IFIR-EVAL: Evaluating Information Retrieval Models for Instruction Following in Specialized Domains

Anonymous ACL submission

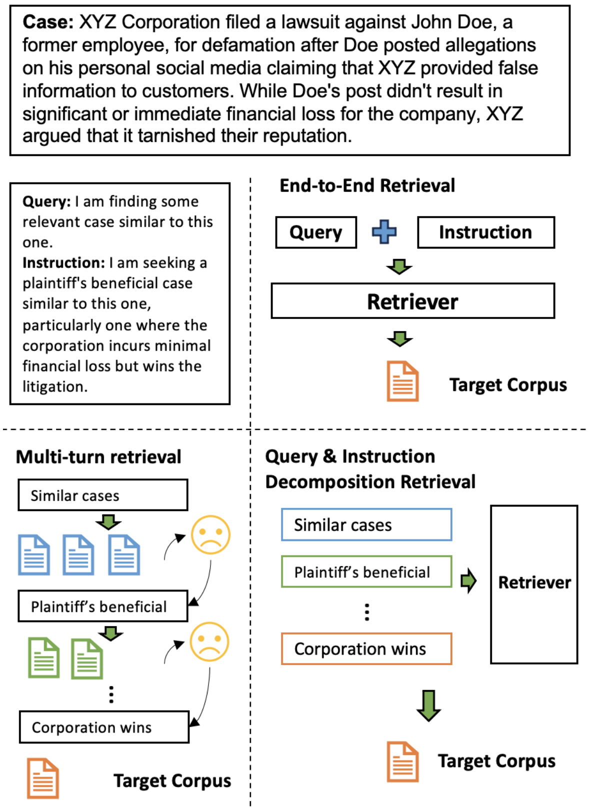
Abstract

Despite the recent success of aligning large language models (LLMs) with human instructions, the ability of information retrievers to follow instructions has not been fully explored. To address this gap, we propose IFIR-EVAL, a comprehensive information retrieval benchmark that spans eight subsets across four expert domains: finance, legal, healthcare, and science-literature. Each subset tackles one or more domain-specific retrieval task in real-world scenarios where user-customized instructions are essential. To enable a comprehensive assessment of retrievers' instruction-following abilities, we also construct instructions with different complexity levels. Realizing the limitations of traditional IR metrics for evaluating instruction-following capability, we propose a new LLM-based evaluation method, INSTFOL. We conduct a comprehensive experiments including a wide range of information retrievers. Our experimental results demonstrate that LLM-based retrievers have good potential to follow instructions. However, current information retrieval systems are still far from achieving optimal performance in handling complex instructions. [1]

1 Introduction

The instruction-following ability has become a cornerstone for LLMs, allowing them to interpret and respond to complex user commands and enabling them to handle a variety of user-specific tasks (Ouyang et al., 2022; OpenAI, 2023; AI@Meta, 2024).

Despite its critical importance, the instruction-following capability is still under-explored in the context of information retrieval (IR). Current information retrievers struggle to meet the nuanced needs of users in real-world applications, especially in fields like legal research, healthcare, and academic study, where precise and context-aware retrieval is crucial (Saxena et al., 2022; Mysore et al., 2022). For example, in legal research, lawyers often search for target cases with detailed instruction (*i.e.,* based on specific legal criteria, context,



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: An example of the instruction-following IR scenarios investigated in this study. In this example, we simulate a real-world legal case search scenario in which the user has an explicit need for specific legal cases as described in the instruction. However, current methods, such as multi-turn retrieval and the decomposition approach, are complex and introduce instability into the search process. Our goal is to identify and evaluate the best *end-to-end* retrieval method that can address this challenge effectively.

[1] Code and dataset will be available at https://anonymous.4open.science/r/IFIR-EVAL-C0E1/

and desired outcomes), as illustrated in Figure 1. Traditional IR systems fails to fully understand and process such user-specific instructions in an *end-to-end* manner. Consequently, users have to decompose their information-seeking needs into several simple search queries and manually filter the retrieved cases, which is time-consuming and inefficient.

Recent pilot studies have investigated the instruction-following capabilities of retrievers; however, significant limitations remain. First, existing evaluation benchmarks utilize instructions that consist of either a single sentence (Su et al., 2023a) or a set of keywords (Zhao et al., 2024). These settings oversimplify real-world requirements, where instructions from users in specialized domains are typically more nuanced and layered (Wang et al., 2023a). This lack of complexity in the evaluation process hinders a comprehensive assessment of the retrievers' capabilities to handle multi-dimensional criteria, leading to an incomplete understanding of their performance in real-world applications. Even though some works do feature complex instructions (Weller et al., 2024; Oh et al., 2024), they lack explicit complexity levels to qualify information retrievers' ability in terms of reasoning or handling intricate tasks. Furthermore, the instructions in these datasets are not tailored to meet the specific needs of specialized domains.

In this work, we introduce IFIR-EVAL, a benchmark designed to evaluate the instruction-following capabilities of current information retrievers, particularly in the context of specialized domains. Specifically, we construct a benchmark consisting of eight subsets covering four specialized domains, namely finance, scientific literature, legal, and biomedical. Moreover, in order to measure retrievers' ability with finer granularity, we construct three complexity levels in four subsets. IFIR-EVAL includes 2424 instruction-following queries, each averaging 6.14 ground-truth passages. To enhance dataset quality, we conduct a comprehensive human validation during the benchmark construction process. Moreover, since traditional methods are inadequate for measuring instruction-following abilities and LLMs are capable of making such assessments, we implement INSTFOL, an LLM-based metric derived from G-Eval (Liu et al., 2023), to precisely measure the instruction-following performance of these systems.

We evaluate a wide range of information retrievers on IFIR-EVAL, including lexical, semantic, and proprietary retrievers. Through our experiments, we derive four key findings: (1) BM25 performs relatively well in expert domains because the instructions contain more glossary terms than the queries alone, providing additional hints. (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) A performance drop is observed when instruction complexity increases. (4) LLM-based retrievers demonstrate more robust performance on traditional metrics and also have a good instruction-following ability compared to other traditional retrievers, such as Contriever (Izacard et al., 2022) and INSTRUCTORs (Su et al., 2023b).

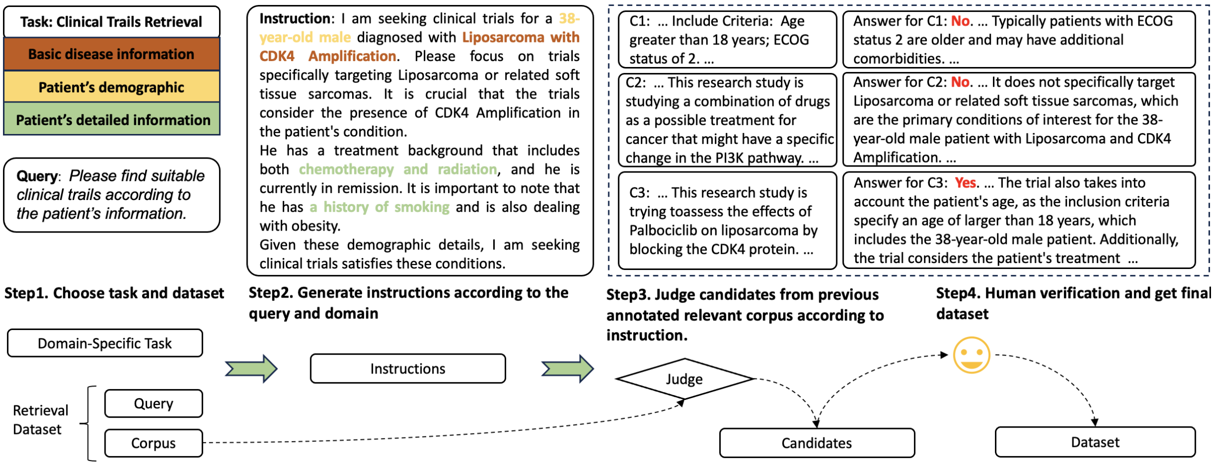
We conclude our main contributions as follows:

- We introduce IFIR-EVAL, 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 LLMs in *end-to-end* retrieval.

2 Related Work

2.1 IR Benchmarks in Specialized Domain

IR plays a crucial role in specialized domains by enabling efficient access to domain-specific knowledge, facilitating evidence-based decision-making, and accelerating research and innovation. In recent years, IR benchmarks tailored to specialized domains have garnered significant attention. For instance, in the legal field, LeCaRDv2 (Li et al., 2023) focuses on similar case matching, while in finance, FiQA (Jangid et al., 2018) addresses investment thesis retrieval. In scientific literature, SciFact (Wadden et al., 2020) targets the retrieval of research papers for verifying specific claims. However, the queries within existing IR benchmarks typically lack specific instructions, which is a critical gap when compared to real-world scenarios


Task: Clinical Trails Retrieval

Basic disease information

Patient's demographic

Patient's detailed information

Query: *Please find suitable clinical trails according to the patient's information.*

Instruction: 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.
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 satisfies these conditions.

C1: … Include Criteria: Age greater than 18 years; ECOG status of 2. …

Answer for C1: No. … Typically patients with ECOG status 2 are older and may have additional comorbidities. …

C2: … This research study is studying a combination of drugs as a possible treatment for cancer that might have a specific change in the PI3K pathway. …

Answer for C2: No. … It does not specifically target Liposarcoma or related soft tissue sarcomas, which are the primary conditions of interest for the 38-year-old male patient with Liposarcoma and CDK4 Amplification. …

C3: … This research study is trying toassess the effects of Palbociclib on liposarcoma by blocking the CDK4 protein. …

Answer for C3: Yes. … The trial also takes into account the patient's age, as the inclusion criteria specify an age of larger than 18 years, which includes the 38-year-old male patient. Additionally, the trial considers the patient's treatment …

Step1. Choose task and dataset
Step2. Generate instructions according to the query and domain
Step3. Judge candidates from previous annotated relevant corpus according to instruction.
Step4. Human verification and get final dataset
Domain-Specific Task
Retrieval Dataset
Query
Corpus
Instructions
Judge
Candidates
Dataset

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.

in specialized domains. For example, when a doctor searches for suitable clinical trials for a patient, they are currently limited to using keywords or short queries instead of specifying detailed patient information. To bridge this gap, we focus on benchmark domain specific tasks which have customized demands.

2.2 Instruction-Following IR

The instruction-following abilities have become a cornerstone for LLMs, enabling them to interpret and respond to complex user commands for specialized tasks (Ouyang et al., 2022; OpenAI, 2023; AI@Meta, 2024). There has been a growing interest in investigating and enhancing instruction-following capabilities of information retrievers. Researchers have proposed several training techniques to enhance the instruction-following abilities of retrievers (Su et al., 2023a; Asai et al., 2023; Wang et al., 2023c). However, due to the lack of instruction-following IR benchmark at the time of model development, they rely on BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) for evaluation, which lack complex instructions and don't reflect customized needs in various domains. To bridge this gap, more recently, several instruction-following IR benchmarks have been proposed (Oh et al., 2024; Weller et al., 2024). Specifically, INSTRUCTIR re-writes query according to characteristics based on existing retrieval dataset, addresses the need to serve individuals from diverse backgrounds. But for one instruction, there is only one golden passage, which does not align with reality FOLLOWIR introduces more complex instructions and corpus setups. They revise instructions from TREC and measure the instruction following ability through re-rank task instead of retrieval. However, they do not adequately consider the relationship between tasks and their respective domains. Moreover, how to adequately evaluate the reasoning capabilities of retriever is under-explored. Consequently, we focus on several expert domains of significant public interest and their representative retrieval tasks. We also explore more effective evaluation methods to measure the instruction-following ability.

3 IFIR-EVAL Benchmark

Recognizing the gap between current IR benchmarks and the need for instruction-following capabilities in specialized domains, we propose IFIR-EVAL. Our goal is to create a comprehensive evaluation benchmark that advances the instruction-following capabilities of retrievers, pushing the boundaries of current IR systems. We present an overview of the IFIR-EVAL construction pipeline in Figure 2, and detail the benchmark construction and quality validation process in the following subsections.

3.1 Expert-domain IR Corpus Collection

As mentioned above, IFIR-EVAL has four domains including finance, science-literature, legal and healthcare. To construct nuanced and complex instructions, we select tasks where customized needs are common and essential, as shown in Ta-

Domain	Dataset	# Q	# Instruct	# Corpus	# G	Retrieval Task Description
Finance	FiQA (Jangid et al., 2018)	639	1,718	57,638	3.53	Financial suggestion. Retrieve suggestions fit for customized financial demands.
Scientific Literature	SciFact-open (Wadden et al., 2022)	45	149	500,000	4.84	Given customized demands, retrieve relevant passages or evidence.
	NFCorpus (Boteva et al., 2016)	86	86	3,633	2.81	
Legal	AILA (Bhattacharya et al., 2019)	40	85	2,914	2.01	Given a legal case, retrieve cases which satisfy both customized demands and similarity.
	FIRE (Mandal et al., 2017)	168	168	1,745	3.36	Given a legal case, retrieve cases to support judicial decision.
Healthcare	TREC-PM (Roberts et al., 2017, 2018)	59	172	241,006	15.61	Given a patient's demographic, retrieve suitable clinical trails.
	TREC-CDS (Roberts et al., 2015)	43	43	633,955	10.84	Given a patient's demographic, retrieve diagnosis, treatment and clinical trails.

Table 1: List of existing information retrieval datasets adopted within the IFIR-EVAL benchmark. # G is the average golden passage number of one instruction.

ble 1. However, building a comprehensive retrieval dataset from scratch across multiple domains can be a time-intensive process. To balance efficiency and data quality, we choose to construct our IFIR-EVAL by building upon existing, well-established retrieval datasets.

3.2 Instruction Generation

A key insight in generating instructions is to create additional conditions relevant to the domain-specific tasks. The instructions, which complement the original query, help differentiate some corpora from those previously annotated as relevant to the query. As illustrated in Figure 2, we generate and paraphrase instructions by utilizing LLM to add extra information according to the original query and domain-specific tasks. Detailed instructions can be found in Appendix A.1.

To measure the retrievers' ability to follow instructions with varying complexity levels, we generate instructions with different tiers in each domain, which we will detail as follows:

Finance For personal finance inquiries, we construct instructions that simulate someone seeking help in the finance field. We add basic demographic information such as gender and age, and expand the query with their goals and basic financial status. We create instructions at three levels: At the first level, the instruction is simple, *e.g.*,, "Please help me to find a financial suggestion for the query." The second level includes additional personal information such as age, occupation, and financial status. The third level builds upon level two by incorporating additional financial goals, such as "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 For the scientific literature retrieval task, we construct instructions to simulate a person working in a relevant area trying to find passages related to specific 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 instruction might state, "Please help me to 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 conditions like research topics and objectives.

Legal For the legal case retrieval task, 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 such as lawyers. For the first type, we use the context around the previously annotated reasoning process and paraphrase it with GPT-4o. 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 Given the two different datasets and corresponding tasks in the biomedical field, the TREC CDS Track provides a summary accompanied by a detailed description, which we use directly as the instruction, adding information including family medical history, etc. 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 about 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.

3.3 Selecting Corpus Candidates According to Instructions

Due to the long context and large number of potential responses, which may lead to performance drops in LLM, we don't give all the corpus at one time. Instead, we judge the corpus relevance one by one. We also ask LLM[1] to generate reasons for including and excluding each candidate.

3.4 Human Validation

And to ensure the quality of our dataset, a human verification stage is implemented in the validation process. We extract 10% of the queries with instructions from each domain. To align with real world demands, we ask human annotators to give a naturalness score of instruction ranging from 1 to 5. And to validate the corpus, human annotators need to check both the golden passages and excluded passages with the same range of score. The higher the score, the better the match. We illustrate these scores in Table 2, which indicates the high quality of our dataset.

Metric	Score
Naturalness of instructions	4.17
Golden passage matching score	4.52
Excludeded passage score	4.32

Table 2: Human Validation Results. The naturalness of instructions is evaluated based on how well the instructions align with real-world demands. The golden passage matching score assesses whether the golden passage correctly matches the corresponding instruction. The excluded passage score evaluates how effectively irrelevant passages are rejected.

3.5 Dataset Analysis

As illustrated in Table 1, we construct the dataset across several domains. In finance field, we have 3.54 average corpus number for a single query, and at least 2.81 in science, 2.01 in legal and 10.84 in biomedical. In each field, we have domain-specific instructions for representative tasks, which guarantee the diversity of instructions and the task to be representative and challenging.

4 Experiments

4.1 Problem Formulation

We formally define the two tasks as follows:

Task 1 Measuring Instruction-Following Capabilities: To measure retrievers' instruction-following ability, we propose a new LLM-based evaluation metric named InstFol. Given retrieved corpus C_q from a query, and retrieved corpus C_{inst} from a query accompanied by an instruction, we allow the LLM to decide each corpus C_k in C_q and C_{inst}'s matching score to the instruction. We then measure the difference between the average scores of C_q and C_{inst} to evaluate the instruction-following ability. The implementation details of the InstFol metric are illustrated in the A.2.

Task 2 Measuring Retrieval Capabilities: Given a query Q, and instruction I, and the golden passages G, and a retriever R, we measure the retrieved passages from golden passages use normalized discounted cumulative gain (nDCG) to measure the performance.

4.2 Baseline

We compare different retrievers on our dataset, with size ranges from 110M to 7B. In our experiment, we categorize the models into two types: non-instruction-tuned models and instruction-tuned models. Detailed experiment settings are provided in Appendix A.2.

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) **ColBERT** (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 pretrained on MSMARCO.

[1] The LLM we use is GPT-4o.

	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
BM25	0.25	10.42	0.49	1.66	0.43	-1.72	0.10	-0.58	0.55	-0.12	0.47	18.81	0.07	-3.37	0.34	3.58
Contriever	0.13	-0.45	0.29	-39.07	0.36	-3.59	0.08	0.26	0.51	0.18	0.09	9.20	0.04	-10.22	0.21	-6.24
Colbert	0.07	-5.10	0.14	-2.21	0.16	3.05	0.07	-0.01	0.39	2.11	0.02	11.42	0.00	1.43	0.12	1.53
GTR-base	0.33	-8.47	0.47	-0.45	0.47	-19.48	0.05	-0.04	0.52	1.38	0.27	10.94	0.12	-14.58	0.32	-4.39
GTR-large	0.39	-12.31	0.50	-2.92	0.52	-45.47	0.07	-0.05	0.49	4.61	0.28	8.25	0.09	-70.01	0.33	-16.84
GTR-xl	0.40	-20.94	0.52	-10.64	0.51	-14.17	0.05	-0.56	0.54	0.42	0.23	6.98	0.15	-77.08	0.34	-16.57
INSTRUCTOR-base	0.39	9.86	0.45	-7.91	0.48	2.20	0.06	-0.81	0.51	1.07	0.17	13.38	0.09	-36.72	0.31	-2.70
INSTRUCTOR-large	0.49	4.79	0.46	1.97	0.56	4.13	0.07	-0.30	0.51	0.74	0.15	-11.14	0.17	2.92	0.34	0.45
INSTRUCTOR-xl	0.48	2.36	0.48	1.18	0.53	-11.51	0.07	0.00	0.53	0.90	0.17	0.04	0.19	-3.58	0.35	-1.52
E5-mistral-7b-instruct	0.42	9.96	0.45	-0.57	0.51	22.09	0.10	0.08	0.50	2.67	0.32	6.46	0.08	0.58	0.34	5.90
GritLM-7B	0.49	10.47	0.47	11.87	0.57	38.34	0.10	0.14	0.59	0.96	0.44	7.00	0.36	-3.27	0.43	9.36
OpenAI-v3-small	0.46	6.92	0.58	-0.60	0.56	-14.00	0.08	-0.25	0.53	1.78	0.41	1.03	0.24	-6.51	0.41	-1.66
OpenAI-v3-large	0.54	-0.57	0.59	-10.18	0.58	-8.98	0.11	-0.36	0.57	0.03	0.52	5.23	0.30	-5.29	0.46	-2.87

Table 3: Results of retrievers on IFIR-EVAL with metrics being nDCG@20 and INSTFOL. The best-performing entries are highlighted in red, and model's with worst performance are in blue. The deeper the color, the better the performance—for example, dark red indicates better performance, while dark blue indicates worse performance.

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; and (4) **Proprietary Retriever**, including OpenAI's Text-Embedding-v3-Large and Text-Embedding-v3-Small.

4.3 Main Results

We evaluate various retrievers' ability to follow instructions. The result is shown in Table 3.

Discussion About Non-instruction-tuned models BM25 demonstrates better performance compared to ColBERT and Contrievers, suggesting possible lexical bias in the datasets. Moreover, GTR models outperform BERT-based models. Unlike ColBERT and Contriever, which are trained solely on the MS-MARCO dataset, GTR models also utilize the Community QA and Natural Question datasets. These datasets, which are more closely aligned with human interactions, may contribute to the superior performance of GTR models.

Discussion About Instruction-tuned models The instruction-tuned models, particularly the INSTRUCTOR models, exhibit relatively good performance on datasets. Furthermore, the finetuned LLMs for retrieval tasks outperform other retrievers. For example, E5-mistral-7b-instruct performs well on IFIR-EVAL. One exception is the TREC-CDS dataset, where detailed patient descriptions require more reasoning ability rather than just lexical or semantic matching. GritLM-7B outperforms other open-source retrievers. Notably, GritLM-7B, which is of the same size as E5-mistral-7b-instruct, shows relatively strong performance in scientific literature subsets where E5-mistral-7b-instruct struggles. This performance difference may be attributed to the fact that GritLM-7B includes the Semantic Scholar Open Research Corpus (Lo et al., 2020) in its training data, which likely enhances its ability to handle scientific content effectively. Meanwhile, retrievers built on LLM backbones demonstrate more robust and generalizable instruction-following capabilities compared to other models, offering valuable insights into potential methods for enhancing the instruction-following abilities of retrievers. The proprietary retriever, OpenAI-v3-Large achieves the best performance on nDCG metrics among all models. However, both OpenAI-v3-small and OpenAI-v3-large do not demonstrate superior performance on INSTFOL compared to other retrievers. Unfortunately, the technical details of the OpenAI retrievers, including their training processes, are confidential, which limits our ability to fully understand or analyze the factors contributing to their performance.

Overall The current training methodology that integrates instructions is not yet a perfect solution for handling long instructions across various domains. From the relatively good performance of BM25 on both metrics, we can deduce that lexical search may serve as an auxiliary tool for complex instructions in specific domains. The INSTRUCTOR models show minimal improvement over their backbone, the GTR models, and in some cases, even perform worse. This may indicate that

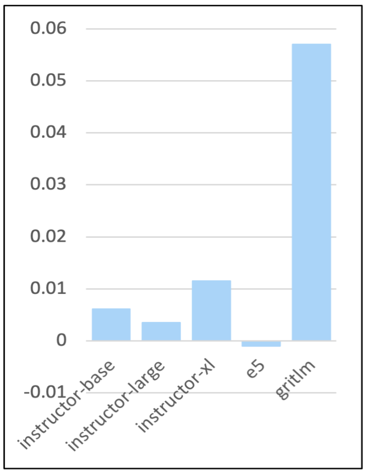


Figure 3: nDCG improvement when giving instructions. The instruction specifies domains like "scientific" and object like "for retrieval".

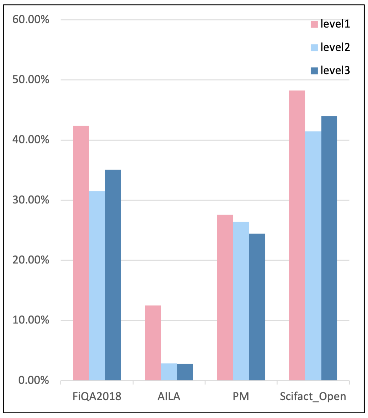


Figure 4: Average nDCG perfomance on different levels of instructions in different domains.

the INSTRUCTOR models could be overfitting on specific datasets or are better suited to shorter instructions. Meanwhile, although some LLM-based retriever do not perform well in traditional metrics like nDCG, they exhibit a superior and stable instruction-following ability compared to other retrievers.

4.4 Analysis

Scaling Up Model Size Leads to Better 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 models and INSTRUCTOR demonstrate improved nDCG metrics. Additionally, E5-mistral-7b-instruct and 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 1B threshold. Given the strong performance of E5-mistral-7b-instruct and GritLM-7B in instruction-following ability, it can be inferred that the current retrieval system can be further enhanced by LLMs finetuned for retrieval tasks.

Current Instruction Retriever is Inadequate for Long Instructions 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 method can enhance performance across various domains. Accordingly, we incorporate these instructions as prompts in both the query and embedding processes, as described in these works. We format the input as “[Prompt] [Query] [Instruction]". The prompt is actually the instruction in these works which give 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 these is an enhancement compared to embedding with no prompt. The results are shown in Figure 3, with detailed outcomes available in the Appendix B.1. We observe that adding instructions did not significantly impact the final performance, with the maximum improvement being only 0.05 on average results. 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 To test the instruction-following abilities of different retrieval systems with finer granularity, we constructed instructions with varying levels of reasoning difficulty by incorporating both implicit and explicit conditions. We selected the FiQA, Scifact-open, AILA, and TREC-PM datasets, constructing different levels of instructions within these datasets for convenience. As shown in Figure 4, there is a noticeable performance degradation with level2 and level3 instructions compared to level1. Interestingly, in some datasets, level3 performs better than level2. This improvement is attributed to the fact that level3 instructions are longer and contain more explicit conditions, providing additional lexical and semantic hints about possible candidates, unlike level2, which includes some explicit instructions but fewer

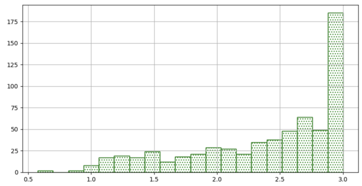


Figure 5: Scores of non-golden passages retrieved by Contreiver on the NFCorpus dataset.

	Not Annotated Ratio	Golden Passage Average Score
FiQA	0.96	2.89
Scifact-open	1.00	2.97
NFCorpus	0.87	2.90
AILA	0.97	2.19
FIRE	0.56	2.60
TREC-PM	0.95	2.84
TREC-CDS	0.96	2.95

Table 4: Analysis of the Contreiver's INSTFOL scores. The Not Annotated Ratio records the proportion of passages with unexpectedly high values that are not annotated in the original golden passages. The Golden Passage Average Score reflects the average INSTFOL score of retrieved golden passages for the corresponding instruction.

explicit conditions. Overall, from these metrics, we observe that retrievers finetuned from LLMs exhibit robust and superior performance compared to other models. This excellent performance is particularly evident in their ability to adapt to complex instructions and to maintain high accuracy across diverse datasets, underlining the effectiveness of LLM-based architectures in handling nuanced natural language processing tasks. Detailed results can be found in the Appendix B.2.

Validation of LLM-based Evaluation As shown in Figure 5, some retrieved passages have relatively high scores. As we derive our dataset from an existing retrieval dataset and selected possible candidates from the golden passages, we check whether these higher-scoring passages were not annotated or there was an overestimation problem in our method. After human validation, we find that the passages do satisfy the instructions. We then analyze how many passages were not annotated in the original dataset and investigated their proportion among all over-estimated passages. Additionally, we check the average scores of retrieved golden passages annotated by us. As shown in Table 4, a large number of overestimated passages were not annotated in the seed dataset, and the scores of the instructions' corresponding golden passages are close to 3, indicating that these passages do satisfy the instructions. These results also validate our dataset quality.

4.5 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, (2) Dense with Specialized Knowledge, (3) Highly Customized Instructions, as illustrated in Appendix A.3. In the legal domain, a large number of instructions exceed 1,000 tokens. Current retrievers are typically trained with a maximum token length of 512, which cannot perfectly handle these lengthy instructions. For instructions that require specialized knowledge, especially in the science and healthcare domains, common training data do not cover all the expert knowledge needed in specialized domains. And for the highly customized instructions, such as those in the finance and healthcare domains, users or doctors have several prioritized goals and needs that traditional retrievers may not recognize. However, using an LLM as the retriever backbone can utilize its general capabilities and long-context abilities. LLMs are also adept at intent recognition, making them potential candidates for the backbone in these *end-to-end* retrieval scenarios.

5 Conclusion

In this paper, we propose a benchmark, IFIR-EVAL, designed to evaluate the instruction-following abilities of current information retrievers in *end-to-end* retrieval scenarios. While existing benchmarks do evaluate complex instructions, they often do not construct instructions that align with the demands of specialized domains. Our benchmark, therefore, focuses on domain-specific instructions, reflecting the diverse needs across various fields. Unlike previous work, we also design instructions with varying difficulty levels to assess the performance of current retrievers under different challenges.

Our experiments reveal that current instruction-tuned models struggle with long, complex instructions. And 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* retrieval scenarios in specialized domains.

Limitations

There are three major limitations of our benchmark. 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 domain-specific retrievers like BioBERT (Lee et al., 2020) with general retrievers. A domain-specific retriever may perform better than general retrievers due to additional training data. Third, we only evaluate current retrieval systems using the *end-to-end* method. As mentioned earlier, other methods can also be used to solve complex instruction-based problems.

References

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

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.

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.

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.

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.

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*.

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.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983.

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*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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, 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.

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, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

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. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*.

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

Appendix

A Implementation details

A.1 Details of Instructions in Each Domain

Details of instructions are shown in Table 5. For the AILA and FIRE 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.

For instructions with different complexity levels, examples are shown in Table 6. For clearer demonstration, we describe the content of each level again. As the level increases, so do the conditions. Target corpus candidates 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 and requires 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.

A.2 Experiment Settings

Embedding Considering the long context of some corpora, we opt for a sliding window of 512 tokens with an overlap of 128 tokens, and utilize the mean pooling method to generate the embeddings. For the E5-mistral-7b-instruct, we compared the mean pooling method to the last token pooling method and found a significant difference in performance. Consequently, we use the last token pooling method for E5-mistral-7b-instruct, which demonstrates better overall performance. Due to hardware limitations, for the LLM-based retrievers, we use fp16 to reduce GPU memory usage. When querying, we concatenate the query and instruction with a space character.

LLM-Based Evaluation Method Given C_{q}, which is the retrieved corpus set for a query, and C_{inst}, which is the retrieved corpus set for the query combined with instructions, we evaluate each passage or corpus in both sets using LLM[1] with some evaluation criteria. This approach is inspired by G-Eval (Liu et al., 2023) and TREC's principles of data collection. Details of prompts can be found in Appendix **??**. We then obtain two sets of weighted scores, S_q and S_{inst}. And we

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 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 INSTFOL metric to evaluate the retriever's instruction-following ability. The max, which is a constant, denotes the corpus that satisfies all requirements and has the highest relevance level. K is a constant representing the number of passages.

$$\textsc{InstFol} = \frac{Average(S_i) - Average(S_q)}{\text{max} - Average(S_q)}$$

When calling the API to evaluate the INSTFOL, we use top_p = 0.7 and top_logprobs = 10. We set the temperature to 0.0 to reduce the overestimation by the LLM. And the prompt for evaluation is shown in Figure 6

A.3 Error Analysis

The example of error analysis is illustrated in Table 7.

B Details of Experimental Results

B.1 Detailed Results of Retrievers with Instructions as Prompts

As shown in Table 8, we present the instructions as prompts, as described in these papers. However, we use these instructions as prompts to differentiate from our own 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.

B.2 Detailed results of different retrievers on different levels.

The detailed result for each domain is shown in Table 9. The result of INSTFOL is shown in Table 10.

[1] The LLM used for INSTFOL is GPT-3.5-0125

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 infomation 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.

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.

	FiQA	SciFact-Open	NFCorpus	AILA	FIRE	TREC-PM	TREC-CDS	Average
INSTRUCTOR-base	39.2	45.14	48.23	5.89	50.6	17.39	9.11	30.79
	39.33	44.52	48.94	5.93	49.92	23.16	8.0	31.4
INSTRUCTOR-large	48.76	46.4	56.43	7.03	51.01	14.95	16.7	34.47
	49.28	46.85	56.67	7.24	51.63	16.56	15.46	34.81
INSTRUCTOR-xl	48.37	48.46	53.04	7.09	52.89	16.89	18.83	35.08
	48.91	48.93	54.36	7.17	53.3	20.39	20.55	36.23
E5-mistral-7b-instruct	42.49	44.94	50.67	9.7	50.25	31.48	7.93	33.92
	41.57	44.56	50.2	8.66	50.25	33.97	7.52	33.82
GritLM-7B	49.06	46.83	57.46	10.01	59.32	43.89	35.69	43.18
	61.78	59.38	63.17	9.74	59.09	51.2	37.82	48.88

Table 8: Detailed results of adding instructions as prompt. The first line is without instruction as prompt, the second is with instructions as prompt.

[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 6: Prompt for instruction generation on AILA dataset.

From Table 10, we can find that current information retrievers are not good at long instructions and instructions with highly dense expert knowledge.

C Details of Data Construction Pipeline

We first ask LLM[1] to generate a instruction according to the reality demands , with the prompts on dataset NFCoprus shown in Figure 7. And then we check whether previous annotated candidates for the query satisfies the generated instruction, with the prompts shown in Figure 8.

For the reason datasets, which include 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 9.

[1] The LLM we use is GPT-4o

	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	0.06	0.03	0.505	0.437	0.482	0.568	0.434	0.481
Contriever	0.146	0.121	0.111	0.144	0.018	0.012	0.112	0.084	0.077	0.306	0.248	0.33
Colbert	0.078	0.043	0.107	0.111	0.052	0.012	0.012	0.023	0.037	0.132	0.117	0.165
GTR-base	0.422	0.215	0.337	0.096	0.017	0.0	0.269	0.28	0.268	0.511	0.446	0.458
GTR-large	0.479	0.279	0.391	0.117	0.023	0.048	0.293	0.312	0.219	0.538	0.465	0.512
GTR-xl	0.53	0.33	0.325	0.073	0.032	0.023	0.255	0.239	0.21	0.595	0.508	0.461
INSTRUCTOR-base	0.424	0.361	0.387	0.11	0.024	0.0	0.119	0.208	0.197	0.481	0.437	0.441
INSTRUCTOR-large	0.531	0.454	0.472	0.11	0.024	0.048	0.144	0.157	0.147	0.48	0.404	0.515
INSTRUCTOR-xl	0.558	0.435	0.445	0.122	0.012	0.042	0.19	0.181	0.135	0.536	0.457	0.47
E5-mistral-7b-instruct	0.491	0.409	0.359	0.16	0.037	0.045	0.404	0.304	0.232	0.514	0.47	0.369
GritLM-7B	0.527	0.424	0.52	0.176	0.02	0.047	0.451	0.418	0.447	0.507	0.416	0.491
OpenAI-v3-small	0.529	0.41	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	0.512	0.159	0.056	0.062	0.552	0.524	0.479	0.622	0.57	0.587

Table 9: 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	-3.3	2.2	32.4	-0.1	-0.5	-1.1	0.7	10.2	45.5	-0.1	-1.4	6.5
Contriever	-0.3	-0.7	-0.3	0.6	-0.2	0.4	-1.6	2.3	26.9	-7.2	-26.3	-83.8
Colbert	-6.1	-21.1	11.9	-0.3	-0.1	0.4	0.5	3.2	30.6	-2.0	-2.9	-1.7
GTR-base	-0.3	-11.7	-13.4	-0.1	0.2	-0.2	-0.5	5.3	28.0	0.8	0.5	-2.6
GTR-large	-0.6	-13.6	-22.8	0.0	-0.2	0.0	-0.9	2.4	23.3	0.3	-0.5	-8.5
GTR-xl	-0.6	-7.7	-54.5	0.5	-1.5	-0.6	-0.7	-1.1	22.7	-0.2	2.3	-34.0
INSTRUCTOR-base	0.0	1.4	28.2	0.0	-1.1	-1.3	-0.4	6.2	34.3	-0.4	0.4	-23.7
INSTRUCTOR-large	0.1	1.6	12.7	0.1	-0.5	-0.5	-0.6	-6.1	-26.7	-2.4	-0.2	8.6
INSTRUCTOR-xl	-0.0	-2.3	9.4	-0.4	-1.0	1.4	-0.2	-1.0	1.3	-0.6	4.3	-0.2
E5-mistral-7b-instruct	0.9	8.2	20.8	0.1	-0.6	0.8	-0.0	-0.4	19.8	0.9	1.0	-3.6
GritLM-7B	-0.4	0.9	30.9	0.5	-0.3	0.2	-0.3	2.7	18.6	-0.5	-1.0	37.1
OpenAI-v3-small	0.0	-0.9	21.6	0.4	0.0	-1.2	-0.3	0.3	3.1	1.0	0.8	-3.7
OpenAI-v3-large	0.0	0.1	-1.9	0.3	-1.0	-0.4	0.1	-1.4	17.0	0.7	0.2	-31.4

Table 10: Detailed INSTFOL results of different retrievers on different levels.

[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 7: Prompt for generate 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 8: Prompt for evaluate corpus on NFCorpus dataset.

[system prompt]
You are an expert in 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 9: Prompt for instruction generation on AILA dataset.