	39.88	-	45.19	-	31.82	-	31.74	-	36.60	-	50.59	-	35.06	-
`gtr-t5-base`	37.79	-	40.55	-	31.36	-	33.77	-	32.95	-	47.78	-	37.18	-
`gtr-t5-large`	41.43	-	43.97	-	35.62	-	37.35	-	33.93	-	51.83	-	42.29	-
`all-MiniLM-L6-v2`	38.66	26.90	40.05	25.56	42.90	35.85	29.20	18.46	35.24	22.32	44.54	28.15	35.90	29.14
`e5-small-v2`	41.36	34.47	43.24	34.12	38.83	32.83	32.20	28.55	37.89	30.82	55.32	50.49	39.70	35.26
`e5-base-v2`	43.45	38.80	43.78	38.95	42.00	38.37	35.76	33.06	40.22	30.68	58.33	54.18	43.48	40.13
`e5-large-v2`	44.75	39.45	44.60	38.93	45.44	40.01	36.96	32.86	41.90	35.13	58.92	54.27	44.28	40.00
`gte-base-en-v1.5`	44.06	34.25	44.72	30.97	40.15	35.34	37.89	34.73	45.63	37.18	61.64	49.23	40.50	33.11
`gte-large-en-v1.5`	46.58	33.87	46.38	30.97	47.00	40.23	41.04	34.25	47.75	32.91	63.47	41.57	41.58	32.66
`bge-base-en-v1.5`	42.58	30.63	43.33	27.70	38.85	31.66	33.56	26.89	43.78	31.58	55.54	38.95	42.46	35.05
`bge-base-en-v1.5`	44.34	28.01	44.91	24.78	41.20	32.05	37.35	24.09	45.31	25.71	58.69	31.38	42.58	34.77
`text-embedding-3-small`	47.89	45.28	48.10	46.65	44.43	41.11	45.71	41.06	42.74	37.81	61.36	57.99	48.87	46.74
`gte-Qwen2-1.5B-instruct`♠	49.14	51.81	46.81	51.30	48.87	49.98	50.91	51.34	45.97	45.20	62.85	64.61	50.40	53.48
`NV-Embed-v1`♠	50.29	51.19	49.76	51.07	50.70	51.14	44.07	44.57	40.71	38.96	68.27	69.59	52.62	54.52
`e5-mistral-7b-instruct`♠	50.85	54.43	48.52	**54.57**	48.78	50.03	**52.94**	52.50	**48.35**	**51.10**	66.52	68.76	52.29	54.83
`GritLM-7B`♠	48.46	**55.26**	43.01	54.50	51.46	**53.45**	50.70	**53.14**	45.38	48.22	64.28	**70.83**	53.62	**57.52**
`monot5-base-msmarco`	43.51	38.27	47.03	42.24	35.44	27.00	41.45	38.41	36.19	31.31	59.34	56.72	40.27	34.25
`mxbai-rerank-large-v1`	42.84	22.91	44.33	24.94	32.98	15.53	42.50	19.14	43.25	24.64	56.54	26.23	41.78	23.78
`jina-reranker-v2-base`	50.88	48.59	**52.31**	50.74	45.22	42.96	49.32	46.10	46.78	42.87	61.35	59.28	51.03	48.24
`bge-reranker-v2-m3`	46.59	40.45	48.63	43.64	40.01	33.38	49.18	42.38	41.12	30.85	64.17	60.36	41.47	34.67
`bge-reranker-v2-gemma`	**52.20**	38.16	51.26	42.40	**52.09**	30.89	47.74	30.47	45.69	28.96	**68.34**	46.41	**54.00**	39.86
`gpt-3.5-turbo`♠	47.84	49.02	47.80	49.70	41.59	41.02	47.47	48.38	42.21	41.97	64.45	64.52	49.93	52.27

Table 2: **Main Results (nDCG@10)** Rows marked with ✓ indicate results with instruction input, while rows marked with ✗ indicate results without instruction input. Models labeled with ♠ are instruction fine-tuned. The last group represents the re-ranking models, all of which re-rank the top 100 results from `text-embedding-3-small`.

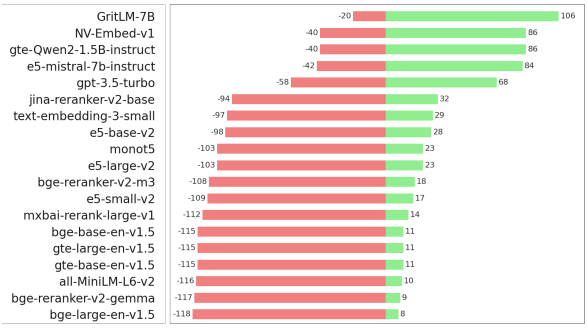

Figure 4: With the addition of instruction, the number of tasks that obtain performance improvement (green part) and reduction (red part). We can see that instruction-tuned models show more improvements while non-instruction-tuned models reduce on most tasks.

4.2 Evaluation

Following previous work, we use **nDCG@10** as the evaluation metric. The overall score is defined as the average score across all queries. We also report the average nDCG@10 for each of the following domains: Web, Academic, Code, Medical, Legal, and Finance. The specific tasks included in each field can be found in Table 5 - 9.

For all models, we consider two settings: (i) *no instruction*, which retrieves passages without task instruction or with a simple web search instruction when required, and (ii) + *instruction*, which retrieves passages with instruction input paired with the query. The performance changes after using instructions indicate the model's ability to understand them. Note that for non-instruction-tuned models, we also test their performance under the instruction setting for reference, even though they may not be optimized to understand instructions.

All re-ranking models use `text-embedding-3-small` as the first-stage retriever and re-rank the top-100 passages. Passages are truncated to the maximum input length of each model.

5 Evaluation Results

5.1 Main Results

Table 2 reports the evaluation results of the tested models. Table 5-9 reports the detailed results of each individual task. We observe that instruction-tuned embedding models with instruction input achieve the best overall performance. `GritLM-7B` achieves an average score of 48.40, with a 6.80 nDCG improvement when instructions are added.

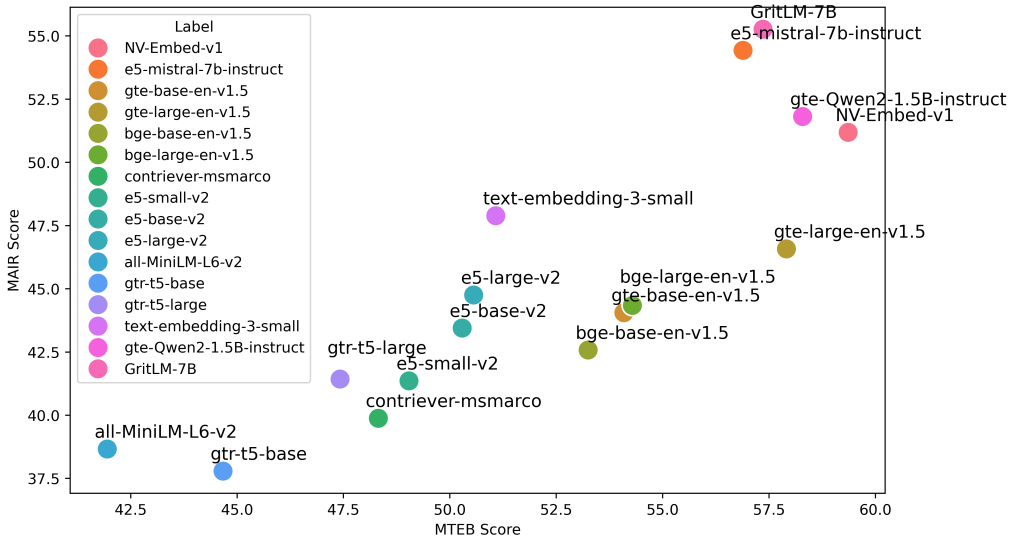

Figure 5: Score between MTEB (Retrieval) and MAIR.

`NV-Embed` also shows a clear improvement with the addition of instructions, though not as significant as `GritLM-7B`. This difference is likely due to the LLM-generated instruction-tuning data enhancing the models' ability to understand instructions (Wang et al., 2023a).

Non-instruction-tuned models experience a performance drop when instructions are added. Some, like `text-embedding-3-small`, show a slight decrease, while others, such as `bge-base-en-v1.5`, exhibit a more significant decline.

For the re-ranker, `bge-reranker-v2-gemma` achieves the best results, outperforming all embedding models when no instructions are provided, but it shows a notable decline in performance when instructions are added. RankGPT based on `gpt-3.5-turbo` achieves results close to `bge-reranker-v2-m3` without instruction input but demonstrates a 1.21 nDCG improvement when instructions are included. This suggests that prompted LLMs can intuitively transfer general instruction-following capabilities to ranking tasks.

5.2 Gain of Instruction

Figure 4 shows that with the addition of instructions, the model's performance improves on a number of the 126 tasks while decreasing on others. We observe that the instruction-tuned models show improvement on more than half of the tasks, with `GritLM-7B` performing the best, achieving improvements on 106 out of 126 tasks. `NV-Embed-v1`, `gte-Qwen2-1.5B-instruct`, and `e5-mistral-7b-instruct` achieve similar results and outperform `gpt-3.5-turbo`. In contrast, non-instruction-tuned models show a performance decrease on more tasks when instructions are added.

5.3 Compare with MTEB

Figure 5 shows the relationship between models' performance on MTEB (Retrieval) and MAIR. We can see that the two benchmarks share a similar trend. Among these models, `text-embedding-3-small` achieves better results on MAIR than on MTEB. For example, on MTEB, `gte-large-en-v1.5` outperforms `text-embedding-3-small` by about 7 points and outperforms `e5-mistral-7b-instuct` by 1 point. However, it performs worse than them on MAIR. This is probably because `text-embedding-3-small` has better generalization, as it performs better on massive unseen tasks in MAIR, while `gte-large-en-v1.5` is more optimized towards tasks in MTEB. We also observe that single-task models such as `contriever-msmarco` perform poorly on MAIR, which indicates that MAIR requires more generalization ability.

5.4 Analysis on IFEval

To evaluate the model on challenging instruction-following tasks, we designed the IFEval task (Zhou et al., 2023) within MAIR. IFEval consists of 8 different instruction-following subtasks, such as `format` (selecting responses in a specific format), `keywords` (including specific words), and `length` (adhering to length restrictions). The retrieval task

Figure 6: Results (nDCG@10) on IFEval sub-tasks.

in IFEval is to select the answer that correctly follows the instructions from among the 100 candidates. Specifically, building on the original IFEval data (Zhou et al., 2023), we used `gpt-3.5-turbo` to generate 100 candidate answers for each question, ensuring that only 10 fully follow the given instructions. IFEval is challenging for retrieval models because (i) the instructions are out-of-training-distribution; (ii) all candidate answers are semantically relevant to the question, thus, the model must focus on the instructions to identify the correct answer.

Figure 6 demonstrates the performance of `GritLM-7B`, `bge-reranker-v2-gemma`, and RankGPT with `gpt-3.5-turbo`, `gpt-4o-mini`, and `gpt-4o` on the 8 subtasks of IFEval. The results show that existing instruction-tuned retrieval models still perform poorly on challenge instruction-followed ranking task. For example, the `GritLM-7B` achieves an nDCG@10 score of less than 60 on 7 out of the 8 tasks. In contrast, advanced LLM `gpt-4o` achieves an nDCG@10 score of over 80 on 6 out of the 8 tasks. It indicates again the shortcomings of instruction-tuned retrieval models in handling complex information requirements, and utilizing advanced language models as supervisors might be an effective strategy. For full results on IFEval, refer to Tables 13 and 10.

6 Related Work and Background

6.1 Instruction tuning for Retrieval

Instruction tuning has been a effective technique in the development of LLMs (Chung et al., 2024; Wang et al., 2023b, 2022b). In this process, models are trained on diverse instruction-response pairs, empowering them with the ability to adaptively perform unseen tasks based on instructions (Chen et al., 2024; Wei et al., 2022). This has inspired emergent research on training instruction-tuned retrieval models. Asai et al. (2022); Su et al. (2022) are the earliest works in this direction, where they propose training text embedding models with instructions alongside the query, enabling the models to perform well on unseen tasks using instructions. Recent research has started to finetune larger text embedding models with instructions, such as Mistral-7B and Mixtral-7x8B (Muennighoff et al., 2024), while Wang et al. (2023a) further proposes using LLMs to generate instruction-tuning data for training retrievers. These instruction-tuned embeddings have climbed to the top of existing IR leaderboards like MTEB (Muennighoff et al., 2022). Meanwhile, some paper explore prompt-based method to instruct general-purpose LLMs for re-ranking tasks (Sun et al., 2023).

6.2 IR Benchmark

Benchmarking has been a crucial part of IR system development. Traditional IR benchmarks typically evaluate models on narrow tasks, like question-answering, or on a certain domain (Campos et al., 2016; Karpukhin et al., 2020). Recently, with the advance of pre-trained language models and language model-based retrievers, such as monoBERT (Nogueira and Cho, 2019) and GTR (Ni et al., 2021), increasing attention has been paid to constructing multi-task IR benchmarks and evaluating retrievers on out-of-domain tasks. A typical effort is BEIR (Thakur et al., 2021), which consists of 18 tasks across 4 domains. MTEB (Muennighoff et al., 2022) further extends BEIR by adding other embedding-related non-retrieval tasks like clustering and classification. KILT (Petroni et al., 2020) collects 11 knowledge-intensive language tasks. However, given the recent progress of instruction-tuned retrievers, existing benchmarks cannot comprehensively evaluate models' abilities since the number of tasks and the scope of tasks are limited. Some recent papers, such as FollowIR (Weller et al., 2024)

and InstructIR (Oh et al., 2024), propose benchmarks that specifically focus on evaluating models' instruction-following abilities. FollowIR rewrites the original narratives in three TREC shared tasks to include some exclusionary instructions, making some relevant passages become irrelevant. InstructIR utilizes LLM to generate synthetic user backgrounds and further rewrites the relevant passages using LLM to fit the background. However, these tasks are synthetic and only include limited tasks, making it difficult to robustly evaluate the performance of models on real-world unseen instructions. In comparison, MAIR is a large-scale multi-task IR benchmark consisting of 126 distinct tasks. Most tasks are long-tail and without training data, and detailed instructions for each task are manually annotated.

7 Conclusion

In this paper, we introduce a novel massive instructed IR benchmark called MAIR for evaluating instruction-tuned retrieval models. Compared with existing related benchmarks, MAIR has a more comprehensive coverage of various IR tasks, with different types of queries, documents and relevance criteria. Specifically, MAIR comprises 126 distinct IR tasks, with 805 newly annotated instructions. All tasks and instructions are manually selected and annotated to ensure the benchmark can assess the instruction-following capabilities of different IR models in practice.

Based on MAIR, we benchmark various retrieval models, including both instruction-tuned and non-instruction-tuned models. Our results demonstrate that MAIR poses greater challenges than existing benchmarks, and provides a more comprehensive evaluation.

Limitation

The limitations of this work include the lack of study of retrieval in a multilingual setting. Our benchmark focuses only on the English language and considers only text retrieval. Therefore, we plan to explore multilingual IR settings in the future. Another limitation is the lack of study on prompt sensitivity. It is well known that LLMs are sensitive to prompt words. We plan to annotate more instructions in the future to study how LLM performance is impacted by prompt words.

Ethics Statement

We acknowledge the importance of the ACM Code of Ethics and fully agree with it. We ensure that this work is compatible with the provided code in terms of publicly accessible datasets and models. Risks and harms associated with large language models include the generation of harmful, offensive, or biased content. The new benchmark is composed of various previous datasets and is therefore licensed under their respective data licenses.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Jafar Afzali, Aleksander Mark Drzewiecki, and Krisztian Balog. 2021. Pointrec: A test collection for narrative-driven point of interest recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search. *ArXiv*, abs/2407.18940.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen tau Yih. 2022. Task-aware retrieval with instructions. In *Annual Meeting of the Association for Computational Linguistics*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *ArXiv*, abs/2108.07732.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the fire 2019 aila track: Artificial intelligence for legal assistance. In *Fire*.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Christian Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of touché 2020: Argument retrieval. In *Conference and Labs of the Evaluation Forum*.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Arun Tejasvi Chaganty, Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, and Filip Radlinski. 2023. Beyond single items: Exploring user preferences in item sets with the conversational playlist curation dataset. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. 2021. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. In *ICLR*.

Xiuying Chen, Hind Alamro, Li Mingzhe, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022a. Target-aware abstractive related work generation with contrastive learning. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema N Moussa, Matthew I. Beane, Ting-Hao 'Kenneth' Huang, Bryan R. Routledge, and William Yang Wang. 2021a. Finqa: A dataset of numerical reasoning over financial data. *ArXiv*, abs/2109.00122.

Zhiyu Chen, SHIYANG LI, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022b. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Zhiyu Chen, Shuo Zhang, and Brian D. Davison. 2021b. Wtr: A test collection for web table retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *JMLR*, pages 1–53.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Paslaru Bontas Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *International Conference on Language Resources and Evaluation*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *ArXiv*, abs/1907.09190.

Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, Jiafeng Guo, et al. 2022. Pre-training methods in information retrieval. *Foundations and Trends® in Information Retrieval*, 16(3):178–317.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *ArXiv*, abs/2108.05540.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. 2024. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. *ArXiv*, abs/2405.01121.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9:459–479.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Xiaodong Song, and Jacob Steinhardt. 2021a. Measuring coding challenge competence with apps. *ArXiv*, abs/2105.09938.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. Cuad: An expert-annotated nlp dataset for legal contract review. *ArXiv*, abs/2103.06268.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874.

William R. Hersh, Aaron M. Cohen, Jianji Yang, Ravi Teja Bhupatiraju, Phoebe M. Roberts, and Marti A. Hearst. 2007. Trec 2007 genomics track overview. In *Text Retrieval Conference*.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Conference on Empirical Methods in Natural Language Processing*.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. *Proceedings of the 20th Australasian Document Computing Symposium*.

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. *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 29–32.

Hamel Husain, Hongqiu Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *ArXiv*, abs/1909.09436.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *ArXiv*, abs/2311.11944.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *ArXiv*, abs/2310.06770.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *ArXiv*, abs/2004.04906.

Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2021. A test collection for ad-hoc dataset retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Anastassia Kornilova and Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *ArXiv*, abs/1910.00523.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *ArXiv*, abs/1706.04115.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023a. Lecardv2: A large-scale chinese legal case retrieval dataset. *ArXiv*, abs/2310.17609.

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

Tengteng Lin, Qiaosheng Chen, Gong Cheng, Ahmet Soylu, Basil Ell, Ruoqi Zhao, Qing Shi, Xiaxia Wang, Yu Gu, and Evgeny Kharlamov. 2022. Acordar: A test collection for ad hoc content-based (rdf) dataset retrieval. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tianyang Liu, Canwen Xu, and Julian McAuley. 2023. Repobench: Benchmarking repository-level code auto-completion systems. *ArXiv*, abs/2306.03091.

Antoine Louis, Gerasimos Spanakis, and G. van Dijck. 2021. A statutory article retrieval dataset in french. In *Annual Meeting of the Association for Computational Linguistics*.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *ArXiv*, abs/2407.03618.

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. *Companion Proceedings of the The Web Conference 2018*.

Niklas Muennighoff, Qian Liu, Qi Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro von Werra, and S. Longpre. 2023. Octopack: Instruction tuning code large language models. *ArXiv*, abs/2308.07124.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *ArXiv*, abs/2402.09906.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas A. Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David P. Schnurr, Felipe Petroski Such, Kenny Sai-Kin Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *ArXiv*, abs/2201.10005.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers. *ArXiv*, abs/2112.07899.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *ArXiv*, abs/1901.04085.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. Instructir: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2020. Kilt: a benchmark for knowledge intensive language tasks. In *North American Chapter of the Association for Computational Linguistics*.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. *ArXiv*, abs/2403.17859.

Yujia Qin, Shi Liang, Yining Ye, Kunlun Zhu, Lan Yan, Ya-Ting Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Marc H. Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *ArXiv*, abs/2307.16789.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. 2020. Trec-covid: rationale and structure of an information retrieval shared task for covid-19. *Journal of the American Medical Informatics Association : JAMIA*, 27:1431 – 1436.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen tau Yih, Joëlle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Conference on Empirical Methods in Natural Language Processing*.

Chris Samarinas and Hamed Zamani. 2024. Procis: A benchmark for proactive retrieval in conversations. *ArXiv*, abs/2405.06460.

Zhengliang Shi, Shen Gao, Xiuyi Chen, Yue Feng, Lingyong Yan, Haibo Shi, Dawei Yin, Zhumin Chen, Suzan Verberne, and Zhaochun Ren. 2024a. Chain of tools: Large language model is an automatic multi-tool learner. *arXiv preprint arXiv:2405.16533*.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024b. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ivan Srba, Branislav Pecher, Matús Tomlein, Róbert Móro, Elena Stefancova, Jakub Simko, and Mária Bieliková. 2022. Monant medical misinformation dataset: Mapping articles to fact-checked claims. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv*, abs/2212.09741.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *ArXiv*, abs/1803.05355.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Annual Meeting of the Association for Computational Linguistics*.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Conference on Empirical Methods in Natural Language Processing*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023a. Improving text embeddings with large language models. *ArXiv*, abs/2401.00368.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, pages 13484–13508.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language. *ArXiv*, abs/2104.01112.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*.

Orion Weller, Dawn J Lawrie, and Benjamin Van Durme. 2023. Nevir: Negation in neural information retrieval. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Marco Wrzalik and Dirk Krechel. 2021. Gerdalir: A german dataset for legal information retrieval. In *NLLP*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *ArXiv*, abs/2309.07597.

Jason Yang, Ariane Mora, Shengchao Liu, Bruce J. Wittmann, Anima Anandkumar, Frances H. Arnold, and Yisong Yue. 2024. Care: a benchmark suite for the classification and retrieval of enzymes. In *NeurIPS*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *ArXiv*, abs/1809.08887.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, He Yang Er, Irene Z Li, Bo Pang, Tao Chen, Emily Ji, Shreya Dixit, David Proctor, Sungrok Shim, Jonathan Kraft, Vincent Zhang, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2019. Sparc: Cross-domain semantic parsing in context. *ArXiv*, abs/1906.02285.

Wenhao Zhang, Mengqi Zhang, Shiguang Wu, Jiahuan Pei, Zhaochun Ren, Maarten de Rijke, Zhumin Chen, and Pengjie Ren. 2024. Excluir: Exclusionary neural information retrieval. *ArXiv*, abs/2404.17288.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shanshan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, Teng Su, Zhilin Yang, and Jie Tang. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *ArXiv*, abs/2303.17568.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *ArXiv*, abs/2311.07911.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2022. Docprompting: Generating code by retrieving the docs. In *International Conference on Learning Representations*.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat seng Chua. 2022. Towards complex document understanding by discrete reasoning. *Proceedings of the 30th ACM International Conference on Multimedia*.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

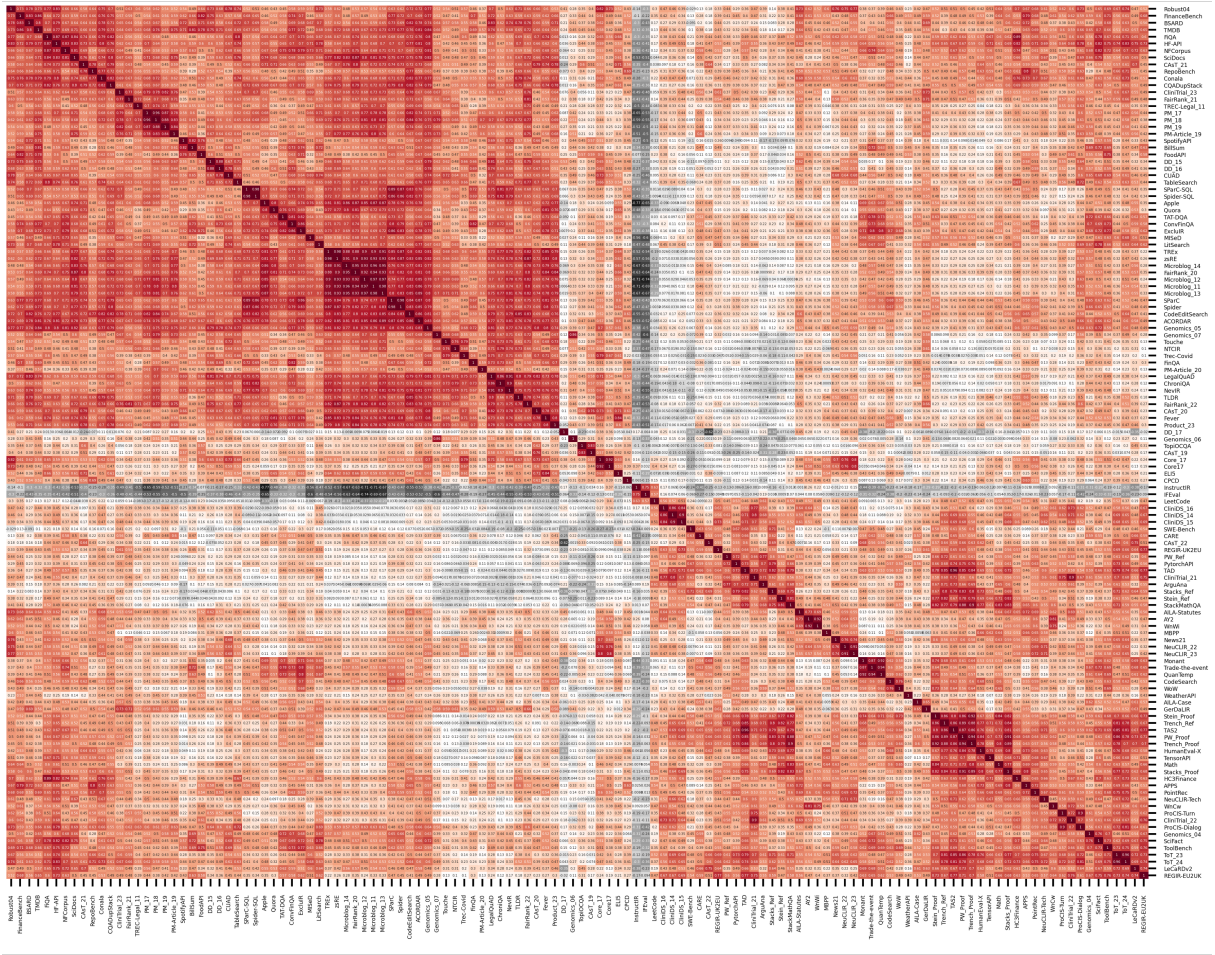

Figure 7: Task correlation heat map. The correlation between two tasks is measured by the Pearson correlation coefficient of different models' performance on these tasks. This coefficient ranges from -1 to 1. A value close to 1 indicates that models which perform well on Task A are likely to perform well on Task B. Conversely, a value close to -1 indicates that models which perform well on Task A are likely to perform poorly on Task B.

A Data Card

The data and evaluation code are available at https://github.com/sunnweiwei/MAIR. Please refer to https://github.com/sunnweiwei/MAIR/blob/main/MAIR-Example.pdf for data processing details and examples of MAIR.

Table 3: List of tasks in MAIR.

Task	Reference	Query	Doc	Task Type	Domain
Competition-Math	Hendrycks et al. (2021c)	Math question	Answer	Question Answering	Academic
ProofWiki_Proof	Welleck et al. (2021)	Math Statement	Proof	Proof Retrieval	Academic
Stacks_Proof	Welleck et al. (2021)	Math Statement	Proof	Proof Retrieval	Academic
Stein_Proof	Welleck et al. (2021)	Math Statement	Proof	Proof Retrieval	Academic
Trench_Proof	Welleck et al. (2021)	Math Statement	Proof	Proof Retrieval	Academic
ProofWiki_Reference	Welleck et al. (2021)	Math Statement	Reference	Reference Retrieval	Academic
Stacks_Reference	Welleck et al. (2021)	Math Statement	Reference	Reference Retrieval	Academic
Stein_Reference	Welleck et al. (2021)	Math Statement	Reference	Reference Retrieval	Academic
Trench_Reference	Welleck et al. (2021)	Math Statement	Reference	Reference Retrieval	Academic
TAD	Chen et al. (2022a)	Paper Abstract	Citation	Reference Retrieval	Academic
TAS2	Chen et al. (2022a)	Paper Abstract	Citation	Reference Retrieval	Academic
StackMathQA	Chen et al. (2022a)	Math Question	Answer	Question Answering	Academic
SciDocs	Cohan et al. (2020)	Paper Title	Citation	Reference Retrieval	Academic
SciFact	Wadden et al. (2020)	Claim	Document	Fact Checking	Academic
FairRanking_2020	Fair-TREC	Key Words	Academic Articles	Web Search	Academic
LitSearch	Ajith et al. (2024)	Question	Articles	Literature Search	Academic
APPS	Hendrycks et al. (2021a)	Code Problem	Solution	Code Retrieval	Code
CodeEditSearch	Muennighoff et al. (2023)	Commit	Code Diff	Code Retrieval	Code
CodeSearchNet	Husain et al. (2019)	Function Header	Function	Code Retrieval	Code
HumanEval-X	Zheng et al. (2023)	Code Problem	Solution	Code Retrieval	Code
LeetCode	LeetCode	Code Problem	Solution	Code Retrieval	Code
MBPP	Austin et al. (2021)	Code Problem	Solution	Code Retrieval	Code
Conala	Zhou et al. (2022)	NL Command	Command Doc	Code Retrieval	Code
TLDR	Zhou et al. (2022)	CNL Command	Command Doc	Code Retrieval	Code
RepoBench	Liu et al. (2023)	Code Context	Next Function	GitHub Agent	Code
SWE-Bench	Jimenez et al. (2023)	GitHub Issue	Related File	GitHub Agent	Code
FoodAPI	Shi et al. (2024a)	Question	Food API	Tool Retrieval	Code
TensorAPI	Patil et al. (2023)	Question	Tensor API	Tool Retrieval	Code
HF-API	Patil et al. (2023)	Question	HuggingFace API	Tool Retrieval	Code
PyTorchAPI	Patil et al. (2023)	Question	PyTorch API	Tool Retrieval	Code
SpotifyAPI	Shi et al. (2024a)	Question	Spotify API	Tool Retrieval	Code
TMDB	Shi et al. (2024a)	Question	TMDB API	Tool Retrieval	Code
WeatherAPI	Shi et al. (2024a)	Question	Weather API	Tool Retrieval	Code
ToolBench	Qin et al. (2023)	Question	Tool	Tool Retrieval	Code
Apple	FinQabench	Question	Paragraph	Question Answering	Finance
ConvFinQA	Chen et al. (2022b)	Dialog	Table & Paragraph	Dialog Retrieval	Finance
FinQA	Chen et al. (2021a)	Question	Paragraph	Question Answering	Finance
FinanceBench	Islam et al. (2023)	Question	Pages	Question Answering	Finance
HC3Finance	HC3Finance	Question	Answer	Question Answering	Finance
TAT-DQA	Zhu et al. (2022)	Question	Table & Paragraph	Question Answering	Finance
Trade-the-event	Trade-the-event	Title	Article	Summary Retrieval	Finance
FiQA	Maia et al. (2018)	Question	Answer	Question Answering	Finance
AILA2019-Case	Bhattacharya et al. (2019)	Situation	Prior Case	Case Retrieval	Legal
AILA2019-Statutes	Bhattacharya et al. (2019)	Situation	Statute	Statute Retrieval	Legal
BSARD	Louis et al. (2021)	Question	Statute	Statute Retrieval	Legal
BillSum	Kornilova and Eidelman (2019)	Summary	Article	Summary Retrieval	Legal
CUAD	Hendrycks et al. (2021b)	Instruction	Highlight	Contract Review	Legal
GerDaLIR	Wrzalik and Krechel (2021)	Legal Case	Prior Cases	Case Retrieval	Legal
LeCaRDv2	Li et al. (2023a)	Legal Case	Prior Cases	Case Retrieval	Legal
LegalQuAD	Hoppe et al. (2021)	Question	Statute	Statute Retrieval	Legal
REGIR-EU2UK	Chalkidis et al. (2021)	EU Directive	UK Legislation	Statute Retrieval	Legal
REGIR-UK2EU	Chalkidis et al. (2021)	UK Legislation	EU Directive	Statute Retrieval	Legal
TREC-Legal_2011	TREC Legal	Request	Communications	Case Retrieval	Legal
NFCorpus	Boteva et al. (2016)	Question	Medical Document	Medical QA	Medical
Trec-Covid	Roberts et al. (2020)	Medical Question	Answer	Medical QA	Medical
Monant	Srba et al. (2022)	Medical Claim	Document	Question Answering	Medical
CliniDS_2014	TREC CDS 2014	Medical Case	Articles	Medical IR	Medical
CliniDS_2015	TREC CDS 2015	Medical Case	Articles	Medical IR	Medical
CliniDS_2016	TREC CDS 2016	Health Record	Articles	Medical IR	Medical

Table 4: List of tasks in MAIR.

Task	Reference	Query	Doc	Task Type	Domain
ClinicalTrials_2021	TREC CDS 2021	Health Record	Clinical Trials	Medical IR	Medical
ClinicalTrials_2022	TREC CDS 2022	Health Record	Clinical Trials	Medical IR	Medical
ClinicalTrials_2023	TREC CDS 2023	Patient Description	Clinical Trials	Medical IR	Medical
Genomics-AdHoc_2004	Genomics	Question	Document	Medical IR	Medical
Genomics-AdHoc_2005	Genomics	Question	Document	Medical IR	Medical
Genomics-AdHoc_2006	Genomics	Question	Document	Medical IR	Medical
Genomics-AdHoc_2007	Genomics	Question	Entity & Passages	Medical IR	Medical
PM_(17,18,19)	Precision Medicine	Patient Data	Clinical Trials	Medical IR	Medical
PM-Article_(19,20)	Precision Medicine	Patient Data	Articles	Medical IR	Medical
CARE	Yang et al. (2024)	Reaction	Proteins Documents	Medical IR	Medical
ELI5	Fan et al. (2019)	Question	Passages	Question Answering	Web
Fever	Thorne et al. (2018)	Claim	Passages	Fact Checking	Web
AY2	Hoffart et al. (2011)	Entity Mention	Entity Page	Entity Linking	Web
WnCw	Guo and Barbosa (2018)	Entity Mention	Entity Page	Entity Linking	Web
WnWi	Guo and Barbosa (2018)	Entity Mention	Entity Page	Entity Linking	Web
TREx	ElSahar et al. (2018)	Entity & Relation	Entity Page	Slot Filling	Web
zsRE	Levy et al. (2017)	Entity & Relation	Entity Page	Slot Filling	Web
WoW	Dinan et al. (2018)	Dialog	Passage	Dialog Retrieval	Web
ArguAna	Wachsmuth et al. (2018)	Claim	Document	Argument Retrieval	Web
CQADupStack	Hoogeveen et al. (2015)	Question	Duplicate Question	Query Retrieval	Web
Quora	Quora	Question	Duplicate Question	Query Retrieval	Web
TopiOCQA	Adlakha et al. (2021)	Dialog	Passage	Dialog Retrieval	Web
Touche	Bondarenko et al. (2020)	Question	Passages	Argument Retrieval	Web
ACORDAR	Lin et al. (2022)	Request	Dataset	Dataset Retrieval	Web
CPCD	Chaganty et al. (2023)	Dialog	Music	Recommendation	Web
ChroniclingAmericaQA	Piryani et al. (2024)	Question	News	News Retrieval	Web
NTCIR	Kato et al. (2021)	Key Words	Dataset	Dataset Retrieval	Web
PointRec	Afzali et al. (2021)	Question	POI	Recommendation	Web
ProCIS-Dialog	Samarinas and Zamani (2024)	Dialog	Passage	Dialog Retrieval	Web
ProCIS-Turn	Samarinas and Zamani (2024)	Utterance	Passage	Dialog Retrieval	Web
QuanTemp	V et al. (2024)	Numerical Claim	Document	Fact Checking	Web
WebTableSearch	Chen et al. (2021b)	Question	Table	Table Retrieval	Web
CAsT_(19,20,21,22)	TREC CAsT	Dialog	Passage	Dialog Retrieval	Web
DD_2015	Domain2015	Topic	Document	Web Search	Web
DD_2016	Domain2016	Topic	Document	Web Search	Web
DD_2017	Domain2017	Topic	Document	Web Search	Web
FairRanking_2021	Fair-TREC	WikiProject	Wikipedia Page	Web Search	Web
FairRanking_2022	Fair-TREC	WikiProject	Wikipedia Page	Web Search	Web
NeuCLIR-Tech	NeuCLIR	Question	Document	Web Search	Web
NeuCLIR_2022	NeuCLIR	Question	Document	Web Search	Web
NeuCLIR_2023	NeuCLIR	Question	Document	Web Search	Web
ProductSearch_2023	TREC Product Search	Key Words	Product	Product Search	Web
ToT_2023	TREC ToT	Description	Wikipedia Page	Tip-of-the-Tongue	Web
ToT_2024	TREC ToT	Description	Wikipedia Page	Tip-of-the-Tongue	Web
Microblog	TREC Microblog	Topic	Tweet	Tweet Retrieval	Web
MISeD	Golany et al. (2024)	Dialog	Meeting Transcript	Meeting Retrieval	Web
SParC	Yu et al. (2019)	Question	Table	Table Retrieval	Web
SParC-SQL	Yu et al. (2019)	SQL Statement	Table	Table Retrieval	Web
Spider	Yu et al. (2018)	Question	Table	Table Retrieval	Web
Spider-SQL	Yu et al. (2018)	SQL Statement	Table	Table Retrieval	Web
ExcluIR	Zhang et al. (2024)	Question	Passage	Negative Retrieval	Web
FollowIR_Core17	Weller et al. (2024)	Question	Document	Negative Retrieval	Web
FollowIR_News21	Weller et al. (2024)	Question	News	Negative Retrieval	Web
FollowIR_Robust04	Weller et al. (2024)	Question	News	Negative Retrieval	Web
InstructIR	Oh et al. (2024)	Question	Passage	Web Search	Web
NevIR	Weller et al. (2023)	Question	Passage	Negative Retrieval	Web
IFEval	Zhou et al. (2023)	Question	Answer	Question Answering	Web

Table 5: Results (nDCG@10) on MAIR tasks. The last group represents the re-ranking models, which all re-rank the top 100 results of `text-embedding-3-small`.

Model	Math		PW_Proof		PW_Ref		Stacks_Proof		Stacks_Ref		Stein_Proof		Stein_Ref		Trench_Proof		Trench_Ref		TAD	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	50.82	-	49.71	-	12.23	-	28.74	-	24.18	-	59.01	-	23.47	-	63.16	-	22.13	-	26.16	-
`contriever-msmarco`	47.67	-	38.04	-	11.76	-	13.19	-	20.31	-	50.51	-	20.76	-	43.16	-	19.56	-	18.71	-
`gtr-t5-base`	60.22	-	33.65	-	12.11	-	16.33	-	16.68	-	51.58	-	23.73	-	38.12	-	21.83	-	20.89	-
`gtr-t5-large`	70.21	-	44.40	-	11.96	-	20.69	-	17.80	-	60.47	-	26.36	-	46.13	-	23.23	-	22.17	-
`all-MiniLM-L6-v2`	79.14	76.00	51.29	44.79	12.47	12.39	38.39	31.50	20.32	21.14	67.69	63.22	33.45	32.96	63.62	62.70	25.59	25.04	25.79	21.39
`e5-small-v2`	72.35	68.75	42.48	37.58	10.50	10.21	26.08	21.90	21.66	19.48	63.65	54.20	25.01	25.75	56.85	48.41	21.20	18.79	24.59	20.20
`e5-base-v2`	83.67	80.09	44.40	49.99	10.41	11.80	28.64	30.69	23.44	21.94	64.03	62.86	31.16	28.04	59.89	57.95	22.30	22.34	27.16	24.72
`e5-large-v2`	81.94	83.30	59.67	58.26	10.45	10.47	33.52	29.14	22.87	21.63	68.10	65.83	28.21	28.43	64.78	61.59	22.75	21.35	30.16	26.60
`gte-base-en-v1.5`	73.63	70.75	43.97	39.20	14.91	14.40	22.55	25.47	23.11	20.16	58.93	56.08	24.24	25.40	48.15	44.79	21.81	21.57	31.09	28.20
`gte-large-en-v1.5`	86.75	78.10	56.04	48.00	13.63	10.86	28.61	25.12	22.72	22.66	65.93	57.78	31.20	25.63	63.07	54.52	24.83	23.65	33.12	30.14
`bge-base-en-v1.5`	64.68	61.02	44.58	42.30	13.49	13.41	22.02	22.53	22.33	20.80	59.52	52.29	23.93	23.44	51.69	44.27	22.66	21.63	28.65	25.75
`bge-large-en-v1.5`	74.72	65.73	40.59	35.73	13.15	12.73	25.21	23.81	22.83	19.88	63.04	55.05	28.36	28.46	49.98	46.75	23.21	21.13	32.66	27.57
`text-embedding-3-small`	82.71	78.51	53.04	47.37	12.75	11.09	33.45	29.79	23.67	22.94	63.62	58.62	28.49	25.89	64.95	59.71	25.88	24.52	27.44	26.47
`e5-mistral-7b-instruct`	82.54	87.39	68.05	74.26	14.14	14.53	44.08	46.44	25.48	25.42	73.73	72.56	29.37	34.41	75.33	74.84	23.87	25.09	29.68	29.29
`NV-Embed-v1`	86.12	86.37	67.22	71.76	14.11	15.33	51.51	49.84	26.15	26.27	72.80	72.76	34.02	37.58	71.73	71.07	27.83	29.55	31.10	30.34
`GritLM-7B`	90.98	93.18	68.53	75.03	13.63	15.09	53.13	53.97	26.50	26.19	76.05	75.53	33.23	32.09	76.43	74.75	27.70	29.76	36.61	39.02
`gte-Qwen2-1.5B-instruct`	90.62	90.17	63.21	70.14	12.15	13.38	42.49	46.98	24.90	22.42	67.91	68.42	29.44	27.36	74.64	74.08	26.29	26.12	35.38	34.60
`monot5-base-msmarco`	69.58	57.96	33.63	17.62	10.06	8.20	26.36	10.64	14.43	12.04	62.74	43.21	22.58	22.82	41.62	24.69	20.16	14.60	16.96	8.40
`bge-reranker-v2-m3`	81.77	76.17	39.32	23.22	15.62	5.94	30.76	21.69	15.72	5.30	67.84	51.65	25.59	21.21	51.75	38.10	22.54	15.28	21.61	16.83
`bge-reranker-v2-gemma`	91.86	87.26	66.25	29.24	17.95	15.56	48.08	22.66	24.01	15.20	74.86	53.92	36.47	23.12	78.94	41.69	28.97	17.47	34.63	11.88
`jina-reranker-v2-base`	78.89	81.25	51.65	47.23	12.19	12.75	34.12	31.67	19.66	21.42	66.66	59.81	24.12	25.79	63.45	62.79	22.62	22.24	30.25	28.64
`mxbai-rerank-large-v1`	63.35	30.41	31.90	12.47	10.78	4.90	17.22	12.50	13.38	13.67	42.36	14.63	20.39	18.96	27.39	7.03	19.21	11.19	15.52	8.04
`gpt-3.5-turbo`	83.28	80.46	58.09	56.31	12.35	15.02	31.82	31.79	22.76	24.69	65.82	56.94	25.28	28.38	58.00	54.20	23.85	25.44	27.41	23.86

Model	TAS2		StackMathQA		APPS		CodeEditS		CodeSearch		Conala		HumanEval-X		LeetCode		MBPP		RepoBench	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	30.45	-	29.67	-	11.66	-	41.39	-	41.37	-	13.06	-	18.64	-	27.49	-	7.02	-	7.06	-
`contriever-msmarco`	27.87	-	17.34	-	9.38	-	36.10	-	46.05	-	17.18	-	22.72	-	21.10	-	58.26	-	7.44	-
`gtr-t5-base`	25.80	-	22.00	-	14.59	-	37.01	-	53.53	-	15.64	-	36.80	-	24.06	-	67.92	-	14.09	-
`gtr-t5-large`	29.59	-	33.47	-	19.72	-	42.80	-	59.53	-	20.41	-	51.41	-	27.08	-	74.61	-	14.73	-
`all-MiniLM-L6-v2`	34.90	27.40	47.08	44.48	18.51	16.58	40.23	10.90	53.85	42.16	13.97	3.37	50.12	45.46	24.71	36.03	69.66	68.54	11.56	4.87
`e5-small-v2`	33.26	27.70	42.33	41.66	11.08	14.33	39.67	16.05	65.87	59.20	14.94	6.13	53.21	47.71	21.87	26.16	68.82	70.43	11.79	9.72
`e5-base-v2`	33.46	29.03	46.73	48.29	21.77	23.37	45.49	21.13	71.23	64.96	17.92	11.06	64.46	62.73	26.08	43.11	72.58	71.26	8.72	9.53
`e5-large-v2`	40.05	32.12	57.94	58.48	27.37	27.50	45.55	25.54	70.48	67.71	18.24	8.76	65.85	63.14	26.15	34.45	81.25	79.93	10.30	7.89
`gte-base-en-v1.5`	39.30	36.09	32.01	30.22	15.92	9.42	33.46	5.12	48.42	37.39	22.29	5.93	39.12	33.99	23.34	42.62	66.50	63.85	12.53	5.17
`gte-large-en-v1.5`	40.65	36.68	57.51	53.20	18.34	10.17	30.33	3.72	54.51	30.94	26.88	4.40	45.34	29.27	24.53	45.70	73.66	64.25	11.43	1.95
`bge-base-en-v1.5`	37.44	29.52	35.24	31.76	16.82	14.66	37.21	6.63	76.61	60.01	20.74	14.44	46.21	36.55	25.95	37.30	69.59	68.76	9.92	2.96
`bge-large-en-v1.5`	41.60	32.66	36.52	33.06	17.06	14.50	40.25	7.87	77.86	59.09	18.96	10.92	43.62	34.10	28.78	43.21	73.80	60.00	8.48	4.93
`text-embedding-3-small`	40.37	37.80	44.45	41.91	32.05	31.01	46.25	31.01	78.25	74.70	20.50	21.36	85.31	81.66	29.47	47.31	72.01	73.28	14.78	12.61
`e5-mistral-7b-instruct`	44.36	40.52	55.86	57.77	37.01	36.29	54.45	57.15	66.60	67.11	26.99	36.63	77.50	78.10	35.08	47.37	83.81	85.35	14.81	18.83
`NV-Embed-v1`	47.24	43.72	56.08	59.06	40.05	41.69	62.64	63.61	67.25	64.51	31.91	35.02	79.32	83.02	36.09	49.87	85.00	86.10	17.99	18.90
`GritLM-7B`	47.84	51.29	60.23	62.39	45.55	48.38	62.72	67.32	66.34	67.07	26.17	32.52	83.86	84.22	33.94	52.72	82.89	88.58	21.34	20.75
`gte-Qwen2-1.5B-instruct`	44.71	44.43	51.66	51.05	35.56	46.33	41.53	41.50	78.98	80.56	27.24	31.36	90.39	90.30	33.86	50.21	84.91	88.23	10.35	19.92
`monot5-base-msmarco`	20.80	7.37	23.62	14.60	19.19	19.76	49.62	43.25	59.49	38.28	21.54	20.87	27.04	11.21	17.61	18.11	66.00	63.39	14.39	10.71
`bge-reranker-v2-m3`	29.43	27.33	23.88	26.50	18.32	20.34	53.74	36.72	70.70	49.42	18.19	16.32	27.34	9.90	26.69	30.34	69.49	64.63	7.41	4.22
`bge-reranker-v2-gemma`	45.72	19.05	62.17	36.88	41.11	39.19	57.09	40.89	79.56	46.95	31.51	23.28	64.41	28.68	32.71	46.88	88.89	77.86	14.29	13.88
`jina-reranker-v2-base`	41.45	38.61	51.92	46.60	39.71	36.46	55.39	46.07	79.70	77.18	21.04	20.69	51.02	49.09	30.45	45.86	75.67	72.75	18.94	16.88
`mxbai-rerank-large-v1`	26.10	10.49	13.70	9.42	17.23	18.88	57.89	17.83	68.68	20.26	27.76	11.34	42.15	21.74	20.43	18.02	78.39	67.39	10.88	8.82
`gpt-3.5-turbo`	35.88	34.51	24.71	24.70	38.50	38.65	56.58	54.08	77.71	78.84	27.51	28.43	68.43	68.50	20.50	42.01	83.74	84.86	14.55	14.13

Model	TLDR		SWE-Bench		Apple		ConvFinQA		FinQA		FinanceBench		HC3Finance		TAT-DQA		Trade-event		AY2	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	10.76	-	56.69	-	63.76	-	61.36	-	61.07	-	18.20	-	14.04	-	85.22	-	96.25	-	10.40	-
`contriever-msmarco`	15.59	-	32.58	-	64.48	-	60.63	-	55.40	-	12.96	-	17.75	-	78.99	-	92.12	-	9.04	-
`gtr-t5-base`	15.17	-	40.47	-	62.56	-	51.33	-	38.49	-	19.92	-	19.33	-	74.25	-	93.01	-	12.56	-
`gtr-t5-large`	19.63	-	41.49	-	63.53	-	58.76	-	49.33	-	22.57	-	26.74	-	63.87	-	93.36	-	13.93	-
`all-MiniLM-L6-v2`	15.60	7.39	35.18	32.09	63.02	41.43	40.52	14.83	35.94	30.85	10.71	5.66	11.27	10.16	82.80	37.47	90.44	71.98	9.31	7.48
`e5-small-v2`	12.92	5.54	38.64	35.39	63.39	43.16	69.76	63.97	64.99	61.30	21.66	15.82	17.76	13.20	83.12	83.95	97.03	95.95	13.65	13.05
`e5-base-v2`	15.63	10.17	46.07	46.91	63.26	49.62	70.57	67.05	65.11	59.99	36.16	34.81	22.47	16.51	83.17	83.57	96.21	95.03	12.90	9.64
`e5-large-v2`	17.58	11.34	42.71	37.69	63.36	52.62	71.71	63.01	64.91	67.73	32.19	27.71	22.91	17.90	86.80	82.81	97.80	95.45	12.98	12.15
`gte-base-en-v1.5`	13.53	3.28	49.60	44.96	64.27	26.16	73.54	59.39	75.64	56.11	35.07	25.70	29.44	25.07	80.61	76.86	95.59	95.39	17.88	16.36
`gte-large-en-v1.5`	13.75	2.05	47.93	37.28	63.58	20.62	66.90	45.84	63.39	29.35	27.71	6.35	45.30	33.49	84.34	66.22	96.95	94.58	15.44	12.93
`bge-base-en-v1.5`	11.83	2.02	44.41	41.18	62.27	31.66	61.22	36.93	55.74	36.19	34.87	23.44	20.88	10.78	80.60	76.43	94.98	84.83	13.77	9.27
`bge-large-en-v1.5`	11.13	3.39	42.20	41.49	63.77	23.16	59.52	35.14	58.35	35.69	39.99	13.99	29.97	12.20	81.02	28.57	96.52	88.19	16.85	12.19
`text-embedding-3-small`	19.80	17.15	40.67	33.66	68.38	44.60	67.29	66.71	55.14	59.83	50.39	47.39	28.94	26.31	83.36	83.55	96.89	96.76	15.87	14.40
`e5-mistral-7b-instruct`	24.48	19.68	53.54	51.35	67.05	57.86	69.00	72.78	62.76	66.78	51.49	60.69	43.85	47.96	85.70	85.61	98.23	97.79	12.25	61.44
`NV-Embed-v1`	27.62	27.05	36.92	35.60	68.34	62.89	69.90	70.98	58.66	61.34	40.13	43.62	64.53	67.67	88.49	89.01	92.30	92.95	14.02	28.62
`GritLM-7B`	14.25	24.23	45.98	49.93	61.49	61.38	71.33	74.04	66.77	67.00	52.97	56.27	37.33	53.73	90.17	93.84	97.63	98.46	14.87	34.79
`gte-Qwen2-1.5B-instruct`	17.15	24.88	37.20	20.44	63.95	60.26	73.43	74.36	65.26	68.66	52.03	53.64	31.30	32.80	79.05	83.68	97.26	97.06	13.42	23.77
`monot5-base-msmarco`	25.77	21.33	25.39	17.70	63.74	60.73	64.63	63.96	62.06	63.97	41.25	41.78	24.86	21.05	87.53	79.09	96.75	93.91	10.87	9.17
`bge-reranker-v2-m3`	19.32	15.26	24.14	20.96	65.64	61.04	80.50	71.53	74.42	69.65	47.81	44.92	23.16	22.49	88.20	85.54	96.68	91.28	18.29	22.50
`bge-reranker-v2-gemma`	28.56	15.21	36.90	25.00	66.89	29.63	78.70	56.90	73.92	73.30	62.87	44.43	32.70	14.31	89.18	56.23	97.35	70.12	26.86	24.55
`jina-reranker-v2-base`	22.48	12.65	46.73	43.48	64.27	47.82	61.08	69.19	59.64	68.38	51.52	52.67	24.73	25.20	88.70	81.33	97.19	97.15	22.45	22.13
`mxbai-rerank-large-v1`	30.59	16.28	17.10	8.12	62.96	14.66	62.98	23.79	65.82	52.74	27.91	9.42	28.25	10.23	64.92	28.43	97.13	51.12	9.10	8.93
`gpt-3.5-turbo`	26.58	26.19	38.97	39.19	57.25	61.71	77.04	76.03	76.10	78.58	53.03	54.91	28.09	24.14	92.65	89.70	96.63	97.93	21.31	31.20

Table 6: Results (nDCG@10) on MAIR tasks. The last group represents the re-ranking models, which all re-rank the top 100 results of `text-embedding-3-small`.

Model	ELI5		Fever		TREx		WnCw		WnWi		WoW		zsRE		AILA-Case		AILA-Statutes		BSARD	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	16.22	-	58.37	-	70.80	-	6.24	-	9.11	-	45.67	-	79.84	-	13.20	-	11.03	-	14.17	-
`contriever-msmarco`	45.91	-	90.62	-	84.10	-	7.44	-	9.31	-	58.93	-	96.52	-	10.27	-	9.10	-	10.89	-
`gtr-t5-base`	41.35	-	83.23	-	70.04	-	9.28	-	5.93	-	57.99	-	89.39	-	4.69	-	9.84	-	14.22	-
`gtr-t5-large`	41.28	-	86.75	-	77.34	-	9.71	-	6.13	-	58.62	-	91.97	-	4.83	-	12.96	-	18.58	-
`all-MiniLM-L6-v2`	43.27	27.33	74.50	44.85	71.75	0.43	7.86	6.20	6.66	7.02	54.75	50.82	84.69	0.00	6.54	5.76	11.71	7.24	8.04	1.17
`e5-small-v2`	43.96	42.91	88.16	83.88	86.31	13.87	9.49	7.91	10.03	11.71	59.22	61.69	98.50	10.54	7.60	6.43	10.41	7.52	10.10	7.84
`e5-base-v2`	40.66	41.66	91.09	83.02	87.30	49.68	7.19	9.63	10.82	8.86	60.02	57.62	98.00	51.05	8.27	5.65	10.59	8.38	14.93	12.38
`e5-large-v2`	40.99	36.94	93.12	78.54	87.16	59.93	10.14	10.65	10.61	10.16	64.65	60.77	98.63	53.41	11.85	9.15	8.41	7.75	16.77	10.51
`gte-base-en-v1.5`	41.47	29.09	92.62	54.05	82.11	21.11	9.92	8.81	7.09	6.19	58.69	35.84	96.15	20.54	11.53	9.77	13.74	12.13	9.02	6.57
`gte-large-en-v1.5`	45.59	33.31	93.48	55.39	82.16	5.87	11.86	10.86	9.83	7.95	61.33	42.19	97.63	11.29	9.03	7.54	18.93	18.07	12.90	4.75
`bge-base-en-v1.5`	46.49	32.25	85.63	67.25	85.86	4.32	11.57	6.72	8.58	7.31	65.30	61.59	96.75	0.96	4.73	5.02	11.03	4.53	11.74	8.44
`bge-large-en-v1.5`	44.12	20.41	86.64	29.08	83.48	1.13	13.44	9.00	9.37	8.21	64.93	61.54	97.55	0.00	6.68	4.65	11.07	8.51	14.46	1.96
`text-embedding-3-small`	44.39	42.39	92.87	90.57	75.61	71.20	10.91	10.84	8.66	10.82	67.01	62.38	94.29	84.56	10.44	9.43	13.65	11.72	30.96	21.46
`e5-mistral-7b-instruct`	40.83	52.05	91.78	93.28	83.45	83.14	10.23	23.95	14.25	50.23	60.49	66.54	97.50	96.65	11.62	11.82	22.26	34.95	33.28	31.70
`NV-Embed-v1`	41.88	43.52	92.92	94.19	88.63	86.15	10.68	11.58	15.13	24.77	63.59	65.72	99.13	98.76	8.54	9.32	22.26	27.49	32.89	33.52
`GritLM-7B`	22.99	44.27	68.88	90.35	40.52	83.27	12.56	14.04	13.93	24.83	62.73	66.41	66.66	96.82	10.23	9.99	32.68	38.82	36.63	29.79
`gte-Qwen2-1.5B-instruct`	42.38	46.06	81.28	91.69	63.38	72.90	12.37	11.69	8.10	11.07	60.53	63.86	90.94	92.06	9.30	10.18	20.82	17.42	33.09	34.23
`monot5-base-msmarco`	44.36	42.22	85.59	84.43	76.34	76.00	9.85	9.67	7.14	5.18	54.82	28.08	93.49	91.37	6.42	6.31	9.90	11.57	24.52	21.53
`bge-reranker-v2-m3`	40.43	40.51	92.51	90.49	86.15	85.56	7.83	7.88	11.81	11.83	68.61	68.03	96.79	93.27	6.53	10.48	24.04	22.30	36.91	34.28
`bge-reranker-v2-gemma`	43.17	34.33	93.07	86.15	85.10	73.75	12.86	10.51	16.97	17.31	64.37	57.25	99.63	90.33	9.23	7.48	33.41	34.82	34.56	14.07
`jina-reranker-v2-base`	44.71	40.56	93.38	89.00	87.23	84.77	10.55	11.37	15.60	16.42	61.61	63.16	98.89	95.22	10.68	11.73	16.41	15.96	28.33	23.85
`mxbai-rerank-large-v1`	39.27	28.71	88.57	56.03	84.47	22.75	3.32	3.96	7.89	9.75	45.52	16.14	98.52	32.97	4.89	3.98	9.06	7.33	31.47	7.18
`gpt-3.5-turbo`	46.08	43.79	88.91	82.33	82.36	79.34	12.01	10.20	17.71	27.31	61.65	63.98	97.08	91.25	9.52	8.79	13.38	21.49	34.05	33.14

Model	BillSum		CUAD		GerDaLIR		LeCaRDv2		LegalQuAD		REGIR-EU2UK		REGIR-UK2EU		ArguAna		CQADupStack		FiQA	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	80.23	-	23.03	-	66.29	-	58.98	-	57.73	-	47.31	-	64.97	-	32.77	-	28.87	-	24.82	-
`contriever-msmarco`	67.57	-	21.85	-	3.83	-	47.67	-	37.07	-	40.67	-	44.34	-	32.16	-	29.99	-	35.54	-
`gtr-t5-base`	59.18	-	20.91	-	19.78	-	41.16	-	51.74	-	42.62	-	44.27	-	32.70	-	28.93	-	34.50	-
`gtr-t5-large`	65.49	-	24.57	-	19.40	-	48.40	-	62.61	-	45.02	-	40.69	-	34.26	-	29.82	-	43.28	-
`all-MiniLM-L6-v2`	53.80	21.47	23.03	8.55	10.26	4.32	43.01	36.93	39.66	16.34	41.49	38.92	31.53	31.01	35.66	34.96	37.02	22.78	37.08	18.96
`e5-small-v2`	69.43	58.14	22.92	15.05	8.11	4.93	46.16	44.05	48.19	45.90	45.32	41.75	30.43	32.31	27.49	29.37	34.61	30.96	36.91	36.22
`e5-base-v2`	74.06	68.17	22.94	19.51	10.22	7.91	42.18	41.51	55.18	50.17	52.16	48.83	39.79	41.67	31.07	33.40	38.29	31.68	39.63	35.52
`e5-large-v2`	67.43	58.26	27.94	20.03	12.76	11.93	51.15	47.97	62.88	56.16	47.17	46.83	36.80	36.26	31.67	31.34	38.51	23.77	42.33	36.74
`gte-base-en-v1.5`	73.92	69.56	25.97	21.17	17.46	10.23	50.28	50.32	43.97	37.21	61.61	62.29	43.11	44.95	41.91	40.61	33.99	30.14	47.21	32.28
`gte-large-en-v1.5`	75.31	68.91	28.90	26.87	17.16	6.47	58.35	54.53	52.32	25.71	59.21	58.03	49.19	50.31	51.20	50.89	39.73	15.60	66.01	36.79
`bge-base-en-v1.5`	77.92	69.76	19.81	14.44	2.48	1.46	45.50	22.72	39.89	21.46	56.10	54.19	39.42	43.00	41.81	41.30	35.07	23.36	43.37	20.28
`bge-large-en-v1.5`	78.26	66.30	19.70	11.24	7.42	2.50	52.70	26.41	46.89	3.36	56.98	50.56	47.58	46.27	41.59	41.07	38.25	22.65	48.48	11.63
`text-embedding-3-small`	71.61	61.39	24.65	22.26	39.69	34.18	57.78	58.00	66.31	59.02	58.95	56.88	46.55	46.65	37.18	35.64	44.12	39.84	48.58	43.37
`e5-mistral-7b-instruct`	82.89	82.53	25.82	27.36	40.98	37.25	68.13	68.10	67.67	61.85	75.26	71.67	61.91	65.49	44.11	45.13	41.41	43.18	59.68	62.67
`NV-Embed-v1`	58.22	61.95	26.55	25.29	27.98	25.59	53.74	53.60	71.23	71.15	57.71	57.65	50.91	51.53	49.07	50.23	46.25	48.52	68.98	71.48
`GritLM-7B`	84.17	83.35	15.97	31.05	29.81	31.71	63.87	60.14	58.78	72.67	76.59	75.22	66.08	67.15	40.66	43.39	35.04	50.22	44.10	65.99
`gte-Qwen2-1.5B-instruct`	86.00	85.36	28.14	29.66	38.42	39.90	69.37	65.72	62.67	65.98	69.34	65.79	55.67	59.61	41.70	44.19	37.36	37.66	46.34	51.73
`monot5-base-msmarco`	85.28	83.52	26.38	12.89	18.66	13.59	52.68	48.45	77.94	75.20	47.15	49.30	29.13	28.32	21.62	3.18	40.64	37.13	44.02	38.08
`bge-reranker-v2-m3`	85.72	87.00	36.41	29.83	28.07	24.28	46.01	27.82	81.21	80.73	65.57	44.04	45.72	36.35	34.28	32.12	35.47	33.99	46.26	45.57
`bge-reranker-v2-gemma`	89.66	53.73	34.43	17.64	33.01	14.38	53.49	18.24	80.67	62.23	55.46	37.91	24.72	32.28	59.54	50.79	42.42	33.80	51.97	27.87
`jina-reranker-v2-base`	89.52	83.83	36.00	30.27	29.36	25.72	52.32	48.71	80.48	79.16	63.38	58.80	47.81	48.16	35.62	37.43	36.91	36.69	52.48	37.34
`mxbai-rerank-large-v1`	88.39	33.70	30.40	12.41	17.73	7.32	36.33	21.70	82.02	48.43	41.60	17.62	46.67	17.77	9.39	7.69	37.75	21.35	47.78	19.59
`gpt-3.5-turbo`	81.92	80.93	41.50	43.51	23.96	33.96	58.45	55.88	74.31	73.08	51.94	50.88	48.40	47.95	24.24	24.10	38.34	37.04	43.03	41.79

Model	NFCorpus		Quora		SciDocs		SciFact		TopiOCQA		Touche		Trec-Covid		ACORDAR		CPCD		ChroniQA	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	32.72	-	73.84	-	12.46	-	65.14	-	29.71	-	59.03	-	58.34	-	27.28	-	1.15	-	78.77	-
`contriever-msmarco`	33.94	-	86.62	-	15.77	-	70.34	-	28.77	-	42.55	-	59.64	-	33.51	-	2.09	-	55.74	-
`gtr-t5-base`	32.65	-	88.68	-	12.08	-	58.61	-	30.41	-	48.05	-	56.05	-	24.71	-	1.98	-	39.91	-
`gtr-t5-large`	33.41	-	89.05	-	12.77	-	61.93	-	30.11	-	53.06	-	56.68	-	31.57	-	4.77	-	40.85	-
`all-MiniLM-L6-v2`	31.43	24.01	85.68	44.54	19.68	3.34	67.39	57.86	34.91	34.95	35.06	29.40	46.07	37.24	26.34	15.32	0.26	0.19	36.13	5.89
`e5-small-v2`	32.93	31.91	87.04	86.66	17.61	8.37	70.24	66.30	36.43	34.61	39.49	41.75	65.19	69.28	25.35	18.55	4.71	2.31	56.09	47.41
`e5-base-v2`	35.01	33.72	86.35	83.77	18.25	10.01	73.92	68.10	33.28	30.04	41.10	40.26	60.67	66.81	31.94	18.91	4.60	1.90	52.83	51.03
`e5-large-v2`	37.35	35.76	86.86	85.37	19.56	10.40	76.30	66.33	35.28	34.99	34.80	33.59	55.24	57.36	28.59	23.31	3.98	3.70	59.53	49.61
`gte-base-en-v1.5`	35.85	33.12	89.23	71.93	19.09	6.17	75.22	70.28	36.04	31.60	45.78	33.69	73.06	59.27	27.62	14.86	1.03	0.14	48.87	23.69
`gte-large-en-v1.5`	37.15	34.40	89.49	32.68	24.86	12.38	82.58	78.11	36.35	31.14	46.05	29.61	77.38	56.39	30.33	23.92	1.02	0.30	48.95	26.65
`bge-base-en-v1.5`	35.98	27.84	89.83	27.35	19.79	4.02	74.33	66.67	27.52	24.76	41.40	33.20	67.08	49.59	33.33	18.90	1.03	0.17	38.68	18.79
`bge-large-en-v1.5`	38.12	27.30	89.27	8.10	20.59	2.04	75.24	65.16	29.70	26.29	45.54	8.52	64.50	24.44	29.00	10.27	0.62	0.24	42.52	9.98
`text-embedding-3-small`	37.65	37.23	89.97	87.68	21.28	19.56	73.15	69.64	43.34	47.05	47.52	45.88	78.22	73.92	31.54	28.46	2.66	2.47	47.70	39.94
`e5-mistral-7b-instruct`	39.64	38.88	84.80	88.79	20.72	15.88	76.55	76.69	31.09	39.51	46.80	54.97	77.84	86.81	37.06	37.53	2.97	4.92	53.88	48.35
`NV-Embed-v1`	38.62	39.21	82.66	83.29	21.87	22.52	78.74	78.32	35.72	36.74	45.29	45.35	81.69	82.53	35.08	34.58	2.20	4.06	64.72	65.47
`GritLM-7B`	36.92	40.14	84.92	89.15	19.58	24.55	76.78	79.55	38.41	44.79	21.53	45.39	31.51	66.82	32.85	35.04	6.81	8.18	34.05	61.32
`gte-Qwen2-1.5B-instruct`	34.41	37.06	89.82	89.01	21.49	22.18	76.19	80.41	49.29	50.73	43.16	57.27	54.24	77.67	33.83	31.94	3.19	5.10	55.65	50.29
`monot5-base-msmarco`	35.93	31.26	85.59	84.51	16.72	12.94	73.28	70.81	72.72	65.26	49.14	39.04	84.80	77.72	36.06	38.07	3.78	3.16	70.81	66.58
`bge-reranker-v2-m3`	34.32	35.78	88.53	89.39	18.03	14.95	73.83	71.80	35.54	34.25	56.52	54.67	82.98	78.92	33.67	31.90	2.21	3.21	71.89	70.89
`bge-reranker-v2-gemma`	37.76	26.59	88.29	79.21	19.82	10.08	77.84	58.63	26.36	23.49	58.43	38.18	88.05	71.85	34.26	25.70	1.81	2.54	71.51	55.10
`jina-reranker-v2-base`	36.98	33.42	88.85	56.84	18.23	15.02	78.29	76.56	47.20	56.46	58.81	52.99	84.06	79.43	35.16	33.23	4.51	4.23	71.42	68.23
`mxbai-rerank-large-v1`	38.31	26.40	69.78	29.51	17.97	9.17	76.33	53.19	26.62	22.66	57.83	38.73	87.91	75.78	34.57	13.78	2.07	4.43	65.07	34.51
`gpt-3.5-turbo`	32.66	34.70	75.74	82.63	18.49	18.41	66.73	71.01	43.33	70.70	44.89	44.28	83.66	84.30	31.64	33.31	5.76	5.23	62.34	61.11

Table 7: Results (nDCG@10) on MAIR tasks. The last group represents the re-ranking models, which all re-rank the top 100 results of `text-embedding-3-small`.

Model	Monant		NTCIR		PointRec		ProCIS-Dialog		ProCIS-Turn		QuanTemp		TableSearch		CARE		MISeD		SParC	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	72.51	-	25.28	-	21.08	-	27.73	-	20.62	-	91.82	-	11.31	-	73.55	-	60.31	-	47.28	-
`contriever-msmarco`	72.95	-	16.66	-	43.99	-	37.33	-	26.20	-	87.67	-	13.23	-	51.98	-	63.61	-	75.03	-
`gtr-t5-base`	75.99	-	12.06	-	35.61	-	47.47	-	33.00	-	87.25	-	12.12	-	35.47	-	55.83	-	26.55	-
`gtr-t5-large`	75.27	-	12.07	-	43.43	-	50.13	-	36.37	-	87.30	-	14.05	-	33.51	-	60.80	-	51.90	-
`all-MiniLM-L6-v2`	74.20	70.12	15.06	10.19	27.58	17.97	41.03	19.41	32.02	19.35	79.52	74.10	12.01	7.27	41.09	14.63	57.14	53.14	62.26	18.16
`e5-small-v2`	77.68	75.59	12.88	8.67	45.11	42.47	26.69	7.31	22.28	11.06	88.65	88.39	9.97	8.46	57.97	48.39	54.83	51.81	66.65	32.80
`e5-base-v2`	77.96	73.53	10.62	9.65	46.62	40.75	33.14	13.51	26.83	18.40	87.81	86.19	10.39	11.15	57.57	30.97	59.75	56.68	76.01	68.74
`e5-large-v2`	77.76	71.65	15.21	8.08	47.95	40.37	34.26	18.95	26.77	15.61	88.80	88.65	11.92	10.45	68.91	58.33	59.30	55.90	68.33	58.89
`gte-base-en-v1.5`	77.80	73.85	13.18	8.32	46.15	29.94	50.90	35.47	36.29	32.93	89.55	88.61	11.83	10.81	66.93	55.85	56.18	54.99	59.17	17.45
`gte-large-en-v1.5`	74.40	70.75	15.62	6.98	48.55	48.55	52.75	49.65	38.08	38.59	89.23	79.79	17.89	11.64	65.02	33.28	58.23	53.33	53.02	16.83
`bge-base-en-v1.5`	77.64	69.94	10.55	5.88	40.61	30.51	48.97	12.37	34.71	18.35	88.03	78.92	12.22	9.02	66.25	40.59	57.24	54.32	66.82	32.55
`bge-large-en-v1.5`	78.33	62.96	11.75	6.87	44.07	42.58	55.48	24.86	38.17	23.81	87.97	80.23	12.45	6.97	68.01	31.37	56.85	50.12	68.17	23.48
`text-embedding-3-small`	81.31	75.39	16.73	14.25	49.31	53.96	53.08	50.42	39.89	34.62	91.37	88.87	12.90	12.96	51.85	39.61	60.29	57.43	76.23	66.47
`e5-mistral-7b-instruct`	77.94	77.68	19.07	16.93	54.14	64.98	37.88	61.33	29.50	35.15	90.06	90.31	14.65	14.24	75.72	72.50	62.62	60.15	77.33	77.14
`NV-Embed-v1`	82.11	81.53	21.79	19.76	52.84	58.52	53.32	53.51	36.66	31.90	90.81	90.84	15.34	16.33	45.82	30.68	62.66	61.03	77.28	77.67
`GritLM-7B`	78.47	82.15	3.12	16.56	59.89	67.21	52.17	56.42	35.83	40.51	91.07	92.24	12.34	13.94	61.76	54.16	59.50	61.75	76.83	82.11
`gte-Qwen2-1.5B-instruct`	88.94	82.33	16.13	19.68	48.65	57.97	49.91	60.55	38.74	42.37	92.31	91.52	14.06	16.30	68.27	61.34	61.56	59.08	75.31	69.42
`monot5-base-msmarco`	79.01	70.46	23.24	20.40	38.06	29.71	32.98	44.46	23.54	31.08	90.79	87.76	12.43	13.53	24.32	17.61	58.36	50.45	75.30	69.40
`bge-reranker-v2-m3`	69.58	68.37	19.55	19.38	50.33	23.51	46.82	24.05	27.11	16.17	88.38	87.02	16.60	12.67	45.44	19.69	65.98	57.57	72.50	72.80
`bge-reranker-v2-gemma`	78.58	46.06	20.05	11.85	54.63	60.24	61.89	24.37	38.94	35.89	91.44	66.49	14.24	11.34	43.74	16.74	63.73	44.78	77.69	69.41
`jina-reranker-v2-base`	80.24	66.65	28.98	16.01	52.22	54.28	67.07	62.81	48.31	44.09	91.10	86.98	15.04	13.79	49.07	49.39	61.78	55.33	78.93	78.56
`mxbai-rerank-large-v1`	78.14	21.52	20.66	12.74	43.49	23.33	33.42	24.67	17.21	13.05	90.69	21.82	14.11	11.89	51.55	14.38	61.73	47.73	76.75	5.30
`gpt-3.5-turbo`	72.20	71.15	16.71	17.39	54.45	59.76	52.41	58.97	37.05	34.50	88.84	89.70	13.27	12.68	51.87	47.52	61.24	62.17	58.73	63.83

Model	SParC-SQL		Spider		Spider-SQL		LitSearch		CAsT_19		CAsT_20		CAsT_21		CAsT_22		Core_17		Microblog_11	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	80.83	-	41.91	-	65.45	-	78.41	-	17.41	-	9.97	-	16.43	-	55.32	-	28.40	-	53.42	-
`contriever-msmarco`	71.95	-	60.85	-	54.95	-	52.72	-	24.28	-	24.08	-	21.28	-	64.39	-	41.96	-	54.82	-
`gtr-t5-base`	11.44	-	27.02	-	14.53	-	50.31	-	14.23	-	20.85	-	19.30	-	60.46	-	41.79	-	52.97	-
`gtr-t5-large`	34.34	-	46.81	-	34.84	-	53.84	-	14.19	-	25.89	-	21.15	-	61.94	-	41.03	-	55.08	-
`all-MiniLM-L6-v2`	41.19	27.85	51.81	15.28	22.97	15.81	63.23	46.63	17.98	13.12	19.44	4.57	23.51	12.85	52.98	45.27	37.10	33.15	46.70	2.32
`e5-small-v2`	71.04	47.88	62.38	27.13	50.65	33.50	49.71	29.59	23.48	23.56	22.24	20.43	22.98	22.16	51.92	47.81	44.01	38.75	52.54	15.84
`e5-base-v2`	75.91	73.85	62.27	51.99	55.78	51.13	61.80	34.37	16.13	21.79	19.83	23.53	19.59	19.94	42.29	49.65	40.21	42.21	58.45	22.78
`e5-large-v2`	73.77	56.66	59.63	45.59	56.07	47.04	63.43	47.24	22.13	25.27	23.85	26.46	20.06	19.64	50.91	48.52	43.00	45.84	54.80	16.69
`gte-base-en-v1.5`	51.64	2.11	49.76	23.01	34.70	8.54	68.51	60.15	21.40	18.97	22.63	7.86	20.22	11.47	53.84	41.97	43.28	37.59	54.44	2.48
`gte-large-en-v1.5`	36.72	3.22	44.17	20.85	26.53	6.06	72.30	69.52	22.67	22.04	27.86	12.86	23.00	10.58	56.80	42.70	49.46	46.33	56.53	6.10
`bge-base-en-v1.5`	36.51	21.21	58.52	25.28	33.67	19.31	56.80	35.44	15.46	13.70	18.83	4.42	19.94	10.53	47.06	40.46	36.98	36.92	59.37	1.51
`bge-large-en-v1.5`	48.62	6.48	57.88	20.63	37.21	7.86	67.34	43.48	14.98	9.85	20.22	3.63	20.20	9.28	46.77	37.51	39.69	30.60	58.55	1.17
`text-embedding-3-small`	66.45	54.88	58.51	52.07	48.93	36.35	67.66	58.97	21.26	27.91	28.16	23.73	28.11	25.78	66.03	67.70	47.97	51.45	48.18	41.72
`e5-mistral-7b-instruct`	82.28	79.33	60.33	61.65	65.10	62.62	69.64	78.32	21.42	24.60	23.52	28.04	25.23	28.97	64.35	66.50	42.19	54.28	60.03	59.86
`NV-Embed-v1`	82.30	73.78	61.77	60.81	58.96	56.91	77.02	77.44	20.02	21.82	27.30	29.74	24.77	26.02	62.84	63.08	50.66	55.96	47.37	49.89
`GritLM-7B`	77.51	81.81	55.75	61.57	52.36	59.60	71.97	74.13	16.39	32.92	19.77	34.88	22.70	26.57	43.97	58.34	31.92	55.81	44.63	57.88
`gte-Qwen2-1.5B-instruct`	57.33	51.83	54.74	58.07	44.10	41.66	74.04	76.27	20.52	29.48	23.31	28.64	26.25	27.32	67.95	69.92	44.91	56.16	49.67	47.03
`monot5-base-msmarco`	76.39	71.94	59.92	55.49	59.57	54.95	76.77	70.66	49.72	38.01	40.47	31.06	27.57	20.49	46.88	33.96	53.43	42.51	56.16	54.77
`bge-reranker-v2-m3`	78.55	60.99	64.77	52.68	62.04	43.36	81.47	78.33	28.49	31.57	30.68	30.69	18.70	17.14	30.34	11.21	50.86	50.07	56.90	39.92
`bge-reranker-v2-gemma`	81.01	72.33	60.60	53.73	60.54	52.58	82.03	32.97	16.05	15.79	23.19	14.88	20.03	14.24	32.31	6.23	54.59	52.02	55.75	18.80
`jina-reranker-v2-base`	77.59	77.07	59.89	57.59	58.21	55.21	80.24	73.30	33.54	34.40	32.83	29.73	26.32	24.92	58.41	53.31	53.66	55.97	60.36	51.95
`mxbai-rerank-large-v1`	76.17	1.42	57.50	7.26	60.34	3.03	83.73	11.71	19.71	18.58	25.76	19.79	18.70	13.41	21.24	8.94	57.88	39.66	56.06	17.07
`gpt-3.5-turbo`	59.83	65.32	33.03	51.81	37.88	41.74	70.70	71.28	24.82	24.39	29.82	28.88	25.92	27.25	56.50	59.73	53.40	48.37	49.93	45.91

Model	Microblog12		Microblog13		Microblog14		PM_17		PM_18		PM_19		$PM_{Article}$_19		$PM_{Article}$_20		CliniDS_14		CliniDS_15	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	46.14	-	68.58	-	75.94	-	33.92	-	38.14	-	49.22	-	50.12	-	6.75	-	37.61	-	41.92	-
`contriever-msmarco`	48.78	-	63.62	-	83.07	-	17.26	-	16.86	-	26.99	-	35.13	-	2.90	-	24.92	-	25.27	-
`gtr-t5-base`	50.12	-	60.27	-	81.68	-	12.93	-	15.75	-	27.63	-	37.09	-	1.48	-	30.85	-	24.92	-
`gtr-t5-large`	49.70	-	66.29	-	84.24	-	13.48	-	19.20	-	25.60	-	39.07	-	1.31	-	35.47	-	26.26	-
`all-MiniLM-L6-v2`	42.06	1.69	51.33	2.59	77.05	6.83	15.24	2.16	19.45	0.00	27.59	1.97	44.99	5.74	1.23	0.00	42.92	39.92	33.39	33.34
`e5-small-v2`	46.68	7.56	63.90	25.21	81.05	51.68	13.89	1.31	20.71	2.56	25.73	5.92	40.51	14.86	4.62	1.82	23.86	24.21	23.88	23.12
`e5-base-v2`	45.58	16.87	64.71	33.28	82.30	57.80	20.12	2.07	23.88	1.22	38.38	8.78	42.43	23.89	6.77	0.00	26.13	28.59	24.85	25.49
`e5-large-v2`	47.28	9.29	63.95	29.65	81.88	48.35	20.58	8.88	19.07	8.69	36.32	14.19	44.67	25.07	5.11	0.30	27.31	31.11	24.69	23.72
`gte-base-en-v1.5`	45.22	2.74	64.06	7.11	87.95	15.87	14.44	8.17	19.59	9.09	32.12	21.18	44.25	25.77	2.11	1.64	49.54	47.74	36.90	36.23
`gte-large-en-v1.5`	52.19	6.36	60.18	17.11	82.36	33.20	23.83	5.82	28.01	1.04	39.78	13.03	50.06	22.12	4.86	0.21	57.89	47.39	42.53	43.09
`bge-base-en-v1.5`	50.74	1.00	59.71	1.36	82.02	5.04	22.40	2.23	23.37	2.03	35.51	21.27	46.59	19.21	1.68	0.00	41.26	41.22	33.72	36.42
`bge-large-en-v1.5`	51.35	0.00	66.74	0.58	85.97	9.84	27.89	1.93	32.82	0.83	33.57	4.44	48.70	7.21	4.51	0.78	44.24	43.19	39.00	34.37
`text-embedding-3-small`	50.50	38.64	57.71	54.38	76.55	74.65	22.74	6.37	25.96	11.29	37.83	23.38	46.91	42.13	4.54	4.80	42.20	40.20	29.28	29.90
`e5-mistral-7b-instruct`	48.61	44.76	72.57	71.69	80.37	80.47	22.39	26.96	27.03	35.55	38.29	45.00	53.36	54.26	1.79	8.07	45.97	54.80	42.05	50.77
`NV-Embed-v1`	48.26	47.13	66.26	63.62	71.55	73.32	21.65	23.07	22.69	24.50	24.39	25.47	24.03	25.74	3.32	2.76	43.17	45.28	37.16	40.89
`GritLM-7B`	33.39	45.68	48.26	68.97	74.29	81.66	29.82	33.16	31.00	34.92	44.18	55.36	48.25	51.86	3.58	5.00	52.02	59.59	50.03	56.38
`gte-Qwen2-1.5B-instruct`	40.46	39.26	56.69	57.93	74.93	75.20	22.64	18.92	27.42	24.46	25.55	19.88	54.00	50.91	3.37	4.64	51.35	54.73	40.01	45.49
`monot5-base-msmarco`	51.92	51.87	71.21	72.47	81.69	79.30	28.29	17.20	30.23	23.87	38.28	20.74	39.89	30.58	3.62	7.72	26.37	24.62	20.42	21.15
`bge-reranker-v2-m3`	53.91	38.73	74.99	50.48	76.92	64.32	26.33	5.05	32.33	10.59	44.43	16.68	46.12	29.26	3.74	3.63	38.80	28.54	31.49	25.44
`bge-reranker-v2-gemma`	55.34	18.34	71.36	26.23	76.08	39.37	30.50	11.71	41.76	22.89	46.58	31.14	55.32	34.11	6.37	4.74	55.09	35.17	41.33	24.55
`jina-reranker-v2-base`	53.44	47.28	76.04	60.07	83.18	70.84	30.10	13.68	39.06	23.43	47.82	39.86	48.63	46.68	4.37	3.59	43.60	41.26	36.44	30.54
`mxbai-rerank-large-v1`	60.47	17.82	75.97	18.07	81.93	36.03	31.08	9.89	42.65	14.45	37.23	26.73	48.45	38.55	4.17	4.70	29.92	38.73	29.64	32.04
`gpt-3.5-turbo`	53.91	50.45	61.66	61.72	77.19	72.59	25.91	25.49	29.83	28.88	42.05	39.51	47.02	47.95	6.40	5.67	43.18	45.06	29.46	31.33

Table 8: Results (nDCG@10) on MAIR tasks. The last group represents the re-ranking models, which all re-rank the top 100 results of `text-embedding-3-small`.

Model	CliniDS_16		CliniTrial_21		CliniTrial_22		CliniTrial_23		DD_15		DD_16		DD_17		FairRank_20		FairRank_21		FairRank_22	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	20.79	-	22.95	-	16.70	-	44.86	-	35.31	-	29.33	-	27.22	-	37.19	-	60.13	-	81.42	-
`contriever-msmarco`	6.71	-	22.90	-	21.59	-	27.10	-	35.73	-	28.42	-	35.98	-	40.43	-	52.85	-	80.39	-
`gtr-t5-base`	11.68	-	21.27	-	18.00	-	23.49	-	46.36	-	33.66	-	32.17	-	39.54	-	43.81	-	81.19	-
`gtr-t5-large`	12.26	-	24.10	-	23.78	-	25.27	-	45.33	-	34.99	-	31.65	-	37.01	-	43.22	-	80.00	-
`all-MiniLM-L6-v2`	11.95	9.33	25.25	25.12	22.06	19.64	22.19	5.24	34.98	24.87	28.31	18.89	34.13	32.33	38.00	9.78	51.68	27.66	74.17	24.82
`e5-small-v2`	5.57	5.47	18.05	12.79	16.94	11.57	22.00	14.68	38.19	30.02	34.35	28.62	31.17	29.74	42.86	28.54	41.17	25.82	67.21	27.60
`e5-base-v2`	7.69	7.17	23.60	25.76	21.95	20.18	29.39	24.80	43.51	40.25	34.51	31.46	33.57	32.74	42.42	34.81	32.42	29.40	64.69	61.53
`e5-large-v2`	10.14	10.51	20.67	16.43	15.21	11.00	27.69	19.05	41.06	41.12	35.06	35.62	32.24	32.73	43.04	19.64	21.99	15.64	51.06	51.12
`gte-base-en-v1.5`	21.96	22.41	38.27	32.32	35.09	24.88	26.23	9.28	46.32	38.11	34.94	30.17	33.44	34.80	41.43	16.78	44.48	20.79	78.15	29.18
`gte-large-en-v1.5`	26.47	28.12	39.50	32.57	39.55	28.12	32.92	13.81	45.93	40.25	36.96	32.98	36.57	36.47	43.52	13.34	50.36	21.85	77.27	32.62
`bge-base-en-v1.5`	18.99	16.69	27.39	26.44	26.80	24.55	23.12	16.85	48.53	20.98	32.38	16.69	35.20	33.50	41.42	13.33	44.65	25.07	54.81	29.64
`bge-large-en-v1.5`	17.52	18.96	29.66	26.51	28.20	24.21	23.24	11.77	48.83	18.77	34.45	22.92	35.62	35.46	42.79	4.72	40.76	17.82	73.60	26.31
`text-embedding-3-small`	15.75	19.23	32.22	33.46	26.09	28.68	29.07	25.04	48.24	44.10	31.41	33.27	31.24	31.93	43.53	40.65	52.09	48.78	84.05	79.83
`e5-mistral-7b-instruct`	17.01	32.32	30.46	35.71	31.25	42.79	37.39	40.16	49.92	50.30	33.38	33.73	31.84	35.08	42.27	43.51	57.07	58.26	86.39	84.77
`NV-Embed-v1`	13.65	18.47	29.11	29.70	31.77	33.62	31.02	32.61	46.95	45.02	35.01	31.54	31.97	30.76	43.53	43.93	50.66	43.03	85.70	84.68
`GritLM-7B`	23.08	31.11	38.33	37.23	39.99	40.02	36.78	38.81	51.02	52.28	36.56	36.01	34.01	36.33	39.89	41.57	54.98	58.30	84.26	82.67
`gte-Qwen2-1.5B-instruct`	15.70	22.71	26.37	25.41	26.65	24.00	28.38	24.54	49.37	51.04	33.49	37.36	30.91	31.19	39.96	42.54	53.11	57.04	82.98	85.22
`monot5-base-msmarco`	9.95	11.95	21.67	19.44	29.63	26.13	18.77	15.64	49.44	46.68	34.75	28.45	31.57	31.80	38.41	37.57	38.81	18.30	78.70	68.54
`bge-reranker-v2-m3`	18.11	12.56	23.66	18.24	25.06	22.43	33.50	19.29	61.98	53.39	41.06	35.73	48.16	42.31	40.67	37.25	53.44	45.45	89.69	85.74
`bge-reranker-v2-gemma`	21.64	17.96	34.67	23.48	34.15	24.53	33.33	21.17	63.22	57.46	41.34	36.61	47.49	45.43	40.46	20.99	54.86	50.20	85.00	80.10
`jina-reranker-v2-base`	20.43	19.73	33.21	33.34	32.73	34.08	37.24	31.54	60.85	60.04	41.75	40.48	37.14	40.79	42.68	37.58	61.51	59.84	88.50	88.82
`mxbai-rerank-large-v1`	11.21	12.24	13.73	12.96	7.86	10.17	31.13	20.68	42.35	49.17	33.19	29.42	37.11	37.97	42.88	22.02	44.70	18.57	89.56	66.41
`gpt-3.5-turbo`	15.14	15.77	30.67	31.95	24.67	27.76	28.83	30.37	56.42	40.07	38.86	29.76	44.66	35.95	37.54	36.00	52.43	54.24	89.19	86.28

Model	Genomics_04		Genomics_05		Genomics_06		Genomics_07		TREC-Legal		NeuCLIR-Tech		NeuCLIR_22		NeuCLIR_23		Product_23		ToT_23	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	58.97	-	30.38	-	36.89	-	42.61	-	53.87	-	17.81	-	46.20	-	40.53	-	48.35	-	29.72	-
`contriever-msmarco`	53.76	-	35.30	-	19.88	-	23.61	-	40.12	-	26.79	-	53.32	-	47.46	-	48.00	-	43.41	-
`gtr-t5-base`	47.50	-	35.65	-	16.40	-	20.16	-	45.90	-	21.79	-	49.71	-	44.59	-	45.46	-	36.09	-
`gtr-t5-large`	49.90	-	36.04	-	19.61	-	23.66	-	47.26	-	23.13	-	52.14	-	46.55	-	50.66	-	49.09	-
`all-MiniLM-L6-v2`	55.65	39.82	36.00	20.73	17.96	12.24	20.45	10.50	49.94	15.65	33.20	28.58	52.13	53.09	41.54	45.00	42.24	24.73	29.12	29.23
`e5-small-v2`	56.23	46.55	36.10	29.26	20.32	18.49	25.23	12.40	22.38	12.24	32.97	16.41	55.79	48.88	48.08	42.02	49.61	44.07	26.34	23.79
`e5-base-v2`	55.77	52.06	42.07	36.65	22.47	22.96	26.21	9.90	30.08	20.34	32.62	19.02	54.54	48.76	51.15	44.87	45.79	39.42	30.78	30.12
`e5-large-v2`	59.86	53.28	45.17	38.17	23.13	21.85	30.90	22.05	22.52	10.80	34.82	19.95	54.86	49.80	48.61	47.04	50.25	45.24	40.78	39.01
`gte-base-en-v1.5`	62.20	55.59	45.10	32.30	23.15	18.49	32.71	12.66	53.70	11.66	33.61	23.75	56.84	49.77	47.01	47.84	49.93	37.03	37.39	36.33
`gte-large-en-v1.5`	61.43	59.73	47.25	31.32	24.47	18.09	30.60	13.98	47.01	23.70	42.14	38.01	56.98	57.78	52.15	52.17	47.87	4.76	49.55	42.42
`bge-base-en-v1.5`	60.91	54.78	42.65	27.62	22.82	17.90	29.17	13.43	29.81	29.52	36.67	28.63	55.65	49.23	48.82	45.97	45.94	36.97	25.47	26.90
`bge-large-en-v1.5`	59.42	48.08	41.29	23.05	22.19	10.38	31.51	3.60	49.01	24.68	35.78	33.00	52.27	50.80	46.97	47.26	49.12	15.05	40.58	35.36
`text-embedding-3-small`	59.29	59.87	42.15	40.08	20.67	17.17	27.40	24.64	49.19	9.37	42.20	36.96	55.61	56.89	50.56	52.14	55.77	50.66	52.30	48.68
`e5-mistral-7b-instruct`	58.98	57.89	45.94	46.52	20.92	20.47	33.58	32.03	56.16	54.23	45.56	40.94	55.70	62.90	52.50	56.44	54.05	55.20	56.21	68.15
`NV-Embed-v1`	56.42	55.66	42.88	44.36	20.31	20.90	30.57	32.31	43.19	45.55	42.11	41.52	57.36	61.53	54.32	53.28	54.78	55.12	55.55	59.50
`GritLM-7B`	58.40	58.01	44.11	50.00	23.40	24.33	32.02	31.17	38.15	54.02	41.80	41.11	48.29	62.84	45.73	55.97	43.47	57.36	61.47	69.04
`gte-Qwen2-1.5B-instruct`	56.84	57.28	40.35	41.54	49.51	51.05	45.94	43.19	28.41	45.39	39.02	41.13	53.11	59.91	48.70	56.01	59.12	59.17	53.31	63.90
`monot5-base-msmarco`	55.44	52.11	45.64	41.55	23.64	25.18	28.08	31.23	54.47	46.93	31.80	15.77	52.19	43.55	50.88	37.90	53.18	48.01	23.16	6.15
`bge-reranker-v2-m3`	59.16	56.46	45.72	42.36	24.92	22.21	33.57	28.16	51.13	34.77	29.71	22.12	58.04	52.67	52.80	48.79	53.24	51.48	35.26	12.08
`bge-reranker-v2-gemma`	60.57	56.99	48.93	36.70	26.72	20.67	33.33	18.70	55.60	40.85	40.27	32.28	62.50	59.81	55.01	53.41	54.13	36.89	63.00	18.47
`jina-reranker-v2-base`	61.02	58.53	45.93	46.99	52.34	49.51	54.01	53.20	64.38	52.93	36.68	30.30	61.15	59.70	55.84	55.23	53.30	50.09	61.58	59.34
`mxbai-rerank-large-v1`	52.29	37.92	48.57	33.41	52.38	39.08	58.62	32.65	37.45	16.36	37.50	28.76	58.73	54.44	55.90	47.81	56.01	38.98	33.38	10.33
`gpt-3.5-turbo`	59.07	60.30	45.28	44.61	23.59	23.99	25.99	27.20	52.86	52.30	33.89	34.96	58.34	60.58	52.04	53.14	55.24	52.40	56.40	58.25

Model	ToT_24		FoodAPI		HF-API		PytorchAPI		SpotifyAPI		TMDB		TensorAPI		ToolBench		WeatherAPI		ExcluIR	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	25.46	-	49.92	-	17.96	-	67.05	-	53.98	-	36.12	-	25.43	-	61.28	-	80.92	-	74.84	-
`contriever-msmarco`	40.94	-	50.38	-	29.74	-	56.50	-	53.05	-	33.70	-	40.96	-	63.52	-	73.08	-	81.06	-
`gtr-t5-base`	38.27	-	52.83	-	26.92	-	46.77	-	55.66	-	31.15	-	34.43	-	63.51	-	73.22	-	77.94	-
`gtr-t5-large`	47.75	-	58.02	-	26.32	-	46.58	-	56.77	-	44.60	-	38.58	-	75.00	-	76.92	-	82.33	-
`all-MiniLM-L6-v2`	27.39	21.64	37.78	35.24	23.68	16.18	39.58	36.41	38.25	28.05	35.52	13.50	44.50	37.09	40.08	41.12	75.47	75.08	74.01	58.62
`e5-small-v2`	26.67	16.36	58.74	53.99	24.41	19.34	48.58	43.92	56.07	51.67	40.98	33.72	37.69	41.08	68.68	61.19	78.42	79.04	79.53	77.69
`e5-base-v2`	32.76	19.44	50.74	49.31	27.98	19.14	54.26	53.61	53.46	49.91	42.75	38.82	47.24	40.10	65.61	57.91	74.09	75.00	83.04	81.12
`e5-large-v2`	38.13	23.81	53.20	51.60	31.54	20.77	61.35	57.93	49.67	48.95	42.82	27.83	36.80	41.78	65.38	62.18	72.94	72.15	82.88	81.80
`gte-base-en-v1.5`	37.41	29.78	59.20	56.44	36.94	21.31	56.45	53.93	53.46	51.05	47.45	33.98	39.20	28.34	67.31	63.23	77.86	77.41	82.19	72.83
`gte-large-en-v1.5`	46.95	33.79	60.48	48.47	34.33	31.30	46.72	39.45	55.77	49.89	46.56	34.91	41.63	47.19	74.22	72.52	74.66	67.74	82.01	68.79
`bge-base-en-v1.5`	34.13	25.68	60.98	62.70	29.00	15.37	54.82	50.76	55.48	58.76	44.52	25.74	45.78	34.63	69.07	66.69	80.42	79.20	83.51	77.79
`bge-large-en-v1.5`	39.06	30.48	57.53	57.29	30.81	12.55	59.26	58.25	59.47	50.46	48.72	31.05	38.99	37.85	65.39	60.16	78.00	80.34	82.83	69.02
`text-embedding-3-small`	48.33	41.18	66.19	70.12	37.38	30.65	62.07	56.56	57.29	50.36	45.87	45.61	46.68	45.45	77.91	74.81	76.34	76.83	81.85	81.51
`e5-mistral-7b-instruct`	52.86	60.10	70.41	75.79	40.36	40.84	65.93	65.26	61.69	65.81	58.93	58.69	39.26	53.16	78.39	79.36	80.81	81.15	84.92	86.34
`NV-Embed-v1`	53.70	57.64	66.01	66.07	43.80	45.75	64.07	66.11	58.70	60.55	42.45	43.57	55.85	59.70	79.75	80.66	77.72	78.80	87.19	86.45
`GritLM-7B`	58.69	67.59	67.45	71.10	45.10	46.58	60.28	64.38	63.93	69.91	57.17	58.07	54.03	52.77	81.15	83.88	80.33	78.05	83.12	89.42
`gte-Qwen2-1.5B-instruct`	47.06	56.45	63.74	64.89	39.12	42.38	61.78	65.97	58.69	63.69	56.89	57.20	44.81	50.44	75.34	76.79	76.94	76.49	84.19	83.82
`monot5-base-msmarco`	27.88	12.58	78.41	77.82	38.15	29.55	26.55	20.80	73.41	74.82	51.56	53.17	39.67	34.78	73.33	51.73	74.49	76.68	82.68	81.73
`bge-reranker-v2-m3`	39.05	11.29	69.08	75.52	36.83	25.22	35.74	17.76	70.64	69.77	51.38	46.09	37.48	27.63	81.34	79.30	80.18	78.30	86.63	86.08
`bge-reranker-v2-gemma`	61.96	14.08	70.62	66.03	42.45	16.52	75.21	64.37	70.08	36.02	56.78	47.90	45.09	31.84	87.41	55.00	84.66	77.32	79.97	73.75
`jina-reranker-v2-base`	61.01	55.29	81.42	79.02	37.95	31.28	64.47	64.79	71.86	65.09	56.65	53.13	42.24	35.21	85.19	81.33	79.66	79.20	86.03	86.48
`mxbai-rerank-large-v1`	41.51	7.90	59.49	34.25	32.66	10.72	31.87	13.11	71.98	52.37	51.85	26.55	46.76	29.60	55.09	26.26	76.90	72.01	84.68	51.25
`gpt-3.5-turbo`	50.26	53.37	57.10	66.58	34.71	36.10	71.73	68.42	56.66	59.36	54.15	61.16	40.19	43.13	75.09	78.08	80.57	82.06	82.42	76.43

Table 9: Results (nDCG@10) on MAIR tasks. The last group represents the re-ranking models, which all re-rank the top 100 results of `text-embedding-3-small`.

Model	**Core17**		**News21**		**Robust04**		**InstructIR**		**NevIR**		**IFEval**	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	11.51	-	20.79	-	18.57	-	36.31	-	66.76	-	28.28	-
`contriever-msmarco`	18.18	-	27.33	-	26.64	-	50.45	-	62.97	-	25.51	-
`gtr-t5-base`	22.73	-	22.49	-	26.90	-	49.04	-	63.14	-	22.59	-
`gtr-t5-large`	22.80	-	26.62	-	29.07	-	50.87	-	59.95	-	23.46	-
`all-MiniLM-L6-v2`	20.46	17.50	26.86	25.76	25.96	23.92	48.28	74.06	57.43	41.58	26.06	37.39
`e5-small-v2`	22.84	20.45	28.35	24.02	23.85	19.90	49.82	80.25	62.58	60.99	21.75	36.96
`e5-base-v2`	21.62	21.03	22.92	22.17	24.32	21.09	48.31	84.64	63.96	55.68	24.47	40.33
`e5-large-v2`	22.97	29.12	25.27	26.02	29.03	23.89	49.85	83.76	66.60	62.25	22.49	39.28
`gte-base-en-v1.5`	21.75	21.44	25.63	22.60	27.57	25.52	47.75	86.08	58.48	41.88	25.96	42.18
`gte-large-en-v1.5`	27.61	26.58	25.16	26.22	30.68	29.14	48.93	87.13	58.95	45.56	25.30	41.23
`bge-base-en-v1.5`	19.00	20.63	21.23	20.60	23.22	19.50	52.36	87.25	63.74	48.69	24.42	36.48
`bge-large-en-v1.5`	24.05	18.26	21.41	22.37	27.65	19.16	53.16	85.91	64.42	36.28	22.43	37.65
`text-embedding-3-small`	27.57	30.04	26.22	25.60	32.19	28.15	53.01	86.80	69.06	61.44	21.50	31.91
`e5-mistral-7b-instruct`	21.85	33.41	29.35	31.74	30.44	36.05	42.38	85.30	64.59	62.30	24.98	32.15
`NV-Embed-v1`	26.49	32.81	33.11	30.94	33.04	35.13	49.66	66.14	69.55	69.54	22.45	27.68
`GritLM-7B`	14.47	34.94	21.87	31.03	29.93	38.27	29.89	79.68	59.49	71.53	22.03	38.73
`gte-Qwen2-1.5B-instruct`	21.52	32.65	24.36	31.77	30.06	33.79	50.32	82.87	65.28	65.38	21.85	42.76
`monot5-base-msmarco`	28.99	18.98	26.21	14.77	32.51	24.70	48.47	55.21	76.94	75.70	24.71	31.17
`bge-reranker-v2-m3`	28.17	26.80	27.31	27.06	33.91	31.80	50.55	78.45	81.92	79.06	20.86	21.23
`bge-reranker-v2-gemma`	28.28	29.02	31.41	27.62	39.11	30.19	52.99	90.40	80.84	65.00	19.42	53.73
`jina-reranker-v2-base`	26.68	31.53	29.13	30.20	32.99	34.68	48.87	78.45	79.29	75.07	11.17	20.80
`mxbai-rerank-large-v1`	32.60	25.82	30.64	14.58	33.05	19.53	49.45	86.81	83.53	48.47	17.13	27.74
`gpt-3.5-turbo`	29.09	33.43	29.62	30.27	29.06	29.73	52.22	78.64	76.96	72.50	15.15	29.04

Table 10: Results of each sub task in SWE-Bench, CQADupStack, and IFEval.

Task	SWE-Bench																			
Sub-Task	**astropy**		**django**		**matplotlib**		**mwaskom**		**pallets**		**psf**		**pydata**		**pylint-dev**		**scikit-learn**		**sphinx-doc**	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	41.08	-	69.01	-	34.98	-	76.95	-	74.11	-	60.27	-	58.86	-	55.51	-	51.68	-	48.50	-
contriever-msmarco	36.66	-	55.60	-	19.46	-	24.45	-	51.04	-	43.86	-	30.10	-	29.30	-	37.73	-	13.90	-
gtr-t5-base	48.85	-	64.13	-	16.13	-	25.87	-	56.41	-	65.62	-	33.72	-	28.46	-	37.54	-	50.13	-
gtr-t5-large	52.69	-	57.08	-	16.41	-	31.84	-	60.92	-	55.43	-	35.79	-	36.10	-	42.01	-	39.22	-
all-MiniLM-L6-v2	36.33	33.47	61.65	53.21	17.90	17.69	21.99	27.21	62.14	58.69	35.26	36.33	31.23	26.02	27.47	26.77	41.26	38.26	30.12	16.34
e5-small-v2	45.01	42.40	64.01	62.60	23.68	20.18	16.18	22.66	51.60	51.82	40.97	36.79	35.25	37.75	30.85	24.20	44.10	34.46	48.07	47.15
e5-base-v2	47.35	62.63	75.06	81.33	33.28	34.48	14.30	25.47	64.68	66.82	60.25	57.81	49.72	50.48	34.99	31.71	56.44	49.78	51.95	46.11
e5-large-v2	48.14	42.52	64.00	66.00	25.00	28.49	50.15	32.77	57.47	51.49	42.73	23.97	28.68	24.83	37.82	36.75	48.26	46.79	41.97	27.37
gte-base-en-v1.5	54.81	58.98	68.48	66.49	30.63	27.74	28.49	18.83	79.37	66.90	48.90	49.34	51.39	50.58	51.09	46.20	54.01	38.63	45.97	39.09
gte-large-en-v1.5	54.75	40.84	67.80	62.08	27.74	22.81	36.22	23.11	75.13	73.04	48.44	49.00	56.54	35.46	35.19	34.19	67.06	31.30	48.18	39.32
bge-base-en-v1.5	53.71	57.13	61.00	65.82	29.55	30.62	30.35	19.87	60.66	55.53	43.23	26.05	43.69	39.79	34.41	30.99	58.14	60.07	51.73	42.16
bge-large-en-v1.5	55.38	57.58	68.05	68.02	17.55	16.14	15.78	15.41	59.79	50.02	52.82	46.43	30.04	33.62	34.79	34.94	54.93	55.34	52.06	52.41
text-embedding-3-small	51.70	46.43	45.95	37.75	21.37	19.89	22.51	21.60	63.47	44.69	46.36	41.93	45.52	36.90	36.60	34.04	36.96	24.70	40.70	37.02
e5-mistral-7b-instruct	63.52	54.50	70.08	57.13	32.85	38.86	34.19	35.92	76.15	75.93	43.94	40.42	48.46	41.79	50.40	57.80	67.55	64.93	60.69	55.83
NV-Embed-v1	41.60	39.52	45.68	42.94	22.64	27.33	32.88	27.50	66.29	66.24	32.89	34.20	25.94	22.44	37.34	32.02	48.77	50.71	28.07	27.37
GritLM-7B	58.34	62.96	62.62	74.61	19.13	34.25	33.32	47.87	60.50	67.43	34.01	42.39	43.36	34.64	47.17	39.25	63.54	69.27	57.52	54.43
gte-Qwen2-1.5B-instruct	42.50	15.87	46.77	33.56	13.56	11.31	8.77	6.13	70.20	40.99	32.62	15.45	40.44	17.72	42.22	23.09	35.01	13.87	37.70	28.34
monot5-base-msmarco	18.13	19.55	19.21	11.02	13.20	19.11	15.36	16.88	47.35	26.64	39.91	24.53	28.90	21.82	16.77	15.39	27.29	9.18	36.75	20.46
bge-reranker-v2-m3	17.88	15.84	7.89	24.74	14.03	15.00	17.13	17.82	29.74	27.22	32.18	33.12	13.10	16.81	22.18	16.41	55.69	27.14	38.51	14.97
bge-reranker-v2-gemma	41.96	30.39	45.27	37.08	10.95	28.28	25.77	21.38	49.29	22.03	37.51	35.81	32.00	22.10	29.84	10.97	66.94	28.79	45.33	29.23
jina-reranker-v2-base	61.72	58.95	57.58	60.86	20.09	26.51	57.52	35.83	59.24	56.49	42.13	38.86	41.76	34.75	36.65	34.25	56.46	55.17	46.22	41.36
mxbai-rerank-large-v1	28.12	10.40	22.04	6.13	3.87	3.33	14.03	14.20	16.27	8.35	9.20	21.77	15.41	7.12	12.05	3.22	25.12	5.26	30.04	9.61
gpt-3.5-turbo	40.75	40.29	69.45	70.86	33.20	30.00	25.17	18.89	66.38	66.63	56.08	54.92	41.80	15.68	29.16	21.41	17.93	46.72	30.49	45.30

Task	SWE-Bench	CQADupStack																		
Sub-Task	**sympy**	**Android**		**English**		**Gaming**		**Gis**		**Math**		**Physics**		**Programmers**		**Stats**		**Tex**		
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	53.70	-	35.28	-	33.87	-	68.49	-	39.20	-	0.00	-	22.44	-	31.01	-	12.06	-	23.42	-
contriever-msmarco	19.19	-	38.95	-	29.14	-	45.09	-	42.62	-	5.32	-	12.89	-	55.70	-	14.71	-	24.48	-
gtr-t5-base	29.15	-	36.67	-	42.44	-	39.83	-	40.00	-	1.91	-	20.00	-	44.00	-	15.31	-	31.31	-
gtr-t5-large	33.75	-	41.36	-	30.18	-	43.23	-	37.46	-	9.23	-	17.32	-	43.10	-	12.46	-	30.67	-
all-MiniLM-L6-v2	28.60	23.74	68.86	49.41	26.49	17.62	45.96	19.34	60.83	30.44	6.80	6.02	24.38	20.00	63.39	29.94	26.33	14.41	23.87	16.31
e5-small-v2	32.29	19.31	56.59	45.72	34.20	31.31	36.28	49.97	60.26	53.95	4.19	0.91	12.43	19.46	68.56	55.52	18.26	7.61	30.18	25.00
e5-base-v2	28.75	23.03	53.07	44.61	31.51	33.32	51.22	33.56	55.65	49.46	8.90	8.39	16.31	13.15	69.75	54.77	20.39	22.26	24.31	27.20
e5-large-v2	30.05	34.44	63.63	31.08	34.81	32.44	43.96	29.56	62.20	43.16	7.82	3.56	16.31	16.31	78.82	28.08	14.84	7.70	28.61	10.00
gte-base-en-v1.5	31.14	30.64	49.53	50.91	26.31	21.41	42.46	46.01	46.31	34.46	13.44	8.03	10.09	0.00	60.15	50.81	30.93	24.29	19.46	19.20
gte-large-en-v1.5	21.59	1.68	51.99	19.83	44.56	20.00	42.25	9.69	57.88	13.33	12.82	0.00	15.62	17.88	68.07	19.95	30.04	14.65	28.95	12.62
bge-base-en-v1.5	31.03	34.09	45.28	26.31	31.37	21.13	45.86	20.47	56.66	41.04	8.99	9.54	8.33	7.12	57.50	38.56	26.35	16.54	19.87	16.31
bge-large-en-v1.5	29.67	32.39	50.56	18.18	43.56	19.87	44.24	13.56	60.79	49.48	14.40	7.25	18.33	20.00	66.26	33.93	25.04	17.36	25.91	14.31
text-embedding-3-small	39.93	25.00	65.94	66.97	42.79	35.09	64.55	54.98	54.87	50.62	15.20	15.79	29.33	28.32	78.86	63.56	30.04	24.34	19.87	20.63
e5-mistral-7b-instruct	43.91	35.91	51.34	59.36	44.71	46.20	61.68	59.68	69.46	62.80	15.13	12.14	18.15	18.56	56.29	66.54	24.34	22.77	35.57	37.36
NV-Embed-v1	23.63	24.50	59.52	71.27	42.33	47.02	57.76	60.47	74.31	76.49	16.11	18.68	29.87	29.64	75.94	72.45	32.08	33.10	14.31	17.04
GritLM-7B	25.21	31.78	51.93	70.31	16.03	46.10	40.24	64.87	62.87	78.33	8.28	21.07	22.66	31.02	55.41	79.12	13.99	29.41	28.97	37.04
gte-Qwen2-1.5B-instruct	34.85	16.14	58.52	49.94	40.00	40.61	53.09	52.89	57.07	60.95	8.42	13.26	19.38	15.65	60.72	60.30	21.97	22.08	31.31	29.87
monot5-base-msmarco	24.21	12.19	48.55	53.30	42.89	43.01	68.19	57.61	65.00	58.61	7.99	4.31	16.13	16.13	61.07	56.95	26.35	23.37	37.11	37.15
bge-reranker-v2-m3	18.99	25.53	50.89	48.96	34.64	40.62	67.57	60.91	56.31	56.31	12.32	8.61	16.13	12.44	44.94	44.25	11.98	13.90	36.18	32.26
bge-reranker-v2-gemma	27.37	21.49	58.13	45.20	46.31	39.32	61.60	59.71	66.18	49.68	13.06	6.99	24.29	20.00	55.04	51.30	25.04	17.95	38.33	26.31
jina-reranker-v2-base	43.73	43.59	41.80	54.65	40.18	32.62	54.78	56.40	59.46	56.18	12.60	9.42	20.44	23.87	47.58	59.09	19.61	16.11	37.89	23.87
mxbai-rerank-large-v1	16.47	4.31	59.09	36.18	35.70	24.55	57.70	18.78	58.33	35.18	16.11	17.71	15.00	20.18	45.84	24.76	24.08	7.47	33.87	9.85
gpt-3.5-turbo	27.09	36.37	55.66	46.65	38.59	46.31	53.23	47.64	63.51	73.93	9.75	7.64	22.36	30.60	60.82	73.21	27.01	24.40	24.31	14.32

Table 11: Results of each sub task in SWE-Bench, CQADupStack, and IFEval.

Task	CQADupStack						IFEval													
Sub-Task	**Unix**		**WebMasters**		**Wordpress**		**format**		**keywords**		**punctuation**		**change_case**		**length_constraints**		**combination**		**content**	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	32.80	-	31.60	-	16.31	-	26.81	-	26.67	-	42.69	-	42.85	-	20.91	-	33.11	-	34.10	-
contriever-msmarco	46.18	-	19.83	-	25.00	-	19.87	-	22.40	-	29.02	-	39.50	-	22.20	-	24.50	-	32.82	-
gtr-t5-base	41.33	-	24.31	-	10.00	-	18.85	-	21.27	-	32.62	-	22.13	-	19.07	-	33.82	-	31.01	-
gtr-t5-large	47.52	-	18.72	-	26.57	-	19.33	-	17.93	-	34.26	-	25.36	-	21.29	-	32.31	-	31.73	-
all-MiniLM-L6-v2	55.63	33.63	32.50	30.00	9.20	6.31	22.42	40.41	19.33	57.56	30.13	33.27	39.31	27.34	21.75	24.56	26.03	33.95	32.69	42.10
e5-small-v2	28.74	32.18	40.59	32.64	25.00	17.20	18.22	31.97	10.43	63.87	36.18	36.34	36.28	30.03	18.37	34.38	23.91	26.51	26.14	38.95
e5-base-v2	59.31	33.33	44.78	32.64	24.31	27.46	21.75	36.57	13.01	64.93	36.19	33.80	39.29	30.31	19.09	32.13	19.82	26.08	28.14	44.45
e5-large-v2	53.87	28.18	38.32	38.33	18.87	16.88	17.23	33.94	19.73	65.39	37.78	33.50	36.38	37.36	17.19	32.84	23.32	25.22	33.19	45.90
gte-base-en-v1.5	51.31	44.08	35.70	46.13	22.20	16.31	23.83	51.98	19.08	49.93	32.32	20.50	28.70	27.03	20.37	24.15	25.52	26.84	31.71	49.29
gte-large-en-v1.5	57.62	35.00	40.41	19.31	26.62	5.00	22.58	40.72	17.85	50.29	32.78	40.37	28.12	34.55	21.62	25.67	23.12	25.45	33.35	53.54
bge-base-en-v1.5	44.38	22.88	45.18	40.73	31.07	19.64	21.11	44.40	22.68	54.94	33.71	22.29	28.48	23.68	20.45	23.74	23.39	26.75	30.20	46.03
bge-large-en-v1.5	48.48	35.30	35.59	29.95	25.81	12.62	17.88	40.04	18.76	57.36	31.77	20.72	31.58	20.38	18.89	27.19	19.56	25.70	30.54	51.96
text-embedding-3-small	67.62	61.55	36.88	33.87	23.51	22.35	23.47	34.09	23.11	48.87	32.44	33.47	20.28	17.98	16.94	23.08	13.41	20.78	28.73	38.65
e5-mistral-7b-instruct	60.77	53.93	38.20	48.69	21.31	30.18	23.04	32.74	11.51	58.66	31.98	33.90	31.34	21.96	14.60	22.76	38.96	24.51	30.24	32.65
NV-Embed-v1	68.93	72.62	46.23	53.39	37.64	30.09	21.65	25.74	22.51	42.90	28.55	37.48	27.25	27.58	17.89	19.91	17.77	24.17	31.32	36.96
GritLM-7B	62.88	68.80	45.06	59.81	12.20	16.74	20.75	34.36	5.51	65.11	42.05	33.82	29.92	40.99	12.87	29.17	32.33	49.73	23.13	45.11
gte-Qwen2-1.5B-instruct	48.79	56.49	40.44	36.95	8.56	12.87	18.13	44.52	14.95	58.09	32.27	32.59	28.90	34.94	16.40	23.86	24.68	42.86	25.95	56.47
monot5-base-msmarco	49.93	41.33	35.03	28.18	29.46	25.62	22.18	24.75	20.54	44.52	20.97	26.82	32.71	44.53	25.06	26.08	29.90	22.54	33.89	32.23
bge-reranker-v2-m3	50.21	43.91	20.00	23.01	24.48	22.74	19.62	13.89	19.95	38.65	32.48	27.92	21.82	13.10	21.96	28.75	17.89	12.51	31.01	23.39
bge-reranker-v2-gemma	54.48	41.43	35.80	29.64	30.77	18.05	22.09	46.06	16.68	74.87	23.97	33.63	16.45	22.64	18.02	39.14	17.29	79.48	26.14	89.00
jina-reranker-v2-base	46.09	51.32	29.66	32.37	32.80	24.32	11.55	24.03	4.01	31.94	12.07	20.72	15.91	8.69	17.05	26.22	17.32	18.45	7.53	21.08
mxbai-rerank-large-v1	47.49	29.64	36.05	28.02	23.77	3.87	21.38	34.19	9.76	35.70	15.37	16.52	6.42	20.66	20.11	32.83	25.37	12.26	23.88	27.02
gpt-3.5-turbo	61.19	39.64	26.49	25.71	17.20	14.46	17.59	30.77	18.99	46.48	15.21	31.70	10.23	21.13	18.85	30.91	5.85	16.79	16.45	32.19

Table 12: Results of each sub task in SWE-Bench, CQADupStack, and IFEval.

Task	IFEval	
Sub-Task	**startend**	
Instruction?	✗	✓
BM25	14.55	-
contriever-msmarco	26.57	-
gtr-t5-base	20.55	-
gtr-t5-large	22.68	-
all-MiniLM-L6-v2	26.62	40.72
e5-small-v2	18.86	43.20
e5-base-v2	27.78	59.70
e5-large-v2	14.72	48.01
gte-base-en-v1.5	34.70	62.87
gte-large-en-v1.5	31.31	61.94
bge-base-en-v1.5	27.17	33.74
bge-large-en-v1.5	23.24	48.42
text-embedding-3-small	14.01	37.18
e5-mistral-7b-instruct	35.03	35.23
NV-Embed-v1	17.17	19.75
GritLM-7B	28.59	31.65
gte-Qwen2-1.5B-instruct	28.39	53.10
monot5-base-msmarco	16.90	38.37
bge-reranker-v2-m3	7.83	22.97
bge-reranker-v2-gemma	11.29	72.85
jina-reranker-v2-base	2.29	7.08
mxbai-rerank-large-v1	7.27	16.82
gpt-3.5-turbo	7.88	16.35

Table 13: Results (nDCG@10) on IFEval sub task. AVG means average results.

Model	**AVG**		**detectable_format**		**keywords**		**punctuation**		**change_case**		**length_constraints**		**combination**		**detectable_content**		**startend**	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	28.28	-	26.81	-	26.67	-	42.69	-	42.85	-	20.91	-	33.11	-	34.10	-	14.55	-
text-embedding-3-small	21.50	31.91	23.47	34.09	23.11	48.87	32.44	33.47	20.28	17.98	16.94	23.08	13.41	20.78	28.73	38.65	14.01	37.18
e5-mistral-7b-instruct	24.98	32.15	23.04	32.74	11.51	58.66	31.98	33.90	31.34	21.96	14.60	22.76	38.96	24.51	30.24	32.65	35.03	35.23
NV-Embed-v1	22.45	27.68	21.65	25.74	22.51	42.90	28.55	37.48	27.25	27.58	17.89	19.91	17.77	24.17	31.32	36.96	17.17	19.75
GritLM-7B	22.03	38.73	20.75	34.36	5.51	65.11	42.05	33.82	29.92	40.99	12.87	29.17	32.33	49.73	23.13	45.11	28.59	31.65
gte-Qwen2-1.5B-instruct	21.85	42.76	18.13	44.52	14.95	58.09	32.27	32.59	28.90	34.94	16.40	23.86	24.68	42.86	25.95	56.47	28.39	53.10
bge-reranker-v2-gemma	19.42	53.73	22.09	46.06	16.68	74.87	23.97	33.63	16.45	22.64	18.02	39.14	17.29	79.48	26.14	89.00	11.29	72.85
RankGPT																		
gpt-3.5-turbo	15.15	29.04	17.59	30.77	18.99	46.48	15.21	31.70	10.23	21.13	18.85	30.91	5.85	16.79	16.45	32.19	7.88	16.35
gpt-4o-mini	-	52.68	-	61.60	-	65.96	-	38.24	-	52.70	-	37.70	-	35.94	-	54.82	-	53.24
gpt-4o	-	86.09	-	95.43	-	82.29	-	51.99	-	84.83	-	70.70	-	87.04	-	92.40	-	98.66