Retrieval-Augmented Generation implementation for Wikipedia API

Creating specific knowledge assistant, Winter 2024

Warsaw University of Technology

Stanisław Kurzatkowski
01150709@pw.edu.pl

Filip Mieszkowski
01171165@pw.edu.pl

Jakub Szypuła
01142147@pw.edu.pl

1 Main goals of the project

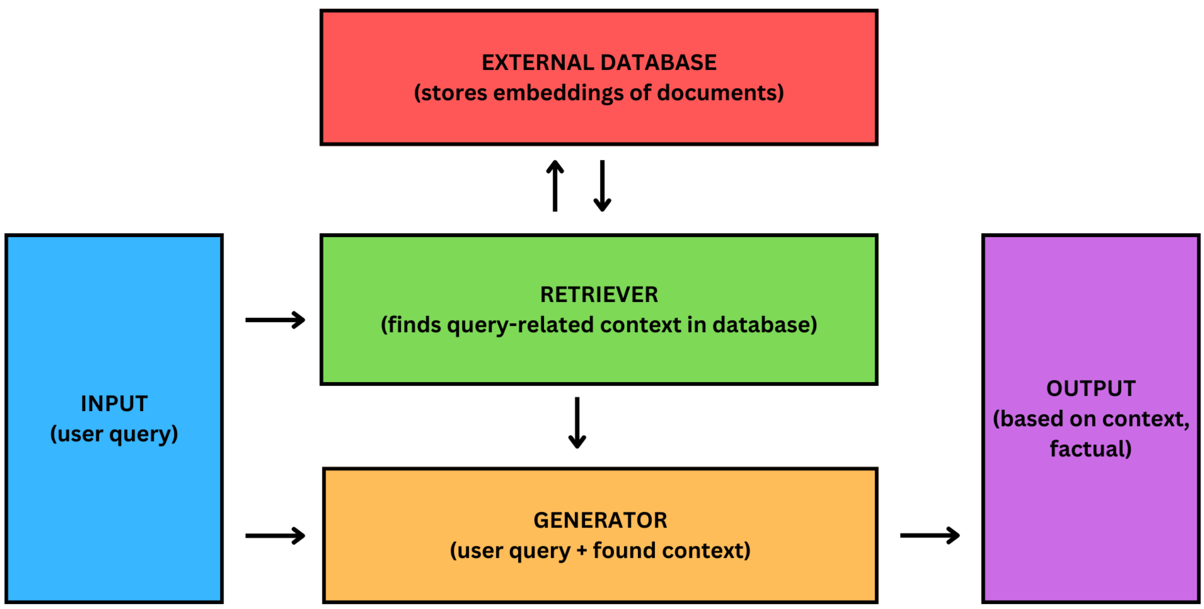
Since the beginning of the boom of the Large Language Models (LLMs) in the NLP field, the halucination of the models has been a major concern. Also, even if the models are able to provide true information, they are not able to provide the source of the information. Retrival-Augmented Generation (RAG) is a solution to both of these problems. The main goal of this project is to provide a specific knowledge assistant using pre-trained LLMs (like ChatGPT or ollama) and RAG architecture. The main issues that we are going to address are:

- Maximizing the factuality of the models answers
- The ability to provide the information even if the model was not trained on data describing the events of interest
- Extracting as much valuable information as possible from the database (including tables, not only plain text)

2 Brief explanation of the RAG architecture

2.1 The main idea

The main idea of the RAG architecture is to use a retriever to find the relevant information from the knowledge base and then use the generator to generate the answer. The knowledge base is usually a vector database, which can be quickly searched using the retriever. After the information is found it is passed to the generator together with the question. This way both the question and possibly the answer are passed to the LLM, which can generate the answer, even if it does not know the answer on its own.



2.2 Project approach

In the project the Wikipedia API was used as the knowledge base. Base on the Wikipedia articles a vector database is created. As the encoding model we decided to use $ada - 002$ by OpenAI. We store the embeddings in the $faiss$ index, which is a library for efficient similarity search and clustering of dense vectors. We decided not to store the articles in the index. Rather, the metadata about the articles is stored. This approach allows for creating vast databases without the need of storing the actual Wikipedia articles physically on any local machine. Therefore, the potential user is not bothered by any memory limits and can freely create the database which they need.

3 The topic of the knowledge base

Selecting the topic of the knowledge base is a crucial part of the project. The topic should be broad enough to allow for creating a massive database, which should include as many Wikipedia articles as possible. On the other hand, the topic should be specific and possibly not too broad, to allow for the retriever to find the relevant information.

3.1 Star Wars

In the first version of the project, it was decided to use Star Wars universe as the topic of the knowledge base. The topic is broad enough to allow for creating a database with thousands of records. Unfortunately, RAG architecture wasn't working significantly better than the LLMs on their own. This was caused by the fact that the used LLM was already trained on the data and in many cases could provide sufficient answer without the RAG enforcement.

3.2 New Topics

To perform the validation of the RAG architecture, it was decided to move to a topic about which the generator ($llama3.1 : 8b$) knows nothing or just a little. The $llama3.1 : 8b$ has its knowledge cutoff in March 2023. Therefore, the following topics were selected as the ones staying outside the LLM training data and also broadly described on Wikipedia.

- Russian policy in 2023 - 2024 (focusing on elections, summits, government, ...etc.)
- Summer Olympics 2024

4 Creating the database

The process of creating the external database for RAG involves two major steps - finding the titles of the articles related to the topics specified by the user, extracting the data from selected articles in the form of text and tables.

4.1 Article selection process

The selection process exploits the fact that every Wikipedia article belongs to some "Wikipedia Categories". These categories create a structure, which starts with the most general ones. Mentioned "general categories" are parent nodes from their "Wikipedia subcategories". This structure follows, allowing for "Wikipedia subcategories" to have their own subcategories. To get the articles, a recursive search is performed which visits all the subcategories of the parent node, which must be provided by the user, and saves the titles of the articles. Due to the fact, that the general category does not have to correspond to the topic user is interested in, it is possible to provide multiple categories. In case when the user wants to narrow the range of topics, the list of keywords can be provided. If needed, both can be used together, multiple parent categories and keyword filtering.

4.2 Data extraction and preparation

With the titles of the articles selected, the database creation starts. All articles are scanned, section by section, with only one section at the time being in the memory of the local machine. The metadata for the section is created, which is the python dictionary object storing the: page title, section title, subsection title or subsubsection title. The text data is chunked by length, with the proper overlap between the chunks to prevent cutting the crucial information "in half". Additionally, every chunk of text coming from the section is merged with the page title, section title, subsection title ... etc. This solution assures that the chunk of text is contextualized as much as possible (especially when the title contains a date and the body of the section in the description of the events that took place at the time). Each chunk is embedded into a vector and with the corresponding metadata object states, a pair representing the original text based on which it was created.

Wikipedia articles bring a lot of information, but not only in the form of plain text but also tables. These tables often contain valuable information, absent in text. Therefore, the chunking process and title contextualizing is also performed for them. Every chunk of the table has a constant number of rows and consists of:

- context for the retriever to associate the table data with the user question: merged titles of page section and subsection where the table was found
- context for the generator: "Start of the table data."
- context for the retriever and the generator: first row of the table (it always consists of keywords and allows the generator to interpret the body of the table)
- information: body of the table, rows of table with values separated with tabulation and rows with the newline
- context for the generator: "End of the table data."

A chunk of the table is embedded and a corresponding metadata object is created.

5 Results

The validation phase was mainly focused on determining whether the following objectives were achieved in the project and if so, how:

- The ability to provide the answer based on the new information from the database
- The ability to admit that the information is not known
- The comparison of the answer with the answer provided by the LLM on its own
- Time needed to provide the answer
- high factuality of the answer

Results are presented in the Tables 1 and 2.

5.1 Time needed to provide the answer

The time needed to provide the answer is crucial in the real-time applications. As presented in the Figure 1, the most time-consuming part is the generation of the answer by the LLM. This might be due to the long model input caused by the query - fusion used in this project. In this example the number of retrieved documents is equal to 5.

6 Discussion of the results

Questions were asked in two separate runs, each with the database corresponding to the topic of the questions. As can be seen in the tables, the overall RAG performance is satisfactory if only the information is in the database. The model is able to provide the information even if it does not know the answer on its own. The time needed to provide the answer is acceptable, as can be seen in the table. The most time-consuming part is the generation of the answer by the LLM, which is not dependent on the RAG architecture.

7 Conclusions

The RAG architecture is a promising solution to the halucination of the LLMs, as well as improving the factuality of the answer. The architecture is working correctly, if the information is in the database.

Validation process was mostly performed on human - generated questions. The procedure came down to randomly selecting the article from the database and by viewing it making two to three questions. This was caused by the fact that the database does not contain every information about a given subject, so generating the lists of questions via larger LLMs with access to the internet often led to asking about the things from beyond the database. If one wants to perform the validation automatically, on a large scale, the following approaches are proposed

- ARES (Automated RAG Evaluation System) [11]
- VERA (Validation and Evaluation of Retrieval-Augmented Systems) [8]
- RAGAS (Retrieval Augmented Generation Assessment) [4]
- (authors) validation based on random section selecting from random article within the database and using another instance of the model to generate the question based on it. Metadata of a section could be stored to serve as an answer.

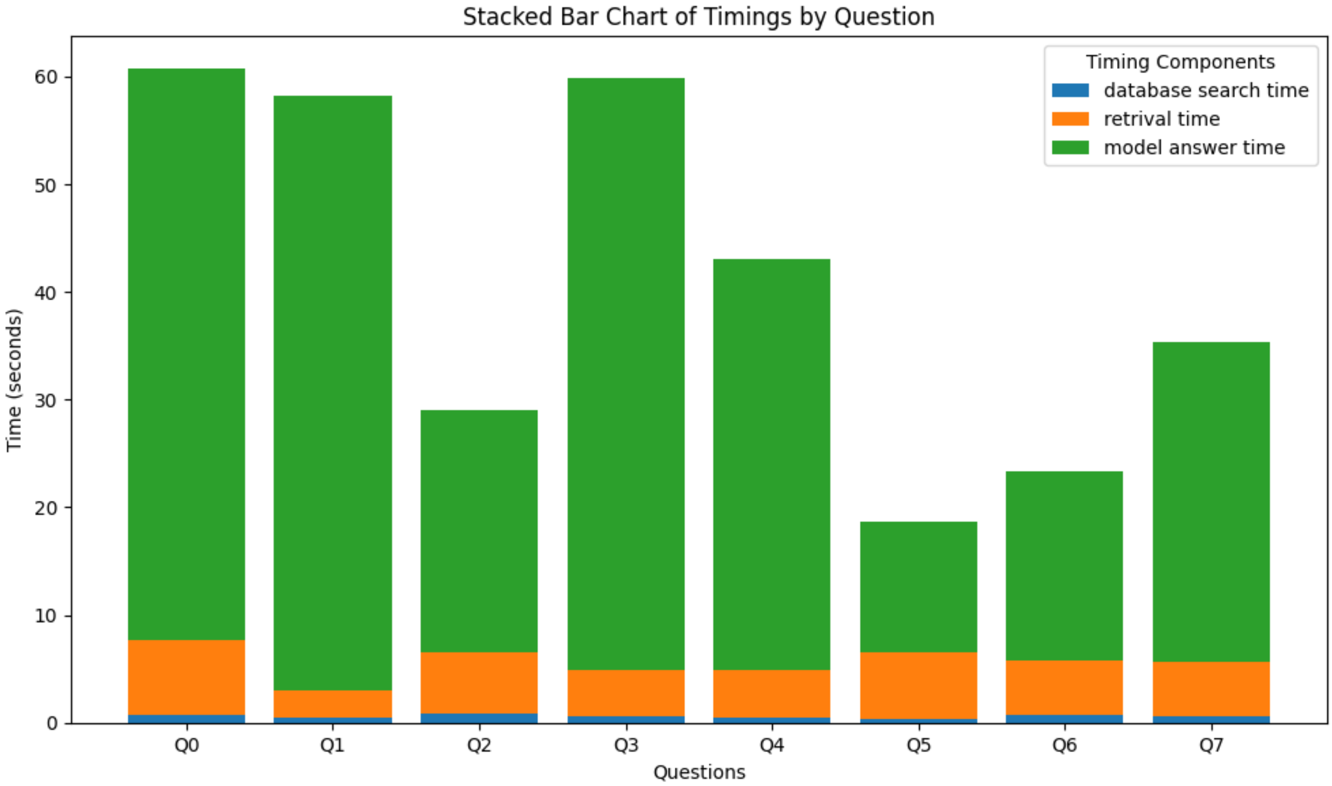


Figure 1. Graph representing the contributions to the time consumption of different actions

Question	Answer with RAG	answer without RAG
When, where and how has Yevgeny Prigozhin died?	According to the context, Yevgeny Prigozhin, a Russian oligarch, mercenary, Wagner Group co-founder and leader, died on August 23, 2023. The plane he was traveling in crashed near Kuzhenkino, Bologovsky District, Tver Oblast. The circumstances of his death are disputed, with the Wagner Group attributing it to being shot down by air defenses, while the US government and aviation experts suggest that the plane was downed by a bomb onboard or other sabotage.	I don't have information about the death of Yevgeny Prigozhin.
Who is the Minister of Defence of Russia?	The defence minister mentioned in the context is Sergei Shoigu.	The current Minister of Defense of Russia is Sergei Shoigu. He has been serving in this position since November 6, 2012. However, it's worth noting that in March 2020, President Vladimir Putin appointed General Valery Gerasimov as the Chief of the General Staff and First Deputy Minister of Defense, making him effectively the second-in-command of the Russian military.
Who won 2024 Russian presidential election?	According to the context, Vladimir Putin won the 2024 Russian presidential election with 87.28% of all votes cast (including invalid ballots), receiving a total of 76,277,708 votes or 88.48%.	I don't have any information about the outcome of a 2024 Russian presidential election. The information I can access is current up to 01 March 2023, and there's no available data on such an event. For more recent or updated information, please check with reliable news sources or official announcements from relevant authorities.
Who won 2024 USA presidential election?	The context provided does not mention the results of the 2024 USA presidential election. It only discusses the 2024 Russian presidential election and Vladimir Putin's victory in that election. I do not have any information about the outcome of the 2024 US Presidential Election based on this context.	I don't have the ability to predict the future or know information that hasn't yet occurred. The 2024 United States presidential election has not taken place, and I do not have any information on who may win it. If you're looking for information on past elections or general information about presidential elections in the US, I'd be happy to help with that!

Table 1: Results of the questions

<table>
<tr><th>Question</th><th>Answer with RAG</th><th>Correct answer</th></tr>
<tr><td>What was the key outcome of the Russia-Africa Summit held in July 2023?</td><td>According to the context provided, many agreements were signed with participating African countries on July 28, 2023. The summit concluded with a final declaration and an official plan for implementing a Partnership Forum for 2023-2026, along with other documents. Therefore, the key outcome of the Russia-Africa Summit held in July 2023 was the signing of multiple agreements between Russia and African countries, as well as the adoption of a final declaration and a partnership framework for the next few years.</td><td>The summit aimed to strengthen ties between Russia and African countries, focusing on economic cooperation, trade, and military support.</td></tr>
<tr><td>How many teams will participate in the men's football tournament at the 2024 Summer Olympics, and how is qualification achieved?</td><td>A total of 16 teams will participate in the men's football tournament at the 2024 Summer Olympics. Qualification is achieved as follows:
* The host nation, France, earned automatic qualification.
* From the six continental confederations: + AFC (Asian Football Confederation): 3 automatic slots + 1 additional slot through a play-off match against CAF
+ CAF (Africa): 3 automatic slots + 1 additional slot through a play-off match against AFC
+ UEFA (Union of European Football Associations): 3 slots
+ CONCACAF (Confederation of North, Central America and Caribbean Association Football) and CONMEBOL (South American Football Confederation): 2 slots each
+ OFC (Oceania Football Confederation): 1 slot
+ CONMEBOL : 2 slot</td><td>The men's football tournament at the 2024 Summer Olympics in Paris will feature 16 teams.
France automatically qualifies as the host nation. Each regional confederation organizes qualification tournaments or designates existing competitions to determine its representatives. The allocation is as follows:
+ AFC (Asia) 3
+ CAF (Africa) 3
+ CONCACAF (North and Central America, Caribbean) 2
+ CONMEBOL (South America) 2
+ OFC (Oceania) 1
+ UEFA (Europe) 3
+ AFC-CAF play-off 1</td></tr>
<tr><td>How many medals in total did Italy acquire in swimming competitions in 2024 Olympics? How many of them were gold?</td><td>Italy acquired a total of 6 medals in swimming competitions at the 2024 Olympics.
As for how many gold medals, Italy acquired, according to the table data provided:
Rank Nation Gold Silver Bronze Total
7 Italy 2 1 3 6
So, Italy won 2 gold medals.</td><td>
Medal summary [edit]
Medal table [edit]
* Host nation (France)
<table>
<tr><th>Rank</th><th>Nation</th><th>Gold</th><th>Silver</th><th>Bronze</th><th>Total</th></tr>
<tr><td>1</td><td>United States</td><td>8</td><td>13</td><td>7</td><td>28</td></tr>
<tr><td>2</td><td>Australia</td><td>7</td><td>9</td><td>3</td><td>19</td></tr>
<tr><td>3</td><td>France*</td><td>4</td><td>1</td><td>2</td><td>7</td></tr>
<tr><td>4</td><td>Canada</td><td>3</td><td>2</td><td>3</td><td>8</td></tr>
<tr><td>5</td><td>Hungary</td><td>3</td><td>1</td><td>1</td><td>5</td></tr>
<tr><td>6</td><td>China</td><td>2</td><td>3</td><td>7</td><td>12</td></tr>
<tr><td>7</td><td>Italy</td><td>2</td><td>1</td><td>3</td><td>6</td></tr>
<tr><td>8</td><td>Sweden</td><td>2</td><td>0</td><td>0</td><td>2</td></tr>
<tr><td>9</td><td>Great Britain</td><td>1</td><td>4</td><td>0</td><td>5</td></tr>
<tr><td>10</td><td>Germany</td><td>1</td><td>1</td><td>1</td><td>3</td></tr>
<tr><td>11</td><td>South Africa</td><td>1</td><td>1</td><td>0</td><td>2</td></tr>
</table></td></tr>
<tr><td>Who won the gold medal in Taekwondo in Women's 67kg category?</td><td>Viviana Márton from Hungary won the gold medal in the Taekwondo Women's 67 kg category at the 2024 Summer Olympics. She defeated Aleksandra Perišić of Serbia with a score of 7-4 in the gold medal match.</td><td>
Results [edit]
Gold medal match [edit]
<table>
<tr><th colspan="5">Gold medal match</th></tr>
<tr><td>12</td><td>Viviana Márton (HUN)</td><td>7</td><td>4</td><td></td></tr>
<tr><td>3</td><td>Aleksandra Perišić (SRB)</td><td>1</td><td>2</td><td></td></tr>
</table></td></tr>
</table>

Table 2: Results of the questions

References

[1] Sebastian Borgeaud et al. "Improving Language Models by Retrieving from Trillions of Tokens". In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 2206–2240. URL: https://proceedings.mlr.press/v162/borgeaud22a.html.

[2] Sebastian Borgeaud et al. "Improving language models by retrieving from trillions of tokens". In: *CoRR* abs/2112.04426 (2021). arXiv: 2112.04426. URL: https://arxiv.org/abs/2112.04426.

[3] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. "Reciprocal rank fusion outperforms condorcet and individual rank learning methods". In: *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2009.

[4] A. Gan, L. Zhu, and P. Sharma. "RAGAS: Metrics for Reference-Free Evaluation of Retrieval-Augmented Generation Pipelines". In: *arXiv preprint arXiv:2309.15217* (2023). URL: https://arxiv.org/abs/2309.15217.

[5] Kelvin Guu et al. "REALM: retrieval-augmented language model pre-training". In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org, 2020.

[6] Yikun Han, Chunjiang Liu, and Pengfei Wang. *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge*. 2023. arXiv: 2310.11703 [cs.DB]. URL: https://arxiv.org/abs/2310.11703.

[7] Gautier Izacard and Edouard Grave. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". In: *CoRR* abs/2007.01282 (2020). arXiv: 2007.01282. URL: https://arxiv.org/abs/2007.01282.

[8] J. Kang, R. Lopez, and S. Patel. "VERA: A Framework for Validation and Evaluation of Retrieval-Augmented Systems Using Cross-Encoder Mechanisms". In: *arXiv preprint arXiv:2409.03759* (2024). URL: https://arxiv.org/pdf/2409.03759.

[9] Urvashi Khandelwal et al. *Generalization through Memorization: Nearest Neighbor Language Models*. 2020. arXiv: 1911.00172 [cs.CL]. URL: https://arxiv.org/abs/1911.00172.

[10] OpenAI. "GPT-4 Technical Report". In: *CoRR* abs/2303.08774 (2023). DOI: 10.48550/ARXIV.2303.08774. arXiv: 2303.08774. URL: https://doi.org/10.48550/arXiv.2303.08774.

[11] A. Rammohan, R. Pradeep, and J. Ganesh. "ARES: Automated RAG Evaluation System for Context Relevance, Answer Faithfulness, and Answer Relevance". In: *arXiv preprint arXiv:2311.09476* (2023). URL: https://arxiv.org/abs/2311.09476.

[12] Machel Reid and Nikolay Savinov. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *ArXiv* abs/2403.05530 (2024). URL: https://api.semanticscholar.org/CorpusID:268297180.

[13] Hanan Samet. "K-Nearest Neighbor Finding Using MaxNearestDist". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.2 (2008), pp. 243–252. DOI: 10.1109/TPAMI.2007.1182.

[14] Shangyu Wu et al. *Retrieval-Augmented Generation for Natural Language Processing: A Survey*. 2024. arXiv: 2407.13193 [cs.CL]. URL: https://arxiv.org/abs/2407.13193.

[15] Shangyu Wu et al. "Retrieval-Augmented Generation for Natural Language Processing: A Survey". In: (July 2024). DOI: 10.48550/arXiv.2407.13193.