# PoliChat: Retrieval Augmented Generation on University Documents and Regulations

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.** University regulations are often complex and difficult to navigate. To address this, we developed PoliChat, a Retrieval-Augmented Generation (RAG)-based chatbot that provides accurate and transparent access to regulatory information. Validated at Wrocław University of Science and Technology, PoliChat integrates real-time retrieval with citation mechanisms to enhance reliability. As part of our research, we prepared and annotated a dataset of university regulations to evaluate information retrieval and answer generation performance. We examine key factors that affect RAG performance in regulatory domains, including model size, document length, summarization, retrieved context size, and prompting strategies. We introduce Analyze&Answer, a prompting method that improves response coherence and citation accuracy.

**Keywords:** Information Retrieval · Retrieval-Augmented Generation · Natural Language Processing

## 1 Introduction

Navigating university regulations and procedures can be overwhelming due to their complexity, volume, and technical nature. For students and staff, finding specific information in official documents is often time-consuming and frustrating. Although consulting administrative staff may seem like an easier solution, their availability and ability to keep up with constantly evolving policies are inherently limited. To address these challenges, we developed a Retrieval-Augmented Generation (RAG)-based chatbot, designed to simplify access to regulatory information of a large university-like organization. Special attention is paid to accuracy and transparency. The chatbot has been implemented and experimentally validated in the case of the Wroclaw University of Technology and Science, a very large university in the south-west of Poland.

The system's key feature is its real-time retrieval of accurate university information with explicit source citations. Unlike Large Language Models (LLMs), which rely on static training data and may produce outdated responses, RAG systems dynamically integrate external sources. High-quality citations are a crucial objective of our system, because they allow users to verify the information directly. To achieve this, we experimented with RAG pipelines designed to re-evaluate each retrieved document and incorporate chosen documents into the

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

answer. This approach improves explainability by showing users which documents are used, ensuring confidence in the information provided.

Initial feedback from the deployment of this real-use-case scenario highlights its value: Students can efficiently navigate complex regulations, while staff use it as a reliable reference tool for university policies. By bridging the gap between users and institutional knowledge, the RAG-based chatbot offers a scalable, explainable, and transparent solution.

Our contributions can be summarized as follows:

- We manually prepared and corrected an evaluation dataset [1] based on complex university documents. The dataset is annotated with relevant passages (IR evaluation) as well as correct answers and citations (generator evaluation)
- We systematically examined the impact of key RAG design choices – model size, document length, retrieval strategies, summarization, prompting methods, and citation – to determine their impact on response and citation accuracy in a real-case environment of a complex regulatory domain.
- We proposed a prompting strategy, **Analyze & Answer (A&A)**, that significantly improves response quality and explainability in RAG settings while ensuring better grounding of the answer in retrieved documents.

## 2 Research Questions

We experimented with different model sizes, retrieval strategies, prompt designs, and citation mechanisms. While RAG enhances generative models with real-time retrieval, its effectiveness depends on understanding these elements, especially in complex regulatory domains like university policies. To this end, our study systematically investigates the following research questions:

1. **How does document length influence both retrieval effectiveness and response generation quality?** Since university regulations often contain long, dense documents, we explore whether longer documents degrade retrieval precision and affect answer quality.
2. **For long documents, how does summarization affect response quality, and to what extent does the choice of summarization method influence results?** Summarization can reduce retrieval noise, but it may also omit critical information.
3. **How does the amount of retrieved context provided in the generation prompt affect response quality?** In RAG systems, retrieved passages serve as external memory, but an excessive amount of context may lead to dilution of relevant information or model confusion.
4. **How does the choice of prompting strategy (e.g., "Analyze & Answer" vs. basic prompting) influence response and citation quality?**

[1] `https://huggingface.co/datasets/clarin-pl/polichat-rag-evaluation`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

Prompt engineering plays a crucial role in guiding the model's reasoning process. We compare a carefully designed *Analyze & Answer* (A&A) prompting strategy with a more basic approach to assess their impact on coherence, accuracy, and informativeness.

5. **To what extent does the presence of citations affect response quality in a RAG-based system?** While citations enhance trustworthiness and explainability, they may also introduce biases in response formulation.

Similar questions were explored in [13], but two key differences justify our study: (1) they focused on best practices for English QA benchmarks, while we evaluate a real-world Polish dataset in a specific domain, and (2) their pipeline lacked citations, whereas ours explicitly handles citation-based retrieval and response generation.

## 3 Related Work

Integration of RAG systems with domain-specific document repositories has attracted significant attention in recent years. This setting applies the power of the LLM generative capabilities and incorporates Information Retrieval to provide domain-specific and up-to-date information for proper answer generation.

**Information Retrieval** (IR) of relevant documents is a crucial part of any RAG system, as it fuels the proper generation of answers by long-context LLMs. Most encoders rely on BERT [5] architectures used as sentence encoders [16], with 512-token length input as the limiting factor. Recent advances introduced multilingual long-context retrievers, which can encode documents of up to 8k tokens [2,17,22]. Strong re-ranking models for the Polish language were distilled from larger multilingual models [4] or trained on multilingual datasets [2,17,22]. There are established IR benchmarks for Polish lanugage e.g. PIRB [3] and BEIR-PL [20], and multilingual MMTEB [11], which cover various domains, but still lack a good representation of long or regulatory documents in Polish.

**Citation-Enhanced Generation** – multiple approaches integrated citation mechanisms into RAG to improve the reliability of the generated answers [8,14]. Efficient Citer [18] explored fine-tuning smaller models via distillation to improve citation generation while maintaining efficiency. Self-RAG [1] and RARR [7] leveraged self-reflection and retrieval-on-demand strategies to refine answer attribution. Additionally, reinforcement learning with fine-grained rewards has proven beneficial in training models to generate accurate in-text citations [10]. ALCE [8] proposed a systematic method of evaluation citation quality based on TRUE NLI model [9]. However, this approach has limitations in assessing citation precision, relies on the use of computationally heavy LLM, and accumulates prediction errors. To address these issues, we have taken a different approach by creating a human-annotated dataset that enables citation evaluation without the need for an external model.

**Prompting** plays a crucial role in RAG, with various advanced strategies proposed beyond simple zero-shot prompts. Chain-of-Thought (CoT) [12]

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

enhances reasoning by breaking tasks into smaller steps, improving accuracy and transparency. Chain-of-Verification (CoV) [6] refines responses iteratively by generating verification questions to minimize factual errors. Chain-of-Note (CoN) [21] structures note-taking to summarize retrieved documents, aiding information aggregation. Chain-of-Knowledge (CoK) [19], guides step-by-step reasoning using evidence triplets and explanation hints similar to CoT. Unlike these approaches, our system focuses solely on relevant documents and refuses to answer when no relevant information is found, whereas most systems rely on LLMs' internal knowledge in such cases.

## 4 Dataset

### 4.1 University Database for RAG system

The dataset consists of 2,350 documents, including regulations, bylaws, and official texts issued by the university's rector, deans, or senate. Some digitized documents contain typing errors and coherence issues, with tables posing the greatest parsing challenge due to their linearized format. Covering the period from May 29, 1995 to December 20, 2024, document lengths range from 11 to 129,584 tokens. The average length is 1,327 tokens, with a median of 432 tokens. The strong focus on student affairs and university organization, often repeated in slightly different versions across departments, also challenges retriever models due to high semantic overlaps between documents.

### 4.2 Information Retrieval Dataset

We constructed a 100-question evaluation set using three methods: (1) **Human-written**: While grounded in realistic scenarios and phrased naturally, most discarded due to high complexity or answer ambiguity; (2) **LLM-generated (user intent)**: Using GPT-4o, Claude, and Bielik, prompted to imitate plausible questions from students, candidates, or employees; (3) **LLM-generated (document-based)**: Using Llama3-70B on sampled database chunks; often mirrored source text too closely, resulting in unnatural or trivial questions.

From almost 500 initial suggestions, only questions ensuring answerability within the corpus, unambiguous binary judgment, and natural phrasing remained. Thus, the final set, though small, reflects a high-effort curation process with extensive filtering and manual refinement.

Each question was manually annotated with relevant document chunks using configuration of two models: *bge-m3*[2] and *bge-reranker*[3] as retriever and re-ranker, respectively, we obtained the top20 results from the IR model for each question from our dataset and then manually annotated the relevant ones. Additionally, annotators searched the database with a GUI tool that allowed for keyword search and highlighted matches.

[2] BAAI/bge-m3

[3] BAAI/bge-reranker-v2-m3

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

Using both methods ensured more complete annotations: the retriever could miss relevant documents, while manual keyword searches depended on annotators anticipating all phrasings. Some queries also produced overly broad keywords, yielding hundreds of chunks—too many for manual review. This process produced two evaluation datasets: one with 8,201 passages for 512-token chunks and another with 2,873 passages for 4K-token chunks. The latter was mostly reannotated from parent documents, except for rare long documents where overlaps were checked.

### 4.3 Answer Generation Dataset

For Answer Generation Evaluation, we used the same questions as in IR evaluation but with detailed annotations. Each question was paired with document chunks categorized as relevant (containing necessary information and expected citations) or distractor (irrelevant and to be ignored).

To support our evaluation scheme, we manually annotated two word lists per question: *Include* words (ensuring completeness by requiring their presence in correct answers, with some alternatives, e.g., `["four", "4"]`), and *Exclude* words (incorrect or irrelevant terms). Exclude words were identified in provided chunks and annotators' knowledge. Inclusion and exclusion were checked using lemmatization.

To evaluate answer quality, we used multiple metrics:

- *Inclusion Accuracy* — checks if all required Include words are present, ensuring factual completeness.
- *Exclusion Accuracy* — verifies that misleading or irrelevant Exclude words are absent.
- *Citation F1* — measures correct citation of relevant document chunks while avoiding distractors.
- *Answer Length Analysis* — tracks character and word count to distinguish between verbosity and precision.

Some documents exceeded 32k tokens, surpassing our GPU memory limit. In such cases, full-document evaluation was impossible, so we also tested the model's performance on document summaries.

## 5 Methodology

### 5.1 Information Retrieval

We evaluated top-performing IR models for Polish, MMTEB and PIRB, using long-context retrievers and re-rankers for 4K-token passages (handling up to 8K tokens). For each 4K-token passage, we generated abstractive and extractive summaries to assess whether they retained critical information. Summaries enabled the use of 512-token context encoders. Generative summaries were produced with Llama3.3-70B using a hand-crafted prompt, while extractive summaries were generated via the Cohesive Coalition Algorithm [15].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

### 5.2 Generation

We evaluated five LLMs on the Answer Generation Dataset:

1. *Llama 3.1-8B*[4],
2. *Llama 3.1-70B*[5],
3. *Llama 3.3-70B*[6] with enhanced reasoning and instruction-following;
4. *Bielik*[7], a high-performing open Polish model;
5. *Command-R-Plus*[8], a 104B model specialized for RAG and citation-grounded answers, denoted as *cohere*.

We evaluated three answer generation strategies:

- *cite* – the model receives numbered passages (or documents) and generates an answer with proper citations;
- *nocite* – the same fragments are provided but concatenated without numbering, with no citation requirement;
- *A&A* – the model receives numbered fragments, first analyzes them in relation to the question, then provides the answer (see Sec. 5.3).

### 5.3 A&A Prompting

*A&A* is a two-stage answer generation strategy, designed as a specialized variant of Chain of Thought (CoT) for RAG tasks. Unlike standard CoT, it prioritizes grounding responses with citations and precise extraction of relevant information from retrieved documents. The details can be checked on our repository [9].

**Analyze** – The model first reviews retrieved documents, identifying relevant ones and extracting key information needed for a well-supported response. This step filters out irrelevant content, improving accuracy and citation quality.

**Answer** – The model then generates a response based solely on the extracted information. Separating analysis from answer generation enhances conciseness, citation correctness, and prevents hallucination.

### 5.4 RAG Pipeline Evaluation

The evaluation of the RAG pipeline begins with the retrieval phase, where the system retrieves the top relevant passages for each query. These passages are then provided as context to the generator. The relevance of each retrieved passage is assessed to determine whether it should be cited in the final answer. The correctness of the generated answer is evaluated using Inclusion and Exclusion Accuracy, as well as the accuracy of the provided citations.

[4] meta-llama/Llama-3.1-8B-Instruct

[5] meta-llama/Llama-3.1-70B-Instruct

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

[7] speakleash/Bielik-11B-v2.3-Instruct

[8] CohereForAI/c4ai-command-r-plus

[9] `https://github.com/CLARIN-PL/polichat-evaluation`

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

For 512-token fragments, we evaluated the pipeline using the retrieved fragments in the top $k \in \{5, 10, 20\}$. Increasing the number of retrieved fragments improves the likelihood that the relevant passage is included in the context provided to the model. However, this also introduces more irrelevant fragments, requiring the LLM to better understand and differentiate between relevant and irrelevant fragments. Additionally, this setting demands improved citation accuracy as the LLM must select from a larger set of documents.

For 4K-token chunks, we follow the same procedure, but limit retrieval to the top 5 fragments. This constraint ensures that the total input length to the generator remains below 32K-tokens, helping to limit GPU memory usage and reduce computation time.

**Table 1.** Information retrieval results for 512-token passages.

| retriever | reranker | MRR@10 | NDCG@10 | Recall@100 | Acc@5 |
|---|---|---|---|---|---|
| e5-large | bge-reranker | 88.32 | 82.03 | 94.82 | 95.00 |
| e5-large | pol-reranker | 80.95 | 78.45 | 94.82 | 90.00 |
| e5-large | jina-reranker | 80.31 | 77.31 | 94.82 | 91.00 |
| e5-large | plt5-large | 78.15 | 73.46 | 94.82 | 89.00 |
| e5-large | – | 65.56 | 59.31 | 94.82 | 78.00 |
| mmlw-large | bge-reranker | 86.82 | 79.88 | 91.20 | 94.00 |
| mmlw-large | – | 65.70 | 59.48 | 91.20 | 82.00 |
| gte-base | bge-reranker | 87.33 | 79.66 | 89.96 | 94.00 |
| gte-base | – | 57.94 | 49.92 | 89.96 | 71.00 |
| bge-m3 | bge-reranker | 85.82 | 78.33 | 89.49 | 93.00 |
| bge-m3 | – | 70.34 | 61.31 | 89.49 | 81.00 |
| jina-v3 | bge-reranker | 84.79 | 77.65 | 88.58 | 92.00 |
| jina-v3 | – | 59.12 | 52.05 | 88.58 | 69.00 |
| e5-base | bge-reranker | 86.15 | 78.56 | 87.74 | 93.00 |
| e5-base | – | 62.74 | 55.36 | 87.74 | 77.00 |
| bm25 | bge-reranker | 83.87 | 77.43 | 86.47 | 90.00 |
| bm25 | – | 53.78 | 49.16 | 86.47 | 68.00 |
| mmlw-base | bge-reranker | 85.22 | 77.30 | 85.36 | 91.00 |
| mmlw-base | – | 64.00 | 56.21 | 85.36 | 78.00 |

## 6 Experiments

We experimentally evaluated the solutions with respect to IR, answer generation, and, finally, the overall RAG pipeline.

### 6.1 Information Retrieval Results

To evaluate the effectiveness of retrieval strategies, we examined how document length and retrieval models influence retrieval quality in our RAG pipeline.

**Table 2.** Information retrieval results for 4K-token passages.

| retriever | reranker | MRR@10 | NDCG@10 | Recall@100 | Acc@5 |
|---|---|---|---|---|---|
| bm25 | bge-reranker | 79.93 | 76.39 | 93.91 | 91.00 |
| bm25 | jina-reranker | 63.84 | 58.42 | 93.91 | 78.00 |
| bm25 | gte-reranker | 59.89 | 53.48 | 93.91 | 75.00 |
| bm25 | – | 51.69 | 48.88 | 93.91 | 67.00 |
| bge-m3 | bge-reranker | 78.85 | 73.64 | 89.41 | 89.00 |
| bge-m3 | – | 57.61 | 50.80 | 89.41 | 76.00 |
| jina-v3 | bge-reranker | 75.48 | 71.45 | 84.90 | 88.00 |
| jina-v3 | – | 47.98 | 41.75 | 84.90 | 63.00 |
| gte-base | bge-reranker | 71.20 | 60.90 | 69.15 | 83.00 |
| gte-base | – | 37.27 | 30.84 | 69.15 | 52.00 |

**Impact of Retrieval Model on IR Performance** Table 1 presents IR results for 512-token passages. The best retriever in terms of Recall@100 was *multilingual-e5-large*[10] (94.8), indicating its superior ability to retrieve relevant passages. However, in terms of NDCG@10, which measures ranking quality, *bge-m3* achieved the best results, showing that its top-retrieved passages were more relevant. Given that our approach includes a re-ranking phase, maximizing recall in the initial retrieval phase was critical. The best IR pipeline combined *multilingual-e5-large* with *bge-reranker-v2-m3* as the re-ranker, yielding an NDCG@10 score of 82.03 and Acc@5 of 95.00, which is sufficient for our final system.

**Impact of Document Length on IR Performance** Table 2 presents the IR results for documents up to 4K tokens. Given the need to retrieve longer passages, we tested only models capable of handling long documents. Surprisingly, BM25 outperformed the other retrievers demonstrating the effectiveness of traditional lexical matching methods for long documents. The lowest Recall@100 score (69.15) was observed for *gte-multilingual-base*, suggesting that this model struggles with long Polish documents despite its strong performance on 512-token passages. The best-performing re-ranker remained *bge-reranker-v2-m3*, which significantly improved NDCG@10 scores across all retrievers, achieving the highest score (76.39) with BM25 and an Acc@5 of 91.00. However, the results for 4K-token fragments were lower than for 512-token passages, despite the smaller document collection size.

**Impact of Summarization on IR Performance** Tables 3 and 4 show IR performance on abstractive and extractive summaries. A significant drop was observed compared to both 4K-token fragments and 512-token passages. This suggests that while summarization condenses information, it may remove key lexical or semantic cues necessary for effective retrieval. The highest Recall@100 was 80.22 for abstractive summaries and 82.25 for extractive summaries, while NDCG@10 reached 58.47 and 58.98, respectively.

[10] intfloat/multilingual-e5-large-instruct

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

**Table 3.** Information retrieval result for abstractive summaries of 4K-token passages.

| retriever | reranker | MRR@10 | NDCG@10 | Recall@100 | Acc@5 |
|---|---|---|---|---|---|
| e5-large | bge-reranker | 65.08 | 58.47 | 80.22 | 78.00 |
| e5-large | – | 55.68 | 46.34 | 80.22 | 71.00 |
| bm25 | bge-reranker | 64.01 | 58.04 | 78.62 | 78.00 |
| bm25 | – | 47.98 | 42.52 | 78.62 | 65.00 |
| bge-m3 | bge-reranker | 65.58 | 58.83 | 78.15 | 78.00 |
| bge-m3 | – | 57.28 | 48.24 | 78.15 | 74.00 |
| gte-base | bge-reranker | 63.31 | 57.18 | 76.54 | 74.00 |
| gte-base | – | 45.06 | 38.67 | 76.54 | 61.00 |
| mmlw-large | bge-reranker | 65.03 | 57.35 | 75.85 | 76.00 |
| mmlw-large | – | 58.63 | 46.61 | 75.85 | 71.00 |

**Table 4.** Information retrieval result for extractive summaries of 4K-token passages.

| retriever | reranker | MRR@10 | NDCG@10 | Recall@100 | Acc@5 |
|---|---|---|---|---|---|
| jina-v3 | bge-reranker | 66.41 | 58.98 | 82.25 | 80.00 |
| jina-v3 | – | 51.03 | 42.18 | 82.25 | 65.00 |
| e5-large | bge-reranker | 68.09 | 59.49 | 79.62 | 81.00 |
| e5-large | – | 51.24 | 43.83 | 79.62 | 68.00 |
| bge-m3 | bge-reranker | 64.82 | 57.03 | 77.51 | 79.00 |
| bge-m3 | – | 48.64 | 40.87 | 77.51 | 60.00 |
| bm25 | bge-reranker | 68.01 | 60.16 | 77.17 | 81.00 |
| bm25 | – | 46.65 | 44.12 | 77.17 | 62.00 |
| mmlw-large | bge-reranker | 64.94 | 56.34 | 73.70 | 77.00 |
| mmlw-large | – | 53.14 | 44.74 | 73.70 | 65.00 |

### 6.2 Answer Generation Results

To evaluate answer quality, we analyzed the impact of model size, document length, retrieved context, and prompting strategies on response accuracy and citation quality.

**Impact of Passages Length on Answer Quality** Table 5 presents Answer Generation results across different LLMs for 512-token passages. The highest total score was achieved by Command-R-Plus (87.32) with A&A prompting and Llama3.3-70B (83.94) with basic prompt. Longer documents reduced the correctness across all models, as shown in Table 6, with the highest score of 84.34 achieved with A&A prompting and Command-R-Plus, and also 81.59 for the basic prompt by Llama3.3-70B. Despite the support for a long context, performance still declines, highlighting the benefit of providing more fine-grained information. The primary factor contributing to the decline in the total score is the citation F1 score, implying that models struggle to provide proper grounding when answering based on long passages. However, the smaller model Llama3.1-8B shows a significant increase in include ratio (74.50 → 80) and citation recall (55.62 → 80.75) when using longer documents with A&A prompting.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21

**Impact of Summarization on Response Quality** Tables 7 and 8 show that answer quality declines when using summaries instead of full 4K-token passages. The total score dropped significantly from 84.34 to 66.27 for abstractive summaries and to 70.42 for extractive summaries. Across all models, correctness scores were consistently lower for abstractive summaries compared to extractive summaries, confirming that crucial information is lost in paraphrasing.

**Impact of Prompting Strategy on Answer and Citation Quality**, the A&A strategy improved citation F1 and answer correctness across all models except Llama3.1 (both sizes) on 512-token passages. Compared to the basic prompt, responses have higher include scores and improved citation recall, but excessive detail lowered exclude scores. Citation precision increased, as the model was double-checking the sources. For longer documents (in 4K-token evaluation), while answers included all necessary details and citations, they also contained too much irrelevant information resulting in even more visible exclude score drops. Overall, applying A&A prompting leads to a consistent improvement in total scores in all settings while also providing additional analysis of documents, that may be helpful to the user.

**Impact of Citations on Response Quality** varied by passage length and citation strategy (Tables 5, 6). In a basic prompt, citations significantly reduced correctness for every tested model on 512-token passages, with similar trends in 4K-token evaluation. However, the A&A mostly improved correctness beyond basic prompting, e.g., for Llama3.3-70B (87. 00% → 88. 14%). However, Llama3.1, regardless of the model size, does not respond well to the citations, except for 4k-token passages, where the basic prompt may have slightly better results but then drop again with A&A. Note that citation-heavy prompts almost doubled the response length.

**Table 5.** Answer Generation on 512-token passages provided as a context to the model.

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | nocite | 57.70 | 86.54 | 86.83 | 86.26 | 0.00 | 0.00 | 0.00 | 266 | 35 |
| bielik | basic | 78.72 | 84.52 | 89.67 | 79.38 | 67.12 | 80.17 | 63.03 | 586 | 78 |
| bielik | A&A | 82.96 | 86.68 | 87.92 | 85.43 | 75.53 | 90.33 | 70.78 | 415 | 56 |
| cohere | nocite | 57.11 | 85.66 | 79.50 | 91.82 | 0.00 | 0.00 | 0.00 | 106 | 14 |
| cohere | basic | 81.21 | 84.56 | 80.25 | 88.86 | 74.53 | 77.52 | 76.62 | 149 | 21 |
| cohere | A&A | 87.32 | 88.66 | 88.50 | 88.82 | 84.65 | 88.08 | 85.97 | 223 | 31 |
| llama31-70b | nocite | 58.75 | 88.12 | 83.33 | 92.92 | 0.00 | 0.00 | 0.00 | 134 | 18 |
| llama31-70b | basic | 80.91 | 87.57 | 84.25 | 90.89 | 67.60 | 87.75 | 60.03 | 183 | 25 |
| llama31-70b | A&A | 82.44 | 85.79 | 83.67 | 87.91 | 75.74 | 85.75 | 72.70 | 177 | 24 |
| llama31-8b | nocite | 57.54 | 86.30 | 80.00 | 92.61 | 0.00 | 0.00 | 0.00 | 130 | 17 |
| llama31-8b | basic | 73.00 | 85.15 | 82.08 | 88.22 | 48.70 | 66.50 | 41.67 | 222 | 31 |
| llama31-8b | A&A | 73.11 | 81.56 | 74.50 | 88.61 | 56.22 | 65.28 | 55.62 | 259 | 36 |
| llama33-70b | nocite | 58.00 | 87.00 | 83.67 | 90.34 | 0.00 | 0.00 | 0.00 | 164 | 22 |
| llama33-70b | basic | 83.94 | 85.78 | 87.33 | 84.22 | 80.28 | 87.17 | 79.20 | 315 | 43 |
| llama33-70b | A&A | 85.95 | 88.14 | 88.67 | 87.61 | 81.56 | 87.33 | 80.98 | 259 | 37 |

**Table 6.** Answer Generation on 4K-token passages provided as a context to the model.

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | nocite | 56.74 | 85.12 | 81.83 | 88.40 | 0.00 | 0.00 | 0.00 | 259 | 35 |
| bielik | basic | 73.68 | 84.87 | 87.17 | 82.57 | 51.31 | 62.17 | 48.32 | 548 | 73 |
| bielik | A&A | 78.43 | 83.50 | 89.08 | 77.93 | 68.28 | 76.90 | 68.45 | 844 | 110 |
| cohere | nocite | 54.80 | 82.20 | 72.58 | 91.82 | 0.00 | 0.00 | 0.00 | 103 | 14 |
| cohere | basic | 75.67 | 81.47 | 70.67 | 92.27 | 64.08 | 66.97 | 65.83 | 148 | 21 |
| cohere | A&A | 84.34 | 86.33 | 82.75 | 89.91 | 80.38 | 82.35 | 83.78 | 307 | 43 |
| llama31-70b | nocite | 59.02 | 88.53 | 83.83 | 93.23 | 0.00 | 0.00 | 0.00 | 128 | 17 |
| llama31-70b | basic | 79.48 | 89.31 | 86.00 | 92.62 | 59.82 | 79.67 | 51.87 | 172 | 23 |
| llama31-70b | A&A | 81.16 | 87.07 | 90.08 | 84.06 | 69.36 | 69.08 | 78.05 | 652 | 89 |
| llama31-8b | nocite | 54.89 | 82.33 | 72.33 | 92.33 | 0.00 | 0.00 | 0.00 | 134 | 18 |
| llama31-8b | basic | 70.43 | 83.76 | 79.00 | 88.51 | 43.79 | 59.33 | 37.92 | 217 | 29 |
| llama31-8b | A&A | 76.42 | 80.99 | 80.00 | 81.98 | 67.30 | 62.80 | 80.75 | 996 | 137 |
| llama33-70b | nocite | 58.14 | 87.22 | 83.42 | 91.01 | 0.00 | 0.00 | 0.00 | 149 | 20 |
| llama33-70b | basic | 81.59 | 88.22 | 89.67 | 86.77 | 68.32 | 81.83 | 64.62 | 287 | 40 |
| llama33-70b | A&A | 82.73 | 87.76 | 90.58 | 84.94 | 72.66 | 79.48 | 75.42 | 646 | 89 |

**Table 7.** Answer Generation results on abstractive summaries of 4K-token passages.

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | nocite | 46.98 | 70.47 | 47.50 | 93.44 | 0.00 | 0.00 | 0.00 | 218 | 29 |
| bielik | basic | 57.17 | 66.81 | 45.03 | 88.59 | 37.90 | 45.92 | 35.75 | 432 | 57 |
| bielik | A&A | 65.97 | 71.56 | 52.17 | 90.95 | 54.80 | 64.92 | 52.12 | 400 | 53 |
| cohere | nocite | 45.92 | 68.89 | 40.92 | 96.86 | 0.00 | 0.00 | 0.00 | 90 | 12 |
| cohere | basic | 61.30 | 68.20 | 42.33 | 94.07 | 47.51 | 52.08 | 47.37 | 131 | 19 |
| cohere | A&A | 66.27 | 70.66 | 49.00 | 92.32 | 57.48 | 65.33 | 55.32 | 208 | 29 |
| llama31-70b | nocite | 43.98 | 65.97 | 35.33 | 96.61 | 0.00 | 0.00 | 0.00 | 99 | 14 |
| llama31-70b | basic | 59.92 | 69.58 | 43.00 | 96.16 | 40.59 | 51.50 | 36.20 | 127 | 17 |
| llama31-70b | A&A | 58.08 | 67.37 | 40.17 | 94.57 | 39.49 | 47.17 | 37.45 | 130 | 18 |
| llama31-8b | nocite | 44.06 | 66.10 | 34.83 | 97.36 | 0.00 | 0.00 | 0.00 | 84 | 11 |
| llama31-8b | basic | 48.93 | 65.59 | 35.20 | 95.98 | 15.60 | 21.00 | 13.83 | 146 | 20 |
| llama31-8b | A&A | 55.84 | 63.97 | 34.83 | 93.11 | 39.57 | 45.43 | 38.25 | 211 | 29 |
| llama33-70b | nocite | 42.59 | 63.89 | 32.17 | 95.61 | 0.00 | 0.00 | 0.00 | 94 | 13 |
| llama33-70b | basic | 58.02 | 66.12 | 40.67 | 91.57 | 41.82 | 47.17 | 40.62 | 197 | 27 |
| llama33-70b | A&A | 61.50 | 68.98 | 44.33 | 93.63 | 46.53 | 51.98 | 46.12 | 188 | 27 |

**Table 8.** Answer Generation results on extractive summaries of 4K-token passages.

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | nocite | 48.27 | 72.41 | 52.13 | 92.69 | 0.00 | 0.00 | 0.00 | 226 | 30 |
| bielik | basic | 61.64 | 71.44 | 53.43 | 89.46 | 42.01 | 49.92 | 40.12 | 474 | 63 |
| bielik | A&A | 68.80 | 73.67 | 60.63 | 86.71 | 59.04 | 70.92 | 55.28 | 461 | 61 |
| cohere | nocite | 48.50 | 72.74 | 49.88 | 95.61 | 0.00 | 0.00 | 0.00 | 80 | 11 |
| cohere | basic | 65.81 | 73.63 | 51.13 | 96.13 | 50.17 | 55.92 | 49.00 | 102 | 15 |
| cohere | A&A | 70.42 | 74.24 | 56.13 | 92.36 | 62.78 | 69.58 | 61.78 | 163 | 23 |
| llama31-70b | nocite | 48.09 | 72.13 | 49.80 | 94.46 | 0.00 | 0.00 | 0.00 | 102 | 14 |
| llama31-70b | basic | 64.90 | 73.32 | 54.80 | 91.85 | 48.05 | 61.75 | 42.87 | 148 | 20 |
| llama31-70b | A&A | 67.03 | 73.31 | 55.47 | 91.15 | 54.47 | 66.25 | 49.53 | 160 | 22 |
| llama31-8b | nocite | 47.33 | 71.00 | 45.55 | 96.44 | 0.00 | 0.00 | 0.00 | 96 | 13 |
| llama31-8b | basic | 58.13 | 69.07 | 46.38 | 91.76 | 36.23 | 48.17 | 31.67 | 178 | 24 |
| llama31-8b | A&A | 64.06 | 71.74 | 51.37 | 92.10 | 48.71 | 56.75 | 46.62 | 280 | 38 |
| llama33-70b | nocite | 48.09 | 72.13 | 49.63 | 94.63 | 0.00 | 0.00 | 0.00 | 118 | 16 |
| llama33-70b | basic | 67.17 | 72.77 | 54.80 | 90.74 | 55.97 | 61.08 | 55.53 | 271 | 37 |
| llama33-70b | A&A | 65.20 | 71.47 | 52.63 | 90.31 | 52.65 | 58.33 | 52.12 | 209 | 29 |

### 6.3 RAG Pipeline Results

In this section, we present the results of the full RAG pipeline, where models generate answers based on retrieved documents. Table 9 shows model performance with top k∈5, 10, 20 retrieved passages. As more documents are provided, performance declines, particularly in citation F1, due to the challenge of selecting correct sources. However, Analyze & Answer (A&A) mitigates this drop, improving citation accuracy with larger context sizes. Table 10 compares full 4K-token passages vs. summaries. Summarization significantly lowers performance, indicating information loss. A&A improves overall scores, making it particularly effective for longer texts or incomplete contexts like summaries.

## 7 Conclusions

In this work, we evaluated model performance on IR and RAG tasks, analyzing the impact of model size, document length, summarization, citations, and retrieved context size. Larger models performed better, but retrieval quality remained critical. Longer documents reduced recall, and extractive summaries retained more relevant information than abstractive ones.

While citations improved transparency, they sometimes reduced correctness, and longer retrieved contexts made citation selection more challenging. However, in real-world regulatory and administrative domains, where users need direct access to official sources, citation accuracy is valuable, as it enables users to verify and navigate referenced documents themselves. The Analyze&Answer strategy addressed these challenges, improving overall performance, citation accuracy, and readability. These findings highlight the importance of retrieval strategies, effective prompting, and model scalability in optimizing RAG systems.

DOI: 10.1007/978-3-031-97570-7_21

**Table 9.** RAG pipeline results on top $k \in \{5, 10, 20\}$ most relevant 512-token passages. The value of $k$ is added to the prompt name.

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | basic_top20 | 68.75 | 78.38 | 80.83 | 75.94 | 49.49 | 64.42 | 51.89 | 784 | 104 |
| bielik | basic_top10 | 71.60 | 80.92 | 81.33 | 80.51 | 52.95 | 66.73 | 51.04 | 734 | 98 |
| bielik | basic_top5 | 78.14 | 83.78 | 86.50 | 81.07 | 66.84 | 79.60 | 63.72 | 632 | 84 |
| bielik | A&A_top20 | 68.08 | 80.62 | 75.67 | 85.57 | 43.01 | 52.69 | 46.89 | 637 | 86 |
| bielik | A&A_top10 | 72.01 | 79.81 | 76.67 | 82.94 | 56.41 | 67.14 | 54.18 | 513 | 69 |
| bielik | A&A_top5 | 75.10 | 80.38 | 79.00 | 81.76 | 64.55 | 76.70 | 61.33 | 514 | 70 |
| cohere | basic_top20 | 72.46 | 81.10 | 71.50 | 90.71 | 55.15 | 55.78 | 63.52 | 175 | 28 |
| cohere | basic_top10 | 77.22 | 83.42 | 75.50 | 91.35 | 64.79 | 67.70 | 70.61 | 173 | 26 |
| cohere | basic_top5 | 81.47 | 85.20 | 82.00 | 88.40 | 74.00 | 76.63 | 77.98 | 170 | 25 |
| cohere | A&A_top20 | 76.55 | 85.16 | 84.00 | 86.32 | 59.32 | 58.81 | 74.28 | 265 | 40 |
| cohere | A&A_top10 | 79.30 | 85.26 | 83.33 | 87.20 | 67.36 | 66.78 | 76.01 | 232 | 35 |
| cohere | A&A_top5 | 80.27 | 83.46 | 80.67 | 86.24 | 73.91 | 75.62 | 79.07 | 210 | 30 |
| llama31-70b | basic_top20 | 73.15 | 85.45 | 79.50 | 91.40 | 48.55 | 78.94 | 41.45 | 158 | 22 |
| llama31-70b | basic_top10 | 75.47 | 85.78 | 82.33 | 89.23 | 54.83 | 84.16 | 46.10 | 190 | 26 |
| llama31-70b | basic_top5 | 76.99 | 86.03 | 83.83 | 88.23 | 58.90 | 83.60 | 50.65 | 193 | 26 |
| llama31-70b | A&A_top20 | 78.91 | 86.97 | 85.00 | 88.94 | 62.79 | 74.85 | 60.77 | 219 | 31 |
| llama31-70b | A&A_top10 | 79.64 | 87.66 | 86.50 | 88.82 | 63.59 | 76.15 | 61.65 | 198 | 28 |
| llama31-70b | A&A_top5 | 79.83 | 85.60 | 84.17 | 87.02 | 68.29 | 81.18 | 64.57 | 193 | 27 |

**Table 10.** RAG pipeline results for the most relevant 4K-token passages and summaries (*abs* for abstractive, *ext* for extractive summaries and *4K* for whole 4k-token passages)

| model | prompt | total | ans_corr | include | exclude | f1 | precision | recall | chars | words |
|---|---|---|---|---|---|---|---|---|---|---|
| bielik | basic_abs | 57.92 | 67.01 | 46.00 | 88.02 | 39.75 | 47.08 | 38.92 | 489 | 64 |
| bielik | basic_ext | 59.89 | 68.84 | 54.13 | 83.56 | 41.97 | 48.23 | 40.92 | 624 | 82 |
| bielik | A&A_abs | 61.65 | 68.69 | 47.67 | 89.72 | 47.56 | 55.58 | 46.18 | 386 | 51 |
| bielik | A&A_ext | 63.49 | 72.59 | 59.47 | 85.71 | 45.29 | 52.67 | 44.02 | 454 | 60 |
| bielik | basic_4k | 66.60 | 76.27 | 73.67 | 78.87 | 47.27 | 54.82 | 49.35 | 679 | 90 |
| bielik | A&A_4k | 68.47 | 79.86 | 76.67 | 83.04 | 45.70 | 57.42 | 43.85 | 455 | 62 |
| cohere | basic_abs | 60.17 | 68.47 | 45.50 | 91.44 | 43.57 | 45.30 | 47.47 | 168 | 24 |
| cohere | A&A_abs | 61.46 | 68.86 | 49.37 | 88.36 | 46.65 | 47.35 | 53.72 | 251 | 35 |
| cohere | basic_ext | 61.62 | 70.30 | 49.13 | 91.46 | 44.27 | 45.50 | 48.52 | 157 | 22 |
| cohere | A&A_ext | 64.84 | 72.03 | 54.80 | 89.25 | 50.47 | 53.25 | 52.75 | 207 | 29 |
| cohere | basic_4k | 71.57 | 78.97 | 65.67 | 92.27 | 56.77 | 57.88 | 60.68 | 135 | 20 |
| cohere | A&A_4k | 75.12 | 82.79 | 75.63 | 89.96 | 59.78 | 60.00 | 65.95 | 185 | 27 |
| llama31-70b | basic_abs | 58.70 | 69.00 | 43.17 | 94.83 | 38.10 | 48.08 | 34.85 | 162 | 22 |
| llama31-70b | A&A_abs | 59.50 | 67.94 | 43.17 | 92.72 | 42.61 | 48.58 | 43.07 | 186 | 26 |
| llama31-70b | basic_ext | 60.12 | 70.28 | 51.97 | 88.59 | 39.80 | 53.00 | 35.43 | 170 | 23 |
| llama31-70b | A&A_ext | 61.66 | 69.61 | 51.13 | 88.10 | 45.75 | 54.17 | 44.55 | 181 | 25 |
| llama31-70b | basic_4k | 73.05 | 85.84 | 78.17 | 93.52 | 47.47 | 67.00 | 40.23 | 148 | 20 |
| llama31-70b | A&A_4k | 76.10 | 84.38 | 79.83 | 88.94 | 59.53 | 70.63 | 56.27 | 168 | 24 |

## Acknowledgements

Financed by: (1) CLARIN ERIC (2024–2026), funded by the Polish Minister of Science (agreement no. 2024/WK/01); (2) CLARIN-PL, the European Regional Development Fund, FENG programme (FENG.02.04-IP.040004/24); (3) statutory funds of the Department of Artificial Intelligence, Wroclaw Tech; (4) the EU project "DARIAH-PL", under investment A2.4.1 of the National Recovery and Resilience Plan. The views expressed are those of the authors and do not necessarily reflect those of the EU or the European Research Executive Agency.

## References

1. Akari Asai, e.a.: Self-rag: Learning to retrieve, generate, and critique through self-reflection (2023), https://arxiv.org/abs/2310.11511
2. 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
3. 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 LREC-COLING 2024. pp. 12761–12774. ELRA and ICCL, Torino, Italia (May 2024)
4. Dadas, S., Grębowiec, M.: Assessing generalization capability of text ranking models in polish (2024), https://arxiv.org/abs/2402.14318
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. ACL, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
6. Dhuliawala, Shehzaad, e.a.: Chain-of-verification reduces hallucination in large language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the ACL 2024. pp. 3563–3578. ACL, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.212
7. Gao, Luyu, e.a.: RARR: Researching and revising what language models say, using language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st ACL (Volume 1: Long Papers). pp. 16477–16508. ACL, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.910
8. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 EMNLP. pp. 6465–6488. ACL, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.398
9. Honovich, O., et al.: TRUE: Re-evaluating factual consistency evaluation. In: Feng, S., Wan, H., Yuan, C., Yu, H. (eds.) Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering. pp. 161–175. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.dialdoc-1.19, https://aclanthology.org/2022.dialdoc-1.19

10. Huang, C., Wu, Z., Hu, Y., Wang, W.: Training language models to generate text with citations via fine-grained rewards. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2926–2949. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.161
11. Kenneth Enevoldsen, e.a.: Mmteb: Massive multilingual text embedding benchmark (2025), https://arxiv.org/abs/2502.13595
12. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
13. Li, S., Stenzel, L., Eickhoff, C., Bahrainian, S.A.: Enhancing retrieval-augmented generation: A study of best practices. arXiv preprint arXiv:2501.07391 (2025)
14. Menick, J., Trebacz, M., Mikulik, V., Aslanides, J., Song, F., Chadwick, M., Glaese, M., Young, S., Campbell-Gillingham, L., Irving, G., McAleese, N.: Teaching language models to support answers with verified quotes (2022), https://arxiv.org/abs/2203.11147
15. Okulska, I.: Team up! cohesive text summarization scoring sentence coalitions. In: Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II. p. 388–399. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-61534-5_35
16. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 EMNLP-IJCNLP. pp. 3982–3992. ACL, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410
17. Saba Sturua, e.a.: jina-embeddings-v3: Multilingual embeddings with task lora (2024), https://arxiv.org/abs/2409.10173
18. Tahaei, Marzieh, e.a.: Efficient citer: Tuning large language models for enhanced answer quality and verification. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the NAACL 2024. pp. 4443–4450. ACL, Mexico City, Mexico (Jun 2024). https://doi.org/10.18653/v1/2024.findings-naacl.277
19. Wang, J., Sun, Q., Li, X., Gao, M.: Boosting language models reasoning with chain-of-knowledge prompting. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd ACL (Volume 1: Long Papers). pp. 4958–4981. ACL, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.271
20. Wojtasik, K., Wołowiec, K., Shishkin, V., Janz, A., Piasecki, M.: BEIR-PL: Zero shot information retrieval benchmark for the Polish language. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 LREC-COLING 2024. pp. 2149–2160. ELRA and ICCL, Torino, Italia (May 2024)
21. Yu, Wenhao, e.a.: Chain-of-note: Enhancing robustness in retrieval-augmented language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 EMNLP. pp. 14672–14685. ACL, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.813
22. Zhang, Xin, e.a.: mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In: Dernoncourt, F., Preoţiuc-Pietro, D., Shimorina, A. (eds.) Proceedings of the 2024 EMNLP: Industry Track. pp. 1393–1412. ACL, Miami, Florida, US (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-industry.103

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_21