# Comparatively Assessing Large Language Models for Query Expansion in Information Retrieval via Zero-Shot and Chain-of-Thought Prompting

Daniele Rizzo, Alessandro Raganato[*,†] and Marco Viviani[†]

*Department of Informatics, Systems, and Communication (DISCo), University of Milano-Bicocca, Milan, Italy*

**Abstract**

In our research, we aim to assess the effectiveness of Large Language Models (LLMs) in performing query expansion in the context of Information Retrieval (IR). Some recent solutions proposed and studied in the literature to perform this task have proven effective considering specific LLMs, datasets, or prompt engineering techniques. In this paper, we intend to deepen this analysis with a more comprehensive and up-to-date view of their effectiveness, by comparing the results obtained from such solutions in the context of Zero-Shot (ZS) and Chain-of-Thought (CoT) learning, so as to be agnostic with respect to Few-Shot (FS) learning that requires additional training data from the dataset considered for evaluations, and using a variety of LLMs also of the latest generation. Results obtained across various LLMs generally demonstrate the superiority of utilizing recent LLM-based solutions for query expansion when employed in a prompt engineering scenario based on Zero-Shot learning. This showcases the intrinsic effectiveness of such recent LLMs even characterized by a modest number of parameters.

**Keywords**

Information Retrieval, Query Expansion, Large Language Models, Prompt Engineering, Natural Language Processing

## 1. Introduction

In an *Information Retrieval System* (IRS), the proper formulation of a *query* has a substantial impact on the effectiveness of the system in retrieving relevant search results. However, it is well known in the literature how *uncertainty* and *vagueness* are challenges encountered when formulating a query due to imprecise or ambiguous user input [1]. Users may be unsure about the exact terms or concepts they want to retrieve information on and may express their information needs in broad terms or use ambiguous language, making it challenging for the system to accurately interpret their intentions. In this way, both uncertainty and vagueness can lead to retrieval difficulties, as the system may struggle to understand and match the user's query with relevant documents.

To tackle these hurdles effectively, it is often necessary to reformulate the query, ensuring it aligns more closely with the user's information needs. This often involves employing a specialized strategy such as *query expansion*, which entails integrating additional related terms to encompass potential interpretations of the user's intent. Many approaches have been proposed over the years to address this tasks [2].

Nowadays, in particular, with the rapid progress of the so-called generative AI, *Large Language Models* (LLMs) have been effectively applied to the query expansion problem in IR through the application of *prompt engineering* techniques [3, 4]. In particular, the *Query-to-Document* (Q2D) model [4] capitalizes on the generative capabilities of LMMs to generate a pseudo-document from the original query, serving as a dependable reference instead of the conventional *Pseudo-Relevance Feedback* (PRF) documents for

*IIR 2024: The 14th Information Retrieval Workshop, September 05–06, 2024, Udine, Italy*

[*]Corresponding author.

[†]These authors contributed equally.

d.rizzo20@campus.unimib.it (D. Rizzo); alessandro.raganato@unimib.it (A. Raganato); marco.viviani@unimib.it (M. Viviani)

0000-0002-7018-7515 (A. Raganato); 0000-0002-2274-9050 (M. Viviani)

query expansion. Such a model was tested in the original paper in a *Few-Shot* (FS) learning setting [4], on a single LLM. PRF and Q2D models have been further compared in [3], where different prompts in *Zero-Shot* (ZS) and *Chain-of-Thought* (CoT) learning settings have been tested in addition to the FS learning setting, by employing a couple of open-source LLMs. The mentioned paper also proposes an approach denoted as Q2E that, unlike Q2D, does not consider the pseudo-document to expand the query, but only *keywords* generated by the considered LLMs.

Such solutions have proven their effectiveness with respect to specific and distinct LLMs, datasets, or prompt engineering techniques; however, it is our opinion that they need to be further tested comparatively against the use of different LLMs (both with a large and small number of parameters, both proprietary and open-source) and against prompt learning techniques that are agnostic to additional labeled data with respect to the datasets used for evaluations but rely only on the potential of the pre-trained models. Hence, in our study, we examine the capabilities of several recent LLMs ranging from 7 to 20 billion parameters, in both Q2D and Q2E operations, with respect to multiple IR subtask considering different datasets (and contexts), in both a ZS and a CoT setting.

The findings align closely with existing literature, affirming the efficacy of the Q2D approach for query expansion tasks using LLMs guided by prompt engineering techniques. A notable observation, especially when compared to prior comprehensive studies, is that contemporary LLM models consistently outperform their predecessors in Zero-Shot learning scenarios. This underscores the inherent effectiveness of these recent LLMs, even when operating with a relatively limited number of parameters.

The remainder of the article is structured as follows: Section 2 delves into previous literature pertaining to the task under consideration; Section 3 delineates the models and prompt engineering techniques employed with LLMs for executing the task; Section 4 expounds upon and deliberates the outcomes of the comparative evaluation; and finally, Section 6 encapsulates the study's findings and identifies avenues for future research.

## 2. Background and Related Work

*Query expansion* in *Information Retrieval* (IR) is a technique used to improve the effectiveness of search queries by automatically supplementing them with additional relevant terms [2]. The aim is to capture a broader range of relevant documents and increase the chances of retrieving the most relevant information, thus enhancing the performance of IRSs [5]. The operations involved in expanding a query can vary in nature and may or may not involve the user and their feedback in an iterative process. In this work, we exclusively focus on query expansion techniques with no user interaction.

At its core, query expansion enhances IRSs by broadening query terms into additional terms that convey the same concept or information need, thereby increasing the probability of a lexical match with documents in the corpus. Early research on query expansion predominantly centered on *Pseudo-Relevance Feedback* (PRF) [6, 7]. In PRF, the top-ranked documents are treated as *pseudo-relevant documents*, and terms that frequently occur in these documents but are not present in the original query are extracted. These additional terms are then added to the original query to refine and broaden its scope, aiming to retrieve more relevant documents in subsequent searches. PRF-based methods are particularly practical as they do not necessitate the construction of a domain-specific knowledge base and can be applied to any corpus. Recent advancements in query expansion have capitalized on neural networks to either generate or select expansion terms [8, 9], typically through model training or fine-tuning approaches.

It is in this scenario that the capabilities associated with pre-trained generative AI models open up intriguing possibilities for query expansion. Although the literature is still quite limited in this regard, some work is demonstrating the effectiveness of LLMs for the task of expanding queries using only *prompt engineering* techniques [10]. In this regard, there are actually three main techniques that are used nowadays [11], previously mentioned in the Introduction, which are worth briefly detailing in this section. Zero-Shot learning refers to the ability of a model to perform a task even when it has not been explicitly trained on examples of that task. In other words, the model is able to generalize from its

training data to perform new tasks it has not seen before. Few-Shot learning extends the idea of ZS learning by allowing the model to be further fine-tuned on a very small number of examples (a "few shots") for a given task. Finally, Chain-of-Thought learning emphasizes the ability of a language model to maintain context and coherence over longer passages of text. It is about the model's capability to follow and understand a chain of related ideas or thoughts within a conversation or text.

Not directly related to the query expansion task, but equally useful to discuss in this section as a starting point, is the approach presented in [12]. Here the authors propose *Hypothetical Document Embeddings* (HyDE) for dense retrieval, where ZS learning instructs an LLM (i.e., GPT-3, `text-davinci-003`) to generate a *hypothetical document d* from a query $q$, by employing the following prompt: "*Write a paragraph that answers the question*". The document $d$, encoded into an embedding vector, may capture relevance patterns but also contain non-relevant information or hallucinations. Hence, this vector is employed to identify a neighborhood in the corpus embedding space, where similar real documents are retrieved based on vector similarity. This second step ground the generated document to the actual corpus, with the encoder's dense bottleneck filtering out the incorrect or non-relevant information. A notable approach that has some similarities with the previous one but this time developed specifically to perform query expansion is *Query-to-Document* (Q2D) [4]. This solution is based on the generation of documents from an LLM (i.e., i.e., GPT-3, `text-davinci-003`) as a reliable proxy of the original queries, enhancing retrieval accuracy with no use of PRF. The approach, in particular, given a query $q$, employs FS learning to generate a *pseudo-document d* that is later employed to expand $q$. The prompt comprises the brief instruction: "*Write a passage that answers the given query:*" and $k$ labeled pairs randomly sampled from a training set (in the paper, $k = 4$). Subsequently, $q$ is expanded to a new query $q^+$ by concatenating $q$ with the pseudo-document $d$. The approach was tested against both sparse and dense retrieval and has proven to be effective w.r.t. to both of them and the considered baselines. However, in both cases, the solution relies on FS learning – it is therefore necessary to have training samples that can be used for the generation of the pseudo-document – and, however, results refers just to GPT-3.

Further elaborated in [3], the performance of the above-mentioned Q2D LLM-based query expansion technique has been assessed when utilizing the FLAN open-source model (specifically, `Flan-T5` and `Flan-UL2`) [13] to generate a hypothetical document from a query to operate a Q2D expansion, by using different prompt engineering techniques, such as ZS, FS, and CoT, and prompts. Additionally, a Q2E approach has been proposed, which is similar to the Query2Doc FS learning but with examples of query expansion terms instead of documents. Also its ZS and CoT versions have been tested in the paper. The results of this work illustrate that in general Q2D is superior to Q2E and that in each case performing prompt engineering by CoT gives outperforming results compared to ZS, for the LLM considered. The fact remains, despite the breadth of comparative evaluations against the different considered configurations, their effectiveness is only evaluated against the single FLAN model.

## 3. Methodology

As we have seen, the literature works presented above each have some drawback stemming either from the use of a single LLM or from the utilization of training data in the prompt engineering phase, which furthermore has also shown not to yield better results. Given that our objective is to conduct a comparative evaluation across multiple LLMs, and we do not consider any data referencing the datasets we will use for evaluations (thus neither FS learning nor PRF), in this approach, we explore the query expansion problem in both a ZS and a CoT scenario.

### 3.1. Query Expansion

Similar to what was done in [4] and [3], we consider the *template* for query expansion in a sparse retrieval scenario as follows:

$$q^+ = \text{concat}\left(\{q \times n\}, \text{prompt}_x\right)$$

where $q^+$ is the expanded query, $n$ is the number of times the original query $q$ is repeated,[1] and $\text{prompt}_x$ returns the expansion terms for $q$, which may consist of the pseudo-document in the case of using the Q2D model or a set of expansion keywords in the case of using the Q2E model, both of which were previously illustrated in Section 2. The $x$ symbol refers to the specific configuration used to generate pseudo-document or keywords, with respect to the prompt engineering technique considered. Thus, for example, $x = \mu/\phi$ indicates the use of the $\mu$ model driven by an LLM using $\phi$ learning for the generation of the pseudo-document for query expansion. The list of configurations used is shown in the next section.

### 3.2. Prompt Engineering

In the context of the Zero-Shot learning scenario, the prompts utilized for generating the pseudo-document (*i*) and the expansion keywords (*ii*) are as follows:

(*i*) *Write a passage that answers the following query:* [query]
(*ii*) *Write a list of keywords for the following query:* [query]

In the context of the Chain-of-Thought learning scenario, the prompt utilized for generating the pseudo-document (*iii*) is as follows:

(*iii*) *Answer the following query:* [query]
*Give the rationale before answering*

These prompts were executed by each of the LLMs detailed subsequently in the next section, concerning the various configurations outlined below:

- Q2D/ZS: the Zero-Shot version of the Q2D model [4] based on prompt (*i*);
- Q2E/ZS: the Zero-Shot version of the Q2E model [3] based on prompt (*ii*);
- Q2D/CoT: the Chain-of-Thought version of the Q2D model [4] based on prompt (*iii*).

## 4. Comparative Evaluation

Experimental comparative evaluation is conducted in this study with regard to the utilization of various LLMs, encompassing both proprietary and recent open-source models (Section 4.1). Evaluation is conducted on a variety of datasets, encompassing a range of tasks and domains (Section 4.2), utilizing standard metrics for assessing Information Retrieval Systems (IRSs) effectiveness (Section 4.3), and benchmarked against established baselines (Section 4.4). Subsequently, detailed results of this comparative evaluation are presented (Section 4.5). All experiments have been carried out within a sparse retrieval setting, using BM25 [14, 15] as implemented by *pyTerrier* [16] with its default parameters ($b = 0.75$, $k_1 = 1.2$, $k_3 = 8.0$).[2]

### 4.1. Large Language Models

A series of recent LLMs has been considered, which spans both open-source and proprietary models, ranging from 7B to 20B parameters. Specifically, we include:

- GPT-4 [17]: it is the well-known and proprietary large multimodal model (accepting image and text inputs, emitting text outputs) developed by OpenAI, updated with text up to June 13th 2023. In this paper, we use its `GPT-4-0613` version;[3]

[1] Since LLM output may be verbose, this is therefore a necessary ploy to preserve the importance of the terms of the original query in the expanded query. In this work we considered $n = 5$, as already done in previous works [3, 4].
[2] https://pyterrier.readthedocs.io/
[3] https://platform.openai.com/docs/models/overview

- Mistral 7B [18]: developed by Mistral AI, Mistral 7B is an open-source LLM that leverages *Grouped-Query Attention* (GQA) [19], and *Sliding Window Attention* (SWA) [20]. In this paper, we employ: (*i*) `Mistral-7B-Instruct-v0.1`,[4] i.e., the instruct fine-tuned 7B LLM, trained on a variety of publicly available English conversation datasets; and (*ii*) `Mistral-7B-Instruct-v0.2`,[5] i.e., the second iteration of the previous model, trained with a large context window, i.e. 32k tokens, mainly on English data;
- QWEN [21]: it is an open-source LLM series that encompasses distinct models with varying parameter counts. The model series include the base pre-trained language models and chat models fine-tuned with human alignment techniques, i.e., *Supervised Fine-Tuning* (SFT), *Reinforcement Learning with Human Feedback* (RLHF), etc. In this work we employed `Qwen1.5-7B-Chat`,[6] i.e., the instruct fine-tuned 7B multilingual model, supporting contexts up to 32K tokens;
- Meta Llama 3 [22]: it is a family of LLMs coming in two sizes – 8B and 70B parameters – in pre-trained and instruction tuned variants. In this work, we use `Meta-Llama-3-8B-Instruct`,[7] i.e., the recent instruction-tuned 8B model, released in April 2024, optimized for dialogue use cases and aligned with human preferences for helpfulness and safety;
- Gemma [23]: it is a family of open LLMs based on Google's Gemini models [24]. In this work we use `gemma-1.1-7b-it`,[8] i.e., the recent open-source instruction-tuned 7B English LLM, trained on a combination of diverse data sources, i.e., Web documents, code, and mathematics, totaling 6 trillion tokens.

Additionally, we include a comparison with the top-performing system from [3], i.e., the `Flan-UL2` model [25], which is equipped with 20B parameters.[9]

### 4.2. Datasets

The datasets used pertain to assess the effectiveness of LLMs for the considered task with respect to distinct subtasks of Information Retrieval. In particular, the considered datasets are MS MARCO [26], developed for Passage Retrieval, and a subset of those contained in BEIR [27, 28], a benchmark dataset for Zero-Shot evaluation of IR models across different domain/task combinations, encompassing Medical IR, Entity Retrieval, Fact Checking, etc.

1. MS MARCO [26]:[10] the *MicroSoft MAchine Reading COmprehension dataset* is a collection of datasets focused on deep learning in search. In this article, the employed dataset is that referred to *Passage Retrieval* (PR). Based on the passages and questions available in the *Question Answering* (QA) dataset,[11] a PR task is formulated. With a pool of 8.8 million passages, the aim is to rank them according to their relevance. Relevance labels are derived from passages marked as containing the answer in the QA dataset;
2. NFCorpus [29]:[12] it is an extensive English retrieval dataset tailored for *Biomedical Information Retrieval*. It encompasses 3,244 natural language queries, sourced from the NutritionFacts.org Website. Alongside these queries are 169,756 automatically extracted relevance judgments, pertaining to 9,964 medical documents. These documents, characterized by their terminology-rich language, primarily originate from *PubMed*;

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[6] https://huggingface.co/Qwen/Qwen1.5-7B-Chat
[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[8] https://huggingface.co/google/gemma-1.1-7b-it
[9] Comparison with FLAN is made only in terms of *Recall* (details on the use of this measure are provided in Section 4.3), the only measure used in [3] to perform experimental evaluations.
[10] https://microsoft.github.io/msmarco/
[11] https://microsoft.github.io/msmarco/#qna
[12] https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/

3. FiQA-2018: it is related to the *Financial Opinion Mining and Question Answering* evaluation lab.[13] Task 2, in particular, consists of opinion-based question-answering. The employed version include financial data crawled by *StackExchange* posts under the "Investment" topic from 2009-2017. A random selection of 500 and 648 queries is extracted from the original training split to serve as development and test sets, respectively;
4. Touché-2020 [30]:[14] Touché has been the first evaluation lab on *Argument Retrieval* that was held at CLEF 2020. In the lab, two tasks are conducted: (1) aiding individuals in locating arguments on socially significant subjects, and (2) assisting individuals with arguments concerning everyday personal choices. Task 1 is based on the *args.me* corpus [31], while Task 2 is based on the *ClueWeb12* document collection [32];
5. SciFact [33]:[15] the dataset is related to *Scientific Claim Verification*, a task involving the selection of abstracts from research literature that either support or refute a given scientific claim, along with the identification of rationales for each decision. The dataset is constituted by 1.4K expert-written scientific claims paired with annotated evidence-containing abstracts, including labels and rationales;
6. SciDocs [34]:[16] it contains a corpus of 30K held-out pool of scientific papers. These documents can be utilized for the *Citation Prediction* task, as suggested in the BEIR benchmark. In this task, the model aims to retrieve cited papers (output) for a given paper title (input);
7. DBpedia-Entity: the DBpedia-Entity collection [35] has been used as a standard test collection for *Entity Search* for many years. Here, the DBpedia-Entity v2 collection [36] is employed,[17] which uses a more recent DBpedia dump and a unified candidate result pool from the same set of retrieval models;
8. TREC-COVID [37]:[18] it is a test collection that captures the information needs of researchers in *Biomedical Information Retrieval* using the scientific literature during a pandemic. It employs the document set provided by CORD-19 [33];
9. NQ [38]:[19] the *Natural Questions* corpus is a *Question Answering* data set. Questions consist of real anonymized, aggregated queries issued to the Google search engine. The public release consists of 307,373 training examples with single annotations; 7,830 examples with 5-way annotations for development data; and a further 7,842 examples with 5-way annotated sequestered as test data;
10. Climate-Fever [39]:[20] it is a dataset for the *Fact Checking* of climate change-related claims. The dataset is formed by collecting 1,535 claims from the Web. For every claim, the top five relevant evidence candidate sentences are algorithmically retrieved from Wikipedia using natural language understanding (NLU). Subsequently, humans annotate each sentence as supporting, refuting, or insufficient to validate the claim. This database of 7,675 annotated claim-evidence pairs is referred to as the Climate-Fever dataset.

### 4.3. Evaluation Metrics

Some of the most commonly used evaluation metrics in Information Retrieval (IR) are employed to assess the effectiveness of the system in retrieving relevant results. Specifically, the following metrics are considered:

- *Recall* at 1K (Recall@1K): it computes the recall, or the proportion of relevant documents retrieved, at various cutoff points (in our case, at 1K, as in the previous literature works);

[13] https://sites.google.com/view/fiqa/home
[14] https://touche.Webis.de/data.html
[15] https://allenai.org/data/scifact
[16] https://github.com/allenai/scidocs
[17] http://tiny.cc/dbpedia-entity
[18] https://ir.nist.gov/trec-covid/
[19] https://ai.google.com/research/NaturalQuestions
[20] http://climatefever.ai/

- *Mean Reciprocal Rank* at 10 (MRR@10): it measures the average reciprocal rank of the top 10 retrieved relevant documents;
- *Normalized Discounted Cumulative Gain* at 10 (nDCG@10): it normalizes the DCG score by the ideal DCG score at 10, providing a measure of ranking quality considering both relevance and position of retrieved documents.

### 4.4. Baselines

Classical PRF-based query expansion methods have been considered as baselines. In particular, as illustrated in [40]:

- *Bose-Einstein* weighting (1) (BE1): this refers to a weighting scheme inspired by the *Bose-Einstein* (BE) statistics [41], often used in IR to assign weights to terms in a query-document context. It typically involves incorporating term frequency and document length normalization to improve retrieval effectiveness;
- *Bose-Einstein* weighting (2) (BE2): similar to BE1, Bose-Einstein weighting (2) is another variant of the weighting scheme inspired by Bose-Einstein statistics. It may involve different formulations or adjustments tailored to specific retrieval tasks or datasets;
- *Kullback-Leibler* weighting (KL): this weighting scheme utilizes the *Kullback-Leibler* (KL) divergence [42], a measure of dissimilarity between two probability distributions. In the context of IR, KL weighting is often employed to compute the similarity between the language model of the query and that of the documents in the collection, facilitating more effective retrieval by considering the relevance of documents based on their language models.

### 4.5. Results

In Tables 1, 2, and 3, the results of the comparative evaluation are presented, in terms of Recall@1k, MRR@10, and nDCG@10 respectively.

Based on previous studies, particularly [4] and [3], it has already been demonstrated that Q2D performs well compared to the query expansion task, and this is also demonstrated by our evaluations against the baselines considered. However, [3] showed that the FS learning scenario was not optimal for the Q2D solution, demonstrating a clear superiority of the CoT scenario. However, this article considered only the FLAN LLM.

Based on our experimental evaluation, it becomes evident that across nearly all datasets and a comprehensive range of LLMs – regardless of their parameter count, source availability (open-source or proprietary), and sophistication – contemporary pre-trained models exhibit remarkable effectiveness for our task, especially when applied in a Zero-Shot scenario. This underscores the robustness and adaptability of modern LLMs, highlighting their efficacy across diverse settings and configurations.

## 5. Data Availability

The datasets used in this work are publicly accessible, as indicated in Section 4.2. The data generated by the seven LLMs considered in the context of this research across the three prompt settings – namely Query-to-Document in Zero-Shot (Q2D/ZS), Query-to-Entity in Zero-Shot (Q2E/ZS), and Query-to-Document with Chain-of-Thought (Q2D/CoT) – are made publicly available at the following address: https://github.com/ikr3-lab/QueryExpansionLLMs.

## 6. Conclusion and Future Works

Our research delved into the effectiveness of Large Language Models (LLMs) for query expansion in Information Retrieval (IR) contexts, particularly focusing on the application of prompt engineering techniques. We conducted a comprehensive analysis, comparing the performance of recent LLM-based

| Dataset | BM25 | Classical QE | | | Flan-UL2-20B | | | Mistral-7B-v0.1 | | | Mistral-7B-v0.2 | | | Llama-3-8B | | | Gemma-7B | | | Qwen_7B | | | GPT-4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BE1 | BE2 | KL | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT |
| **1** | 73,62 | 75,11 | 75,28 | 74,47 | 75,73 | 73,79 | 79,58 | 79,97 | 73,19 | 80,57 | 85,97 | 79,28 | 82,33 | 85,39 | 81,76 | 83,27 | 79,05 | 79,17 | 77,6 | 81,63 | 78,4 | 81,5 | 84,99 | 78,09 | 86,41 |
| **2** | 36,06 | 54,38 | 55,66 | 54,61 | 59,81 | 44,12 | 52,63 | 60,54 | 59,63 | 55,6 | 61,07 | 59,58 | 59,61 | 60,49 | 58,73 | 60,22 | 59,45 | 59,31 | 57,87 | 60,67 | 58,92 | 60,42 | 60,79 | 58,49 | 58,29 |
| **3** | 77,42 | 83,41 | 81,84 | 83,17 | 78,26 | 77,31 | 80,08 | 81,93 | 76,99 | 81,21 | 83,41 | 81,84 | 83,17 | 83,25 | 81,5 | 82,58 | 82,03 | 82,21 | 83,22 | 84,03 | 80,89 | 83,83 | 83,82 | 79,85 | 83,91 |
| **4** | 85,05 | 86 | 86,47 | 86,11 | 83,44 | 85,02 | 85,51 | 85,78 | 83,68 | 84,76 | 84,62 | 84,51 | 84,86 | 84,71 | 84,6 | 84,94 | 84,43 | 85,06 | 84,08 | 84,04 | 83,1 | 84,28 | 84,58 | 85,05 | 84,7 |
| **5** | 97 | 97,67 | 97,67 | 97,33 | 97,57 | 97,17 | 97,57 | 99 | 98 | 98,33 | 99 | 99 | 99 | 99 | 99 | 98,67 | 98,33 | 98,67 | 99 | 98,67 | 99 | 99,33 | 99 | 98 | 98,33 |
| **6** | 55,04 | 57,47 | 58,3 | 56,62 | 59,78 | 57,7 | 58,51 | 59,46 | 58,52 | 58,36 | 60,98 | 60,56 | 60,32 | 60,42 | 60,42 | 60,68 | 60,21 | 60,43 | 60,19 | 61,04 | 60,54 | 61,11 | 60,73 | 59,66 | 59,52 |
| **7** | 63,61 | 64,9 | 64,53 | 64,3 | 65,47 | 63,92 | 65,77 | 69,17 | 64,15 | 68,3 | 71,75 | 68,08 | 71,28 | 71,51 | 69,39 | 70,6 | 69,26 | 68,32 | 69,17 | 72,09 | 69,01 | 71,04 | 70,85 | 64,98 | 70,2 |
| **8** | 36,77 | 38,45 | 40,66 | 38,49 | 38,05 | 43,12 | 43,43 | 40,32 | 37,62 | 40,66 | 42,14 | 42 | 41,61 | 42,9 | 44,12 | 42,3 | 41,9 | 43,1 | 43,02 | 42,6 | 43,56 | 42,39 | 41,21 | 42,54 | 41,69 |
| **9** | 78,96 | 81,09 | 80,55 | 80,32 | 84,71 | 79,11 | 85,46 | 89,42 | 81,36 | 87,51 | 92,05 | 88,43 | 91,1 | 92,8 | 89,24 | 92,38 | 88,09 | 85,26 | 87,29 | 90,97 | 88,09 | 89,87 | 91,55 | 80,81 | 90,45 |
| **10** | 57,63 | 60,22 | 60,29 | 59,31 | 47,66 | 46,44 | 47,42 | 69,41 | 65,26 | 67,86 | 71,28 | 68,73 | 70,48 | 70,87 | 68,91 | 69,26 | 69,76 | 66,77 | 67,44 | 72,59 | 71,1 | 71,91 | 69,73 | 60,73 | 68,11 |
| **AVG** | 66,12 | 69,87 | 70,13 | 69,47 | 69,05 | 66,77 | 69,60 | 73,50 | 69,84 | 72,32 | 75,23 | 73,20 | 74,38 | 75,13 | 73,77 | 74,49 | 73,25 | 72,83 | 72,89 | 74,83 | 73,26 | 74,57 | 74,73 | 70,82 | 74,16 |

Table 1: Comparative evaluations in terms of Recall@1K.

| Dataset | BM25 | Classical QE | | | Mistral-7B-v0.1 | | | Mistral-7B-v0.2 | | | Llama-3-8B | | | Gemma-7B | | | Qwen_7B | | | GPT-4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BE1 | BE2 | KL | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT |
| **1** | 79,44 | 78,75 | 80,01 | 79,01 | 88,18 | 81,78 | 86,59 | 90,54 | 88,95 | 86,98 | 90,31 | 85,61 | 90,45 | 80,04 | 79,85 | 84,3 | 82,87 | 86,05 | 84,88 | 95,81 | 81,67 | 96,12 |
| **2** | 53,44 | 52,74 | 52,57 | 52,59 | 54,88 | 53,95 | 53,77 | 57,04 | 56,86 | 57,49 | 56,42 | 55,55 | 57,04 | 56,84 | 55,3 | 55,93 | 57,31 | 57,16 | 57,39 | 57,49 | 55,19 | 56,23 |
| **3** | 31,03 | 29,37 | 28,2 | 30,22 | 34,03 | 29,25 | 31,74 | 34,12 | 31,22 | 33,78 | 34,24 | 31,58 | 34,23 | 34,34 | 33,1 | 34,7 | 32,85 | 30,72 | 31,47 | 35,96 | 32,67 | 36,98 |
| **4** | 62,28 | 63,54 | 65,52 | 64,46 | 72,35 | 68,07 | 74,08 | 68,15 | 73,3 | 72,03 | 77,86 | 75,12 | 76,41 | 73,1 | 72,07 | 75,24 | 74,52 | 68,79 | 75,58 | 77,86 | 65,71 | 77,68 |
| **5** | 63,24 | 60,42 | 59,87 | 62,85 | 65,06 | 63,5 | 65,36 | 65,51 | 65,16 | 66,51 | 65,86 | 65,09 | 64,2 | 65,52 | 64,94 | 65,26 | 64,46 | 64,79 | 64,15 | 65,8 | 63,85 | 66,09 |
| **6** | 25,37 | 25,28 | 24,96 | 25,49 | 25,61 | 24,33 | 26,24 | 27,07 | 25,91 | 27,17 | 26,67 | 25,82 | 27,21 | 26,47 | 25,57 | 26,9 | 26,3 | 25,79 | 26,37 | 27,06 | 26,05 | 26,83 |
| **7** | 51,7 | 50,47 | 48,71 | 49,47 | 62,87 | 50,65 | 59,15 | 63,39 | 59,15 | 63,5 | 62,9 | 59,37 | 64,5 | 58,8 | 56,4 | 60,58 | 62,3 | 57,36 | 62,11 | 65,95 | 55,49 | 62,59 |
| **8** | 81,72 | 83,54 | 84,72 | 82,4 | 80,34 | 74,4 | 84,22 | 78,97 | 85,98 | 78,12 | 84,56 | 87,6 | 79,95 | 81,19 | 86,25 | 86,47 | 80,01 | 86 | 82,12 | 77,12 | 82,23 | 81,78 |
| **9** | 19,99 | 20,3 | 19,88 | 20,01 | 29,28 | 22,35 | 27,98 | 34,59 | 29,1 | 33,84 | 36,29 | 29,03 | 35,19 | 28,78 | 24,05 | 28,15 | 31,34 | 27,41 | 30,59 | 37,12 | 21,4 | 36,2 |
| **10** | 17,07 | 18,08 | 18,11 | 18,22 | 24,78 | 20,91 | 24,55 | 27,36 | 23,67 | 26,67 | 26,96 | 22,96 | 25,43 | 25,49 | 22,2 | 23,48 | 28,46 | 24,59 | 27,1 | 25,23 | 19,03 | 25,1 |
| **AVG** | 48,53 | 48,25 | 48,26 | 48,47 | 53,74 | 48,92 | 53,37 | 54,67 | 53,93 | 54,61 | 56,21 | 53,77 | 55,46 | 53,06 | 51,97 | 54,10 | 54,04 | 52,87 | 54,18 | 56,54 | 50,33 | 56,56 |

Table 2: Comparative evaluations in terms of MRR@10.

| Dataset | BM25 | Classical QE | | | Mistral-7B-v0.1 | | | Mistral-7B-v0.2 | | | Llama-3-8B | | | Gemma-7B | | | Qwen_7B | | | GPT-4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BE1 | BE2 | KL | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT | Q2D/ZS | Q2E/ZS | Q2D/CoT |
| **1** | 47,95 | 50,86 | 51,45 | 50,56 | 58,46 | 52,78 | 55,16 | 63,32 | 56,31 | 57,3 | 63,11 | 57,93 | 59,76 | 55,65 | 52,08 | 53,25 | 58,54 | 55,14 | 56,48 | 65,08 | 54,22 | 67,65 |
| **2** | 32,22 | 33,49 | 33,36 | 33,1 | 34,33 | 33 | 33,52 | 34,93 | 35,45 | 34,95 | 34,65 | 35,14 | 34,55 | 34,81 | 34,73 | 34,36 | 35 | 34,69 | 34,96 | 35,43 | 34,47 | 34,02 |
| **3** | 25,26 | 24,36 | 23,57 | 24,86 | 27,67 | 24,14 | 26,01 | 28,7 | 25,36 | 27,69 | 28,68 | 25,75 | 28,11 | 28,11 | 27,37 | 28,66 | 27,69 | 25,5 | 26,01 | 29,71 | 26,45 | 30,17 |
| **4** | 34,28 | 35,62 | 37,78 | 35,63 | 42,54 | 39,42 | 41,73 | 38,87 | 41,53 | 41,42 | 44,15 | 42,79 | 42,91 | 41,11 | 40,11 | 41,04 | 42,08 | 41,41 | 44,22 | 44,67 | 38,05 | 43,91 |
| **5** | 67,22 | 65,15 | 64,66 | 67,16 | 68,8 | 67,26 | 69,31 | 69,82 | 69,54 | 70,57 | 70,08 | 69,43 | 68,71 | 69,59 | 69,1 | 69,06 | 68,82 | 68,51 | 68,43 | 70,06 | 68,02 | 70,42 |
| **6** | 14,71 | 15,1 | 14,85 | 15,05 | 14,77 | 14,11 | 14,87 | 15,77 | 15,05 | 15,67 | 15,39 | 14,95 | 15,53 | 15,25 | 14,95 | 15,33 | 15,17 | 15 | 15,18 | 15,61 | 15,08 | 15,37 |
| **7** | 26,59 | 26,59 | 26,11 | 25,76 | 32,6 | 26,01 | 31,41 | 33,82 | 30,89 | 33,84 | 33,24 | 31,77 | 35,29 | 30,59 | 29,43 | 31,34 | 33,21 | 29,64 | 32,79 | 34,84 | 29,03 | 34,23 |
| **8** | 57,61 | 58,56 | 58,94 | 58,16 | 57,55 | 55,61 | 59,91 | 62,21 | 63,89 | 61,65 | 65,16 | 65,37 | 62,16 | 59,39 | 67,28 | 66,84 | 62,7 | 66,16 | 64,91 | 60,72 | 62,43 | 62,78 |
| **9** | 23,09 | 23,54 | 23,29 | 23,23 | 33,07 | 25,73 | 31,67 | 38,51 | 32,67 | 37,61 | 40,23 | 32,82 | 39,1 | 32,6 | 27,7 | 32,11 | 35,05 | 31,27 | 34,54 | 40,88 | 24,52 | 39,89 |
| **10** | 12,52 | 13,45 | 13,56 | 13,54 | 18,94 | 15,88 | 18,21 | 20,37 | 18,18 | 19,78 | 20,35 | 17,21 | 19,2 | 18,94 | 16,5 | 17,58 | 21,61 | 18,72 | 20,67 | 19 | 14,24 | 18,7 |
| **AVG** | 34,15 | 34,67 | 34,76 | 34,71 | 38,87 | 35,39 | 38,18 | 40,63 | 38,89 | 40,05 | 41,50 | 39,32 | 40,53 | 38,60 | 37,93 | 38,96 | 39,99 | 38,60 | 39,82 | 41,60 | 36,65 | 41,71 |

Table 3: Comparative evaluations in terms of nDCG@10.

solutions across various learning scenarios, including Zero-Shot (ZS) and Chain-of-Thought (CoT) learning, while aiming to remain agnostic to Few-Shot (FS) learning that necessitates additional training samples (though limited in number). Our study builds upon prior research, which showcased the potential of LLMs, notably the Query-to-Document (Q2D) model, in improving query expansion tasks. By extending these investigations to encompass a broader spectrum of LLMs, datasets, and learning scenarios, we offer a more nuanced understanding of their efficacy. Our findings affirm the superiority of recent LLM-based solutions, particularly in Zero-Shot learning scenarios, underscoring their robustness and adaptability across diverse contexts.

Looking ahead, several avenues for future research emerge. A deeper exploration of the peculiarities of individual domains and the typical form of queries within those domains could be undertaken. Additionally, investigating the interplay between different prompt designs and LLM architectures could uncover synergies for further enhancing performance. Furthermore, examining the generalization capabilities of LLMs across various languages could extend the applicability of these models in real-world IR applications.

## Acknowledgments

We acknowledge the support of: the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU; CINECA under the ISCRA initiative, for the availability of high-performance computing resources;[21] CSC – IT Center for Science, Finland,[22] for the availability of high-performance computing resources; the Italian MUR under the PRIN 2022 Project KURAMi: "Knowledge-based, explainable User empowerment in Releasing private data and Assessing Misinformation in online environments" (20225WTRFN);[23] the University of Milano-Bicocca under the ATEQC 2024 Project PriQuaDeS: "Next-generation Privacy- and Quality-preserving Decentralized Social Web Applications".

## References

[1] H. R. Turtle, W. B. Croft, Uncertainty in information retrieval systems, in: Uncertainty management in information systems: from needs to solutions, Springer, 1997, pp. 189–224.

[2] H. K. Azad, A. Deepak, Query expansion techniques for information retrieval: a survey, Information Processing & Management 56 (2019) 1698–1735.

[3] R. Jagerman, H. Zhuang, Z. Qin, X. Wang, M. Bendersky, Query expansion by prompting large language models, 2023. arXiv:2305.03653.

[4] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9414–9423. URL: https://aclanthology.org/2023.emnlp-main.585. doi:10.18653/v1/2023.emnlp-main.585.

[5] V. Gupta, A. Dixit, Recent query reformulation approaches for information retrieval system-a survey, Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 16 (2023) 94–107.

[6] G. Amati, Probability models for information retrieval based on divergence from randomness (2003).

[7] J. J. Rocchio, Relevance feedback in information retrieval, in: G. Salton (Ed.), The Smart retrieval system - experiments in automatic document processing, Englewood Cliffs, NJ: Prentice-Hall, 1971, pp. 313–323.

[8] Z. Zheng, K. Hui, B. He, X. Han, L. Sun, A. Yates, Bert-qe: Contextualized query expansion for document re-ranking, 2020. arXiv:2009.07258.

[21] https://www.hpc.cineca.it/hpc-access/access-cineca-resources/iscra-projects/iscra-general-informations/

[22] https://csc.fi/en/

[23] https://kurami.disco.unimib.it/

[9] Z. Zheng, K. Hui, B. He, X. Han, L. Sun, A. Yates, Contextualized query expansion via unsupervised chunk selection for text retrieval, Information Processing & Management 58 (2021) 102672. URL: https://www.sciencedirect.com/science/article/pii/S0306457321001576. doi:https://doi.org/10.1016/j.ipm.2021.102672.
[10] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering in large language models: a comprehensive review, 2023. arXiv:2310.14735.
[11] L. Reynolds, K. McDonell, Prompt programming for large language models: Beyond the few-shot paradigm, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–7.
[12] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, 2022. arXiv:2212.10496.
[13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).
[14] S. E. Robertson, K. S. Jones, Relevance weighting of search terms, Journal of the American Society for Information science 27 (1976) 129–146.
[15] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec-3, Nist Special Publication Sp 109 (1995) 109.
[16] C. Macdonald, N. Tonellotto, Declarative experimentation ininformation retrieval using pyterrier, in: Proceedings of ICTIR 2020, 2020.
[17] J. Achiam, et al., GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[19] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, S. Sanghai, Gqa: Training generalized multi-query transformer models from multi-head checkpoints, arXiv preprint arXiv:2305.13245 (2023).
[20] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[21] J. Bai, et al., QWEN technical report, arXiv preprint arXiv:2309.16609 (2023).
[22] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[23] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[24] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).
[25] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, D. Bahri, T. Schuster, S. Zheng, et al., Ul2: Unifying language learning paradigms, in: The Eleventh International Conference on Learning Representations, 2022.
[26] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human-generated machine reading comprehension dataset (2016).
[27] E. Kamalloo, N. Thakur, C. Lassance, X. Ma, J.-H. Yang, J. Lin, Resources for brewing beir: Reproducible reference models and an official leaderboard, arXiv preprint arXiv:2306.07471 (2023).
[28] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, arXiv preprint arXiv:2104.08663 (2021).
[29] V. Boteva, D. Gholipour, A. Sokolov, S. Riezler, 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 38, Springer, 2016, pp. 716–722.
[30] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, et al., Overview of touché 2020: argument retrieval, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11,

Springer, 2020, pp. 384–395.
[31] A. Bondarenko, P. Braslavski, M. Völske, R. Aly, M. Fröbe, A. Panchenko, C. Biemann, B. Stein, M. Hagen, Comparative web search questions, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 52–60.
[32] J. Callan, The lemur project and its clueweb12 dataset, in: Invited talk at the SIGIR 2012 Workshop on Open-Source Information Retrieval, 2012.
[33] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, arXiv preprint arXiv:2004.14974 (2020).
[34] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, arXiv preprint arXiv:2004.07180 (2020).
[35] K. Balog, R. Neumayer, A test collection for entity search in dbpedia, in: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, 2013, pp. 737–740.
[36] F. Hasibi, F. Nikolaev, C. Xiong, K. Balog, S. E. Bratsberg, A. Kotov, J. Callan, Dbpedia-entity v2: a test collection for entity search, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1265–1268.
[37] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: constructing a pandemic information retrieval test collection, in: ACM SIGIR Forum, volume 54, ACM New York, NY, USA, 2021, pp. 1–12.
[38] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al., Natural questions: a benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 453–466.
[39] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, M. Leippold, Climate-fever: A dataset for verification of real-world climate claims, arXiv preprint arXiv:2012.00614 (2020).
[40] G. Amati, C. J. Van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Transactions on Information Systems (TOIS) 20 (2002) 357–389.
[41] V. Yukalov, Theory of cold atoms: Bose–einstein statistics, Laser Physics 26 (2016) 062001.
[42] T. Van Erven, P. Harremos, Rényi divergence and kullback-leibler divergence, IEEE Transactions on Information Theory 60 (2014) 3797–3820.