Could You Elaborate? A Multi-Perspective Evaluation of Query Reformulation in Retrieval-Augmented Generation

Stanisław Markowski[1 [0009−0001−3585−0975]], Konrad Wojtasik[1 [0000−0002−5715−5201]], Adrian Berdowski[1 [0009−0007−5687−8516]], Inez Okulska[1 [0000−0002−1452−9840]], and Maciej Piasecki[1 [0000−0003−1503−0993]]

Wrocław University of Science and Technology

Abstract. We propose a novel method for evaluating query reformulation in Retrieval-Augmented Generation (RAG) without relying on costly manual annotations of relevant documents. The approach measures both semantic similarity and downstream retrieval performance. We benchmark multiple reformulation techniques and large language models on a Polish university-regulations dialog dataset with an official document corpus as the retrieval base. The resulting evaluation pipeline enables multi-perspective assessment and can be readily adapted to new domains, document collections, reformulation methods, or settings with annotated relevance data. Our code is available on Github[1].

Keywords: Retrieval-Augmented Generation · Conversational Query Reformulation · Information Retrieval.

1 Introduction

LLM-based retrieval systems face challenges because conversational queries are often ambiguous, context-dependent, and shaped by references to earlier turns, including anaphora and deixis [14].

Query reformulation addresses this by converting context-dependent questions into concise self-contained representations of the user's intent. Unlike traditional ad-hoc retrieval, where queries are explicit, conversational queries evolve over multiple turns. Reformulation condenses this history into a single statement that retrieval, re-ranking, and generation components can process reliably.

In order to address those challenges, we prepared a conversational dataset on university documents and regulations, and an evaluation pipeline to check the models' performance on this task. To our knowledge, it is the first dataset of this kind for the Polish language. The main research questions were which model excels at the query reformulation task, which query reformulation method is the best choice, and how to perform a comprehensive evaluation for this task.

The main challenge was to prepare reliable evaluation metrics without access to gold-standard annotations for the target task, that is, a set of documents that

[1] https://anonymous.4open.science/r/anonymous_submission-063C.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

are expected to be retrieved by a particular query. Such an annotation is a very expensive process that requires extensive manual labor. In addition, we wanted these metrics to gauge semantic similarities as well as practical implications in a retrieval setting, to observe the outcomes from many perspectives. As a result, we propose the cosine similarity metric, the Reranker Support score, and the LLM Citation score that mimic the usage of retrieved documents in a downstream task. Access to reliable metrics that can be used without a gold-standard set of expected documents was a niche we wanted to explore in this work.

We address this gap by introducing metrics that operate without document-level relevance annotation, requiring only contextual and gold-standard queries, and a RAG corpus with answers.

Our main contributions:

- Developed the first conversational query reformulation dataset for the Polish language, focused on university regulations and documents.
- Proposed a comprehensive evaluation pipeline that can be extended to other query reformulation tasks involving annotated gold-standard, full-context queries. We highlight the limitations of certain evaluation methods and compare their results.
- Evaluated several strong multilingual and Polish LLMs on the query reformulation task using different methods, providing insights and identifying the most effective approaches and models.

The overview of the evaluation pipeline is presented in Figure 1.

2 Related Work

2.1 Query reformulation

Query reformulation can be achieved, for example, by utilizing an off-the-shelf large language model (open- or closed-source) without any expensive fine-tuning, by means of in-context learning. The early methods were proposed and tested on the GPT-2 [16] model, where the model de-contextualized the query based on previous queries providing few-shot and zero-shot results [19]. This setting is similar to our query-only method. Modern LLMs enabled more sophisticated prompts that instruct the model to clarify the user's last question, based on the entire conversation, and/or summarize the conversation to facilitate the following clarification [13].

The other approach to the query reformulation task is to generate pseudo-documents [10] or extend the query with pseudo-documents [17]. Both approaches are conceptually very similar and intend to use LLMs' knowledge in order to perform a better search. We extend this idea to generate a pseudo-response, which we call *the Vanilla* method because it only requires the model to generate a plausible answer in conversation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

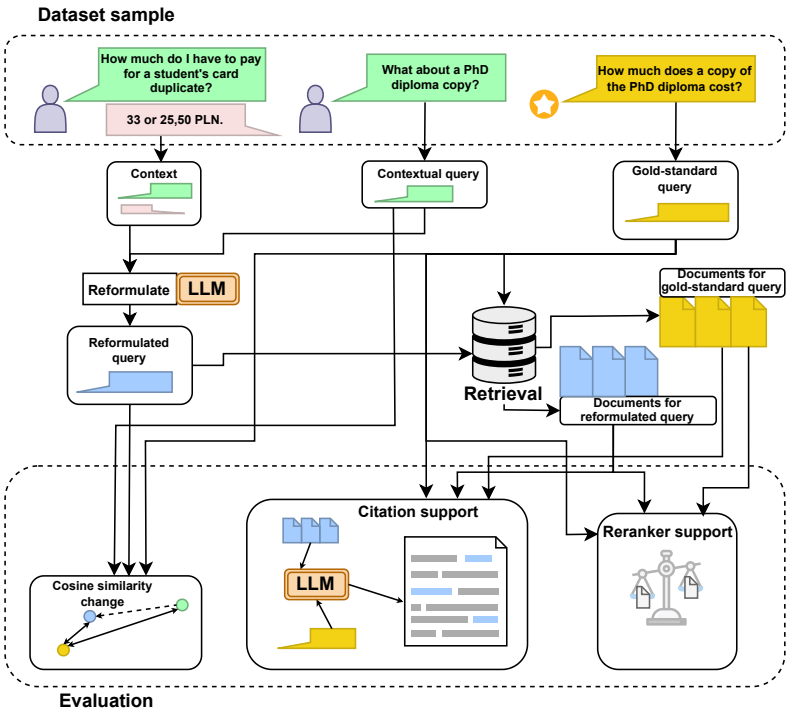

Fig. 1. Overview of the evaluation pipeline and its components, which involves: 1) reformulating query, 2) retrieving documents relevant to the reformulated query and its gold-standard version, 3) evaluating results in three ways: comparing cosine similarity, measuring number of used document citations in generated response, and estimating how well each document supports the gold-standard query using reranker model.

2.2 Query reformulation datasets

The existing query reformulation datasets, consisting of dialogs, are constructed with manual annotation and automatic generation [14]. One of the existing datasets is CANARD [8], a widely used dataset for training conversational query reformulation models. It contains manually rewritten versions of conversational queries into standalone queries for an effective search of specific Wikipedia sections. QReCC [1] is another prominent dataset focusing on conversational question answering, providing rich conversational contexts and the corresponding query reformulations tailored for information retrieval (IR) within a large collection of web pages. The TREC Conversational Assistance Track (CAsT) [7] datasets contribute multi-turn conversational challenges with benchmarks that emphasize natural query reformulations for search tasks. The ProMISe dataset [3] was designed to understand the user's intent to seek information and proactively suggest user questions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

2.3 Reformulation evaluation

The most common approaches to evaluate the quality of query reformulation can be divided into lexical similarity metrics or end-to-end retrieval performance metrics. The lexical similarity is measured with exact match, which measures if two sequences of tokens in two queries are identical after standard preprocessing steps, including stemming or lemmatization, and removal of stopwords and punctuation. An extension to this metric is token-level or word-level F1-score, to measure how complete the rewriting is with respect to human annotation. Since we can assume that reformulation is a specific case of paraphrasing, we can apply other lexical metrics, such as ROUGE [12], METEOR [2] and BLEU [15], which were also applied to compare reformulated queries [1]. Except for lexical similarity, it is possible to measure semantic similarity by applying embedding models trained for IR tasks, which are capable of representing queries in an appropriate space for adequate comparison. Sentence encoding models are a great fit for this task. In our evaluation, we apply the widely used [4] model [5], which provides multilingual support.

End-to-end retrieval performance is a complementary evaluation that assesses how well the reformulated query improves downstream retrieval effectiveness when used in an actual retrieval system. After reformulation, the query is used to retrieve documents (either passages or entire documents), and standard IR metrics are used, such as Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Recall at various cutoff thresholds [9].

However, the practical impact can only be measured if an annotated dataset is available for the IR task, which, in the case of new document collections or domains, can be both time-consuming and costly to create. In this work, we address the limitations of such evaluation by introducing novel metrics that allow us to assess the effectiveness of query reformulation without requiring manually annotated IR datasets.

3 Methodology

Our methodology consists of 3 evaluation methods: Semantic Similarity Comparison, Reranker Support Scores evaluation, and LLM Citation Score in the RAG setting with citation as a downstream task evaluation. Together, these metrics provide a clearer view of the reformulation methods and their real performance.

3.1 Semantic Similarity Comparison

The entire experimentation pipeline consists of embedding dataset gold-standard queries and contextual queries using *bge-m3* [4], followed by the reformulation of contextual queries using one of the backbone models and reformulation methods. The reformulated queries (or hypothetical documents; see Section 3.5) are embedded using the same *bge-m3* model. Then, we measure the cosine similarity of embeddings between the reformulated and gold-standard queries. Finally,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

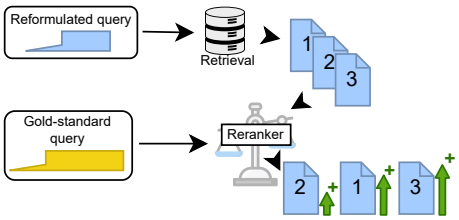

Fig. 2. Overview of the reranker support score pipeline and its components, which involves: 1) retrieving documents based on the reformulated query and retrieval pipeline, and 2) calculating retrieved documents' support for gold-standard query using *BAAI/bge-reranker-v2.5-gemma2-lightweight*.

we examine how similarity changes after applying the reformulation, specifically whether the reformulated query is closer to the gold-standard query than the original contextual question.

3.2 Reranker Support Scores

For each backbone LLM and reformulation method, we retrieved documents using the reformulated query and evaluated their support for the original gold-standard query with the *BAAI/bge-reranker-v2.5-gemma2-lightweight* [4] reranker, which scores relevance between the gold query and retrieved documents. According to the PIRB benchmark [6], this is the strongest reranker for Polish. We used two retrieval pipelines: a retriever+reranker setup (*BAAI/bge-m3+Qwen/Qwen3-Reranker-4B* [20]) retrieving 100 documents and reranking to the top 5; and a lightweight retriever-only pipeline limited to 5 documents. An example is shown in Figure 2.

$$\text{RerankerSupportScore}(q_r, q_g) = \frac{1}{k}\sum_{i=1}^{k} R(q_g, d_i),$$

where q_g is the gold-standard (original) query, q_r is the reformulated query, d_i is the i-th retrieved document, $R(q_g, d_i)$ is the relevance score assigned by the reranker between the gold query q_g and document d_i.

3.3 LLM Citation Score

We use strong LLMs as the final evaluation method. For each reformulation method, we provide the model with the five top retrieved documents, obtained either from the retriever alone or from the retriever–reranker pipeline. The model is then asked to answer the gold-standard query based solely on the information in the provided documents and to cite the sources in its response. If none of the documents contain information relevant to the query, the model must refuse to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

answer. We assume that strong LLMs can reliably perform this task and effectively understand the provided documents. A higher number of documents cited in an answer indicates that more relevant information was retrieved, suggesting that the corresponding reformulation method performs better on downstream tasks. To ensure a diverse set of evaluators, we conducted this evaluation using several different models: *Llama-3.3-70B-Instruct*, *c4ai-command-r-plus*, *Llama-3.1-70B-Instruct*, *Llama-PLLuM-70B-chat*. We present averaged results to reduce biases caused by models that tend to cite either too many or too few documents. Also, we include *Average Citation Score*, which is defined as:

$$\text{Avg. Cit. Score} = \frac{\text{num\_cit} - \frac{1}{M}\sum_{m=1}^{M}\text{num\_cit}_m}{\frac{1}{M}\sum_{m=1}^{M}\text{num\_cit}_m} \tag{1}$$

where, *num_ cit* is a number of citations in the answer generated from the document retrieved by an evaluated method, *num_ cit*$_m$ is a number of citations in the answer generated from the document retrieved by the method m and M is a total number of reformulation methods.

The final *Average Citation Score* is averaged across all test queries. This score indicates if the evaluated method performs better than average in all reformulation methods.

3.4 Backbone models

The backbone models tested in the conversational query reformulation task:

- *speakleash/Bielik-11B-v2.2-Instruct*
- *CohereForAI/c4ai-command-r-plus*
- *CYFRAGOVPL/Llama-PLLuM-70B-chat*
- *meta-llama/Llama-3.3-70B-Instruct*
- *openai/gpt-oss-120b*
- *CYFRAGOVPL/pllum-12b-chat*

Regardless of the backbone model, each one of them was prompted with the following parameters: temperature 0.1, max tokens 4096. All models are available on the Huggingface platform.

3.5 Reformulation methods

We divided the evaluated reformulation methods into two groups: *paraphrase*, and *hypothetical document* methods, the latter group inspired by HyDE [10]. Note that all reformulation methods require only one prompt, except for summarization (summarize + reformulate). It is also important to point out that, from the perspective of possible usages of reformulation outputs, both groups play different roles: paraphrase methods produce questions that are ready for classical RAG, while hypothetical document methods produce documents that should be treated differently when incorporating them into the pipeline, as they are used as synthetic passages that represent potential answers, rather than as queries. That is why it is important to use an adequate reranker model, as most available rerankers are trained to rank query-document pairs. We applied the LLM reranker with an adapted prompt to overcome this issue.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

Query Paraphrase Methods

Full dialog The backbone model receives the entire conversation with a prompt to reformulate the last question, or keep it as it is if the question is unrelated to the conversation or contains all relevant information for search.

Summarization The backbone model receives the full conversation and is first prompted to generate a summary that captures the contributions of both participants. In a subsequent step, the model is instructed to evaluate whether the user's last message is topically related to the preceding dialog. If the message is related, it is reformulated to explicitly include the topic of the conversation. If it is unrelated, it is carried forward without modification. An additional constraint specifies that if the last message is an instruction, it must be reproduced verbatim.

Questions-only The backbone model receives only user questions (indicated in the conversation by role: "user"), with a prompt to reformulate the last question, or keep it as it is if the question is unrelated to the conversation or includes all the information from the previous question's context.

Hypothetical Document Methods

Vanilla The backbone model receives the entire conversation, along with a prompt to generate a hypothetical text that can answer the user's last question. The prompt emphasizes that the text must be self-contained, i.e., it should use named entities rather than demonstrative pronouns referring to them, for example, *the internship* instead of *it*. The prompt also includes three-shot examples.

Masked vanilla Similarly to the vanilla reformulation, with an additional requirement to hide all key elements behind a [MASK] word, to avoid hallucinations, for example, masking people, job positions, addresses, places, and other proper names. Masking must be applied only to crucial words that did not appear in the conversation. The word [MASK] is filtered out before proceeding with the pipeline. The main rationale behind the masked vanilla method is to avoid searching for passages based on named entities and other key elements that might be inaccurate in the context of our document corpus.

Document The backbone model receives the entire conversation along with a prompt to create an example office text or statute passage that might contain answers for the user's last question. The model is instructed to find keywords in the conversation that are important for the answer and write them in the example text.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

User (Turn 1): What are the rules regarding professional internships?
Assistant: The form, location, schedule, and duration of professional internships are determined by the dean in accordance with the approved study program.

User (Turn 2): And who organizes them?
Reformulated query: Who organizes professional internships for students?
Assistant: Professional internships are organized by the university in cooperation with workplaces.

User (Turn 3): Is it mandatory at the undergraduate level?
Reformulated query: Is the professional internship mandatory for first-cycle (undergraduate) studies?
Assistant: Yes, the professional internship is mandatory for first-cycle studies and should last at least four weeks.

Fig. 3. Example of a conversation in the dataset. The conversation contains both user and reformulated questions. The user questions are contextual and require reformulation, in order to become standalone, full-context queries.

Masked document Similar to document reformulation, with additional encouragement to make up proper names, but they need to be hidden behind a [MASK] word. The rationale is the same as in the masked vanilla reformulation, that is, we want to check if it is better to mask named entities and other information in order to avoid searching with inaccuracies in the generated hypothetical document.

3.6 Dataset

Our experiments were carried out on a dataset of university institutional regulations [18]. Based on this corpus, we constructed 310 university conversational scenarios designed to reflect realistic interactions with a Retrieval Augmented Generation (RAG) system [11].

Using a randomly sampled collection of 2,000 aforementioned university documents' chunks, we constructed long-context inputs and used three different LLMs (ChatGPT-5, Claude 4.0, and Bielik 2.0) to generate a set of 320 candidate three-turn conversations, each paired with answers grounded in the source documents. To ensure diversity in conversational flow, the models were instructed to follow seven distinct schemas, varying the degree of topical relation between turns. For instance, in some schemas, question 3 referred back to question 1 or 2, while in others it was deliberately unrelated. This design provided a controlled range of scenarios in which query reformulation might be required or unnecessary.

For cases where a user question acted as a follow-up to a previous turn, an additional version of that query (Q_{ref}) was created, reformulated into a

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

clear stand-alone form. This allowed the dataset to support both reformulation experiments and direct retrieval tasks.

Each generated dialog was subsequently reviewed by two human annotators and a super-annotator, who verified that the conversation remained semantically consistent with the source documents, contextual queries were correctly labeled, and follow-up turns preserved coherence and logical continuity with preceding utterances.

Finally, each conversation consists of multi-turn exchanges in which user queries are paired with reference answers grounded in the corresponding university documents. Approximately 70% of the user queries were annotated as contextual, i.e., underspecified or context-dependent questions that require reformulation in order to be effectively processed by a retrieval system. An example of the conversation from the dataset is presented in Figure 3.

Augmentation The base dataset consisted of three-turn conversations. To extend the range of conversational depth and enable evaluation of reformulation methods in more complex dialog settings, we applied augmentation using the LLaMA 3.3[2] model. The model was prompted to generate preceding dialog turns that smoothly introduce and lead to the original three-turn conversations, ensuring thematic consistency and natural conversational flow. This process resulted in two extended variants of the dataset:

- **6-turn** version, comprising the original conversation plus three additional preceding turns,
- **11-turn** version, comprising the original conversation plus eight additional preceding turns.

This augmentation strategy allowed us to systematically evaluate query reformulation methods across dialogs of varying complexity, from short three-turn exchanges to longer, contextually rich eleven-turn conversations. The extended conversations were reviewed by a human annotator to ensure the quality of the dialogs and make the necessary corrections.

4 Experiments

4.1 Semantic Similarity Comparison Results

It seems reasonable to compare cosine similarity scores of HyDE-derived and paraphrase reformulation methods separately, as these methods produce outputs from different distributions, that is, either realistic answers or documents (the former group) or short, retrieval-ready questions (the latter group). This is reflected in our results shown in Table 1, as HyDE-based methods on average decreased similarity to gold queries.

In the paraphrase methods, we observed that the full dialog method generally brings the reformulated question embedding closest to the gold-standard query,

[2] meta-llama/Llama-3.3-70B-Instruct

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

with the most significant improvements when using PLLuM 70B or GPT 120B as presented in Table 1. The similarity improvement diminishes as we increase the length of the conversation. PLLuM 70B also achieved the best results in the question-only method, followed by the GPT 120B model in second place. Overall, the scores were slightly lower compared to those obtained with full dialogs. The summarization-based method led to a significant drop in performance for nearly all models except LLaMA 3.3 and GPT 120B, indicating that these models are capable of generating relevant conversation summaries that can be effectively used for query reformulation, in contrast to smaller models like Bielik and PLLuM 12B, where the performance drop was the most significant.

HyDE-derived methods consistently worsened similarity scores, as they generate texts from a distribution different from that of the gold-standard queries. This may lead to the mistaken conclusion that such methods should not be used, as they appear to degrade performance compared to contextual queries. The smallest negative impact was observed for the vanilla HyDE method, which might also lead to the incorrect assumption that it is the most effective among the HyDE-derived approaches. However, when considering the Reranker Support Scores and LLM Citation Scores, which will be discussed later, it becomes evident that the similarity metric is not adequate for comparing HyDE-derived methods. The results may indicate only that masking leads to a slight decrease in similarity when masks are applied.

We also examine the mean cosine similarity scores aggregated by each model, the reformulation method, and the number of conversation turns. Across all turns and methods, PLLuM 70B achieved the highest mean similarity. We present aggregated scores in Table 2.

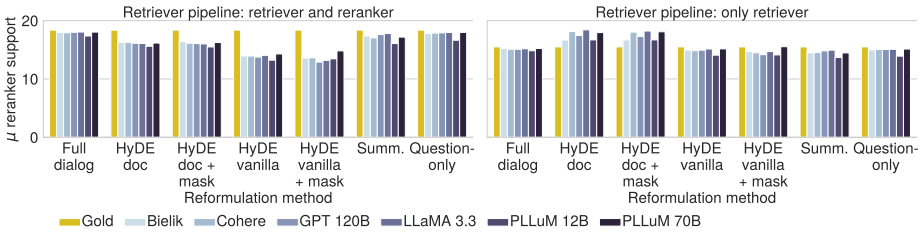

Fig. 4. Mean support of top 5 retrieved documents, aggregated by backbone LLM, reformulation method, and retrieval method. We can notice significant improvement in the retriever-only setting when using HyDE-document, comparable with full-dialog in retriever+reranker setting.

4.2 Reranker Support Scores Results

After evaluating the retrieval results using the reranker scores shown in Figure 4, we observed a similar performance pattern in all reformulation methods.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

		Number of conv. turns			
	Backbone	< 5	5-10	>=10	Any number
Full dialog	Bielik	0.096	0.079	0.074	0.083
	Cohere	0.104	0.092	0.089	0.095
	PLLuM 70B	**0.126**	**0.113**	**0.108**	**0.116**
	LLaMA 3.3	0.108	0.087	0.080	0.092
	GPT 120B	0.115	0.107	0.093	0.105
	PLLuM 12B	0.074	0.060	0.059	0.064
Summarization	Bielik	0.011	-0.005	-0.026	-0.006
	Cohere	0.063	0.029	0.018	0.037
	PLLuM 70B	0.073	0.060	0.046	0.060
	LLaMA 3.3	**0.118**	**0.101**	**0.097**	**0.105**
	GPT 120B	0.088	0.077	0.071	0.079
	PLLuM 12B	-0.118	-0.145	-0.149	-0.137
Questions-only	Bielik	0.082	0.066	0.054	0.068
	Cohere	0.099	0.087	0.082	0.090
	PLLuM 70B	**0.118**	**0.111**	**0.107**	**0.112**
	LLaMA 3.3	0.098	0.083	0.079	0.087
	GPT 120B	0.110	0.100	0.090	0.100
	PLLuM 12B	0.002	-0.064	-0.107	-0.055
HyDE doc	Bielik	-0.090	-0.096	-0.102	-0.096
	Cohere	-0.092	-0.103	-0.101	-0.098
	PLLuM 70B	-0.077	-0.084	-0.095	-0.085
	LLaMA 3.3	**-0.074**	**-0.083**	**-0.087**	**-0.081**
	GPT 120B	-0.085	-0.092	-0.098	-0.092
	PLLuM 12B	-0.099	-0.105	-0.112	-0.105
HyDE doc + mask	Bielik	-0.100	-0.109	-0.110	-0.106
	Cohere	-0.099	-0.113	-0.106	-0.106
	PLLuM 70B	**-0.078**	**-0.091**	**-0.096**	**-0.088**
	LLaMA 3.3	-0.089	-0.095	-0.099	-0.094
	GPT 120B	-0.096	-0.098	-0.107	-0.100
	PLLuM 12B	-0.111	-0.122	-0.119	-0.117
HyDE vanilla	Bielik	-0.079	-0.088	-0.088	-0.085
	Cohere	-0.077	-0.091	-0.097	-0.088
	PLLuM 70B	**-0.057**	**-0.066**	**-0.072**	**-0.065**
	LLaMA 3.3	-0.061	-0.074	-0.079	-0.071
	GPT 120B	-0.072	-0.076	-0.082	-0.076
	PLLuM 12B	-0.099	-0.116	-0.118	-0.111
HyDE vanilla + mask	Bielik	-0.100	-0.102	-0.107	-0.103
	Cohere	-0.087	-0.091	-0.092	-0.090
	PLLuM 70B	**-0.056**	**-0.070**	**-0.073**	**-0.066**
	LLaMA 3.3	-0.089	-0.097	-0.107	-0.097
	GPT 120B	-0.093	-0.098	-0.101	-0.097
	PLLuM 12B	-0.102	-0.098	-0.109	-0.103
Any method	Bielik	-0.026	-0.036	-0.043	-0.035
	Cohere	-0.013	-0.027	-0.029	-0.023
	PLLuM 70B	**0.007**	**-0.004**	**-0.011**	**-0.002**
	LLaMA 3.3	0.002	-0.011	-0.017	-0.009
	GPT 120B	-0.005	-0.012	-0.019	-0.012
	PLLuM 12B	-0.065	-0.084	-0.093	-0.081

Table 1. Mean change in cosine similarity with gold-standard query after applying reformulation, aggregated by backbone model, reformulation method, and number of conversation turns. This indicates if the reformulated query is more similar to gold-standard than the contextual query.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

	Reformulator model					
Cosine similarity	Bielik	Cohere	PLLuM 70B	LLaMA 3.3	GPT 120B	PLLuM 12B
< 5 conv. turns	0.771	0.784	**0.804**	0.799	0.792	0.732
5-10 conv. turns	0.766	0.775	**0.799**	0.791	0.791	0.718
>=10 conv. turns	0.764	0.778	**0.797**	0.791	0.789	0.714
Reranker support score						
Only retriever	15.39	15.74	15.93	**15.95**	15.54	14.86
Retriever+reranker	16.19	16.11	**16.38**	16.16	16.05	15.41

Table 2. Mean-aggregated scores for each backbone LLM across all reformulation methods: a) cosine similarity between reformulated and gold-standard query at different conversation lengths; b) mean reranker support score for documents retrieved by two pipelines - retriever+reranker and only retriever.

The full-dialog reformulation achieved slightly higher scores than the question-only approach, while the summarization method showed a noticeable drop in performance for all models except LLaMA 3.3 and GPT 120B. Query paraphrase methods outperform HyDE-derived methods, but only in the retriever with reranker pipeline. However, when the retriever-only pipeline is considered, all methods give similar performance, with an exception for the HyDE document and masked document, which achieve better results than gold-standard using the same pipeline. To our surprise, HyDE document retriever-only results turned out to be comparable to full dialog in a heavier retriever+reranker setting. These findings highlight the significant potential of HyDE-derived methods, which was not captured by the Semantic Similarity Comparison. As shown in Figure 4, the results suggest that the reranker worsens the document ordering for the document- and vanilla-based HyDE. This effect could potentially be mitigated through improved prompt engineering or employing a specialized model for document–document comparison, as most available rerankers are optimized for query–document ranking. The results also indicate that the HyDE vanilla method performs worse than the HyDE document variant in both retrieval settings. The aggregated results for each model, presented in Table 2, indicate that the optimal model choice depends on the pipeline configuration. For pipelines that include both a retriever and a reranker, PLLuM 70B and Bielik achieve the best performance and for pipelines that rely solely on a retriever, LLaMA 3.3 and PLLuM 70B.

4.3 LLM Citation Scores Results

We perform the evaluation on LLaMA 3.3 query reformulations and generate citation-based responses using four distinct LLMs. Each LLM produced an answer based on the top five documents retrieved by each reformulation method and retrieval pipeline, as shown in Table 3. The evaluation approach, which relies on document citations from strong LLMs, revealed that the full-dialog reformulation method performed nearly as well as retrieval using the gold-standard

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

query. The HyDE document method in the retriever-only setting ranked second, consistent with the Reranker Support Score results. This finding suggests that the HyDE document approach can serve as an effective and efficient method for retrieving relevant documents, as it requires only a retriever and avoids the computational cost of reranking. These results also indicate that our reranker configuration was suboptimal for document–document comparison, despite the use of a dedicated prompt for this task.

Method	**Retriever**	**Avg. Cit. Score**	**Avg. Num. Citations**
Gold-standard Query	retriever&reranker	**0.59**	**1.57**
Full dialog	retriever&reranker	0.54	1.48
HyDE doc	only retriever	0.38	1.36
Gold-standard Query	only retriever	-0.29	0.85
Full dialog	only retriever	-0.24	0.84
HyDE vanilla + mask	only retriever	-0.29	0.75
HyDE doc	retriever&reranker	-0.24	0.69
HyDE vanilla + mask	retriever&reranker	-0.46	0.57

Table 3. Comparison of reformulation methods with citation-based metrics. The provided results are an average across 4 different results from strong LLMs answers on a sample of 100 queries.

5 Conclusions

The main results indicate that the PLLuM 70B and Llama 3.3 70B models are the best choices for the query reformulation task in a conversational setting within the university documents domain. There is a significant drop in performance when using smaller models, such as Bielik or PLLuM 12B, particularly for summarization-based query reformulation.

We also found that the full dialog query reformulation method performs best overall. However, it requires a retrieval pipeline with an additional reranking component. In contrast, HyDE document reformulation performs exceptionally well when only a retriever is used in the pipeline, achieving competitive results even compared to the full dialog method with a retriever + reranker pipeline.

Our results indicate that evaluation based solely on semantic similarity is not sufficient, especially when comparing methods that generate hypothetical documents or answers. In such cases, similarity-based metrics can be misleading, even when using embeddings designed to position documents and queries closely within the same embedding space. Therefore, Reranker Support Scores and LLM Citation Score metrics are necessary to more accurately estimate performance in downstream tasks. Based on these results, we can confidently conclude that full dialog and HyDE document are the most promising query reformulation methods.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

Acknowledgements

(1) CLARIN-PL project financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education, agreement number 2024/WK/01. (2) CLARIN, the European Regional Development Fund, FENG programme (FENG.02.04-IP.040004/24); (3) statutory funds of the Department of Artificial Intelligence, Wroclaw Tech;

References

1. Anantha, R., Vakulenko, S., Tu, Z., Longpre, S., Pulman, S., Chappidi, S.: Open-domain question answering goes conversational via question rewriting. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 520–534. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.44, https://aclanthology.org/2021.naacl-main.44/
2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909/
3. Butala, Y., Garg, S., Banerjee, P., Misra, A.: ProMISe: A proactive multi-turn dialogue dataset for information-seeking intent resolution. In: Graham, Y., Purver, M. (eds.) Findings of the Association for Computational Linguistics: EACL 2024. pp. 1774–1789. Association for Computational Linguistics, St. Julian's, Malta (Mar 2024), https://aclanthology.org/2024.findings-eacl.124/
4. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2024)
5. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation (2024), https://arxiv.org/abs/2402.03216
6. Dadas, S., Perełkiewicz, M., Poświata, R.: PIRB: A comprehensive benchmark of Polish dense and hybrid text retrieval methods. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 12761–12774. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.1117
7. Dalton, J., Xiong, C., Callan, J.: Trec cast 2019: The conversational assistance track overview (2020), https://arxiv.org/abs/2003.13624
8. Elgohary, A., Peskov, D., Boyd-Graber, J.: Can you unpack that? learning to rewrite questions-in-context. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5918–5924. Association for Computational Linguis-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7

tics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1605, https://aclanthology.org/D19-1605/
9. Enevoldsen, K., et al.: MMTEB: Massive multilingual text embedding benchmark. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=zl3pfz4VCV
10. Gao, L., Ma, X., Lin, J., Callan, J.: Precise zero-shot dense retrieval without relevance labels. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1762–1777. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.99, https://aclanthology.org/2023.acl-long.99/
11. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., Wang, H.: Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 **2**(1) (2023)
12. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/
13. Mo, F., Ghaddar, A., Mao, K., Rezagholizadeh, M., Chen, B., Liu, Q., Nie, J.Y.: Chiq: Contextual history enhancement for improving query rewriting in conversational search (2024), https://arxiv.org/abs/2406.05013
14. Mo, F., Mao, K., Zhao, Z., Qian, H., Chen, H., Cheng, Y., Li, X., Zhu, Y., Dou, Z., Nie, J.Y.: A survey of conversational search. ACM Trans. Inf. Syst. **43**(6) (Sep 2025). https://doi.org/10.1145/3759453, https://doi.org/10.1145/3759453
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040/
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019), https://api.semanticscholar.org/CorpusID:160025533
17. Wang, L., Yang, N., Wei, F.: Query2doc: Query expansion with large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 9414–9423. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.585, https://aclanthology.org/2023.emnlp-main.585/
18. Wojtasik, K., Berdowski, A., Okulska, I., Piasecki, M.: Polichat: Retrieval augmented generation on university documents and regulations. In: International Conference on Computational Science. pp. 273–288. Springer (2025)
19. Yu, S., Liu, J., Yang, J., Xiong, C., Bennett, P., Gao, J., Liu, Z.: Few-shot generative conversational query rewriting. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1933–1936. SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3397271.3401323, https://doi.org/10.1145/3397271.3401323
20. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_7