THE TARDIS OPERATIONAL MANUAL: CREATING AND EVALUATING A SYNTHETIC DATASET FOR RETRIEVAL-AUGMENTED GENERATION

ALEXANDRE THUROW BENDER[1], ULISSES BRISOLARA CORRÊA[2],
RICARDO MATSUMURA ARAUJO[3]

[1] *Universidade Federal de Pelotas - atbender@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas - ulisses@inf.ufpel.edu.br*
[3] *Universidade Federal de Pelotas - ricardo@inf.ufpel.edu.br*

1. INTRODUCTION

Large language models (LLMs) have transformed natural language processing, leading to significant improvements in tasks like text generation, translation, and summarization (ZHAO et al., 2023). These advancements have made LLMs widely applicable across different fields, offering benefits such as greater coherence in generated text and the ability to understand context more effectively. As a result, interest in AI technologies has surged, highlighting the potential of LLMs to enhance productivity and creativity.

Despite their strengths, LLMs also face notable pitfalls and limitations (MALLEN et al., 2022). They rely heavily on the training data used to develop them, which can introduce biases and inaccuracies in their outputs (KANDPAL et al., 2022). Furthermore, LLMs are limited by their parametric knowledge – information encoded in the weights of the model during its training that remains unchanged unless the model is retrained. This creates a clear distinction between parametric knowledge and non-parametric knowledge, the latter of which can be updated dynamically from external sources.

To address these limitations, retrieval-augmented generation (RAG) has emerged as a promising approach. By combining the generative power of LLMs with external retrieval mechanisms, RAG systems have been shown to improve the relevance and accuracy of generated content (LEE; CHANG; TOUTANOVA, 2019). This integration allows models to access up-to-date information, mitigating the risk of producing outdated or misleading responses. However, evaluating RAG systems poses unique challenges, as it requires a thorough assessment of both the generative and retrieval components.

In this study, we introduce the TARDIS Operational Manual, a synthetic knowledge base and question-answering dataset created to assess retrieval-augmented generation systems. We evaluate the quality of this synthetic dataset by comparing it to a non-synthetic dataset, using agentic evaluation as the assessment method.

2. METHODOLOGY

Synthetic datasets are often seen as a more cost-effective but less reliable alternative to those based on real-world data. However, the extent of this tradeoff is frequently not fully understood, with the potential to reduce costs by up to 90% when using synthetic datasets. Moreover, advances in large language models have significantly improved the quality of modern synthetic datasets.

We would like to thank CAPES, FAPERGS, and the NVIDIA Corporation for the partial funding of this work.

The primary goal of the RAG evaluation datasets is to compile information with specific details that cannot be obtained without referencing the original documents. This information usually includes restaurant menus, frequently asked questions, technical specifications, or databases containing specialized domain knowledge. Many of these sources are private databases.

For this reason, we created a synthetic knowledge database focused on a fictional operational manual for the TARDIS[1]. This fictional knowledge database was developed by prompting large language models to generate a list of topics for the manual. Following this, the topic titles were used in another prompt to elaborate on each topic, ensuring that at least 1,000 words were provided for each. The topics were generated using OpenAI model *gpt-4o* and the content was created using *gpt-4o-mini*.

With the simulated private knowledge database containing topic items and their descriptions, we can utilize tools such as RAGAS (ES et al., 2023) for test set generation. This allows us to chunk the knowledge database, create questions for each contextual chunk, and develop these questions into reasoning or multi-step questions, among others. In this approach, seed questions are generated by a large language model and undergo evaluation iterations using a critic LLM, which is typically a more complex and expensive model. This process helps evolve the questions from basic to more complex types. For the experiments, we use *gpt-3.5-turbo* for the generator LLM and *gpt-4o-mini* for the critic LLM.

The dataset consists of questions generated through this process, along with the context chunks used to create each question. These context chunks contain the information needed to answer the questions. Additionally, we have the ground truth, which is the answer provided by a large language model based on the specific context. The dataset also includes a metadata field that describes the topic from which the context was extracted.

The ground truth in the generated dataset is produced using the context and a large language model. Therefore, we are naturally interested in evaluating the quality of this provided answer, as it directly reflects the overall quality of the dataset. Metrics for evaluation are frequently discussed, but in recent years, there has been an increase in agentic approaches. These methods assess the answers using LLMs and prompting. The RAGAS library offers several unified metrics that are applicable to RAG pipelines. One of the advantages of these approaches is that we can evaluate them using the context, allowing us to assess the quality of our ground truth.

Faithfulness refers to how well an answer $as(q)$ aligns with the context $c(q)$, meaning the claims made in the answer should be derivable from the context. To evaluate faithfulness, we first use an LLM to extract a set of statements $S(as(q))$, breaking down longer sentences into shorter, clearer assertions.

The final faithfulness score F is calculated using the formula:

$$F = \frac{|V|}{|S|}$$

Here, $|V|$ represents the number of statements that the LLM supports, while $|S|$ is the total number of statements extracted. This score indicates the proportion of statements in the answer that are validated by the context.

[1] Time and Relative Dimension in Space (TARDIS), the time machine from the series Doctor Who.

Answer Relevance indicates that the answer $as(q)$ is relevant if it directly addresses the question appropriately. The assessment of answer relevance focuses on completeness and avoids penalizing for factuality. To evaluate relevance, a large language model is prompted to generate n potential questions q_i based on $as(q)$.

We then obtain embeddings for all questions using the *text-embedding-ada-002* model from the OpenAI API. For each q_i, we compute the similarity $sim(q, q_i)$ with the original question q as the cosine similarity between the corresponding embeddings. The answer relevance score AR for question q is calculated as:

$$AR = \frac{1}{n}\sum_{i=1}^{n} sim(q, q_i) \tag{1}$$

This score evaluates how closely the generated answer aligns with the initial question or instruction.

3. RESULTS AND DISCUSSION

The TARDIS Operational Manual document collection comprises 100 topics, each with a description of at least 1,000 words. Each document provides a detailed series of instructions related to the topic. This collection simulates private company knowledge bases, which typically contain domain-specific information relevant to a company's area of expertise.

Table 1 shows example topics from the generated dataset (the 1,000-word content for each has been omitted for brevity).

Generated Topics Sample
How to Calibrate the Chameleon Circuit
Configuring the TARDIS Navigation System
How to Repair the TARDIS Console Components

Table 1: Examples of the topics generated to comprise the TARDIS Operational Manual knowledge database. Each topic contains a 1,000-word document elaborating on the subject.

We can also inspect some of the questions that were generated, these are shown below in Table 2. The questions here maintain the Doctor Who theme, involving fantasy concepts and equipment from the series universe.

Generated Questions Sample
What is the purpose of the Dimensional Stabilization Field in the TARDIS?
What steps can Time Lords take to troubleshoot time distortions in their TARDIS?
What is the significance of the Control Room in the operation of a TARDIS?

Table 2: Examples of the questions generated for different document contexts.

We evaluate the TARDIS dataset using two key metrics: faithfulness and answer relevancy. These metrics assess how well the answer aligns with the actual context, which is important as we are evaluating the dataset itself and the quality of its ground truth in comparison to chunked contexts from the documents. The results are presented in Table 3. To establish a baseline for these metrics, we compare the TARDIS dataset to a standard non-synthetic dataset, Amnesty QA. [2].

Dataset	Faithfulness	Answer Relevancy
TARDIS	0.946 ± 0.18	0.862 ± 0.32
Amnesty QA	0.592 ± 0.35	0.925 ± 0.22

Table 3: Evaluation results.

The results show significantly higher faithfulness for the TARDIS dataset, demonstrating that its answers are more consistently grounded in the chunked context (i.e., the document presumed to contain the answer). However, the answer relevancy for the TARDIS dataset is lower than that of the Amnesty dataset, revealing that the responses are not always complete. This suggests a greater discrepancy between the factual questions and the potential generated questions used in this evaluation.

4. CONCLUSION

This work presents the TARDIS Operational Manual, a synthetic dataset created to assess retrieval-augmented generation systems. We evaluated its quality using an agentic approach and compared it to another baseline dataset. Although both synthetic datasets and agentic evaluations have their limitations, they offer a valuable method for situations where resources to evaluate retrieval-augmented generation systems are limited. Future research could incorporate more metrics, additional comparison baselines, and a broader selection of critic and generator models.

REFERENCES

ES, S.; JAMES, J.; ESPINOSA-ANKE, L.; SCHOCKAERT, S. Ragas: Automated evaluation of retrieval augmented generation. **arXiv preprint arXiv:2309.15217**, 2023.

KANDPAL, N.; DENG, H.; ROBERTS, A.; WALLACE, E.; RAFFEL, C. Large language models struggle to learn long-tail knowledge. arxiv. **arXiv preprint arXiv:2211.08411**, 2022.

LEE, K.; CHANG, M.-W.; TOUTANOVA, K. Latent retrieval for weakly supervised open domain question answering. **arXiv preprint arXiv:1906.00300**, 2019.

MALLEN, A.; ASAI, A.; ZHONG, V.; DAS, R.; KHASHABI, D.; HAJISHIRZI, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. **arXiv preprint arXiv:2212.10511**, 2022.

ZHAO, W. X.; ZHOU, K.; LI, J.; TANG, T.; WANG, X.; HOU, Y.; MIN, Y.; ZHANG, B.; ZHANG, J.; DONG, Z. et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, 2023.

[2] $https://huggingface.co/datasets/explodinggradients/amnesty_qa$