

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/381263242>


A RAG-based Medical Assistant Especially for Infectious Diseases

Conference Paper · April 2024
DOI: 10.1109/CICT60155.2024.10544639

CITATIONS
0

READS
98


6 authors, including:



Stewart Kirubakaran
Karunya University

48 PUBLICATIONS 171 CITATIONS


SEE PROFILE



Mahimai Raja J
Karunya University

4 PUBLICATIONS 0 CITATIONS

SEE PROFILE



Ruban Gino Singh Arul Peppin Raj
University of Alberta

1 PUBLICATION 0 CITATIONS

SEE PROFILE

A RAG-based Medical Assistant Especially for Infectious Diseases

Stewart Kirubakaran S
Division of CSE
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
stewartkirubakaran@gmail.com

Jasper Wilsie Kathrine G
Division of CSE
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
kathrine@karunya.edu

Grace Mary Kanaga E
Division of DSCS
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
grace@karunya.edu

Mahimai Raja J
Division of CSE
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
mahimairaja@karunya.edu.in

Ruban Gino Singh A
Division of CSE
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
rubangino@karunya.edu.in

Yuvaraajan E
Division of CSE
Karunya Institute of Technology and Sciences
Coimbatore, Tamil Nadu, India
yuvaraajane@karunya.edu.in

Abstract— Infectious diseases like COVID-19 have gained international attention recently. Furthermore, there are significantly fewer doctors per capita in densely populated nations like India, which hurts those in need. Under such circumstances, natural language processing techniques might make it feasible to create an intelligent and engaging chatbot system. The primary objective of the effort is to develop an interactive solution that is entirely open source and can be easily installed on a local computer using the most recent data. Even though there are numerous chatbots on the market, proposed solutions highlight the need to provide individualized and sympathetic responses. Getting Back While the data is stored in the graph database as nodes and relationships, and the knowledge graph is constructed on top of it, augmented generation is utilized to extract the pertinent content from the data. To improve the generator's context, pertinent sections are collected during the question-answering process. This reduces hallucinations and increases the correctness of abstractions by providing external knowledge streams. Furthermore, the research study employs a text-to-speech model that was replicated from a physician's voice recording to narrate the produced responses, thereby augmenting user confidence and interaction. Academic institutions and healthcare organizations can benefit from this work by better understanding the value and effectiveness of applying NLP techniques to infectious disease research.

Keywords— natural language processing, chatbot, COVID-19, large language model, retrieval augmented generation, knowledge graph

I. INTRODUCTION

The SARS-CoV-2 virus is the cause of the COVID-19 pandemic, which has killed over a million people globally. The World Health Organization (WHO) announced the death toll on March 11, 2020. As oxygen levels drop due to the virus, oxygen requirement has increased by 6-8% per day in India for the treatment of hospitalized COVID-19 patients. This paper demonstrates [1] an AI chatbot to help with communication and offer answers for COVID-19 prevention and treatment. The chatbot can interpret user inquiries and respond appropriately with the use of Natural Language Processing (NLP), Retrieval Augmented

Generation, and Text-To-Speech. As WHO's Facebook Messenger chatbot that dispels COVID-19 misinformation demonstrates, chatbots are operational around the clock and have important uses in healthcare. According to the statistic report by Bhavin Jankharia in the article published by Times of India (2022), the doctor to people ratio will be at a quite low rate that is 0.76 in the next decade as illustrated in fig. 1. This motivates the development of innovative methods, such as intelligent chatbots, to manage challenging scenarios like a pandemic.

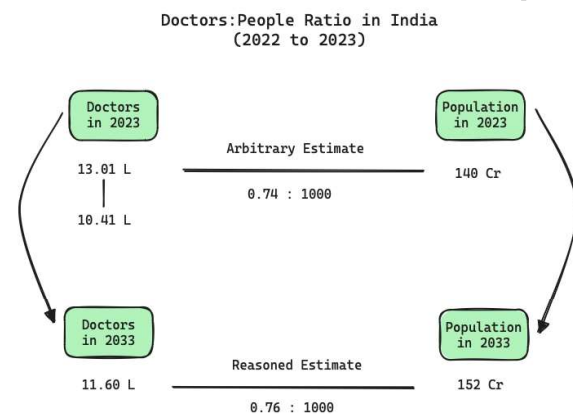


Fig. 1. Doctor to Population Ratio in India

The built chatbot uses artificial intelligence (AI) to interact effectively to modify behavior for COVID-19 prevention and health promotion and the works compares the performance of the RAG application using three different retrievers and analyzed using the Retrieval Augmented Generation Assessments (RAGAs).

A) Data Collection

The subset of AI known as NLP is concerned with how computers and human language interact. Enabling robots to comprehend, interpret, and produce language like that of humans in a meaningful and contextually relevant

manner is the main objective of NLP. This multidisciplinary field uses cognitive psychology, computer science, and linguistics to create models and algorithms that can process and understand data in natural language. NLP is used in a wide range of applications, such as speech recognition, chatbots, sentiment analysis, and language translation. It is an essential tool for improving human-computer interaction and bridging the gap between human language and machine understanding.

B) Chatbot

A chatbot is a computer program that mimics human conversation and offers an interactive communication interface as depicted in Fig 2. Chatbots, which make use of AI and NLP technology, can comprehend user input and respond in a way that mimics human communication. These virtual assistants are used on a variety of platforms, including social media, messaging applications, and websites, to carry out particular activities, respond to user inquiries, and have real-time conversations with users. Chatbots come in a variety of forms, from complex machine-learning models that constantly learn from interactions to rule-based systems with preprogrammed responses. Chatbots are incredibly useful tools that are being used extensively in customer service, automation, and information retrieval. They improve user experiences and streamline communication procedures in a variety of industries.

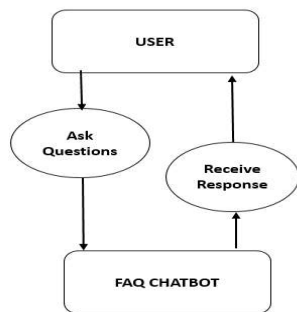


Fig. 2. Chatbot Block Diagram

C) Covid-19

The new coronavirus SARS-CoV-2 that produced the COVID-19 pandemic became a global health emergency that had a significant effect on communities all over the world. The virus emerged in late 2019 and spread quickly, causing societal unrest, severe disease, and previously unheard-of public health issues. To slow the spread of the virus and safeguard public health, governments, healthcare institutions, and communities launched several initiatives, such as lockdowns, social distancing, and vaccination drives. In addition to its direct effects on health, COVID-19 has affected many facets of daily life, including work habits, educational attainment, and international economic dynamics.

II. RELATED WORKS

Research on NLP-based chatbots explores various methodologies to enhance conversational user experience, including optimizations to improve comprehension, coherence, factual grounding, and context-alignment, though reducing perplexity on open-domain dialogues remains an active challenge. S. Narynov et al. (2021) [1] developed a psychological chatbot using Rasa and NLU to provide Kazakh-speaking patients with accessible mental health support, achieving 72% response accuracy in initial experiments however, enhancing understanding and pertinence in human-computer conversations remains an open challenge. M. Ganesan et al. (2020) [2] conducted a seminal survey on various platforms, design techniques, and NLP aspects that drive the performance of AI chatbots in providing services by learning from past experiences and training data, through avenues to boost natural language understanding through reduced perplexity on open-domain conversations remain underexplored. J. Skrebeca et al. (2021) [3] discuss increasingly indispensable chatbots during COVID-19 supporting 24/7 customer service for e-commerce and personalized education using AI, though accurately comprehending varied user intents and responding coherently remains constrained without strides in language model enhancements. Machine Learning and NLP emerged as a powerful tool for building a reliant chatbot to select appropriate responses by matching input statements, with studies by H. Koundinya et al. (2021) [4]. A. Sunithanandhini et al. (2023) [5] developed a medical chatbot using machine learning and NLP to assist patients by responding to health inquiries, though evaluating its ability to comprehend diverse real-world symptomatology and provide coherent clinical guidance on open-domain dialogues.

Aiming to enhance health access during COVID-19, S.Chakraborty et al. (2020) [6] leverage deep learning to build a medical chatbot providing multi-pronged pandemic support via spreading disease awareness and facilitating prevention guidelines, though optimizing consumer adoption warrants assessing naturalistic fluency and coherence with quantitative dialogue metrics yet unexplored. Surveying Generative Pretrained Transformer (GPT) medical chatbots, P. Nandini et al. (2023) [7] demonstrate the potential to enhance patient healthcare access but outline current comprehension limitations absent large-scale evaluations, underscoring future work tailoring transformer architectures with clinical corpora to drive perplexity reductions yielding accuracy gains. Leveraging AI and NLP, K. Anjum et al. (2023) [8] developed MedBot - a medical chatbot for affordable preliminary diagnosis and health advice to minimize unnecessary hospital visits, though quantitative evaluations on the accuracy and coherence of open-domain dialogues remain imperative to optimize reliability and adoption. Surveying AI chatbots delivering affordable personalized healthcare, A. Wahal et al. [9] (2022) highlight potential benefits in accessibility albeit acknowledge lingering diagnostic accuracy constraints without large-scale evaluations, presenting opportunities to advance safety and efficacy through future optimized neural architectures designed leveraging standardized medical corpora. Proposing a modular chatbot

framework for adaptable disease management assistance, S. Montagna et al. (2023) [10] outline backends quantifying patient adherence and frontends providing motivational notifications, though optimizing personalized engagement warrants mitigating vocabulary gaps to enhance comprehension. This body of work collectively underscores chatbots developed using NLP and generative pre-trained transformers exhibit multifaceted capabilities across diverse use cases and applications.

III. METHODOLOGY

The Paper describes an efficient approach using an NLP-based chatbot solution based on RAG models with TTS in 7 steps as illustrated in Fig 3. Large Language Model and Retrieval Augmented Generation are utilized by the chatbot to interpret user input and provide pertinent responses. RAG models are composed of two components: (i) a retriever $p\eta(z|x)$ with parameters η that returns the top K results from the knowledge graph based on the given query x and (ii) a generator $p\theta(y_i|x, z, y_{1:i-1})$ that generates a current token based on a context of the previous $i-1$ tokens, the original input x and a retrieved passage z . Moreover, here the context stored as the knowledge graph acts as the latent space. A graph database makes it possible to retrieve pertinent content more quickly based on user queries. Before analysis, the voice recognition module transcribes speech to text. To improve user experience and believability, the chatbot uses a synthetic voice that is modeled after a doctor's speech patterns to vocalize its textual responses. By synchronizing the spoken responses with the text displayed on the screen, this voice synthesis provides information in two useful ways.

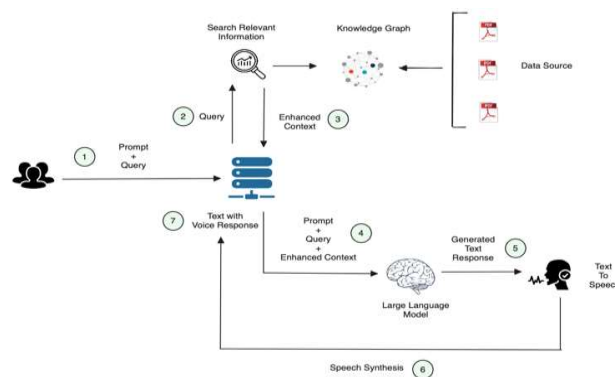


Fig. 3. Architecture Diagram

A) Quantized LLM

More extensibility and interoperability across computing environments and versions are the goals of GGUF quantized models. GGUF supports several model types and removes breaking changes between versions. In the project, the conversational chatbot is powered by a GGUF quantized model. A big language model can be effectively deployed on low-power CPUs and GPUs for scalable and accessible usage by using this quantized GGUF format. The GGUF's unification and extensibility make it

easier to integrate the potent pre-trained model into the chatbot system, enabling it to perform natural language creation and processing.

B) Graph Database

A knowledge graph embeds business logic directly into the data structure by representing important entities in a domain and their interactions. This allows for more sophisticated querying and analysis by tying data to context. Complex pattern-based searches, such as locating all entities connected to a specific node, are made possible by the graph structure. The extracted knowledge graph is securely stored in a graph database for further retrieval. The information is more relevant when it is arranged relationally in a graph database for medical data. By querying relevant items and relationships, the knowledge graph provides the chatbot with domain-specific intelligence that enables it to comprehend user requests and formulate informative responses. Its graph structure makes context-based retrieval more effective.

C) Retrieval Augmented Generation

To increase response accuracy, retrieval augmented generation (RAG) blends external information sources with massive language models. Without requiring the LLM to be retrained, RAG ingests real-time factual data into the context window to provide an efficient response. To supplement the limited knowledge of the LLM, pertinent context is collected from this database upon inquiry receipt. After that, the LLM produces a response based on particular, up-to-date evidence. This method goes beyond the generalization of the LLM to deliver contextual solutions. The solution proposed leverages RAG to add facts that are gathered from COVID-19 documents to the chatbot's Knowledge Base. Pulling pertinent sections from the graph database enhances the background information when answering queries from users. RAG gives the chatbot access to current medical knowledge. In essence, RAG enables ongoing knowledge updates for the chatbot without requiring expensive LLM retraining. The graph database can easily add more documents as new COVID-19 data becomes available, enhancing the chatbot's topic expertise. Using RAG approaches makes it possible to utilize the large language model's reasoning ability as well as the dynamic, all-inclusive factual knowledge that is relationally stored. Because of this synergy, the chatbot can give more complex explanations because it bases its answers on up-to-date data that is taken from the knowledge graph. For the chatbot to be able to store current, contextualized, and reliable COVID-19 information, RAG is essential.

D) Text To Speech Synthesis

XTTS, an open multilingual text-to-speech model created by Coqui supports real-time voice cloning and attains production-level speech quality. With just a short sample of a target voice, XTTS can synthesize natural-sounding speech in 13 languages—with more to follow. One of XTTS's unique features is the ability to perform cross-

language voice cloning, which preserves the vocal identity of a voice sample in one language while synthesizing speech in another. The proposed idea uses XTTS to create a vocal response for the chatbot by cloning a doctor's voice from a recording. The user experience is improved since the simulated doctor's voice used for the audio responses is trustworthy and lifelike, matching the written output. The dual-modality engagement of the chatbot through text and cloned speech is powered by XTTS's high-quality and versatile multilingual voice cloning.

IV. RESULTS AND DISCUSSION

A) Data Collection

The raw data used in the project is in the form of PDF or text documents from Elsevier, the New York Times, and various trusted sources. These documents cover a wide range of topics relevant to the infectious disease particularly, COVID-19. To extract the relevant data from these documents, both automated and manual techniques are used. This rigorous hybrid data collection process combines the scalability of automation with the discernment of human analysis. This allows the project to leverage large volumes of data while ensuring relevance, accuracy, and transparency. The resulting curated data serves as a solid foundation for deriving actionable insights to build a reliable chat solution.

B) Data Preparation

The ingested data undergoes pre-processing within an ETL (extract, transform, load) pipeline to structure it for downstream analytics. Text extraction and tokenization modules leverage NLP to parse textual documents into machine-readable tokens. An enterprise-grade large language model (LLM), GPT-4, synthesizes tokens into contextualized data to deduce entities and semantic relationships. A Complex knowledge graph is engineered as illustrated in Fig 4 to represent probabilistic correlations between identified entities and concepts.

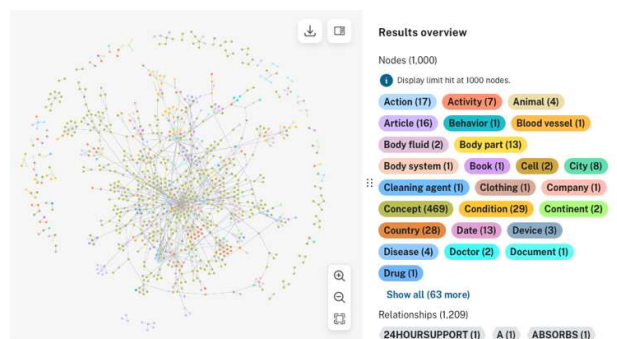


Fig. 4. Knowledge Graph

These rich graphs are then fed into a horizontally scalable graph database. By storing the data in the format nodes and edges we could significantly reduce the memory usage and retrieving from a graph database is more efficient than other databases.

C) RAG Integration

A 4-bit quantized version of Mistral-7B LLM, leverages innovations like block floating point formatting to reduce memory requirements 8x with minimal accuracy loss. Its modular architecture uses separate retriever and generator modules to enable fast, accurate responses. The retriever rapidly indexes knowledge graphs to ground responses in factual data. Integrated with an enterprise chatbot, this Mistral RAG pipeline achieves high reliability and contextual accuracy while optimizing latency, cost, and perplexity - unlocking new potential for more efficient conversational AI.

In this paper, three types of retrieval techniques were effectively demonstrated; naïve RAG, auto-merging retriever-based RAG, and ensemble retriever-based RAG. Naïve RAG consists of a simple retrieval and a synthesis pipeline. However, advanced RAG techniques like Auto-Merging retriever-based RAGs differ in their retrieval part, where the longer documents are split into even smaller chunks and later, they make the retrieval more accurate to fetch the relevant documents. The proposed work uses the Reciprocal Rank Fusion (RRF) algorithm an unsupervised ranking method, underneath the working of the ensemble RAG. The RRF score is computed using capsin

$$RRFscore(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)},$$

and, based on the RRF Score the results were reranked. By combining the advantages of several algorithms, the Ensemble Retriever can outperform most of the single algorithms.

D) Testing and Validation

After the RAG pipelines were built, the synthetic testing data along with the ground truth were prepared using GPT-4 carefully. Using the prepared evaluation dataset with carefully crafted headers with question, answer, context, and ground truth, the work evaluated the systems built using the Retrieval Augmented Generation Assessments (RAGAs) technique. Here, the work uses faithfulness, context precision, context recall, answer relevancy, answer similarity, and answer correctness as a metric to evaluate the built RAG systems. As the evaluation's findings are shown in the fig. 5

Retrieval Augmented Generation - Evaluation

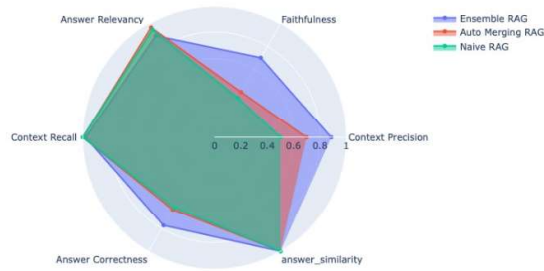


Fig. 5. Evaluation of Various Retrievers

Ensemble-based retriever demonstrated a higher answer correctness of 0.775 on the synthetic evaluation dataset. With a context precision score of 0.697, the auto-merging retriever-based RAG system outperforms the naïve RAG system in terms of comprehension of the context and generating solutions as illustrated in Fig. 5.

Ensemble retriever-based RAG and auto-merging retriever-based RAG exhibited commendable performance with comparatively higher scores than naïve RAG as depicted in Table 1.

Table 1. Evaluation Results

| Architecture | Evaluation Metrics | | |
|------------------|--------------------|------------------|--------------------|
| | Faithfulness | Answer Relevancy | Answer Correctness |
| Ensemble RAG | 0.70 | 0.8918 | 0.7750 |
| Auto-Merging RAG | 0.40 | 0.9634 | 0.6353 |
| Naïve RAG | 0.35 | 0.9439 | 0.6166 |

^a Note: The ground truth of the data is synthetically generated.

Based on the obtained evaluation metric it is seen clearly that the ensemble RAG is performing better than auto-merging RAG and naïve RAG.

V. CONCLUSION

Proposed research shows how conversational agents driven by AI may help the general public receive important health information. Utilizing advancements in NLP, knowledge representation, and speech synthesis, proposed COVID-19 chatbot offers a user-friendly, multilingual interface for obtaining reliable medical advice. Inside, a sizable language model combined with retrieval-enhanced generation produces well-informed responses based on up-to-date scientific information gleaned from reliable sources. Context-specific reasoning is strengthened by relationally organizing this knowledge in a graph database. Moreover, responding to a cloned doctor's voice creates an impression of professionalism and empathy. The chatbot may have natural discussions with users to teach them about COVID-19 prevention, diagnosis, and treatment through text and speech. This open-source remedy could support overworked healthcare facilities and assist isolated

areas. In the future, improvements such as multimodal features may increase user involvement even more. This work highlights the enormous potential for accessible, tailored chatbots to benefit public health by assembling cutting-edge AI to battle misinformation and connect individuals to reliable COVID-19 information.

VI. FUTURE WORK

Subsequent research should improve the chatbot system's scalability and adaptability in various linguistic and cultural situations, to guarantee its efficaciousness in catering to a wider audience. Its application could be further increased by adding real-time data integration techniques to improve the chatbot's reactivity to changing infectious illness scenarios and continuously update the knowledge graph. To further enhance the system, investigating methods to incorporate user feedback loops and improve the model of sympathetic response production would be beneficial. Also, adding multi-modal chat could leverage the performance.

REFERENCES

- [1.] S. Narynov, Z. Zhumanov, A. Kumar, M. Khassanova and B. Omarov, "Development of Chatbot Psychologist Applying Natural Language Understanding Techniques," 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, Republic of, 2021, pp. 636-641, doi: 10.23919/ICCAS52745.2021.9649825.
- [2.] M. Ganesan, D. C., H. B., K. A. S. and L. B., "A Survey on Chatbots Using Artificial Intelligence," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262366.
- [3.] J. Skrebeca, P. Kalniete, J. Goldbergs, L. Pitkevica, D. Tihomirova and A. Romanovs, "Modern Development Trends of Chatbots Using Artificial Intelligence (AI)," 2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS), Riga, Latvia, 2021, pp. 1-6, doi: 10.1109/ITMS52826.2021.9615258.
- [4.] H. K. K., A. K. Palakurthi, V. Putnala and A. Kumar K., "Smart College Chatbot using ML and Python," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020.9262426.
- [5.] S. A. AP, C. M, B. D, K. T., and M. J., "Advanced Chatbots for Home Patients using AI," 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2023, pp. 838-845, doi: 10.1109/ICOEI56765.2023.10125886.
- [6.] S. Chakraborty et al., "An AI-Based Medical Chatbot Model for Infectious Disease Prediction," in IEEE Access, vol. 10, pp. 128469-128483, 2022, doi: 10.1109/ACCESS.2022.3227208.
- [7.] N. P. K. S., S. S. T. T. N., Y. Yuvraaj and V. D. A., "Conversational Chatbot Builder – Smarter Virtual Assistance with Domain Specific AI," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-4, doi: 10.1109/INCET57972.2023.10170114.
- [8.] K. Anjum, M. Sameer and S. Kumar, "AI Enabled NLP based Text to Text Medical Chatbot," 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 2023, pp. 1-5, doi: 10.1109/ICIPTM57143.2023.10117966.
- [9.] Wahal, M. Aggarwal and T. Poongodi, "IoT based Chatbots using NLP and SVM Algorithms," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp.

- 484-489, doi: [12.] Karan A et al. Hum Resour Health. 2021 Mar 22;19(1):39
10.1109/ICIEEM54221.2022.9853095.
- [10.] S. Montagna, S. Mariani and M. F. Pengo, "A Chatbot-based Recommendation Framework for Hypertensive Patients," 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 2023, pp. 730-733, doi: 10.1109/CBMS58004.2023.00309.
- [11.] Es, S., James, J., Espinosa-Anke, L., and Schockaert, S., "RAGAS: Automated Evaluation of Retrieval Augmented Generation", arXiv e-print, 2023. doi:10.48550/arXiv.2309.15217.