Zurich University
of Applied Sciences

School of
Engineering

Specialisation Project 2 / Vertiefungsarbeit 2 (VT2)

ZHAW School of Engineering

CAI, Centre for Artificial Intelligence

Evaluation of Retrieval Augmented Generation Systems utilising Large Language Models

---

Author:
Samuel Stucki

Supervisor:
Prof. Dr. Mark Cieliebak

Secondary Supervisor:
Dr. Jan Deriu

January 31, 2024

Abstract

This thesis aimed to establish a Retrieval Augmented Generation (RAG) pipeline including a comprehensive performance evaluation tailored for a domain-specific corpus. The primary objective was to empower users to evaluate their respective datasets. The contributions of this thesis are threefold: Firstly, the creation of a small corpus comprising 153 QA pairs extracted from the video game Baldur's Gate 3 (BG3), acquired from a YouTube Let's Plays series. Secondly, the execution of various experiments at different stages of the RAG pipeline, highlighting potential enhancements and identifying potential pitfalls. Lastly, an End-to-End RAG evaluation pipeline, allowing users to apply this methodology to their data. This pipeline leverages Haystack for constructing the pipeline and Ragas for evaluation. Notably, Ragas metrics, except for "Context Relevancy," demonstrated that they can accurately quantify failures even without labelled data, a feat not achievable using conventional metrics. The experimental results revealed that, for the BG3 corpus, the most effective pipelines employed a hybrid search approach and an asymmetric search model using a chunk size of approximately 128 and a TopK of 5. These configurations achieved answer rates of 92.16% and 83%, along with average BERTScores of 0.898 and 0.896, respectively.

Preface

This thesis was supported by Prof. Dr. Mark Cieliebak and Dr. Jan Deriu of the Centre for Artificial Intelligence (CAI) at the Zurich University for Applied Sciences (ZHAW). I would like to thank them for their invaluable input, without which this thesis would not have been possible. I would also like to thank the CAI itself for providing me access to their OpenAI API paid account, covering the costs of the experiments in the process.

Contents

Chapter 1

Introduction

Over the past decade, the landscape of Artificial Intelligence (AI) has undergone a transformative shift, primarily fueled by substantial increases in computational power and the availability of extensive datasets. This evolution is characterized by a departure from traditional statistical methods towards the widespread adoption of Neural Networks (NN), especially Deep Neural Networks (DNN). This transition has significantly enhanced the capabilities of AI systems in various domains. Notably, in the sub-field of Natural Language Processing (NLP), these advancements have played a pivotal role, empowering machines to better understand and interpret the intricacies of human language in both written text and spoken speech.

Now with the advent of the Transformer architecture proposed by Vaswani et al. [1] in 2017 and the subsequent releases of Large Language Models (LLM) like BERT [2], GPT-3.5 "ChatGPT" or GPT-4 [3] by OpenAI, and LLaMA 2 [4] by Meta AI, the research community possesses the potential to attain results that were once considered beyond reach. However, LLMs are prone to so-called hallucination errors during text generation.[5] This refers to statements or complete texts which contain misleading or incorrect information, often observed (but not limited to) in translation tasks and mathematical calculations. While research is ongoing on the exact causes and mitigations of hallucinations, some general drivers of the errors are insufficient knowledge about the world, lack of reasoning, and overconfidence. The latter two drivers can be sufficiently controlled using prompt engineering by forcing the model to adhere to specific guidelines and answering patterns. The former can be supplemented by applying so-called Retrieval Augmented Generation (RAG), especially for question answering on, for example, company internal data.

RAG is a novel method developed in 2021 by Lewis et al. [6] that combines a retrieval and a generation system. The retrieval part fetches chunks of documents based on the given query and then passes these along in the prompt to a generation model like GPT-4, which then formulates an answer based on the query and the retrieved context. This approach allows models to use knowledge outside their original training corpora. It even enables them to work with domain-specific data that it would otherwise not be able to answer, e.g. critical company data inaccessible outside its network.

1.1 Motivation

With various technologies working inside the RAG pipeline, the source of potential errors in the final answers is challenging to pinpoint. This sparked the idea of an evaluation pipeline designed to assist users in determining whether applying RAG to their data is possible. Furthermore, conducting a series of experiments at various stages in the pipeline should provide users with insights into how the system can be effectively evaluated.

During discussions, it was discovered that many corpora used for question-answer evaluations were released before 2021 and might have served as training data for LLMs. As most companies do not disclose details about their data collection and application processes, we had to presume this to be the case. Operating under this assumption, any assessments based on these corpora would inevitably require a degree of caution, given that LLM-generated answers could potentially be influenced by memory, introducing bias. This prompted a subsequent discussion on identifying suitable data for evaluation corpora, during which dialogues in video games were identified as a viable alternative.

In August 2023, the award-winning game Baldur's Gate 3 (BG3) was released by Larian Studios[1] that provided ample source material to generate a small dataset of questions and answers related to specific concepts such as quests, locations, lore, and more. Notably, the content was not utilised as training data for LLMs due to its later launch date compared to most well-known LLMs. Since Larian Studios does not offer direct access to its dialogues, an alternative approach was adopted, involving the use of YouTube (YT) transcripts from users playing the game in a Let's Play format. This approach allowed for the extraction of additional information about the game, as users often explained ongoing events, described landscapes, shared their thoughts, and provided additional insights into the available choices.

When it comes to evaluation, the prevailing approach [7] involves utilising high-performing LLMs, commonly referred to as judges, for assessing the outputs of other LLMs, as traditional metrics have shown to be unsuitable in this task.[8][9] This is accomplished by establishing metrics that are designed to offer precise insights into particular aspects of the pipeline, which the judge subsequently quantifies using text prompts. A small number of libraries are now available of which we used Ragas [10] due to its simple application.

The contributions of this thesis are thus threefold: a small dataset containing QA-pairs based on the video game Baldur's Gate 3 and sourced from a YT Let's Plays series (available on HuggingFace[2]), an End-to-End RAG evaluation pipeline for users to apply to their data, applying Haystack [11] for the RAG pipeline and Ragas for evaluation (available on GitHub[3]), and lastly several experiments conducted on

[1] https://larian.com/
[2] https://huggingface.co/datasets/stucksam/BG3-QA-Dataset
[3] https://github.com/stucksam/vt2-rag-eval

different points in the pipeline to showcase potential improvements and pitfalls.

1.2 Literature Review

Retrieval Augmented Generation (RAG) has emerged as a prevalent strategy [12][6], empowering Large Language Models to operate effectively with domain-specific data, ground their generation on factual information [13][14], and provide a certain degree of transparency through proper citation of sources [15].

Initial attempts to assess LLM output with LLM-as-a-judge applied the LLMs out of the box. An illustrative case is [7] wherein Zheng et al. examined the utilisation and limitations of LLM-as-a-judge. The evaluation encompassed aspects such as position bias, verbosity, self-enhancement biases, and limited reasoning ability for LLM based chat assistant evaluation. The study then validated the models by comparing their judgments with human evaluations. The findings revealed that judges, exemplified by GPT-4, achieved an agreement rate exceeding 80% on controlled and crowdsourced human preferences, mirroring agreements between humans.

For more knowledge-intensive tasks, specialized datasets like ALCE [16] and Hagrid [17], along with evaluation guidelines such as LongEval[18] for establishing standardized human assessment protocols for long-form summarization, have been proposed. Recent proposals have also introduced libraries such as Ragas [10] and ARES [19] for the automatic evaluation of generated output, crucial to this thesis. Research also advanced in refining specific aspects within the RAG pipeline itself.

Xu et al. [20], affiliated with NVIDIA, conducted a study investigating the impact of context windows in retrieval tasks. Their research revealed that a Large Language Model (LLM)with a 4K context window, employing simple Retrieval Augmented Generation, can achieve performance comparable to a fine-tuned LLM with a 16K context window utilising positional interpolation for long-context tasks. Notably, a retrieval-augmented Llama2-70B with a 32K context window exhibited superior performance compared to both GPT-3.5-turbo-16k and Davinci003, as assessed by the geometric mean of ROUGE scores (specifically, ROUGE-1/2/L) [21] on QMSum [22], F1 scores for Qasper [23] and NarrativeQA [24], and others.

Chen et al. [25] established a benchmark encompassing six distinct LLMs, wherein their performance across four distinct tasks in Retrieval Augmented Generation was assessed using metrics such as accuracy, rejection rate, and error detection rate. Subsequent analysis revealed that, although ChatGPT consistently exhibited superior overall performance as the answer model, all evaluated models encountered notable challenges in tasks related to Negative Rejection, Information Integration, and Counterfactual Robustness.

Asai et al. [26] introduced the Self-RAG framework, designed to enhance the quality and actuality of LLMs through retrieval and self-reflection. This framework involves training an LLM that dynamically retrieves passages as needed, generating and subsequently reflecting upon the retrieved information using specialized tokens known as reflection tokens. These tokens allow for the LLM to be controllable during the inference phase and empower it to adapt its behaviour to diverse task requirements. Self-RAG demonstrated superior performance compared to both ChatGPT and retrieval-augmented Llama2-chat across various tasks, including Open-domain Question Answering, reasoning, and fact verification.

Balaguer et al. [27] conducted a comprehensive analysis at Microsoft, wherein they compared the performance of RAG against fine-tuned models. The research involved the generation of a synthetic QA dataset, derived primarily from high-quality agricultural data that was filtered and validated, applying GPT-4 for question generation and utilising RAG to generate answers. The filtered dataset served as the basis for fine-tuning several smaller LLMs. Evaluation metrics were custom-designed and applied using GPT-4. The results revealed a notable 6% increase in accuracy through fine-tuning. Furthermore, an additional 5% improvement was observed when combining the smaller LLMs with RAG. This proved that combining RAG and fine-tuning yields the best performance. Due to cost considerations, a general recommendation was made to utilise RAG and leverage already fine-tuned LLMs whenever possible. Additionally, the study confirmed GPT-4 as the best-performing judge.

1.3 Outline

This thesis begins with an insightful introduction that addresses the motivation behind the research and incorporates a literature review to establish the context for the study. Following this, the Foundations Chapter 2 delves into the theoretical underpinnings and key concepts essential for understanding the subsequent analyses. Chapter 3 provides a comprehensive exploration of the evaluation dataset, elucidating its generation process and the methodologies employed for quality assessment. The Experimental Setup Chapter 4 outlines the general setup and offers detailed insights into each executed experiment, ensuring transparency and reproducibility. Subsequently, the Results Chapter 5 offers a comprehensive overview of the Ragas metric distributions and presents detailed findings from all conducted experiments. The final Chapter 6, Discussion and Outlook, engages in a critical analysis of the results, drawing connections to the established foundations, and providing a forward-looking perspective on potential future research directions and implications.

Chapter 2

Foundation

2.1 Embeddings

Data occurs in diverse formats, yet machines are inherently only capable of understanding numbers. Images, sounds, or text must first undergo an initial encoding process, transforming them into numerical formats, for them to be usable. This process of converting data into numerical vectors characterizes embeddings. The significance of this encoding lies in its ability to be comparable to other encodings generated by the same embedding model, as proximity in the vector space correlates to semantic similarity [28].

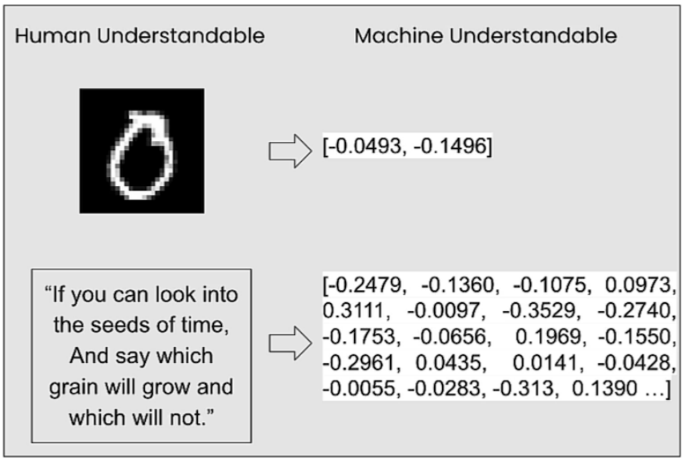

Figure 2.1: Example of encoding of real-world data into embeddings. Figure taken from [29]

2.1.1 Embedding Models

This thesis will apply so-called sentence-transformer[30] pre-trained embedding models, which will be discussed here. The output from BERT[2] and other transformer networks includes an embedding for each token present in an input text. To create a fixed-sized sentence embedding, the model employs mean pooling. This entails taking the average of the output embeddings for all tokens, resulting in a fixed-size vector. The sentences in text form are aligned in a way that sentences conveying similar meanings are positioned closely in vector space. [30]

Two distinctly different models are applied in this thesis, bi-encoder models and cross-encoder models. Bi-encoders generate a sentence embedding for a given sentence, producing embeddings u and v for independently passed sentences A and B through a BERT model. The similarity between these embeddings is assessed using cosine similarity. In contrast, cross-encoders process both sentences simultaneously within a Transformer network, yielding an output value between 0 and 1 that signifies the similarity of the input sentence pair. Notably, a cross-encoder does not generate sentence embeddings, and it does not allow for the individual passage of sentences. [30]

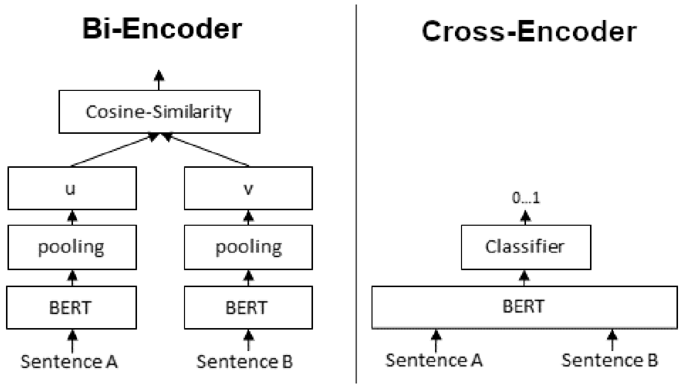

Figure 2.2: Visualization of a bi-encoder and a cross-encoder. Figure taken from sentence-transformer documentation[30].

Cross-encoders are utilised when a predetermined set of sentence pairs needs scoring. On the other hand, bi-encoders, are applied when a sentence embedding in a vector space is required for efficient comparisons, which is the case for Retrieval Augmented Generation (RAG). [30]

2.2 Vector Indices and Databases

Vector indices and databases are purposely crafted for handling high-dimensional vector embeddings representing real-world data. These storage systems play a crucial role in executing semantic similarity searches between query vectors and indexed vectors, facilitating the retrieval of the most suitable data. Traditional scalar-based databases cannot work well with vectors, due to their scale and are thus less applied in the context of RAG. The primary distinction between vector indices and vector databases lies in their support for CRUD operations (create, read, update, delete). While indices are immutable data storages, databases allow for a more fluid approach to their application. The results of the retrieval processes may vary depending on the chosen methodology: Sparse Search (or Keyword Search), Dense

Search (or Semantic Search), or a Hybrid approach combining both. [31]

This thesis applies an index as its storage, given the static nature of the data, rendering frequent updates unnecessary. Real-world applications of RAG are more likely to utilise databases to cater to their dynamic requirements. The decision to employ FAISS by Meta [32] as the storage and search application was motivated by the positive experiences previously made by the team.

2.2.1 Sparse / Keyword Search

Sparse search is the traditional approach to Information Retrieval, with a prominent example being the Best Matching 25 (BM25) algorithm proposed by Stephenson et al. [33]. This algorithm incorporates multiple facets and assesses their relative significance. The first consideration is Term Frequency (TF), which evaluates the frequency of terms or words within a specific document. Rather than a simple summation, the algorithm employs a saturation function to impact excessively high TFs. The second consideration is Inverse Document Frequency (IDF), which evaluates the significance of words within the entire document collection. Less common words are assigned greater weight than their more prevalent counterparts [34]. The formalization of this equation is presented in 2.1.

$$IDF(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5) + 1}\right) \tag{2.1}$$

Thirdly, field-length normalization plays a pivotal role in standardizing TF based on the document's length, thus mitigating bias toward lengthier documents. The fourth factor, Query Term Saturation, denoted by the variable k_1, controls the saturation function for TF. High values of k_1 intensify the saturation function, diminishing the influence of Term Frequency in the process. Lastly, Document Length Saturation, often denoted as b, regulates document length normalization, with smaller b values enhancing the normalization's aggressiveness. [34] The BM25 algorithm is formalized by equation 2.2, where D and Q represent a document and a query, respectively. The variable q_i signifies each term in the query, n denotes the number of terms in the query, $f(q_i, D)$ quantifies the frequency of term q_i in document D, and $avgDl$ refers to the average document length in the collection.

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) * \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgDl})} \tag{2.2}$$

The BM25 algorithm is used in this thesis during one experiment applying a hybrid retrieval setup, which is explained in Section 2.2.3.

2.2.2 Dense / Semantic Search

Dense Search refers to the semantic similarity search previously mentioned in this chapter. This method facilitates multi-modal and multi-lingual searches, improving resilience against typographical errors. Nonetheless, it introduces the potential drawback of overlooking crucial keywords, reducing the quality of results in the process. [35] Various methodologies exist for computing the distance between two vectors with cosine similarity and dot-product being often applied in NLP. Figure 2.3 provides a visual representation of cosine similarity between two vectors.

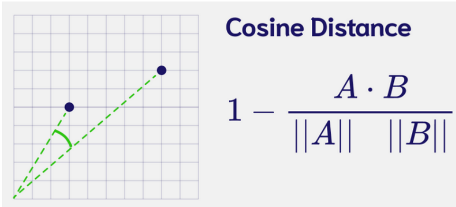

Figure 2.3: Cosine distance calculation between two points. Figure taken from [29].

The second distance measure, dot-product, is illustrated in Figure 2.4.

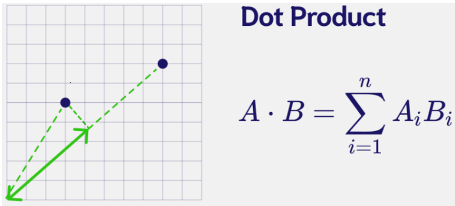

Figure 2.4: Dot-Product calculation between two points. Figure taken from [29].

The thesis predominantly employed cosine similarity, as most experiments leverage sentence-level transformer embedding models. A comparison to dot-product is provided in an experiment, outlined in Chapter 4, to visualize the differences. As with sparse search, there are a multitude of methodologies for performing the semantic search on the indexed embeddings. Within the FAISS storage framework, three options, namely Flat, Hierarchical Navigable Small World (HNSW), and Inverted Index (IVFx) are available. Given the limited size of the evaluation dataset, the decision was made to employ the Flat option.

Symmetric vs. Asymmetric Semantic Search

An essential consideration is the choice between symmetric and asymmetric semantic search. In symmetric semantic search, the query and the entries in the corpus are roughly equivalent in length and content. In symmetric tasks, there is a potential flexibility in swapping the query and the corpus entries. On the other hand, asymmetric semantic search typically involves a concise query, such as a question or keywords, seeking a longer paragraph that answers the query. In asymmetric tasks, swapping the query and the corpus entries often lacks relevance. [30]

Depending on the evaluation dataset, different embedding models are recommended. The evaluation dataset used in this thesis, described in Chapter 3, is of the asymmetric type. However, due to a previous misunderstanding, most experiments (including the benchmark, utilising all-mpnet-base-v2[1]) used a symmetric semantic search model instead of the better-suited asymmetric model. To rectify this error, a comparison was provided in one of the experiments to visualize the impact the model could have had on the RAG results. msmarco-distilbert-base-v4[2] was chosen as the asymmetric semantic search embedding model due to its performance shown on the leaderboard[3] in the sentence-transformer repository.

2.2.3 Hybrid Search

Hybrid search denotes the process of conducting retrieval by employing two or more distinct algorithms to generate a comprehensive search result. In general, this entails the integration of both dense and sparse search methodologies, utilising various approaches, as illustrated by Benham et al. [36]. Haystack offers three distinct strategies: Concatenation of results, Reciprocal Rank Fusion (RRF), and Merging of results [37]. The Concatenation approach involves appending the retrieved documents to a list and subsequently forwarding them through the processing pipeline. This option is frequently employed in conjunction with a re-ranking model to facilitate sorting. RRF reorders the documents based on their shared occurrences in both retrievers, intending to elevate the most relevant documents to the top. The final approach involves merging the documents based on the scores assigned to each by the individual retrievers, as depicted in Figure 2.5.

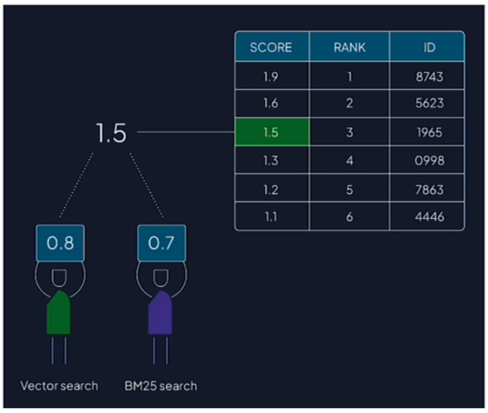

Figure 2.5: Retrieval using a hybrid approach of sparse and dense search. Figure taken from [29].

Hybrid search is applied in one of the experiments performed on the evaluation corpus and is elaborated upon in Section 4.4.

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[2]https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4
[3]https://www.sbert.net/docs/pretrained-models/msmarco-v3.html

2.3 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) has emerged as a promising tool for mitigating LLM hallucination by combining the intrinsic knowledge of the Large Language Models with external knowledge, such as company databases. [38][39] The pipeline consists of three different steps that are visualized in Figure 2.6. The initial step, denoted as *Ingestion* or *Indexing*, involves the collection and cleaning of data in the form of documents, subsequently partitioned into concise segments. These segments or chunks are then subjected to encoding by a model, generating embeddings that are indexed within a database. The advantage of using chunks instead of complete documents is that it enables more granular vector embeddings that result in better search results. Regarding the evaluation dataset, Ingestion would generally be only performed once, however, in a production environment, it is plausible that new documents are added to the database on a more frequent basis. In this thesis, ingestion is applied multiple times, owing to the varied experiments that employ diverse combinations of chunks, vector index properties, and embedding models.

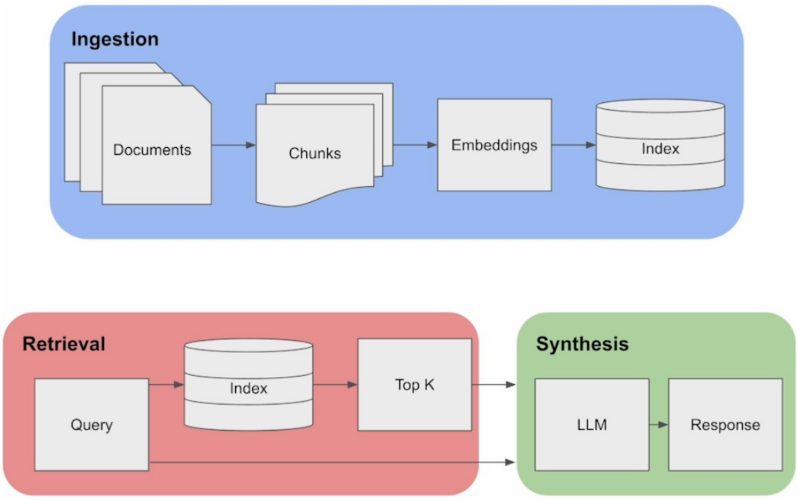

Figure 2.6: RAG pipeline showcasing the three phases of operation. Figure taken from [29].

The second and third steps are generally performed in unison. In the initial phase, known as *Retrieval*, Information Retrieval (IR) is applied to the indexed database, utilising the vector embedding of the user's query as its input. Subsequently, the query vector is employed to compute the similarity between itself and the vectors representing document chunks. Determining the optimal number of chunks to retrieve, denoted as TopK, is pivotal in this stage for the LLM to generate a precise response in the subsequent *Synthesis* or *Generation* step. During Synthesis, the original raw text of the user's query, along with the retrieved document chunks or contexts, is presented to the LLM in a unified prompt. This approach enables the LLM to produce accurate and domain-specific answers that would otherwise be unattainable, as illustrated in Figure 2.7.

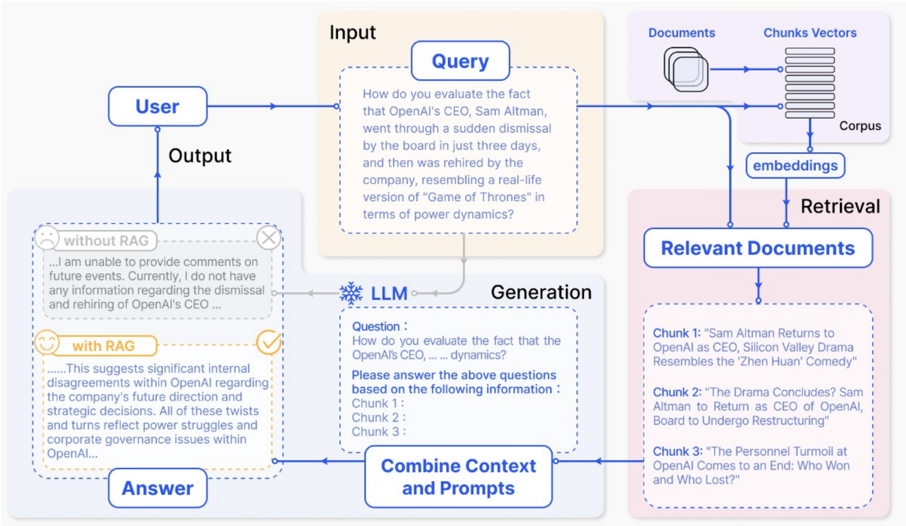

Figure 2.7: RAG pipeline example using actual data. Figure taken from [39].

2.3.1 Haystack

Retrieval Augmented Generation pipelines can be easily created using libraries such as LangChain[40], LLamaIndex[41], or Haystack[11]. Due to the seamless integration of the different components and subjective better abstraction layers, a decision was made to apply Deepset's Haystack library for this thesis.

2.4 Prompt Engineering

Prompt engineering, a recently developed concept [42], involves providing a set of instructions to a LLM to enable it to comprehend and generate appropriate responses, which can be understood as programming the model [43].

- Zero-shot prompts: This approach refrains from providing examples to the model to help it produce the desired user output. It is commonly employed due to its minimal tuning effort.

- Few-shot prompts: In contrast, this strategy does provide a few examples along with the initial prompt. By offering specific examples, the model gains a better understanding of the user's requirements, making it suitable for tasks that involve reasoning.

- Chain of Thought Programming (CoT): Proposed by Wei et al. [44], CoT instructs the model to first engage in reasoning about its final response before generating it. CoT encourages the LLM to solve complex problems with reduced errors.

The prompts utilised in this thesis primarily consisted of Zero-shot and Few-shot prompts. These prompts, crafted by the developers of the Ragas library, were generally maintained in their original form, except for adjustments made during the generation of the evaluation dataset. These adjustments were specifically tailored to align with the transcript of the dataset. Further details are provided in Chapter 3.

2.5 Metrics

The final component of the RAG involves an optional assessment of the outcomes generated by the pipeline. It is recommended [7] to employ high-performing Large Language Models as "judges" to assess the responses generated by other LLMs within the industry. To implement this approach, establishing specific metrics becomes imperative, serving as the foundational basis for the judge's evaluation. Facilitated by prompt engineering, this assessment process has become viable for evaluating both retrieval and generated answers, even in the absence of explicit labels.

Several noteworthy frameworks have surfaced for application, including Ragas [10], ARES [19], TruLens-Evaluation[4], and Galileo[5]. These frameworks offer similar metrics centred on faithfulness, answer relevancy, and context relevancy. Consequently, based on its ease of use, a deliberate decision was made to utilise Ragas for this thesis. Future work may consider incorporating the metrics of other frameworks to obtain a more differentiated overview of the RAG performance. In offering a benchmark against conventional metrics, the custom-made *Source Context Accuracy* and BERTScore[28] are additionally presented for retrieval- and answer-oriented metrics, respectively.

The rest of this section will delve into the definitions and computations of the metrics applied in this thesis. Unless explicitly specified, the OpenAI GPT-3.5-16k-0613 model was utilised as the LLM for scoring. Any prompts utilised by the Ragas library are provided in Appendix A.

2.5.1 Retrieval Oriented Metrics

Traditional retrieval metrics, such as Recall@k or Mean Reciprocal Rank (MRR), can be employed in the evaluation of the Information Retrieval component. However, this approach requires the availability of labels for each question-answer pair and for each chunk in the corpus. Such a requirement may prove impractical for certain users or corpora due to time and cost constraints.

Apart from the reference metric *Source Context Accuracy*, designed for the dataset utilised in this thesis to address particular intricacies, the metrics outlined here generally eliminate the need for labeling. Instead, they rely on the Large Language Model to generate scores, irrespective of the source text domain. This circumvents the challenges associated with the manual annotation of Q&A pairs and corpus

[4]https://www.trulens.org/
[5]https://docs.rungalileo.io/galileo/

chunks, making the evaluation process more efficient, scalable, and applicable in a production environment.

Source Context Accuracy

The evaluation dataset, as will be explained later in Chapter 3, lacks a relevance label for each Q&A pair and associated chunk combination. Instead, it provides the source context, comprising up to five sentences that generated the given question. Consequently, a metric has been formulated to assess whether the source context is present in the TopK results of the RAG pipeline, where k represents the number of retrieved and evaluated contexts.

The metric is deliberately designed to be straightforward, assigning values of 0 if none of the sentences from the source context are found in any of the k retrieved contexts, and 1 otherwise. This metric is formalized in Equation 2.3, where s represents a sentence, SC represents the source context, and RC represents the k retrieved contexts. Subsequent evaluations utilise the overall accuracy of the Source Context Accuracy as the primary measure.

$$SourceContextAccuracy_i = \begin{cases} 1, & \text{if any } s_j \text{ from } SC_i \text{ in any } RC \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

Context Precision@k

The initial metric employed by Ragas for retrieval is $Context\,Precision@k$. Despite sharing a similar name, the metric was not inspired by the original $Precision@k$ metric. The metric of Ragas depends on LLMs to dynamically label contexts and evaluate whether all the items relevant to the ground truth within these contexts are ranked higher. Ideally, all relevant chunks should be situated at the top ranks. The calculation requires access to the query and the retrieved contexts corresponding to the query, yielding values that fall within the range of 0 to 1. [10]

$$Precision@k = \frac{true\,positives@k}{(true\,positives@k + false\,positive@k)} \tag{2.4}$$

$$Context\ Precision@k = \frac{\sum precision@k}{total\,number\,of\,relevant\,items\,in\,the\,top\,k\,results} \tag{2.5}$$

Context Relevancy

The second retrieval-oriented metric is *Context Relevancy*, which assesses the relevance of the retrieved context. This metric quantifies relevance by estimating the value $|S|$, representing the number of sentences in the retrieved context that are deemed relevant to answering the query. The calculation, as formalized in Equation

2.6, requires access to both the query and the retrieved contexts associated with that query. The resulting values fall within the range of 0 to 1. [10]

$$Context\ Relevancy = \frac{|S|}{Total\,number\,of\ sentences\,in\,retrieved\,context} \tag{2.6}$$

Context Recall

The final metric presented by the Ragas library is termed *Context Recall*, which evaluates the extent to which each sentence in the Ground Truth (GT) answer can be attributed to one of the retrieved contexts. As with Context Precision@k, the name resembles the original Recall@k metric, but the underlying concepts differ. The values of this metric range between 0 and 1, and its computation necessitates access to all retrieved contexts and the Ground Truth answer, formalized in Equation 2.7. Since the metric requires access to a GT, it may not be applicable in a production environment due to missing labeling. [10]

$$Context\ Recall = \frac{|GT\ answers\,that\,can\,be\,attributed\,to\,context|}{|Number\,of\ sentences\,in\,GT|} \tag{2.7}$$

2.5.2 Answer Oriented Metrics

As discussed with retrieval-oriented metrics, conventional answer-oriented metrics, such as BLEU[45], ROUGE[21], or BERTScore[28], can be employed in this step. However, their practicality diminishes in a production environment characterized by unseen user-defined queries. BERTScore is provided as a reference point in the discourse on Ragas metrics. The library defines two metrics for data without a GT and two further metrics if users possess labelled data.

BERTScore

BERTScore, introduced by Zhang et al.[28], serves as a metric for evaluating the performance of text generation models, particularly those engaged in tasks such as machine translation or summarization. The metric relies on contextual embeddings obtained from pre-trained BERT[2] models for the generated content and the reference texts. Subsequently, it calculates the cosine similarity between these embeddings.

BERTScore was developed in direct response to two limitations observed in n-gram-based metrics, such as BLEU [45], ROUGE [21], or METEOR [46]. These metrics face challenges in detecting paraphrases and capturing remote relationships in sentences. BERTScore addresses these shortcomings by computing the sum of cosine similarities between pairs of sentences. This approach allows BERTScore to overcome the limitations inherent in n-gram-based metrics, providing a more nuanced evaluation of the quality of generated text. [28]

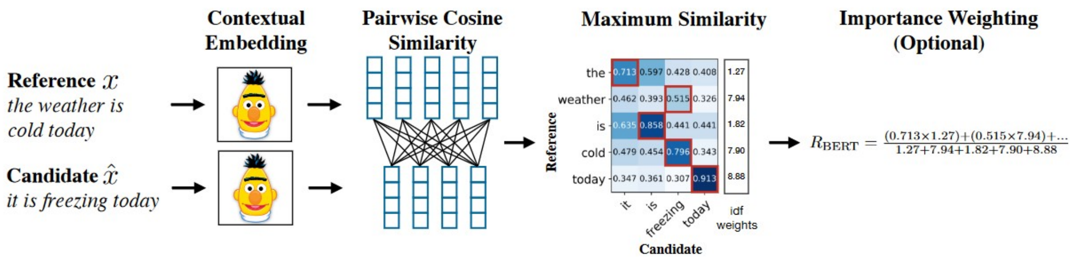

Figure 2.8: BERTScore calculation process for recall BERT, with greedy matching highlighted in red. Figure taken from [28]

The metric is computed through the initial generation of contextual embeddings for both the reference and candidate sentences, employing BERT or one of its variants. Subsequently, a pairwise comparison between the reference tokens, denoted as x_i, and the candidate tokens, represented as $\hat{x}_j$, is conducted utilising cosine similarity. The incorporation of pre-normalized vectors in this process simplifies the calculation to an inner product $x_i^\top \hat{x}_j$. [28]

First, all tokens in x are compared to all tokens in $\hat{x}$ for calculating recall, formalized in equation 2.8

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \tag{2.8}$$

After, each token in $\hat{x}$ is compared to each token in x for calculating precision, formalized in equation 2.9

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} x_i^\top \hat{x}_j \tag{2.9}$$

Lastly, the calculated precision and recall are combined to calculate the F1 score, as seen in equation 2.10.

$$F1_{BERT} = 2 * \frac{P_{BERT} * R_{BERT}}{P_{BERT} + R_{BERT}} \tag{2.10}$$

The BERT model used for evaluation was the RoBERTa-Large model by Liu et al [47].

Faithfulness

The initial metric for evaluating Ragas answers is termed *Faithfulness*, which assesses the factual consistency of a generated answer based on the retrieved context. An answer is considered faithful if all statements made in the answer can be deduced from the provided context. The scoring process involves employing Natural

Language Inference (NLI) through prompting a LLM, formalized in Equation 2.11. Computation needs access to the question, retrieved contexts, and the generated answer. The scoring range is between 0 and 1, with higher scores suggesting higher factual consistency. [10]

$$Faithfulness = \frac{|Number\,of\,claims\,that\,can\,be\,inferred\,from\,given\,context|}{|Total\,number\,of\,claims\,in\,the\,generated\,answer|} \tag{2.11}$$

Answer Relevance

The second metric oriented towards answer evaluation is *Answer Relevance*, which assesses the relevance of the answer in relation to the provided question. In this context, factuality is not the primary consideration; rather, answers containing incomplete, redundant, or unnecessary information are penalized. To compute this metric, the LLM is instructed to generate an appropriate question for the produced answer through multiple iterations (default n=3). Subsequently, the mean cosine similarity between these generated questions and the original question is calculated utilising an embedding model.

The fundamental concept behind this methodology is rooted in the notion that if the generated answer effectively corresponds to the initial question, the LLM should be capable of generating questions from the answer that closely align with the original question. The resulting score can assume values within the range of 0 to 1, with higher scores indicating greater relevancy. The default embedding model employed by the library is the OpenAI text-embedding-ada-002 model, although it can be modified to any model available on HuggingFace[48]. For the experiments, the BAAI/bge-base-en-v1.5[6] model developed by Xiao et al. [49] from the Beijing Academy of Artificial Intelligence was selected, owing to its strong performance demonstrated on the MMTB[7][50] benchmark. [10]

Answer Semantic Similarity

The two metrics depending on a Ground Truth for their calculation are now explained. The first metric, denoted as *Answer Semantic Similarity*, quantifies the semantic similarity between the generated answer and the GT, producing values between 0 and 1. To compute this score, a cross-encoder model is employed; however, the documentation does not provide a direct equation for its calculation. Analogous to Answer Relevance, the BAAI/bge-base-en-v1.5 model, proposed by Xiao et al. [49] was selected as the embedding model instead of the default OpenAI text-embedding-ada-002 model. [10]

[6]https://huggingface.co/BAAI/bge-base-en-v1.5
[7]https://huggingface.co/spaces/mteb/leaderboard

Answer Correctness

The final metric focusing on answer accuracy is *Answer Correctness*, designed to assess the precision of the generated response in relation to the GT. This metric considers two specific facets in a weighted framework to determine the overall score: factual similarity and semantic similarity (Answer Semantic Similarity) between the generated answer and the GT. While the library does not explicitly provide a direct formula, a closer examination of the code reveals the application of a weighted average involving a LLM generated F1 score and the Answer Semantic Similarity score.

Regarding the F1 score, custom prompts were devised to establish True Positive (TP), False Positive (FP), and False Negative (FN) instances, subsequently classified by the LLM. [10]

- TP: Statements that are present in both the answer and the ground truth
- FP: Statements present in the answer but not found in the ground truth
- FN: Relevant statements found in the ground truth but omitted in the answer

$$F1 = \frac{TP}{(TP + 0.5 * (FP + FN))} \tag{2.12}$$

For the semantic similarity, the weights W sum to exactly 1, and thus the score does not need to be normalized over W. Default values for W weigh the F1 score with 0.75 and the Answer Similarity with 0.25. The resulting scores range from between 0 and 1, with higher results signifying a better accuracy between the generated answer and GT. [10]

$$AnswerCorrectness_i = W_i * F1_i + (1 - W_i) * SemanticSimilarityScore_i \tag{2.13}$$

Chapter 3

Data

During discussions regarding potential evaluation corpora for use in the RAG pipeline, it was discovered that the majority of the popular Q&A evaluation datasets in NLP have been in existence since before the release of Chat-GPT in 2022 [51][52][53][54]. Given the confidentiality associated with the training of LLMs, encompassing both the training data and its utilisation throughout the training process, it is reasonable to presume that these evaluation datasets have, in some form, been encountered and assimilated by LLMs, potentially providing them with an advantage during the answering process.

This realization prompted a subsequent discussion, determining that video games could serve as a viable alternative for an evaluation dataset. While comprehensive information about certain games, including their histories, quests, and descriptions, are readily available online through wikis, other games do not offer access to their dialogues. This is the case of Baldur's Gate 3 (BG3), released in 2023 and awarded Game of the Year [55], developed by Larian Studios[1]. Given that the release date is well after the publication of any LLM employed in this thesis and no significant web-scraping of the content is possible, we are confident that the game-specific lore has never been exposed to an LLM, making it a suitable candidate for application in the RAG pipeline.

3.1 BG3 Q&A Dataset

Baldur's Gate 3 is a role-playing video game based on the renowned fantasy role-playing tabletop game Dungeons & Dragons (DnD), published by Wizards of the Coast[2]. The video game allows for both single- and multiplayer experiences in a non-linear storytelling fashion, enabling players and their avatars to experience the world at their own pace and individual approach.

Given the unavailability of the game's lore online, it had to be assembled via an indirect route, which, in this instance, involved transcripts of YouTube (YT) videos.

[1]https://larian.com/
[2]https://company.wizards.com/en

Specifically so-called Let's Plays, where a user records themselves playing the game and optionally provides commentary during the recording. A decision was made to use the videos of YouTuber WolfHeartFPS[3] as the primary source for transcripts. This decision was grounded in several factors: the user possessed a high-quality recording setup for audio, exhibited a profound understanding of the game's lore, and consistently uploaded lengthy segments lasting up to 2 hours and 30 minutes per episode.

To ensure a diverse dataset, transcripts of the initial five episodes were downloaded, resulting in approximately 630 minutes of content, equivalent to an effective text length of around 80 thousand tokens. Table 3.1 provides the exact breakdown of each episode for reference.

LP Part	Length in minutes	Tokens
1	159	20597
2	125	15465
3	90	11382
4	142	16595
5	114	16501
Total:	630	80540

Table 3.1: Length of YouTube Let's Plays and number of tokens occurring in them.

3.1.1 Cleaning

The YouTube transcripts were automatically generated by the platform and subsequently acquired through a third-party platform[4]. The quality of the transcriptions, particularly concerning named entities related to the game, was occasionally questionable, necessitating individual corrections within each transcript file. An illustrative case is the misidentification of the character named "Asterion" by the YouTube transcription tool, with various erroneous renditions such as "Sterion", "a sterion", "Astarian", and numerous others. Additionally, some sentences were cojoined into excessively lengthy text strings, requiring segmentation into their smaller constituents.

Before delving into the intricacies of Q&A pair generation, the complex and, at times, seemingly random nature of the transcript sentences within the dataset needs to be discussed. Given that the transcript was not created by a human, distinguishing between the various sounds occurring during a dialogue requires investing considerable time and effort into the improvement of the script. This means that, at first glance, it is not possible to identify the speaker (whether it is the YouTuber or game characters), the number of participants in the dialogue (whether it involves only the YouTuber, a combination of different game characters, etc.), and the specific

[3] https://www.youtube.com/@WolfheartFPS
[4] https://downsub.com/

individuals engaged in the discourse (such as Asterion, Karlach, or Shadowheart).

These complexities reduced the possibilities regarding the detail of the Q&A pairs, as the precise identification of the speaker relies on dialogue immediately preceding the given passage where the character was mentioned, rather than extracting information directly from the passage itself.

3.1.2 Q&A Pair Generation

Instead of creating the QA pairs entirely by hand, a synthetic approach utilising LLMs was considered. The objective was to generate approximately 150 QA pairs to facilitate a sufficiently comprehensive analysis of the RAG pipeline. Initially, a fully automated approach was explored using the Ragas[5] library. However, during the process, it was observed that the information in the transcript data was either insufficiently dense or excessively complex, primarily due to the aforementioned challenges of the YT transcript. Consequently, the synthesis tools encountered difficulties in generating coherent pairs, resulting in only a limited number of pairs produced using this method.

> Your task is to formulate up to 5 questions from the given context of a YouTube transcript satisfying the rules given below:
> 1. The question should make sense to humans even when read without the given context.
> 2. The question should be able to be fully answered from the given context.
> 3. The question should be framed from a part of the context that contains important information.
> 4. The question should be of moderate difficulty.
> 5. The question must be reasonable and must be understood and responded to by humans.
> 6. Avoid framing questions that contain 'the speaker'.
> 7. Do not use phrases like 'provided context', 'mentioned in the transcript section' etc in the question.
> 8. Avoid framing questions using the word 'and' that can be decomposed into more than one question.
> 9. The question should not contain more than 15 words, make use of abbreviations wherever possible.
>
> Desired format:
> Question: -II-
>
> Transcript Section: {section}

Following the partial failure of the automated pipeline, a second approach was developed by enhancing the prompts of the Ragas synthetic data generation tool and manually initiating prompts for the LLMs. This strategy leveraged GPT-4 through

[5] https://docs.ragas.io/en/stable/concepts/testset_generation.html

the API to generate up to five questions based on a given chunk of the transcript documents. The chunk size was deliberately set at approximately 1024 tokens to ensure a sufficiently varied sample size for generating diverse questions. The prompt for formulating questions is provided above this paragraph.

Subsequently, a manual inspection was conducted to assess their practical relevance to the objectives of this thesis. Queries that proved to be incomprehensible, lacking coherence, taken out of context, or generally exhibiting subpar quality were removed. Following this initial refinement phase, a transition to GPT-3.5 16k occurred, which was subsequently prompted to list the specific segments within the original chunk for which GPT-4-generated questions could provide answers. The context prompt formulation is provided below.

> Please extract relevant sentences from the provided context that can potentially help answer the following question. Adhere to the following rules:
> 1. While extracting candidate sentences you're not allowed to make any changes to sentences from the given context.
> 2. The sentences should not contain 'the speaker said', 'the speaker did' etc.
> 3. The sentences should be sufficient to answer the given question.
> 4. Select at most 5 sentences.
>
> Question:{question}
> Context:{context}
>
> Desired format:
> Sentences: -II-

Following the context generation phase, a manual inspection of the sentences was conducted to ensure high quality and that the question could be answered using the provided chunk, or at least part thereof. For the answer generation, two distinct sub-approaches were applied for the answer-generation process.

1. Answers were manually crafted by referencing the transcript and the original YouTube video, incorporating visual cues for accuracy. Additional verification was performed by cross-checking wikis maintained by dedicated users.

2. Answers were autonomously generated by employing GPT-3.5 16k, utilising the generated question and context.

These approaches were denoted as "G1" and "G2", respectively, and annotated as such in the metadata of the dataset. The prompt used for formulating the answers is provided below.

Answer the question using the information from the given context, satisfying the rules given below:
1. The answer should not contain 'the speaker said', 'the speaker did' etc.
2. The answer should solely be generated by the given context.
3. The answer should not contain more than 25 words.

Question:{question}
Context:{context}

Desired format:
Answer: -II-

Throughout the process, the LLMs were prompted with temperature = 0, frequency penalty = 0, and presence penalty = 0. This configuration was employed to guarantee the reproducibility of the generated text and to ensure that it originated exclusively from the provided source chunk. The resulting dataset comprises 153 questions, which were generated using three distinct approaches, as outlined in Table 3.2.

Generation Type	Number of Pairs
G1: Handwritten answers	76
G2: Automated answers	58
R1: Ragas synth pipeline	19
Total:	153

Table 3.2: Total number of generated QA pairs created with breakdown for each of the three generation types used.

To facilitate a better understanding of the dataset, a selection of QA pairs, along with their corresponding source contexts, are provided in Table 3.3.

QA ID	Question	Ground Truth	Source Context
Q_G1_0	What is the format of the Let's Play series for Baldur's Gate 3?	Longer episodes with sub episodes within and chapter markers. The episodes will be released every other day or so.	Now, with this Let's Play series, I'm thinking about putting out longer episodes with sub episodes within, with chapter markers to make it easy for you to take a break and then come back at a later time. And I'll probably release an episode every other day or so. This series will, of course, be kept in its own separate playlist.
Q_G1_7	What weapon proficiencies does the Elf race have in the game?	Proficiency with longswords, shortswords, shortbows or longbows.	And we get proficiency with a longsword, shortsword, shortbow, or longbow. But Ranger is going to give us all weapon proficiencies, so I don't have to worry about that
Q_G2_4	What is the consequence if targets fail the Constitution Saving Throw after being hit by a weapon coated with Drow poison?	Targets who fail the Constitution Saving Throw after being hit by a weapon coated with Drow poison become Poisoned and fall Asleep.	The Drow poison coat your active weapon with poison. Targets must succeed a Constitution Saving Throw or become Poisoned and fall Asleep. Okay, really powerful poison right there. The Grove. Volo is gone. He went to the goblin camp. And let's see if we can go talk to Kagha and Eddy.
Q_G2_31	What is the speaker's opinion about playing a front line ranger?	The speaker finds playing a front line ranger to be cool and interesting because of the companions and spells that they have.	That is a rage-filled ranger on a path of vengeance. I love it. I think it's so cool to play a front line ranger, it's just different. And Fighters are always a little bit too bland to me but then like Rangers with their Companions and like a few of the spells that they have, it really makes that front line position more interesting in my opinion.
Q_R1_8	What is the consequence of accepting the offer of eternal life?	The consequence of accepting the offer of eternal life is that the person becomes a slave and must obey their master's commands.	Given that my choices were eternal life or bleed to death on the street, I took him up on the offer. It was only afterward I realized just how long eternity could be. A spawn is less than a slave. We have no choice but to obey our master's commands. They speak and our bodies react. It's all part of the deal.

Table 3.3: Selection of dataset samples from each category of generation.

Chapter 4

Experimental Setup

Due to the complex nature of the RAG pipeline, it is recommended to calibrate it according to the specific data it will process. This section aims to serve as a fundamental building block for other users by conducting a series of experiments at various stages of the pipeline. These experiments aim to illustrate potential fine-tuning steps that should be considered. While the experiments may not comprehensively address all possible parameters, they nevertheless aim to provide a satisfactory overview of the challenges inherent in Retrieval Augmented Generation.

The points at which experiments were conducted are highlighted in orange in Figure 4.1. By performing these experiments on the BG3 Q&A evaluation dataset, the thesis additionally evaluates the quality and applicability of the corpus within the RAG pipeline.

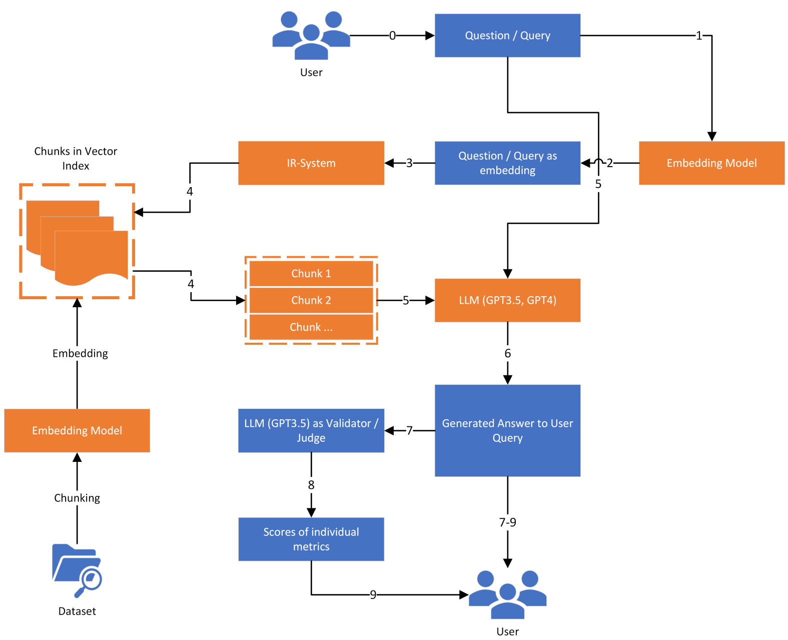

Figure 4.1: The thesis's RAG pipeline is illustrated with orange-outlined boxes representing the specific point where experiments were carried out.

First, a detailed description of the general or default configuration of the RAG is provided, followed by a description of the experiments that were conducted and the specific modifications implemented in the pipeline for the experiments to execute. In instances where no adjustments are specified, assume that the standard general pipeline was employed.

4.1 General Setup

Most details regarding the general setup have been mentioned in Chapter 2; nonetheless, to enable a comprehensive overview, the specific details and configurations are additionally provided here.

Model	Context Window Size	Cost Input	Cost Output
ada v2	8191	$0.0001	$0.0001
ChatGPT-16k	16384	$0.001	$0.002
GPT-4	8192	$0.03	$0.06

Table 4.1: Context window and pricing of models by OpenAI[1] on time of writing. Costs are per 1k tokens and are subject to change.

The pipeline has been fully implemented within Haystack [11] using Python. The vector index selected is FAISS[32] with embedding dimensions set at 768. Similarity distance is computed using cosine-similarity, and the index type is configured as "Flat". The embedding model employed for generating embeddings of both chunks and queries during retrieval is the sentence-transformer model all-mpnet-base-v2[1].

In the answering phase, GPT-3.5-16k-0613 is utilised, employing a temperature of 0.5 while keeping the remaining hyperparameters at their default settings. The formulation of the answer prompt, used to respond to a given query with the retrieved contexts, is provided by Haystack's Prompt-Hub[2]. The specific prompt used, deepset/question-answering-with-references, is listed below. The prompt aims to force the model to strictly adhere to the provided context and appropriately cite the sources for its response. Additionally, the model is programmed to stay factual by generating a pre-defined string in cases where the question cannot be answered based on the retrieved context. The default chunk size for the documents was selected to be approximately 128, and the TopK parameter was configured to 5, taking into consideration cost constraints associated with the utilisation of OpenAI models. The specific pricing details at the time of writing are outlined in Table 4.1. The experiments were conducted utilising a single NVIDIA T4 GPU to facilitate a rapid retrieval process.

[01] https://openai.com/pricing
[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[2] https://prompthub.deepset.ai/

> Create a concise and informative answer (no more than 50 words) for a given question based solely on the given documents. You must only use information from the given documents.
> Use an unbiased and journalistic tone. Do not repeat text. Cite the documents using Document[number] notation.
> If multiple documents contain the answer, cite those documents like ‘as stated in Document[number], Document[number], etc.
> If the documents do not contain the answer to the question, say that ‘answering is not possible given the available information.
>
> {join(documents, delimiter=new_line, pattern=new_line+’Document[$idx]: $content’, str_replace={new_line: ’ ’, ’[’: ’(’, ’]’: ’)’})}
> Question: {query}; Answer:

For the evaluation process, the question, answer, retrieved contexts, and ground truth were submitted to the Ragas library. Subsequently, it employed metrics outlined in Section 2.5 to assess the performance of the pipeline. The judge for all instances involving an LLM was GPT-3.5-16k-0613, with varying embedding models applied based on the specific metric. During the calculation for Answer Relevancy and Answer Similarity, BAAI/bge-base-en-v1.5[3] was utilised. In the case of BERTScore, RoBERTa-large[47] served as the underlying model.

4.2 Chunk-size & TopK

The initial experiment aimed to assess variations resulting from differences in both chunk size and the number of retrieved contexts. Four distinct chunk sizes were evaluated, starting with 128 tokens and subsequently doubling in size, with the largest chunk comprising 1024 tokens. Two alternatives for the TopK parameter were deemed sufficient and were thus set to 5 and 10.

The first stage of this experiment involved generating vector indices corresponding to the diverse chunk sizes. Given the intricate data structure of the YT transcript, a decision was made to implement the chunk splits on a sentence rather than a word basis. Additionally, it was determined that there would be no overlap between the chunks. Although the chunks did not precisely match the predefined length, they remained well within acceptable parameters to not disrupt the experiments, as illustrated by Table 4.2.

The resulting vector index for the 128-chunk size was subsequently used in all other experiments as the standard comparison, also termed ”benchmark”. It was additionally used to analyze the general distribution of the Ragas metrics. Each chunk size option was applied with the two different TopK variants to visualize the impact the number of chunks and length of each chunk can have on an LLMs output. As

[3]https://huggingface.co/BAAI/bge-base-en-v1.5

Target Chunk Size	Mean	Standard Deviation	Min	Max	#Chunks
128	137.00	9.41	128	188	541
256	265.18	10.31	256	331	279
512	521.87	11.59	512	583	141
1024	1032.32	6.51	1024	1054	69

Table 4.2: Distribution of token chunk sizes in vector indices for each size.

such, eight retrieval and answer runs were performed in total through the RAG pipeline.

4.3 Distance Calculation

The second experiment assessed the FAISS data storage system by employing two distinct distance calculation methods. Specifically, cosine similarity, generally recommended for applications involving sentence-transformers, and dot-product, which is recommended in Dense Passage Retrieval, were investigated [11]. An additional vector index had to be generated for dot-product, given that the existing indices were configured for cosine similarity. For both experiments, the all-mpnet-base-v2 embedding model was utilised. The comparison was conducted with a chunk size set at 128 and a TopK value of 5.

4.4 Dense vs. Hybrid Search

The third experiment assessed the capabilities of hybrid search compared to dense search. The utilisation of both sparse and dense search may yield significantly differing results in terms of retrieved contexts when contrasted with the standard dense search setup. Instead of employing FAISS as the vector index, the InMemoryStore[4] storage had to be applied, as FAISS exclusively supports dense search. Consistent with the FAISS configuration, an embedding dimension of 768 was selected.

For sparse search, BM25 was employed, specifically the BM25Okapi algorithm. In the case of dense search, the all-mpnet-base-v2 embedding model utilised in the general configuration was employed. Both retrievers used a TopK value of 10, and the results were subsequently concatenated, filtering the results down to the top 15 selections from the initial 20 retrieved chunks. Lastly, the ms-marco-MiniLM-L-6-v2[5] cross-encoder model served as a reranker to further refine the concatenated chunks, selecting the top 5 among them. These five contexts were then forwarded to the LLM for answering. The retrieval pipeline is illustrated in Figure 4.2.

[4]https://docs.haystack.deepset.ai/reference/document-store-api#inmemorydocumentstore
[5]https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

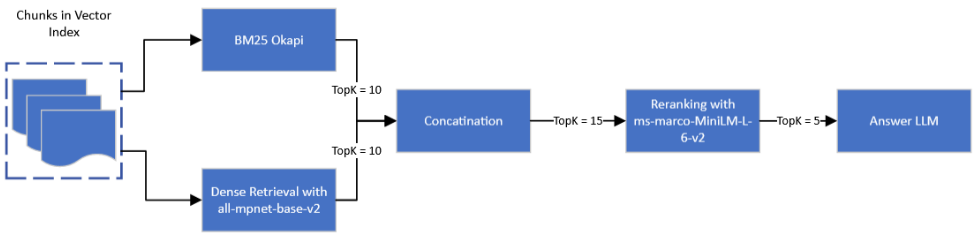

Figure 4.2: Hybrid retrieval pipeline for visualization.

4.5 Embedding Model

While the all-mpnet-base-v2 embedding model used in the general setup had the best overall results based on the BERT rankings[6], it was not exclusively trained on QA corpora. Thus, a decision was made to employ a QA-based embedding model to assess the differences between a generalized and a task-specific model. For this purpose, the multi-qa-mpnet-base-dot-v1[7] model was selected due to its performance on several benchmarks[8]. Compared to the general embedding model, this model applies dot-product for distance calculation, and an additional vector index was created.

Moreover, as detailed in Section 2.2.2, there was a misconception regarding the nature of the aforementioned models, as they are symmetric search models, not optimized for the structure of the BG3 evaluation dataset, unlike asymmetric search models. To rectify this error, a third run was conducted utilising the msmarco-distilbert-base-v4[9] asymmetric search model. Comparing these three embedding models should yield valuable insights into the significance of employing task-specific models.

4.6 Answer LLM

The last experiment focused on the answering process, assessing disparities in responses produced by a weaker and cost-effective model, such as GPT-3.5 16k, as compared to a more powerful and pricier model, such as GPT-4 within RAG. The execution of the experiment involved utilising a chunk size of 128 and a TopK parameter of 5. This decision was driven by the significant cost implications that would arise otherwise, as depicted in Table 4.3.

[6] https://www.sbert.net/docs/pretrained_models.html
[7] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1
[8] https://www.sbert.net/docs/pretrained-models/msmarco-v3.html#performance
[9] https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4

Chunk Size	TopK	GPT-3.5 16k p.P	GPT-3.5 16k all	GPT-4 p.P.	GPT-4 all
128	5	$0.001	$0.153	$0.03	$4.59
128	10	$0.0014	$0.2142	$0.0431	$6.5943
256	5	$0.0014	$0.2142	$0.0431	$6.5943
256	10	$0.0027	$0.4131	$0.0815	$12.4695
512	5	$0.0027	$0.4131	$0.0815	$12.4695
512	10	$0.0053	$0.8109	$0.1583	$24.2199
1024	5	$0.0053	$0.8109	$0.1583	$24.2199
1024	10	$0.0104	$1.5912	$0.3114	$47.6442

Table 4.3: Input Cost per Q&A pair (p.P) and complete dataset (all) for each chunk and TopK combination in the RAG process using OpenAI models. This excludes the cost of the output and the evaluation process, which may increase the costs significantly. Calculation is (ChunkSize * TopK + AnswerFormulatePrompt) / 1000 * InputCostLLM * Samples.

Chapter 5

Results

This chapter is designed to offer insights into the results obtained from the experiments outlined in Chapter 4, employing the evaluation dataset detailed in Chapter 3. The primary findings and assumptions will be further elaborated upon in Chapter 6. All runs of the RAG pipeline, along with the corresponding metric scores, can be accessed on the GitHub repository[1]. Furthermore, a parameterized Python file has been provided to enable users to implement the general RAG pipeline utilised in this thesis with their datasets.

In general, for each experiment, the results will be initially analyzed based on retrieval-oriented metrics, with answer-oriented metrics examined subsequently, as the latter highly depends on the accuracy of the former. Any supplementary results will be presented last. Throughout the analysis, it was observed that certain metrics lacked insight or were occasionally erroneous. Notably, the *Context Relevancy* exhibited significant underperformance in comparison to its counterparts, a matter that will be elaborated upon in the following section.

5.1 Ragas Metrics

This initial section aims to conduct a brief analysis of the metrics associated with Ragas to enhance comprehension regarding their application in subsequent experiments. For assessment purposes, the benchmark utilised across all experiments involved the results obtained from the RAG run with chunk size set at 128 and TopK equal to 5.

The benchmark successfully generated responses for 117 out of the total 153 questions, yielding an answer rate of 76.47%. The metrics for the retrieval part were analysed for two different scopes, the complete dataset and the answered entries only. These analyses are presented in Figures 5.1 and 5.2, respectively. A notable observation was the non-normal distribution of these metrics, with precision and recall tending to approach extremes of 0.0 and 1.0. This phenomenon explained the high standard deviation evident in both metrics during the experiments. Table 5.1

[1] https://github.com/stucksam/vt2-rag-eval

provides an instance of a sample question that was answered with near-perfect scores.

Additionally, it was observed that certain entries exhibited scores exceeding >0.95 for both recall and precision; however, these entries were not answered by the LLM. An example of such an entry can be found in Table 5.2. In this instance, the correct source context was at rank 1, yet due to a missing component in the context that was explicitly queried, the LLM chose not to answer the question. This failure may have arisen from the approach taken during chunk creation, wherein exclusively non-overlapping sentences were used.

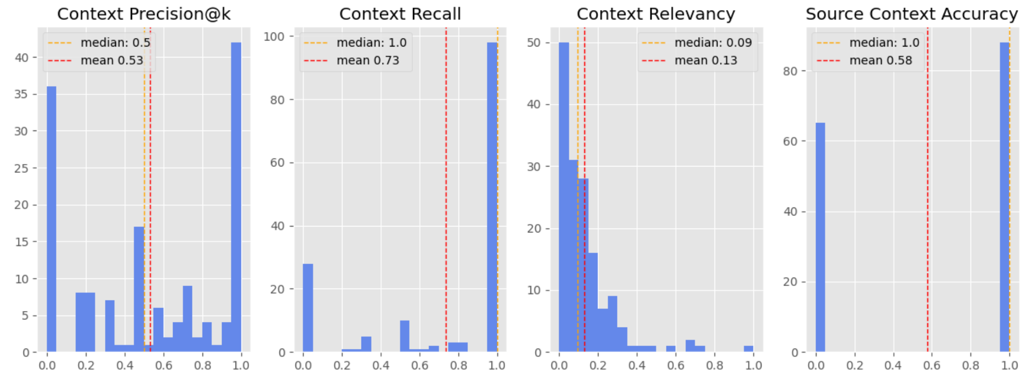

Figure 5.1: Distribution of retrieval metrics for all QA pairs (n=153) in benchmark result. The X-axis represents the scores of each metric, while the Y-axis denominates the number of samples.

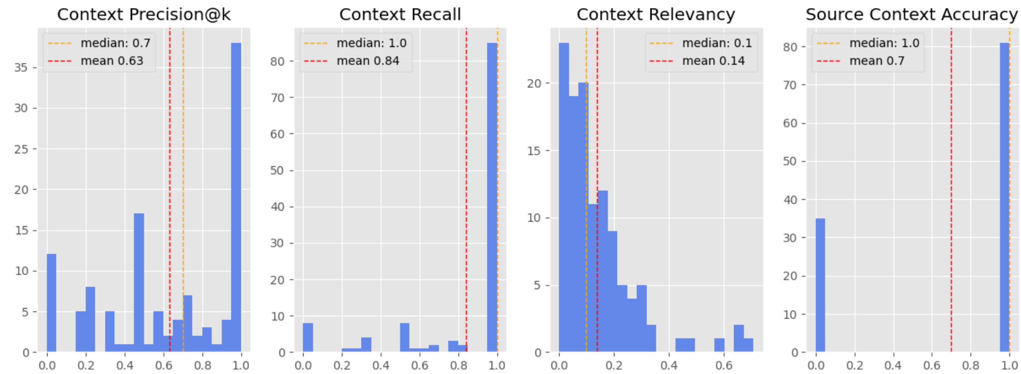

Figure 5.2: Distribution of retrieval metrics for answered QA pairs (n=117) in benchmark result. The X-axis represents the scores of each metric, while the Y-axis denominates the number of samples.

The metric "Context Relevancy" is designed to offer insights into the relevance of the retrieved context. However, owing to its consistently low scores and high standard deviation, interpreting the metric becomes challenging. Notably, the metric penalizes any inclusion of irrelevant sentences. Consequently, chunks of text that may address the query but consist of only a single relevant sentence receive lower scores compared to those containing multiple relevant sentences, but can not fully address the query. Upon closer examination, it was identified that the library intends to progressively integrate this metric with the "Context Precision@k" metric. This has not been documented yet by the library, instead, a presently inactive deprecation warning has been incorporated into the code. For the sake of completeness, the metric is presented in the results without a significant discussion.

Finally, the custom Source Context Accuracy metric. Although it may initially seem counterintuitive that approximately 35 answered samples received an accuracy score of 0, this phenomenon is explainable when one considers that the source context is not mandatory for answering a query. This may be due to the repetition of essential information within the transcript, allowing queries to be addressed without direct reliance on the source context. Notably, the fact that nearly 70% of the answered questions had the source context present in the TopK results provides confidence that such repetitions are in the minority. This observation further demonstrates that the retrieval system is generally proficient in retrieving relevant information chunks, even without clear solutions.

QA ID	Generated Answer	Ground Truth	Rank	Contexts
Q_G2_53	The speaker is concerned about the amount of content for the let's play series, as they mention that doing everything in the game would result in a series that is potentially 200 hours long (Document[2]). They also mention that they are unsure about reading all of the books in the game (Document[1]).	The speaker is concerned about the amount of content for the let's play series, as doing everything could potentially result in 200 hours of gameplay, which they find ridiculous.	1	I'll definitely be reading through all of those. I'm actually going to pick up all the books that I come across and then send them to camp. I may read a few of them to you all, but at the same time, I still haven't figured out like what the best style is for a let's play series. I'm very new to the let's play style videos. I'm not sure if that's would be something that I should read all the books or what. There's definitely a few of them that I read in Early Access that are quite important, but I will read. I'm not sure about reading all of them. Okay, so we have a door right here and a door right here.
			2	I think it has to be like that. As much as I want to do every single quest and talk to every character, a Let's Play series would be like 250 hours long. Refugees, adventurers, no one in years, and suddenly we're overwhelmed. Well met, and thank you for beating back those goblins. Most brave of you. Is there anything you need? Act fast if you do. The ritual will be complete before too long. Are you really locking down the Grove? I know it's drastic, but more monsters seem to terrorize this region every day. And what about the people here? We Druids will be safe. As for those that took refuge here, well, may Silvanus guard them as they continue their travels.
			3	I know there's a basement here too. Little kind of struggling as to how much content I should do on the let's play series. Because I mean if I do like everything like, potentially 200 hours of a let's play, that's ridiculous. So we're gonna pretend that I didn't see that latch right there, okay. Your image, yay. Anders didn't have any armor that I could take. I thought I might get a heavy armor out of him. All right, let's go ahead and loot Trend and then we are out of here. Gotta push on with like the main quest for a little bit. Surprised that I haven't leveled up yet, like I'm leveling slower in the full game than Early Access.
			4	Hello and welcome to possibly the first ever Baldur's Gate 3 full game Let's Play series. I can't even express to you all how excited and happy I am right now. I'm actually shaking a little bit to finally be playing this, especially after covering this game The Early Access phase of this game for over four years now. So thank you all so very much for choosing to come along with me on this epic adventure. Now, with this Let's Play series, I'm thinking about putting out longer episodes with sub episodes within, with chapter markers to make it easy for you to take a break and then come back at a later time. And I'll probably release an episode every other day or so.
			5	That's going to make Inventory management so much better. We got a violin right here. Let's give it a try see if we can get Shadowheart going here. Okay that didn't work, what now. It was worth a try right? I find myself looting everything but it's kind of important in this game because you need the camp supplies and you also probably need gold too. Fresh water there must be a settlement somewhere nearby. I will focus in on all of the books in one of my future live streams in a few videos but I don't think I'm going to read all of them in this particular let's play. I will talk about the lore a little bit to keep but things like letters we definitely have to read Psy I love you there I said it and if you meet me tomorrow I'll say it again.

Table 5.1: Example of a high-performing retrieval from the benchmark with Context Precision = 0.99 and Recall = 1.0. The question was "What is the speaker's concern regarding the amount of content for the Let's Play series?".

QA ID	Question	Ground Truth	Rank	Context
Q_G2_50	What is the effect of the rage wild magic?	The rage wild magic effect causes flowers and vines to sprout around the Barbarian, creating difficult terrain for others.	1	Enter a rage that releases all the magic rolling inside of you. So let's see what the random magical effect is. I'm gonna move over here a little bit. Oh nice, I ran through the potion on the ground. Wow, look at that. That's badass. Whoa, to the end of your rage, flowers and vines sprout around you. Vine growth, difficult terrain for everyone other than the Barbarian. Yeah, so you got to be a little bit careful because this is also difficult to reign for my Ranger. Although Rangers, I think at level eight, get a feature that will allow them to move through difficult terrain at normal speeds. That's pretty badass. Okay, let's go ahead and, hmm, I think I want to stay here because I want to soak up damage.
			2	Diligence and care. Hey, we got Karlach back up and let's just start backing up. I'm tempted to throw another potion in there, try to help these guys out a little bit. Anders has got 21 HP. 60 with the Eldritch blast, it would not come back too. Or do I throw a potion? The thing is even if I throw a potion, I feel like he's so powerful if you still have spell slots for divine Smite. I'm scared either way. Oh, that was great. Okay, so here we go, finally wild. Yes, here we go. Oh, I have like no HP right now. Do I take that potion and run or do we get to see what the rage wild magic is all about? Okay, let's do it.
			3	Let's try this out, this is about to be crazy. Consume, oh my gosh, I'm gonna F5 this, I might want to come back to this. Okay, here we go. Okay, infernal Fury lasts until long rest. The hatred and pain of a captured Soul fuel Karlach's infernal machine. For weapon and unarmed attacks deal an additional one to four fire damage when she's raging or when her hit points drop below 25 percent. Or you could just use the ever burn blade, you know what I'm saying, a sword that I have to get that fire damage. Or you could dip your weapon. Okay, not bad, I thought it would be more powerful though, especially at the cost of a soul coin.
			4	But fair enough, it looks pretty awesome. Let's go kick some ass, I want to try out this wild magic stuff. Don't have a bonus action though to rage, so let's just go ahead and do a regular attack. Damn, oh my God, if I had rage, I would have survived that. Holy crap, okay we gotta get serious now, that was crazy. You gotta be kidding me, Larian, you closed this window, no way. Okay, Defender was the people, him just can't open doors. Are there any other windows in this place? Doesn't look like it. Man, I love tactical combat like this, so much fun and I'm so happy that Larian changed up some of these encounters from Early Access.
			5	I don't think that's gonna hit. I don't think that hits friendlies, it's been a while since I've played. Oh my gosh, that is, that was incredible right there. That is a rage-filled ranger on a path of vengeance. I love it. All right, you're next, you're next up buddy. Come here, I still want the missile snaring to go off, let me see if I have that selected. I do have it on ask, okay. All right, I'm gonna send my Bear right here in the middle and we're gonna do a goading roar to test it out a little bit. They all succeeded on their save so that did nothing, didn't taunt any of the enemies.

Table 5.2: Example of a high-performing retrieval benchmark that was not able to be answered by the LLM with Context Precision@k = 0.99 and Context Recall = 1.0 .

The response metrics were subsequently analyzed and depicted in Figure 5.3. The majority of the metrics performed according to expectations, with only Answer Correctness and Faithfulness displaying a notable number of outliers in their scores. Faithfulness, the metric gauging the extent to which the generated answer can be inferred from the context, consistently yielded high scores for the generated responses, with the median score reaching 1.0. One outlier was examined more closely, with an example provided in Table 5.4. The LLM accurately answered the example, achieving a BERTScore of 0.904. Nevertheless, it received a Faithfulness score of only 0.5 and an Answer Correctness score of 0.75. The source context was ranked fifth, which the LLM correctly identified. Consequently, the low score is presumed to be a result of the trailing phrase "as stated in Document[5]", which could be interpreted as a statement by the library. The model was unable to verify this and penalized the answer accordingly.

The Answer Correctness metric may be susceptible to a similar issue due to its reliance on the weighing of two distinct concepts—F1 and Answer Similarity—against each other, with default values of [0.75, 0.25] for the weights of the respective concepts. It is noteworthy that when the LLM did respond to a query, the answers generally exhibited high quality, consistently achieving a BERTScore of ≥0.8.

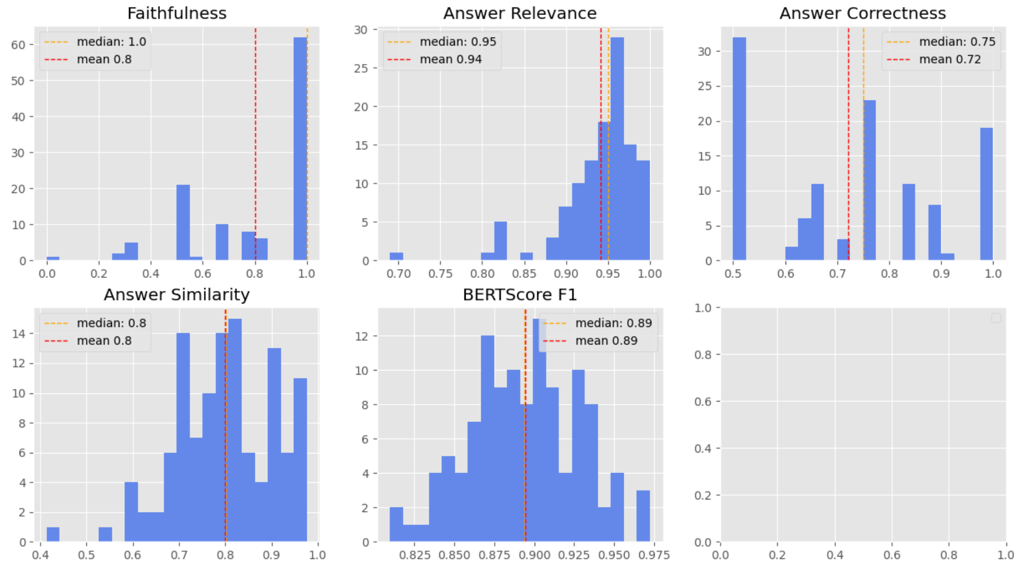

Figure 5.3: Distribution of answer metrics for answered QA pairs (n=117) in benchmark result. The X-axis represents the scores of each metric, while the Y-axis denominates the number of samples.

The understanding of the Ragas metrics gained here should help in interpreting the results of the experiments. An exhaustive list of the benchmark scores discussed in this Section can be found in Table 5.3.

Metric	Scope	Mean	Standard Deviation	Median
Context Precision@k	Complete	0.533	0.39	0.50
Context Precision@k	Answered	0.632	0.34	0.70
Context Recall	Complete	0.735	0.40	1.0
Context Recall	Answered	0.839	0.30	1.0
Context Relevancy	Complete	0.129	0.15	0.09
Context Relevancy	Answered	0.137	0.13	0.1
Source Context Accuracy	Complete	0.575	0.49	1.0
Source Context Accuracy	Answered	0.698	0.46	1.0
Faithfulness	Answered	0.757	0.26	1.0
Answer Relevancy	Answered	0.94	0.04	0.95
Answer Correctness	Answered	0.749	0.19	0.75
Answer Similarity	Answered	0.809	0.11	0.8
BERTScore F1	Answered	0.896	0.03	0.89

Table 5.3: Results of the RAG pipeline for the benchmark setup. For retrieval-based metrics two scopes are provided: Complete, which means the complete corpus was considered with n=153, and Answered, which means only the answered QA pairs were considered with n=117.

QA ID	Ground Truth	Generated Answer	Rank	Contexts
Q_G2_11	Calumvor's Kiss is the name of the plant identified as a deadly poison.	The plant identified as a deadly poison is Calumvor's Kiss, as stated in Document[5].	1	Chronic spells, I know they do not. It is my responsibility to cleanse myself. Perhaps they were trying to free you, a vain notion. I am one of many and will not be a burden to my queen. I'm picking up all of these plants so we can do some more alchemy crafting. Dragon egg mushroom, I'm sure that's used for something. There we go, recipe unlocked: ashes of dragon egg mushroom. Okay, we need three of those. Go ahead and create ashes of Balsam. Recipe unlocked: potion of greater healing. Awesome. The more you discover and the more you craft, the more recipes you unlock. So we need ashes of Balsam for that and then also any salt.
			2	Alright, so we might as well start off with a Magic Missile and then we'll go to the non-lethal attacks. Flourish. Man, my Ranger's missing everything. A lot of misses right now. Okay, let me close this door. Things did not go as planned. So let's go ahead and loot this chest now. Alright, we have a cabinet key, gold, alright, a gem, cabinet key. Oh, it's a soul coin book. The Drow poison coat your active weapon with poison. Targets must succeed a Constitution Saving Throw or become Poisoned and fall Asleep. Okay, really powerful poison right there. I've got a long road ahead. The evil thing's gone. Now where was I? Okay, alright, now the interesting thing is, is the entire Grove going to be hostile to me? My wolf companion is still alive.
			3	You see, I have this condition, very different from the parasite we share, but just as deadly. I think I might have a dark leaf in my bag or some medicinal berries. Thank you for the offer, but the treatment for my condition is very specific. What it comes down to is this: every so often, I need to get my hands on a powerful magical item and absorb the weave inside. And what happens if you don't consume these artifacts? I'll spare you the finer details, but it begins with a simple biological deterioration, muscle spasms, disorientation, a slight ringing in the ears. And if it's there for too long, it's deadly.
			4	What do I currently have on? Nothing, so let's put that on my Ranger. And I can pop a healing word or even take a potion, then you get that one to six poison damage. And when you coat regular poison on your weapon, it's only one to four. Okay, I think we're good. Wooden Shield, carrot, gotta make sure that I do pick up all the camp supplies. What if Netty's still back there? What did I tell you? Let me talk to Amino real fast, see if he has anything to say. The right of thorns has ceased, it seems you will be allowed to stay but behave. Yes sir.
			5	I pity you got me instead of him. He understands these things, studied them. Still, we have options. All right, let's see what we can do. Now, what's that plant? Will it help? It might, but first things first, tell me about your symptoms. Have you noticed anything strange happening? Looking closer, you realize you've seen this thorn in the Wilds. Calumvor's Kiss, a briar from the Dalelands and a deadly poison. I wonder if I got that dialogue because I'm a ranger. Put the briar down and we'll talk. I want to help you, but I can't unless you work with me. So, has anything unusual happened to you? Hmm, I'm sorry, I don't trust you.

Table 5.4: Example of a sample from the benchmark with a Faithfulness and Answer Correctness score of 0.5 and 0.75 respectively and a BERTScore of 0.904. The query for the sample was "What is the name of the plant identified as a deadly poison?".

5.2 Chunk-size & TopK

The chunk-size and TopK experiments were conducted to determine the most effective combination of these parameters for the corpus.These runs also assessed the suitability of the BG3 Q&A dataset for integration into the RAG pipeline.

Retrieval Metrics

Table 5.5 and 5.6 present the scores of the retrieval metrics of each chunk size option for TopK set 5 and 10, respectively, displaying a stark decline in retrieval quality as chunk sizes increase. The 128-chunk size successfully provided answers for 117 queries. Conversely, the largest option with 1024 chunks only addressed 86 out of the total 153 queries.

Upon analyzing the QA pairs, it was observed that only 14 pairs remained unanswered by all four chunk options when TopK was set to 5. This number decreased to 8 pairs when TopK was set to 10. Both Context Precision and Recall additionally exhibited a high but stable standard deviation. This observation aligns with Figure

5.2, which illustrates the distribution of scores for the benchmark option.

Chunk Size	Context Precision@k	Context Recall	Context Relevancy	Source Context Accuracy	Support
128	**0.533±0.39**	0.734±0.4	**0.128±0.15**	**57.52%**	153
256	0.467±0.41	0.736±0.39	0.121±0.1	54.25%	153
512	0.276±0.39	**0.74±0.39**	0.082±0.07	50.32%	153
1024	0.12±0.29	0.737±0.4	0.053±0.05	40.52%	153
128	**0.635±0.34**	0.840±0.30	**0.137±0.13**	**70.09%**	**117**
256	0.560±0.39	0.847±0.29	0.093±0.09	65.79%	114
512	0.371±0.41	**0.859±0.28**	0.057±0.07	64.86%	106
1024	0.181±0.35	0.832±0.32	0.038±0.04	53.84%	86

Table 5.5: Average retrieval metric scores for TopK = 5, broken down for the complete dataset (above) and answered samples only (below).

The benchmark had the best Context Precision@k and Source Context Accuracy, indicating that the relevant segments were more prominently ranked and included the original context. It achieved scores of 0.635 and 70.09% for 117 answered questions, respectively. The 512 option had the best Context Recall, signifying if the retrieved contexts can be used to answer the given query, scoring 0.859.

Chunk Size	Context Precision@k	Context Recall	Context Relevancy	Source Context Accuracy	Support
128	**0.507±0.36**	0.836±0.31	**0.101±0.12**	**65.36%**	153
256	0.455±0.38	0.8±0.34	0.064±0.07	**65.36%**	153
512	0.269±0.36	0.804±0.34	0.037±0.04	60.78%	153
1024	0.142±0.3	**0.866±0.31**	0.011±0.01	52.29%	153
128	**0.568±0.33**	0.904±0.22	**0.101±0.13**	**75.39%**	**124**
256	0.52±0.37	0.878±0.25	0.064±0.06	72.5%	120
512	0.311±0.37	0.850±0.29	0.037±0.04	70.09%	118
1024	0.207±0.35	**0.92±0.23**	0.013±0.01	61.36%	88

Table 5.6: Average retrieval metric scores for TopK = 10, broken down for the complete dataset (above) and answered samples only (below).

Increasing the number of retrieved chunks from 5 to 10 yielded positive results in that a larger number of QA pairs were answered across all four chunk sizes. Specifically, the 128-variant exhibited an increased answer rate of 81.01%, compared to the previous rate of 76.47%. Similar to the smaller TopK variant, this larger variant consistently demonstrated superior Context Precision and Source Context Accuracy across the entire dataset and for answered samples. However, it is noteworthy that the Context Precision scores experienced a decline to 0.507 and 0.568, respectively. Recall conversely increased for three out of the four variants on the answered samples. Notably, the 1024-chunk option displayed the highest Recall at 0.92. This observation aligns with the intuitive notion that more and larger chunks can provide a more extensive overall context, potentially facilitating improved answers.

In general, the 128-chunk options outperformed their larger chunk counterparts, showcasing a substantial advantage in terms of the number of queries successfully answered by the subsequent LLM. Notably, there was an intriguing observation of a systematic decline in Source Context Accuracy as the chunk sizes increased, considering that the chance of accidentally retrieving the correct chunk increases the larger chunk sizes are, as illustrated in Table 4.2. This phenomenon may be attributed to the more granular search possibilities discussed in Chapter 2 when utilising smaller chunk sizes.

Answer Metrics

The evaluation of the generated answers revealed that, although the metric scores exhibited overall similarity, the benchmark predominantly yielded the highest answer rate, with 3/5 and 4/5 of the highest scores for the TopK 5 and 10 variants, respectively. The scores for the TopK 10 variants exhibited a distribution pattern akin to that of the TopK 5 runs, with only marginal performance enhancements observed. Notably, in terms of Faithfulness, the scores exhibited a decrease, reaching up to 0.07 points or 11% in the case of the 1024-chunk variant.

Chunk Size	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
128	**0.802±0.24**	0.940±0.05	0.722±0.17	**0.8±0.11**	**0.895±0.03**	**117**
256	0.76±0.29	0.94±0.05	0.718±0.17	0.797±0.1	0.892±0.03	114
512	0.722±0.29	**0.943±0.05**	**0.748±0.19**	0.793±0.12	0.892±0.04	106
1024	0.653±0.36	0.931±0.06	0.720±0.16	0.776±0.12	0.888±0.03	86

Table 5.7: Average answer metric scores for chunking options and TopK = 5 with their individual support.

Chunk Size	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
128	**0.757±0.26**	**0.94±0.04**	0.749±0.19	**0.809±0.11**	**0.896±0.03**	**124**
256	0.722±0.30	0.934±0.05	**0.755±0.16**	0.8±0.11	0.893±0.03	120
512	0.703±0.32	0.936±0.05	0.741±0.17	0.794±0.12	0.894±0.04	118
1024	0.577±0.39	0.935±0.06	0.728±0.18	0.776±0.12	0.893±0.4	88

Table 5.8: Average answer metric scores for chunking options and TopK = 10 with their individual support.

The results obtained from this experiment suggest that the initial retrieval step plays a pivotal role in influencing the performance of the RAG pipeline. Consequently, optimizing this step, which can be executed at various points, is thus of utmost importance.

Applicability of BG3 dataset

Conducting a concise analysis of the unanswered QA pairs throughout all eight RAG runs is imperative to instil confidence in the applicability of the evaluation dataset. In total, six QA pairs remained unanswered: four originating from the G1 generation type (manual) and two from the G2 generation type (LLM only).

The samples are listed in Table 5.9. Upon close examination, the first two entries appear challenging due to their excessive reliance on "the speaker's character", a phrase open to various interpretations. The questions originated from the character creation screen, wherein the YouTuber customized the appearance and skills of their in-game character while providing insights into the character's background. However, these questions could just as easily have arisen during a dialogue with a Non-Player Character (NPC). Consequently, we must dismiss them as too imprecise. The third sample encountered similar issues to the first two but emerged from a different section of the Let's Play.

QA ID	Question	Ground Truth	Source Context
Q_G1_4	How did the speaker's character survive after the incident where their family died?	The character lived the next 15 years as an outlander trained hard to ensure that they would never be weak again.	I spent the next 15 years of my life as an Outlander, living in isolation but ever on the lookout for shadow Druids. My only friends were that of the creatures of the forest, and one day I stumbled across a lone baby wolf in a fern patch. I named her Fern. She would join me on my quest.
Q_G1_9	What class is the speaker's character?	The character is a Ranger.	We're going to be a strength-based, heavy armor-wearing Ranger. Rangers are unrivaled scouts and trackers, honing a deep connection with nature in order to hunt their favorite prey.
Q_G1_25	What is the consequence if the characters do not escape or get cleansed?	They will become ghaik , also called mind flayers.	Unless we escape, unless we are cleansed, bodies and minds will be tainted and twisted. Within days, we will be ghaik mind flayers.
Q_G1_77	What is the Amulet of lost voices and who does the speaker give it to?	The amulet provides a level 3 necromancy spell speak with the dead which allows the player to speak with the dead. The player gives the Amulet to Gale.	The Amulet of lost voices gives us speak with the dead, a level 3 necromancy spell. Let's get on with it. We'll give that to Gale, our Necromancer.
Q_G2_12	What is the Old Temple of Soluna?	The Old Temple of Soluna is significant because Halsin's research suggests that the parasite is connected to a goblin camp located there.	Halsin's research suggests that the parasite is connected to a goblin camp located in Old Temple of Soluna. Nettie couldn't help us. Instead, she tried to poison us with deadly venom, hoping to kill us before we turned into mind flayers. The adventurers were with Halsin when he disappeared. They might know what happened to him. All I can say for sure is they all went to the old temple of Soluna, and Master Halsin didn't make it back. This is Halsin's journal. Extraordinary happenings. While meditating in the forest, Nettie and I were ambushed by a pack of goblins led by a drow.
Q_G2_33	What are the consequences of Kagha's death?	Kagha's death was significant because it occurred before the Tiefling were forced out of the Grove.	Kagha died before the Tiefling were forced out of the Grove. We defeated the shadow Druids and stopped the ritual, we should report our success to Zevlor. To think Kagha turned to the ways of shadow, a corruption ran so deep. Okay, definitely encourage a few of you guys to try to persuade Kagha after you confront her and rat her out. Kagha's dead, truly? I'm sorry it had to come to this but she left us with no choice, thank you.

Table 5.9: Dataset samples that could not be answered in all chunk and TopK experiment runs.

The fourth sample was very clear in its wording but contained a compound question within a single statement. This may have impeded the retrieval of the relevant information during the dense search process, affecting the accuracy of the response. Furthermore, the responses to the fifth and sixth samples appear disconnected from the corresponding questions, suggesting a potential error during the original question generation or in the correction of typographical errors.

Therefore, the failures are typically of suboptimal quality and must be removed or modified to enhance the overall quality of the dataset. Meanwhile, the remaining portion of the dataset is generally manageable, validating the applicability of the corpus to the RAG pipeline in the process.

5.3 Distance Calculation

The second experiment was generally designed to assess FAISS and the default embedding model employed in the pipeline by comparing two distinct distance calculation methods: cosine similarity and dot product. Given that this primarily affects the retrieved contexts, only the retrieval metrics are of particular concern. The expectation was that the distance calculation method should not significantly influence the results, as the model is proficient in utilising both approaches.

The results showed that the embedding model performed better across all retrieval metrics using the dot-product on the entire dataset. However, in contrast, the subsequent LLM could only answer 114 questions as opposed to 117 achieved through cosine-similarity. The reasons behind these disparities remain unclear. Notably, five queries were successfully answered by cosine-similarity, which the dot-product run failed to provide. Surprisingly, during the analysis of these samples, it was observed that four out of the five queries shared the same retrieved items in identical ranks, and the source context was embedded within them. Nonetheless, the LLM refrained from responding in the dot-product pipeline, whereas the cosine-similarity pipeline did respond, yielding correct answers. Table 5.12 illustrates an example of such a scenario. Although no explanation can be given for this discrepancy, it accounts for the conversely higher average retrieval metric scores in answered samples. Consequently, future investigations should verify whether this anomaly is an isolated incident or indicative of a more extensive issue with the dataset or pipeline. The fifth sample (Q_G1_20) was erroneously answered by the cosine-similarity pipeline, providing "dark-brown" instead of the correct answer, "light-brown".

Distance Calculation	Context Precision@k	Context Recall	Context Relevancy	Source Context Accuracy	Support
Cosine Distance	0.5325±0.38	0.734±0.4	0.128±0.15	**57.51%**	153
Dot Product	**0.5332±0.39**	**0.738±0.4**	**0.129±0.156**	**57.51%**	153
Cosine Distance	**0.635±0.34**	**0.840±0.30**	**0.137±0.13**	**70.09%**	**117**
Dot Product	0.623±0.35	0.837±0.31	0.129±0.14	68.42%	114

Table 5.10: Average retrieval metric scores for TopK = 5 for cosine similarity and dot-product distance calculations during retrieval, broken down for the complete dataset (above) and answered samples only (below).

In comparison to retrieval, the generated answers exhibited greater variability in results. Answers based on the contexts retrieved by cosine distance had better

Distance Calculation	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
Cosine Distance	**0.802±0.24**	0.940±0.05	0.722±0.17	**0.8±0.11**	**0.895±0.03**	**117**
Dot Product	0.785±0.27	**0.942±0.04**	**0.734±0.17**	0.797±0.11	0.893±0.04	114

Table 5.11: Average answer metric scores for TopK = 5 for cosine similarity and dot-product distance calculations during retrieval.

Faithfulness, Answer Similarity, and BERTScore. Conversely, the dot-product retrieved contexts enabled the LLM to produce answers with better Answer Relevancy and Correctness.

Rank	Contexts	Answer Cosine Distance	Answer Dot-Product
1	The Cap of Curing. When you inspire an ally using Bardic Inspiration, they also regain 1d6 hit points. So that is a Bard-specific hat right there. I wonder if I should put it on anybody right now. We're gonna put it on Shadowheart. Not gonna do anything except be a cosmetic, but whoa, Shadowheart looks awesome. Yeah, we gotta get the hell out of here, man. And I love the song. I don't love it that much, though. I keep thinking about Thunderwave. That's not part of this RP playthrough. So, but then again, Gale could do the Thunderwave. And I'm not, you know, RPing Gale. If Gale's angry and he wants to do something, I can't stop him, right? No, never mind.		
2	Chronic spells, I know they do not. It is my responsibility to cleanse myself. Perhaps they were trying to free you, a vain notion. I am one of many and will not be a burden to my queen. I'm picking up all of these plants so we can do some more alchemy crafting. Dragon egg mushroom, I'm sure that's used for something. There we go, recipe unlocked: ashes of dragon egg mushroom. Okay, we need three of those. Go ahead and create ashes of Balsam. Recipe unlocked: potion of greater healing. Awesome. The more you discover and the more you craft, the more recipes you unlock. So we need ashes of Balsam for that and then also any salt.		
3	Protect a creature from attacks, increase its armor class by two and last until a long rest. I wonder if they, you can still cheese that, where you just put the sword on, you use it and then you take the sword off. Sure, you probably can. I don't know how I feel about that. It's like a cheese but it's not a cheese too. All right Karlach, over there, get you at the door. I'm gonna back up. I wish I could ready in action right now. Wow, there's a lot of damage. All right, Karlach's almost dead again. You know what, I bet you could run in here and kill like, I guarantee I can.	The Cap of Curing is a Bard-specific hat that allows the wearer to regain 1d6 hit points when inspiring an ally using Bardic Inspiration. It serves as a cosmetic item and does not have any additional effects. (Document[1])	Answering is not possible given the available information.
4	We unlocked level two spell slots, which gives us a few new spells. Hold person's really good, protection from poison, warding bond, warden ally that gained resistance to all damage, now plus one to their armor class and saving throws, and lasts until a long rest. It's pretty nice that wasn't in Early Access, let's go ahead and prepare that. We'll get rid of cure wounds for now, and I'll prepare warding bonds. We also have spiritual weapon too, I feel like I want to use that, so let me get rid of inflict wounds, even though inflict wounds is a really good, you know what, I'll get rid of sanctuary for now, it's only temporary, you can switch these spells in between combat encounters when you're playing prepared casters, let's pick up spiritual weapon, summon a floating spectral weapon that attacks your enemies alongside you, and it costs a bonus action.		
5	Diligence and care. Hey, we got Karlach back up and let's just start backing up. I'm tempted to throw another potion in there, try to help these guys out a little bit. Anders has got 21 HP. 60 with the Eldritch blast, it would not come back too. Or do I throw a potion? The thing is even if I throw a potion, I feel like he's so powerful if you still have spell slots for divine Smite. I'm scared either way. Oh, that was great. Okay, so here we go, finally wild. Yes, here we go. Oh, I have like no HP right now. Do I take that potion and run or do we get to see what the rage wild magic is all about? Okay, let's do it.		

Table 5.12: Content of retrieval task for Q_G2_14 for both distance calculations and the respective generated answer. The question was "What is the Cap of Curing and its function?".

While the differences in this experiment were not significant, testing and applying such incremental improvements in the RAG pipeline can have a multiplying effect in summation with other improvements and should thus not be ignored.

5.4 Dense vs. Hybrid Search

The disparities between the benchmark and the experimental pipeline were particularly evident in this experiment. While the benchmark utilised FAISS as the vector index, the hybrid approach necessitated the use of an in-memory database to facilitate both dense and sparse searches on the chunks. Given that the BG3 evaluation dataset encompasses a substantial number of game-specific keywords, such as names, items, locations, or concepts, it is theorized that a dense-only search may generally yield suboptimal results. The combination of both sparse and dense searches may lead to significantly improved outcomes, as sparse search can focus specifically on keywords mentioned in the query without sacrificing the advantages provided by dense search due to the hybrid setup.

Search Method	Context Precision@k	Context Recall	Context Relevancy	Source Context Accuracy	Support
Dense Search	0.533±0.39	0.734±0.4	0.128±0.15	57.51%	153
Hybrid Search	**0.781±0.32**	**0.839±0.27**	**0.137±0.11**	**77.12%**	153
Dense Search	0.635±0.34	0.840±0.30	0.137±0.13	70.09%	117
Hybrid Search	**0.779±0.32**	**0.851±0.27**	**0.137±0.11**	**78.01%**	**141**

Table 5.13: Updated average retrieval metric scores for TopK of 5 and chunk size of 128 for comparison between dense search and hybrid search, broken down for the complete dataset (above) and answered samples only (below).

The initial assessment of both the retrieval and answer segments of the hybrid RAG pipeline was nonetheless surprising. The hybrid pipeline demonstrated remarkable success and managed to achieve the highest results for Context Precision@k and Source Context Accuracy, as well as the highest answer rate among all experiments conducted in this thesis, initially accounting for 139 QA pairs or 90.84%. Notably, the hybrid pipeline successfully answered 27 QA pairs that remained unanswered by the dense approach. Conversely, the dense approach answered four samples that the hybrid setup did not address, warranting further analysis.

In three instances (Q_G1_3, Q_G1_36, Q_G2_51), the accurate source context was present in the retrieved results, signifying an issue with the LLM losing focus, consistent with observations from the previous experiment. The fourth case revealed an error in the filtering system applied during evaluation, designed to exclude responses lacking answers. This response featured the subtext "the documents do not provide", which the LLM occasionally employed instead of the default string defined in the answer prompt to refer to a portion of the question that it could not answer. Nevertheless, the LLM accurately addressed the remaining portion of the question. An illustrative example is provided in Table 5.14.

Upon rectifying this filtering issue, the effective answer rate increased to 141 or 92.16%, as displayed in Table 5.13. Subsequently, all other experiment results were checked for this issue and retroactively updated with the correct results.

Question	LLM Answer	Ground Truth
What are inspiration points and how can they be used in the game?	Inspiration points in the game can be used to re-roll dice in dialogue if an ability check fails. They are gained by doing something outlander-ish. The player can expand their hotbar, use a keychain to store keys, an Alchemy pouch for potions, and a camp supply pack for supplies and food. However, **the documents do not provide information on how inspiration points can be used in any other way.**	When a character performs an action that's true to their Background, they receive Inspiration. Inspiration points may be used to reroll an ability check.

Table 5.14: Example of a generated answer that is mostly accurate but had the keywords for not providing information. Sample ID is Q_G1_21.

What was additionally interesting in the hybrid pipeline, was that when comparing the scores derived from the complete dataset with those obtained solely from answered questions, there was no significant variation. In prior experiments, including the benchmark analysis, there was a notable tendency for scores to exhibit a considerable increase when excluding unanswered samples. This phenomenon could potentially be attributed to the high response rate, which minimally affects the overall average.

The scores presented in Table 5.15 validate that the answers to the queries were generally of superior quality compared to those in the benchmark. Although the benchmark exhibited higher Answer Relevancy, the hybrid approach demonstrated greater Faithfulness, Answer Correctness, Answer Similarity, and BERTScore. This proves the assumed strength of employing both sparse and dense search on the BG3 dataset.

Search Method	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
Dense Search	0.802±0.24	**0.940±0.05**	0.722±0.17	0.8±0.11	0.895±0.03	117
Hybrid Search	**0.822±0.23**	0.898±0.1	**0.756±0.17**	**0.818±0.1**	**0.898±0.04**	**141**

Table 5.15: Average answer metric scores for TopK of 5 and chunk size of 128 for comparison between dense search and hybrid search.

It would be interesting to compare different reranking models in this step, alongside an asymmetric search model as opposed to the default symmetric one. However, owing to time and cost constraints, such experiments were considered out of scope for this thesis.

5.5 Embedding Model

The fourth experiment was designed to display the impact of an embedding model on the overall RAG pipeline. Three distinct models were evaluated, encompassing two symmetric search models and one asymmetric. The symmetric models selected for examination were the default all-mpnet-base-v2 and the QA-fine-tuned multi-qa-mpnet-base-dot-v1. On the other hand, the asymmetric search model chosen was the msmarco-distilbert-base-v4. The anticipation was that the asymmetric search model would surpass its symmetric counterparts in performance due to the inherent asymmetric structure of the evaluation dataset.

Embedding Model	Context Precision@k	Context Recall	Context Relevancy	Source Context Accuracy	Support
all-mpnet-base-v2	0.533±0.39	0.734±0.4	0.128±0.15	57.51%	153
multi-qa-mpnet-base-dot-v1	0.546±0.41	0.755±0.36	**0.129±0.14**	55.56%	153
msmarco-distilbert-base-v4	**0.603±0.41**	**0.794±0.34**	0.118±0.14	**60.13%**	153
all-mpnet-base-v2	0.635±0.34	0.840±0.30	0.137±0.13	**70.09%**	117
multi-qa-mpnet-base-dot-v1	0.652±0.36	0.830±0.28	**0.141±0.12**	66.09%	115
msmarco-distilbert-base-v4	**0.693±0.37**	**0.876±0.26**	0.129±0.14	68.50%	**127**

Table 5.16: Average retrieval metric scores for TopK of 5 and chunk size of 128 for comparison between different embedding models used in index store, broken down for the complete dataset (above) and answered samples only (below).

The retrieval metrics presented in Table 5.16 demonstrate comparable scores and distributions for symmetric searches, mirroring those observed in the distance calculation experiment outlined in Section 5.3. However, in contrast to these results, the asymmetric approach exhibited superior performance across most retrieval metrics, achieving an impressive 83% answer rate. This result aligns with the discussion that was had in Section 2.2.2, in which it was theorized that the BG3 evaluation dataset conforms to an asymmetric structure, thereby potentially benefiting from the utilisation of an embedding model tailored to such asymmetry. An example of a sampled correctly answered by the asymmetric model and not by the benchmark is provided in Table 5.18 For answered samples, the all-mpnet model demonstrated the highest Source Context Accuracy overall. This observation may explain the apparent contradiction where the QA-embedding model had better retrieval scores but struggled to answer a larger number of QA samples than the all-mpnet model.

Embedding Model	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
all-mpnet-base-v2	**0.802±0.24**	**0.940±0.05**	**0.722±0.17**	0.8±0.11	0.895±0.03	117
multi-qa-mpnet-base-dot-v1	0.772±0.27	0.880±0.15	0.708±0.18	0.795±0.12	0.892±0.04	115
msmarco-distilbert-base-v4	0.795±0.28	0.893±0.11	0.718±0.17	**0.801±0.11**	**0.896±0.03**	**127**

Table 5.17: Average answer metric scores for TopK of 5 and chunk size of 128 for comparison between the different embedding models used in index store.

The response metrics presented in Table 5.17 indicate that asymmetric search enabled the LLM to similar high-quality answers as the benchmark. Notably, the benchmark exhibited significantly higher Answer Relevancy than the other two models, attaining a score of 0.941, in contrast to 0.893 and 0.880.

This experiment demonstrates that the selection of embedding models has a substantial impact on retrieval accuracy. Therefore, the careful consideration and appropriate choice of an embedding model, preferably tailored to the specific dataset, is of utmost importance when implementing a RAG pipeline.

Rank	Contexts Symmetric Search (Benchmark)	Contexts Asymmetric Search
1	Someone set you on Karlach's tail. I'd like to know who this source of yours is. I can say only this. Karlach's not the only one who's had a villain's knife held to their throats. What's that supposed to mean? The truth will out before you know it. One night soon, when we make camp, the veil will be lifted and I'll pay my pennants. Penance, should I be worried? You're not in any danger, I promise. I can't say the same about me. Okay, so I think I actually want to. We did Wyll's quest for now. Spoke ominously about a Penance he must face. We should take a long rest when the opportunity strikes.	Still, when the time comes, call for the blade. I won't belong to answer. Go to my camp. A splendid plan. We'll talk more there. We've picked up another companion named Wyll. Recruited Wyll, a monster hunter known as the blade of Frontiers, is on a mission to kill Karlach, a powerful devil who's a danger to the entire Sword Coast. So we know from character creation that Karlach is also an origin character and a possible companion. There's just so much content available at the Grove. It's pretty crazy. I mean, I could probably come over here and spend like two hours going through all the content that's over here. But for the sake of the let's play, you know, I'll probably run through his side on my way out.
2	He might be able to stabilize things if I can find him. There are some Tiefling holed up in a Druids Grove nearby, we might look there. Sounds like a good lead. Hopefully our guy will be among them. Tune up would do this old turbo world of good. Infernal war machines are these gigantic, gigantic vehicles of War. The Devils used in the blood war to fight against demons and they're powered by Soul coins. And she has an infernal engine as her heart which is really interesting. I'm very curious about the lore behind that Karlach's story now. Karlach seems just like a fun character to have in your group like this is, this is cool. More money did, I didn't expect to like Karlach as much as I do right now.	I made my way to Avernus to stop her. She fled from my reach, even climbed aboard the Mind Flayer ship as it screeched through the Hells. I followed in close pursuit. I can't bear to imagine the lives Karlach might be taking, the damage she might be doing. Yeah, and he went through all of that to track down Karlach. Holy cow. Who is the source of yours? A powerful friend with a keen interest in privacy. I'm sworn to say no more. All right, what else is on your mind? I've noticed your stone eye. Did you lose it in battle? A most vicious one, in fact. It's made from pure bloodstone, carved from the Galelan Mountains just north of the Moonsea.
3	I made my way to Avernus to stop her. She fled from my reach, even climbed aboard the Mind Flayer ship as it screeched through the Hells. I followed in close pursuit. I can't bear to imagine the lives Karlach might be taking, the damage she might be doing. Yeah, and he went through all of that to track down Karlach. Holy cow. Who is the source of yours? A powerful friend with a keen interest in privacy. I'm sworn to say no more. All right, what else is on your mind? I've noticed your stone eye. Did you lose it in battle? A most vicious one, in fact. It's made from pure bloodstone, carved from the Galelan Mountains just north of the Moonsea.	We have an owl bear egg here too. And we are going to respectfully exit the owl bear cave and continue our search for Karlach. ER, them all, this will be no different. I've always had a soft spot for the confident ones, they always keep moving, stranger. Quietly. What happened here? I told you to go. Let's appear at his collar with a perception check. What are you doing? Get back. Okay, I'm gonna try an intimidation check now. If I wanted to hurt either of you, I'd have done it already. Wow, this is a tough check right here. Amazing, perhaps that's true. Your scent is thick with blood. Still, you should go.
4	Okay, I think I like Karlach. Let's do it. Cornered me outside the tall house just up the hill. Don't they've gone far after the scorching I gave him. Hang on though, looks like you've got enough backup at your side. Not sure there's room for me. I'll catch up with you when it's time to camp for now. But don't get to any of the fun stuff without me, got it? Okay. But why, why, why? A great uncle today. I just figured you've been witness to a pantomime, I'm sorry to say. And I've played my part all too poorly. I respect Wyll for admitting it that he was wrong.	Karlach meets the criteria, pet. Trust me. Al, this is a great scene so far. I like Mizora a lot and I really love the dynamic between Wyll and Karlach. And look at the choices that we have. If we kill Karlach now. Better not lay a damn finger on Karlach. Get to the point, devil. What do you want? The point, oh yes. Thanks for the reminder. The lightning storms of Death Strike his flesh. His soul passes through each layer of the health, gaining their essence and their torment. Oh my God. Hell's have you done? Almost broken a price paid. You know the terms, get used to the new form, Pat. There's no going back.
5	We have an owl bear egg here too. And we are going to respectfully exit the owl bear cave and continue our search for Karlach. ER, them all, this will be no different. I've always had a soft spot for the confident ones, they always keep moving, stranger. Quietly. What happened here? I told you to go. Let's appear at his collar with a perception check. What are you doing? Get back. Okay, I'm gonna try an intimidation check now. If I wanted to hurt either of you, I'd have done it already. Wow, this is a tough check right here. Amazing, perhaps that's true. Your scent is thick with blood. Still, you should go.	And if any of mummy's little friends want to pick up where the others left off, they'll find nothing but a pile of Ash. Damn, Zariel won't get near you again, we'll make sure of that. Can't he, couldn't even lay a finger. Cool, all right, we confronted the paladins chasing Karlach. She's joined our search for a cure for the parasite. Oh shit, what happened to, hell happened to him man, what's going on over here? Is he, that's pretty cool. To rage, can I control her? She's destroying everything. I literally don't have control over her anymore. Unbridled wrath, releasing a decade of pent-up rage in a blaze of infernal Fury.

Table 5.18: Retrieved contexts for QA sample Q_G2_39 that was answered by the asymmetric search model (right) but not by the symmetric (benchmark) model (left). The query for this sample was "Who is hunting down Karlach?" and the LLM correctly generated "Wyll, a monster hunter known as the blade of Frontiers, is hunting down Karlach, a powerful devil who is a danger to the Sword Coast (Document[1]).".

5.6 Answer LLM

The last experiment aimed to investigate whether the answers of retrieval results vary depending on the employed answer LLM. Two models were selected: the default GPT-3.5 16k-0613 and its more robust counterpart, GPT-4. The expectation was that any distinctions would be marginal, given that GPT-3.5 16k has demonstrated its ability to generate high-quality responses, as evidenced by the hybrid and embedding model experiments in Sections 5.4 and 5.5, respectively. These experiments further established that the quality of responses is predominantly influenced by the retrieval phase rather than the LLM.

What was initially observed was that GPT-4 tended to disregard the prompt for generating answers. This observation was particularly notable in situations where it was not possible to provide an answer based on the retrieved contexts. Instead of applying the specified "Answering is not possible given the available information." it used a different answer, often starting with "The documents do not provide information..." and completing the sentence with specifics of the original question. This deviation from the expected behaviour was unexpected, considering that the smaller GPT-3.5 generally adhered to the prompt, utilising identical hyperparameters as GPT-4.

Answer LLM	Faithfulness	Answer Relevancy	Answer Correctness	Answer Similarity	BERTScore (F1)	Support
GPT-3.5 16k	0.802±0.24	**0.940±0.05**	0.722±0.17	0.8±0.11	**0.895±0.03**	117
GPT-4	**0.829±0.2**	0.933±0.05	**0.744±0.18**	**0.808±0.11**	0.892±0.04	**121**

Table 5.19: Average answer metric scores for TopK of 5 and chunk size of 128 for comparison between using GPT-3.5 16k or GPT-4 as the answer LLM.

Further, as seen in Table 5.19, while it could answer a larger number of questions with higher answer scores than GPT-3.5, it did not do so in a satisfying manner at times. An example was the QA-ID Q_G2_41, provided in Table 5.20, where identical five contexts were retrieved. While GPT-3.5 correctly refused to provide an answer due to insufficient contextual information, GPT-4 attempted to provide an explanation. These differences warrant further exploration in future work. Consequently, a recommendation can be proposed for GPT-3.5-16k, while stressing the necessity for additional fine-tuning when deploying GPT-4 as an answer model.

Answer Model	Question	Ground Truth	Generated Answer
GPT-3.5-16k	What is the cause of Shadowheart's unhealing wound?	The cause of the speaker's unhealing wound is Lady Shar's influence, which is also the reason for the terrible pain they experience.	Answering is not possible given the available information.
GPT-4	What is the cause of Shadowheart's unhealing wound?	The cause of the speaker's unhealing wound is Lady Shar's influence, which is also the reason for the terrible pain they experience.	Shadowheart's unhealing wound is not related to the tadpoles, but the exact cause is not specified in the documents (Document[2]).

Table 5.20: Answer difference by GPT-3.5 and GPT-4 for QA-ID Q_G2_46.

Chapter 6

Discussion and Outlook

The contributions of this thesis are threefold. Firstly, it introduces a small dataset of 153 QA pairs derived from the video game Baldur's Gate 3 and sourced from a YT Let's Plays series by the YouTuber WolfHeartFPS[1]. Secondly, a series of experiments were conducted at various stages of the RAG pipeline illustrating potential enhancements and identifying pitfalls that may arise during the implementation process. Lastly, the thesis introduces an End-to-End RAG evaluation pipeline for users to apply to their respective datasets. The pipeline utilises Haystack[11] for creating the pipeline and incorporates Ragas[10] for the evaluation process. Each of these contributions will be elaborated upon here.

The Baldur's Gate 3 (BG3) evaluation dataset was created on the premise of demonstrating that domain-specific, previously unknown data, can be applied to Large Language Models when using Retrieval Augmented Generation. It is duly acknowledged that the dataset exhibits imperfections and may necessitate significant additional refinement and calibration, particularly in light of the chaotic nature of the YouTube transcript, for the corpus to be deemed suitable as a benchmark in subsequent research. Nevertheless, the dataset serves as an initial point of consideration on how LLMs should be evaluated in the future, given the secrecy involved surrounding their training. The dataset is available on HuggingFace[2] and the GitHub repository[3].

The experiments in this thesis were evaluated by applying the Ragas library, which provided the basis for all interpretations made therein. Although not all metrics within the library were deemed helpful, a majority demonstrated value in quantifying issues during retrieval and answer generation, all without dependence on a labelled dataset. The incorporation of BERTScore and Source Context Accuracy into the evaluation process served to establish a comparative benchmark for the Ragas metrics.

[1]https://www.youtube.com/@WolfheartFPS
[2]https://huggingface.co/datasets/stucksam/BG3-QA-Dataset
[3]https://github.com/stucksam/vt2-rag-eval

During the evaluation of these experiments on the BG3 evaluation dataset, some noteworthy findings were made. First, the token chunk size and TopK experiment revealed that employing smaller chunks in larger quantities yields superior context retrievals compared to utilising larger chunks in smaller quantities. Specifically, the experiment with a chunk size of 128 and a TopK value of 10 demonstrated the best scores in Context Precision@k, Context Relevancy and Source Context Accuracy. Furthermore, it excelled in four out of the five answer metrics and exhibited the highest answer rate within this experimental framework. The second experiment, which employed different distance calculation techniques, indicated that the differences in accuracy were marginal. This suggests that such optimisations are likely to be runtime optimisations rather than accuracy optimisations.

The most profound influence on the overall system performance was observed with the implementation of a Hybrid RAG pipeline, as opposed to a dense search-only setup, in experiment three. This setup demonstrated the highest answer rate within this thesis, achieving a notable 92.16%, surpassing the benchmark's 76.47%. This setup may not benefit other datasets as much as the BG3 corpus, which, owing to the prevalence of game-specific keywords throughout the transcripts, was expected to perform better. Furthermore, it is plausible that the pipeline had an advantage due to the larger number of contexts that were retrieved, with each search methodology retrieving ten contexts, which were subsequently filtered down to five.

An additional important consideration involves the embedding model employed in both the creation of the vector index/database and the subsequent execution of dense searches. Experiment four conclusively demonstrated that the asymmetric search embedding model outperformed its symmetric counterparts by a substantial margin, achieving an impressive second-highest answer rate of 83% on the evaluation corpus. This underscores the significant performance enhancements attainable through the application of a matching embedding model to the target data.

In the final experiment, an examination of the disparities in answer LLMs was conducted, comparing the performance of GPT-3.5 with that of GPT-4. While not exhaustive in its scope, this experiment assessed whether a larger and more potent Large Language Model yields comparable answers. The findings revealed that GPT-4 tended to neglect the instructions provided by the answer prompt, particularly in instances where an answer could not be generated due to missing information in the retrieved context. In contrast, the default GPT-3.5 model consistently performed well throughout the thesis, making it a commendable recommendation as the answer LLM.

The last contribution lies in the implementation of the End-To-End General Retrieval Augmented Generation pipeline that was utilised in this thesis in the form of a parameterized Python file. All setups are available within the code, accessible on the GitHub repository[4], intending to enable further research in this area.

[4]https://github.com/stucksam/vt2-rag-eval

Future work may delve into the discovery in the distance and hybrid experiments, wherein an infrequent error (n=7) occurred. This anomaly was characterized by correct contexts being retrieved, but the LLM refused to answer the query. Remarkably, such occurrences were absent in the benchmark pipeline. Subsequent research may also explore the incorporation of non-OpenAI models, assessing disparities when employed both as the answer model and as the judge within the Ragas framework. Moreover, different metric libraries should be investigated to provide a more exhaustive overview of the RAG pipeline's performance.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., Jun. 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[3] OpenAI, "GPT-4 Technical Report," *CoRR*, vol. abs/2303.08774, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, July 2023. [Online]. Available: https://arxiv.org/abs/2307.09288

[5] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, "Siren's song in the ai ocean: A survey on hallucination in large language models," *CoRR*, vol. abs/2309.01219, 2023. [Online]. Available: https://arxiv.org/abs/2309.01219

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020,

pp. 9459–9474. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[7] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," *CoRR*, vol. abs/2306.05685, 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[8] E. Kamalloo, N. Dziri, C. Clarke, and D. Rafiei, "Evaluating open-domain question answering in the era of large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5591–5606. [Online]. Available: https://aclanthology.org/2023.acl-long.307

[9] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy, "Evaluating correctness and faithfulness of instruction-following models for question answering," *CoRR*, vol. abs/2307.16877, 2023. [Online]. Available: https://arxiv.org/abs/2307.16877

[10] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "Ragas: Automated evaluation of retrieval augmented generation," *CoRR*, vol. abs/2309.15217, 2023. [Online]. Available: https://arxiv.org/abs/2309.15217

[11] M. Pietsch, T. Möller, B. Kostic, J. Risch, M. Pippi, M. Jobanputra, S. Zanzottera, S. Cerza, V. Blagojevic, T. Stadelmann, T. Soni, and S. Lee, "Haystack: the end-to-end nlp framework for pragmatic builders," 2019. [Online]. Available: https://github.com/deepset-ai/haystack

[12] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *CoRR*, vol. abs/2002.08909, 2020. [Online]. Available: https://arxiv.org/abs/2002.08909

[13] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," *CoRR*, vol. abs/2104.07567, Apr 2021. [Online]. Available: https://arxiv.org/abs/2104.07567

[14] S. Huo, N. Arabzadeh, and C. L. A. Clarke, "Retrieving supporting evidence for llms generated answers," *CoRR*, vol. abs/2306.13781, June 2023. [Online]. Available: https://arxiv.org/abs/2306.13781

[15] G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, and T. Scialom, "Augmented language models: a survey," *CoRR*, vol. abs/2306.13781, June 2023. [Online]. Available: https://arxiv.org/abs/2302.07842

[16] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," *CoRR*, vol. abs/2305.14627, May 2023. [Online]. Available: https://arxiv.org/abs/2305.14627

[17] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin, "Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution," *CoRR*, vol. abs/2307.16883, July 2023. [Online]. Available: https://arxiv.org/abs/2307.16883

[18] K. Krishna, E. Bransom, B. Kuehl, M. Iyyer, P. Dasigi, A. Cohan, and K. Lo, "LongEval: Guidelines for human evaluation of faithfulness in long-form summarization," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1650–1669. [Online]. Available: https://aclanthology.org/2023.eacl-main.121

[19] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "Ares: An automated evaluation framework for retrieval-augmented generation systems," *CoRR*, vol. abs/2311.09476, Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.09476

[20] P. Xu, W. Ping, X. Wu, L. McAfee, C. Zhu, Z. Liu, S. Subramanian, E. Bakhturina, M. Shoeybi, and B. Catanzaro, "Retrieval meets long context large language models," *CoRR*, vol. abs/2310.03025, 2024. [Online]. Available: https://arxiv.org/abs/2310.03025

[21] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out.* Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[22] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "QMSum: A new benchmark for query-based multi-domain meeting summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 5905–5921. [Online]. Available: https://aclanthology.org/2021.naacl-main.472

[23] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N. A. Smith, and M. Gardner, "A dataset of information-seeking questions and answers anchored in research papers," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4599–4610. [Online]. Available: https://aclanthology.org/2021.naacl-main.365

[24] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette, "The NarrativeQA Reading Comprehension Challenge,"

Transactions of the Association for Computational Linguistics, vol. 6, pp. 317–328, 05 2018. [Online]. Available: https://doi.org/10.1162/tacl_a_00023

[25] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," *CoRR*, vol. abs/2309.01431, Sep 2023. [Online]. Available: https://arxiv.org/abs/2309.01431

[26] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023. [Online]. Available: https://arxiv.org/abs/2310.11511

[27] A. Balaguer, V. Benara, R. L. de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, and R. Chandra, "Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture," *CoRR*, vol. abs/2401.08406, 2024. [Online]. Available: https://arxiv.org/abs/2401.08406

[28] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[29] "Deeplearning.ai," 2023. [Online]. Available: https://www.deeplearning.ai

[30] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[31] E. Cardenas, "Vector library vs vector database," 2022. [Online]. Available: https://weaviate.io/blog/vector-library-vs-vector-database

[32] Meta, "FAISS - GitHub Repository," 2023. [Online]. Available: https://github.com/facebookresearch/faiss

[33] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," in *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 2010, pp. 1–8.

[34] E. Gomede, "Understanding the bm25 ranking algorithm," 2023. [Online]. Available: https://medium.com/@evertongomede/understanding-the-bm25-ranking-algorithm-19f6d45c6ce

[35] L. Monigatti, "Improving Performance in RAG pipelines with hybrid search," Nov 2023. [Online]. Available: https://towardsdatascience.com/improving-retrieval-performance-in-rag-pipelines-with-hybrid-search-c75203c2f2f5

[36] R. Benham and J. S. Culpepper, "Risk-reward trade-offs in rank fusion," in *Proceedings of the 22nd Australasian Document Computing Symposium*, ser. ADCS '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3166072.3166084

[37] I. Nguyen, "Haystack - hybrid document retrieval," Aug 2023. [Online]. Available: https://haystack.deepset.ai/blog/hybrid-retrieval

[38] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[39] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *CoRR*, vol. abs/2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[40] H. Chase, "LangChain," Oct. 2022. [Online]. Available: https://github.com/langchain-ai/langchain

[41] J. Liu, "LlamaIndex," Nov. 2022. [Online]. Available: https://github.com/jerryjliu/llama_index

[42] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *CoRR*, vol. abs/2206.07682, Jun 2022. [Online]. Available: https://arxiv.org/abs/2206.07682

[43] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," *CoRR*, vol. abs/2302.11382, Feb. 2023. [Online]. Available: https://arxiv.org/abs/2302.11382

[44] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan 2022. [Online]. Available: https://arxiv.org/abs/2201.11903

[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318.

[46] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*

for Machine Translation and/or Summarization, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[47] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[48] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[49] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-pack: Packaged resources to advance general chinese embedding," *CoRR*, vol. abs/2309.07597, 2023. [Online]. Available: https://arxiv.org/abs/2309.07597

[50] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "Mteb: Massive text embedding benchmark," *arXiv preprint arXiv:2210.07316*, 2022. [Online]. Available: https://arxiv.org/abs/2210.07316

[51] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," *CoRR*, vol. abs/1611.09268, 2016. [Online]. Available: http://arxiv.org/abs/1611.09268

[52] Y. Yang, W.-t. Yih, and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2013–2018. [Online]. Available: https://aclanthology.org/D15-1237

[53] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf

[54] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1601–1611. [Online]. Available: https://aclanthology.org/P17-1147

[55] TheGameAwards.com, “The Game Awards - Game of the year 2023,” https://thegameawards.com/nominees/game-of-the-year, Dez. 2023, [Online; Accessed 29.12.2023].

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. 5

BG3 Baldur's Gate 3. 1, 6, 22, 28, 32, 39, 41, 45–47, 50, 51

CoT Chain of Thought Programming. 15

DNN Deep Neural Networks. 5

GT Ground Truth. 18, 20, 21

HNSW Hierarchical Navigable Small World. 12

IDF Inverse Document Frequency. 11

IR Information Retrieval. 11, 14, 16

IVFx Inverted Index. 12

LLM Large Language Model. 5–8, 14–17, 20–22, 24, 26, 30, 31, 35, 37, 41, 43–45, 47–52, 61, 62

MRR Mean Reciprocal Rank. 16

NLI Natural Language Inference. 19

NLP Natural Language Processing. 5, 12, 22

NN Neural Networks. 5

NPC Non-Player Character. 42

RAG Retrieval Augmented Generation. 1, 5–8, 10, 11, 13–17, 22, 24, 28, 29, 31–34, 38, 39, 41, 43–46, 48, 50–52, 60, 61

RRF Reciprocal Rank Fusion. 13

TF Term Frequency. 11

YT YouTube. 6, 22, 24, 30, 50

Appendix A

Foundation Details

A.1 Ragas Prompts

A.1.1 Context Precision@k

Verify if the information in the given context is useful in answering the question.

question: What are the health benefits of green tea?
context:
This article explores the rich history of tea cultivation in China, tracing its roots back to the ancient dynasties. It discusses how different regions have developed their unique tea varieties and brewing techniques. The article also delves into the cultural significance of tea in Chinese society and how it has become a symbol of hospitality and relaxation. verification:
{{reason: The context, while informative about the history and cultural significance of tea in China, does not provide specific information about the health benefits of green tea. Thus, it is not useful for answering the question about health benefits., verdict:No}}

question: How does photosynthesis work in plants?
context:
Photosynthesis in plants is a complex process involving multiple steps. This paper details how chlorophyll within the chloroplasts absorbs sunlight, which then drives the chemical reaction converting carbon dioxide and water into glucose and oxygen. It explains the role of light and dark reactions and how ATP and NADPH are produced during these processes.
verification:
{{reason: This context is extremely relevant and useful for answering the question. It directly addresses the mechanisms of photosynthesis, explaining the key components and processes involved., verdict:Yes}}

question: {question}
context:
{context}
verification:

A.1.2 Context Recall

Given a context, and an answer, analyze each sentence in the answer and classify if the sentence can be attributed to the given context or not. Output json with reason.

question: What can you tell me about albert Albert Einstein?
context: Albert Einstein (14 March 1879 – 18 April 1955) was a German-born theoretical physicist,widely held to be one of the greatest and most influential scientists of all time. Best known for developing the theory of relativity, he also made important contributions to quantum mechanics, and was thus a central figure in the revolutionary reshaping of the scientific understanding of nature that modern physics accomplished in the first decades of the twentieth century. His mass–energy equivalence formula E = mc2, which arises from relativity theory, has been called ”the world's most famous equation”. He received the 1921 Nobel Prize in Physics ”for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect”, a pivotal step in the development of quantum theory. His work is also known for its influence on the philosophy of science. In a 1999 poll of 130 leading physicists worldwide by the British journal Physics World, Einstein was ranked the greatest physicist of all time. His intellectual achievements and originality have made Einstein synonymous with genius.
answer: Albert Einstein born in 14 March 1879 was German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time. He received the 1921 Nobel Prize in Physics ”for his services to theoretical physics. He published 4 papers in 1905. Einstein moved to Switzerland in 1895
classification:
{
{{ ”statement_1”:”Albert Einstein, born on 14 March 1879, was a German-born theoretical physicist, widely held to be one of the greatest and most influential scientists of all time.”,
”reason”: ”The date of birth of Einstein is mentioned clearly in the context.”,
”Attributed”: ”Yes”
}},
{{
”statement_2”:”He received the 1921 Nobel Prize in Physics 'for his services to theoretical physics.”,
”reason”: ”The exact sentence is present in the given context.”,
”Attributed”: ”Yes”
}},

{{
"statement_3": "He published 4 papers in 1905.",
"reason": "There is no mention about papers he wrote in the given context.",
"Attributed": "No"
}},
{{
"statement_4":"Einstein moved to Switzerland in 1895.",
"reason": "There is no supporting evidence for this in the given context.",
"Attributed": "No"
}}
}
question: who won 2020 icc world cup?
context: Who won the 2022 ICC Men's T20 World Cup?
The 2022 ICC Men's T20 World Cup, held from October 16 to November 13, 2022, in Australia, was the eighth edition of the tournament. Originally scheduled for 2020, it was postponed due to the COVID-19 pandemic. England emerged victorious, defeating Pakistan by five wickets in the final to clinch their second ICC Men's T20 World Cup title.
answer: England
classification:
{
{{
"statement_1":"England won the 2022 ICC Men's T20 World Cup.",
"reason": "From context it is clear that England defeated Pakistan to win the World Cup.",

"Attributed": "Yes"
}}
}

question:{question}
context:{context}
answer:{answer}
classification:

A.1.3 Faithfulness

Long-Form Answer Prompt

Create one or more statements from each sentence in the given answer.
question: Who was Albert Einstein and what is he best known for?
answer: He was a German-born theoretical physicist, widely acknowledged to be one of the greatest and most influential physicists of all time. He was best known for developing the theory of relativity, he also made important contributions to the development of the theory of quantum mechanics. statements in json:
{{
statements: [
Albert Einstein was born in Germany.,
Albert Einstein was best known for his theory of relativity.
]
}}
question: Cadmium Chloride is slightly soluble in this chemical, it is also called what?
answer: alcohol
statements in json:
{{
statements: [
Cadmium Chloride is slightly soluble in alcohol.
]
}}
question: Were Bach and Tchaikovsky of the same nationality?
answer: Sorry, I can't provide answer to that question.
statements in json:
{{
statements: []
}}
question:{question}
answer: {answer}
statements in json:

NLI Statements Message

Natural language inference. Only use Yes or No as verdict.

Context:
John is a student at XYZ University. He is pursuing a degree in Computer Science. He is enrolled in several courses this semester, including Data Structures, Algorithms, and Database Management. John is a diligent student and spends a significant amount of time studying and completing assignments. He often stays late in the library to work on his projects.
statement_1: John is majoring in Biology.
statement_2: John is taking a course on Artificial Intelligence.
statement_3: John is a dedicated student.
statement_4: John has a part-time job.
Answer:
{
{{
statement_1: John is majoring in Biology.,
reason: John's major is explicitly mentioned as Computer Science. There is no information suggesting he is majoring in Biology.,
verdict: No
}},
{{ statement_2: John is taking a course on Artificial Intelligence.,
reason: The context mentions the courses John is currently enrolled in, and Artificial Intelligence is not mentioned. Therefore, it cannot be deduced that John is taking a course on AI.,
verdict: No
}}
{{
statement_3: John is a dedicated student.,
reason: The context states that he spends a significant amount of time studying and completing assignments. Additionally, it mentions that he often stays late in the library to work on his projects, which implies dedication.,
verdict: Yes
}},
{{
statement_4: John has a part-time job.,
reason: There is no information given in the context about John having a part-time job.,
verdict: No
}}
}
Context:
Photosynthesis is a process used by plants, algae, and certain bacteria to convert light energy into chemical energy.
statement_1: Albert Einstein was a genius.
Answer:

```
{
{{
statement_1: Albert Einstein was a genius.,
reason: The context and statement are unrelated
verdict: No
}}
}

Context:
Albert Einstein was a German-born theoretical physicist who is widely held
to be one of the greatest and most influential scientists of all time.
statement_1: Nil
Answer:
{
{{
statement_1: Nil,
reason: The statement is invalid,
verdict: No
}}
}
context: {context}
statements: {statements}

Answer:
```

A.1.4 Answer Relevancy

Question Generation

Generate a question for the given answer and Identify if answer is noncommittal

Answer:
Albert Einstein was born in Germany.
Context:
Albert Einstein was a German-born theoretical physicist who is widely held to be one of the greatest and most influential scientists of all time
Output:
{{question:Where was Albert Einstein born?,noncommittal:false}}

Answer:
It can change its skin color based on the temperature of its environment.
Context:
A recent scientific study has discovered a new species of frog in the Amazon rainforest that has the unique ability to change its skin color based on the temperature of its environment.
Output:
{{question:What unique ability does the newly discovered species of frog have?,noncommittal:false}}

Answer:
Everest
Context:
The tallest mountain on Earth, measured from sea level, is a renowned peak located in the Himalayas.
Output:
{{question:What is the tallest mountain on Earth?,noncommittal:false}}

Answer:
I don't know about the groundbreaking feature of the smartphone invented in 2023 as am unware of information beyong 2022.
Context:
In 2023, a groundbreaking invention was announced: a smartphone with a battery life of one month, revolutionizing the way people use mobile technology.
Output:
{{question:What was the groundbreaking feature of the smartphone invented in 2023?, noncommittal:true}}
Answer: {answer}
Context: {context}
Output:

A.1.5 Answer Correctness

Correctness Prompt

Extract following from given question and ground truth
Question: What powers the sun and what is its primary function?
Answer: The sun is powered by nuclear fission, similar to nuclear reactors on Earth, and its primary function is to provide light to the solar system.
Ground truth: The sun is actually powered by nuclear fusion, not fission. In its core, hydrogen atoms fuse to form helium, releasing a tremendous amount of energy. This energy is what lights up the sun and provides heat and light, essential for life on Earth. The sun's light also plays a critical role in Earth's climate system and helps to drive the weather and ocean currents.
Extracted statements:
{
{{
statements that are present in both the answer and the ground truth: [The sun's primary function is to provide light],
statements present in the answer but not found in the ground truth: [The sun is powered by nuclear fission, similar to nuclear reactors on Earth],
relevant statements found in the ground truth but omitted in the answer: [The sun is powered by nuclear fusion, not fission, In its core, hydrogen atoms fuse to form helium, releasing a tremendous amount of energy, This energy provides heat and light, essential for life on Earth, The sun's light plays a critical role in Earth's climate system, The sun helps to drive the weather and ocean currents]
}}
}
Question: What is the boiling point of water?
Answer: The boiling point of water is 100 degrees Celsius at sea level.
Ground truth: The boiling point of water is 100 degrees Celsius (212 degrees Fahrenheit) at sea level, but it can change with altitude.
Extracted statements:
{{{
statements that are present in both the answer and the ground truth: [The boiling point of water is 100 degrees Celsius at sea level],
statements present in the answer but not found in the ground truth: [],
relevant statements found in the ground truth but omitted in the answer: [The boiling point can change with altitude, The boiling point of water is 212 degrees Fahrenheit at sea level]
}}}

Question:{question}
Answer: {answer}
Ground truth: {ground_truth}
Extracted statements:

Appendix B

Code & Manual

The code of this thesis is available on a GitHub repository at https://github.com/stucksam/vt2-rag-eval. The Baldurs Gate 3 QA-Dataset is available on both the GitHub repo as well as HuggingFace https://huggingface.co/datasets/stucksam/BG3-QA-Dataset.