CA-GAR: Context-Aware Alignment of LLM Generation for Document Retrieval

Heng Yu[1*], **Junfeng Kang**[1*], **Rui Li**[1], **Qi Liu**[1,2†], **Liyang He**[1],
Zhenya Huang[1,2], **Shuanghong Shen**[2], **Junyu Lu**[1,2]
[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yh112358_1321,kangjf,ruili2000,heliyang,lujunyu}@mail.ustc.edu.cn
{shshen}@iai.ustc.edu.cn, {qiliuql,huangzhy}@ustc.edu.cn

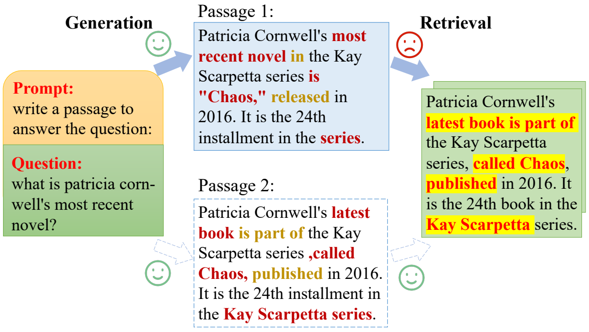
Abstract

Information retrieval has evolved from traditional sparse and dense retrieval methods to approaches driven by large language models (LLMs). Recent techniques, such as Generation-Augmented Retrieval (GAR) and Generative Document Retrieval (GDR), leverage LLMs to enhance retrieval but face key challenges: GAR's generated content may not always align with the target document corpus, while GDR limits the generative capacity of LLMs by constraining outputs to predefined document identifiers. To address these issues, we propose **C**ontext-**A**ware **G**eneration-**A**ugmented **R**etrieval (**CA-GAR**), which enhances LLMs by integrating corpus information into their generation process. CA-GAR optimizes token selection by incorporating relevant document information and leverages a Distribution Alignment Strategy to extract corpus information using a lexicon-based approach. Experimental evaluations on seven tasks from the BEIR benchmark and four non-English languages from Mr.TyDi demonstrate that CA-GAR outperforms existing methods.

1 Introduction

Information retrieval (IR) has become a critical component of natural language processing (NLP). The field has evolved from traditional sparse retrieval methods based on lexicon, such as TF-IDF and BM25 (Robertson and Zaragoza, 2009), to dense retrieval (Lee et al., 2019; Karpukhin et al., 2020) approaches powered by deep learning and pre-trained models like BERT (Devlin et al., 2019).

The advent of Large Language Models (LLMs) has significantly enhanced information retrieval by leveraging their advanced natural language understanding and generalization capabilities (Brown et al., 2020; Gao et al., 2023b; Zhu et al., 2023). Current research in this domain primarily revolves around two paradigms: (1) Generation-Augmented Retrieval (GAR), which enhances retrieval performance by utilizing LLMs' generative abilities to refine queries (Mao et al., 2021; Gao et al., 2023a; Wang et al., 2023; Jagerman et al., 2023); and (2) Generative Document Retrieval (GDR), which employs LLMs to generate document identifiers directly through constrained decoding, thereby encoding corpus-specific information into the retrieval process (Cao et al., 2021; Tay et al., 2022). However, both approaches present challenges. In GAR, a common strategy involves generating auxiliary text (e.g., query rewriting (Wang et al., 2023; Jagerman et al., 2023; Shen et al., 2024) or hypothetical document embeddings such as HyDE (Gao et al., 2023a)) to assist retrieval. However, the generated content may not always align well with the characteristics of the target document corpus, leading to suboptimal retrieval performance. As illustrated in Figure 1, while the language model can generate both *Passage 1* and *Passage 2* as plausible outputs, *Passage 2* is better suited for retrieval within the given document corpus. Although GDR ensures that the generated content exists within the document corpus through constrained decoding, effectively avoiding the retrieval mismatches observed in GAR, this approach inherently limits the



Prompt: write a passage to answer the question:

Question: what is patricia cornwell's most recent novel?

Patricia Cornwell's most recent novel in the Kay Scarpetta series is "Chaos," released in 2016. It is the 24th installment in the series.

Patricia Cornwell's latest book is part of the Kay Scarpetta series ,called Chaos, published in 2016. It is the 24th installment in the Kay Scarpetta series.

Patricia Cornwell's latest book is part of the Kay Scarpetta series, called Chaos, published in 2016. It is the 24th book in the Kay Scarpetta series.

Figure 1: A figure illustrating how GAR's generated text may mismatch with the target document corpus.

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 5836–5849
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

capacity of language models. (Li et al., 2024b).

Based on the observations above, our research question is: **How can we ensure that content generated by LLMs is well-suited for retrieval within a target document corpus while fully leveraging the generative capabilities of LLMs?** In retrieval tasks, the number of documents in the corpus is typically vast, far exceeding the context window that LLMs can process. Consequently, directly providing the entire document corpus as input to the model is infeasible. This presents a key challenge: **How can we effectively integrate corpus information into LLMs to enhance their retrieval performance?**

To address this challenge, we introduce a novel approach called **C**ontext-**A**ware **G**eneration-**A**ugmented **R**etrieval (**CA-GAR**), which incorporates corpus information into the generation process of LLMs. Specifically, at the core of our approach is the optimization of the model's autoregressive generation process by leveraging relevant document information from the corpus to influence token selection. To achieve this, we propose a **Distribution Alignment Strategy**, which utilizes a lexicon-based method to extract corpus information. This strategy approximates the optimization of the model's autoregressive generation process, ensuring that the generated content is better aligned with the target document corpus.

In summary, our contributions are as follows:

- We introduce a new approach called CA-GAR, which effectively combines the generative capabilities of LLMs with contextual information from the target document corpus.
- We propose a Distribution Alignment Strategy that utilizes a lexicon-based method to extract information from the corpus, optimizing autoregressive generation for improved alignment with the target document corpus.
- Our method outperforms existing approaches in retrieval tasks, as shown by experiments on seven BEIR benchmark tasks and four non-English languages from Mr. TyDi.

2 Related Work

Docuemnt Retrieval Information retrieval encompasses multiple domains and tasks (Liu et al., 2021; Zhuang et al., 2022; Sun et al., 2024). Our work primarily focuses on document retrieval. Early approaches were dominated by traditional lexicon-based methods, such as TF-IDF and BM25. These methods were later extended by sparse retrieval techniques that integrate neural networks with BM25, such as DeepCT (Dai and Callan, 2020) and docT5query (Nogueira et al., 2019). Recently, with the rise of pre-trained language models, dense retrieval (Lee et al., 2019; Karpukhin et al., 2020) has emerged as a promising alternative. By capturing semantic representations of text, dense retrieval effectively addresses the semantic mismatch problem inherent in sparse retrieval methods. Researchers have made significant advancements in this area through various techniques, including negative sample mining (Xiong et al., 2021; Qu et al., 2021), knowledge distillation (Qu et al., 2021; Lin et al., 2021; Ren et al., 2021; Liao et al., 2024), loss function optimization (Liao et al., 2024), and multi-vector representations (Zhang et al., 2022; Kang et al., 2025). However, dense retrieval still faces a critical challenge: achieving optimal performance often requires large-scale supervised training on domain-specific datasets, which limits its ability to generalize performance advantages across datasets from different domains (Thakur et al., 2021).

Large Language Models Large Language Models (LLMs) have demonstrated remarkable generalization capabilities, bringing significant transformations to the field of natural language processing (Brown et al., 2020; Ouyang et al., 2022; Hoffmann et al., 2022; Yuan et al., 2024). In the area of information retrieval, researchers have leveraged the instruction-following and in-context learning abilities of LLMs (Min et al., 2022; Sanh et al., 2022; Wei et al., 2022) to effectively address retrieval tasks by simply providing task definitions and a few retrieval examples (Li et al., 2024a). Building on this foundation, further studies have explored the use of LLMs for query enhancement, including query rewriting (Shen et al., 2024) and the expansion of relevant document information (Gao et al., 2023a). Notably, these approaches do not require additional training. Moreover, researchers have proposed parameter-efficient fine-tuning techniques to adapt portions of an LLM's parameters, enabling the model to function directly as an embedding model for retrieval tasks (Wang et al., 2024; BehnamGhader et al., 2024; Lee et al., 2024). This approach significantly reduces the computational and training resources required while preserving the LLMs' powerful natural language understanding capabilities.

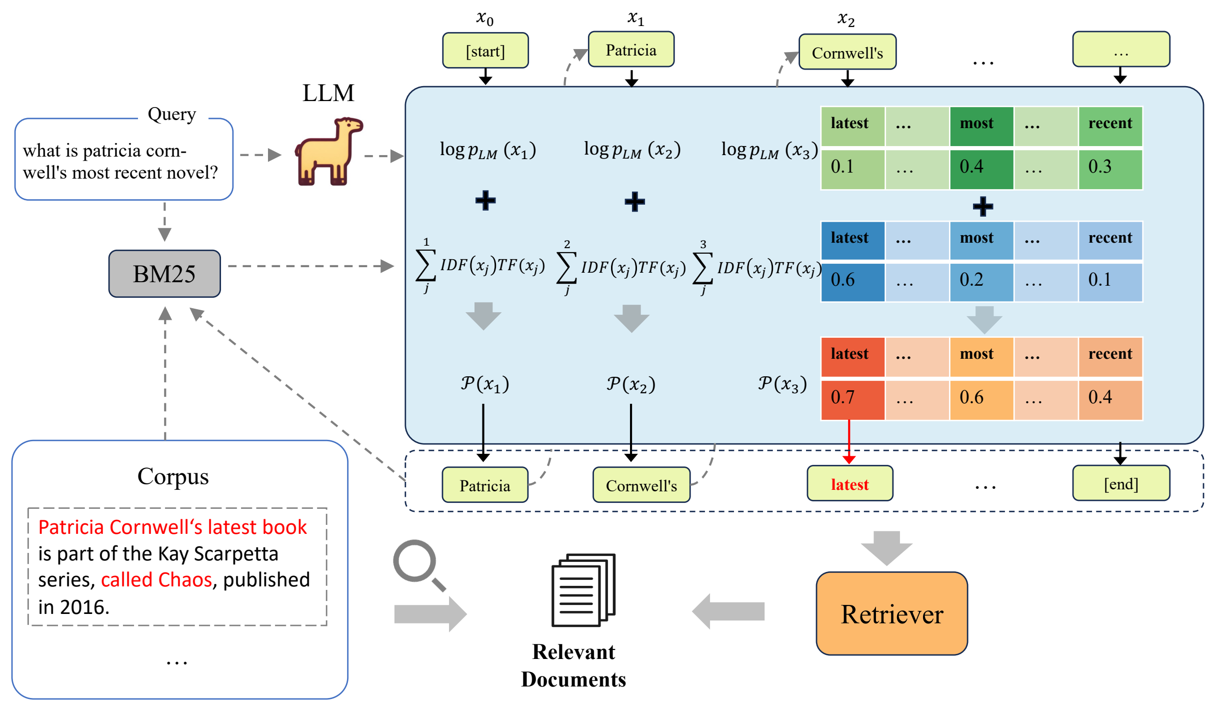


Figure 2: A figure illustrating our CA-GAR method.

Generative Retrieval Generative retrieval leverages the generative capabilities of models to directly facilitate document retrieval. Early approaches focused on enabling generative models to produce document identifiers through constrained decoding, which were then mapped to relevant documents (Cao et al., 2021; Tay et al., 2022). Subsequent research expanded on this paradigm by exploring techniques such as query augmentation and the design of more effective identifiers (Yang et al., 2023; Sun et al., 2023). A key limitation of these methods is their reliance on task-specific training over designated datasets. With the advent of LLMs, researchers have begun investigating their use as generative models for document retrieval. Some studies propose directly employing LLMs to generate pseudo-documents (Gao et al., 2023a; Mackie et al., 2023; Wang et al., 2023), which are treated as queries in dense retrieval frameworks to locate the final relevant documents. This approach bypasses the need for additional training, instead relying on the generalization capabilities of LLMs. However, LLMs often exhibit content bias, leading to pseudo-documents that poorly align with the target corpus and hinder retrieval performance. To address these challenges, our method focuses on improving the alignment of LLM-generated pseudo-documents with the target document collection. This approach aims to facilitate more accurate and effective downstream document retrieval, improving the overall performance of information retrieval systems.

3 Methodology

In this section, we first introduce the task definition of GAR. Then, we explain how our proposed CA-GAR method is designed.

3.1 Preliminaries

The GAR task consists of two primary steps. First, a large language model (LLM) is prompted with a specific instruction INSTRUCT to generate a text x based on a given query q:

$$x = \text{LLM}(\text{INSTRUCT}, q). \tag{1}$$

Second, the generated text x is utilized as input for a similarity function $\texttt{Sim}$, which measures the similarity between x and each document d. The most similar document d^* from the target corpus D is then identified as:

$$d^* = \arg\max_{d \in D} \texttt{Sim}(x, d). \tag{2}$$

A key limitation of this approach lies in the potential misalignment between the generated text x and the distribution of the target corpus. This misalignment is often caused by the content bias and stochasticity inherent in large language models, which can result in x being poorly aligned with the characteristics of the target corpus. As a result, the retrieval performance may be negatively affected. Therefore, a critical challenge is to ensure that the generated text x is more closely aligned with the underlying distribution of the target corpus to enhance retrieval effectiveness.

3.2 Context-Aware GAR

Optimization Objective To fully leverage the generative capabilities of LLMs and ensure that the generated text x based on a given query q aligns more closely with the target document, thereby improving its effectiveness in retrieval tasks, we introduce the following optimization objective:

$$\max \mathcal{P}(x) = \sum_{i=1}^{n} \log p_{LM}(x_i) + \beta \cdot \texttt{Sim}(x, d^*), \tag{3}$$

where $\sum_{i=1}^{n} \log p_{LM}(x_i)$ represents the likelihood of the token x_i generated by the LLM, expressed as the sum of the logarithmic probabilities of each individual token, β serves as a tuning factor, controlling the balance between different components of the objective function and $\texttt{Sim}(x, d^*)$ quantifies the degree of similarity or alignment between x and d^*, providing a measure of their relationship.

By integrating these components, the optimization objective encourages the generation of text that is not only coherent and likely under the LLM but also contextually aligned with the target document, thereby improving its effectiveness in downstream retrieval tasks.

Distribution Alignment Strategy However, the document with the highest similarity score is not necessarily the actual ground truth. Therefore, we take multiple candidate documents into consideration. In document retrieval, after measuring the similarity between a query q and a document d using a similarity function $\texttt{Sim}$, we typically obtain the top-k ranked documents, denoted as D_k. This set consists of the k documents with the highest similarity scores relative to q. We define F_k as follows:

$$F_k(x, D) = \frac{1}{k} \sum_{d' \in D_k}^{k} \texttt{Sim}(x, d'), \tag{4}$$

where F_k is no longer a document set but a real-valued function. It represents the average similarity score between a generated text x and the top-k most relevant documents in D.

To more effectively align F_k with information at the x-level and ensure its consistency with the target document collection, we adopt the BM25 method, which demonstrates exceptional generalization ability in zero-shot scenarios. As a lexicon-based retrieval method, BM25 exhibits strong analytical interpretability in similarity computation, with its scoring mechanism explicitly decomposable into the product of inverse document frequency (IDF) and term frequency (TF). Furthermore, BM25 offers significant advantages in interpretability, enabling more precise capture of the characteristics of the target document collection, thereby enhancing retrieval effectiveness and reliability. Therefore, we define the $\texttt{Sim}$ as BM25, allowing Equation 3 to be reformulated as follows:

$$\begin{aligned} \max \mathcal{P}(x) = \sum_{i=1}^{n} \log p_{LM}(x_i) + \\ \beta \cdot \frac{1}{k} \sum_{d' \in D_k}^{k} \text{IDF}(x) \cdot \text{TF}_{d'}(x), \end{aligned} \tag{5}$$

where IDF represents the Inverse Document Frequency, which measures the importance of a term by evaluating how unique or rare it is across the document collection. TF represents the Term Frequency, which quantifies how often a term appears in a specific document.

Referring to the autoregressive decoding method of LLMs, we approximate Equation 5 using Equation 6 as follows:

$$\begin{aligned} \max \mathcal{P}(x_i | x_{<i}) = \log p_{LM}(x_i | x_{<i}) + \\ \beta \cdot \frac{1}{k} \sum_{d' \in D_{k,i}}^{k} \sum_{j=1}^{i} \text{IDF}(x_j) \cdot \text{TF}_{d'}(x_j), \end{aligned} \tag{6}$$

where $D_{k,i}$ represents the top-k documents in D that are most relevant to $x_{0,1,\ldots,i}$. This dynamic selection mechanism allows for an adaptive adjustment of the generation process, ensuring that the most pertinent information is incorporated at each step.

Document Retrieval Based on the generated content x' and the original query q, we construct a new query q' by concatenation:

$$q' = q + x'. \tag{7}$$

We then employ either sparse or dense retrieval methods to retrieve relevant documents.

For sparse retrieval, we utilize the BM25 algorithm, which ranks documents D based on their relevance to the query q' by incorporating TF and IDF. Alternatively, for dense retrieval, we encode both the query q' and document d into vector representations $\mathbf{v_{q'}}$ and $\mathbf{v_d}$ using neural encoders. The similarity score between a query and a document is then computed as the inner product of their respective vectors:

$$\langle E_q(q'), E_d(d)\rangle = \langle \mathbf{v_{q'}}, \mathbf{v_d}\rangle. \tag{8}$$

Regardless of whether sparse or dense retrieval is used, the final retrieval results consist of the top-k documents with the highest similarity scores to the query q'.

4 Experiments

In this section, we provide a detailed explanation of the implementation of CA-GAR and present the experimental results on datasets. Furthermore, we demonstrate the performance improvements achieved by our approach over both BM25 and dense retrieval models.

4.1 Experimental Setup

Datasets and metrics In our main experiments, following previous works (Gao et al., 2023a; Feng et al., 2024), we selected seven low-resource datasets from the BEIR (Thakur et al., 2021) benchmark, covering a diverse range of domains, including biomedical, finance, and scientific research. Additionally, these datasets span various retrieval tasks, such as argument retrieval, citation prediction, and fact-checking. Specifically, the selected datasets include Arguana (Wachsmuth et al., 2018), Scifact (Wadden et al., 2020), NFCorpus (Boteva et al., 2016), Scidocs (Cohan et al., 2020), FiQA (Maia et al., 2018), Trec-Covid (Voorhees et al., 2020), and Touché (Bondarenko et al., 2020). We use nDCG@10 as the evaluation metric.

To evaluate the effectiveness of our approach in a multilingual setting, we conduct experiments on non-English datasets by selecting four low-resource languages from the Mr.TyDi (Zhang et al., 2021) benchmark: Bengali, Swahili, Telugu, and Thai. For performance assessment, we employ nDCG@10 as the evaluation metric.

For different datasets, we employed tailored prompt instructions to generate more appropriate content. In subsequent comparative experiments, we consistently used the same prompt to ensure fairness and reliability. Detailed instructions can be found in the appendix A.

Implementation details In this study, we utilize LLaMA3-8B-Instruct (Dubey et al., 2024) as the large language model for content generation. For retrieval, we employ BM25 as the sparse retrieval model, implemented using the BM25S (Lù, 2024) library, which is both highly efficient and simple. For English datasets, we apply the corresponding English stemmer to enhance retrieval performance. However, for multilingual datasets, we adopt a unified approach by not applying stemming to ensure consistency across different languages. The dense retrieval models include Contriever (Izacard et al., 2021), Contriever-ft (fine-tuned on MS MARCO), and BGE (Xiao et al., 2024). For non-English datasets, we similarly adopt BM25 as the sparse retrieval model without applying a stemmer. The dense retrieval models for these datasets include mContriever and mContriever-ft (fine-tuned on MS MARCO). In our experiment, we configure the LLaMA3-8B-Instruct model with a temperature of 1.0, a top-p of 1.0, and a top-k of 50. The parameter k in F_k is set to 10 and the β parameter is selected from {0.25,0.5,0.75,1.0} based on the best performance, and it is set to 0.75. The BM25 parameters k_1 and b are configured as 1.5 and 0.75, respectively. All the experiments are conducted with the single A800 GPU with 80GB VRAM.

Baselines First, we consider the traditional term frequency-based sparse retrieval method, BM25, which has been widely recognized for its effectiveness in zero-shot scenarios, learned sparse retrieval method DeepCT (Dai and Callan, 2020) and docT5query (Nogueira et al., 2019). Next, we examine Contriever for English and mContriever for non-English languages, along with their fine-tuned versions on the MS MARCO dataset, referred to as Contriever-ft and mContriever-ft, respectively. Additionally, we incorporate several dense retrieval models, including DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2021), latent interaction-based ColBERT (Khattab and Zaharia, 2020) and ColBERTv2 (Santhanam et al., 2022). Furthermore, we include BGE, which has demonstrated strong performance on English datasets.

Furthermore, we compare our approach with HyDE. To ensure a fair comparison in terms of LLM selection, we use LLaMA3-8B-Instruct to generate hypothetical documents. Following the

Model	Avg	Arguana	Scifact	NFCorpus	Scidocs	FiQA	Trec-Covid	Touché
Baselines (*Prior Work*)								
DeepCT	30.2	30.9	63.0	30.1	12.4	19.1	40.6	15.6
docT5query	40.9	34.9	67.5	32.8	16.2	29.1	71.3	34.7
DPR	19.1	17.5	31.8	18.9	7.7	11.2	33.2	13.1
ANCE	35.3	41.5	50.7	23.7	12.2	29.5	65.4	24.0
ColBERT	36.4	23.3	67.1	30.5	14.5	31.7	67.7	20.2
ColBERTv2	42.9	46.3	69.3	33.8	15.4	35.6	73.8	26.3
BM25	40.1	39.7	66.5	32.5	15.8	23.6	65.6	36.7
Contriever	32.8	37.9	64.9	31.7	13.7	24.5	36.3	20.8
Contriever-ft	40.3	44.6	67.7	32.8	16.5	32.9	59.6	27.8
BGE	48.6	63.5	74.6	**38.1**	22.6	44.3	72.3	24.8
HyDE$_{\textit{llama3-8b}}$								
w/ Contriever	34.6	34.1	63.9	31.4	13.8	26.5	46.3	26.4
w/ BGE	48.7	66.0	74.4	36.9	20.1	41.0	**75.6**	26.8
CA-GAR$_{\textit{llama3-8b}}$ (*Our Work*)								
w/ BM25	$42.2^{\uparrow 2.1}$	40.5	67.7	33.1	15.4	23.7	66.4	**48.3**
w/ Contriever	$34.9^{\uparrow 2.1}$	38.1	64.5	30.8	14.1	24.9	45.8	26.5
w/ Contriever-ft	$41.8^{\uparrow 1.5}$	45.2	67.8	33.2	17.2	33.2	60.3	35.9
w/ BGE	$\mathbf{50.0}^{\uparrow 1.4}$	**69.3**	**75.5**	35.4	**22.8**	**44.4**	72.1	30.2

Table 1: **Low-resource retrieval performance on a selection BEIR tasks** (measured by nDCG@10).

Model	Avg	bn	sw	te	th
Baselines (*Prior Work*)					
BM25	30.3	19.1	48.8	14.9	38.4
mContriever	19.5	21.3	23.9	10.7	21.9
HyDE$_{\textit{llama3-8b}}$	24.1	32.4	24.8	14.2	25.1
mContriever-ft	48.2	46.8	57.6	44.2	44.0
CA-GAR$_{\textit{llama3-8b}}$ (*Our Work*)					
w/ BM25	$34.9^{\uparrow 4.6}$	25.3	50.8	20.7	42.9
w/ mContriever	$24.7^{\uparrow 5.2}$	32.9	25.4	14.8	25.6
w/ mContriever-ft	$\mathbf{59.2}^{\uparrow 11.0}$	**55.5**	**62.1**	**64.9**	**54.2**

Table 2: **Multi-lingual retrieval performance on a selection Mr.TyDi languages** (measured by nDCG@10).

HyDE methodology, we employ Contriever and mContriever as the retrieval model, enabling a comparison between the approaches.

4.2 Low-Resource Retrieval

In Table 1, we present the performance of CA-GAR across seven tasks selected from the BEIR dataset. The results indicate that, on average, CA-GAR achieves a noticeable improvement in nDCG@10 compared to each baseline. In particular, the improvement over BM25 is especially pronounced. This can likely be attributed to our approach of leveraging BM25 to retrieve relevant documents during the decoding process of the large language model (LLM). Furthermore, compared to HyDE, CA-GAR achieves a certain degree of improvement when using Contriever as the retriever.

Our method achieves notable improvements on the Touché dataset, likely due to BM25's superior performance compared to other retrievers, including Contriever, Contriever-ft, and BGE. Since BM25 significantly influences LLM-generated responses, its strong baseline on Touché contributes to the observed performance gains. In contrast, improvements on FiQA and Scidocs are modest, primarily because BM25 performs poorly on these datasets, limiting its impact on LLM's autoregressive decoding. Additionally, these datasets involve complex domains (e.g., finance and scientific citation prediction), where LLM-generated responses tend to be of lower quality. A similar trend is observed for NFCorpus, which focuses on biomedical information retrieval. In such cases, more advanced prompting strategies or LLM architectures may be

Model	Avg	Arguana	Scifact	NFCorpus	Scidocs	FiQA	Trec-Covid	Touché
BM25								
CA-GAR$_{llama3\text{-}8b}$	42.2	40.5	67.7	33.1	15.4	23.7	66.4	48.3
w/o CA	$41.6^{\downarrow 0.6}$	39.8	66.9	32.1	14.8	23.5	66.9	47.3
Contriever								
CA-GAR$_{llama3\text{-}8b}$	34.9	38.1	64.5	30.8	14.1	24.9	45.8	26.5
w/o CA	$34.2^{\downarrow 0.7}$	37.6	64.3	29.7	13.1	23.9	45.5	25.3
Contriever-ft								
CA-GAR$_{llama3\text{-}8b}$	41.8	45.2	67.8	33.2	17.2	33.2	60.3	35.9
w/o CA	$40.9^{\downarrow 0.9}$	44.5	64.8	32.9	16.3	31.9	59.8	35.8
BGE								
CA-GAR$_{llama3\text{-}8b}$	50.0	69.3	75.5	35.4	22.8	44.4	72.1	30.2
w/o CA	$49.0^{\downarrow 1.0}$	68.7	73.8	35.3	21.7	43.9	70.5	29.3

Table 3: **Ablation study on a selection BEIR tasks** (measured by nDCG@10).

Model	Avg	bn	sw	te	th
BM25					
CA-GAR$_{llama3\text{-}8b}$	34.9	25.3	50.8	20.7	42.9
w/o CA	$34.2^{\downarrow 0.7}$	24.9	50.1	19.8	42.1
mContriever					
CA-GAR$_{llama3\text{-}8b}$	24.7	32.9	25.4	14.8	25.6
w/o CA	$23.9^{\downarrow 0.8}$	32.2	24.5	14.0	24.7
mContriever-ft					
CA-GAR$_{llama3\text{-}8b}$	59.2	55.5	62.1	64.9	54.2
w/o CA	$58.2^{\downarrow 1.0}$	54.9	61.2	64.1	52.7

Table 4: **Ablation study on a selection Mr.TyDi languages** (measured by nDCG@10).

required for further improvements.

4.3 Multi-Lingual Retrieval

In Table 2, we present the performance results on four low-resource languages selected from the Mr.TyDi dataset. Our findings indicate that CA-GAR consistently outperforms the baseline across four languages. The performance improvement can be attributed not only to the robust multilingual retrieval capabilities of BM25, which enhance the final content generated by the LLM, but also to the inherently strong multilingual proficiency of LLaMA 3, which further contributes to the observed performance gains.

Additionally, for the stronger retrieval model, mContriever-ft, the generated content tends to be richer in information. As a result, more powerful models are likely to achieve greater performance improvements, as they are less affected by biases introduced by extraneous information. Compared to HyDE, CA-GAR with Contriever also demonstrates a noticeable improvement, suggesting that BM25 plays an effective role in influencing the LLM's autoregressive decoding process.

4.4 Ablation Study

To further validate the effectiveness of our CA-GAR approach, we conducted an ablation study on selected datasets from BEIR and Mr.TyDi. The detailed experimental results are presented in Table 3 and 4. Specifically, we compared CA-GAR with a variant that does not employ the context-aware mechanism, denoted as w/o CA. In this baseline, the LLM generates content in a straightforward manner, without incorporating context-aware strategies to guide the autoregressive decoding process. This comparison allows us to assess the impact of context-aware strategies on the model's overall performance and the quality of the generated outputs.

Our findings indicate that, compared to CA-GAR, w/o CA exhibits a noticeable decline in overall performance. This performance degradation is particularly pronounced for relatively stronger models such as Contriever-ft, mContriever-ft and BGE. The results suggest that when the generated content exhibits certain biases or misalignments, the adverse effects are more pronounced in these high-performing models. This observation further highlights the critical role of CA-GAR in guiding and refining the generated content.

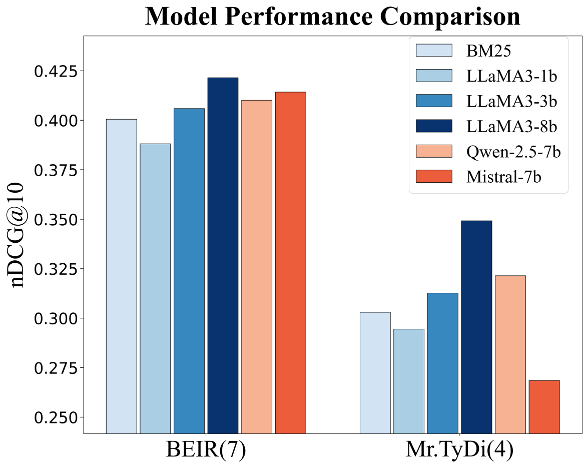


Figure 3: A figure presenting the retrieval performance of different models using BM25 across seven tasks from the BEIR(7) dataset and four languages from the Mr.TyDi(4) dataset.

5 Analysis

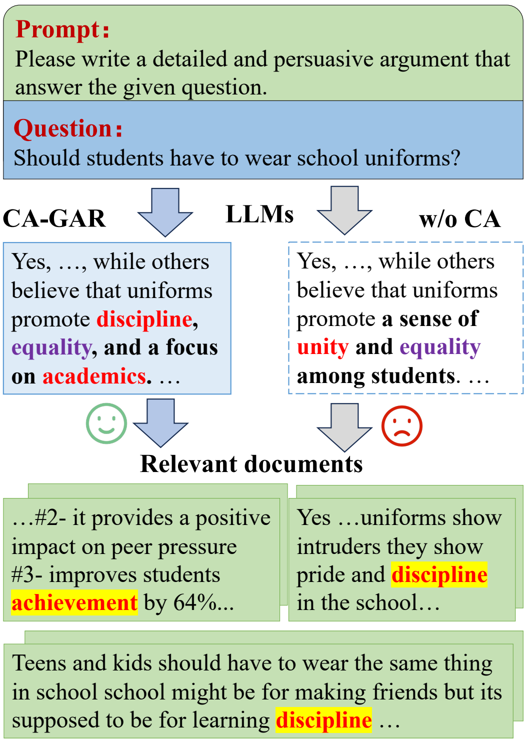
5.1 LLM Base Model Comparsion

To investigate the impact of model size, we compare three LLaMA3-Instruct variants with 1B, 3B, and 8B parameters. Figure 3 shows their performance when integrated with BM25.

Our results demonstrate that larger models generally achieve superior performance due to their enhanced generative capabilities. In contrast, smaller models, the 1B variant, often introduce substantial deviations in generated content, adversely affecting retrieval effectiveness. The 3B model performs on par with the baseline, exhibiting a slight improvement. These findings highlight the critical role of generative model quality in retrieval performance, as lower-quality models may introduce biases or hallucinated content that degrade results.

Additionally, we selected other LLMs of comparable scale, Qwen-2.5-7B (Yang et al., 2024) and Mistral-7B (Jiang et al., 2023), and conducted experiments on the dataset. It can be observed that among the seven datasets selected from BEIR, LLaMA3-8B-Instruct achieves the highest average performance, while LLaMA3-8B-Instruct and Mistral-7B perform similarly, with LLaMA3-8B-Instruct being slightly better. In contrast, Qwen-2.5-7B performs relatively poorly.

Furthermore, in multilingual settings, LLaMA3-8B-Instruct's superior capabilities enhance retrieval performance, while Mistral-7B's weaker multilingual abilities lead to greater deviations, resulting in lower performance than BM25.



Please write a detailed and persuasive argument that answer the given question.

Should students have to wear school uniforms?

Yes, …, while others believe that uniforms promote **discipline, equality, and a focus on academics.** …

Yes, …, while others believe that uniforms promote **a sense of unity and equality among students.** …

…#2- it provides a positive impact on peer pressure #3- improves students **achievement** by 64%...

Yes …uniforms show intruders they show pride and **discipline** in the school…

Teens and kids should have to wear the same thing in school school might be for making friends but its supposed to be for learning **discipline** …

Figure 4: Case Study: An example query from Touché – A comparison of content generated by CA-GAR and w/o CA.

5.2 Case Study

To further demonstrate the effectiveness of our approach, we selected a specific question from the Touché dataset as a case study. As illustrated in Figure 4, when responding to the same prompt for the question *"Should students have to wear school uniforms?"*, both our method, CA-GAR, and w/o CA generate relevant responses that address the question and included the keyword *"equal"*. However, there are notable differences in the choice of other words. CA-GAR incorporated *"discipline"* and *"academics"*, whereas w/o CA used *"unity"* .

From a generative standpoint, both outputs are acceptable. However, for subsequent retrieval processes, the content generated by CA-GAR proves to be superior. This is because it includes semantically related terms such as *"academics"* and *"achievement"*, as well as the identical term *"discipline"*, which are more aligned with the relevant documents in the corpus. This indicates that our context-aware approach successfully influences the autoregressive decoding process of large language models, resulting in generated content that is more aligned with the target retrieval corpus.

β	Arguana	Scifact	NFCorpus	Scidocs	FiQA	Trec-Covid	Touché	Avg
0.25	39.88	67.43	33.10	15.39	23.72	66.54	47.69	41.96
0.5	40.02	67.45	33.14	15.38	23.62	66.28	48.29	42.03
0.75	40.50	67.70	33.14	15.38	23.69	66.43	48.29	**42.16**
1.0	39.70	67.64	33.14	15.38	23.62	66.43	48.19	42.01

Table 5: Performance results for different values of β on a selection BEIR tasks.

5.3 Hyperparameter Analysis

In our study, we address the integration of log probabilities and BM25 scores by introducing a scalar weighting factor, denoted as β. This factor is pivotal in modulating the influence of BM25-derived guidance relative to the model's intrinsic generation probabilities. To identify the optimal value of β, we conducted a series of experiments across multiple datasets, systematically varying β to assess its impact on model performance.

The results are summarized in Table 5. From the results, it is evident that the value of $\beta = 0.75$ yields the highest average performance across all datasets. This suggests that a balanced approach, where the influence of BM25 scores is moderately weighted, enhances the model's effectiveness in generating relevant outputs.

6 Conclusion

This paper introduces CA-GAR, a novel generative augmentation retrieval method that optimizes LLMs generation using a distribution alignment strategy. By leveraging a lexicon-based approach with BM25, CA-GAR ensures better alignment between generated content and target documents. Experiments on low-resource and multilingual datasets validate its effectiveness, with advanced LLMs further enhancing performance. A case study illustrates CA-GAR's impact on content generation, highlighting its ability to maximize LLMs' generative potential.

Limitations

Our approach relies on LLMs to generate content while incorporating context-awareness to influence the autoregressive decoding process during generation. However, this method may not be well-suited for scenarios requiring low latency. Nevertheless, advancements in hardware and the development of optimization algorithms for model inference are expected to significantly reduce the computational cost and latency of content generation. Moreover, future research could explore alternative approaches to influencing the autoregressive decoding process of LLMs, beyond the use of BM25. This opens the possibility for more generalized and adaptable methods to enhance generation quality and relevance.

Acknowledgments

This research was supported by grants from the National Key Research and Development Program of China (Grant No.2024YFC3308200), the National Natural Science Foundation of China (62337001), the Key Technologies R & D Program of Anhui Province (No.202423k09020039) and the Fundamental Research Funds for the Central Universities.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *CoRR*, abs/2404.05961.

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. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1533–1536. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9571–9583, Bangkok, Thailand. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio 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.

Junfeng Kang, Rui Li, Qi Liu, Zhenya Huang, Zheng Zhang, Yanjiang Chen, Linbo Zhu, and Yu Su. 2025. Distribution-driven dense retrieval: Modeling many-to-one query-document relationship. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 11933–11941. AAAI Press.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. Making text embedders few-shot learners. *CoRR*, abs/2409.15700.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024b. From matching to generation: A survey on generative information retrieval. *CoRR*, abs/2404.14851.

Zihan Liao, Hang Yu, Jianguo Li, Jun Wang, and Wei Zhang. 2024. D2LLM: Decomposed and distilled large language models for semantic search. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14798–14814, Bangkok, Thailand. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online. Association for Computational Linguistics.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021. EKT: exercise-aware knowledge tracing for student performance prediction. *IEEE Trans. Knowl. Data Eng.*, 33(1):100–115.

Xing Han Lù. 2024. BM25S: orders of magnitude faster lexical search via eager sparse scoring. *CoRR*, abs/2407.03618.

Iain Mackie, Ivan Sekulic, Shubham Chatterjee, Jeffrey Dalton, and Fabio Crestani. 2023. GRM: generative relevance modeling using relevance-aware sample estimation for document retrieval. *CoRR*, abs/2306.09938.

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.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *CoRR*, abs/1904.08375.

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 *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*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Victor Sanh, Albert Webson, Colin Raffel, Stephen 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. Multitask 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.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946, Bangkok, Thailand. Association for Computational Linguistics.

Ruijun Sun, Hanqin Tao, Yanmin Chen, and Qi Liu. 2024. HACAN: a hierarchical answer-aware and context-aware network for question generation. *Frontiers Comput. Sci.*, 18(5).

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. 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*.

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.

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 (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. 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)*, pages 7534–7550, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. 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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pages 641–649. ACM.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Auto search indexer for end-to-end document retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6955–6970, Singapore. Association for Computational Linguistics.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Binbin Jin, Haoyang Bi, Enhong Chen, and Shijin Wang. 2022. A robust computerized adaptive testing approach in educational question retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 416–426. ACM.

A Instructions

Prompts for Arguana

Please write a counter argument for the passage.
Passage: {PASSAGE}
Counter Argument:

Prompts for Scifact

Please write a scientific paper passage to support or refute the claim.
Claim: {CLAIM}
Passage:

Prompts for NFCorpus

Please write a medically accurate passage to answer the question.
Question: {QUESTION}
Passage:

Prompts for Scidocs

Please write a scientific paper abstract that is cited by the given scientific paper title.
Paper title: {PAPER TITLE}
Paper abstract:

Prompts for FiQA

Please write a financial article passage to answer the question.
Question: {QUESTION}
Passage:

Prompts for Trec-Covid

Please write a passage that answer the question on COVID-19.
Question: {QUESTION}
Passage:

Prompts for Touché

Please write a detailed and persuasive argument that answer the given question.
Question: {QUESTION}
Argument:

Prompts for Mr.TyDi

Please write a passage in {LANGUAGE} to answer the question in detail.
Question: {QUESTION}
Passage:

B Analysis for efficiency

Our method introduces a slight increase in latency due to the relevance-guided generation process. In our experiments on the Scifact dataset, which consists of a corpus of 5k documents and 300 queries, we utilized a single NVIDIA A800 GPU and employed the LLaMA-3-8B model for document generation. The latency per query for each method is summarized in Table 6.

Method	Latency
HyDE	2.059s
CA-GAR	2.933s

Table 6: Latency per query on Scifact

The observed latency overhead primarily arises from the incorporation of document-level signals during the generation process. When the corpus size increases, applying CA-GAR over the entire corpus can lead to substantial latency. However, in practical applications, it is feasible to first retrieve a small subset of documents relevant to each query, which can effectively guide the generation process.

To evaluate this strategy, we conducted experiments on the large-scale Touché dataset, which contains 382k documents. The results, summarized in Table 7:

Method	Latency	nDCG@10
HyDE	2.498s	26.8
CA-GAR		
w/ 1k subset	2.635s	29.5
w/ 3k subset	3.134s	29.8
w/ 5k subset	3.621s	29.9
w/ full corpus	79.923s	30.2

Table 7: Latency and nDCG@10 on Touché

The results indicate that utilizing a pre-filtered subset of documents (e.g., the top 1k or 3k documents ranked by BM25) significantly reduces latency while maintaining competitive performance metrics. This approach enhances the practicality and scalability of CA-GAR in real-world retrieval systems. In conclusion, while there is a noted efficiency overhead associated with CA-GAR, this does not detract from its core contributions.