MINT: A Benchmark for Evaluating Instructed Information Retrieval

Anonymous EMNLP submission

Abstract

Large Language Models (LLMs) have demonstrated powerful capabilities in following human instructions to tackle diverse, complicated tasks. In the Information Retrieval (IR) field, some recent studies investigate IR instruction-tuning over LLMs to explore the development of advanced LLM-based IR systems. However, there are almost no capable benchmarks that can effectively assess those LLM-based IR systems as well as those traditional IR systems upon following human instructions when retrieving documents. Especially, some existing benchmarks like BEIR or KILT have only a dozen or so tasks and most of them are relatively homogeneous, making them usually fail to provide a thorough assessment. Therefore, to address the above problem, we propose MINT (Massive Instructed Text Retrieval Benchmark), a heterogeneous IR benchmark, including 100 distinct IR tasks in 6 domains collected from a massive range of sources. To comprehensively assess the instruction-following ability of different IR systems, we recruit experts to manually annotate 776 detailed and diverse instructions. Based on MINT, we benchmark the state-of-the-art instruction-tuned text embedding models and re-ranking models. Our experiments reveal that instruction-tuned models usually achieve superior performance compared with those Non-instruction-tuned models on MINT. Besides, it can also observed that adding the instructions can lead to the performance improvement of most instruction-tuned models across most tasks. Our results also suggest that current instruction-tuned text embedding models or re-ranking models still lack effectiveness in specific long-tail tasks. And MINT will be publicly available at github.com (hidden for blind reviewing).

1 Introduction

Large Language Models (LLMs) have shown impressive ability in following user instructions to perform a wide range of natural language processing (NLP) tasks (Ouyang et al., 2022). This capability mainly stems from fine-tuning the LLMs on a large set of human-annotated or automatically generated instruction-following data (Taori et al., 2023; Chiang et al., 2023). By doing so, these LLMs gain strong generalization ability and achieve remarkable zero-shot performance on any new tasks and domains. The exciting result of instruction-tuned LLMs has also attracted a lot of attention in other fields including the Information Retrieval (IR) community (Su et al., 2022b; Asai et al., 2022b).

IR technique aims to find a small set of candidates from a large corpus based on relevance or other query requests (Yates et al., 2021; Fan et al., 2022). It is widely regarded as a crucial part of many AI applications, from Web Search (Zhu et al., 2023) to various retrieval-augmented tasks (Gao et al., 2023). Traditional IR models are commonly trained on a single task and domain, thus exhibiting poor generalization to other IR tasks and domains (Zhao et al., 2024). For example, most advanced IR models are trained on MS MARCO (Nguyen et al., 2016) which is the largest training dataset in IR for now containing tens of thousands of queries and millions of web documents. These IR models have been shown to perform poorly on out-of-distribution dataset (Thakur et al., 2021).

Inspired by the success of instruction-tuned LLMs, recent work has explored directly applying these instruction-tuned LLMs for IR (Sun et al., 2023a) or training a unified IR model across multiple datasets with instructions (Oh et al., 2024a; Weller et al., 2024b). They claim that such models have a significant improvement in aligning users' intents. However, most instructions are only simple task descriptions like "Retrieve a Wikipedia paragraph that answers this question" for the Natural Questions dataset (Kwiatkowski et al., 2019). It's unclear whether these models are truly unified, generalized, and follow user instructions strictly to

yield more aligned search targets.

Existing benchmarks such as BEIR (Thakur et al., 2021), KILT (Petroni et al., 2020), InstructIR (Oh et al., 2024b), FollowIR (Weller et al., 2024b) and MTEB (Muennighoff et al., 2022), collected a series of different IR tasks to evaluate their generalization ability, leading to extensive research on general IR models. However, these benchmarks (1) either contain a relatively small number of tasks or (2) the tasks are too similar and have a small coverage of the broad IR scope. As shown in Table 1, FollowIR and InstructIR only contain 3 and 1 tasks, respectively. And the instructions in InstructIR are LLM-synthetic rather than real instructions. BEIR and KILT contain more tasks, 18 and 11, respectively. But the query-doc relevance criteria for these tasks are homogeneous, .

To address the above issues, we introduce MINT, a large-scale IR benchmark consists of 100 diverse retrieval tasks with 776 distinct instructions for evaluating of model generalization to unseen tasks. The tasks in MINT are collected from exiting IR datasets, such as those in (i) SIGIR resource track papers, (ii) tasks in exiting benchmarks, (iii) publicly accessible shared tasks in TREC (`trec.nist.gov`), and (iv) recent LLM benchmarks. These tasks cover a very wide range of information retrieval application scenarios, including very diverse query and document types, and unique relevance criteria. In MINT, We have selected representative tasks from them, which constitute a benchmark consisting of 100 unique tasks. We cleaned, merged, and sampled these data that ultimately includes 8,839 queries, and 5,232,822 documents. We validated that the sampled benchmark obtained highly correlated results with full testing at a smaller scale, achieving a trade-off between evaluation accuracy and evaluation cost. Then, we manually annotate instructions for these IR tasks. Following previous work, the instruction specifies the types of queries and documents, as well as the relevance criteria. We finally annotated 774 distinct instruction, which representing 774 different query-document-relevance combinations. Some tasks include query-level instruction, i.e., each query has a different relevance criterion. These annotations make MINT especially challenge and suitable for evaluating instruction-tuned retriever in terms of whether they can follow instruction to complete unseen tasks.

Based on MINT, we benchmark various different types of retrieval models, including sparse retriever, single-task text embedding models (Ni et al., 2021; Izacard et al., 2021), non-instruction-tuned multi-task text embedding models (Li et al., 2023b; Xiao et al., 2023; Wang et al., 2022a), instruction-tuned embedding models (Wang et al., 2023a; Lee et al., 2024), and re-ranking models (Sun et al., 2023b). We evaluate on both *no instruction* and *with instruction*, and found instruction-tuned embedding models shows clear improvement when instructions are added. `e5-mistral-7b-instruct` achieves the best overall score, with an average nDCG@10 of 50.14. Furthermore, both `e5-mistral-7b-instruct` and `NV-Emebd-v1` show notable improvement when instructions are added.

In conclusion, this paper makes the following contributions: (i) We introduce MINT, the largest and most comprehensive IR benchmarks with 100 distinct tasks and 776 annotated instructions. (ii) We benchmark many retrieval models, including both sparse retrieval, text embedding, and LLM-based rerankers. (iii) Our experiment shows that instruction-tuned models achieve the best results on MINT. (iV) We will open-source the datasets and evaluation toolkit.

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 100 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

	MINT (this work)	BEIR Thakur et al. (2021)	KILT Petroni et al. (2020)	FollowIR Weller et al. (2024a)	InstructIR Oh et al. (2024b)
Number of tasks	100	18	11	3	1
Number of domains	6	4	1	1	1
Number of instructions	776	14	5	104	9,906
Number of test queries	8,839	47,233	50,736	104	9,906
Number of collections	396	18	1	3	1
Total number of docs	5,232,822	38,506,129	5,903,530	98,312	16,072

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

from the following four well-known sources, Summarized in Table 2:

- **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. And the list of articles is in the AppendixA.2

- **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 tailored to specific applications, enriching the task diversity in our benchmark. 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 25 tasks and across 5 domains from TREC tracks. The specific track names and years are listed in AppendixA.1

- **LLM Evaluation Datasets**: In addition to the above datasets, we also integrate several public LLM evaluation datasets released in the LLM era such as SWE-BENCH (Jimenez et al., 2023a). Including those datasets can help extending our benchmark to diverse potential instruction-following scenarios. Finally, We collect 10 tasks across 3 domains from the LLM Evaluation Datasets.

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

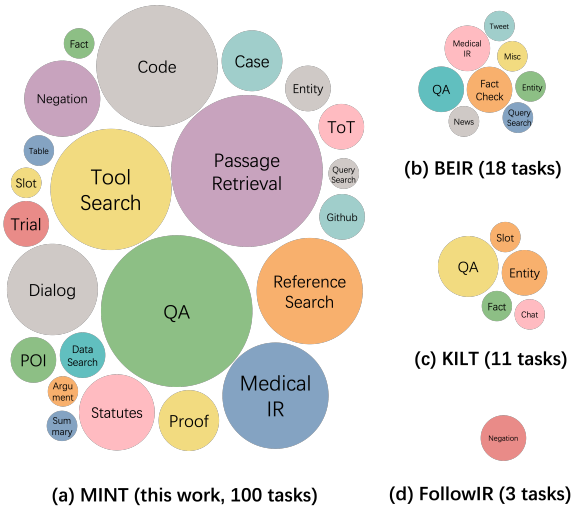

Figure 1: Compared to other datasets, MINT covers a more diverse types of task. Bubble size represents the number of tasks of each type.

Finally, we collect 100 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 100 distinct tasks, we find the data distributions are quite imbalanced among different tasks. Besides, since the scales of some

	# Task	# Domains	# Docs
IR Resources	10	2	298,386
IR benchmarks	55	6	2,771,794
TREC Tracks	25	5	2,542,369
LLM Datasets	10	3	426,002

Table 2: Data statistics of different data collection sources.

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.

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 8,839 diverse queries and 5.2 million documents (500 million tokens based on cl32k)[2], providing a computationally efficient yet representative test-bed for IR models.

[2]The final query number is less than 10,000 because some tasks (e.g., tasks from TREC Tracks) may partly share the same queries

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. (2022a), the basic format of the instruction describes the query, the passages, and the relevance criterion. For example, instruction of `CodeEditSearch` (Muennighoff et al., 2023) is "*Given a GitHub commit message, retrieve the relevant code diff files associated with this commit and implement the changes described.*", and Appendix A list more instruction examples.

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 8,839 queries from 100 distinct tasks, with 776 instructions annotated for each task. There are in total 396 document collections and 5,232,822 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.

3 Dataset Analysis

The table lists the full list of the tasks in MINT and their input / output, and task type and domain. Tasks in MINT mainly come from six domains: (i) Web refers to retrieving information from the general web. 20 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 12 tasks, ranging from code retrieval to tool retrieval, and to GitHub 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

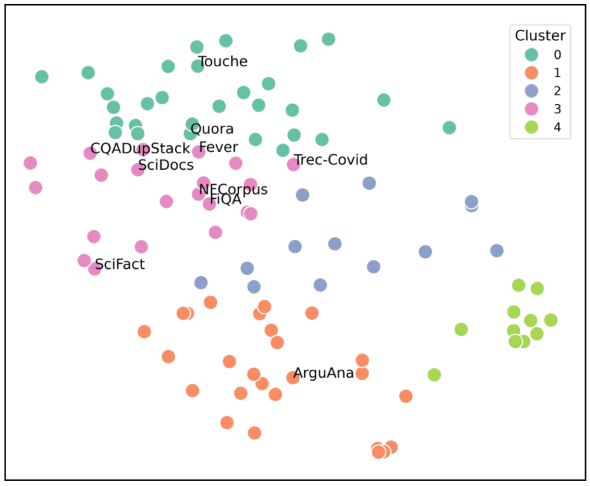

Figure 2: Visualization of the correlation among 100 tasks in MINT, with annotations for tasks from BEIR. MINT includes more diverse tasks. Task similarity is determined based on the performance correlation of all baseline models. We employ KMeans for clustering and t-SNE for visualization.

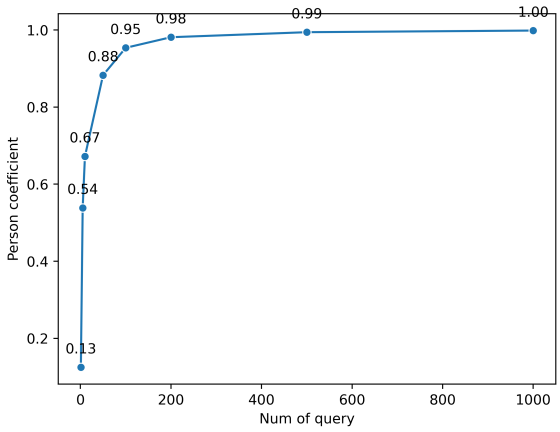

Figure 3: The performance correlation of 8 baseline models with different sampled numbers of queries. Sampling 100 queries achieves a good trade-off between correlation and cost.

in TREC-CDS, which wide range of retrieval and matching tasks in biomedical.

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

Task Correlation To measure the task diversity of MINT, 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 MINT, we calculated the correlation between all tasks, and got a correlation matrix in $\mathbf{M} = \mathbb{R}^{100\times100}$, $\mathbf{M}[i, j]$ denote correlation between task i and j. We plot this matrix in Figure 6. To better visualize the matrix, we use t-SNE to visualize the task correlation matrix in 2D.

Figure 2 illustrates the t-SNE visualization of the correlation matrix, where each task is represented as a point, and the distance between two tasks approximates their correlation. We have annotated the tasks from BEIR and found that most of these tasks cluster together in the top left of the graph, indicating that the performance of different models on these tasks is highly correlated. ArguAna is an exception within BEIR, as it is situated far from the other tasks. In contrast, tasks in NAME display significantly higher diversity, covering a larger area more evenly. The specific pairwise correlations are shown in the heatmap of Figure.6, revealing that many tasks exhibit negative correlations. This suggests that these tasks have significantly different definitions, leading to models that perform well on one task performing 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.

Model	Avg		Web		Academic		Legal		Medical		Finance		Code	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	36.05	-	34.51	-	35.62	-	50.63	-	37.48	-	52.81	-	24.72	-
`contriever-msmarco`	35.47	-	38.99	-	30.26	-	33.00	-	37.60	-	50.59	-	25.44	-
`gtr-t5-base`	35.02	-	37.12	-	29.94	-	35.48	-	36.63	-	47.78	-	27.95	-
`gtr-t5-large`	38.10	-	39.21	-	34.23	-	39.27	-	38.37	-	51.83	-	32.05	-
`all-MiniLM-L6-v2`	36.12	29.29	36.49	27.70	39.64	33.99	32.09	24.53	37.68	32.44	49.93	39.32	27.68	25.85
`e5-small-v2`	37.19	32.49	37.46	30.98	37.94	32.97	33.65	29.85	37.58	33.99	55.32	50.49	29.61	28.51
`e5-base-v2`	39.31	36.52	38.23	34.28	40.46	38.55	37.38	34.67	39.80	37.83	58.33	54.18	33.36	32.73
`e5-large-v2`	40.49	36.78	38.91	35.03	44.01	39.37	38.44	34.26	39.76	36.08	58.92	54.27	34.51	32.61
`gte-base-en-v1.5`	39.99	32.90	39.95	30.54	37.87	33.49	39.44	36.20	46.63	39.40	61.64	49.23	30.02	26.50
`gte-large-en-v1.5`	42.64	32.53	41.90	29.88	45.05	38.03	42.93	35.82	48.37	40.28	63.47	41.57	30.77	25.34
`bge-base-en-v1.5`	38.58	29.44	38.75	26.66	37.46	31.33	35.26	28.18	43.02	36.09	55.54	38.95	31.60	28.17
`bge-base-en-v1.5`	40.08	27.06	40.12	24.28	39.23	31.17	39.16	25.23	43.81	31.17	58.69	31.38	31.53	27.49
`text-embedding-3-small`	43.64	42.14	42.37	41.60	42.65	39.73	47.79	42.92	44.83	43.36	61.36	57.99	37.11	37.63
`e5-mistral-7b-instruct`♠	46.09	50.14	42.21	48.65	47.14	47.85	55.38	54.90	46.66	50.13	66.52	68.76	40.41	45.01
`NV-Embed-v1`♠	46.53	48.13	43.89	45.35	48.68	49.12	46.17	46.65	45.85	46.86	68.27	69.59	41.86	45.69
`monot5-base-msmarco`	38.77	34.30	39.92	36.09	32.39	23.80	43.51	40.30	41.98	38.50	59.34	56.72	28.73	24.39
`bge-reranker-v2-m3`	42.06	37.64	42.92	39.57	36.94	30.05	51.70	44.27	42.95	39.24	64.17	60.36	29.45	25.56
`bge-reranker-v2-gemma`	48.29	36.47	46.33	40.15	49.80	30.65	50.01	31.82	48.86	34.37	68.34	46.41	41.73	31.60
`gpt-3.5-turbo`♠	43.91	45.12	42.89	44.58	39.40	38.76	49.71	50.72	43.12	44.10	64.45	64.52	38.45	40.80

Table 3: **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, which all re-rank the top 100 results of `text-embedding-3-small`.

4 Experimental Setup

4.1 Models

Sparse Retrieval These are rule-based systems that measure relevance by computing term overlap. We benchmark `bm25`, implemented in Pyserini.

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. Additionally, we evaluated `text-embedding-3-small` through the OpenAI API[3].

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 and is trained on public instruction-tuning data (Lee et al., 2024). `NV-Embed-v1` is the state-of-the-art on MTEB leaderboard by the time of writing[4].

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, and (ii) `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

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

[4] https://huggingface.co/spaces/mteb/leaderboard

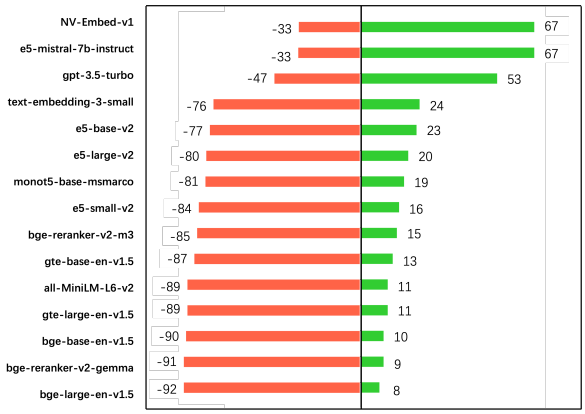

Figure 4: With the addition of instruction, the number of tasks that obtain performance improvement (green part) and reduction (red part). We can see that instruction-tuned models show more improvements while non-instruction-tuned models reduce on most tasks.

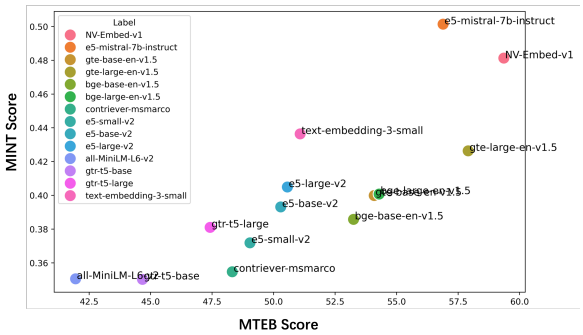

Figure 5: Score between MTEB and MINT.

the `gpt-3.5-turbo` backend (Sun et al., 2023b).

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 4.

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 3 reports the evaluation results of the tested models. Table 9 reports the detailed results of each individual task. We observe that instruction-tuned embedding models with instruction input achieve the best overall performance. `e5-mistral-7b-instruct` achieves an average score of 50.14, with a 4.05 nDCG improvement when instructions are added. `NV-Embed` also shows a clear improvement with the addition of instructions, though not as significant as `e5-mistral-7b-instruct`. 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 re-ranker, `bge-reranker-v2-gemma` achieves the best results, which outperform all embedding models when no instruction, but show a notable performance decline when instructions are added. RankGPT based on `gpt-3.5-turbo` gets results close to `bge-reranker-v2-m3` when no instruction input, but shows a 1.21 nDCG improvement when instructions are added. This indicates that prompted LLMs can intuitively transfer general instruction following the ability to rank tasks.

5.2 Gain of Instruction

Figure 4 illustrates with the addition of instruction, the model shows a performance increase on how many of the 100 tasks and a decrease on how many. We observe that the instruction-tuned models show improvement on more than half of tasks, while `NV-Embed-v1` and `e5-mistral-7b-instruct` perform the same and better than `gpt-3.5-turbo`. We also found that `text-embedding-3-small` receives the least impact from adding instruction among non-instruction-tuned models. This shows that `text-embedding-3-small` are more robust.

5.3 Compare with MTEB

Figure 5 shows the relationship between models' performance on MTEB and MINT. We can see that the two benchmarks share a similar trend. Among these models, `text-embedding-3-small` achieves better results on MINT 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 MINT. This is probably because `text-embedding-3-small` has better generalization, as it performs better on massive unseen tasks in MINT, 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 MINT, which indicates that MINT requires more generalization ability.

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. (2022a); Su et al. (2022a) 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., 2023b).

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 (Welleck et al., 2021) and InstructIR (Oh et al., 2024b), 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, MINT is a large-scale multi-task IR benchmark consisting of 100 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 large-scale IR benchmark called MINT for evaluating instruction-tuned retrieval models. The most notable property of MINT compared to existing tasks is that it contains more diverse IR tasks. Specifically, MINT consists of 100 different IR tasks, with 776 instructions newly annotated. We benchmark different types of retrieval models on MINT, including instruction-tuned and non-instruction-tuned. Our results show that our benchmark can better evaluate instruction-tuned models and is more challenging compared to existing benchmarks due to the very diverse tasks.

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*.

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

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022b. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

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*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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.

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*.

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. 2023a. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023b. 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.

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

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

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*.

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*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

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. 2024a. Instructir: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dong Sheng Li, Weiming Lu, and Yue Ting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *ArXiv*, abs/2303.17580.

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. 2022a. One embedder, any task: Instruction-finetuned text embeddings. *ArXiv*, abs/2212.09741.

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

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023a. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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.

V. Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims.

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. Supernaturalinstructions: 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. 2024a. Followir: Evaluating and teaching information retrieval models to follow instructions. *ArXiv*, abs/2403.15246.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024b. 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.

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.

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*.

Table 4: List of tasks in MINT.

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 Retrieva	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
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. (2023b)	GitHubIssue	Related File	GitHub Agent	Code
Apple	FinQabench	Question	Paragraph	Question Answering	Finance
ConvFinQA	Chen et al. (2022b)	Dialog	Table & Paragraph	Conversational 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
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	Conversational Retrieval	Web
AILA2019-Case	Bhattacharya et al. (2019)	Situation	Prior Case	Case Retrieval	Legal
AILA2019-Statutes	Bhattacharya et al. (2019)	Situation	Statute	Statue Retrieval	Legal
BSARD	Louis et al. (2021)	Question	Statute	Statue 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
ArguAna	Wachsmuth et al. (2018)	Claim	Document	Argument Retrieval	Web
CQADupStack	Hoogeveen et al. (2015)	Question	Duplicate Question	Query Retrieval	Web
FiQA	Maia et al. (2018)	Question	Answer	Question Answering	Finance
NFCorpus	Boteva et al. (2016)	Question	Medical Document	Medical QA	Medical
Quora	Quora	Question	Duplicate Question	Query Retrieval	Web
SciDocs	Cohan et al. (2020)	Paper Title	Citation	Reference Retrieval	Academic
SciFact	Wadden et al. (2020)	Claim	Document	Fact Checking	Academic
TopiOCQA	Adlakha et al. (2021)	Dialog	Passage	Conversational Retrieval	Web
Touche	Bondarenko et al. (2020)	Question	Passages	Argument Retrieval	Web
Trec-Covid	Roberts et al. (2020)	Medical Question	Answer	Medical QA	Medical

Table 5: List of tasks in MINT.

Task	Reference	Query	Doc	Task Type	
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
Monant	Srba et al. (2022)	Medical Claim	Document	Question Answering	Medical
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)	Dialoge	Passage	Conversational Retrieval	Web
ProCIS-Turn	Samarinas and Zamani (2024)	Utterance	Passage	Conversational Retrieval	Web
QuanTemp	Venktesh et al. (2024)	Numerical Claim	Document	Fact Checking	Web
WebTableSearch	Chen et al. (2021b)	Question	Table	Table Retrieval	Web
CAsT	TREC CAsT	Dialog	Passage	Conversational Retrieval	Web
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
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
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_2020	Fair-TREC	Key Words	Academic Articles	Web Search	Academic
FairRanking_2021	Fair-TREC	WikiProject	Wikipedia Page	Web Search	Web
FairRanking_2022	Fair-TREC	WikiProject	Wikipedia Page	Web Search	Web
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
TREC-Legal_2011	TREC Legal	Request	Communications	Case Retrieval	Legal
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
APISearch	Shen et al. (2023)	Question	API	Tool Retrieval	Code
ToolBench	Qin et al. (2023)	Question	Tool	Tool Retrieval	Code
ExcluIR	Zhang et al. (2024)	Question	Passage	Negative Retrieval	Web
FollowIR_Core17	Weller et al. (2024a)	Question	Document	Negative Retrieval	Web
FollowIR_News21	Weller et al. (2024a)	Question	News	Negative Retrieval	Web
FollowIR_Robust04	Weller et al. (2024a)	Question	News	Negative Retrieval	Web
InstructIR	Oh et al. (2024b)	Question	Passage	Web Search	Web
NevIR	Weller et al. (2023)	Question	Passage	Negative Retrieval	Web
IREval	Zhou et al. (2023)	Question	Answer	Question Answering	Web

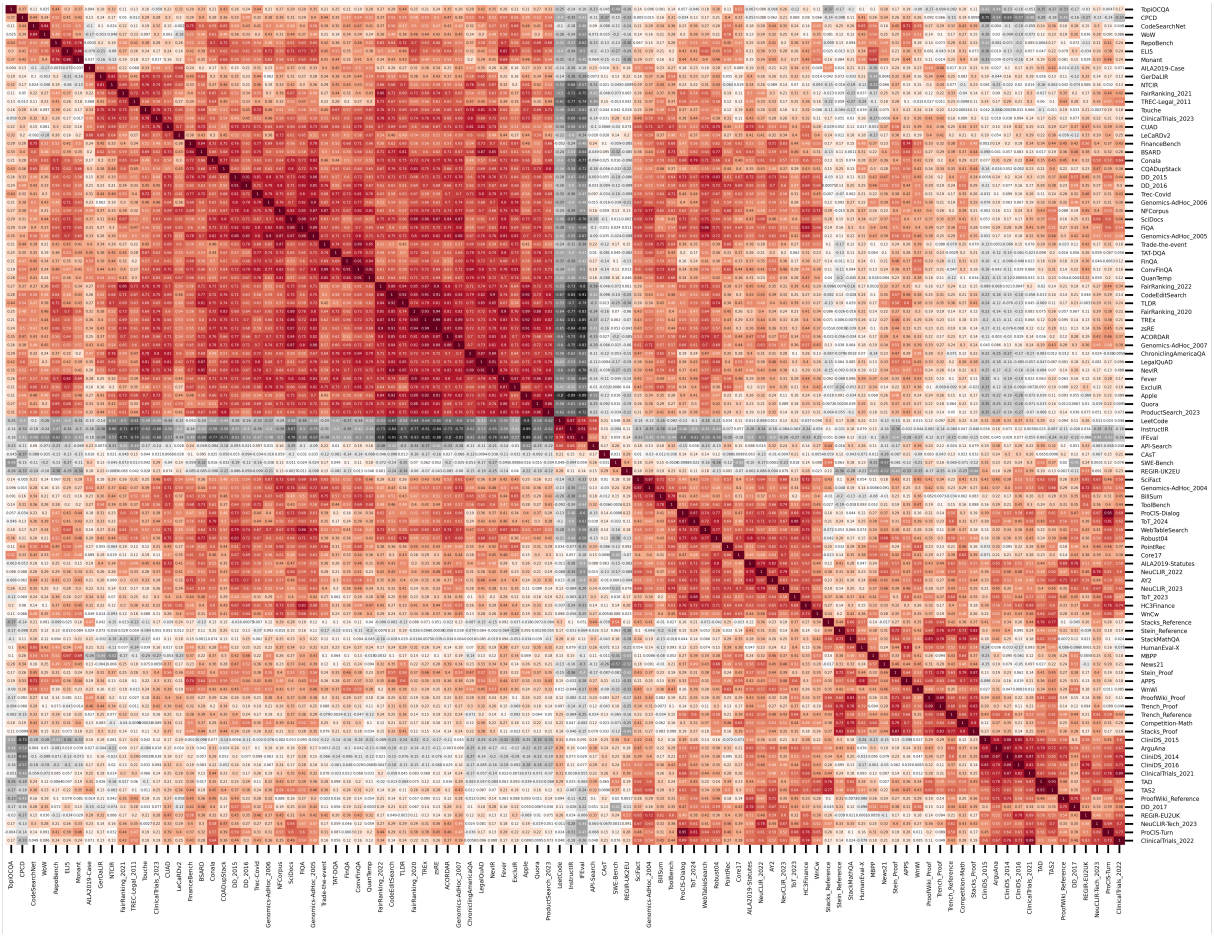

Figure 6: Task correlation heat map. The correlation between two tasks is measured by the Pearson correlation coefficient of different models' performance on these tasks. This coefficient ranges from -1 to 1. A value close to 1 indicates that models which perform well on Task A are likely to perform 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

A.1 TREC

A.1.1 TREC Clinical Decision Support Track 2014

This track studied the literature retrieval task in clinical decision support systems. It used case reports as queries to retrieve relevant biomedical articles.

- The queries are case reports describing the patient's symptoms, medical history, etc., but omitting information about diagnosis, tests, or treatment. For example, "A 58-year-old African-American woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back."
- The target document collection is an open-access subset of PubMed Central (PMC), an online repository of freely available full-text biomedical literature. Specifically, a snapshot of the PMC open access subset taken on January 21, 2014, is used, containing a total of 733,138 articles.
- Relevance is judged based on whether the document provided useful information of the specified clinical information type (diagnosis, test, or treatment) for the given case report. The relevance is divided into 3 levels:
  - Definitely relevant(2): The document provided information of the specified type (diagnosis, test, treatment) and is relevant to the particular patient case described in the topic.

  - Possibly relevant(1): The document is not immediately informative on its own, but may be relevant in the context of a broader literature review.
  - Not relevant(0): The document either did not provide information of the specified type or is not applicable to the patient's case.

- We annotate instructions like this: "Given a medical case record that consists of a summary and a full description, retrieve biomedical articles from PubMed Central that provide useful information for answering the following clinical question: What is the patient's diagnosis? / What tests should the patients receive? / How should the patient be treated?"

- The following is an example: [A table]

A.1.2 TREC Clinical Decision Support Track 2015

The same as A.1.1

A.1.3 TREC Clinical Decision Support Track 2016

The same as A.1.1

A.1.4 TREC Clinical Trials Track 2021

This track studies the task of retrieving clinical trials. It uses simulated patient case descriptions from electronic health records as queries to retrieve clinical trial descriptions published on ClinicalTrials.gov.

- The queries in the TREC 2022 Clinical Trials Track are brief patient case descriptions resembling admission notes, including details like age, medical history, symptoms, and test results. Each multi-sentence paragraph or list describes a realistic patient case derived with a specific disease in mind, but without revealing the disease. The 50 topics serve as queries to retrieve relevant clinical trials based on whether the described patient meets the trial's eligibility criteria.

- The document collection is a snapshot from April 27, 2021, of all clinical trials available on ClinicalTrials.gov, which contained 375,581 trial descriptions. Each trial is identified by its National Clinical Trial (NCT) number, which serves as the document ID. The core part of each trial description relevant for matching to patient cases is the eligibility criteria section, which lists the inclusion and exclusion criteria that determine if a patient qualifies for the trial or not.

- Relevance is determined based on whether the patient meets the inclusion criteria and exclusion criteria for the trial. The relevance judgments in this task are divided into 3 levels:

  - Eligible(2): The patient case description meets all the inclusion criteria for the clinical trial and does not violate any exclusion criteria. The patient would be eligible to enroll in the trial.
  - Excluded(1): The patient has the condition that the trial is targeting, but violates one or more of the exclusion criteria, making them ineligible for the trial.
  - Not Relevant(0): The patient case description does not contain enough information to determine eligibility, or clearly does not match the trial's criteria. The patient is simply not relevant for that particular trial.

- We annotate instructions like this: "Given a patient health record, retrieve clinical trials that are suitable for that patient. Rials are 'relevant' if the patient met the inclusion criteria and did not meet any exclusion criteria; Rials are 'partially relevant' if the patient met the inclusion criteria but is excluded by one or more exclusion criteria."

- The following is an example: [A table]

A.1.5 TREC Clinical Trials Track 2022

The same as A.1.4.

A.1.6 TREC Clinical Trials Track 2023

The same as A.1.4, but the documents are updated to 2023.

A.1.7 TREC Conversational Assistance Track 2019

Similar to A.1.10, The document collection is compiled from the following sources:
MS MARCO Passage Ranking collection
TREC CAR paragraph collection v2.0

A.1.8 TREC Conversational Assistance Track 2020

Similar to A.1.10, The document collection is compiled from the following sources:
MS MARCO Passage Ranking collection
TREC CAR paragraph collection v2.0

A.1.9 TREC Conversational Assistance Track 2021

Similar to A.1.10, The document collection is compiled from the following sources:
KILT Wikipedia
MS MARCO (Documents)

A.1.10 TREC Conversational Assistance Track 2022

This track studies the Conversational Passage Retrieval (ConvPR) task. The system needs to retrieve relevant document passages based on the user's query and previous conversational context and generate natural language responses.

- The queries are formed by the user's natural language utterances, containing their information needs. The utterances evolve and change as the conversation progresses, reflecting the dynamic nature of the user's information needs.
- The document collection includes documents from MS MARCO V2 and Wikipedia articles from the KILT dump. Each document is split into canonical passages of up to 250 words. The collection contains around 17 million passages in total.
- The relevance judgments are based on whether the passage contains useful information to satisfy the user's current utterance. The relevance levels are as follows:
  - Not Relevant(0): The passage does not relate to the current conversational context at all.
  - Partially Relevant(1): The passage is somewhat related to the conversational context, but the relevance is marginal or too tangential.
  - Relevant(2): The passage is relevant to the conversational context, but it is not entirely clear why this particular passage is retrieved.
  - Highly Relevant(3): The passage directly follows from and relates to the conversational context, and it is clear why this passage is retrieved.
  - Perfect Match(4): The passage perfectly matches the information need expressing in the conversational context.
- We annotate instructions like this: "Given a dialogue history, retrieve passages to answer the user's question in the last turn."
- The following is an example: [A table]

A.1.11 TREC Dynamic Domain Track 2015

This track studies dynamic, exploratory search tasks, simulating the interaction between professional searchers and search systems. It uses 2-4 word phrases as initial queries to retrieve documents from specific domain collections.

- The initial query is a 2-4 word phrase representing a topic, and the system needs to retrieve relevant documents for that topic.
- The documents are from three domains: Ebola, Illicit Goods, and Local Politics. They include web pages, PDF files, and tweets. The data sources include government agencies, news organizations, social media, etc.
- Relevance is determined based on how well the text passages from documents are associated with the subtopics of the main topic. The relevance is divided into 4 levels:
  - Marginally relevant(1): A passage judged as marginally relevant means it is only slightly or peripherally related to the subtopic. It may contain some tangentially relevant information but does not directly address the core aspects of that subtopic.
  - Relevant(2): A relevant passage means the content is definitely relevant and on-topic for that particular subtopic. It contains useful information that helps satisfy and answer aspects of the information need represented by that subtopic.
  - Highly relevant(3): A highly relevant passage is one that is extremely important and central to fully covering and addressing that subtopic. It likely contains key details, facts, or analysis that are critical for comprehensively answering the subtopic.
  - Key results(4): Passages judged as key results are the most essential and vital pieces of information required to satisfy that subtopic completely. These are the critical passages without which the subtopic cannot be fully answered or resolved satisfactorily.
- We annotate instructions like this "Given a domain and a topic associated with multiple sub-topics, the goal is to find passages relevant to all sub-topics. The feedback field includes the retrieved passages from an initial system, along with the sub-topics and their corresponding relevance scores. Your task is to retrieve passages related to the sub-topics that are not retrieved by the initial system. Ensure that you exclude any passages already retrieved by the initial system."
- The following is an example: [A table]

A.1.12 TREC Dynamic Domain Track 2016

The same as A.1.11

A.1.13 TREC Dynamic Domain Track 2017

The same as A.1.11, but the documents are updated to the New York Times.

A.1.14 TREC Fair Ranking Track 2019

This track studied the fair ranking retrieval task. It used queries from the academic search engine Semantic Scholar to re-rank academic paper abstracts.

- The query data came from the query logs of Semantic Scholar, containing the query string, displayed documents, and number of clicks for each document. After filtering and sampling, 652 training queries and 635 evaluation queries are provided.
- The documents are academic paper abstracts from Semantic Scholar, including paper ID, DOI, title, abstract, author IDs, and citation information. The document data came from the Semantic Scholar Open Corpus provided by the Allen Institute for Artificial Intelligence.
- Relevance is determined based on whether a document had at least one click in the query logs. If it had a click, it is considered relevant(1), otherwise non-relevant(0).
- We annotate instructions like this "Given a query, retrieve relevant academic articles. Ensure fair exposure to different groups of authors, where the group definitions could based on demographics, stature, topic, and others."
- The following is an example: [A table]

A.1.15 TREC Fair Ranking Track 2020

Similar to A.1.14, only the query and corpus are updated

A.1.16 TREC Fair Ranking Track 2021

Similar to A.1.14, only the query and corpus are updated

A.1.17 TREC Genomics Track 2004

This track aims to improve information retrieval in the biomedical field, with participants developing algorithms to search 160,000 full-text articles from 59 journals based on real biologists' questions. The 2004 Genomics Track included two tasks: an ad hoc retrieval task, which involved searching the MEDLINE database based on real biologists' information needs, and a full-text document categorization task, which simulated the curation process of the Mouse Genome Informatics system. In this paper, we only use data from ad-hoc tasks. Because classification tasks are not conducive to the construction of retrieval benchmarks

ad hoc retrieval task The ad hoc retrieval task mimics a situation where a biomedical researcher has an information need and searches a document collection (the MEDLINE subset) to find relevant medical documents.

- The queries originated from interviews with 43 biologists, yielding 74 initial information needs. These needs are reviewed and modified to create 50 final query topics. Each query topic contained fields like a title, statement of the information need, and background context.
- The document collection is a 10-year subset of the MEDLINE bibliographic database maintained by the U.S. National Library of Medicine, containing a large amount of literature related to genomics.
- The relevance is divided into 3 levels.
  - Definitely Relevant (2): The document is a clear, definite match for the information need expressed in the topic description. The content of the document directly and completely addresses the stated information need.
  - Possibly Relevant (1): The document appears related and potentially relevant to the information needed, but its relevance is not definite. The document may partially address the information needed or address it in a tangential way. There is some uncertainty about whether the document truly satisfies the stated information need.
  - Not Relevant (0): The document does not appear to be relevant or address the stated information need at all. There is no meaningful relationship between the document's content and what the topic is asking for.
- We annotate instructions like this "Given a biomedical question, retrieve documents from the MEDLINE bibliographic database to answer the question. The following is the context of this search: Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport."
- The following is an example: [A table]

A.1.18 TREC Genomics Track 2005

The same as A.1.17

A.1.19 TREC Genomics Track 2006

The same as A.1.17

A.1.20 TREC Genomics Track 2007

The same as A.1.17

A.1.21 TREC Legal Track 2010

This Track simulates the actual conditions of document review in civil litigation, aiming to find documents responsive to production requests.

- The queries are modeled on requests for production in civil litigation, specifying categories of documents that must be identified. Each request for production corresponds to a topic. There are 4 queries total - 3 subject-matter queries (301-303) and 1 privilege query (304) focused on identifying privileged documents.

  Query 301 - Documents related to onshore or offshore oil and gas drilling or extraction activities.

  Query 302 - Documents related to responses to oil and gas spills, blowouts, releases, or pipeline eruptions.

  Query 303 - Documents related to lobbying activities regarding legislation, rules, regulations, etc.

  Query 304 - Documents that could be withheld from production on grounds of privilege, attorney work-product, or other applicable protection.

- The document collection is derived from about 1.3 million emails captured from Enron by the Federal Energy Regulatory Commission (FERC) during its investigation of Enron's bankruptcy. The emails contain parent emails and attachments in various formats.

- The relevance is binary - either relevant (1) or not relevant (0) for each document/message assessed.

- We annotate instructions like this "Given a request for production in civil litigation, retrieve all documents or communications from the database that are responsive to this request while minimizing the number of unresponsive documents that are retrieved. The database consists of email messages captured by the Federal Energy Regulatory Commission (FERC) from Enron, in the course of its investigation of Enron."

- The following is an example: [A table]

A.1.22 TREC Legal Track 2011

The same as A.1.21

A.1.23 TREC NeuCLIR track 2022

This track studies cross-language information retrieval (CLIR) and multi-lingual information retrieval (MLIR) tasks. It uses English queries to retrieve news documents in Chinese, Persian, and Russian, as well as Chinese technical documents like dissertation abstracts. Notably, all documents in the collection are News documents that have been machine-translated into English.

- The queries are human-created, containing an English title, description and narrative to simulate real information needs.

- The documents include news articles in Chinese, Persian and Russian, as well as Chinese technical documents like dissertation abstracts covering fields such as chemistry, physics, computer science and more. The news articles are from newswire sources, while the technical documents are from a Chinese scientific literature database. Document lengths vary from a few sentences to a few hundred words.

- The basis for judging relevance is whether a document contains core information related to the query topic. The relevance is divided into 4 levels:

  - Highly relevant(3): document is fully relevant to the topic (i.e., it contains facts that would be included in the lead paragraph of a report on the topic)
  - No documents(2): there are no documents with relevant judgment

  - Somewhat relevant(1): document is somewhat relevant to the topic (i.e., it contains facts that would be included elsewhere in a report on the topic)
  - Nor relevant(0): document is not relevant to the topic (i.e., it does not contain information that would be included in a report written about the topic)

- Instruction is related to the query, and each query has different instructions.
- The following is an example: [A table]

A.1.24 TREC NeuCLIR track 2023

The TREC NeuCLIR track 2023 has two tasks: 1) Cross-language and multilingual news retrieval across three languages, which is the same as A.1.23 and 2) A pilot task for retrieving Chinese technical/academic content in a cross-language setting using English queries.

Technical Documents Pilot Task The pilot technical document task examines feasibility over diverse technical domains by retrieving Chinese academic abstracts using English queries, with optional monolingual runs using translated Chinese queries.

- Queries in this task are created by graduate students across various technical disciplines, with the topics focused on their specific research areas.
- Documents in this task is a large set of nearly 400,000 Chinese academic paper and thesis abstracts covering a wide range of technical disciplines like engineering, science, medicine, etc. published in the 2010-2020 time period. These technical abstracts formed the target corpus that systems had to search over using the provided English topics.

  All documents have been machine-translated into English.
- The relevance judgment is the same as A.1.23
- Instruction is related to the query, and each query has different instructions.
- The following is an example: [A table]

A.1.25 TREC Product Search Track 2023

This track studies the product search retrieval task. It uses synthetically generated queries to retrieve product descriptions and metadata documents from Amazon's online product catalog.

- The queries are generated in two ways: 1) 500 queries are generated using GPT-4 based on prompts, and 2) substrings are extracted from product titles and descriptions as queries. In total, 998 evaluation queries are generated.
- The documents are product descriptions and metadata, including reviews, dimensions, colors, and other attributes, as well as 1-10 64x64 thumbnail images crawled from Amazon's online catalog(not used in our setting). The data is sourced from the ESCI Product Re-ranking dataset, but only English products are retained and enriched with additional metadata from Amazon.
- The basis for judging relevance levels is the degree of match between the product and the query. Relevance is divided into 4 levels:
  - Perfectly relevant(3): The product is exactly what the user wants.
  - Highly relevant(2): The product could match the user query, but it may be a substitute for the original query intent. It may have a slightly different style, brand, or type, but a user would be satisfied if they received this product.
  - Related(1): The product seems related to the query but not the item the user seeks. Products in this category could complement the user's intended product.
  - Irrelevant(0): The product has nothing to do with the query.

- We annotate instructions like this "Given a search query, retrieve products that are relevant to the query."
- The following is an example: [A table]

A.1.26 TREC Product Search Track 2024

The same as A.1.25

A.1.27 TREC Tip-of-the-Tongue Track 2023

This track studied the tip-of-the-tongue (ToT) known-item retrieval task, using queries sampled from an existing ToT dataset to retrieve a corpus of Wikipedia pages associated with the "audiovisual works" category.

- The queries originated from the "I Remember This Movie..." forum, where users post ToT requests for movies. These queries tend to be verbose and include complex phenomena such as information about the movie itself (semantic memories), the context in which the searcher last encountered the movie (episodic memories), mentions of uncertainty, exclusion criteria, relative comparisons, and false memories.
- The document corpus is constructed from Wikipedia and contained pages associated with the "audiovisual works" category. This corpus included the correct answer for all training, development, and test queries. The documents consisted of multiple fields, including the page source code, parsed infoboxes, etc.
- The relevance is essentially a binary judgment - either a document is the one correct answer for that query topic (1), or it is not relevant(0).
- We annotate instructions like this "You are an expert in movies, assisting someone in recalling the name of a movie that's on the tip of their tongue. For the given tip-of-the-tongue information requests, retrieve Wikipedia pages of potential movie titles that match the description provided."
- The following is an example: [A table]

A.1.28 TREC Tip-of-the-Tongue Track 2024

The same as A.1.27

A.2 SIGIR

A.2.1 ACORDAR

The paper studies the task of ad hoc content-based dataset retrieval. Keyword queries are used to retrieve datasets, where the target documents are the metadata and data content of the datasets.

- The queries come in two forms: synthetic and TREC queries. Synthetic queries are created by human annotators by summarizing the content of a dataset they browsed, aiming to match the dataset's content. TREC queries are general topic queries from the TREC test collections.
- The documents are datasets containing metadata (e.g. title, description, author) and data content (RDF triples). These datasets are from 543 open data portals across multiple domains, in the RDF data format. The datasets vary greatly in data size and schema size.
- Relevance is judged based on whether the information contained in a dataset would be used when writing a report or developing an application on the topic of the query. The relevance is divided into 3 levels:
  - Irrelevant(0): the dataset is completely unrelated to the query topic, and none of its information would be used.

  - Partially relevant(1): the dataset has some content related to the query topic, and some of its information may be used, but not all.
  - Highly relevant(2): the dataset is highly related to the query topic, and much of its contained information would be heavily used. It can largely satisfy the information needs of the query.

- We annotate instructions like this "Given a question, identify and retrieve datasets where the keywords in the query may be found in either the metadata or the actual data of the dataset."

A.2.2 ChroniclingAmericaQA

This paper studies the open-domain question-answering task, using natural language questions as queries to retrieve relevant answers from the Chronicling America collection of historical American newspaper pages.

- The queries used are natural language questions, such as "How many subscribers did Jackson lose on Saturday?". These questions are automatically generated based on the content of the newspaper articles.

- The document collection used for retrieval is the Chronicling America collection of historical American newspapers. It is a large digital archive containing over 21 million pages of newspapers published between 1756 and 1963, covering information ranging from major historical events to everyday life details. The documents are available in both scanned image and OCR text formats.

- The ChroniclingAmericaQA dataset uses a binary relevance classification:

  - Relevant(1): The answer comes from the given historical newspaper document collection and is semantically relevant to the question, providing an answer to the query.
  - Not Relevant(0): The answer does not come from the given document collection, or even if it does, it is semantically unrelated to the question and cannot answer the query.

- We annotate instructions like this "Given a question, retrieve newspaper pages that contain the answer to the question from the American historical newspaper dataset Chronicling America."

A.2.3 CPCD

This paper studies a new retrieval task called Conversational Item Set Curation. In this task, a user engages in a conversation with a system to curate a small set of items (e.g. a music playlist) from a larger collection. The system needs to retrieve relevant items (e.g. songs) based on the conversation history and recommend them to the user.

- The query is formed from the conversation history, comprising all previous user utterances. For example, the user may first state a theme ("Create a rock playlist for a long drive"), and then provide additional details and preferences in subsequent turns ("Do you have anything from old Stevie Wonder?").

- The documents are the items in the collection to be retrieved, which in this work are songs. Each song is represented as a text document in the format "<title> by <artist> from <album>". The song data comes from expert-curated playlist descriptions.

- Relevance is judged based on the preferences expressed by the user during the conversation. A song is considered relevant if it matches the theme, genre, artist, etc. mentioned by the user. The relevance is divided into 4 levels:

  - Highly relevant(3): Matches all preferences expressed by the user
  - Relevant(2): Matches most of the preferences expressed
  - Partially relevant(1): Matches some of the preferences expressed
  - Not relevant(0): Does not match any of the expressed preferences

- We annotate instructions like this "The query is a dialogue history in which the user seeks music recommendations. In each turn, the system recommends some music, and the user provides feedback indicating whether they like or dislike the recommendations. Based on the user's preferences and requests, retrieve music that the user is likely to enjoy as the recommendation response."

A.2.4 Monant

This paper studies the task of detecting medical misinformation. Specifically, given a fact-checked medical claim, it aims to retrieve news articles or blog posts that contain that claim.

- The queries used are fact-checked medical claims, e.g. "Omega-3 fatty acids decrease triglycerides".

- The documents being retrieved are medical news articles and blog posts. These articles come from multiple reliable and unreliable website sources and include text content, images, videos, etc. The data is collected from the web using web crawlers.

- The basis for relevance levels is whether the article contains the given claim, and what stance the article takes towards that claim (supporting, contradicting, or neutral). The relevance levels are:
  - Present(1): The article contains a sentence or paragraph that is literally or semantically similar to the claim
  - Not present(2): The article does not contain the claim
  - Supporting(3) The article supports the claim
  - Contradicting(4) The article contradicts the claim
  - Neutral(5) The article has a neutral stance toward the claim

- We annotate instructions like this "Given a medical claim, retrieve articles that either support or contradict the claim."

A.2.5 NTCIR

The paper introduces a new test collection for ad-hoc dataset retrieval. The task is to retrieve relevant datasets from open government data portals (Data.gov and e-Stat) given a query derived from real information needs.

- The queries are keyword queries transformed from real information needs, aiming to capture users' actual search intents.

- The documents being retrieved are datasets from the government's open data portals, where each dataset contains metadata and one or more data files, mainly from Data.gov (US) and e-Stat (Japan) covering various domains.

- Relevance judgments are based on how useful a dataset is for satisfying the information need behind the query, with 3 grades:
  - Irrelevant(0): The dataset is completely irrelevant to the query.
  - Partially relevant(1): The dataset contains some relevant content but cannot fully satisfy the need.
  - Highly relevant(2) The dataset can well satisfy the information need of the query.

- We annotate instructions like this "Given a search query, retrieve relevant datasets from Data.gov and e-Stat that contain the requested information."

A.2.6 PointRec

This paper studies a point-of-interest (POI) recommendation task driven by natural language descriptions. User queries are natural language descriptions of their travel needs and preferences, which are used to retrieve from a document collection of 695,000 POI entries.

- The queries are in the form of natural language descriptions, specifying the user's travel destination, preferred categories (e.g. restaurants, nightlife), travel group, specific requirements or constraints on the POIs, etc. For example, "I want to find a not-too-noisy restaurant with good service near London to take my 80-year-old dad for a meal."
- The document collection contains 695,000 POI entries collected from Yelp's online directory, spanning over 2,000 cities across 23 countries. Each POI entry includes basic information like name, address, categories, rating, and review count, and is linked to the GeoNames geographical knowledge base for location data. Relevant text snippets are also extracted from web search results as descriptions. So this is a cross-domain document collection containing both structured data and unstructured text.
- The relevance judgments are primarily based on whether the POI satisfies the destination, category, travel group, and specific requirements or constraints described in the query. The relevance levels are:
  - Highly relevant(3): The POI perfectly matches the location, category, and all constraints specified in the query
  - Relevant(2): The location and category match and the requester would likely be interested
  - Somewhat relevant(1): The location matches, and the requester could possibly be interested
  - Not relevant(0): The POI does not satisfy any of the above conditions
- We annotate instructions like this "Given a user's request, recommend a point of interest (POI) that best matches the user's intent. Use the following metadata for reference:'Group': xx, 'Main Category': xx, 'Sub Categories': [xx, ...], ..."

A.2.7 ProCIS

The paper studies the task of proactive retrieval, where the system proactively retrieves and recommends documents relevant to an ongoing multi-party conversation, without the user explicitly providing a query. The query is the entire conversation history, and the documents are from Wikipedia articles.

The paper studies two different tasks:

Turn-level Proactive Retrieval: In this task, the system needs to decide whether to retrieve documents at each turn of the conversation. If it decides to retrieve, it needs to return a ranked list of documents as contextual suggestions for that turn.

Dialog-level Reactive Retrieval: In this task, the system needs to retrieve relevant documents based on the entire conversation history after the conversation is finished, as contextual suggestions for the whole dialogue.

- The query is the full context history of an ongoing multi-party human-machine conversation. The system needs to understand the conversation context and proactively retrieve documents relevant to the conversation at opportune moments as suggestions.
- The documents are articles from Wikipedia, covering a wide range of topics such as politics, religion, sports, finance, science, etc. These documents can provide useful background information, context, or supplementary knowledge for the ongoing conversation. The data source is Reddit conversation threads, which are cleaned and augmented by linking to Wikipedia articles.
- The basis for relevance levels is the degree to which the document is associated with the conversation, i.e., how much useful background information and context the document provides for the conversation. There are 3 relevance levels:

  - Provides Useful Context (2): The document is directly related to the conversation, providing substantial useful background information or supplementary knowledge that enhances the understanding of the conversation.
  - Partially Relevant (1): The document is somewhat related to the conversation. It provides some context or additional information but is not directly tied to the main topic of the conversation.
  - Irrelevant (0): The document is completely unrelated to the conversation and does not provide any useful context or additional information.

- We annotate instructions like this "Given a multi-party human conversation (such as a Reddit post and its comments), retrieve relevant documents from Wikipedia to provide contextual information. The criteria for relevance are as follows: ..."

A.2.8 QuanTemp

This paper studies the retrieval task for numerical claims for fact-checking. It uses the original claims and questions generated from claim decomposition as queries to retrieve relevant evidence documents from the web.

- The queries include the original numerical claims as well as a series of questions decomposed from the claims using methods like ClaimDecomp and Program-FC. For example, for the claim "The ICE spends 4 times the amount to detain a person for a year than on a student in public school", the decomposed questions could include "How much does ICE spend to detain a person for a year?", "How much is spent on a student in public school?" etc.

- The documents are relevant to web page snippets retrieved from the internet, covering domains like Wikipedia, government websites, statistics websites, etc. These snippets do not include any content from fact-checking websites to prevent the model from learning shortcuts. A total of 423,320 such evidence snippets are collected.

- The relevance levels are classified based on the usefulness of the evidence snippets in verifying the original claim. The relevance levels are as follows:
  - Highly relevant(3): Evidence snippet directly supports or refutes the claim
  - Partially relevant(2): Evidence snippet contains information helpful in verifying parts of the claim
  - Not relevant(1): The evidence snippet is not related to the claim
  - Contradictory(0): Evidence snippet contradicts the claim

- We annotate instructions like this "Given a numerical claim, retrieve documents that can verify the claim."

A.2.9 WebTableSearch

This paper studies the task of web table retrieval. It uses keyword queries to retrieve a set of relevant web tables.

- The query is a set of keywords describing the user's information need, e.g. "Fast cars".

- The documents are HTML tables extracted from the Common Crawl web corpus. These tables are from 61,086 different domain names, covering a broad range of topics. In addition to the table itself, each table has associated context fields such as page title, text before/after the table, and entities linked to DBpedia.

- The relevance levels are judged based on how relevant the query is to the table and its context fields (e.g. page title, surrounding text). There are 3 relevance levels:
  - Not relevant(0): This field is not relevant to the query.

  - Relevant (1): This field provides relevant information about the query.
  - Highly relevant (2): This field provides ideal information about the query.

- We annotate instructions like this "Given a question, retrieve relevant table to answer the question."

Table 6: Results.

Model	Math		ProofWiki_Proof		ProofWiki_Ref		Stacks_Proof		Stacks_Ref		Stein_Proof		Stein_Ref		Trench_Proof		Trench_Ref		TAD	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	54.53	-	50.83	-	10.63	-	25.42	-	21.54	-	61.54	-	25.16	-	67.21	-	20.59	-	23.85	-
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	-
e5-small-v2	78.07	74.42	48.94	43.25	11.91	11.23	32.64	27.73	21.88	21.19	64.99	58.68	28.98	28.20	61.81	56.94	24.22	22.78	25.94	22.69
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	22.64	20.16	58.93	56.08	24.24	25.40	48.15	44.79	20.29	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
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
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		CodeEditSearch		CodeSearchNet		Conala		HumanEval-X		LeetCode		MBPP		RepoBench	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	29.15	-	29.63	-	9.93	-	42.52	-	40.85	-	9.46	-	29.45	-	23.79	-	9.91	-	2.37	-
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	-
e5-small-v2	36.18	30.97	44.62	42.68	20.55	20.64	42.05	19.32	65.99	58.69	16.47	10.29	62.88	58.27	25.35	36.50	70.17	70.75	12.71	9.07
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
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
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-the-event		AY2	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	15.78	-	53.13	-	64.25	-	65.31	-	64.20	-	20.23	-	13.38	-	88.05	-	94.97	-	8.50	-
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	-
e5-small-v2	16.10	10.03	38.16	33.71	64.93	43.06	59.19	48.50	52.02	50.66	27.58	22.96	19.32	16.56	83.09	68.32	94.79	88.23	12.94	11.64
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
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
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 7: Results.

Model	ELI5		Fever		TREx		WnCw		WnWi		WoW		zsRE		AILA2019-Case		AILA2019-Statutes		BSARD	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	14.23	-	65.10	-	71.37	-	5.98	-	7.42	-	42.93	-	83.18	-	12.55	-	10.83	-	14.71	-
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	-
e5-small-v2	43.87	37.54	85.18	73.10	77.89	28.50	9.42	8.32	8.45	9.85	60.33	58.30	92.49	31.70	8.19	7.21	11.92	8.83	16.37	10.16
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
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
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	80.19	-	31.19	-	61.97	-	66.69	-	64.54	-	52.97	-	50.73	-	28.55	-	28.93	-	27.38	-
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	-
e5-small-v2	64.95	47.00	23.53	15.29	19.36	14.48	48.98	46.32	51.39	40.42	48.59	45.85	36.17	36.66	33.44	33.32	38.58	31.19	40.86	32.85
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
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
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	33.18	-	77.39	-	11.34	-	70.79	-	27.88	-	68.44	-	59.41	-	25.74	-	1.16	-	78.09	-
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	-
e5-small-v2	34.00	31.05	87.56	72.96	19.52	10.42	70.26	64.60	38.23	38.87	40.69	39.01	63.16	60.15	27.75	20.78	2.54	1.66	46.64	31.08
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
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
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 8: Results.

Model	Monant		NTCIR		PointRec		ProCIS-Dialog		ProCIS-Turn		QuanTemp		TableSearch		CAsT		CliniDS_14		CliniDS_15	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	59.60	-	33.12	-	21.21	-	23.34	-	16.99	-	91.33	-	8.85	-	15.32	-	36.13	-	40.39	-
contriever-msmarco	72.95	-	16.66	-	43.99	-	37.33	-	26.20	-	87.67	-	13.23	-	15.32	-	24.92	-	25.27	-
gtr-t5-base	75.99	-	12.06	-	35.61	-	47.47	-	33.00	-	87.25	-	12.12	-	15.32	-	30.85	-	24.92	-
gtr-t5-large	75.28	-	12.07	-	43.43	-	50.13	-	36.37	-	87.30	-	14.05	-	15.32	-	35.47	-	26.26	-
e5-small-v2	77.73	73.70	14.89	11.04	40.66	38.13	40.27	25.71	31.40	21.67	86.51	83.78	11.63	9.56	13.44	15.32	36.33	34.78	28.85	28.79
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	9.68	15.32	23.86	24.21	23.88	23.12
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	15.32	15.32	26.13	28.59	24.85	25.49
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	15.32	15.32	27.31	31.11	24.69	23.72
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	15.32	15.32	49.54	47.74	36.90	36.23
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	15.32	15.32	57.89	47.39	42.53	43.09
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	15.32	15.32	41.26	41.22	33.72	36.42
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	15.32	15.32	44.24	43.19	39.00	34.37
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	15.32	15.32	42.20	40.20	29.28	29.90
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	15.32	15.32	45.97	54.80	42.05	50.77
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	15.32	15.32	43.17	45.28	37.16	40.89
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	15.32	15.32	26.37	24.62	20.42	21.15
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	15.32	6.60	38.80	28.54	31.49	25.44
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	15.32	15.32	55.09	35.17	41.33	24.55
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	15.32	15.32	43.18	45.06	29.46	31.33

Model	CliniDS_16		CliniTrial_21		CliniTrial_22		CliniTrial_23		DD_15		DD_16		DD_17		FairRank_20		FairRank_21		FairRank_22	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	17.11	-	22.63	-	17.13	-	44.52	-	29.61	-	24.00	-	22.99	-	35.12	-	54.37	-	76.93	-
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	-
e5-small-v2	11.09	11.34	25.18	23.79	21.70	19.96	24.42	14.98	40.47	32.99	31.36	26.93	32.18	31.33	41.46	26.32	48.31	34.08	75.15	44.08
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
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
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_11		NeuCLIR-Tech_23		NeuCLIR_22		NeuCLIR_23		Product_23		ToT_23	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	59.01	-	29.51	-	14.52	-	23.91	-	46.29	-	24.24	-	50.41	-	43.50	-	46.03	-	25.89	-
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	-
e5-small-v2	57.06	48.75	38.08	30.03	19.65	15.97	24.36	15.85	40.50	12.42	36.12	27.32	54.51	52.95	46.72	46.38	49.21	39.82	35.92	33.90
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
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
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

Table 9: Results.

Model	**ToT_24**		**API-Search**		**ToolBench**		**ExcluIR**		**Core17**		**News21**		**Robust04**		**InstructIR**		**NevIR**		**IFEval**	
Instruction?	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓
BM25	22.15	-	16.87	-	72.45	-	72.71	-	13.97	-	17.83	-	21.90	-	40.30	-	67.62	-	26.84	-
`contriever-msmarco`	40.94	-	16.05	-	63.52	-	81.06	-	18.18	-	27.33	-	26.64	-	50.45	-	62.97	-	25.51	-
`gtr-t5-base`	38.27	-	13.31	-	65.93	-	77.94	-	22.73	-	22.49	-	26.90	-	49.04	-	63.14	-	22.59	-
`gtr-t5-large`	47.75	-	14.90	-	73.40	-	82.33	-	22.80	-	26.62	-	29.07	-	50.87	-	59.95	-	23.46	-
`e5-small-v2`	34.13	26.39	13.50	18.91	62.20	59.04	78.46	72.61	23.62	22.66	27.14	25.12	27.34	23.99	50.37	80.37	63.02	54.67	23.10	35.42
`e5-small-v2`	26.67	16.36	14.40	21.28	68.61	61.19	79.53	77.69	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`	32.76	19.44	16.73	20.92	65.67	57.91	83.04	81.12	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`	38.13	23.81	17.54	20.13	65.51	62.18	82.88	81.80	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`	37.41	29.78	18.06	21.61	67.34	63.09	82.19	72.83	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`	46.95	33.79	15.15	20.67	74.13	72.32	82.01	68.79	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`	34.13	25.68	16.41	20.18	69.07	66.69	83.51	77.79	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`	39.06	30.48	16.45	20.21	65.20	60.16	82.83	69.02	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`	48.33	41.18	17.14	22.32	77.91	74.81	81.85	81.51	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`	52.86	60.10	20.18	29.81	78.57	79.41	84.92	86.34	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`	53.70	57.64	23.05	30.83	79.75	80.66	87.19	86.45	26.49	32.81	33.11	30.94	33.04	35.13	49.66	66.14	69.55	69.54	22.45	27.68
`monot5-base-msmarco`	27.88	12.58	13.65	15.86	73.33	51.73	82.68	81.73	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`	39.05	11.29	12.73	13.61	81.34	79.31	86.63	86.08	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`	61.96	14.08	23.01	19.94	87.41	55.00	79.97	73.75	28.28	29.02	31.41	27.62	39.11	30.19	52.99	90.40	80.84	65.00	19.42	53.73
`gpt-3.5-turbo`	50.26	53.37	18.72	21.86	77.87	78.05	82.42	76.43	29.09	33.43	29.62	30.27	29.06	29.73	52.22	78.64	76.96	72.50	15.15	29.04