Evaluation and Analysis

This project is evaluated on a dataset that contains 100 questions with respect to Apple's Annual Earnings Report for 2022 fiscal year, as well as "ground-truth" answers to the questions and corresponding contexts excerpted from the annual report.

1. Evaluation dimensions and metrics

To thoroughly evaluate the architecture of the project, evaluation is carried out on two dimensions, i.e. the intermediate contexts and the final answers.

Meanwhile, given that the intermediate contexts and the final answers are in the format of natural language where texts consistent of different words may have the same or similar meaning, cosine similarity to the ground truth is used as the evaluation metric.

Specifically, the same encoder (sentence-transformers/all-MiniLM-L6-v2) as used in the Retriever is first used to convert ground-truth answers and generated answers to embeddings, and then cosine similarity between the ground-truth answer and the generated answer is calculated for each question in the dataset. The same procedure is applied to the pre-defined contexts from the dataset and the contexts retrieved in the project.

Additionally, how a source document is chunked can impact the context to be retrieved and thus impact the final answer. Therefore, three scenarios are evaluated and compared: 1) the document is chunked into sentences and the most relevant sentence is selected as the context; 2) if the context resulting from 1) has length less than 256, the next most relevant sentence will be appended to the context until its length is no less than 256; 3) the document is split into 100-token chunks and the most relevant chunk is retrieved as the context.

2. Criteria for evaluation

As cosine similarity ranges from 0 to 1, a cosine similarity score closer to 1 means a generated answer/retrieved context is more similar to the corresponding ground-truth answer/pre-defined context.

Besides, the same version(7B-Chat) of Llama 2 alone without context retrieval is used as a baseline for performance comparison.

3. Results interpretation

	Final answer similarity	Intermediate context similarity
Baseline - no context	0.569551	-
RAG one sentence context	0.625027	0.68319
RAG sentence-based context with minimum length of 256	0.661820	0.722181
RAG 100-tokens context	0.665909	0.771046

The baseline is generation by Llama 2 alone without context retrieval, while Llama 2 was trained between January 2023 and July 2023, which means it could be “knowledgeable” of this annual report of Apple or related information. On top of that, our project still significantly outperforms the baseline as shown in the above table.

As expected, performance varies among different document chunking methods. Splitting the document into sentences and only retrieving the most relevant sentence as the context has the worst result, which makes sense as some sentences might be too short to provide useful information. When we increase the length of context by the other two methods, the results improve by a large margin, in terms of both the final answers and the intermediate contexts.

4. Potential ways for improvement

In this project, different document chunking methods are explored to improve the results and they do improve the results significantly, which is done on the Tokenizer component of the project.

There may be more ways for improvement on the other two components, i.e. the Retriever and the Generator. In the context retrieval process, the k most relevant chunks are selected as the context, and the default value of k is 1, while adjusting the value of k may lead to better results. Besides, a better Generator may also improve the results as the context similarity scores are higher than the answer similarity scores as shown in the above table.