# PandaChat-RAG: Towards the Benchmark for Slovenian RAG Applications

Taja Kuzman
Tanja Pavleska
{taja,tanja}@pc7.io
PC7, d.o.o.
Ljubljana, Slovenia
Jožef Stefan Institute
Ljubljana, Slovenia

Urban Rupnik
Primož Cigoj
{urban,primoz}@pc7.io
PC7, d.o.o.
Ljubljana, Slovenia

## Abstract

Retrieval-augmented generation (RAG) is a recent method for enriching the large language models' text generation abilities with external knowledge through document retrieval. Due to its high usefulness for various applications, it already powers multiple products. However, despite the widespread adoption, there is a notable lack of evaluation benchmarks for RAG systems, particularly for less-resourced languages. This paper introduces the PandaChat-RAG – the first Slovenian RAG benchmark established on a newly developed test dataset. The test dataset is based on the semi-automatic extraction of authentic questions and answers from a genre-annotated web corpus. The methodology for the test dataset construction can be efficiently applied to any of the comparable corpora in numerous European languages. The test dataset is used to assess the RAG system's performance in retrieving relevant sources essential for providing accurate answers to the given questions. The evaluation involves comparing the performance of eight open- and closed-source embedding models, and investigating how the retrieval performance is influenced by factors such as the document chunk size and the number of retrieved sources. These findings contribute to establishing the guidelines for optimal RAG system configurations not only for Slovenian, but also for other languages.

## Keywords

retrieval-augmented generation, RAG, embedding models, large language models, LLMs, benchmark, Slovenian

## 1 Introduction

The advent of large language models (LLMs) has introduced significant advancements in the field of natural language processing (NLP). Although LLMs have shown impressive capabilities in generating coherent text, they are prone to hallucinations [7, 16], i.e., providing false information. Furthermore, they are dependent on static and potentially outdated corpora [9]. Retrieval-augmented generation (RAG) is a method devised to address these challenges by augmenting LLMs with external information retrieved from a provided document collection. Connecting LLMs with a relevant database improves the factual accuracy and temporal relevance of the generated responses. Moreover, RAG contributes to the explainability of the generated answers by providing verifiable sources, which facilitates the evaluation of the system's accuracy [2]. These advantages have spurred quick adoption of RAG systems across various applications. For instance, PandaChat[1] leverages RAG to provide explainable responses with high accuracy in Slovenian and other languages, integrated in customer service bots and platforms that allow LLM-based retrieval of information from texts.

Although RAG benchmarking is a relatively recent endeavor, some initial frameworks have already emerged [3, 5, 7]. However, these benchmarks are only limited to English and Chinese, leaving a gap in the evaluation of RAG systems for other languages. To address this gap, we make the following contributions:

- We present the first benchmark for RAG systems for the Slovenian language. The benchmark is based on the newly developed PandaChat-RAG-sl test dataset[2], which comprises authentic questions, answers and source texts.
- We introduce a methodology for an efficient semi-automated development of RAG test datasets that is easily replicable for the languages included in the MaCoCu [1] and CLASSLA-web corpora collections [10], which include all South Slavic languages, Albanian, Catalan, Greek, Icelandic, Maltese, Ukrainian and Turkish.
- As the first step of RAG evaluation, we evaluate the retriever's performance in terms of its ability to provide relevant sources crucial to retrieve accurate answers to the posed questions. The evaluation encompasses comparison of performance of several open- and closed-source embedding models. Furthermore, we provide insights on the impact of the document chunk size and the number of retrieved sources, to identify optimal configurations of the indexing and retrieval components for robust and accurate retrieval.

The paper is organized as follows: in Section 2, we provide an introduction to the previous research concerning the evaluation of RAG systems; Section 3 introduces the PandaChat-RAG-sl dataset (Section 3.1) and the RAG system architecture (Section 3.2), which is evaluated in Section 4. Finally, in Section 5, we conclude the paper with a discussion of the main findings and suggestions for future work.

## 2 Related Work

Despite the recent introduction of the RAG architecture, several benchmarking initiatives have already emerged [3, 5, 7, 15]. However, since the RAG systems can be applied to various end tasks, the benchmarks focus on different aspects of these systems. Inter


*Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia*

https://doi.org/10.70314/is.2024.scai.538

[1] https://pandachat.ai/
[2] The PandaChat-RAG benchmark and its test dataset are openly available at https://github.com/TajaKuzman/pandachat-rag-benchmark.

alia, current benchmarks assess their performance in text citation [7], text continuation, question-answering with support of external knowledge, hallucination modification, and multi-document summarization [12].

The closest to our work is the evaluation of the RAG systems on the task of Attributable Question Answering [2]. This task involves providing a question as input to the system, which then generates both an answer and an attribution, indicating the source text on which the answer is based. The advantage of this task over the closed-book question-answering task is that it also measures the system's capability to provide the correct source.

The majority of RAG benchmarks assess RAG systems in English [3, 5, 7, 15] or Chinese [5, 12]. Consequently, the generalizability of their findings to other languages remains uncertain. Furthermore, a limitation of many benchmarks is their reliance on synthetic data generated by LLMs [5, 12, 15]. To avoid potential biases introduced by LLMs and to better represent the complexity and diversity of real-world language use, a more reliable evaluation would be based on non-synthetic test datasets. Despite focusing on a different task, recent research [6] has shown that resource-efficient development of non-synthetic and non-machine translated question-answering datasets is feasible by leveraging the availability of general web corpora and genre classifiers.

## 3 Methodology

### 3.1 PandaChat-RAG-sl Dataset

The PandaChat-RAG-sl dataset comprises questions, answers, and the corresponding source texts that encompass the answers. It was created through a semi-automated process involving the extraction of texts from the Slovenian web corpus CLASSLA-web.sl 1.0 [11], followed by a manual extraction of high-quality instances. Since the texts were automatically extracted from a general text collection, the dataset encompasses a diverse range of topics that were not predefined or decided upon.

The CLASSLA-web.sl 1.0 corpus is a collection of texts, collected from the web in 2021 and 2022 [10]. It was chosen due to its numerous advantages: 1) it has high-quality content, with the majority of texts meeting the criteria for publishable quality [17]; 2) it is one of the largest and most up-to-date collections of Slovenian texts, comprising approximately 4 million texts; 3) the texts are enriched with genre labels, facilitating genre-based text selection; and 4) it is developed in the same manner as 6 other CLASSLA-web corpora [10] and 7 additional MaCoCu web corpora in various European languages [1]. This enables easy expansion of the benchmark to other languages, including all South Slavic languages and various European languages, such as Albanian, Catalan, Greek, Icelandic, Ukrainian and Turkish.

The development of the PandaChat-RAG-sl dataset involves the following steps: 1) the genre-based selection of texts from the CLASSLA-web.sl corpus; 2) the extraction of texts that comprise paragraphs ending with a question (80,215 texts); 3) the extraction of questions and answers (paragraphs, following the question); 4) a manual review process to identify high-quality instances. In the genre-based selection phase, we extract texts labeled with genres that are most likely to contain objective questions and answers, that is, *Information/Explanation*, *Instruction* and *Legal*.

In its present iteration, the dataset consists of 206 instances derived from the first 1,800 extracted texts. It is important to note that this effort can easily be continued with further manual inspection of the extracted texts should there be a need to prepare a larger dataset.

**Table 1: Statistics for the PandaChat-RAG-sl dataset.**

| | Number |
|---|---|
| Instances | 206 |
| Unique texts | 160 |
| Words (questions) | 1,184 |
| Words (texts w/o questions) | 83,467 |
| Total words (questions + texts) | 84,651 |

Table 1 provides the statistical overview for the PandaChat-RAG-sl dataset. The dataset consists of 206 instances, that is, triplets of a question, an answer and a source text, derived from 160 texts. The total size of the dataset is 84,651 words, encompassing both the questions and the texts containing the answers.

### 3.2 RAG System

The RAG pipeline encompasses three main components: indexing, retrieval, and text generation. During the indexing phase, the user-provided text collection is transformed into a database of numerical vectors (embeddings) to facilitate document retrieval by the retriever. This process involves segmenting the documents into fixed-length chunks, which are then converted into embeddings using large language models. The choice of the embedding model and the chunk size are critical factors that can significantly impact the retrieval performance of the model. Selecting an appropriate embedding model is essential to ensure that the textual information is converted into a meaningful numerical representation for effective retrieval. Moreover, the chunk size, in terms of the number of tokens, plays a crucial role in determining the informativeness of the embeddings. Incorrect chunk sizes may lead to numerical vectors that lack important information necessary for connecting the question to the corresponding text chunk, thereby compromising retrieval accuracy [12].

When presented with a question, the retrieval component uses the semantic search (also known as dense retrieval) to retrieve the most relevant text chunks. The search is based on determining the smallest cosine distance between the chunk vectors and the question vector. Lastly, during the text generation phase, the retriever provides the large language model (LLM) with a selection of top retrieved sources. The LLM is prompted to provide a human-like answer to the provided question based on the retrieved text sources. The selection of an appropriate number of top retrieved sources is crucial in this phase: including more than just one retrieved source may enhance retrieval accuracy and address situations where the first retrieved source fails to encompass all relevant information, especially in the case when more texts cover the same subject matter. However, increasing the number of sources also leads to a longer prompt provided to the LLM, potentially increasing the costs of using the RAG system.

In this study, we assess the indexing and retrieval components, focusing on the impact of different embedding models, chunk sizes, and the number of retrieved sources on retrieval performance.

*Embedding Models.* The evaluation includes a range of multilingual open-source and closed-source models. The selection of open-source models is based on the Massive Text Embedding

Benchmark (MTEB) Leaderboard[3] [13]. Specifically, we choose medium-sized multilingual models with up to 600 million parameters that have demonstrated strong performance on Polish and Russian – Slavic languages that are linguistically related to Slovenian. The models used in the evaluation are:

- Closed-source embedding models provided by the OpenAI: an older model text-embedding-ada-002 (OpenAI-Ada) [8], and two recently published models: text-embedding-3-small (OpenAI-3-small), and text-embedding-3-large (OpenAI-3-large) [14].
- Open-source embedding models, available on the Hugging Face repository: BGE-M3 model [4], base-sized mGTE model (mGTE-base) [19], and small (mE5-small), base (mE5-base) and large sizes (mE5-large) of the Multilingual E5 model [18].

*Chunk size.* The impact of the chunk size on retrieval performance is assessed by varying chunk sizes of 128, 256, 512, and 1024 tokens, with a default chunk overlap of 20 tokens. In these experiments, the performance is evaluated based on the first retrieved source.

*Number of retrieved sources.* Previous work indicates that increasing the number of retrieved sources improves the retrieval accuracy [12]. In this study, we examine the retrieval accuracy of embedding models, with a chunk size set to 128 tokens, when the models retrieve 1 to 5 sources. In this scenario, if any of the multiple retrieved sources matches the correct source, the output is evaluated as being correct.

The retrieval capabilities of the RAG system are evaluated on the task of Attributed Question-Answering. The evaluation is based on accuracy, measured as the percentage of questions correctly matched with the relevant source.

The experiments are performed using the LlamaIndex library[4]. The chunk size is defined using the SentenceSplitter method in the indexing phase. Number of retrieved sources (*similarity top k*), the embedding model and the prompt for the LLM model are specified as parameters of the chat engine. The closed-source embedding models are used via the OpenAI API, while the experiments with the open-source models are conducted on a GPU machine.

## 4 Experiments and Results

In this section, we present the results of the experiments examining the impact of the chunk size, the number of retrieved sources, and the selection of the embedding model on the retrieval performance of the RAG system.

### 4.1 Chunk Size

Figure 1 shows the impact of the chunk size on the retrieval performance of the RAG systems that are based on different embedding models. The findings suggest that, with the exception of the OpenAI-Ada model, all systems demonstrate the best performance when the text chunk size is set to 128 tokens. Increasing the chunk size hinders the retrieval performance, which is consistent with previous research [12]. These results confirm that smaller chunk sizes enable the embedding models to capture finer details that are essential for retrieving the most relevant text for the given question.

[3]https://huggingface.co/spaces/mteb/leaderboard
[4]https://www.llamaindex.ai/

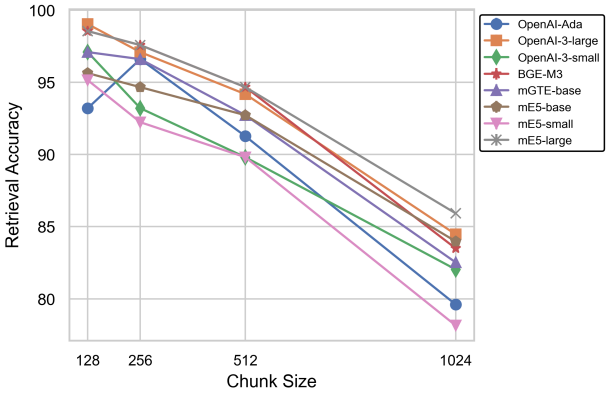


**Figure 1: The impact of the chunk size on the retrieval performance.**

### 4.2 Number of Retrieved Sources

Figure 2 shows the performance of the RAG systems when increasing the number of retrieved sources. The results demonstrate that increasing the number of retrieved sources initially improves the performance, however, after a certain threshold, the performance levels off.

Increasing the number of retrieved sources results in larger inputs to the LLM in the text generation component, incurring higher costs. Using more than two retrieved sources does not significantly improve results in most systems. What is more, with the top two retrieved sources, certain embedding models, namely, BGE-M3 and mE5-large, already reach perfect accuracy. Thus, our findings indicate that using more than the top two retrieved sources is unnecessary.

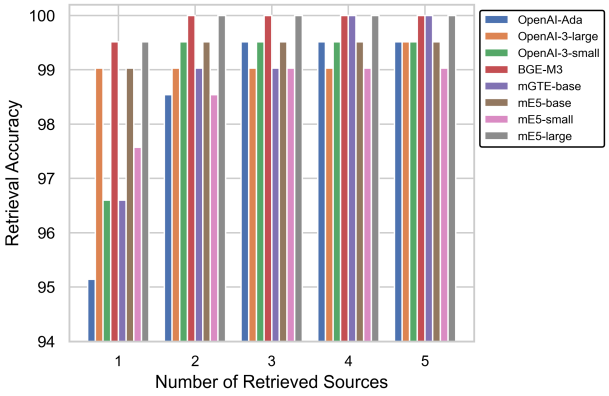


**Figure 2: Impact of the number of retrieved sources on the retrieval performance.**

### 4.3 Embedding Models

We provide the final comparison of the performance of systems that use different embedding models. We use the parameters that have shown to provide the best results in the previous experiments: the chunk size of 128 tokens and top two retrieved sources. As shown in Table 2, the retrieval systems that use the open-source BGE-M3 and mE5-large embedding models achieve the perfect retrieval score. They are closely followed by the closed-source OpenAI-3-small and the mE5-base models which achieve

**Table 2: Performance comparison between the open-source and closed-source embedding models.**

| embedding model | speed (s) | retrieval accuracy |
|---|---|---|
| BGE-M3 | 0.58 | 100 |
| mE5-large | 0.58 | 100 |
| OpenAI-3-small | 0.69 | 99.51 |
| mE5-base | 0.29 | 99.51 |
| OpenAI-3-large | 1.19 | 99.03 |
| mGTE-base | 0.31 | 99.03 |
| OpenAI-Ada | 0.63 | 98.54 |
| mE5-small | 0.15 | 98.54 |

accuracy of 99.5%. While having slightly lower scores, all other retrieval systems still achieve high performance, ranging between 98.5% and 99% in accuracy.

Additionally, Table 2 provides the inference speed of the models measured in seconds per instance. If inference speed is a priority, the mE5-base model emerges as the optimal selection, as it yields high retrieval accuracy of 99.51% and is two times faster than the two best performing models. In cases where users are restricted to closed-source models due to the unavailability of GPU resources, the OpenAI-3-small model stands out as the most suitable option. Its inference speed is comparable to the OpenAI-Ada model, while it achieves a superior retrieval accuracy.

## 5 Conclusion and Future Work

In this paper, a novel test dataset was introduced to assess the performance of the RAG system on Slovenian language. A general methodology for efficient creating of non-synthetic RAG test datasets was established that can be extended to other languages. We evaluated the retrieval accuracy of the RAG system, examining the impact of the embedding models, the document chunk size, and the number of retrieved sources. The assessment of embedding models encompassed eight open-source and closed-source LLM models. It revealed that open-source models, specifically, BGE-M3 and mE5-large, reached perfect retrieval accuracy, demonstrating their suitability for RAG applications on Slovenian texts. Furthermore, the evaluation of the optimal chunk size and the number of retrieved sources showed that smaller chunk sizes yielded superior results. In contrast, increasing the number of retrieved sources enhanced results up to a certain threshold, beyond which the model performance plateaued. Certain models already achieved perfect accuracy when evaluated based on the top two retrieved sources.

While the novel test dataset can be used to evaluate all the components of the RAG system, in this paper, we focused on the evaluation of the indexing and retrieval components. In our future work, we will extend the evaluations to the text generation component with regard to fluency, correctness, and usefulness of the generated answers. Furthermore, we plan to expand the benchmark to encompass a wider range of languages. The plans include extending the dataset and evaluation to South Slavic languages and other European languages that are covered by comparable MaCoCu [1] and CLASSLA-web [10] corpora.

## References


[1] Marta Bañón et al. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: Focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Ghent, Belgium, (June 2022), 303–304. https://aclanthology.org/2022.eamt-1.41.

[2] Bernd Bohnet et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

[3] Shuyang Cao and Lu Wang. 2024. Verifiable Generation with Subsentence-Level Fine-Grained Citations. *arXiv preprint arXiv:2406.06125*.

[4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. (2024). arXiv: 2402.03216 [cs.CL].

[5] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* number 16. Vol. 38, 17754–17762.

[6] Anni Eskelinen, Amanda Myntti, Erik Henriksson, Sampo Pyysalo, and Veronika Laippala. 2024. Building Question-Answer Data Using Web Register Identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors. ELRA and ICCL, Torino, Italia, (May 2024), 2595–2611. https://aclanthology.org/2024.lrec-main.234.

[7] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488.

[8] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. New and improved embedding model. https://openai.com/index/new-and-improved-embedding-model/. [Accessed 26-08-2024]. (2022).

[9] Angeliki Lazaridou et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34, 29348–29363.

[10] Nikola Ljubešić and Taja Kuzman. 2024. CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3271–3282.

[11] Nikola Ljubešić, Peter Rupnik, and Taja Kuzman. 2024. Slovenian web corpus CLASSLA-web.sl 1.0. In Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1882.

[12] Yuanjie Lyu et al. 2024. CRUD-RAG: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. *arXiv preprint arXiv:2401.17043*.

[13] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2014–2037.

[14] OpenAI. 2024. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates/. [Accessed 26-08-2024]. (2024).

[15] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 338–354.

[16] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784–3803.

[17] Rik van Noord, Taja Kuzman, Peter Rupnik, Nikola Ljubešić, Miquel Esplà-Gomis, Gema Ramírez-Sánchez, and Antonio Toral. 2024. Do Language Models Care about Text Quality? Evaluating Web-Crawled Corpora across 11 Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5221–5234.

[18] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

[19] Xin Zhang et al. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. (2024). https://arxiv.org/abs/2407.19669 arXiv: 2407.19669 [cs.CL].