# IFIR: A Comprehensive Benchmark for Evaluating Instruction-Following in Expert-Domain Information Retrieval

**Tingyu Song**♣    **Guo Gan**◇    **Mingsheng Shang**♡♣    **Yilun Zhao**♠

♣ School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
♡ Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
◇ Zhejiang University   ♠ Yale University

https://github.com/SighingSnow/IFIR

## Abstract

We introduce IFIR, the first comprehensive benchmark designed to evaluate instruction-following information retrieval in expert domains. IFIR includes 2,426 high-quality examples and covers eight subsets across four specialized domains: finance, law, healthcare, and science literature. Each subset addresses one or more domain-specific retrieval tasks, replicating real-world scenarios where customized instructions are critical. IFIR enables a detailed analysis of instruction-following retrieval capabilities by incorporating instructions at different levels of complexity. We also propose a novel LLM-based evaluation method to provide a more precise and reliable assessment of model performance in following instructions. Through extensive experiments on 15 frontier information retrievers, including those based on LLMs, our results reveal that current models face significant challenges in effectively following complex, domain-specific instructions. We further provide in-depth analyses to highlight these limitations, offering insights to guide future advancements in retriever development.

## 1 Introduction

The instruction-following ability has become a cornerstone for LLMs (Ouyang et al., 2022; Jiang et al., 2023; Groeneveld et al., 2024; AI@Meta, 2024; Yang et al., 2024), empowering them to interpret and respond to complex user commands and perform a wide range of user-specific tasks. Despite its critical importance, the instruction-following capability remains underexplored in the context of information retrieval (IR).

Current information retrievers struggle to meet the nuanced requirements of users in real-world applications, particularly in specialized fields like law (Ma et al., 2021; Goebel et al., 2024), healthcare (Roberts et al., 2020; Ionescu et al., 2024; Ke et al., 2024), and scientific research (Cohan et al.,



**Case:** XYZ Corporation filed a lawsuit against John Doe, a former employee, for defamation after Doe posted allegations on his personal social media claiming that XYZ provided false information to customers. While Doe's post didn't result in significant or immediate financial loss for the company, XYZ argued that it tarnished their reputation.

**Query:** I am finding some relevant case **similar** to this one.
**Instruction:** I am seeking a **plaintiff's beneficial** case similar to this one, particularly one where the **corporation incurs minimal financial loss but wins the litigation**.
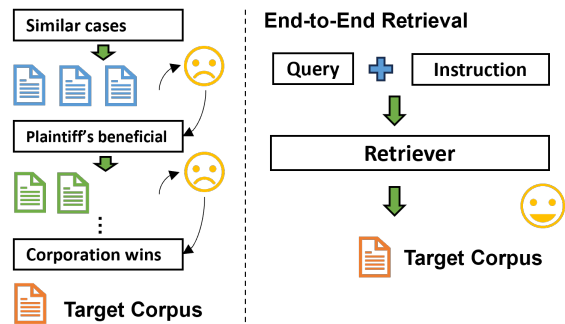
Figure 1: (Top): An illustration of instruction-following IR scenarios explored in this study. The example simulates a legal case search, where the user provides detailed instructions to retrieve relevant legal cases. Current IR systems struggle to handle such complex queries. (Bottom left): As a result, users have to break down their information needs into simpler, iterative search queries and manually filter the retrieved cases, resulting in a time-consuming and inefficient process. (Bottom right): This study focuses on evaluating the progress and limitations of current end-to-end retrieval systems in expert-domain instruction-following IR.

2020; Wang et al., 2023a; Liu et al., 2024), where precise and context-aware retrieval is crucial (Saxena et al., 2022; Mysore et al., 2022; Weller et al., 2025b). For instance, in legal research, lawyers often search for target cases using detailed instructions that incorporate specific legal criteria, contextual information, and desired outcomes, as illustrated in Figure 1. However, traditional IR systems (Johnson et al., 2016; Liu et al., 2021; Chan et al., 2024) lack the ability to fully understand and process such complex user instructions in an *end-to-*

*end* manner. Consequently, users must break down their complex information-seeking needs into several simpler search queries and manually filter the retrieved cases (Liu et al., 2021), which is time-consuming and inefficient.

On the other hand, existing instruction-following IR benchmarks typically employ simplified instructions, such as single sentence (Su et al., 2023a) or a set of keywords (Zhao et al., 2024a). This simplification in evaluation leads to an incomplete assessment of the model's real-world performance. Although concurrent works like FOL-LOWIR (Weller et al., 2024a, 2025a), INSTRUC-TIR (Oh et al., 2024) and BRIGHT (Su et al., 2024) incorporate more complex instructions, they do not establish explicit complexity levels to evaluate retrievers' *fine-grained* abilities in following instructions. Moreover, these studies primarily focus on general domains, leaving the evaluation of instruction-following retrieval in *expert domains* largely underexplored.

In this work, we introduce **IFIR**, a comprehensive benchmark designed to evaluate the **I**nstruction- **F**ollowing capabilities of **I**nformation **R**etrievers, particularly in the context of specialized domains. IFIR includes eight subsets covering four specialized domains: finance, scientific literature, law, and healthcare. To provide a more granular evaluation, we create three levels of instruction complexity for each domain, representing a range of real-world information retrieval scenarios in expert domains. IFIR includes 2,426 instruction-following queries, each averaging 6.14 ground-truth passages. To ensure the dataset's high quality, we conduct a comprehensive human expert validation during its construction. Moreover, recognizing the limitations of traditional evaluation methods in measuring instruction-following IR performance, we introduce a novel LLM-based metric, INSTFOL, designed to more accurately assess how well retrievers follow instructions.

Through extensive experiments on 15 frontier retrievers, including those based on LLMs, we derive three key findings: (1) BM25 performs relatively well because the instructions in expert domains contain more glossary terms. (2) Instruction-tuned retrievers like INSTRUCTOR (Su et al., 2023b) do not perform significantly better than their base models, *i.e.*, GTRs (Ni et al., 2022). This demonstrates that current instruction-tuned retrievers may not be suitable for complex instructions. (3) Most evaluated models experience performance declines as

the complexity of the instructions increases. (4) LLM-based retrievers demonstrate more robust performance on both nDCG and INSTFOL, highlighting their potential in managing more complex retrieval tasks in specialized domains.

We conclude our main contributions as follows:

- We introduce IFIR, a comprehensive IR benchmark to evaluate the instruction-following ability of information retrievers across specialized domains, meeting their specific demands. The experiments provide insights into *end-to-end* retrieval in specialized domains.

- We propose INSTFOL, the first LLM-based evaluation method to measure the instruction-following ability of information retrievers.

- We conduct extensive experiments encompassing a wide range of retrievers, deriving key findings about their instruction-following abilities. Our experimental results reveal the potential of LLM-based retrievers in instruction-following retrieval.

## 2 Related Work

### 2.1 Expert-domain IR Benchmarks

Information retrieval plays a crucial role in expert domains by enabling efficient access to domain-specific knowledge, facilitating evidence-based decision-making, and accelerating research. In recent years, there has been a growing emphasis on developing IR benchmarks in specialized domains, such as law (Xiao et al., 2019; Li et al., 2023), finance (Jangid et al., 2018; Chen et al., 2021; Zhao et al., 2024b), scientific literature (Wu et al., 2024; Ajith et al., 2024), and healthcare (Roberts et al., 2020; Tamine and Goeuriot, 2021; Xiong et al., 2024). However, existing benchmarks primarily employ oversimplified queries, lacking the depth and specificity required in real-world specialized domains. For instance, in healthcare, practitioners often formulate complex, context-rich queries that integrate patient-specific information (*e.g.,* medical history and treatment plans) to retrieve relevant clinical passages. Such queries go beyond simple keyword search and require models to handle complex, user-customized needs. This gap underscores the need for a new benchmark that thoroughly evaluates the ability of retrievers to handle instruction-following IR tasks in specialized domains.
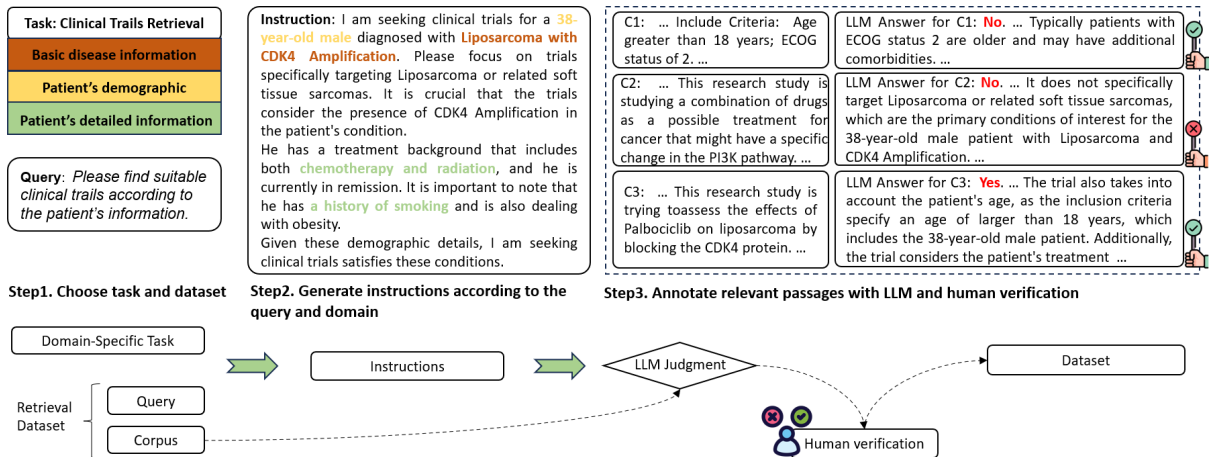
Figure 2: Dataset Construction Pipeline: We derive a specific task according to the dataset, which then guides the generation of instructions based on the original query and task conditions. An LLM is used to assess whether the corpora are relevant to these instructions. As illustrated in the figure, different colors in the "Task" section correspond to the conditions outlined in the "Instruction" section.

## 2.2 Instruction-Following IR

Several studies have proposed novel training techniques to improve retrievers' instruction-following capabilities (Su et al., 2023a; Asai et al., 2023; Wang et al., 2023c). However, due to the lack of instruction-specific IR benchmarks during model development, these approaches have typically been evaluated using traditional benchmarks like BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023), which contain queries without complex instructions. To address this gap, new instruction-following IR benchmarks have been proposed. Specifically, INSTRUCTIR (Oh et al., 2024) reformulates queries in existing retrieval datasets by incorporating user-aligned instruction. However, it limits each query to a single relevant passage, which does not reflect the complexity of real-world scenarios where multiple relevant passages may exist. While FOLLOWIR (Weller et al., 2024a) introduces more complex instructions and passage setups, it focuses on the reranking task. Moreover, these studies primarily focus on instruction-following retrieval in general domains, leaving the evaluation of instruction-following IR in *expert domains* largely underexplored.

## 3 IFIR Benchmark

We introduce IFIR, a comprehensive benchmark designed to assess and enhance retrievers' instruction-following capabilities in expert domains. The construction process of IFIR, illustrated in Figure 2, employs a semi-automated, human-in-the-loop pipeline that ensures both scala-

bility and high quality. Specifically, we expand the queries in existing specialized-domain IR benchmark (detailed in §3.1) by incorporating detailed instructions that closely mirror real-world scenarios in relevant expert domains (§3.2). Each example is thoroughly validated by domain experts (details of annotator selection and assigned tasks are provided in Table 4 in Appendix), ensuring that the instructions are contextually relevant and represent real-world challenges (§3.2). Additionally, the corresponding relevant passages are carefully verified for completeness and accuracy (§3.3). In the following subsections, we provide a detailed description of each step in the construction pipeline.

## 3.1 Retrieval Dataset Collection

To facilitate comprehensive evaluations of instruction-following capabilities across diverse expert domains, IFIR spans four specialized domains: finance, scientific literature, law, and healthcare. These domains are chosen for their significant demand for precise information retrieval and the complexity of their often nuanced queries. For each domain, we extend one or two well-established traditional IR benchmarks (see Table 1), adapting them to instruction-following IR tasks. This ensures that IFIR captures the real-world challenges for each expert domain.

## 3.2 Instruction Annotation

We now describe the process of augmenting the original IR benchmark queries with instructions that mirror real-world demands and challenges spe-

10188

| Domain | Adopted Datasets | # Qry | Corpus Size | # RP | Designed Tasks Reflecting Real-world Challenges |
|---|---|---|---|---|---|
| Finance | FiQA (Jangid et al., 2018) | 1,718 | 57,638 | 3.54 | Retrieve financial suggestions based on user-specific needs to support informed decision-making. |
| Scientific Literature | SciFact-open (Wadden et al., 2022) | 152 | 500,000 | 4.84 | Retrieve relevant scientific literature tailored to specific scientific research needs. |
| | NFCorpus (Boteva et al., 2016) | 86 | 3,633 | 2.81 | |
| Law | AILA (Bhattacharya et al., 2019) | 85 | 2,914 | 2.01 | Retrieve legal cases that satisfy customized demands. |
| | FIRE (Mandal et al., 2017) | 168 | 1,745 | 3.36 | Retrieve legal cases to support the judicial decision. |
| Healthcare | TREC-PM (Roberts et al., 2017, 2018) | 172 | 241,006 | 15.61 | Retrieve relevant clinical trials based on patient's demographics (*e.g.,* age, gender, medical history) |
| | TREC-CDS (Roberts et al., 2015) | 43 | 633,955 | 10.84 | |

Table 1: Overview of the adopted datasets and designed instruction-following IR tasks in IFIR, along with basic statistics for each domain. "**# RP**" represents the average relevant passage number per instruction-following query.

cific to each domain. Details of instruction generation are shown in Appendix A.1. These instructions introduce complex, often implicit conditions that significantly increase the difficulty of retrieval tasks, as they require the model to not only identify relevant passages but also interpret and follow specific instructions tied to nuanced, task-specific information retrieval. To ensure high-quality instruction annotation, we employ a *human-in-the-loop* pipeline. We first use LLMs (*i.e.,* GPT-4o `2024-05-13` version) to generate instructions at varying levels of complexity for each query. These instructions are then reviewed and refined by domain experts to ensure they are precise and aligned with the real-world demands of the respective domains. We detail the annotation process for each domain as follows:

**Finance** In the finance domain, we focus on the instruction-following IR task centered around personal finance inquiries, simulating scenarios where users seek guidance in making informed financial decisions. We design three complexity levels of instructions in queries: The first level involves simple instructions, *e.g.,* "Please help me to find a financial suggestion for the query {$query}." The second level includes additional personal information such as age, occupation, and financial status. The third level builds upon the second level by incorporating specific financial goals. For example, "As a 40-year-old accountant with a steady income and moderate savings, I am seeking advice on the best business structure for taxes when combining full-time work with running a small side business. I am looking for insights on how to optimize tax efficiency while balancing the demands of my full-time job and side business."

**Scientific Literature** In the science literature field, we focus on the science passage searching queries, simulating a person working in a relevant

area(*e.g.,* teacher, student, etc.) trying to find passages related to scientific claims or problems. We recognize that query instructions can vary in research topics (*e.g.,* society, history, biomedical, etc.) and research objectives (*e.g.,* influence, reasoning process). We use the Scifact-open dataset to generate three different levels of instructions. The first level of instruction might state, "Please help me find relevant evidence to support the scientific claim." The second level uses previously annotated "SUPPORT" and "CONTRADICT" tags to generate instructions like "Please help me to find supporting evidence for this scientific claim." The third level includes more customized requirements like research topics and objectives.

**Law** In the legal field, we focus primarily on the legal case retrieval task, simulating someone(*i.e.,* a lawyer) attempting to find relevant references from previous legal cases. We have two types of instructions. One type is to retrieve prior cases that support the reasoning process for the current case, which originates from the FIRE2017 dataset. The other, derived from AILA2019, is designed to retrieve similar cases according to the demands of legal professionals. For the first type, we construct instructions based on the context in the passage that requires citation. For the second type, we construct three different levels of instructions. The first level, similar to previous domains, is "Please help me to find cases similar to the current legal case." The second level adds conditions including whether the case is beneficial to the defendant or plaintiff. The third level constructs instructions searching for cases relevant to some details of the current case while still satisfying the previous two levels.

**Healthcare** In the healthcare domain, we focus on retrieving healthcare-relevant passages (*e.g.,,* clinical trials, diagnoses), simulating a doctor finding passages suitable for patients. The seed

datasets are derived from TREC-CDS and TREC-PM. Given the two different datasets and corresponding tasks in the biomedical field, TREC-CDS provides a summary accompanied by a detailed description, which we directly use as the instruction. Inspired by the TREC-CDS track, we expand the basic information provided in the TREC-PM track and construct three levels of instructions. The first level contains conditions of the patient's disease and gene variation. The second level adds conditions about the patient's demographics, including age and gender. The third level allows the LLM to create information about the patient's treatment history and family medical history.

Table 1 presents the data statistics of IFIR. It includes a total of 2,426 instruction-following queries across four expert domains. Examples of queries for each domain are provided in Appendix A.2. Each query is carefully reviewed by one of domain experts (Table 4 in Appendix).

### 3.3 Relevant Passage Annotation

We next discuss the process of annotating relevant passages. For each instruction-following query in IFIR, we select relevant passages from its original dataset. The key insight is that if a passage is annotated as relevant to a query, it may also be relevant to an instruction constructed based on that query. Therefore, we need to verify whether these query-relevant passages satisfy the conditions outlined in the corresponding instructions. Specifically, we first use LLM (*i.e.,* GPT-4o) to assess the relevance of each original relevant passage to the instruction. The LLM is tasked with generating justification explanations alongside its relevance assessments (we present prompt in Appendix B). Human annotators then review the relevance of each passage and the justifications provided by the LLM. If a passage is found to be misaligned with the instruction, the annotators exclude it.

### 3.4 Dataset Analysis

Table 1 presents the basic statistics of IFIR. We also conduct a final human evaluation of data quality on 50 examples from each domain subset. For each domain, an external expert (not involved in the data annotation) assesses the quality of each example, providing ratings across several criteria on a scale of 1 to 5. The results, shown in Table 2, indicate consistently high scores ($> 4$) for instruction-following query and relevant passages. These evaluation results indicate the high quality of the bench-

| Evaluation Criteria | Score (1-5) |
|---|---|
| *Instruction-Following Query* | |
| Naturalness | 4.24 |
| Fluency | 4.81 |
| Expertise | 4.87 |
| *Relevant Passage (RP)* | |
| Relevant Passage Agreement | 4.43 |
| *Excluded RP* | |
| Exclusion Agreement | 4.32 |

Table 2: Human Validation Results. Naturalness of instructions evaluates how well the instructions align with real-world demands. The Relevant Passage Agreement Score refers to human annotators' agreement with the LLM on identifying a golden passage, while the Exclusion Agreement Score reflects human annotators' agreement on excluding a passage.

mark proposed by us, further demonstrating its reliability for assessing instruction-following capabilities in relevant tasks.

## 4 Experiment Setup

This section outlines the experimental setup of our study. We first introduce the two automated metrics used for evaluating IFIR, and then discuss the evaluated retrieval systems. Implementation details can be found in Appendix B.

### 4.1 Evaluation Metrics

**nDCG** We use the widely-adopted IR metric, nDCG (Järvelin and Kekäläinen, 2002), to evaluate *retrieval performance*. Specifically, given a query with instruction $Q$, golden passages $G$, the retrieved passages $P$ are compared against the $G$ using nDCG to quantify the accuracy and relevance of the retrieval. While nDCG offers a broad assessment of a model's retrieval capabilities, it does not capture the fine-grained aspects of a model's ability to follow instructions.

**INSTFOL** To address the aforementioned limitation, we introduce a new LLM-based metric, INSTFOL, specifically designed to evaluate *instruction-following capabilities* on IFIR. The core idea is to assess the improvement a retriever demonstrates when instructions are incorporated into the query, compared to when they are not. We adopt the evaluation prompt from G-Eval (Liu et al., 2023), as shown in Figure 8 in the Appendix, and use GPT-4o-mini as the base evaluator to assess the alignment between each retrieved passage and the given instruction. Considering the potential biases and

| | FiQA | | SciFact-open | | NFCorpus | | AILA | | FIRE | | TREC-PM | | TREC-CDS | | **Average** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nDCG | INSTFOL | nDCG | INSTFOL | nDCG | INSTFOL | nDCG | INSTFOL | nDCG | INSTFOL | nDCG | INSTFOL | nDCG | INSTFOL | **nDCG** | INSTFOL |
| *Non-instruction-tuned Models* | | | | | | | | | | | | | | | | |
| GTR-XL | 0.44 | -4.85 | 0.54 | -2.93 | 0.60 | 7.30 | 0.05 | -0.13 | 0.54 | 0.40 | 0.31 | -0.85 | 0.15 | 7.69 | 0.37 | 0.95 |
| BM25 | 0.25 | 1.10 | 0.49 | -0.40 | 0.43 | 1.46 | 0.10 | 0.02 | 0.55 | 0.03 | 0.47 | 3.43 | 0.07 | -2.14 | 0.34 | 0.50 |
| GTR-Large | 0.39 | -6.47 | 0.50 | 0.05 | 0.51 | -3.37 | 0.07 | -0.36 | 0.49 | -5.21 | 0.28 | -1.46 | 0.09 | 11.38 | 0.33 | -0.78 |
| GTR-Base | 0.33 | -4.62 | 0.47 | -0.62 | 0.47 | -0.08 | 0.05 | 0.54 | 0.52 | 0.81 | 0.27 | -0.65 | 0.13 | 7.03 | 0.32 | 0.35 |
| Contriever | 0.13 | 0.52 | 0.29 | -8.22 | 0.36 | 0.17 | 0.08 | 0.09 | 0.51 | 2.82 | 0.09 | 1.24 | 0.04 | 2.46 | 0.21 | -0.13 |
| ColBERT | 0.07 | 0.17 | 0.14 | 0.34 | 0.16 | 0.06 | 0.07 | -0.01 | 0.39 | 1.44 | 0.02 | 1.03 | 0.00 | -0.94 | 0.12 | 0.30 |
| *Instruction-tuned Models* | | | | | | | | | | | | | | | | |
| NV-Embed-v2 | **0.68** | 2.76 | **0.65** | -1.10 | **0.71** | 13.70 | 0.07 | -0.35 | 0.54 | 0.60 | 0.54 | 0.72 | 0.40 | -5.19 | **0.51** | 1.59 |
| GritLM-7B | 0.63 | 3.09 | 0.63 | -0.06 | 0.70 | 15.10 | 0.10 | -0.32 | 0.51 | 4.01 | **0.57** | -0.09 | **0.42** | -0.32 | **0.51** | 3.06 |
| E5-mistral-7B | 0.54 | 4.26 | 0.63 | 0.05 | 0.69 | 14.14 | 0.10 | 0.08 | **0.57** | 6.31 | 0.56 | 0.92 | 0.28 | -4.42 | 0.48 | 3.05 |
| Instructor-XL | 0.48 | 1.03 | 0.49 | -2.36 | 0.53 | 0.35 | 0.07 | -0.30 | 0.53 | 1.96 | 0.17 | -2.06 | 0.19 | 0.18 | 0.35 | -0.17 |
| Instructor-Large | 0.49 | 3.65 | 0.46 | 0.20 | 0.56 | 3.68 | 0.07 | 0.29 | 0.51 | 2.19 | 0.15 | -3.86 | 0.17 | 6.70 | 0.34 | 1.84 |
| Promptriever-7B | 0.22 | **8.95** | 0.34 | **3.69** | 0.60 | **18.17** | 0.09 | -0.31 | 0.52 | 5.18 | 0.35 | **13.26** | 0.09 | 7.07 | 0.32 | **8.00** |
| Instructor-Base | 0.39 | 3.31 | 0.45 | 0.42 | 0.48 | 2.06 | 0.06 | 0.18 | 0.51 | -2.70 | 0.17 | 1.34 | 0.09 | **13.93** | 0.31 | 2.65 |
| *Proprietary Models* | | | | | | | | | | | | | | | | |
| OpenAI-v3-large | 0.54 | 1.57 | 0.59 | -0.48 | 0.58 | 0.31 | **0.11** | -0.03 | **0.57** | -0.03 | 0.52 | 0.18 | 0.30 | -5.72 | 0.46 | -0.60 |
| OpenAI-v3-small | 0.46 | 2.31 | 0.58 | -0.94 | 0.56 | 0.83 | 0.08 | -0.29 | 0.53 | 3.26 | 0.41 | -1.21 | 0.24 | -0.81 | 0.41 | 0.45 |

Table 3: Performance of retrievers on IFIR measured by nDCG@20 and INSTFOL@20 (%). Fine-grained results(*i.e.,* hybrid) can be found at Appendix C. Model performance is ranked based on average results with the nDCG metric.

inaccuracies introduced by LLMs, we use the Probability Normalization technique (as detailed in Appendix B ) to reduce overestimation, which has been proven effective in current works (Liu et al., 2023; Liusie et al., 2024). For each passage in the Top-$K$ retrieval results, the LLM is instructed to produce the relevance score. We then average the scores as the final matching score $S$. The INSTFOL @$K$ for the Top-$K$ retrieved passages is then calculated as:

$$\text{INSTFOL@K} = (S_{\text{inst}} - S_{\text{q}}) \cdot \alpha \qquad (1)$$

where $S_{\text{q}}$ measures how well the passages retrieved by the original query (without instruction) align with the given instruction; $S_{\text{inst}}$ measures how well the passages retrieved by the query (with instruction) meet the same instruction. The factor $\alpha$ is a normalization function that ensures the INSTFOL score ranges between -1 and 1. In practice, we set $\alpha$ as $\frac{1}{3-S_{\text{q}}}$. The implementation details of the INSTFOL metric are provided in Appendix B.

Our in-depth analysis in Section 6 demonstrates that INSTFOL exhibits a high correlation with human expert evaluations, highlighting its reliability.

## 4.2 Evaluated Retrievers

We evaluate a wide range of retrievers, with model sizes ranging from 110M to 7B parameters. These models are categorized into two main types:

**Non-instruction-tuned models** We include the following commonly-used non-instruction-tuned models for the experiments: (1) **BM25** (Robertson

et al., 2009), which is a lexical retriever; (2) **Col-BERT** (Khattab and Zaharia, 2020), which encodes queries and documents separately and introduces a mechanism of delayed interaction to be more effective; (3) **Contriever** (Izacard et al., 2022), which is a BERT-based model trained by contrastive learning; and (4) **GTR** (Ni et al., 2022), which uses the encoder from the T5 model and are pre-trained on MS MARCO (Bajaj et al., 2018).

**Instruction-tuned retrievers** For the instruction-tuned models, we select: (1) **INSTRUCTOR** (Su et al., 2023b), which are finetuned on the GTR family using MEDI datasets, and can be utilized for various tasks including retrieval; (2) **E5-mistral-7b-instruct** (Wang et al., 2023b), which is a retriever based on a Mistral model and trained on synthetic data. (3) **GritLM-7B** (Muennighoff et al., 2024), which is also a Mistral model, trained on the synthetic data from E5-mistral-7b-instruct and MEDI2, capable of performing both generation and retrieval tasks; (4) **Promptriever** (Weller et al., 2024b), is an instruction-trained bi-encoder retriever specialized for instruction-following tasks. The model we choose is Promptriever-7B. (5) **NV-Embed** (Lee et al., 2024) is a retriever trained by a two-stage contrastive instruction-tuning method. (6) **Proprietary Retrievers**, including OpenAI's Text-Embedding-v3-Large and Text-Embedding-v3-Small.

## 5 Experimental Results

This section first presents the key experimental results, followed by an in-depth analysis of model

performance, an error case study, and a reliability evaluation of the proposed IFIR metric.

## 5.1 Main Results

Table 3 presents the main results, from which we derive the following key findings:

**Non-instruction-tuned models** BM25 demonstrates relatively good performance, suggesting possible lexical bias in the datasets. Moreover, GTR models outperform BERT-based models. Unlike ColBERT and Contriever, which are trained solely on the MSMARCO dataset, GTR models also utilize the collected community question-answer pairs and Natural Question (Kwiatkowski et al., 2019) datasets. These datasets, which are more closely aligned with human interactions, may contribute to the superior performance of GTR models.

**Instruction-tuned models** (1) The INSTRUCTOR models show minimal improvement over their backbone (*i.e.,* GTR), and in some cases, even perform worse. This may indicate that the INSTRUCTOR models could be overfitting on specific datasets or are better suited to shorter instructions. (2) GritLM-7B, which is of the same size as E5-mistral-7b-instruct, demonstrates stronger performance on healthcare domains where E5-mistral-7b-instruct encounters difficulties. This performance gap may stem from GritLM-7B's inclusion of training data in extra domains, which likely boosts its capacity to handle healthcare content more effectively. (3) Notably, Promptriever-7b outperforms other open-source retrievers in the INSTFOL metric, while NV-Embed-v2 and GritLM-7B demonstrate outstanding performance in the nDCG metric. Promptriever-7b excels in INSTFOL due to its targeted training on instance-level instructions, enabling it to adjust relevance based on user input. (4) The proprietary retriever, OpenAI-v3-Large, achieves relatively strong performance on nDCG. However, both OpenAI's retrievers do not demonstrate superior performance on INSTFOL compared to other retrievers. Unfortunately, the technical details of them, including their training processes, are confidential, which limits our ability to fully understand the factors contributing to their performance.

**Overall** The current training methodologies that integrate instructions are not yet perfect solutions for handling long instructions across various domains. From the relatively good performance of BM25 on both metrics, we can deduce that lexi-

cal search may serve as an auxiliary tool for complex instructions in specific domains. Meanwhile, although some LLM-based retrievers do not perform well in traditional metrics like nDCG, they exhibit a superior and stable instruction-following ability compared to other retrievers. Additionally, the instruction-trained method introduced by Promptriever is highly intuitive and effective, holding promise for future integration into IR systems.

## 5.2 Analysis

**Scaling up model size leads to better retrieval performance.** From Table 3, we can conclude that the scaling law applies to retrievers as well. Specifically, as model sizes increase from 110M to 1B, both the GTR and INSTRUCTOR models demonstrate improved nDCG metrics. Additionally, LLM-based retrievers(*e.g.,* E5-mistral-7b-instruct, GritLM-7B) exhibit relatively strong performance on average. However, when considering instruction-following ability, the scaling law does not apply when the model size is below the 1B threshold. Given the strong performance of GritLM-7B and Promptriever in instruction-following ability, it can be inferred that the current retrieval system can be further enhanced by LLMs finetuned for retrieval tasks.

**Existing instruction training methods are still limited.** Currently, some retrievers, such as INSTRUCTOR and GritLM-7B, are trained with instructions like "Retrieve document from Wikipedia" or "Classify the question's topic" to fit the varying demands of different domains. We investigate how significantly such training methods can enhance performance across various domains. Accordingly, as described in these works, we incorporate these instructions as prompts in both the query and embedding processes. We format the input as "`[Prompt] [Query] [Instruction]`". The prompt is actually the instruction in these works which gives hints to target tasks and domains, *e.g.,* "Represent the science question for retrieval. " which is different from our instruction. We use instruction in these works as a prompt to check whether this is an enhancement compared to embedding with no prompt. The results are shown in Figure 3, with detailed outcomes available in the Appendix C.2. We observe that adding instructions does not significantly impact the final performance. Additionally, for LLM-based retrievers (*i.e.,* GritLM), performance even declines.
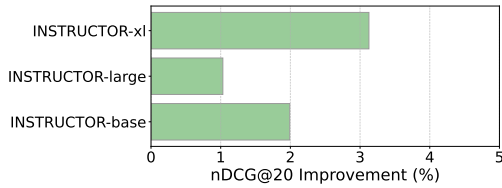
Figure 3: The nDCG improvement when the model is provided with detailed instructions for retrieval.
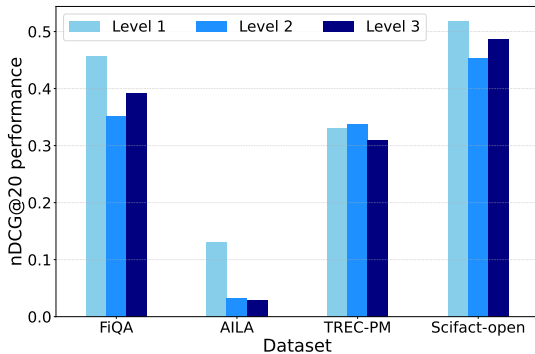


Figure 4: Average nDCG@20 performance on different levels of instructions in different domains.

Therefore, for various domains, merely adding minimal instructions is insufficient. Domain-specific datasets and more complex instructions are required for different domains.

**Increase in instruction complexity results in performance decline.** As discussed in §3.2, to test the instruction-following abilities of different retrieval systems with finer granularity, we construct instructions with varying complexity levels. As shown in Figure 4, there is a noticeable performance degradation with level2 and level3 instructions compared to level1. Interestingly, in some subsets, level3 performs better than level2. This improvement is attributed to the fact that level3 instructions are longer and contain more explicit conditions, providing more hints about possible candidates. Overall, we observe that retrievers fine-tuned from LLMs exhibit robust and superior performance compared to other models. This excellent performance may result from LLMs' general capabilities and long-context abilities. Detailed results can be found in the Appendix C.3.

### 5.3 Error Analysis

We select those instructions with both low nDCG scores and INSTFOL scores and create a taxonomy of these instructions, categorizing them as (1)

Long Instructions. In the legal domain, some instructions exceed 1,024 tokens due to the inclusion of lengthy legal case. Current retrievers are typically trained with a maximum token length of 512, which cannot perfectly handle these lengthy instructions. (2) Dense with Specialized Knowledge. For instructions that require specialized knowledge, especially in the science literature and healthcare domains, common training data do not cover all the expert knowledge needed in specialized domains. (3) Highly Customized Instructions, as illustrated in Appendix B. In finance and healthcare domains, users or doctors have several prioritized goals and needs that traditional retrievers may not recognize.

## 6 Reliability Analysis of INSTFOL

To evaluate the reliability of the proposed LLM-based evaluation metric, INSTFOL, we conduct a thorough analysis. Our focus is on INSTFOL's core component: the LLM's ability to score the relevance between an instruction-following query and a passage. We validate these LLM-based relevance scores by comparing them with human evaluation scores. Specifically, we randomly sample 400 query-passage pairs and assign each to one domain expert for evaluation. Both the human evaluators and the LLM (*i.e.,* GPT-4o) are instructed to rate relevance on a scale of 1 to 5. After collecting the scores from both sources, we evaluate their alignment at the instance level. Notably, we observe a strong Pearson correlation coefficient of 0.704, indicating that INSTFOL can effectively assess a retriever's instruction-following capabilities.

## 7 Conclusion

We introduce IFIR, a novel IR benchmark designed to assess the instruction-following capabilities of retrievers. IFIR focuses on domain-specific instructions, reflecting the diverse needs across various fields. Our experiments reveal that current instruction-tuned models struggle with long, complex instructions. Moreover, as the complexity increases, a noticeable performance decline occurs across all tested retrieval systems. However, LLM-based retrievers demonstrate more robust performance and relatively better results compared to other models. This suggests potential solutions for end-to-end specialized-domain retrieval scenarios.

## Limitations

In this section, we list three limitations of this study, each of which opens avenues for future improvements. First, our dataset has a limited number of queries accompanied by instructions, and we do not provide a training dataset for future works to train their models. Second, we do not compare the performance of domain-specific retrievers like BioBERT (Lee et al., 2020) with general-purpose retrievers. Domain-specific retrievers, equipped with specialized training data, may achieve superior results in niche fields. Future work should explore the integration and evaluation of such specialized models, particularly in domains like healthcare or law, where domain knowledge is crucial. Finally, there are some relevance judgment gaps (Abbasiantaeb et al., 2024) in the selected seed datasets. Since we select relevant passages based on previously annotated passages, this approach may lead to some relevant passages being ignored. Future work could investigate ways to improve relevance judgment, potentially through the use of more crowd-sourced human annotation.

## References

Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can we use large language models to fill relevance judgment holes? *arXiv preprint arXiv:2405.05600*.

AI@Meta. 2024. The llama 3 herd of models.

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

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Fire 2019 aila track: Artificial intelligence for legal assistance. In *Proceedings of the 11th annual meeting of the forum for information retrieval evaluation*, pages 4–6.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

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

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview and discussion of the competition on legal information, extraction/entailment (coliee) 2023. *The Review of Socionetwork Strategies*, 18(1):27–47.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Bogdan Ionescu, Henning Müller, Ana-Maria Drăgulinescu, Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, et al. 2024. Overview of the imageclef 2024: multimedia retrieval in medical applications. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 140–164. Springer.

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

Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Aspect-based financial

sentiment analysis using deep learning. In *Companion Proceedings of the The Web Conference 2018*, pages 1961–1966.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. 2024. Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

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

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Bulou Liu, Yueyue Wu, Yiqun Liu, Fan Zhang, Yunqiu Shao, Chenliang Li, Min Zhang, and Shaoping Ma. 2021. Conversational vs traditional: Comparing search behavior and outcome in legal case retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1622–1626.

Hao Liu, Ali Soroush, Jordan G Nestor, Elizabeth Park, Betina Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, et al. 2024. Retrieval augmented scientific claim verification. *JAMIA open*, 7(1):ooae021.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the fire 2017 irled track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68.

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

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

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings*

*of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen 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*, 27(9):1431–1436.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the TREC 2018 precision medicine track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*, volume 500-331 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. In *The... text REtrieval conference: TREC. Text REtrieval Conference*, volume 26. NIH Public Access.

Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the trec 2015 clinical decision support track. In *TREC*.

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

Shreya Saxena, Raj Sangani, Siva Prasad, Shubham Kumar, Mihir Athale, Rohan Awhad, and Vishal Vaddina. 2022. Large-scale knowledge synthesis and complex information retrieval from biomedical documents. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2364–2369. IEEE.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023a. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2023b. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval.

Lynda Tamine and Lorraine Goeuriot. 2021. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023a. Doris-mae: scientific document retrieval using multi-level aspect-based queries. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 38404–38419.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

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 preprint arXiv:2403.15246*.

Orion Weller, Benjamin Chang, Eugene Yang, Mahsa Yarmohammadi, Sam Barham, Sean MacAvaney, Arman Cohan, Luca Soldaini, Benjamin Van Durme, and Dawn Lawrie. 2025a. mfollowir: a multilingual benchmark for instruction following in retrieval.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025b. Rank1: Test-time compute for reranking in information retrieval.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024b. Promptriever: Instruction-trained retrievers can be prompted like language models. *arXiv preprint arXiv:2409.11136*.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. 2024. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv preprint arXiv:2401.13478*.

Chaojun Xiao, Haoxiang Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. Cail2019-scm: A dataset of similar case matching in legal domain. *ArXiv*, abs/1911.08962.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Xinran Zhao, Tong Chen, Sihao Chen, Hongming Zhang, and Tongshuang Wu. 2024a. Beyond relevance: Evaluate and improve retrievers on perspective awareness. *arXiv preprint arXiv:2405.02714*.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024b. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.

# A IFIR Benchmark

## A.1 Instruction Generation

We first ask LLM (*i.e.,* GPT-4o) to generate an instruction according to the reality demands, with the prompts on dataset NFCoprus shown in Figure 5. And then we check whether previous annotated passages for the query satisfy the generated instruction, with the prompts shown in Figure 6.

For the datasets with multi-levels, including FiQA, SciFact-open, AILA, and TREC-PM, we do not construct a fully complex instruction all at once. Since we have different levels of reasoning, we ask the LLM to detail the instruction level by level, akin to a bottom-up approach. For example, for a given instruction at level 2, such as "Please help me to find the plaintiff's beneficial legal case." we request the LLM to generate a more complex instruction for the next level that also includes the "plaintiff's beneficial" condition. The prompt for instruction generation on dataset AILA is shown in Figure 7.

## A.2 Details of Instructions in Each Domain

Instruction examples are shown in Table 5. For the legal datasets, which belong to the legal domain, the query part consists of only a summary or is omitted due to the length context of legal cases.

Different complexity instruction examples are shown in Table 6. To emphasize, we describe the content of each level again. As the level increases, so do the conditions. Target passages in Level 3 must satisfy the conditions of Levels 1 and 2, and Level 2 candidates must satisfy Level 1 conditions. (1) **FiQA**: The first level simply asks for financial suggestions. The second level includes information about personal financial status. The third level incorporates personal financial purposes. (2) **Scifact-open**: The first level involves asking for science passages relevant to a given science claim. The second level seeks evidence that either contradicts or supports this claim. The third level is tailored for students or researchers who need to find evidence based on customized demands. (3) **AILA**: The first level involves searching for similar cases. The second level requires that the relevant case be beneficial for the plaintiff or defendant. The third level adds more explicit conditions such as the details of the current cases, potential goals, and more similar scenarios. (4) **TREC-PM**: The first level includes information about the patient's disease. The second level adds the patient's demographics, including age and gender. The third level incorporates additional information about the patient's treatment history and family history.

# B Implementation Details

**Embedding** To accommodate the long context of certain passages, we employ a sliding window of 512 tokens with an overlap of 128 tokens, using mean pooling to generate embeddings. For LLM-based retrievers, however, we do not apply mean pooling due to their extended context window. Given hardware constraints, we run LLM-based retrievers in FP16 mode to reduce GPU memory usage. When querying, we concatenate the query and instruction with a space character. Because of the large number of passages and the correspondingly large embeddings, we use LangChain and Elasticsearch in our experiments. In addition, we also provide the BEIR code.

**Implementation Details of INSTFOL** Given $C_q$, which is the retrieved passages set for a query, and $C_{inst}$, which is the retrieved corpus set for the query combined with instructions, we evaluate each passage in both sets using LLM (*i.e.,* `gpt-4o-mini`) with some evaluation criteria. This approach is inspired by G-Eval (Liu et al., 2023) and TREC's principles of data collection. We then obtain two sets of weighted scores, $S_q$ and $S_{inst}$. $S_q$ measures how well the passages retrieved by the original query (without instruction) align with the given instruction; $S_{inst}$ measures how well the passages retrieved by the query (with instruction) meet the same instruction. Each score in $S_q$ and $S_{inst}$ is calculated as follows, where $s_k$ is an element of $S_q$ or $S_{inst}$, and $p_k$ represents the logarithmic probability of each score as determined by the API:

$$\text{weighted\_score} = \frac{\sum_{k=1}^{10}(s_k \times e^{p_k})}{\sum_{k=1}^{10} e^{p_k}}$$

Here, $e^{p_k}$ converts the log probabilities back to standard probabilities for calculation purposes. This formula accounts for the inherent probabilistic nature of LLMs, where predictions for each token are based on a statistical probability distribution influenced by configurations such as temperature and top_p.

We use the average of $S_q$ and $S_{inst}$ to calculate the instruction-following ability through a metric we propose, called INSTFOL. The insight is to consider the maximum improvement a retriever can

```
[system prompt]
You are an expert in science.

[user input]
Given the scientific claim: {claim}, Imagine you are a student or researcher seeking information on a specific topic.
Based on the conditions listed below, construct a detailed retrieval instruction tailored to the claim. You do not need
to incorporate all of the conditions, but ensure your instruction is relevant.
* Research fields
* Research topics
* Research objectives
* Customized needs (For example, experimental subjects, experimental methods, etc.)

The instruction should target a single type of information and be both coherent and logical. It should also be
detailed and specific, presented in the first person, and narrated naturally in one paragraph.

Please return your answer as follows:
Instruction: ...
```

Figure 5: Prompt for generating instruction on NFCorpus dataset.

```
[system prompt]
You are an expert in science.
[user input]
Given an instruction {instruction}, and an corpus {corpus}, check whether the instruction is satisfied by the corpus.
Please only return 'yes' or 'no' and your reason, and return in the following format.
Answer: yes/no Reason: ...
```

Figure 6: Prompt for evaluating corpus on NFCorpus dataset.

```
[system prompt]
You are an expert in the legal domain.
[user input]
Given a legal case:
{case}.
And an instruction: {instruction}.

Please provide a detailed instruction based on the case. Include specific situations from the case to elaborate on the
instruction. Your response should be narrated as if you are examining various cases, and it should be presented in a
single paragraph.
The instruction should not be longer than 2-3 sentences.

For example:
Legal Case: "XYZ Corporation filed a lawsuit against John Doe, a former employee, for defamation after Doe
posted allegations on his personal social media claiming that XYZ provided false information to customers. While
Doe's post didn't result in significant or immediate financial loss for the company, XYZ argued that it tarnished
their reputation."
Instruction: "I'm the plaintiff's lawyer and I'm looking for civil tort cases involving the right to reputation and
lowered social evaluation, particularly where an employee posted on social media that the company made false
statements in providing services but no serious consequences occurred, and it's difficult to prove the lowered social
evaluation."
Your response should be formatted as follows:
Instruction: ...
```

Figure 7: Prompt for instruction generation on AILA dataset.

| ID | Fluent in English | Major | Annotation tasks |
|---|---|---|---|
| 1 | > 10 yrs | Legal | Legal subset annotation |
| 2 | > 7 yrs | Legal | Legal subset annotation |
| 3 | > 10 yrs | Biology | Healthcare subset annotation; Science Literature subset annotation; Annotation Validation |
| 4 | > 7 yrs | Pharmacy | Healthcare subset annotation; Science Literature subset annotation |
| 5 | > 10 yrs | Biomedical engineering | Healthcare subset annotation; Science Literature subset annotation; Annotation Validation |
| 6 | > 7 yrs | Mathematics | Financial subset annotation; Annotation Validation. |
| 7 | > 6 yrs | Finance | Financial subset annotation; |
| 8 | Native Speaker | Finance | Financial Subset Annotation |
| 9 | > 10 yrs | Legal | Legal subset annotation |

Table 4: Human annotator's tasks

achieve. Consider a case with two students, A and B. Student A has a rank of 300 and a previous rank of 500, while student B has a rank of 10 and a previous rank of 40. Traditionally, we would calculate improvement through absolute differences. However, student B has less room to improve his rank. Based on this insight, we propose the INST-FOL metric to evaluate the retriever's instruction-following ability. The factor $\alpha$ is a normalization function that ensures the INSTFOL score ranges between -1 and 1. In practice, we set $\alpha$ as $\frac{1}{3-S_q}$.

$$\text{INSTFOL@K} = (S_{\text{inst}} - S_q) \cdot \alpha \qquad (2)$$

When calling the API to evaluate the INSTFOL, we use top_p = 0.7 and top_logprobs = 5. We set the temperature to 0.0 to reduce the overestimation by the LLM. The prompt for evaluation is shown in Figure 8.

**Error Analysis** The example of error analysis is illustrated in Table 7.

## C Details of Experimental Results

### C.1 Detailed Results of IFIR

In addition to the results obtained using INSTFOL, we also present the outcomes of traditional IR metrics, as shown in Table 8. Furthermore, beyond the *end-to-end* retrieval method, we implement a *hybrid* retrieval approach utilizing BM25. An overall improvement in performance can be observed for GTR-xl, Instructor-xl, and Promptriever compared to the *end-to-end* method. However, as shown in Table 8, performance varies across tasks. This highlights the importance of designing a carefully tailored hybrid retrieval pipeline to meet the demands of real-world scenarios.

### C.2 Detailed Results of Retrievers with Instructions as Prompts

As shown in Table 9, we present the instructions as prompts, as described in these papers. However, we use these instructions as prompts to differentiate from our instructions. The input to the retrievers should be formatted as "[Prompt] [Query] [Instruction]." Additionally, there may be slight differences in the input format due to different models.

### C.3 Detailed results of different retrievers on different levels.

The detailed result for each domain is shown in Table 10. The result of INSTFOL is shown in Table 11. From Table 11, we can find that current information retrievers are not good at long instructions and instructions with highly dense expert knowledge.

| Dataset | query | instruction |
|---|---|---|
| FiQA | Full-time work + running small side business: Best business structure for taxes? | As a 40-year-old accountant with a steady income and moderate savings, I am seeking advice on the best business structure for taxes when combining full-time work with running a small side business. I am particularly interested in understanding the tax implications, legal considerations, and potential benefits of different business structures. Additionally, I am looking for insights on how to optimize tax efficiency while balancing the demands of my full-time job and side business. |
| SciFact-open | A deficiency of folate decreases blood levels of homocysteine. | As an expert in the field of science, I need to find a peer-reviewed research article or a review paper that presents contradicting evidence regarding the relationship between folate deficiency and homocysteine levels in the blood. The passage should offer evidence that opposes the claim stating that a deficiency of folate results in decreased blood levels of homocysteine. |
| NFCorpus | Why are Cancer Rates so Low in India? | I am a student researching the factors contributing to low cancer rates in India, and I am specifically interested in understanding the role of dietary habits. I need to find scientific studies or articles from the fields of oncology, nutrition, and epidemiology that focus on the relationship between Indian dietary patterns and cancer prevention. My objective is to analyze the types of foods commonly consumed in India and their potential protective effects against cancer. To meet my customized needs, I require information on specific dietary components, such as spices, fruits, vegetables, and traditional Indian dishes, that have been associated with lower cancer rates. Additionally, I am interested in any experimental studies or clinical trials investigating the effects of these dietary factors on cancer cells or animal models. |
| AILA | The appellant, once a prime witness in a bribery trial, became a Cabinet Minister and resigned after critical judicial remarks during an appeal that acquitted the first respondent. The High Court questioned the evidence and the appellant's credibility, overturning the initial conviction for accepting bribes. | I represent the appellant and I seek cases involving a defendant who benefitted from a reversal of a conviction due to lack of acceptable evidence and a plausible explanation for the incriminating evidence found in their possession, despite adverse remarks made by the Appellate Judge regarding the credibility of the appellant's testimony in a bribery case where the defendant was acquitted based on insufficient prosecution evidence. |
| FIRE | [A legal case summary] What was the decision and legal principle established in the case referred to as [?CITATION?] in relation to the doctrine of promissory estoppel in the context of government representations and obligations? | Retrieve the prior case referred to as [?CITATION?] and focus on the court's analysis and ruling regarding the application of promissory estoppel against the government, particularly in situations where representations are made by governmental authorities and the subsequent obligations arising from such representations. Pay attention to any discussion on the enforceability of promises made by the government, the limitations of promissory estoppel against the government, and the factors determining the applicability of the doctrine in cases involving governmental representations. |
| TREC-PM | A patient diagnosed with Liposarcoma with CDK4 Amplification. I am looking for possible clinical trials suitable for this patient. | I am seeking clinical trials for a 38-year-old male diagnosed with Liposarcoma with CDK4 Amplification. Please focus on trials specifically targeting Liposarcoma or related soft tissue sarcomas. It is crucial that the trials consider the presence of CDK4 Amplification in the patient's condition. Additionally, the patient's age and gender should be taken into account when selecting suitable clinical trial options. Patient Profile: The patient is a 38-year-old male who has been diagnosed with Liposarcoma with CDK4 Amplification. He has a treatment background that includes both chemotherapy and radiation, and he is currently in remission. It is important to note that he has a history of smoking and is also dealing with obesity. Given these demographic details, I am seeking clinical trials that specifically target Liposarcoma or related soft tissue sarcomas, taking into consideration the presence of CDK4 Amplification. The trials should also consider the patient's age and gender, as well as any potential influences from his treatment background, smoking history, and obesity. |
| TREC-CDS | Given some information about patient. 58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.What is the patient's diagnosis? | A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes. |

Table 5: Examples of instructions in different domains.

10201

| Dataset | level1 | level2 | level3 |
|---|---|---|---|
| FiQA | Please help me to find the financial suggestions for my query. | I am a 40-year-old accountant with a steady income and moderate savings. | As a 40-year-old accountant with a steady income and moderate savings, I am seeking advice on the best business structure for taxes when combining full-time work with running a small side business. I am particularly interested in understanding the tax implications, legal considerations, and potential benefits of different business structures. Additionally, I am looking for insights on how to optimize tax efficiency while balancing the demands of my full-time job and side business |
| SciFact-open | Please find the science passage which related to the claim | Please help me to find the contradict evidence. | As an expert in the field of science, I need to find a peer-reviewed research article or a review paper that presents contradicting evidence regarding the relationship between folate deficiency and homocysteine levels in the blood. The passage should offer evidence that opposes the claim stating that a deficiency of folate results in decreased blood levels of homocysteine. |
| AILA | Please help me find the relevant legal cases. | As a defendant player, I want the case where the defendant is beneficial. | I represent the appellant and I seek cases involving a defendant who benefitted from a reversal of a conviction due to lack of acceptable evidence and a plausible explanation for the incriminating evidence found in their possession, despite adverse remarks made by the Appellate Judge regarding the credibility of the appellant's testimony in a bribery case where the defendant was acquitted based on insufficient prosecution evidence. |
| TREC-PM | I'm looking for clinical trials suitable for a 38-year-old male patient diagnosed with Liposarcoma with CDK4 Amplification. | I am seeking clinical trials for a 38-year-old male diagnosed with Liposarcoma with CDK4 Amplification. Please focus on trials specifically targeting Liposarcoma or related soft tissue sarcomas. It is crucial that the trials consider the presence of CDK4 Amplification in the patient's condition. Additionally, the patient's age and gender should be taken into account when selecting suitable clinical trial options. | I am seeking clinical trials for a 38-year-old male diagnosed with Liposarcoma with CDK4 Amplification. Please focus on trials specifically targeting Liposarcoma or related soft tissue sarcomas. It is crucial that the trials consider the presence of CDK4 Amplification in the patient's condition. Additionally, the patient's age and gender should be taken into account when selecting suitable clinical trial options. Patient Profile: The patient is a 38-year-old male who has been diagnosed with Liposarcoma with CDK4 Amplification. He has a treatment background that includes both chemotherapy and radiation, and he is currently in remission. It is important to note that he has a history of smoking and is also dealing with obesity. Given these demographic details, I am seeking clinical trials that specifically target Liposarcoma or related soft tissue sarcomas, taking into consideration the presence of CDK4 Amplification. The trials should also consider the patient's age and gender, as well as any potential influences from his treatment background, smoking history, and obesity. |

Table 6: Examples for different levels' instruction in various domains.

| Type | Example |
|---|---|
| Long Instruction | [A long legal case] As the defendant player, seek cases where the prosecution's evidence relies heavily on circumstantial evidence and lacks direct proof of intent or direct involvement in the alleged crime, similar to a situation where the accused individuals were convicted based on circumstantial evidence and witness testimonies, despite maintaining their innocence throughout the trial and appeal process. |
| Dense with specialized knowledge | CHEK2 has a significant role in breast cancer As a scientist investigating the claim that 'CHEK2 has a significant role in breast cancer,' I should search for research articles or review papers that provide support evidence on the specific functions of the CHEK2 gene in relation to breast cancer development. |
| Highly customized instructions | I am seeking clinical trials suitable for a 35-year-old female diagnosed with colorectal cancer and exhibiting FGFR1 Amplification. Please prioritize trials that focus on colorectal cancer specifically or a narrower focus related to this patient's condition. Additionally, it is crucial to include trials that directly match the FGFR1 Amplification gene mutation in the patient. The patient's age and gender are also important factors to consider in selecting appropriate clinical trials. Please ensure that the trials selected meet these criteria for optimal patient care and treatment options. |

Table 7: Taxonomy of instructions with low nDCG score and INSTFOL score.

```
[system prompt]
You are an expert in legal domain.

[user input]
Given an instruction: instruction,
and a prior case: corpus,
please evaluate the prior case according to the instruction and Evaluation Criteria and return a JSON object with the score
and reason.

There are 3 relevant levels to evaluate the case regarding the instruction:
1. The prior case is similar to the one in the instruction.
2. The prior case satisfies the instruction at the 'plaintiff' or 'defendant' beneficial level.
3. The prior case totally matches the instruction, including the detailed requirements in the instruction.

Evaluation Criteria:
1. If the prior case only meets the instruction at the first level, the score is 1.
2. If the prior case meets the instruction at the first and second levels, the score is 2.
3. If the prior case meets the instruction at all three levels, the score is 3.
4. If the prior case does not meet any of the levels, the score is 0.

Please give a score between 0 and 3.

**
IMPORTANT: Please make sure to only return in JSON format, with the "score" and "reason" key. No additional words or
explanations are needed.
Please think step by step about the reason and give the score according to the Evaluation Criteria.

Example JSON:

"score": 1,
"reason": "The corpus only matches the instruction in terms of research field and research topics."

**

JSON:
```

Figure 8: Prompt for instruction generation on AILA dataset.

| | FiQA | | SciFact-open | | NFCorpus | | AILA | | FIRE | | TREC-PM | | TREC-CDS | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | nDCG | MRR | nDCG | MRR | nDCG | MRR | nDCG | MRR | nDCG | MRR | nDCG | MRR | nDCG | MRR | nDCG | MRR |
| *End-to-end retrieval* | | | | | | | | | | | | | | | | |
| GTR-XL | 0.44 | 0.40 | 0.54 | 0.49 | 0.60 | 0.57 | 0.05 | 0.04 | 0.54 | 0.50 | 0.31 | 0.27 | 0.15 | 0.12 | 0.37 | 0.34 |
| BM25 | 0.25 | 0.21 | 0.49 | 0.45 | 0.43 | 0.40 | 0.10 | 0.08 | 0.55 | 0.51 | 0.47 | 0.43 | 0.07 | 0.05 | 0.34 | 0.30 |
| GTR-Large | 0.39 | 0.34 | 0.50 | 0.46 | 0.51 | 0.46 | 0.07 | 0.07 | 0.49 | 0.41 | 0.28 | 0.23 | 0.09 | 0.06 | 0.33 | 0.29 |
| GTR-Base | 0.33 | 0.28 | 0.47 | 0.43 | 0.47 | 0.42 | 0.05 | 0.04 | 0.52 | 0.47 | 0.27 | 0.24 | 0.13 | 0.09 | 0.32 | 0.28 |
| Contriever | 0.13 | 0.10 | 0.29 | 0.24 | 0.36 | 0.29 | 0.08 | 0.06 | 0.51 | 0.48 | 0.09 | 0.06 | 0.04 | 0.04 | 0.21 | 0.18 |
| ColBERT | 0.07 | 0.05 | 0.14 | 0.12 | 0.16 | 0.13 | 0.07 | 0.05 | 0.39 | 0.35 | 0.02 | 0.01 | 0.00 | 0.00 | 0.12 | 0.10 |
| NV-Embed-v2 | **0.68** | **0.66** | **0.65** | **0.62** | **0.71** | **0.70** | 0.07 | 0.04 | 0.54 | 0.51 | 0.54 | 0.53 | 0.40 | 0.36 | **0.51** | **0.49** |
| GritLM-7B | 0.63 | 0.61 | 0.63 | 0.60 | 0.70 | 0.69 | 0.10 | 0.06 | 0.51 | 0.46 | **0.57** | **0.54** | **0.42** | **0.38** | **0.51** | 0.48 |
| E5-mistral-7b-inst | 0.54 | 0.51 | 0.63 | 0.59 | 0.69 | 0.67 | 0.10 | 0.06 | 0.57 | **0.54** | 0.56 | 0.52 | 0.28 | 0.23 | 0.48 | 0.45 |
| Instructor-XL | 0.48 | 0.44 | 0.49 | 0.44 | 0.53 | 0.47 | 0.07 | 0.05 | 0.53 | 0.49 | 0.17 | 0.12 | 0.19 | 0.15 | 0.35 | 0.31 |
| Instructor-Large | 0.49 | 0.45 | 0.46 | 0.42 | 0.56 | 0.51 | 0.07 | 0.06 | 0.51 | 0.45 | 0.15 | 0.11 | 0.17 | 0.12 | 0.34 | 0.30 |
| Promptriever-7B | 0.22 | 0.17 | 0.34 | 0.28 | 0.60 | 0.59 | 0.09 | 0.07 | 0.52 | 0.48 | 0.35 | 0.29 | 0.09 | 0.06 | 0.32 | 0.28 |
| Instructor-Base | 0.39 | 0.34 | 0.45 | 0.42 | 0.48 | 0.42 | 0.06 | 0.05 | 0.51 | 0.46 | 0.17 | 0.13 | 0.09 | 0.07 | 0.31 | 0.27 |
| OpenAI-v3-large | 0.54 | 0.51 | 0.59 | 0.54 | 0.58 | 0.55 | **0.11** | **0.08** | **0.57** | 0.54 | 0.52 | 0.46 | 0.30 | 0.26 | 0.46 | 0.42 |
| OpenAI-v3-small | 0.46 | 0.41 | 0.58 | 0.52 | 0.56 | 0.52 | 0.08 | 0.06 | 0.53 | 0.48 | 0.41 | 0.38 | 0.24 | 0.21 | 0.41 | 0.37 |
| *Hybrid retrieval* | | | | | | | | | | | | | | | | |
| GritLM-7B | **0.59** | **0.54** | **0.62** | **0.58** | **0.57** | **0.51** | **0.10** | 0.07 | 0.55 | 0.50 | **0.59** | **0.56** | **0.39** | **0.32** | **0.49** | **0.44** |
| GTR-XL | 0.43 | 0.38 | 0.57 | 0.51 | 0.52 | 0.47 | 0.06 | 0.06 | **0.56** | **0.53** | 0.33 | 0.29 | 0.14 | 0.13 | 0.37 | 0.34 |
| Instructor-XL | 0.46 | 0.41 | 0.54 | 0.51 | 0.52 | 0.46 | 0.09 | **0.07** | 0.56 | 0.52 | 0.23 | 0.20 | 0.20 | 0.16 | 0.37 | 0.33 |
| Promptriever-7B | 0.25 | 0.21 | 0.42 | 0.38 | 0.54 | 0.49 | 0.09 | 0.07 | 0.56 | 0.53 | 0.42 | 0.37 | 0.11 | 0.09 | 0.34 | 0.31 |

Table 8: Performance of retrievers on IFIR measured by nDCG@20 and MRR@20.

| | FiQA | SciFact-open | NFCorpus | AILA | FIRE | TREC-PM | TREC-CDS | Average |
|---|---|---|---|---|---|---|---|---|
| Instructor-Base | 0.392 | 0.451 | 0.482 | 0.059 | 0.506 | 0.174 | 0.091 | 0.308 |
| | 0.393 | 0.445 | 0.489 | 0.059 | 0.499 | 0.232 | 0.080 | 0.314 |
| Instructor-Large | 0.488 | 0.463 | 0.564 | 0.070 | 0.510 | 0.149 | 0.167 | 0.345 |
| | 0.493 | 0.469 | 0.567 | 0.072 | 0.516 | 0.166 | 0.155 | 0.348 |
| Instructor-XL | 0.484 | 0.488 | 0.530 | 0.071 | 0.529 | 0.169 | 0.188 | 0.351 |
| | 0.489 | 0.489 | 0.544 | 0.072 | 0.533 | 0.204 | 0.205 | 0.362 |
| E5-mistral-7b-inst | 0.541 | 0.629 | 0.686 | 0.103 | 0.565 | 0.563 | 0.283 | 0.481 |
| | 0.495 | 0.607 | 0.679 | 0.108 | 0.574 | 0.569 | 0.293 | 0.475 |
| GritLM-7B | 0.632 | 0.631 | 0.698 | 0.096 | 0.511 | 0.575 | 0.423 | 0.509 |
| | 0.567 | 0.612 | 0.681 | 0.093 | 0.516 | 0.571 | 0.425 | 0.495 |

Table 9: Detailed nDCG@20 results of adding instructions as prompt. The first line is without instruction as a prompt, the second is with instruction as a prompt.

| | FiQA | | | AILA | | | TREC-PM | | | Scifact-open | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 |
| BM25 | 0.282 | 0.221 | 0.239 | 0.158 | 0.060 | 0.030 | 0.505 | 0.437 | 0.482 | 0.568 | 0.438 | 0.480 |
| Contriever | 0.146 | 0.121 | 0.111 | 0.144 | 0.018 | 0.012 | 0.112 | 0.084 | 0.077 | 0.306 | 0.252 | 0.325 |
| ColBERT | 0.078 | 0.043 | 0.107 | 0.111 | 0.052 | 0.012 | 0.012 | 0.023 | 0.037 | 0.132 | 0.121 | 0.165 |
| GTR-Base | 0.422 | 0.215 | 0.337 | 0.096 | 0.017 | 0.000 | 0.269 | 0.280 | 0.268 | 0.511 | 0.448 | 0.456 |
| GTR-Large | 0.479 | 0.279 | 0.391 | 0.117 | 0.023 | 0.048 | 0.293 | 0.312 | 0.219 | 0.538 | 0.467 | 0.513 |
| GTR-XL | 0.530 | 0.366 | 0.413 | 0.073 | 0.032 | 0.023 | 0.350 | 0.324 | 0.243 | 0.595 | 0.512 | 0.515 |
| Instructor-Base | 0.424 | 0.361 | 0.387 | 0.110 | 0.024 | 0.000 | 0.119 | 0.208 | 0.197 | 0.481 | 0.429 | 0.449 |
| Instructor-Large | 0.531 | 0.454 | 0.472 | 0.110 | 0.024 | 0.048 | 0.144 | 0.157 | 0.147 | 0.480 | 0.404 | 0.513 |
| Instructor-XL | 0.558 | 0.435 | 0.445 | 0.122 | 0.012 | 0.042 | 0.190 | 0.181 | 0.135 | 0.536 | 0.461 | 0.476 |
| E5-mistral-7b-inst | 0.628 | 0.490 | 0.488 | 0.162 | 0.041 | 0.060 | 0.582 | 0.567 | 0.537 | 0.673 | 0.620 | 0.601 |
| GritLM-7B | 0.705 | 0.594 | 0.581 | 0.185 | 0.014 | 0.017 | 0.618 | 0.578 | 0.527 | 0.699 | 0.620 | 0.583 |
| Promptriever-7B | 0.164 | 0.144 | 0.377 | 0.140 | 0.051 | 0.048 | 0.231 | 0.406 | 0.417 | 0.287 | 0.276 | 0.449 |
| NV-Embed-v2 | 0.759 | 0.657 | 0.594 | 0.116 | 0.026 | 0.029 | 0.563 | 0.547 | 0.493 | 0.714 | 0.634 | 0.625 |
| OpenAI-v3-small | 0.529 | 0.410 | 0.429 | 0.148 | 0.032 | 0.013 | 0.428 | 0.436 | 0.371 | 0.631 | 0.545 | 0.563 |
| OpenAI-v3-large | 0.616 | 0.488 | 0.511 | 0.158 | 0.056 | 0.062 | 0.553 | 0.523 | 0.483 | 0.621 | 0.578 | 0.587 |

Table 10: Detailed nDCG@20 results of different retrievers on different levels.

| | FiQA | | | AILA | | | TREC-PM | | | Scifact-open | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 | Level1 | Level2 | Level3 |
| BM25 | -3.76 | -2.48 | 9.53 | -0.08 | -0.15 | 0.29 | 0.83 | 2.28 | 7.17 | -0.70 | -5.36 | 4.85 |
| Contriever | -2.04 | -0.69 | 4.28 | 0.07 | -0.11 | 0.30 | 0.07 | 1.91 | 1.75 | -3.20 | -11.25 | -10.21 |
| ColBERT | -4.43 | -7.28 | 12.21 | 0.24 | 0.85 | -1.13 | 0.14 | 1.11 | 1.84 | -0.23 | -1.06 | 2.31 |
| GTR-Base | -3.76 | -12.39 | 2.30 | -0.06 | 0.51 | 1.18 | -4.56 | -0.12 | 2.73 | -1.09 | -0.39 | -0.39 |
| GTR-Large | -6.40 | -13.20 | 0.18 | 0.12 | -0.31 | -0.88 | -3.37 | -0.28 | -0.73 | -0.18 | 0.64 | -0.30 |
| GTR-XL | -4.45 | -10.59 | 0.49 | 0.15 | -0.20 | -0.34 | -3.21 | -0.73 | 1.40 | -0.14 | 0.46 | -9.12 |
| Instructor-Base | -1.43 | 1.52 | 9.83 | 0.16 | -0.23 | 0.61 | -4.31 | 1.64 | 6.70 | -0.94 | 0.66 | 1.54 |
| Instructor-Large | -0.34 | 2.49 | 8.79 | -0.22 | 0.04 | 1.04 | -2.38 | -2.77 | -6.42 | -1.15 | -2.07 | 3.83 |
| Instructor-XL | -0.66 | -1.93 | 5.67 | 0.03 | -0.07 | -0.86 | -2.00 | -0.82 | -3.37 | 0.53 | 2.34 | -9.95 |
| E5-mistral-7b-inst | 0.14 | -0.13 | 12.78 | 1.11 | -0.08 | -0.79 | -0.89 | 1.10 | 2.55 | -0.22 | 0.19 | 0.18 |
| GritLM-7B | 0.25 | 1.68 | 7.32 | -0.37 | 0.06 | -0.66 | -0.75 | 0.14 | 0.33 | -0.36 | 0.92 | -0.73 |
| Promptriever-7B | -2.41 | 2.18 | 27.09 | -1.22 | 0.37 | -0.07 | 2.09 | 12.49 | 25.21 | -2.02 | -5.17 | 18.27 |
| NV-Embed-v2 | 0.14 | 1.13 | 7.02 | -0.83 | -0.20 | -0.02 | 0.37 | 0.44 | 1.35 | -0.54 | 0.51 | -3.27 |
| OpenAI-v3-small | -0.60 | -0.62 | 8.16 | 0.12 | 0.10 | -1.09 | -1.93 | -0.80 | -0.88 | 0.26 | 1.72 | -4.79 |
| OpenAI-v3-large | -0.95 | -0.97 | 6.63 | 0.54 | -0.12 | -0.50 | -1.72 | -1.37 | 3.64 | -0.19 | 0.75 | -1.99 |

Table 11: Detailed INSTFOL @20 results of different retrievers on different levels.