



# INTERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning

Yutao Zhu<sup>1</sup>, Peitian Zhang<sup>1</sup>, Chenghao Zhang<sup>1,2\*</sup>, Yifei Chen<sup>1,3\*</sup>, Binyu Xie<sup>1</sup>  
Zhicheng Dou<sup>1†</sup>, Zheng Liu<sup>4</sup>, and Ji-Rong Wen<sup>1</sup>

<sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China

<sup>2</sup>School of Computer Science, Beijing University of Posts and Telecommunications

<sup>3</sup>School of Artificial Intelligence, Nankai University, <sup>4</sup>Beijing Academy of Artificial Intelligence  
yutaozhu94@gmail.com, dou@ruc.edu.cn

## Abstract

Large language models (LLMs) have demonstrated impressive capabilities in various natural language processing tasks. Despite this, their application to information retrieval (IR) tasks is still challenging due to the infrequent occurrence of many IR-specific concepts in natural language. While prompt-based methods can provide task descriptions to LLMs, they often fall short in facilitating comprehensive understanding and execution of IR tasks, thereby limiting LLMs’ applicability. To address this gap, in this work, we explore the potential of instruction tuning to enhance LLMs’ proficiency in IR tasks. We introduce a novel instruction tuning dataset, INTERS, encompassing 21 tasks across three fundamental IR categories: query understanding, document understanding, and query-document relationship understanding. The data are derived from 43 distinct datasets with manually written templates. Our empirical results reveal that INTERS significantly boosts the performance of various publicly available LLMs, such as LLaMA, Mistral, and Phi, in search-related tasks. Furthermore, we conduct a comprehensive analysis to ascertain the effects of base model selection, instruction design, volume of instructions, and task variety on performance. We make our dataset and the models fine-tuned on it publicly accessible at <https://github.com/DaoD/INTERS>.<sup>1</sup>

## 1 Introduction

Recent advancements in large language models (LLMs) have significantly impacted the field of natural language processing (NLP). These LLMs, characterized by their extensive training data and numerous parameters, excel in various tasks through zero-shot or few-shot in-context learning,

thereby demonstrating remarkable generalizability. In the area of information retrieval (IR), the introduction of LLMs has also led to notable developments. Various studies have investigated the integration of LLMs into diverse IR tasks (Wang et al., 2023; Tang et al., 2023; Sun et al., 2023; Ma et al., 2023). Despite these efforts, LLMs have not consistently outperformed smaller models in IR tasks, a finding that diverges from our intuition. This discrepancy may stem from the complexity of IR-specific concepts like queries, relevance, and user intent, which are infrequently encountered in natural language texts and are inherently challenging to comprehend.

Concurrently, the concept of instruction tuning has emerged as a crucial method to enhance the capabilities and controllability of LLMs. Instruction fine-tuned LLMs have shown impressive generalization to new tasks without prior exposure. Despite the availability of numerous instruction tuning datasets, a gap remains in their applicability to IR tasks.<sup>2</sup> This lack of targeted datasets for IR poses additional challenges in applying LLMs effectively in this domain.

To fill the aforementioned gap, in this work, we build a new **INstruction Tuning dataseT foR Search** (INTERS). This dataset is designed to specifically enhance the search capabilities of LLMs. We focus on three key aspects that are common in various search-related tasks: query understanding, document understanding, and the comprehension of the relationship between queries and documents. Specifically, we collect 43 datasets covering 20 distinct search-related tasks. For each task, we manually craft a task description and 12 unique templates. These templates serve as a foundation for generating both zero-shot and few-shot data examples. Finally, we mix all the data examples and

<sup>\*</sup>This work was done when Chenghao Zhang and Yifei Chen were doing internship at Renmin University of China.

<sup>†</sup>Corresponding author.

<sup>1</sup>This work is still in progress. More detailed descriptions and experimental results will be added.

<sup>2</sup>We will use the term “search-related tasks” and “IR tasks” indiscriminately in this paper.

construct INTERS.

We conduct experiments by fine-tuning several open-sourced LLMs using the INTERS dataset. Experimental results show that INTERS consistently enhances the performance of LLMs of different sizes across a spectrum of search tasks. This improvement is observed not only in tasks that are directly learned in the training data (in-domain) but also in tasks that are unseen in the training set (out-of-domain). Furthermore, we delved into more nuanced aspects of model training and adaptation. Our experiments examined the impact of different instruction designs on the LLMs’ performance. We also investigated the role of data volume, considering how the quantity of training data influences the models’ learning and generalization capabilities. Additionally, we pay attention to the effectiveness of few-shot examples, assessing how in-context learning can aid in adaptation to new tasks. These comprehensive experiments provide valuable insights into the optimization of LLMs for enhanced performance in search-related tasks.

Our further experiments investigate the influence of different designs of instructions, the influence of data volume, and the few-shot examples’ effect.

The contributions of this work are threefold:

(1) We carefully analyze and categorize existing search tasks into three groups: query understanding, document understanding, and query-document relationship understanding. This classification provides a structured approach to addressing the diverse aspects of search tasks and forms the basis for targeted model training and improvement.

(2) We collect a new instruction tuning set INTERS, specifically designed for enhancing search tasks. This dataset is comprehensive, containing data from 20 search-related tasks and integrating 43 widely-used datasets. The diversity and richness of INTERS are further ensured through the use of manually written templates and task descriptions.

(3) We conduct a series of experiments to validate the effectiveness of applying instruction tuning to improve LLMs’ search ability. We also include an in-depth analysis of different settings and configurations. This thorough analysis contributes to a deeper understanding of the factors that enhance LLMs’ effectiveness in search-related tasks.

## 2 Related Work

**Large Language Models for Information Retrieval** LLMs possess a remarkable capacity for

language understanding, enabling them to be highly valuable in comprehending user queries and documents. Therefore, many researchers have explored applying LLMs to IR tasks (Zhu et al., 2023). Existing studies can be roughly categorized into two groups. The first group of methods leverages LLMs to enhance IR components. For example, LLMs can be used as query rewriters to understand users’ search intent more accurately, thereby reformulating original queries into more effective ones (Wang et al., 2023; Srinivasan et al., 2022; Tang et al., 2023; Mao et al., 2023). LLMs can also be applied to modeling the relationship between queries and documents for tasks like document ranking (Sun et al., 2023; Zhang et al., 2023b; Ma et al., 2023; Zhuang et al., 2023). The other group of methods treats LLMs as search agents to accomplish a range of search tasks (Nakano et al., 2021; Qin et al., 2023; Liu et al., 2023). A notable method is WebGPT (Nakano et al., 2021), which employs imitation learning to teach an LLM (*i.e.*, GPT-3) to use search engines and answer questions like a human.

Different from existing studies, our research focuses on using instruction tuning to improve the overall performance of LLMs on various search tasks. This involves refining the models’ abilities to interpret and respond to search-related instructions more effectively, thereby improving their utility in complex IR scenarios.

### **Instruction Tuning for Large Language Models**

Instruction tuning aims at fine-tuning pre-trained LLMs on a collection of formatted instances in the form of natural language (Wei et al., 2022; Mishra et al., 2022; Wang et al., 2022). This approach bears a close resemblance to supervised fine-tuning (Ouyang et al., 2022) and multi-task prompt training (Sanh et al., 2022). Instruction tuning’s efficacy lies in its ability to not only enhance LLMs’ performance on tasks they have been directly trained on but also to equip them with the ability to generalize to new, unseen tasks. (Sanh et al., 2022; Wei et al., 2022).

In this work, we leverage instruction tuning to specifically enhance LLMs’ performance on search-related tasks. While our study is inspired by FLAN (Wei et al., 2022), our focus diverges towards search tasks rather than general NLP tasks. Our experiments will show that instruction tuning is also an effective way to improve LLMs’ overall performance on various search tasks.

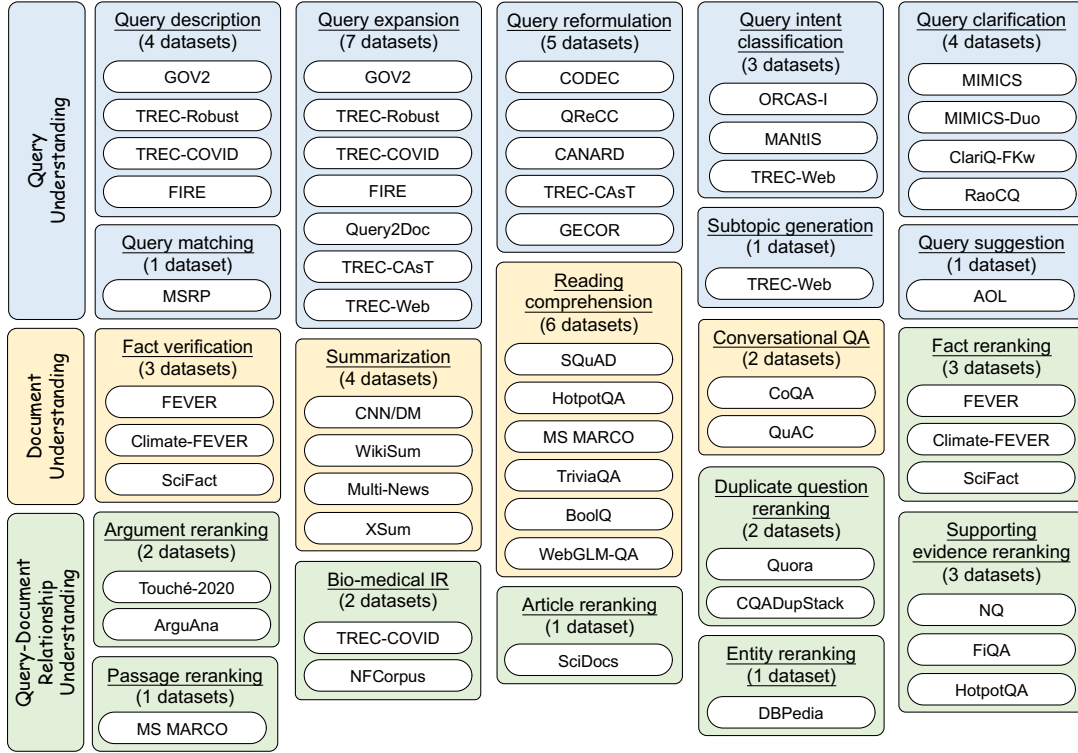


Figure 1: Categories, tasks, and datasets used in INTERS. Different colors indicate the category that the task or dataset belongs to (*i.e.*, blue for query understanding, yellow for document understanding, and green for query-document relationship understanding).

### 3 Instruction Tuning for Search

Instruction tuning has proven to be effective for LLMs in responding to instructions. This method essentially involves training LLMs through supervised learning to execute particular tasks based on provided instructions. A notable benefit of this approach is that, after fine-tuning, LLMs can comprehend and execute instructions not only for similar tasks but also for tasks they have not learned before. However, it is important to note that search tasks, which are the focus of our study, differ significantly from typical NLP tasks in terms of their objectives and structures. Search tasks primarily revolve around two key elements: *queries* and *documents*. Therefore, as shown in Figure 1, we consider collecting tasks and datasets in three categories: query understanding, document understanding, and query-document relationship understanding. We posit that tasks within these categories are instrumental in refining LLMs’ abilities to interpret queries, comprehend documents, and discern their relationships. After selecting the datasets, we manually write templates for each dataset. The final dataset INTERS is then generated by fitting the data samples into these templates.

#### 3.1 Tasks & Datasets

Developing a comprehensive instruction tuning dataset covering a wide range of tasks is very resource-intensive. To address this, we follow the previous studies (Wei et al., 2022; Chung et al., 2022) and choose to convert existing datasets from the IR research community into an instructional format. We consider tasks under the categories of query understanding, document understanding, and query-document understanding.

##### 3.1.1 Query Understanding

In IR, a query is a user-initiated request for information, typically composed of keywords, phrases, or natural language questions. It aims at retrieving relevant information from a retrieval system (*e.g.*, a search engine). The effectiveness of a query is measured by its ability to accurately reflect the user’s intent and retrieve the most relevant documents. During the retrieval process, query understanding is a critical component in determining the efficiency and user satisfaction of the IR systems. Therefore, we collect a group of tasks addressing aspects of query understanding to enhance LLMs’ capability of understanding the semantics of queries and capturing the underlying user search intent. Specif-

ically, we consider the following eight tasks.

- **Query description:** The query description task involves describing the documents potentially relevant to a user-provided query. Queries typically comprise keywords reflecting the user’s information needs. The objective of the task is to articulate the characteristics and content of documents that would be considered pertinent to these keywords, aiding in the understanding and retrieval of relevant information. We use the following four datasets: GOV2,<sup>3</sup> TREC-Robust (Voorhees, 2004, 2005), TREC-COVID (Voorhees et al., 2020), and FIRE 08, 10-12.<sup>4</sup>

- **Query expansion:** The query expansion task involves elaborating an original, brief query into a longer, more detailed version while preserving the original search intent. This process enhances the search engine’s understanding of the user’s needs, leading to more accurate and relevant document retrieval. We use the following seven datasets: GOV2, TREC-Robust, TREC-COVID, FIRE, Query2Doc (Wang et al., 2023), TREC-CAsT (Dalton et al., 2020), and TREC-Web 09-14.<sup>5</sup>

- **Query reformulation:** The query reformulation task enhances user-input queries to be more explicit and comprehensible for search engines. It addresses omissions typical of user queries, which often exclude common sense or contextually implied information. The refined query, therefore, includes all necessary details to guide the search engine towards retrieving the most relevant documents. We use the following datasets: CODEC (Mackie et al., 2022), QReCC (Anantha et al., 2021), CANARD (Elgohary et al., 2019), TREC-CAsT, and GECOR (Quan et al., 2019).

- **Query intent classification:** User queries can have various search intents, such as informational (seeking knowledge about a topic), transactional (aiming to purchase a product), or navigational (looking to find a specific website). Accurately discerning the type of intent behind a query is crucial for search engines to tailor and refine their results effectively. We use the following three datasets: ORCAS-I (Alexander et al., 2022), MAN-IIS (Penha et al., 2019), and TREC-Web 09-14.

- **Query clarification:** The query clarification task addresses unclear or ambiguous user queries by asking for further details or providing clarification

options. This process helps refine the query, resulting in clearer and more precise search terms for improved search engine results. We use the following datasets: MIMICS (Zamani et al., 2020), MIMICS-Duo (Tavakoli et al., 2022), ClariQ-FKw (Sekulic et al., 2021), and RaoCQ (Rao and III, 2018).

- **Query matching:** The query matching task involves determining whether two queries or texts, despite differing in expression, convey the same meaning. This is crucial in search tasks where identifying synonymous queries can enhance the relevance and accuracy of results. We use the dataset: MSRP.<sup>6</sup>

- **Query subtopic generation:** The query subtopic generation task addresses the ambiguity of web searches by identifying and presenting various aspects of the initial query. This approach aids search engines in understanding the query’s breadth, leading to more diverse and relevant search results. We use the dataset: TREC-Web 09-14.

- **Query suggestion:** In search sessions, users often input a series of queries to fulfill a specific information need. The query suggestion task aims to analyze these queries and associated search behaviors to understand the user’s intent and predict the next likely query, thereby enhancing the search experience. We use the AOL dataset.<sup>7</sup>

### 3.1.2 Document Understanding

In IR, a document refers to any piece of information that can be retrieved in response to a query, such as web pages in search engines. Document understanding is the process by which an IR system interprets and comprehends the content and context of these documents. The importance of document understanding lies in its direct impact on the effectiveness and accuracy of information retrieval. Enhanced document understanding leads to better search results, more effective organization of information, and an overall more efficient and user-friendly retrieval process. Therefore, we collect the following four tasks to enhance LLMs’ capability of document understanding.

- **Fact verification:** The fact verification task involves assessing whether a claim is supported or refuted by the given evidence. It requires a clear analysis of the relationship between the claim

<sup>3</sup><https://ir-datasets.com/gov2.html#gov2>

<sup>4</sup><https://www.isical.ac.in/~fire/data.html>

<sup>5</sup><https://trec.nist.gov/data/webmain.html>

<sup>6</sup><https://www.microsoft.com/en-us/download/details.aspx?id=52398>

<sup>7</sup>The AOL dataset has been officially withdrawn. However, as it is the most commonly used dataset for query suggestion, we still include it in INTERS.



and the evidence, with a careful check to determine if there is sufficient information for a conclusive judgment. Such detailed understanding aids search engines in achieving a deeper comprehension of the documents, enhancing their ability to deliver accurate and relevant results. We use the three datasets: FEVER (Thorne et al., 2018), Climate-FEVER (Diggelmann et al., 2020), and SciFact (Wadden et al., 2020).

- **Summarization:** The text summarization task seeks to create a concise summary of one or more lengthy documents, encapsulating all vital information while omitting extraneous details. The summary must accurately reflect the content of the original documents without introducing any new information. Achieving this necessitates a profound understanding of the documents, which can significantly enhance the performance of search engines by providing distilled, relevant content. We use four datasets: CNN/DM (Nallapati et al., 2016), WikiSum (Liu et al., 2018), Multi-News (Fabbri et al., 2019), and XSum (Narayan et al., 2018).

- **Reading comprehension:** The reading comprehension task requires generating an answer to a question using information from a given context. It necessitates a deep understanding of the text’s context and semantics, enabling search engines to more accurately rank the relevance of retrieved documents based on this nuanced comprehension. We use the following six datasets: SQuAD (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), MS MARCO (Nguyen et al., 2016), TriviaQA (Joshi et al., 2017), BoolQ (Clark et al., 2019), and WebGLM-QA (Liu et al., 2023).

- **Conversational question-answering:** Conversational question-answering involves responding to a series of interrelated questions based on a given context. As these questions might build upon shared information, some details may be implicitly understood rather than explicitly stated. By comprehensively understanding and analyzing this dialogue structure, search engines can enhance their interpretation of user queries and their connections to relevant documents, thereby improving result accuracy and relevance. We use these two datasets: CoQA (Choi et al., 2018) and QuAC (Choi et al., 2018).

### 3.1.3 Query-document Relationship Understanding

Query-document relationship understanding in information retrieval is the process of determining

how well the content of a document matches or satisfies the intent behind a user’s query. This involves interpreting the query’s semantics, context, and purpose, and then assessing the relevance of documents based on how closely they correspond to these aspects. It is the core task of information retrieval. In this category, we mainly consider the document reranking task.

- **Document reranking:** In document reranking, the target is to rerank a list of candidate documents according to their relevance to the user’s query. The most relevant documents, those that best cover the user’s information needs, are ranked highest. It is worth noting that candidate documents are often obtained from upstream retrieval systems. Since LLMs cannot process a large number of documents directly (due to their length limit and high resource cost), we do not consider the retrieval task. We use the MS MARCO passage reranking dataset and the datasets in the BEIR (Thakur et al., 2021) benchmark across multiple domains (such as bio-medical, finance, and social media), which includes Touché-2020 (Bondarenko et al., 2020), ArguAna (Wachsmuth et al., 2018), TREC-COVID, NFCorpus (Boteva et al., 2016), SciDocs (Cohan et al., 2020), Quora,<sup>8</sup> CQADupStack (Hoogeveen et al., 2015), DBPedia (Auer et al., 2007), FEVER, Climate-FEVER, SciFact (Wadden et al., 2020), NQ (Kwiatkowski et al., 2019), FiQA (Maia et al., 2018), and HotpotQA.

### 3.2 INTERS Construction

After determining the tasks and datasets we plan to use, we start to construct INTERS. The construction process is illustrated in Figure 2, which can be divided into four steps.

- (1) **Preprocessing.** We download all datasets from publicly available resources, filter out unnecessary attributes and invalid data samples, and then convert them into the JSONL format for further processing.

- (2) **Template collection.** Following the design of FLAN (Wei et al., 2022), we craft 12 distinct templates for *each dataset*. These templates use natural language instructions to describe the specific task associated with each dataset (two example templates are shown in the second part of Figure 2). To improve the diversity of the templates, we integrate up to two “inverse” templates per dataset.

<sup>8</sup><https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

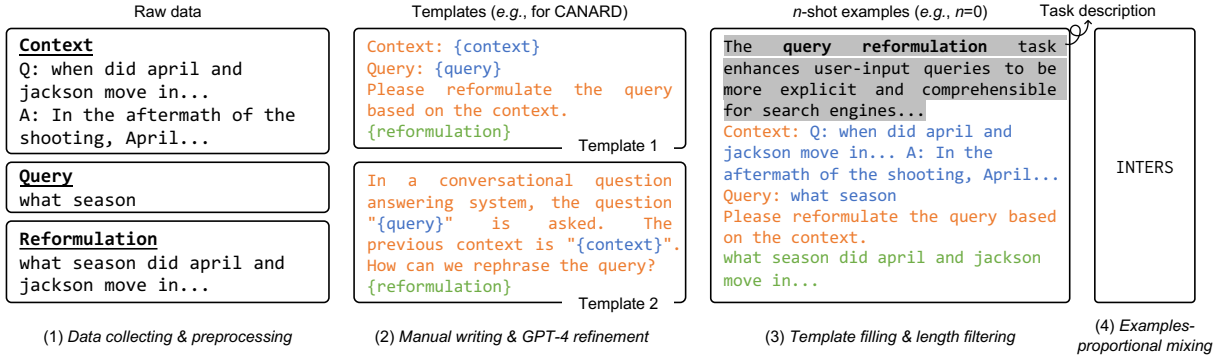


Figure 2: An example of our data construction process: (1) We first collect and preprocess the original dataset to produce raw data. (2) Then, for each dataset, we manually craft 12 distinct templates, which are further refined by GPT-4 to ensure precision and relevance. (3) Next, we compose a comprehensive description for each task, which, in conjunction with the corresponding templates, is employed to generate examples in both zero-shot and few-shot contexts. During the process, a length filter is implemented to exclude examples exceeding a specified length threshold. (4) Finally, to compile the final INTERS, we adopt an examples-proportional mixing strategy (Raffel et al., 2020), ensuring a balanced and diverse collection of instruction data.

For example, for the query expansion task, we include templates that prompt for simplifying a query. Additionally, to enhance the LLMs’ task comprehension, we provide detailed descriptions for *each task*. These task descriptions serve a dual purpose: offering a granular understanding of the task’s objectives and establishing a linkage among datasets under the same task. The efficacy of this design will be demonstrated through our experiments presented in Section 5.1.

**(3) Example generation.** For each data sample, we use the corresponding task description and a randomly selected template to generate  $n$ -shot examples (where  $n \in [0, 5]$  in our experiments). The third part of Figure 2 shows an zero-shot example generated from the CANARD dataset. For few-shot examples (where  $n \geq 1$ ), we insert the  $n$  examples between the task description and the input, where the examples are separated by special tokens (*i.e.*, “\n\n”). All few-shot examples are randomly selected from the training set. To further improve the diversity of the training data, half of the few-shot examples are constructed using the same template as the current sample, while the rest are constructed by randomly selected templates. Moreover, to ensure that the few-shot examples are within the learnable scope of LLMs, we apply a length filter to exclude examples that exceed a predefined length threshold (2,048 tokens in our experiments).

**(4) Example mixture.** To compile INTERS, we randomly select examples from our entire collection until we accumulate a total of 200,000 exam-

ples.<sup>9</sup> To balance the different sizes of datasets, we limit the number of training examples per dataset to 10,000 and adhere to an examples-proportional mixing strategy (Raffel et al., 2020) with a mixing rate maximum of 5,000. Under this scheme, any dataset contributing more than 5,000 examples does not receive extra weighting for the additional samples, thus preventing the dominant influence from larger datasets.

## 4 Experiments

We fine-tune several open-sourced LLMs on our INTERS, and evaluate their performance in different settings. All of these fine-tuned models will be released later.

### 4.1 Backbone Models

We employ four LLMs in different sizes, ranging from 1B parameters to 7B parameters.

- **Falcon-RW-1B** (Penedo et al., 2023) is a language model developed by the Technology Innovation Institute, trained on 600B tokens of English data. The model is designed for researching large language models and the impact of adequately filtered and deduplicated web data on their properties, such as fairness, safety, limitations, and capabilities.

- **Minima-2-3B** (Zhang et al., 2023a) is a novel language model designed to achieve a new compute-performance frontier on common benchmarks by distilling knowledge from a large teacher language

<sup>9</sup>This number is determined to strike a balance between efficacy and training costs.

model (LLaMA-2-7B). The model uses a data mixture of 126 billion tokens from various sources for distillation.

- **Mistral-7B** (Jiang et al., 2023) is a language model engineered for superior performance and efficiency. It leverages mechanisms such as grouped-query attention (Ainslie et al., 2023) and sliding window attention (Beltagy et al., 2020; Child et al., 2019) to outperform other language models in various benchmarks.

- **LLaMA-2-7B** (Touvron et al., 2023) is language model trained on around 2T tokens. It has shown exceptional performance across multiple benchmark tests and has been widely used for LLM research. In our experiments, we find that the LLaMA-2-Chat model performs slightly better than the LLaMA-2-Base after fine-tuning (the result is reported in Section 4.3). Therefore, we use LLaMA-2-Chat in our main experiments and further investigation.

## 4.2 Implementation Details

For all backbone models, we used their publicly available checkpoints on Huggingface. The fine-tuning process was implemented using PyTorch and Colossal-AI frameworks (Li et al., 2023). To optimize memory usage and accelerate training, we applied DeepSpeed ZeRO stage 2 (Rasley et al., 2020) and BFloat16 mixed precision techniques. Additionally, Flash attention (Dao et al., 2022) was used to further improve training efficiency. The training was conducted with a batch size of 32, a learning rate of  $1e-5$ , and a maximum length setting of 2,048 tokens. All models were trained on 8 Tesla A100-40G GPUs. It is important to note that the hyperparameters were set based on empirical observations, as the primary aim was to validate the effectiveness of INTERS. Comprehensive hyperparameter tuning was beyond the scope of this study due to resource limitations.

## 4.3 In-domain Evaluation

We first perform an in-domain evaluation to validate the effectiveness of instruction tuning on search tasks. In this experiment, we split all data into training, validation, and test sets.<sup>10</sup> The models are fine-tuned on the training set and evaluated on the test set. As all tasks and datasets are exposed to the models during training, we call it an in-domain evaluation.

<sup>10</sup>More details will be updated in Appendix.

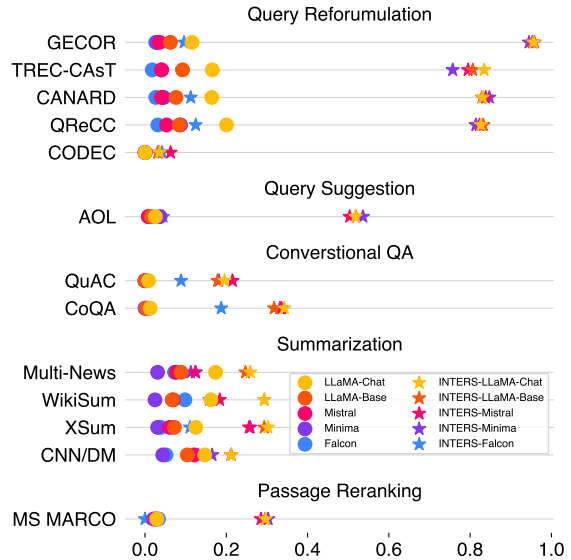
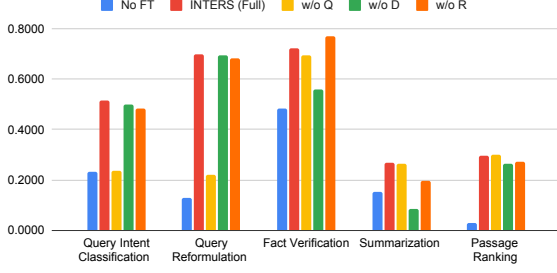


Figure 3: Average performance of all backbone models and fine-tuned models on five selected tasks under zero-shot settings. The full results will be added to Appendix.

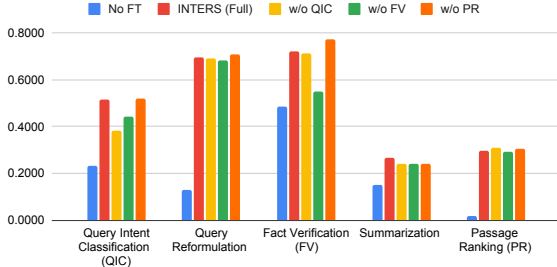
The experimental results are shown in Figure 3. Generally, after fine-tuning on INTERS, all models in various sizes can achieve significantly better performance. This demonstrates the effectiveness and broad applicability of instruction tuning in enhancing LLMs’ search performance. Besides, we have the following observations.

On most datasets, larger models tend to perform better than smaller ones. For instance, LLaMA-7B and Mistral-7B show superior performance compared to Minima-3B and Falcon-1B. Intriguingly, in specific tasks such as query reformulation and summarization, larger models without fine-tuning can even outperform the smaller models after fine-tuning (e.g., LLaMA-7B-Chat > INTERS-Falcon on GECOR). This confirms the inherent advantages of larger-scale parameters in model performance. Notably, in tasks such as query suggestion, the INTERS-fine-tuned Minima model with 3B parameters outperforms other models with 7B parameters. This suggests that fine-tuning smaller models can be a cost-effective strategy for certain specific tasks.

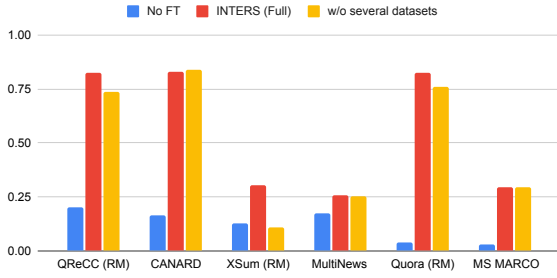
Before fine-tuning, the LLaMA-Chat model, which is already optimized for dialogue scenarios, exhibit superior performance compared to the LLaMA-Base model. This advantage is attributed to LLaMA-Chat’s better capability of understanding instructions and performing tasks. However, after instruction tuning with INTERS, the performance gap diminishes. This shows the broad gen-



(a) Performance of removing different task groups. We use “Q”, “D”, and “R” to denote query understanding, document understanding, and query-document relationship understanding, respectively.



(b) Performance of removing different tasks.



(c) Performance of removing different tasks. “RM” indicates the dataset is removed from the training set and becomes unseen during test.

Figure 4: Out-of-domain evaluation result.

erality of our instruction tuning on various types of LLMs.

#### 4.4 Out-of-domain Evaluation

Instruction fine-tuned LLMs have demonstrated a remarkable zero-shot performance on unseen tasks (Wei et al., 2022; Chung et al., 2022). We also investigate the generalizability of the models after fine-tuned on INTERS. Specifically, we consider the following three scenarios.

- **Group-level generalizability:** In this scenario, we exclude an entire group of tasks (*i.e.*, query understanding, document understanding, and query-document relationship understanding) from INTERS. Then, we fine-tune the models on the remaining data and test them on all datasets. This experiment can help understand how distinct groups

of tasks relate to each other and contribute to overall model performance.

- **Task-level generalizability:** In this scenario, we remove specific tasks (*i.e.*, query intent classification, fact verification, and passage reranking) from INTERS. Similarly, we fine-tune the models on the remaining data and evaluate them on all datasets. The goal is to assess whether fine-tuned models can generalize to unseen tasks effectively.

- **Dataset-level generalizability:** In this scenario, we exclude several datasets (including TREC-Robust, QReCC, MIMICS-Duo, Climate-FEVER, XSum, Quora, and NQ) from INTERS. Then, we fine-tune the models on the remaining data and test them on all datasets. This experiment aims to evaluate the fine-tuned models’ ability to generalize to unseen datasets within the scope of learned tasks.

The experimental results are shown in Figure 4. From the result, we can see:

- (1) In the group-level ablation study (Figure 4a), the models fine-tuned with the full INTERS outperforms those trained on the ablated datasets. This verifies the efficacy of comprehensive fine-tuning in improving search task performance. We can also observe that models trained on a subset of tasks still surpass the performance of the untrained models. For example, the performance of “w/o Q” is higher than “No FT” on the query intent classification task. This result indicates that the different task groups are effectively complementary.

- (2) The result in Figure 4b shows that the model can exhibit task-level generalization. For instance, models fine-tuned without the query intent classification task still outperform the untrained one in this task. This implies that knowledge learned from other search tasks is helpful for understanding query intent. Furthermore, the query reformulation task’s performance also drops when the query intent classification task is removed. This also validates that these tasks can influence each other. Overall, task-level generalization indicates that LLMs fine-tuned on our INTERS can be better applied to other search tasks.

- (3) The result of the third scenario is illustrated in Figure 4c. Compared to the previous two scenarios, this scenario is much easier for the fine-tuned model as all tasks have been learned during training. Nevertheless, we can see that some datasets (such as XSum) are difficult for knowledge transfer from other datasets. We also notice that removing QReCC from training leads to improved performance on CANARD, highlighting the complex



relationship between different datasets. This suggests a need for further exploration into the optimal combination of datasets for instruction tuning.

## 5 Further Analysis

We also conduct a series of experiments to investigate the impact of different settings in INTERS. All the experiments are conducted by fine-tuning the LLaMA-2-Chat-7b model and evaluate its performance in the in-domain setting.

### 5.1 Impact of Task Description

INTERS includes a detailed description for each task, intended to enhance the model’s understanding of the task and create connections among datasets under the same task. To examine its effectiveness, we conduct an experiment by removing the task descriptions from our dataset. The result is presented in Table 1.

The results demonstrate that the use of task descriptions significantly improves model performance across most datasets, both with and without fine-tuning. This strongly supports our hypothesis that detailed task descriptions can help the model understand the tasks better. Besides, the task description appear to enhance the instruction tuning process, leading to substantial improvements in some cases (*e.g.*, a 77.5% performance improvement on TriviaQA). We speculate that these task descriptions not only clarify individual tasks but also facilitate more effective knowledge transfer across different datasets.

### 5.2 Comparison with FLAN

FLAN (Wei et al., 2022; Chung et al., 2022) is a commonly used dataset for fine-tuning LLMs on natural language tasks. We compare its effectiveness on search-related tasks with that of our INTERS. Given the significantly larger size of FLAN, we randomly sample 200k data examples from it for a fair comparison.<sup>11</sup> Besides, to ensure fairness, as FLAN does not include the search-related tasks tested in this experiment, we also remove these tasks from INTERS for comparison (denoted as INTERS-T). By this means, both models trained on FLAN and INTERS are evaluated on tasks not seen during training. The results are shown in Table 2.

The findings reveal that both FLAN and INTERS can enhance LLM performance on the three tasks,

<sup>11</sup><https://huggingface.co/datasets/Open-Orca/FLAN>

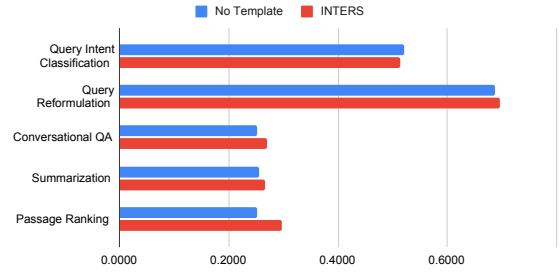


Figure 5: Ablation study result of using no template during training.

demonstrating again the effectiveness of instruction tuning in unlocking LLM potential for search tasks. Notably, INTERS yields a more substantial improvement in search tasks, particularly in reranking tasks. This is consistent with our expectation considering INTERS is specifically tailored for search tasks. Although the tested tasks are unseen in training, other search-related tasks can provide relevant knowledge for these tasks. Finally, we can see training on FLAN achieves better performance on the fact verification task. The potential reason is that this task is very close to other NLP tasks included in FLAN, allowing the model to leverage knowledge from these tasks.

### 5.3 Impact of Template

In the construction of INTERS, a key component is the development of 12 distinct templates for each dataset, aims at guiding the models in task comprehension. It is also interesting to study the influence of these templates on model performance. As an initial exploration, we compare the performance when training with or without these templates. For the no template setup, we remain the keywords to indicate the different parts of the input. For the example shown in Figure 2, we keep only “Context: ... Query: ...” as the input. Besides, we follow FLAN and use the INTERS instructions during zero-shot test (because if we use no template, the model cannot know what task to perform). Future work will include a more thorough discussion on template selection. Figure 5 shows the results—the ablation configuration yield inferior results compared to the full INTERS, indicating the significance of instructional templates in task learning..

### 5.4 Zero-shot vs. Few-shot Performance

LLMs behaves a strong ability on few-shot learning (also known as in-context learning), which enable them to adapt to a wide range of tasks. Given that

	TREC-Web (QE)			RaoCQ (QC)			CNN/DM			BoolQ	TriviaQA	MS MARCO	
	B-1	B-2	R-L	B-1	B-2	R-L	R-1	R-2	R-L	F1	F1	MRR@10	nDCG@10
LLaMA-Base	0.93	0.35	3.42	1.24	0.28	2.66	15.34	6.04	10.40	52.13	5.40	1.80	2.71
- Task Description	1.06	0.46	3.56	1.03	0.26	1.89	11.44	4.32	8.00	51.64	4.64	1.36	2.06
LLaMA-Chat	2.60	1.30	5.81	2.25	0.78	4.68	23.07	7.16	14.70	51.41	27.72	1.91	2.92
- Task Description	2.24	0.98	6.48	2.14	0.60	4.63	20.55	6.56	13.55	50.41	18.91	1.48	2.24
INTERS-LLaMA	<b>45.31</b>	<b>39.82</b>	<b>54.77</b>	<b>21.40</b>	<b>5.81</b>	<b>12.46</b>	31.15	<b>12.50</b>	21.16	<b>82.33</b>	<b>66.36</b>	23.96	29.66
- Task Description	44.67	37.18	52.25	20.14	5.33	12.28	<b>32.02</b>	12.20	<b>21.43</b>	81.53	37.38	<b>24.38</b>	<b>30.07</b>

Table 1: The influence of task descriptions on various models. “B” and “R” stand for “BLEU” and “ROUGE”. “QE” and “QC” indicate the query expansion and query clarification task respectively. All results are multiplied by 100. The best results are in **bold**. Results improved by task descriptions are highlighted in blue.

Task	No FT	FLAN	INTERS-T
Query Intent Classification	23.34	24.40	38.45
Fact Verification	48.43	57.67	54.93
Passage Reranking	2.92	18.09	30.59

Table 2: Performance comparison between INTERS and FLAN on three search-related tasks. We keep the data volume as the same (200k). To make a fair comparison, we show the performance of INTERS by removing the corresponding task from the full dataset.

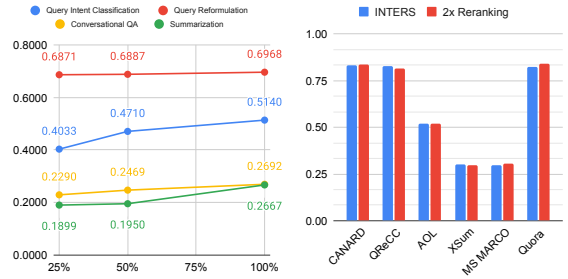


Figure 7: (Left) Performance of different data volumes. (Right) Performance of using more reranking data.

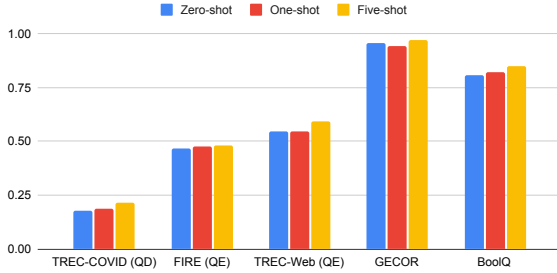


Figure 6: Performance under few-shot settings. “QD” stands for query description, while “QE” represents query expansion.

INTERS comprises a mix of zero-shot and few-shot, it is critical to examine the few-shot performance of the LLMs fine-tuned on INTERS. We choose datasets for few-shot testing that fit within the models’ input length limit (2,048 tokens in our case). The results are shown in Figure 6. Generally, few-shot examples bring a consistent improvement in performance across all datasets, compared to zero-shot INTERS. Few-shot examples are particularly beneficial in tasks with complex output spaces, such as reading comprehension (BoolQ), potentially because these examples help the model better understand the task and output format.

## 5.5 Impact of Data Volumes

The quantity of training data plays a pivotal role in the success of instruction tuning. To explore this, we conduct experiments using only 25% and 50% of the data sampled from INTERS for fine-tuning. Furthermore, we investigate the effects of task-specific data volumes by doubling the data for tasks in the query-document relationship understanding group. The results shown in Figure 7 clearly demonstrate that increasing the volume of instructional data generally enhances model performance. However, the sensitivity to data volume varies across tasks. For instance, the query reformulation task shows consistent performance across data volumes, possibly because this task requires straightforward modifications to the original query, which LLMs can easily learn. On the other hand, increasing the volume of reranking data leads to improved performance in reranking tasks but influences other tasks like summarization (XSum). This highlights the need for further research to optimize the mix and volume of instructional data for diverse tasks.

## 6 Conclusion

In this paper, we investigated the application of instruction tuning to augment the capabilities of

LLMs in performing search tasks. Our instruction tuning dataset INTERS demonstrated its effectiveness in consistently enhancing the performance of various open-sourced LLMs across both in-domain and out-of-domain settings. Our extensive experiments delved into several critical aspects, including the structure and design of instructions, the effects of few-shot learning, and the significance of data volumes in instruction tuning. It is our aspiration that this paper will serve as a catalyst for further research in the realm of LLMs, particularly in their application to IR tasks, and will encourage continued exploration into the optimization of instruction-based methods for enhancing the performance of these models.

## References

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. [GQA: training generalized multi-query transformer models from multi-head checkpoints](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4895–4901. Association for Computational Linguistics.
- Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. [ORCAS-I: queries annotated with intent using weak supervision](#). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3057–3066. ACM.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. [Open-domain question answering goes conversational via question rewriting](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 520–534. Association for Computational Linguistics.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. [Dbpedia: A nucleus for a web of open data](#). In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *CoRR*, abs/2004.05150.
- Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. [Overview of touché 2020: Argument retrieval - extended abstract](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings*, volume 12260 of *Lecture Notes in Computer Science*, pages 384–395. Springer.
- Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. [A full-text learning to rank dataset for medical information retrieval](#). In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*, pages 716–722. Springer.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. [Generating long sequences with sparse transformers](#). *CoRR*, abs/1904.10509.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentaoh Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [Quac: Question answering in context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2174–2184. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *CoRR*, abs/2210.11416.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [Boolq: Exploring the surprising difficulty of natural yes/no questions](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. [SPECTER: document-level representation learning using citation-informed transformers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics.
- Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. [TREC cast 2019: The conversational assistance track overview](#). *CoRR*, abs/2003.13624.

- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. [Flashattention: Fast and memory-efficient exact attention with io-awareness](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Thomas Diggelmann, Jordan L. Boyd-Graber, Janis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [CLIMATE-FEVER: A dataset for verification of real-world climate claims](#). *CoRR*, abs/2012.00614.
- Ahmed Elgohary, Denis Peskov, and Jordan L. Boyd-Graber. 2019. [Can you unpack that? learning to rewrite questions-in-context](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5917–5923. Association for Computational Linguistics.
- Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. [Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1074–1084. Association for Computational Linguistics.
- Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. [Cqadupstack: A benchmark data set for community question-answering research](#). In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS 2015, Parramatta, NSW, Australia, December 8-9, 2015*, pages 3:1–3:8. ACM.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: a benchmark for question answering research](#). *Trans. Assoc. Comput. Linguistics*, 7:452–466.
- Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023. [Colossal-ai: A unified deep learning system for large-scale parallel training](#). In *Proceedings of the 52nd International Conference on Parallel Processing, ICPP '23*, page 766–775, New York, NY, USA. Association for Computing Machinery.
- Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. [Generating wikipedia by summarizing long sequences](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. [Webglm: Towards an efficient web-enhanced question answering system with human preferences](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4549–4560. ACM.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. [Fine-tuning llama for multi-stage text retrieval](#). *CoRR*, abs/2310.08319.
- Iain Mackie, Paul Owoicho, Carlos Gemmell, Sophie Fischer, Sean MacAvaney, and Jeffrey Dalton. 2022. [CODEC: complex document and entity collection](#). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3067–3077. ACM.
- Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www'18 open challenge: Financial opinion mining and question answering](#). In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1941–1942. ACM.
- Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. [Large language models know your contextual search intent: A prompting framework for conversational search](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1211–1225. Association for Computational Linguistics.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [Cross-task generalization via natural language crowdsourcing instructions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.



- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. [Webgpt: Browser-assisted question-answering with human feedback](#). *CoRR*, abs/2112.09332.
- Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence rnns and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. [MS MARCO: A human generated machine reading comprehension dataset](#). In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *NeurIPS*.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. [The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only](#). *CoRR*, abs/2306.01116.
- Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. [Introducing mantis: a novel multi-domain information seeking dialogues dataset](#). *CoRR*, abs/1912.04639.
- Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. [Webcpm: Interactive web search for chinese long-form question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8968–8988. Association for Computational Linguistics.
- Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. [GECOR: an end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4546–4556. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100, 000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.
- Sudha Rao and Hal Daumé III. 2018. [Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2737–2746. Association for Computational Linguistics.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *KDD ’20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. [Multi-task prompted training enables zero-shot task generalization](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

- Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2021. [Towards facet-driven generation of clarifying questions for conversational search](#). In *IC-TIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, pages 167–175. ACM.
- Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Michael Bendersky. 2022. [QUILL: query intent with large language models using retrieval augmentation and multi-stage distillation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pages 492–501. Association for Computational Linguistics.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. [Is chatgpt good at search? investigating large language models as re-ranking agents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics.
- Yanran Tang, Ruihong Qiu, and Xue Li. 2023. [Prompt-based effective input reformulation for legal case retrieval](#). In *Databases Theory and Applications - 34th Australasian Database Conference, ADC 2023, Melbourne, VIC, Australia, November 1-3, 2023, Proceedings*, volume 14386 of *Lecture Notes in Computer Science*, pages 87–100. Springer.
- Leila Tavakoli, Johanne R. Trippas, Hamed Zamani, Falk Scholer, and Mark Sanderson. 2022. [Mimics-duo: Offline & online evaluation of search clarification](#). In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3198–3208. ACM.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models](#). *CoRR*, abs/2104.08663.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and verification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Ellen M. Voorhees. 2004. [Overview of the TREC 2004 robust retrieval track](#). NIST Special Publication. National Institute of Standards and Technology (NIST).
- Ellen M. Voorhees. 2005. [Overview of the TREC 2005 robust retrieval track](#). In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).
- Ellen M. Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2020. [TREC-COVID: constructing a pandemic information retrieval test collection](#). *SIGIR Forum*, 54(1):1:1–1:12.
- Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. [Retrieval of the best counterargument without prior topic knowledge](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 241–251. Association for Computational Linguistics.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. [Fact or fiction: Verifying scientific claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.
- Liang Wang, Nan Yang, and Furu Wei. 2023. [Query2doc: Query expansion with large language models](#). pages 9414–9423.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit

- Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujay Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. [Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. [Finetuned language models are zero-shot learners](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.
- Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. [MIM-ICS: A large-scale data collection for search clarification](#). In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3189–3196. ACM.
- Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. 2023a. [Towards the law of capacity gap in distilling language models](#). *CoRR*, abs/2311.07052.
- Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023b. [Rankinggpt: Empowering large language models in text ranking with progressive enhancement](#). *CoRR*, abs/2311.16720.
- Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. [Large language models for information retrieval: A survey](#). *CoRR*, abs/2308.07107.
- Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. [Open-source large language models are strong zero-shot query likelihood models for document ranking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8807–8817. Association for Computational Linguistics.

Task	Dataset	# Examples	Avg #In	Avg #Out
Query Description	GOV2	900	308.07	57.90
Query Description	TREC-Robust	1,794	280.38	48.18
Query Description	TREC-COVID	300	258.74	33.13
Query Description	FIRE	1,200	290.38	46.58
Query Expansion	GOV2	900	168.71	15.77
Query Expansion	TREC-Robust	1,800	189.72	20.54
Query Expansion	TREC-COVID	300	193.50	17.59
Query Expansion	FIRE	1,200	197.39	18.83
Query Expansion	Query2Doc	62,400	378.88	81.21
Query Expansion	Trec-CAsT	300	182.64	17.39
Query Expansion	TREC-Web	1,506	163.57	12.50
Query Reformulation	CODEC	236	853.89	74.29
Query Reformulation	QReCC	62,395	644.02	15.66
Query Reformulation	CANARD	30,437	666.32	16.43
Query Reformulation	TREC-CAsT	606	444.37	14.40
Query Reformulation	GECOR	4,056	559.53	12.27
Query Clarification	MIMICS	16,734	153.83	21.06
Query Clarification	MIMICS-Duo	5,484	172.27	22.43
Query Clarification	ClariQ-FKw	13,086	142.50	12.47
Query Clarification	RaoCQ	2,759	854.22	15.33
Query Subtopic Generation	TREC-Web	1,506	321.30	74.82
Query Suggestion	AOL	62,400	202.07	5.18
Query Matching	MSRP	25,656	325.13	2.00
Query Intent Classification	MANtIS	6,062	1,109.86	3.81
Query Intent Classification	ORCAS-I	6,000	242.26	3.36
Query Intent Classification	TREC-Web	1,200	224.34	3.66
Fact Verification	FEVER	61,932	547.03	2.29
Fact Verification	Climate-FEVER	8,544	1,133.29	2.88
Fact Verification	SciFact	4,638	618.58	2.34
Conversational QA	CoQA	19,741	1,208.52	80.81
Conversational QA	QuAC	19,874	1,267.01	124.72
Summarization	CNN/DM	21,883	823.92	301.19
Summarization	XSum	31,510	1,057.63	135.22
Summarization	WikiSum	6,874	2,101.13	422.31
Summarization	Multi-News	5,339	3,106.26	285.37
Reading Comprehension	SQuAD	62,336	858.54	5.74
Reading Comprehension	HotpotQA	62,400	595.62	5.46
Reading Comprehension	MS MARCO	40,029	1,314.41	24.82
Reading Comprehension	BoolQ	62,384	652.50	2.00
Reading Comprehension	WebGLM-QA	29,164	1,107.86	140.69
Reading Comprehension	Trivia-QA	34,140	1,312.96	9.32
General Retrieval	MS MARCO	65,909	816.71	4.25
Argument Retrieval	Touché-2020	21,951	992.36	4.46
Argument Retrieval	ArguAna	42,736	1,077.62	4.06
Biomedical Retrieval	TREC-COVID	31,476	1,127.98	4.38
Biomedical Retrieval	NFCorpus	4,508	1,185.16	3.79
Article Retrieval	SciDocs	41,043	1,090.32	3.82
Duplicate Question Retrieval	Quora	43,930	589.70	7.20
Duplicate Question Retrieval	CQADupStack	88,934	1,117.72	4.43
Entity Retrieval	DBPedia	470	909.46	3.59
Fact Retrieval	FEVER	35,201	1,131.90	5.20
Fact Retrieval	Climate-FEVER	57,672	945.14	4.11
Fact Retrieval	SciFact	1,963	1,179.06	7.70
Supporting Evidence Retrieval	NQ	43,963	944.69	5.33
Supporting Evidence Retrieval	FiQA	20,988	1,063.95	5.73
Supporting Evidence Retrieval	Hotpot-QA	63,441	934.56	7.41

Table 3: The statistics of all datasets. “Avg” and “Max” stand for “Average” and “Maximum”, respectively. “#In” and “#Out” represent the number of tokens in the input and output with the LLaMA’s tokenizer.