Multilingual Information Retrieval with Large Language Model-Driven Query Expansion

Jonathan Colaço Carr
McGill University

Jan Tiegges
McGill University

Daniel Kappert
McGill University

Abstract

Information Retrieval (IR) is a critical task in web search and data mining. Recent work has shown the potential of Large Language Models (LLMs) to expand search queries with relevant information, leading to improved IR. However, LLM query expansion has only been considered for English IR tasks. In this paper, we investigate LLM query expansion across five different languages: English, German, French, Spanish and simplified Chinese, and verify the effect of prompting observed in previous literature. Overall, we find LLM query expansion improves upon baseline retrieval methods in almost all cases, with short-Chain-of-thought prompting leading to the best results. Of the two language models considered (GPT 3.5 Turbo and Llama 2 7B), we found that GPT 3.5 Turbo consistently lead to better results, which may be due to the fact that it has more model parameters. We discuss several avenues for future work on multilingual IR[1].

1 Introduction

Information Retrieval (IR) is a core topic in modern data science. Roughly speaking, the goal of information retrieval is to efficiently retrieve documents from a database which are most relevant to a given query. IR is used for a variety of applications, from search engines (Ibrihich et al., 2022) to retrieval augmented generation (Lewis et al., 2020).

Recent work (Jagerman et al., 2023; Wang et al., 2023a; Claveau, 2022) has shown the ability for language models to aid in IR by augmenting the original query. Specifically these studies show that expanding queries with LLM-generated terms (referred to here as "LLM query expansion") improves the relevance of documents retrieved from BM25, a popular IR search strategy based on Term Frequency-Inverse Document Frequency (TF-IDF). Figure 1 highlights the main artefacts of LLM query expansion (discussed further in Section 3).

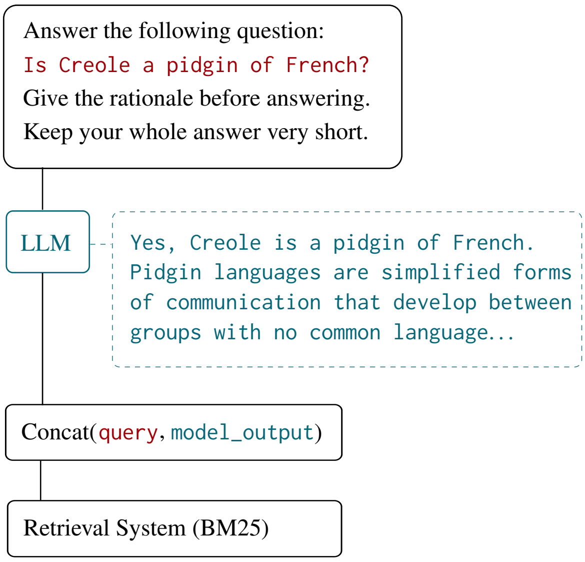
Answer the following question:
Is Creole a pidgin of French?
Give the rationale before answering.
Keep your whole answer very short.
LLM
Yes, Creole is a pidgin of French. Pidgin languages are simplified forms of communication that develop between groups with no common language...
Concat(query, model_output)
Retrieval System (BM25)

Figure 1: High-level overview of LLM query expansion using short Chain-of-Thought prompting. An example query and truncated model output are shown in red and blue, respectively.

However, to the best of our knowledge, prior work on LLM query expansion (Jagerman et al., 2023; Wang et al., 2023a; Claveau, 2022) has only considered IR for English datasets. Furthermore, the effect of prompting during LLM query expansion has only been considered on one family of language models; Google's suite of Flan models (Wei et al.). In this work, we investigate the impact of prompting on LLM query expansion for two language models and test LLM query expansion across five different languages: German (de), English (en), Spanish (es), French (fr) and simplified Chinese (zh). Our research hypotheses are:

1. Different prompting strategies during LLM query expansion will have varying effects depending on the LLM that is used.

2. LLM query expansion can improve IR across multiple languages.

To address (1), we compare two language models,

[1] Code: https://github.com/jantiegges/comp550-final

Open AI's GPT 3.5 Turbo[2] and Meta's Llama 2 7B[3], across six different prompting strategies for LLM query expansion. We compared these results to Jagerman et al.'s prompting experiments performed on Google's Flan models. Our findings show that although prompting effects GPT 3.5 Turbo and Llama 2 differently, both perform best on LLM query expansion when using short Chain-of-Thought prompts. For (2) we perform LLM query expansion with both GPT 3.5 Turbo and Llama 2 across five different languages: German (de), English (en), Spanish (es), French (fr) and simplified Chinese (zh). All experiments were conducted using MIRACL (Zhang et al., 2022), a multilingual IR dataset. We find that the efficacy of LLM query expansion across languages is highly model-dependent. While GPT 3.5 Turbo query expansion outperforms the baseline on all five languages, Llama 2 query expansion does not consistently improve baseline results for French and Chinese IR. This opens up interesting avenues for LLM query expansion and multilingual information retrieval.

2 Related Work

In this section we highlight a few related works on query expansion and multilingual information retrieval.

LLM query expansion. Using large language models to expand queries has been successfully used on English information retrieval tasks (Wang et al., 2023a; Jagerman et al., 2023; Claveau, 2022).

Query2doc (Wang et al., 2023a) used a few-shot prompting strategy that asked an LLM to answer the original query and used the LLM's response in the expanded query. Without any additional fine-tuning, this method improved the performance of BM25 retrieval on two English IR datasets - MS-MARCO (Bajaj et al., 2016) and TREC DL (Craswell et al., 2021).

In follow-up work, Jagerman et al. 2023 investigated other kinds of prompting strategies to obtain LLM-generated query terms, including asking an LLM to provide keywords related to a given query and asking the LLM to explain its reasoning for the answer. With Google's Flan (Wei et al.) models, Jagerman et. al found that Chain-of-Thought prompting led to the best expanded queries for information retrieval, improving on Query2doc's results for the MS-Marco dataset.

Our work expands the previous literature in LLM query expansion in two ways. First, we study whether the differences in prompting strategies observed by Jagerman et al. are also found in GPT 3.5 Turbo and LLama 2. Second, we investigate the effect of LLM query expansion on IR across multiple languages, as opposed to just English.

Multilingual Information Retrieval. We used the MIRACL dataset (Zhang et al., 2022) to test our methodology across five different languages (the MIRACL dataset is described in Section 3). Although various IR methods have already been examined for this dataset (e.g. (Huang et al., 2023; Yoon et al., 2023)), we are not aware of any prior work that uses LLM query expansion for multilingual information retrieval. However, there is one method that uses an LLM for multilingual IR.

The Search-Adaptor algorithm (Yoon et al., 2023) leverages the information contained in the text embeddings of LLMs to improve IR on the MIRACL dataset. Roughly speaking, the Search-Adaptor algorithm computes the similarity between queries and documents in the LLM embedding space. They fine-tune an LLM so that the its embeddings produce the highest overlap between queries and related documents on the training set. Although their method leads to improved performance on the MIRACL baselines, there is a significant cost associated to training algorithm to produce embeddings that are optimized for information retrieval.

In contrast to Search-Adaptor, our strategy does not require any fine-tuning or training on the MIRACL dataset. Also, while Search-Adaptor is designed to achieve the highest IR scores on the MIRACL dataset, our primary focus is to observe how LLM query expansion varies across languages, without necessarily obtaining the highest IR scores.

3 Methodology

Problem Formulation. Figure 1 shows the high-level idea of our method, which loosely follows that of Jagerman et al. 2023. Given an initial query q, the goal of LLM query expansion is to construct an expanded query q' that contains q as well as additional relevant information. Specifically, the expanded query consists of the original query q and the LLM's response to a `prompt` derived from q. The `prompt` may take additional pseudo-relevant

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://huggingface.co/meta-llama/Llama-2-7b

documents (PRDs) as input. PRDs are obtained by performing an initial BM25 search with the original query and selecting the top documents. As in (Jagerman et al., 2023; Wang et al., 2023a), we repeated the initial query five times and added it to the LLM output in order to up-weight the relative frequency of the original query. Therefore, the expanded query is defined as:

$$q' = \text{Concat}(q, q, q, q, q, \text{LLM}(\texttt{prompt}(q)), \quad (1)$$

where Concat denotes string concatenation. Table 1 outlines the six types of prompts that we considered.

ID	Prompt
`Ans`	Write a passage that answers the following query: `{query}`
`Ans-PRD`	Write a passage that answers the given query based on the context: Context:`{doc_1}{doc_2}{doc_3}` Query: `{query}` Passage:
`Kwd`	Write a list of keywords for the following query: `{query}`
`Kwd-PRD`	Write a list of keywords for the given query based on the context: Context:`{doc_1}{doc_2}{doc_3}` Query: `{query}` Keywords:
`CoT`	Answer the following query: `{query}` Give the rationale before answering
`CoT-PRD`	Answer the following query based on the context: Context:`{doc_1}{doc_2}{doc_3}` Query: `{query}` Give the rationale before answering
`CoT-Short`	Answer the following question: `{query}` Give the rationale before answering. Keep your whole answer very short.

Table 1: Prompts considered in our experiments. For prompts containing pseudo-related documents, we used the top three documents obtained with BM25 using the original query.

The first five prompts are similar to those found in Jagerman et. al. The last prompt, `CoT-Short`, was a new prompt that we tried after noticing that `CoT` produced very long outputs. To generate prompts for the non-English languages, we translated the prompt with the best scores on the English dataset using Google translate.

Dataset. To test our two research hypotheses, we used the MIRACL dataset (Zhang et al., 2022). This dataset was designed to benchmark information retrieval algorithms across a variety of languages. The MIRACL dataset has three main components for each of its 18 languages: a large corpus of text passages, a set of queries and a set of labelled query-passage pairs. The label of each query is a binary "judgement" of whether or not the passage is relevant to the query. We used the 'dev' split of the MIRACL dataset for evaluation as in the original paper (Zhang et al., 2022). In addition, we used the Pyserini (Lin et al., 2021) pre-built indices for the MIRACL dataset, which allowed us to speed up the BM25 search.

Evaluation. We evaluated the LLM query expansion techniques with two metrics: Recall@100 and nDCG@10. Recall@100 evaluates the ability of the model to retrieve all relevant documents by measuring the proportion of relevant documents among the top 100 results, thereby measuring the comprehensive retrieval ability of the model. Normalized Discounted Cumulative Gain at rank 10 (nDCG@10) measures the quality and relevance of the ranking of the top 10 results. It prioritizes more relevant documents at higher positions in the search results.

Baselines. We compared our LLM query expansion methods to the scores from a BM25 search with the original query. We are not aware of any other query expansion methods to compare with on the MIRACL dataset. Unfortunately, due to memory constraints we were only able to implement the BM-25 searcher, and were not able to experiment with other IR search methods such as mDPR. Investigating the effect of the search method on LLM query expansion is an important area of future work.

Models. We tested our prompts using two different language models:

- **GPT 3.5 Turbo**[4] is a publicly available language model from OpenAI. While the exact number of model parameters is not officially known, the lowest estimates are around 20 billion (Singh et al., 2023). It cost approximately

[4] https://platform.openai.com/docs/models/gpt-3-5

$10 CAD to run all of our GPT 3.5 Turbo experiments.

- **LLama 2**[5] is a family of LLMs from Meta. We used the Llama model with 7 billion parameters and were able complete all of our experiments using a free trial.

4 Results

4.1 Effect of Prompting across LLMs

Table 2 presents the outcomes of our experiments for the six different prompting strategies and the BM25 baseline. First of all, it's evident that query expansion generally leads to a significant improvement in performance for both GPT 3.5 Turbo and Llama 2. The largest improvement for both models was achieved with the short Chain-of-Thought (CoT) prompt, showing a 10% increase in recall and an even more impressive 47% in nDCG compared to the BM25 baseline. The advantage of `CoT-Short` compared to `CoT` may be that it increases the frequency of key terms in an LLM's response. This would be beneficial for BM25 since it is frequency-based search algorithm.

There are, however, a few notable differences between the trends in prompting strategies for the two models. Generally, GPT 3.5 Turbo shows greater improvements in recall, and Llama 2 shows more consistent improvements across prompts in nDCG. While prompting for answers (`Ans`) is better than prompting for keywords (`Kwd`) for GPT 3.5 Turbo, the reverse is true for Llama 2.

It is also interesting to note that for both models, adding pseudo-related documents (`-PRD`) decreases effectiveness of LLM query expansion. This is contrary to what was observed by Jagerman et. al 2023, where pseudo-related documents lead to improved information retrieval for most prompts. However, this discrepancy may be due to the fact that the experiments were conducted on different datasets.

4.2 Multilingual LLM Query Expansion

To select the prompt used in LLM query expansion across different languages, we used the best prompting strategy for each model obtained from Table 2, which happened to be `CoT-Short` for both models. We then used Google Translate[6] to translate the English `CoT-Short` prompt into French, German, Spanish, and Chinese. Table 3 shows the

	Recall@100	nDCG@10
Baseline		
BM25	0.8190	0.3506
GPT 3.5 Turbo		
`Ans`	0.8984	0.5122
`Kwd`	0.8433	0.3745
`Kwd-PRD`	0.8337	0.3722
`CoT`	0.8866	0.4877
`CoT-PRD`	0.8486	0.4343
`CoT-Short`	**0.9027**	**0.5154**
Llama 2		
`Ans`	0.8424	0.4219
`Kwd`	0.8688	0.4314
`Kwd-PRD`	0.8035	0.3635
`CoT`	0.8334	0.4340
`CoT-PRD`	0.8285	0.4265
`CoT-Short`	0.8890	0.4775

Table 2: Prompting strategies using GPT-3.5 Turbo and LLama 2 on the English portion of the MIRACL dataset. Best scores are in bold.

	Recall@100	nDCG@10
German (de)		
Baseline	0.5724	0.2262
GPT 3.5	**0.7225**	**0.3442**
Llama 2	0.6099	0.2864
English (en)		
Baseline	0.8190	0.3506
GPT 3.5	**0.9027**	**0.5154**
Llama 2	0.8890	0.4775
Spanish (es)		
Baseline	0.7018	0.3193
GPT 3.5	**0.7770**	**0.4034**
Llama 2	0.7409	0.3646
French (fr)		
Baseline	0.6528	0.1832
GPT 3.5	**0.7643**	**0.3016**
Llama 2	0.6089	0.2549
Chinese (zh)		
Baseline	0.5599	0.1801
GPT 3.5	**0.6929**	**0.2872**
Llama 2	0.5007	0.1451

Table 3: LLM query expansion results across all five languages considered. For both GPT 3.5 Turbo and Llama 2, the best prompt from Table 2 was translated into the target language. The best scores for each language are highlighted in bold.

[5] https://huggingface.co/meta-llama/Llama-2-7b

[6] https://translate.google.com

results of LLM query expansion using translated versions of the `CoT-Short` prompt for each of these languages in the MIRACL dataset.

It is clear from Table 3 that GPT 3.5 Turbo delivers the best scores across all languages, with the difference to the Llama model being only marginal for English but considerably larger for other languages. For French and Chinese, Llama performs worse than the baseline. It is also interesting to note that the performance gains over the baseline with the GPT 3.5 Turbo model varies greatly across languages. The largest improvements are seen in German and Chinese, whereas in Spanish, there is a significantly smaller relative improvement. The noticeable difference in performance between GPT 3.5 Turbo and Llama 2 may be due to the fact that the Llama model contains significantly less parameters than GPT 3.5 Turbo (7B vs. >20B).

For our second research hypothesis, it seems that the improvements of LLM query expansion across languages varies significantly depending on which language model is used. GPT 3.5 Turbo is able to consistently provide improved search queries, while Llama 2 does not show the same consistency.

5 Discussion & Conclusion

We accept our first research hypothesis, but note that although the effect of prompting during LLM query expansion is language model-dependent, short Chain-of-Thought reasoning is the best prompting strategy for both GPT 3.5 Turbo and Llama 2. Unlike Jagerman et. al 2023, we found that providing pseudo-relevant feedback for language models degraded IR, rather than improved it. This may be due to the fact that we used a different IR dataset.

For our second research hypothesis, we confirm that LLM query expansion often works in different languages for GPT 3.5 Turbo. However, since expansions from Llama 2 in French and Chinese do not improve upon baselines, we conclude that LLM query expansion does *not necessarily improve IR for all languages*. This may reflect the fact that LLMs are typically optimized to perform well on English data (Wang et al., 2023b). Our results suggest that improved IR via LLM query expansion does not come “for free”, and the LLM must have sufficient capabilities in the language before it can improve query expansion.

Limitations and Future Work. The MIRACL dataset opens up many interesting avenues for future work on multilingual IR. Due to memory constraints, we were not able to experiment with alternative search strategies to BM25 (e.g. mDPR). Therefore, future work could investigate whether the trends in query expansion observed in Section 4.1 are consistent for different search algorithms. Moreover, we were not able to experiment with all 18 languages in the MIRACL dataset. Further experiments could determine whether or not query expansion with GPT 3.5 Turbo is effective for improving IR on all 18 languages. A more detailed analysis of query expansion techniques could also connect the effectiveness of different prompts to morphological differences across languages. For example, the brevity of an LLM’s response may have different effects across languages depending on the language’s ratio of morphemes to words.

Statement of Contribution. All members contributed to report writing. JT created the initial codebase. JCC and DK performed prompting experiments.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Vincent Claveau. 2022. Neural text generation for query expansion in information retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT ’21, page 202–209, New York, NY, USA. Association for Computing Machinery.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M Voorhees, and Ian Soboroff. 2021. Trec deep learning track: Reusable test collections in the large data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2369–2375.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Cross-lingual knowledge transfer via distillation for multilingual information retrieval.

S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir. 2022. A review on recent research in information retrieval. *Procedia Computer Science*, 201:777–782. The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40).

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.

Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Negreanu, and Gust Verbruggen. 2023. Codefusion: A pre-trained diffusion model for code generation.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Data management for large language models: A survey.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jinsung Yoon, Sercan O Arik, Yanfei Chen, and Tomas Pfister. 2023. Search-adaptor: Text embedding customization for information retrieval.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *arXiv preprint arXiv:2210.09984*.