

Zero-Indexing Internet Search Augmented Generation for Large Language Models

Guangxin He^{*†§}, Zonghong Dai^{*‡}, Jiangcheng Zhu[§], Binqiang Zhao[§], Qicheng Hu[§]
 Chenyue Li[†], You Peng[†], Chen Wang^{||}, Binhang Yuan[†]

† HKUST ‡ Fudan University § 01.AI || Tsinghua University

Abstract

Retrieval augmented generation has emerged as an effective method to enhance large language model performance. This approach typically relies on an internal retrieval module that uses various indexing mechanisms to manage a static pre-processed corpus. However, such a paradigm often falls short when it is necessary to integrate the most up-to-date information that has not been updated into the corpus during generative inference time. In this paper, we explore an alternative approach that leverages standard search engine APIs to dynamically integrate the latest online information (without maintaining any index for any fixed corpus), thereby improving the quality of generated content. We design a collaborative LLM-based paradigm, where we include: (i) a PARSER-LLM that determines if the Internet augmented generation is demanded and extracts the search keywords if so with a single inference; (ii) a *mixed ranking strategy* that re-ranks the retrieved HTML files to eliminate bias introduced from the search engine API; and (iii) an EXTRACTOR-LLM that can accurately and efficiently extract relevant information from the fresh content in each HTML file. We conduct extensive empirical studies to evaluate the performance of this Internet search augmented generation paradigm. The experimental results demonstrate that our method generates content with significantly improved quality. Our system has been successfully deployed in a production environment to serve 01.AI’s generative inference requests.

1 Introduction

Large language models (LLM) have demonstrated remarkable capabilities, fueling a new wave of innovative AI applications [1]. To incorporate information not included during their training phase, retrieval augmented generation (RAG) has been developed as an efficient method to enhance LLM performance by integrating externally retrieved information [2, 3]. Usually, RAG systems use an internal retrieval module that employs various indexing mechanisms to manage a static corpus. In this paper, we study an alternative approach, exploring the design and implementation of a *RAG paradigm that leverages standard search engine APIs* (such as Google and Bing), which allows for the flexible and dynamic integration of the most update-to-date online information, thereby improving the quality of content generated by LLMs.

While RAG systems with a static retrieval component, which leverages a fixed corpus of data, are effective for tasks within well-defined knowledge domains, Internet search augmented generation [4, 5, 6, 7, 8, 9] offers distinct advantages. By accessing the *most update-to-date information* available online, internet search augmented generation is particularly beneficial for tasks that require up-to-date data, such as news updates, market trends, or recent scientific discoveries, enabling the model to produce more relevant and timely responses [9]. Moreover, Internet search augmented systems can retrieve information from a vast array of sources across the web, encompassing a *broader range* of topics and perspectives, thereby enhancing the comprehensiveness and diversity of the generated content [6]. In fact, OpenAI has also released its toolkit CHATGPT-SEARCH to integrate information from the Internet to improve the user experience very recently [10].

Building a production-level, general-purpose Internet search augmented generation system is a complex and challenging task. Unlike standard RAG systems that rely on static retrieval methods—such as sparse [11, 12, 13], dense [14, 15, 16], or more advanced learnable indexing [17, 18, 19, 20, 21, 22, 23, 24] mechanisms—to retrieve carefully pre-processed, chunked texts stored as key-value pair from a vector database [25, 26, 27, 28, 29, 30, 31, 32] for context generation, Internet search augmented generation must *dynamically* process content returned by search engine APIs — otherwise any platform that serves the generative inference requests has to re-build the search engine infrastructure of Google or Bing from scratch internally. The Internet search augmented generation paradigm requires an approach that is not only economically viable but also highly efficient, capable of quickly extracting relevant information from large-scale unstructured data from the web to feed the LLM prompt input. The real-time nature of this processing, combined with the need to accurately extract relevant information across diverse and constantly changing content, significantly increases the complexity of deploying such a system at the production scale.

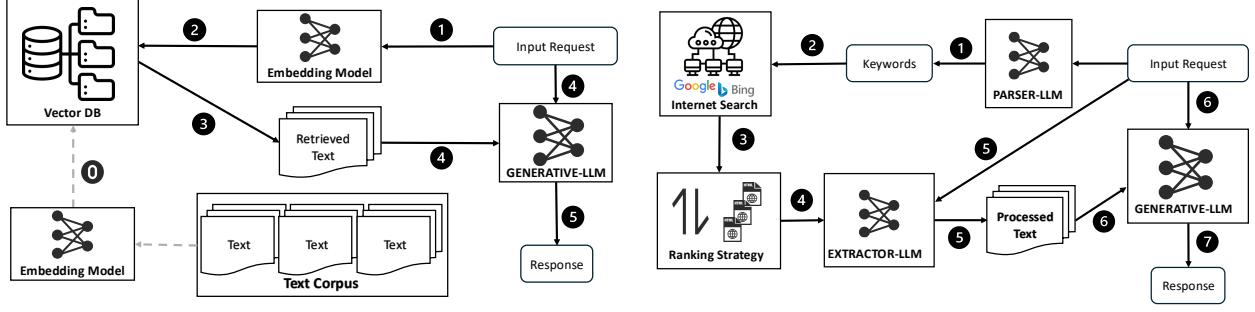
Naive approaches often fall short when attempting to build a production-level Internet search augmented generation system. The raw data retrieved from search engine APIs is usually highly redundant and noisy, which can degrade the quality of the content generated by LLMs. For instance, a simplistic approach might involve applying a set of rule-based text extraction heuristics; however, this method is labor-intensive and struggles to generalize across the diverse and ever-changing landscape of online information. Moreover, despite the advancements in LLMs that now support significantly longer context lengths — such as GPT-4o [33] and LLAMA 3.1 [34], which handle up to 128K tokens, and YI [35] and CLAUDI3 [36], which can process 200K tokens — feeding raw, unprocessed data directly from search engines into these models is inefficient — such approach not only increases the computational load, but also risks diminishing the quality of the generated content [37]. To truly harness the potential of Internet search augmented generation, more sophisticated techniques are needed to filter and process online information before it is used in LLM content generation.

Towards this end, we design and implement an Internet search augmented generation system deployed in 01AI’s production-level inference API system. Concretely, we summarize our key contributions:

Contribution 1. We propose a novel zero-indexing paradigm for Internet search augmented generation, where we introduce a set of core components to collaboratively accomplish the task. Concretely, we implement (i) a PARSER-LLM that determines the necessity of Internet-augmented generation and, if required, extracts search keywords in a single inference pass; (ii) a *mixed ranking strategy* that re-ranks the retrieved HTML files to mitigate biases introduced by the search engine API; and (iii) an EXTRACTOR-LLM that is designed to efficiently and accurately extract relevant information from the fresh content within each HTML file. Unlike traditional methods that rely on indexing or storage modules, our approach retrieves all augmented information via the Google and Bing Search APIs and processes it on the fly, eliminating the need to manage the pre-indexed data. This paradigm allows for real-time, contextually relevant content generation without the overhead of maintaining a static corpus, making it adaptable for a wide range of applications.

Contribution 2. We enumerate the design principle and implementation details for the core components: (i) we introduce how we construct the instruction set to train the PARSER-LLM to equip it with multi-instruction following ability and be able to extract the keywords for the search engine API precisely; (ii) we implement the *mixed ranking strategy* that considers both the snippet and the full content in HTML file to provide more effective re-ranking based on pure semantic relevance; (iii) we apply both supervised fine-tuning and direct preference optimization to train the EXTRACTOR-LLM so that it can accurately extract relevant information from each retrieved HTML file to improve the end-to-end generation quality.

Contribution 3. We conduct a set of concrete empirical evaluations to estimate the design and implementation of our search augmented generation system. The experimental results suggest that: (i) our Internet search augmented generation paradigm generates higher-quality outputs when compared with other RAG paradigms, which is also more cost-efficient in terms of the processed input tokens for the GENERATIVE-LLM; (ii) the EXTRACTOR-LLM can extract the relevant information accurately and precisely for various benchmarks, where it can also robustly reject inappropriate content rather than generate hallucinations when there is no relevant content.



(a) Standard retrieval augmented generation.

(b) Internet search augmented generation.

Figure 1: A comparison between the standard RAG paradigm and our Internet search augmented generation paradigm.

2 Internet Search Augmented Generation

We first introduce our proposed framework to support the paradigm of Internet search augmented generation. Figure 1 compares the standard RAG and the Internet search augmented generation proposed in this paper.

2.1 Paradigm Overview

Before we introduce our Internet search augmented generation, we first briefly review the paradigm of standard RAG as we illustrated in Figure 1(a). Before processing any inference request, standard RAG needs to prepare key-value pairs that need to be stored in a vector database, where a value is a piece of chunked text from the corpus, and the key is the corresponding embedding generated by the embedding model (**Step 0** in Figure 1(a)). When an inference request arrives, the following steps will be conducted:

- **Step 1.** The RAG system utilizes an embedding model (typically the same model used during the indexing phase) to generate an embedding vector for the input request.
- **Step 2.** This embedding vector is then used to query the vector database created during the indexing phase to search for semantically similar text chunks.
- **Step 3.** The RAG system retrieves and selects the most relevant text from the vector database based on some similarity measurements (e.g., cosine similarity).
- **Step 4.** The RAG system combines the original input request with the retrieved text to form a new prompt that integrates additional information.
- **Step 5.** The generative LLM uses this new prompt to generate a response, leveraging the retrieved information to produce more accurate and contextually relevant output.

By contrast, our Internet search augmented generation works in a fully dynamic paradigm by interacting with the external search engine APIs, which requires *zero-indexing* mechanism. Figure 1(b) illustrates our proposed paradigm when processing an inference request:

- **Step 1.** Our system first leverages a multi-functional PARSER-LLM to determine whether the input request requires external information from the Internet to generate the response — If external information is needed, the PARSER-LLM will interpret and extract the keywords based on the input request, which will be provided to the search engine API. Note that these functionalities are completed simultaneously within a single inference pass by the PARSER-LLM.
- **Step 2.** The parsed keywords are then used to query external search APIs, such as Google or Bing, to retrieve relevant information from the Internet.
- **Step 3.** We employ a mixed ranking strategy that incorporates two levels of granularity to rerank retrieved HTML files so that the original bias introduced by the search engines can be reduced.

- **Step 4.** We apply some lightweight pre-processing to insert content tags for the retrieved text so that it can be further processed by the EXTRACTOR-LLM.
- **Step 5.** As another core component of our system, the EXTRACTOR-LLM first analyzes the original input request along with the tagged retrieved content, then extracts the most relevant text portion identified by tags for generation.
- **Step 6.** The system combines the original input request with the extracted text to form an enriched prompt that integrates the additional information from the Internet.
- **Step 7.** The backbone GENERATIVE-LLM uses this new prompt to generate a response, leveraging the information from the Internet to produce high-quality output with the potential most up-to-date information.

When compared with the standard RAG system, the Internet search augmented generation needs to answer the following two essential questions:

Question 1. *How can we accurately interpret the user’s intent to interact with the search engine API, where we dynamically search the Internet and re-organize the retrieved context?*

Question 2. *How can we accurately and efficiently extract contextually relevant information to improve the quality of the generated output by implementing the EXTRACTOR-LLM?*

To answer these questions, we first identify the key challenges in Section 2.2, then we propose our system design and implementation to overcome these challenges in Section 2.3.

2.2 Key Challenges

Given the Internet search augmented generation paradigm, we identify two critical challenges, each addressed by dedicated modules outlined in Section 2.3:

Challenge 1. *Effectively retrieving intent-aligned web pages from the Internet without biased ranking.* Effectively retrieving web pages that align with user intent involves solving two critical sub-problems: (i) how to interpret user requests comprehensively before the search; and (ii) how to mitigate biases in search result rankings after the search. To be more specific, we summarize the key issues below:

- **Comprehensive interpretation of user requests:** a critical module in our system (i.e., the PARSER-LLM) must accurately understand and interpret user queries in a way that captures the full scope of the user’s intent. Ambiguities in user requests, variations in language, and the context-dependent nature of search queries create substantial obstacles. Although some recent RAG systems have implemented a similar component in question answering [6, 38] and query rewriting mechanisms [39, 40, 41], accurately inferring the user’s true intent without oversimplification or misinterpretation remains a persistent challenge in our paradigm.
- **Mitigating bias in search result rankings:** ranking bias in search engines remains an obstacle in our paradigm since reliance on engagement metrics such as click-through rate (CTR) [42] and dwell time [43] often skews result rankings, which prioritize user interaction over the actual relevance of the content — such a ranking bias needs to be removed to improve the generation quality.

Challenge 2. *Accurately and efficiently extracting contextually relevant contents from each retrieved HTML file.* When implementing the EXTRACTOR-LLM, we need to fulfill the following two concrete demands:

- **Accuracy in content extraction:** to ensure that the information retrieved from the internet is genuinely beneficial for output generation, it is crucial that the extracted content is *contextually relevant* to the original input request — in our paradigm, this indicates that the EXTRACTOR-LLM must accurately identify and select the most pertinent parts of the retrieved data, avoiding the inclusion of irrelevant or misleading content, which is particularly challenging given the variability and complexity of web data.
- **Computational efficiency in context extraction:** since large volumes of HTML data must be processed rapidly to maintain real-time response generation when serving the inference API, we need to guarantee com-

putational efficiency of the EXTRACTOR-LLM in our paradigm. Therefore, it is essential to have a lightweight LLM for this functionality, which ensures the system can deliver fast, accurate, and scalable performance by leveraging economic computational resources.

2.3 Core Functionality Design Overview

To implement the paradigm introduced in Section 2.1 and address the two challenges outlined in Section 2.2, we provide an overview of the design of the two associated functionalities:

Intent-aligned Internet Search. To obtain well-processed Internet search results, i.e., reranked HTML files, for further processing, we need to integrate the PARSER-LLM’s functionality from **Step 1** in Figure 1(b) and incorporate the reranking mechanism described in **Step 3** in Figure 1(b). To address the challenges of effectively retrieving intent-aligned web pages, we propose a hybrid approach that combines keyword extraction with reranking strategies. Concretely, the process begins with the PARSER-LLM, which is carefully trained to achieve two primary objectives when preparing user requests for search: (i) to comprehensively understand the user’s intentions and (ii) to extract and output these intentions in the form of a list of keywords, allowing for optimal utilization of search engines. After retrieving search results using the Parser LLM, we implement a *mixed ranking strategy* that combines fine-grained with coarse-grained content ranking. This approach mitigates biases influenced by commercially driven engagement factors. The strategy combines with our prior keyword extraction methodology, ensuring the effectiveness of the internet search retrieval. We enumerate the design and implementation details of this functionality in Section 3.

Content Extraction from Search Results. Given the well-organized Internet search results, we need to efficiently and accurately extract relevant information from the retrieved HTML files; our approach involves a two-step processing pipeline: (i) pre-processing the HTML content, and (ii) identifying relevant information using the EXTRACTOR-LLM with efficient computation cost during extraction. The preprocessing of the HTML files leverages a simple finite state machine to segment the content in the HTML file, enclosing each segment within corresponding tags. To guarantee the extraction accuracy, we implement a two-phase strategy to train the EXTRACTOR-LLM: (i) the first phase uses supervised fine-tuning to train the model’s basic ability to extract relevant information with careful construction of the instruction set, and (ii) the second phase further aligns the extraction with the preference of a judge after training with reinforcement learning technique. Additionally, to enhance computational efficiency, we optimize the EXTRACTOR-LLM in two ways: (i) we choose a relatively light-weight 9B LLM, where the inference latency is low; (ii) we only require the EXTRACTOR-LLM to generate relevant tags instead of directly output the original segmented contents, which significantly decreases the number of output tokens for the EXTRACTOR-LLM, and thus accelerate the extracting processing time. This integrated process ensures that the information extraction is both accurate and efficient, enabling faster and more effective retrieval of relevant information from segmented HTML content. The detailed discussion about how we implement this functionality is illustrated in Section 4.

3 Retrieval from Search Engines

As we introduced in Section 2, effectively organizing the Internet search results is indispensable in our Internet search augmented generation paradigm. In pursuit of the goal of user intent alignment and elimination of the bias introduced by the search engine APIs, we apply the following design: we introduce our training procedure of the PARSER-LLM in Section 3.1 to refine the user’s query before Internet search and the *mixed ranking strategy* in Section 3.2 to optimize the ranking of the results returned from search engine APIs.

3.1 Training the PARSER-LLM

Functionally, the PARSER-LLM takes the original user request as input, where two primary tasks need to be accomplished by a single inference computation: (i) the PARSER-LLM assesses whether Internet search

is necessary to improve the generation quality based on the user query prompt; (ii) If deemed necessary, PARSE-LLM need to extract a list of effective keywords from the user query prompt, where during the keyword extraction workflow, the PARSE-LLM needs to solve the potential issues of time-sensitivity, single-language limitation, and out-of-vocabulary error. Concretely, we enumerate the following problem that needs to be resolved by the training strategy design:

- **Multi-instruction following:** in order to improve the system’s efficiency, we demand the PARSE-LLM to generate the indication of whether Internet search is demand and the corresponding extracted keywords in a single inference computation, the PARSE-LLM has to be able to interpret the inherited relationship between the two tasks and generate integrated results.
- **Time information sensitivity:** the effectiveness of the keyword extraction can vary significantly depending on how time-specific information is integrated into the extraction prompt, where the impact varies based on different type of request. For instance, in cases with positive temporal dependency (e.g., "How's the weather?"), adding explicit time details to the keywords improves search accuracy by narrowing the context. However, for other queries (e.g., "What is the best weight loss method of 2024?"), including temporal information such as "2024" might inadvertently reduce search performance by limiting the scope of relevant results during the search process.
- **Monolingual search limitation:** relying solely on monolingual keyword extraction, such as the language of the user prompts, may not fully capture the essential keywords, especially when dealing with high-context and deductive request. For example, some languages are concise and rich in meaning but prone to ambiguity, making it challenging for the PARSE-LLM to extract the keywords [44]. This limitation highlights the difficulty of interpreting complex user queries and retrieving relevant results effectively when confined to a single language.
- **Out-of-vocabulary error:** when extracting keywords, we also need to carefully handle the out-of-vocabulary error, where some phases were not included during pretraining or aligning phases for the PARSE-LLM — such issue could significantly compromise the results of keyword extraction. For instance, decomposing unfamiliar proper nouns, such as company names that a large language model (LLM) has not encountered—into multiple keywords can lead to the retrieval of irrelevant information. This issue arises because the model lacks prior knowledge of these terms, making it challenging to interpret and process them accurately.

Instruction Set for PARSE-LLM. We construct a multi-task instruction set to train our PARSE-LLM to enable it to perform both retrieval determination and keyword extraction tasks in a single inference pass. These tasks are guided by tailored instructions designed to address the specific challenges outlined above. We utilize GPT-4O, one of the most advanced LLM API available, to generate high-quality ground truths for these tasks, ensuring the instruction set is comprehensive and robust. To fulfill the requirements listed above, we integrate the following implementation:

- **Support multi-instruction following:** to improve efficiency, the PARSE-LLM is trained to generate both the retrieval determination and the extracted keywords simultaneously within a single pass at the inference time. This requires the model to effectively interpret the interdependence of these tasks. Since 6B scale LLMs often exhibit weaker multi-task instruction-following capabilities, we carefully designed and manually tuned an instruction template that enables our smaller Parser-LLM to achieve acceptable zero-shot performance.
- **Time information integration:** recognizing that keyword extraction effectiveness can vary with time sensitivity, we specifically design instructions that guide the PARSE-LLM to handle both positive and negative temporal dependencies in user queries. The data in the instruction set for these cases is generated by GPT-4O, ensuring accurate modeling of scenarios where temporal details either enhance or hinder retrieval effectiveness. This approach enables the model to dynamically adjust its handling of time-specific information based on the context of the query.
- **Enhancing with multi-lingual search:** to address the limitations of monolingual keyword extraction, the

instruction set explicitly incorporates multi-language instructions. This enables the PARSER-LLM to extract keywords from high-context and deductive queries expressed in different languages. Training data includes examples emphasizing the ambiguity and richness of languages like Chinese, ensuring the model effectively captures essential keywords across linguistic boundaries. This multilingual approach reduces the risk of misinterpretation and improves retrieval accuracy.

- **Resolving out-of-vocabulary errors:** out-of-vocabulary errors are mitigated through the construction of synthetic data containing made-up words designed to simulate unfamiliar terms, such as novel proper nouns or company names. By learning from these examples, the PARSER-LLM becomes more adept at handling previously unseen vocabulary without decomposing such terms into irrelevant keywords. This approach ensures the robustness of the keyword extraction process, even in scenarios involving rare or new terms.

Training Details. We choose the YI-6B-CHAT model as the initial model to obtain the PARSER-LLM throughput instruction fine-tuning. Note that compared with the YI-6B-BASE, YI-6B-CHAT has been supervised fine-tuned to obtain basic instruction-following capabilities. We use the prepared instruction set notated as \mathcal{I} to train the base model in a standard auto-regressive manner, with the optimization objective defined as follows:

$$\mathcal{L} = \min_{\theta} \mathbb{E}_{j \sim \mathcal{I}} [-\log p(\Omega_j | \text{Request}_j; \theta)]$$

where \mathcal{L} represents the negative likelihood loss function that needs to be minimized, Ω_j is the expected output of the multi-task instruction for the j -th Request $_j$, θ denotes the model parameters. The expectation $\mathbb{E}_{j \sim \mathcal{I}}$ is calculated by sampling through the instruction set \mathcal{I} .

We train the model for three epochs, with a maximum sequence length of 16K tokens. The learning rate is set to 5×10^{-6} initially, and a cosine learning rate scheduler was employed, and the minimal learning rate is set to 1.5×10^{-6} . To prevent overfitting, the training data is shuffled at the beginning of each epoch. We use the FusedAdam optimizer with the hyper-parameter configured as $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1 \times 10^{-8}$. The global batch size is set to 16. The training computation was conducted on an NVIDIA server equipped with four H800 GPUs by the DeepSpeed [45] parallel training framework.

3.2 Search Result Re-ranking

To eliminate the bias introduced by search engine API, considering the irrelevant factors such as click-through rate and dwell time, we leverage open-source the BGE-M3 model [46] to assess the relevance of search results based on the extracted keywords and generate the ranking score.

We introduce a new *mixed ranking strategy*, where the BGE-M3 model can compute relevance scores for each retrieved item in the reranking pool, which consists of two levels of granularity derived from the HTML file: the snippet and the full content. We consider both the snippet and the full content because they can provide different information for the reranking model: the snippet, as a concise summary provided by the search engine, offers high-level relevance but may lack detailed accuracy; while the full content, extracted from the HTML body, contains comprehensive details but can obscure high-level relevance due to information overload. In our *mixed ranking strategy*, each snippet and the corresponding full content are treated as an independent item, which doubles the number of candidates to $2n$ for n HTML files. This approach ensures that both high-level and detailed relevance are considered in the ranking process. The reranked results are then aggregated, selecting the top-K distinct HTML files based on the highest ranking achieved by either their snippet or full content. Note that the results retrieved from auxiliary keyword lists (e.g., from multilingual queries) are combined with those from the main keyword lists, broadening the candidate pool and enhancing coverage. This strategy ensures that the final selection is optimized for both granularity types, balancing high-level relevance and detailed accuracy.

4 EXTRACT RELEVANT INFORMATION

As we introduced in Section 2, the second core technique functionality in our Internet search augmented generation paradigm is the EXTRACTOR-LLM, which should be able to accurately extract relevant information

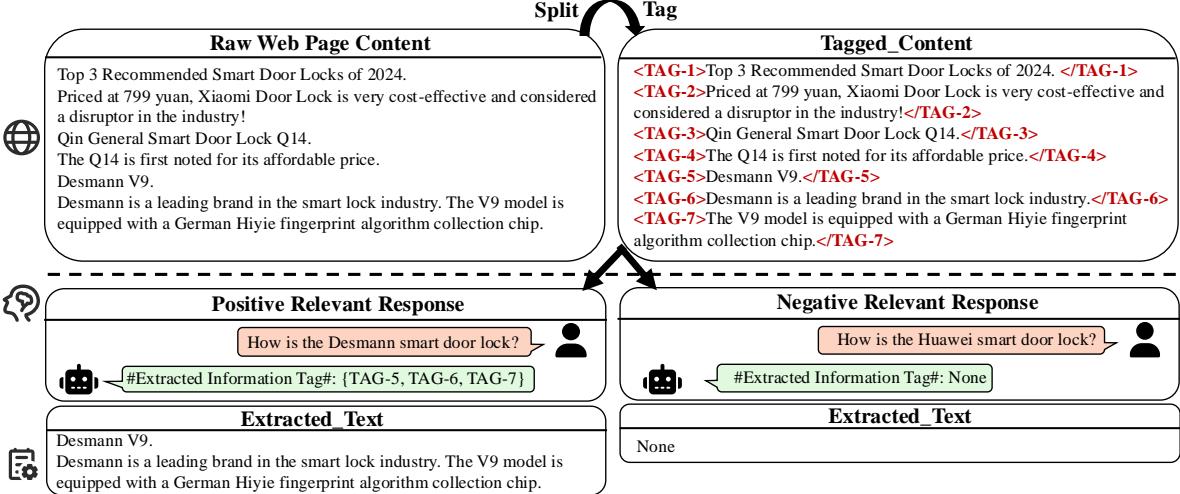


Figure 2: An illustration about pre-processing the retrieved HTML file and Extractor LLM processing procedure. without introducing significant computational overhead at the inference time. Toward this end, in Section 4.1, we first enumerate a preprocessing mechanism for the retrieved HTML file to generate tags for the content so that the EXTRACTOR-LLM only outputs the corresponding tags of the content to save the budget of generated tokens. Then we introduce a novel two-phase training strategy to train the EXTRACTOR-LLM, where the primary goal is to fulfill the functionality of determining if the relevant content exists or not and extracting tags of the relevant and user-preferred texts from the pre-processed contents. We summarize this two-phase training paradigm below:

- **SFT phase:** the first phase of training uses supervised fine-tuning (SFT) to train the EXTRACTOR-LLM’s basic ability to extract relevant information, where the construction of instruction set is explained in Section 4.2 and the SFT training details are in Section 4.3.
- **DPO phase:** the second phase further aligns the extraction with the preference of a judge after training with reinforcement learning technique, i.e., direct preference optimization (DPO)[47], where the data preparation is introduced in Section 4.4 and DPO training details are in Section 4.5.

4.1 Preprocessing the HTML File

Our pre-processing mechanism applies a heuristic to generate splitting tags for the HTML body content. This heuristic, implemented as a finite state machine (FSM), uses special punctuation marks such as ‘.’, ‘!’, and ‘?’ as delimiters to segment the content, enclosing each segment within corresponding tags. For example, the i -th segment is enclosed within the tags $\langle \text{TAG}-i \rangle$ and $\langle / \text{TAG}-i \rangle$. Note that the finite state machine is also in charge of handling some relatively complex situations; for example, when the period ‘.’ is processed, the finite state machine manages to distinguish if the ‘.’ indicates the end of a sentence or a decimal point based on the status whether ‘.’ follows numerical values. Figure 2-(top) shows a concrete example of this pre-processing procedure: The retrieved context from the HTML file introduces three types of smart door locks: Xiaomi, Qin General, and Desmann; our finite state machine is able to split the contents with different tags; for example, the introduction for Xiaomi, Qin General, and Desmann is enclosed with $\{\text{TAG-2}\}$, $\{\text{TAG-3}, \text{TAG-4}\}$, $\{\text{TAG-5}, \text{TAG-6}, \text{TAG-7}\}$, respectively.

During the inference time, once the content is segmented, the EXTRACTOR-LLM performs two main tasks: (i) determining if the content includes information relevant to the input requests; and (ii) extracting the relevant tags and their content. Formally, the EXTRACTOR-LLM takes the Request and each of the tagged content from Internet search API notated as Tagged_Content as inputs. If there is relevant information in the content, the EXTRACTOR-LLM outputs the corresponding tags; otherwise, the EXTRACTOR-LLM outputs None to indicate there is no relevant information in this content. Figure 2-(bottom) shows how the Extractor LLM determines the output for the same Tagged_Content with different generative inference Requests: On the left-hand side, the

inference Request demands the information about Desmann smart door lock, so the Extractor LLM outputs tags {TAG-5,TAG-6,TAG-7}; On the right-hand side, the inference Request demands the information about Huawei smart door lock, so the Extractor LLM returns None since there is no relevant information about Huawei smart door lock from this Tagged_Content.

4.2 Instruction Set Construction for SFT

In the first SFT phase, we aim to obtain a fine-tuned model that can explicitly determine the existence of the relevant information and the corresponding tags (if they exist). For this purpose, the central goal of our instruction set construction is to collect training data that can mirror the process within the Internet search augmented generation paradigm. We find that the essential requests for the constructed instruction set lie in three aspects:

- **Diversity of the requests and contents:** since we expect the Internet search augmented generation paradigm to be generally applicable to any generative inference requests in an API serving system, the EXTRACTOR-LLM should be able to handle various information-extracting scenarios. Thus, the instruction set should include a collection of <Request, Tagged_Content> data pairs from diversified real-world scenarios.
- **Necessary reasoning procedure:** during our preliminary exploration, we found that the EXTRACTOR-LLM would generalize poorly if we directly fed the desired TAGs during the instruction fine-tuning phase without providing additional information. We speculate that this is because some necessary reasoning procedure is missing in this instruction data so that the fine-tuning procedure fails to capture the implicit relation between the Request, the Tagged_Content, and the desired output TAGs. This indicates that some reasoning information about the extracting procedure will be necessary for instruction fine-tuning.
- **Accurate extraction with minimal noise:** when preparing the instruction set, the common practice involves calling existing LLM APIs to significantly automate and accelerate this labor-intensive process. However, such an approach comes at the cost of degraded instruction data quality, as the generated content can be noisy, containing irrelevant, inaccurate, or misleading information. Thus, it is necessary to apply some additional refinement mechanisms to ensure the final instruction set is both accurate and clear.

In order to satisfy these requests, we enumerate how we construct the instruction set as below:

Data Source. To ensure quality and diversity in our instruction set, we collect real-world production data from 01AI's inference API serving system. Concretely, we collect raw API requests during one week. Given that each request is associated with multiple web pages returned by the search engine, to avoid excessive repetition and maintain request diversity, we randomly sampled only two contents from all the returned web pages for each request.

Prepare Instruction Data with LLMs. After collecting the raw data from the production environment, we need to use LLMs to process the input data by the format of <Request, Tagged_Content>, to automate the extraction process of the desired collection of TAGs. To introduce some information for the necessary reasoning procedure, we design the following procedure: we ask the LLM not only to generate the desired collection of TAGs, but also to generate a short Summary of the corresponding content associated with each of the TAGs. Formally, this LLM-aided instruction construction procedure can be defined as follows:

$$\begin{aligned} & \langle \text{Request}, \text{Tagged_Content} \rangle \\ \rightarrow & \langle \text{Request}, \text{Tagged_Content}, \{ \langle \text{Summary}_{k_i}, \text{TAG}_{k_i} \rangle_{i=1,\dots,n} \} \rangle \end{aligned}$$

which means, for each <Request, Tagged_Content> pair, we generate a set of n relevant <Summary_{k_i}, TAG_{k_i}> pairs. Besides such instructions notated as Instruction-A, we find that empirically it is also necessary for the instruction set to include the instructions that only contain the TAGs without the Summary. To obtain such instruction data, for each Instruction-A, we manually remove the Summary to generate Instruction-B, which can be formalized as:

$$\begin{aligned} & \langle \text{Request}, \text{Tagged_Content}, \{ \langle \text{Summary}_{k_i}, \text{TAG}_{k_i} \rangle_{i=1, \dots, n} \} \rangle \\ \rightarrow & \langle \text{Request}, \text{Tagged_Content}, \{ \langle \text{TAG}_{k_i} \rangle_{i=1, \dots, n} \} \rangle \end{aligned}$$

During SFT, we include both **Instruction-A** and **Instruction-B**. Empirically, this mixture of two instruction types works well — we speculate that this is owing to the **Instruction-A** helps to bridge the semantic meaning of the context ($\langle \text{Request}, \text{Tagged_Content} \rangle$) to the associated TAGs, while **Instruction-B** further help to teach the EXTRACTOR-LLM to follow the desired information extracting instruction under the Internet search augmented generation paradigm.

Improve the Robustness of the Instruction Set. We also improve the robustness of the instruction set before, during, and after the processing by the OpenAI API calls.

- **Before** processing by the API, we carefully filter the input pair by the format of $\langle \text{Request}, \text{Tagged_Content} \rangle$, and remove the pairs with **Tagged_Content** less than 100 tokens — such short web page from the search engine API usually contained limited helpful information.
- **During** processing by the API, since the API does not generate deterministic output in multiple inferences given the same input prompt, for each $\langle \text{Request}, \text{Tagged_Content} \rangle$ pair, we process the same pair twice with different random seeds; if the results are consistent, we keep the instruction otherwise we abandon such contradictory result. We find such a policy can remove a lot of low-quality data and decrease the uncertainty and noise introduced by API calls.
- **After** processing by API, since it is important for the EXTRACTOR-LLM to determine whether there is relevant information from the content, we manually check the ratio of the instructions that should return `None` as the results, and ensure the ratio of `None` to be 5% in the instruction set, which is determined based on empirical tests.

4.3 SFT Training Details

We choose the YI-9B-CHAT-16K model as the initial model for instruction fine-tuning. Note that compared with YI-9B-BASE-16K, this model has been supervised and fine-tuned to obtain basic instruction-following capabilities. We use the prepared instruction set notated as \mathcal{I} to train the base model in a standard auto-regressive manner, with the optimization objective defined as follows:

$$\mathcal{L} = \min_{\theta} \mathbb{E}_{j \sim \mathcal{I}} [-\log p(\Omega_j | \langle \text{Request}_j, \text{Tagged_Content}_j \rangle; \theta)]$$

where \mathcal{L} represents the negative likelihood to be minimized, Ω_j is the expected output for the j -th $\langle \text{Request}_j, \text{Tagged_Content}_j \rangle$ pair, θ denotes the model parameters. The expectation $\mathbb{E}_{j \sim \mathcal{I}}$ is taken over the instruction set \mathcal{I} . We train the model for three epochs, with a maximum sequence length of 16K tokens. The learning rate is set to 5×10^{-6} initially, and a cosine learning rate scheduler was employed. To prevent overfitting, the training data is shuffled at the beginning of each epoch. We choose the FusedAdam optimizer with the hyper-parameter configured as $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1 \times 10^{-8}$. The global batch size is set to 16. The training computation was conducted on four NVIDIA H800 GPUs, each with 80GB of memory by the DeepSpeed [45] parallel training system.

4.4 Data Preparation for DPO

In the second DPO phase, we further optimize the EXTRACTOR-LLM’s output TAGs towards an end-to-end preference estimated by OpenAI API (i.e., GPT-4O) considered as a *LLM judge*. Note that we can also replace this role with human preference. We modify the single answer grading prompt provided in a prior study [48] to design the judging instruction. To be more specific, we add more detailed grading explanations and standards while asking the Judge LLM to be more critical and precise when scoring. We denote EXTRACTOR-LLM after SFT as π_{ref} . The data preparation for the DPO phase needs to provide a set of data in the format of $\{\text{Request}, \text{Tagged_Content}, y^+, y^-\}$, where y^+ indicates the extracted TAGs is preferred by the LLM judge, and

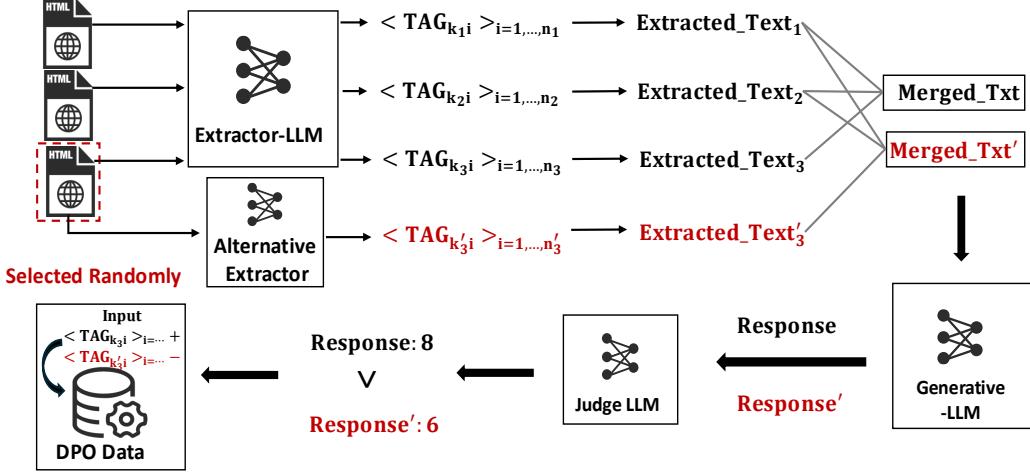


Figure 3: An illustration Of DPO Training Preparation.

y^- indicates they are less favored. We explain how we construct y^+ and y^- from the TAGs extracted by different extractors as below:

- **Step 1. Generate response using π_{ref} as extractor:** given a Request, for each Tagged_Content retrieved by the Internet search engines, we firstly use our EXTRACTOR-LLM after SFT noted as π_{ref} to perform extractions and obtain $\langle \text{TAG}_{k_i} \rangle_{i=1, \dots, n}$. The extracted $\langle \text{TAG}_{k_i} \rangle_{i=1, \dots, n}$ are referred back to the m corresponding texts, obtaining Extracted_Text. Then we ensemble all the extracted text $\langle \text{Extracted_Text}_k \rangle_{k=1, \dots, m}$ from the corresponding tagged contents $\langle \text{Tagged_Content}_k \rangle_{k=1, \dots, m}$ into Merged_Text. Lastly, we combine the Merged_Text and Request together as an enriched prompt, which is sent to the generative LLM to generate a Reponse.
- **Step 2. Permute response by partial replacement:** we use GPT-4O as the alternative extractor model to perform TAG extraction, enabling the judge to make comparisons during the later evaluation process. To implement the replacement, we randomly select one tagged content, denoted as Tagged_Content_i, from all the tagged contents returned by Internet search engines. For this Tagged_Content_i, we replace its corresponding Extracted_Text_i with Extracted_Text'_i, where Extracted_Text'_i is the text extracted by GPT-4O from Tagged_Content_i using the same procedure described in **Step 1**. All other extracted texts, denoted as $\langle \text{Extracted_Text}_k \rangle_{k=1, \dots, i-1, i+1, \dots, m}$, remain unchanged, and the updated Extracted_Text'_i is then merged with all the other unchanged texts $\langle \text{Extracted_Text}_k \rangle_{k=1, \dots, i-1, i+1, \dots, m}$ to construct Merged_Text'. This Merged_Text' is combined with Request to form an enriched prompt and later be fed into the backbone GENERATIVE-LLM to generate a new response, Reponse'. The upper part of Figure 3 illustrates the process of generating both Reponse and Reponse' for a given Request and selected Tagged_Content_i.
- **Step 3. Obtain the preference from the LLM judge:** for each given Request and a chosen Tagged_Content_i, we now have two different responses generated by the backbone GENERATIVE-LLM (i.e., YI-LIGHTNING). Reponse is generated with information extracted entirely by π_{ref} . And Reponse' uses the majority of extractions by π_{ref} , except for the chosen Tagged_Content_i, where the extraction is performed by GPT-4O. We ask the LLM judge to score each response separately based on how well it answers the Request. The better-scored response is chosen as the preferred response, denoted by Response⁺, and the other is denoted by Response⁻.
- **Step 4. Construct the dataset:** up to this point, we have a pair of responses, Response⁺ and Response⁻, for a given Request and a chosen Tagged_Content_i. For each response, we trace back to the corresponding tags noted as $\langle \text{TAG}_{k_i} \rangle_{i=1, \dots, n}$ or $\langle \text{TAG}_{k'_i} \rangle_{i=1, \dots, n'}$ extracted from Tagged_Content_i, depending on which is used to generate this response. And the extracted TAGs corresponds to Response⁺ is constructed as y^+ , while the extracted TAGs corresponds to Response⁻ is y^- . We then construct $\{\text{Request}, \text{Tagged_Content}, y^+, y^-\}$

as a piece of data in the DPO dataset. For each Request, we repeat **Step 2** and **Step 3** only twice to obtain two such pieces of data, so that more Requests can be included in the dataset. This can increase the diversity of different Requests during the DPO training.

4.5 DPO Training Details

Given the training dataset prepared for DPO, we want to obtain a trainable actor model denoted as π_θ for which it can extract TAGs more aligned with the judge’s preference. We initialize π_θ with the reference model π_{ref} , which is the fine-tuned EXTRACTOR-LLM after SFT phase. We use the prepared DPO training dataset noted as \mathcal{D} to train the actor model π_θ , with the optimization objective defined as follows:

$$\mathcal{L} = E_{(x, y^+, y^-) \sim \mathcal{D}} [-\log \sigma(\beta \log \frac{\pi_\theta(y_+|x)}{\pi_{ref}(y_+|x)} - \beta \log \frac{\pi_\theta(y_-|x)}{\pi_{ref}(y_-|x)})]$$

where x consist of a given Request and a chosen Tagged_Content, $\pi_\theta(y_+|x)$ is the probability of predicting y^+ when giving x as input, and $\pi_\theta(y_-|x)$ is the probability of predicting y^- when giving x as input. $\pi_{ref}(y_+|x)$ and $\pi_{ref}(y_-|x)$ are defined in a similarly way. σ is the sigmoid function, and β is the coefficient hyperparameter [49].

Minimizing \mathcal{L} suggests that we expect π_θ to extract TAGs aligned with y^+ and less like y^- when given x . Meanwhile, we don’t want π_θ to deviate too much from π_{ref} , so that it can still extract properly, and doesn’t turn into something that is rated very high, but has a garbled extraction of TAGs.

5 Evaluation

Our empirical study lies in two aspects: (i) the end-to-end performance of the Internet search augmented generation paradigm; and (ii) the capability of the EXTRACTOR-LLM to identify the relevant information. Concretely, we aim to answer the following four questions:

- What is the end-to-end performance of the response quality when comparing the proposed Internet search augmented generation paradigm to other RAG paradigms?
- What is the generative cost in terms of the processed input tokens when contrasting the proposed Internet search-augmented generation paradigm with other RAG approaches?
- Can the EXTRACTOR-LLM extract the relevant information accurately and precisely?
- How robust is the EXTRACTOR-LLM that can reject inappropriate contents rather than generate hallucinations when there is no relevant information?

5.1 End-to-End Evaluation

To evaluate the (i) end-to-end performance regarding the quality of the generated content and (ii) the generative cost in terms of the processed input tokens between the proposed Internet search augmented generation paradigm and other RAG paradigms, we design the following configurations:

Evaluation Task Data. The evaluation dataset is constructed from real-world data. We randomly sampled 500 queries from 01AI’s real-world API request — we confirmed that there is no overlap between this evaluation dataset and the raw data we used to prepare the instruction set introduced in Section 4.2. After deduplication — 463 unique requests remained, accompanied by the content scraped from corresponding web pages returned by search engines.

Baseline Systems. We use *Internet-SAG-Ext* to denote our proposed Internet search augmented generation paradigm. We consider two baseline systems:

- **Internet-SAG-Naive:** in this baseline system, we remove the EXTRACTOR-LLM component in *Internet-SAG-Ext*, and rely on the GENERATIVE-LLM’s ability to comprehensively integrate the information from the search

API returned contents. Concretely, all of the content retrieved from the Internet via search engines is simply concatenated and directly used as reference materials for the GENERATIVE-LLM to respond to the API request.

- **VectorDB-RAG:** we configure a vector database to manage all of the Internet-retrieved content in the evaluation task data. Concretely, we employ the standard RAG paradigm. During the building phase, all content associated with the API request is split into overlapping chunks according to specified chunk size and chunk overlap parameters. The embedding model is then used to generate vector representations of each chunk, which are stored in the vector database. During the retrieval phase, a vector representation of the request is generated, and the top-k most relevant chunks (based on cosine similarity) are retrieved; The top-k relevant chunk is concatenated to form the final reference materials for the GENERATIVE-LLM. We use BGE-m3 [46] from BAAI as the embedding model, FAISS [50] as the vector store, with hyperparameters set as: chunk-size = 512, chunk-overlap = 128, top-k = 12, neighbor-num = 1 (indicating that chunks immediately before and after the relevant chunk should be included, if available).

GENERATIVE-LLM Configuration. We evaluate three models with different scales as the GENERATIVE-LLMs in different RAG systems, including the open-source models QWEN2-72B and YI-34B, as well as the closed-source model YI-LARGE.

Evaluation Metrics. We employ GPT-4o as the judge to evaluate the quality of responses generated by different RAG systems. To maintain consistency, we specifically use the GPT-4o-2024-05-13 version for this evaluation. To address positional bias when applying *LLM judge*, we switch the order of the two responses specified in the prompt-JUDGE and conduct two separate evaluations [48]. Responses that yield consistent results across both evaluations are accepted, while those with inconsistent outcomes are classified into a special positional bias (P-BIA) category. Regarding length and structure biases, the generation prompt prompt-GEN explicitly instructs the model to provide detailed answers and use markdown formatting unless otherwise specified in the API request. For the judgment prompt prompt-JUDGE, used by GPT-4o for pairwise comparison, we reuse the prompt provided in a prior study [48].

Experiment Results. The experimental results of end-to-end performance regarding the quality of the generated response are summarized in Table 1, where each row presents the performance outcomes (WIN, TIE, LOSE, and P-BIA) of the *Internet-SAG-Ext* paradigm compared to baseline paradigm equipped with different GENERATIVE-LLMs. On the other hand, the generative cost in terms of the consumed input tokens by the GENERATIVE-LLM is shown in Table 2.

Table 1: End-to-End evaluation on Internet-SAG-Ext system versus other two baseline systems, includes *Internet-SAG-Naive* and *VectorDB-RAG*, employed with various GEN-LLMs (GENERATIVE-LLMs). The value in each cell denotes the number of WIN, TIE, LOSE, or P-BIA respectively.

Baseline SYS	GEN-LLM	WIN	TIE	LOSE	P-BIA
Internet-SAG-Naive	YI-LARGE	143	22	89	209
	QWEN2-72B	134	20	93	216
	YI-34B	137	22	85	219
VectorDB-RAG	YI-LARGE	134	15	95	219
	QWEN2-72B	258	23	66	116
	YI-34B	185	12	61	205

Discussion. The experimental results indicate that our proposed *Internet-SAG-Ext* paradigm consistently outperforms the baseline systems across different GENERATIVE-LLMs and achieves the lowest cost in terms of the input tokens processed by GENERATIVE-LLMs. When using the YI-LARGE as GERERATIVE-LLM, *VectorDB-RAG* demonstrates a slight advantage over *Internet-SAG-Naive*. This suggests that simply concatenating content returned by the search engine and passing it to the generative model is insufficient. A plausible explanation is that occupying an excessively long sequence window can degrade model performance, as noted in [37]

Table 2: The total input token cost of generative LLM provided by different systems. M means million. The second column indicates the number of total input token cost.

SYS	#Input Token
Internet-SAG-Ext	2.04 M
Internet-SAG-Naive	3.88 M
VectorDB-RAG	2.59 M

concerning inflated window length. On the other hand, when using the QWEN2-72B, *VectorDB-RAG* performs worse than *Internet-SAG-Ext*. In many responses generated by QWEN2-72B within the *VectorDB-RAG* paradigm, we observed that the model autonomously created its own questions based on the references and then answered them, completely ignoring the original request specified in the prompt-GEN. This issue may arise from the pre-configured chunk size and top-K settings in *VectorDB-RAG*, which could have introduced irrelevant information during retrieval. When the proportion of irrelevant information becomes too large, it may trigger such observed issues in QWEN2-72B. When using the YI-34B model, *Internet-SAG-Naive* outperforms *VectorDB-RAG*. As mentioned earlier, the inherent issue with *VectorDB-RAG* lies in its pre-configured chunk size and top-K settings, which may introduce irrelevant information during retrieval. Although we attempted to filter out irrelevant information using instructions within prompt-GEN, the other two models, except for YI-LARGE, exhibited lower resistance to interference and reduced robustness when handling irrelevant information. Nevertheless, *Internet-SAG-Ext* demonstrates superior performance across various scenarios. After the EXTRACTOR-LLM performs information extraction, the text length is reduced, preventing the model from occupying too much of the sequence window and degrading its performance. Additionally, irrelevant information is filtered out, reducing the risk of the model being biased by irrelevant data. Moreover, since most LLM service platforms charge based on the number of tokens processed, the information processed by *Internet-SAG-Ext* requires fewer input tokens for the generative LLM compared to the baselines, providing a cost advantage as well. Specifically, we reduce 21% and 47% input token cost for Generative LLM compared to *VectorDB-RAG* and *Internet-SAG-Naive* respectively.

5.2 EXTRACTOR-LLM Evaluation

In this section, we comprehensively assess the capabilities of our EXTRACTOR-LLM, primarily focusing on extracting relevant information, accurately locating tags, and effectively rejecting irrelevant content.

Benchmark. We discuss three benchmarks to comprehensively evaluate the context retrieval ability of our model in different aspects, including the synthetic, open-source, and real-world benchmarks. We begin by adopting a standard methodology in the retrieval field to construct our *Synthetic Benchmark*, which enables fine-grained evaluations through controlled variations in context length and the depth of the answer where it is inserted in the context. The *Open-source Benchmark* adapts established datasets to assess specific retrieval capabilities, providing extensive test cases. The *Real-world Benchmark* utilizes authentic, request-based data to test the model’s adaptability and robustness under practical application conditions.

- **Synthetic benchmark:** we propose a *Synthetic Benchmark* based on the previous needle-based tasks NeedleInAHaystack [51, 52], a standard methodology for information retrieval. This task performs fine-grained evaluations on the retrieval of specific sentences—referred to as “needles”—that are embedded across two dimensions: varying context lengths and differing depths within the same context. In addition to the original sets of needles from the NeedleInAHaystack benchmark, we designed three additional types of needles, targeting distinct task categories: multi-hop reasoning, multi-question, and multi-answer retrieval tasks.
 - (i) The multi-hop reasoning task draws inspiration from [53]. We developed distinct sets of needles that are interconnected through reasoning dependencies.
 - (ii) The multi-question task, based on [54], evaluates the model’s capability to retrieve relevant information for multiple questions simultaneously. For this purpose, we created multiple questions, where the answer to each question corresponds to a specific subset of needles.
 - (iii) The multi-answer task addresses limitations identified in previous studies [55], which highlight the tendency of

LLMs to exhibit "lazy retrieval". LLMs tend to only retrieve a subset of possible answers while neglecting others. To assess this, we designed tasks that require the model to extract all the correct answers for a single query. Although our synthetic needle-based benchmark benefits from adhering to a widely recognized standard methodology, it has certain limitations. Specifically, the needles are inserted in irrelevant contexts, which isn't the case in a real usage scenario. Another problem lies in the limited amount of manually designed sets of needles.

- **Open-source benchmark:** to address the limitations of *Synthetic Benchmark*, we propose an *Open-source Benchmark* derived from three open-source task-dedicated datasets, each corresponding to a specific task category. Unlike retrieving the needles which are inserted in irrelevant background context, this approach provides a more general assessment of retrieving information within context-relevant passages based on the questions. It also contains extensive test cases, enabling a comprehensive evaluation. To integrate these open-source datasets into our workflow, the original test cases of the datasets are reformatted into a tag-based structure. We introduce the construction of three categories of tasks as follows. (i) Reasoning tasks are derived from Multi-hop QA Dataset [56]. This dataset is designed to evaluate reasoning ability by challenging a model to analyze and combine information from multiple paragraphs to a given question. As introduced in the pre-processing Section 4.1, we enclose each sentence of context in this dataset within the $\langle \text{TG}-i \rangle$ and $\langle / \text{TG}-i \rangle$ tag-pairs. The tags associated with the necessary reasoning sentences provided by the original Multi-hop QA Dataset are utilized as ground-truth tags for our Reasoning tasks. (ii) Multi-answer tasks are made from MulTiple [57] dataset. This dataset assesses the ability of LLMs to answer multiple questions simultaneously, ensuring no question is overlooked. The dataset includes two types of questions: question-dependent questions and document-dependent questions. The former explicitly specifies the number of demanded answers, while the latter challenges LLMs to determine the number of answers itself. For general purposes, document-dependent questions are selected as evaluation queries, assessing whether LLMs are able to retrieve all the answers associated with a given question. Document-dependent questions with auxiliary information of answer numbers are provided to GPT-4o to generate the ground truth. (iii) Multi-question tasks are constructed from MTI BENCH [58]. This benchmark is designed to evaluate LLMs' ability to handle multiple questions with one single inference. We first extract all the original test cases with (question, contents) pairs from the MTI BEBCH dataset, for which can assess the information retrieval ability. Then, we crafted 300 multi-question test cases by combining multiple extracted (question, contents) pairs together. The original dataset only provides the index of paragraphs where the answers are in, even though only a few sentences within the paragraph are truly demanded. Thus, we utilize GPT-4o to refine these answers, extracting only the target sentences while ignoring irrelevant parts, thereby providing more accurate and fine-grained ground truth tags for our multi-question tasks.
- **Real-world benchmark:** we directly extract request-based tests from real-world data, focusing on assessing the models' general performance in application scenarios. We sampled 500 pairs of data from real-world data sources, each consisting of a request and corresponding contents. Subsequently, those contents are tagged by the finite state machine described in Section 4.1. This approach tests our models under diverse realistic scenarios. By leveraging real-world data, the benchmark provides a more accurate measure of the model's robustness and adaptability compared with synthetic data.

Evaluation Metrics. To better align with our three benchmarks, we employ the following metrics regarding the relevance of the data. The data will be considered as *positive relevant* if the context contains information relevant to the user query. Otherwise, it will be considered *negative relevant*. All the test data in the *Synthetic Benchmark* and *Open-source Benchmark* is *positive relevant*, since the tags for this evaluation corresponding to each needle are predetermined. More specifically, we extract $\langle \text{Summary}_{k_i}, \text{TG}_{k_i} \rangle_{i=1, \dots, n}$ for each piece of evaluation data outlined in Section 4.2. The test data adopted for request-based tests in the Real-world benchmark may be *negative relevant*, as there is the possibility of the absence of ground truth. we use Precision, Recall, and F1 score, typical assessment methods for multi-class classification, as our key evaluation metrics. We formulate the measurement of *negative relevant* data as a multi-label matching problem, where Exact Match Ratio considers only fully correct matches of each TAGs, is particularly suitable for scenarios where the ground truth is None. By utilizing this property, the Exact Match Ratio is adopted to examine a model's capacity to reject irrelevant data, made-up, or inconsistent data.

Baseline Models. We select a set of baseline models for comparison, including open-source models such as QWEN2-72B and YI-9B-CHAT, along with the closed-source model GPT-4O. Additionally, we compare these models against our proposed EXTRACTOR-LLM.

Experiment Results and Discussion. This subsection presents the experiment results on synthetic, open-source benchmarks, and real-world benchmarks, respectively. For the needle-based *Synthetic Benchmark*, we calculate the average F1 score across all results, considering every context length and insertion point depth for each needle.

- **Sythetic benchmark result:** The results of *Sythetic Benchmark* tests are summarized in Table 3. As one of the most powerful models, GPT-4O achieves leading performance across all task scenarios, ranking first in base, reasoning, and multi-answer retrieval tasks, and tying with EXTRACTOR-LLM in the multi-question tasks. While EXTRACTOR-LLM delivers slightly lower scores than GPT-4O in other three tasks, it remains a strong competitor, demonstrating robust and consistent performance across all tasks. These results highlight EXTRACTOR-LLM’s reliability in our context retrieval workflow, standing out as a versatile alternative just behind GPT-4O in key metrics.

Table 3: Average F1 scores (higher is better) across different categories of tasks in *Sythetic Benchmark*. “Reasoning” refers to “Reasoning tasks,” “Multi-Ans” to “Multi-answer tasks,” and “Multi-Q” to “Multi-question Tasks.”

Model	Base	Reasoning	Multi-Ans	Multi-Q
YI-9B-CHAT	0.328	0.147	0.284	0.166
QWEN2-72B	0.732	0.346	0.675	0.946
GPT-4O	0.969	0.904	0.986	1.0
EXTRACTOR-LLM	0.925	0.803	0.973	1.0

- **Open-source benchmark result:** The results of the *Open-source Benchmark* tests are summarized in Table 4. GPT-4O demonstrates exceptional performance, achieving the highest score in the multi-question tasks. However, EXTRACTOR-LLM achieves a very close score in the multi-question tasks while surpassing all other models, including GPT-4O, in both reasoning and multi-answer tasks, showcasing its cutting-edge capability to handle complex reasoning and multi-answer extraction challenges with remarkable accuracy. Notably, other models such as YI-9B-CHAT and QWEN2-72B trail behind, further emphasizing the advancements achieved by EXTRACTOR-LLM and GPT-4O.

Table 4: Average F1 scores (higher is better) across different categories of tasks in *Open-source Benchmark*. “Multi-Ans” refers to “Multi-answer tasks,” and “Multi-Q” to “Multi-question Tasks.”

Model	Reasoning	Multi-Ans	Multi-Q
YI-9B-CHAT	0.361	0.328	0.439
QWEN2-72B	0.568	0.441	0.570
GPT-4O	0.591	0.536	0.755
EXTRACTOR-LLM	0.724	0.581	0.687

- **Real-world benchmark result:** The results of *Real-world benchmark* are summarized in Table 5. GPT-4O delivers high performance across all metrics, slightly leading in Recall and achieving competitive scores in other areas. However, EXTRACTOR-LLM excels by surpassing GPT-4O in Precision, F1 score, and Exact Match (EM), highlighting its superior accuracy and robustness in real-world scenarios. In contrast, YI-9B-CHAT’s accuracy is particularly low because it tends to output a large portion of the TAGs from content indiscriminately, resulting in a high Recall but a very low Precision as well as a pretty low Exact Match Ratio. These results emphasize the consistent advancements of EXTRACTOR-LLM showcasing the strong context retrieval abilities under the real-world scenario.

Table 5: *Real-world Benchmark* results across different models, evaluated using metrics including Recall, Precision, F1, and Exact Match Ratio (EM), where higher values indicate better performance.

Model	Recall	Precision	F1	EM
YI-9B-CHAT	0.8383	0.4722	0.6041	0.1729
QWEN2-72B	0.7145	0.6482	0.6797	0.7078
GPT-4O	0.8357	0.6691	0.7431	0.7563
EXTRACTOR-LLM	0.8353	0.7139	0.7698	0.7985

6 Related Work

RAG is a powerful approach that significantly enhances the quality of LLM-generated content by retrieving and integrating relevant information from data sources not visited during the generative LLM training phase. In this section, we focus on the relevant techniques for Internet-level search augmented generation and model tuning techniques under the RAG paradigm instead of comprehensive RAG literature review [2, 3].

6.1 Search Augmented Generation

Classic RAG systems leverage vector databases [25, 26, 27, 28, 29, 30, 31, 32] to retrieve semantically similar chunks of text from a static corpus, which are usually limited by the scope of their pre-indexed datasets. On the other hand, Internet-level dynamic search-augmented generation [4, 5, 6, 7] leverages the most up-to-date information available online to significantly enhance the quality of question-answering [4, 6, 7] and dialogue generation [5] within some specific domains. Furthermore, this approach effectively addresses the challenge of hallucinations in LLM-generated outputs [5, 9] by grounding the generation process in up-to-date, verifiable data retrieved from the web. By integrating real-time information, these systems [4, 5, 6, 7] ensure that the responses are not only relevant and accurate but also reflective of the latest developments, thereby improving the overall reliability of the generative inference.

6.2 Model Tuning in RAG

To enhance the quality of outputs generated by LLMs, various learning approaches and learnable components have been explored within the RAG paradigm to integrate retrieved information effectively [59, 60, 61]. For example, WebGPT [4] fine-tuned the generative LLM, i.e., GPT-3, to answer long-form questions by using a text-based web-browsing environment, enhancing both retrieval and synthesis in an end-to-end manner; recently, RankRAG [62] instruction-tunes an end-to-end generative LLM for the dual purpose of context ranking and answer generation in RAG. Instead of re-training the generative LLM, DRAGIN [40] introduces a module dynamically determining when and what to retrieve based on the LLM’s information needs during the text generation process; similarly, an advanced continual knowledge learning approach has been explored to selectively decide whether to access the Internet or not dynamically [7]. Query re-write module [39, 41] has also been studied to eliminate ambiguity in the original input prompt; for example, RQ-RAG [41] tuned a 7B plugin model in RAG paradigm to dynamically refine search queries through explicit rewriting, decomposition, and disambiguation.

7 Conclusion

In this paper, we have presented a novel paradigm, Internet search augmented generation for LLMs that can integrate real-time information through the use of standard search engine APIs, rather than relying on a static pre-processed corpus. This method addresses the limitations of the traditional RAG paradigm, which often fails to incorporate the most up-to-date information. Concretely, we implement a PARSER-LLM that determines whether Internet-augmented generation is necessary and, if so, extracts the relevant search keywords in a single inference pass; a *mixed ranking strategy* that mitigates biases introduced by search engine APIs by re-ranking

the retrieved HTML files; and an EXTRACTOR-LLM designed to accurately and efficiently extract relevant information from the fresh content within each HTML file. Our extensive empirical evaluations demonstrate that this Internet search augmented generation paradigm significantly enhances the quality of the generated content. The successful deployment of this system in a production environment, specifically within 01.AI’s generative inference framework, further validates the practical applicability of our approach.

References

- [1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [2] Penghao Zhao, Hailin Zhang, Qinhuan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.
- [3] Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.06211*, 2024.
- [4] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [5] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, 2022.
- [6] Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*, 2022.
- [7] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. The web can be your oyster for improving large language models. *arXiv preprint arXiv:2305.10998*, 2023.
- [8] Carl Duffy, Jaehoon Shim, Sang-Hoon Kim, and Jin-Soo Kim. Dotori: A key-value ssd based kv store. *Proceedings of the VLDB Endowment*, 16(6):1560–1572, 2023.
- [9] Weijian Xie, Xuefeng Liang, Yuhui Liu, Kaihua Ni, Hong Cheng, and Zetian Hu. Weknow-rag: An adaptive approach for retrieval-augmented generation integrating web search and knowledge graphs. *arXiv preprint arXiv:2408.07611*, 2024.
- [10] OpenAI. Introducing chatgpt search. <https://openai.com/index/introducing-chatgpt-search/>, 2024.
- [11] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.
- [12] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- [13] Derek Paulsen, Yash Govind, and AnHai Doan. Sparkly: A simple yet surprisingly strong tf/idf blocker for entity matching. *Proceedings of the VLDB Endowment*, 16(6):1507–1519, 2023.
- [14] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

- [15] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335, 2024.
- [16] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.
- [17] Yifan Wang, Haodi Ma, and Daisy Zhe Wang. Lider: an efficient high-dimensional learned index for large-scale dense passage retrieval. *Proceedings of the VLDB Endowment*, 16(2):154–166, 2022.
- [18] Zhou Zhang, Zhaole Chu, Peiquan Jin, Yongping Luo, Xike Xie, Shouhong Wan, Yun Luo, Xufei Wu, Peng Zou, Chunyang Zheng, et al. Plin: A persistent learned index for non-volatile memory with high performance and instant recovery. *Proceedings of the VLDB Endowment*, 16(2):243–255, 2022.
- [19] Xi Zhao, Yao Tian, Kai Huang, Bolong Zheng, and Xiaofang Zhou. Towards efficient index construction and approximate nearest neighbor search in high-dimensional spaces. *Proceedings of the VLDB Endowment*, 16(8):1979–1991, 2023.
- [20] Zhaoyan Sun, Xuanhe Zhou, and Guoliang Li. Learned index: A comprehensive experimental evaluation. *Proceedings of the VLDB Endowment*, 16(8):1992–2004, 2023.
- [21] Pengfei Li, Hua Lu, Rong Zhu, Bolin Ding, Long Yang, and Gang Pan. Dili: A distribution-driven learned index. *Proceedings of the VLDB Endowment*, 16(9):2212–2224, 2023.
- [22] Alexandros Zeakis, George Papadakis, Dimitrios Skoutas, and Manolis Koubarakis. Pre-trained embeddings for entity resolution: an experimental analysis. *Proceedings of the VLDB Endowment*, 16(9):2225–2238, 2023.
- [23] Konstantinos Lampropoulos, Fatemeh Zardbani, Nikos Mamoulis, and Panagiotis Karras. Adaptive indexing in high-dimensional metric spaces. *Proceedings of the VLDB Endowment*, 16(10):2525–2537, 2023.
- [24] Shunkang Zhang, Ji Qi, Xin Yao, and André Brinkmann. Hyper: A high-performance and memory-efficient learned index via hybrid construction. *Proceedings of the ACM on Management of Data*, 2(3):1–26, 2024.
- [25] Miao Cai, Junru Shen, Yifan Yuan, Zhihao Qu, and Baoliu Ye. Bonsaikv: Towards fast, scalable, and persistent key-value stores with tiered, heterogeneous memory system. *Proceedings of the VLDB Endowment*, 17(4):726–739, 2023.
- [26] Hailin Zhang, Penghao Zhao, Xupeng Miao, Yingxia Shao, Zirui Liu, Tong Yang, and Bin Cui. Experimental analysis of large-scale learnable vector storage compression. *Proceedings of the VLDB Endowment*, 17(4):808–822, 2023.
- [27] Ziyi Lu, Qiang Cao, Hong Jiang, Yuxing Chen, Jie Yao, and Anqun Pan. Fluidkv: Seamlessly bridging the gap between indexing performance and memory-footprint on ultra-fast storage. *Proceedings of the VLDB Endowment*, 17(6):1377–1390, 2024.
- [28] Geoffrey X Yu, Markos Markakis, Andreas Kipf, Per-Åke Larson, Umar Farooq Minhas, and Tim Kraska. Treeline: an update-in-place key-value store for modern storage. *Proceedings of the VLDB Endowment*, 16(1), 2022.
- [29] Ishtiyaque Ahmad, Divyakant Agrawal, Amr El Abbadi, and Trinabh Gupta. Pantheon: Private retrieval from public key-value store. *Proceedings of the VLDB Endowment*, 16(4):643–656, 2022.
- [30] Zhiqi Wang and Zili Shao. Mirrorkv: An efficient key-value store on hybrid cloud storage with balanced performance of compaction and querying. *Proceedings of the ACM on Management of Data*, 1(4):1–27, 2023.

- [31] Dingheng Mo, Fanchao Chen, Sijiang Luo, and Caihua Shan. Learning to optimize lsm-trees: Towards a reinforcement learning based key-value store for dynamic workloads. *Proceedings of the ACM on Management of Data*, 1(3):1–25, 2023.
- [32] Yongye Su, Yinqi Sun, Minjia Zhang, and Jianguo Wang. Vexless: A serverless vector data management system using cloud functions. *Proceedings of the ACM on Management of Data*, 2(3):1–26, 2024.
- [33] OpenAI. Openai gpt-4o, 2024.
- [34] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [35] Alex Young, Bei Chen, Chao Li, Chengan Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jian-qun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [36] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024.
- [37] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What’s the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.
- [38] Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. *arXiv preprint arXiv:2310.03184*, 2023.
- [39] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, 2023.
- [40] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*, 2024.
- [41] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- [42] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530, 2007.
- [43] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 113–120, 2014.
- [44] Huilin Chen, Xu Xu, Haiquan Li, Xinyue Yu, Ruting Pan, and Zhaoyang Zhang. A database of ambiguous chinese characters with measures for meaning dominance and meaning balance. *Applied Psycholinguistics*, 45(4):695–716, 2024.
- [45] Microsoft. Extreme speed and scale for dl training and inference. <https://www.deepspeed.ai/>, 2024.
- [46] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.
- [47] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

- [48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.
- [49] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.
- [50] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.
- [51] gkamradt. Llmtest needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.
- [52] Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in 1 million context window?, 2024.
- [53] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.
- [54] Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. Multi-task inference: Can large language models follow multiple instructions at once? *arXiv preprint arXiv:2402.11597*, 2024.
- [55] Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, 2023.
- [56] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.
- [57] Mtqa. <https://github.com/multipledata/MTQA>, 2023.
- [58] Guijin Son, Sangwon Baek, Sangdae Nam, Ilgyun Jeong, and Seungone Kim. Multi-task inference: Can large language models follow multiple instructions at once? 2024.
- [59] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023.
- [60] Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen Meng, and James Glass. Search augmented instruction learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3717–3729, 2023.
- [61] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024.
- [62] Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *arXiv preprint arXiv:2407.02485*, 2024.