



Knowledge Media Institute

Research Companion App

Tech Report
Sep 2024

Author: Oke Onyebuchukwu
KMi Supervisors: Dr Venetia Brown and Audrey Ekuban

Abstract

This paper introduces the Research Companion App, designed to assist students in comprehending research papers through interactive features. Utilising Retrieval Augmented Generation (RAG) and prompt engineering, the app enhances response accuracy, scalability, and flexibility. The app, built with Streamlit and Hugging Face, and leveraging Microsoft Azure OpenAI models, includes functionalities such as summarisation, visual aids generation, quizzes, and a chatbot. Evaluation of the chatbot's performance across different temperature settings reveals that higher temperatures generally yield more consistent and reliable outputs. However, the quality of responses is contingent on the retrieval process. Future improvements aim to refine summary generation and explore better QA evaluators.

Introduction

This paper introduces the Research Companion App, which provides support in understanding research papers. The application uses interactive features to let students more quickly grasp research papers. This can support students who balance their studies with part-time jobs. The app helps all students be more time efficient by letting them concentrate more on applying the material they've read rather than spending time trying to understand it. The app also simplifies advanced topics, making them easier to understand. This can support students who find themselves needing to study research papers from a field they're less familiar with.

The application uses Retrieval-Augmented Generation (RAG) to generate the best responses to any questions you have about the text. RAG works in two phases: retrieval (retrieving relevant information from a large database) and generation (generating a coherent output using a generative model). RAG improves the accuracy, scalability, and flexibility of the product.

The application also employs prompt engineering practices to improve the quality of output. Prompt engineering is the practice of designing and refining the input prompts given to a large language model to achieve specific and desired outputs. The code employs methods such as crafting effective prompts, using examples and providing context and constraints. As part of the production process, prompts were refined to improve the output's accuracy, relevance, and quality.

Prompt refinement was also essential in mitigating hallucinations. Hallucinations are where the language model generates incorrect information, made up or not based on the data it was trained on.

One of the variables in the language model we used was temperature. Temperature typically varies from 0 to 1, and a higher temperature will cause the output to be more random/creative. It will be discussed later whether changing temperature affects the quality of the output given. Next, the features of the application will be discussed.

Overview of Application

The Research Companion Application uses Streamlit [1] (a Python framework for deploying web applications) and Hugging Face [2] (a Machine Learning collaboration platform). Microsoft Azure OpenAI [3] provided the gpt4-o model for chat completion and the text-embedding-ada-002 model for text embedding. The app includes features such as a summariser, a visual aids generator, a quiz, and a chatbot.

Figure 1 displays the interface for file uploading and summarising and snippets of the code for the summarise function. The application allows you to upload any Word doc or PDF saved on your computer. The user is allowed to upload a maximum of 3 files. Clicking the summary button provides a short summary of the document.

Research Companion App

Upload File (Word doc or PDF)

Drag and drop file here
Limit: 200MB per file

[Browse files](#)

Cold War.docx 15.6KB x

Select file

Cold War.docx v

[Delete file](#)

[Summarise](#)

Summary

The Cold War was a period of geopolitical tension between the United States and the Soviet Union, along with their respective allies, the Western Bloc and the Eastern Bloc. This era began in 1947, two years after the end of World War II, and lasted until the fall of the Soviet Union in 1991.

The term "cold war" is used because there was no large-scale fighting directly between the two superpowers. Instead, they supported opposing sides in major regional conflicts, known as proxy wars. The Cold War was rooted in an ideological and geopolitical struggle for global influence, following the roles of the United States and the Soviet Union as Allies in World War II, which led to their victory against Nazi Germany and Imperial Japan in 1945.

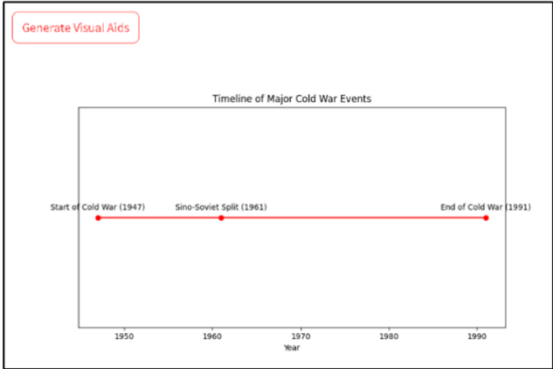
The struggle for dominance during the Cold War was expressed through various means, including the nuclear arms race, conventional military deployment, psychological warfare, propaganda campaigns, espionage, embassies, sports diplomacy, and technological competition like the Space Race.

```
def summarise(text):
    text_was_trimmed = False
    if len(text) > 20000:
        text = text[:20000]
        text_was_trimmed = True
    summary = client.chat.completions.create(
        model=deployment,
        temperature=0,
        messages=[
            {"role": "system", "content": "You are a professional AI assistant who summarises text"},
            {"role": "user", "content": f"Summarise the following, breaking up your answer into nicely readable paragraphs: {text}"}
        ]
    ).choices[0].message.content
    if text_was_trimmed:
        summary += '\n\nNote: Only the first 20000 characters of the text should\'ve been summarised'
    return summary
```

Figure 1 - File upload and summary

The interface and snippets of the code for generating visual aids are shown in Figure 2. Clicking the Generate Visual Aids button generates an appropriate graph displaying information from your selected document

[Generate Visual Aids](#)



Timeline of Major Cold War Events

Event	Year
Start of Cold War	1947
Sino-Soviet Split	1961
End of Cold War	1991

```
background_info = f"""You provide code that lets us generate images based off a prompt using matplotlib. Make sure to only use information from the prompt when generating images, don't use your own knowledge. Include imports if you need to, don't include any introductory text, just the code. Save the figure at the location of plot_path, which is plot.png. Ensure to add padding to the layout using fig.tight_layout(pad=5.0) like in the example given below. If the text suggests there's a correlation between two variables, draw a line of best fit. For example for this prompt ({example_prompt}), you should generate this code: {example_code}"""
prompt = f"Write a script that saves a matplotlib image based on the following context: {text}"
image_generation_code = client.chat.completions.create(
    model=deployment,
    temperature=0,
    messages=[
        {"role": "system", "content": background_info},
        {"role": "user", "content": prompt}
    ]
).choices[0].message.content
# Define the pattern to match from 'import' to 'close(fig)'
pattern = r"(import[\s\S]*?close\(\fig\))"
match = re.search(pattern, image_generation_code)
if match:
    image_generation_code = match.group(1)
else:
    raise Exception("Couldn't extract image_generation_code from model")
exec(image_generation_code)
return "plot.png"
```

Figure 2 - Visual aids generator

Figure 3 shows the quiz interface and code snippets for generating feedback. The quiz provides detailed feedback on each answer to improve your understanding of the document.

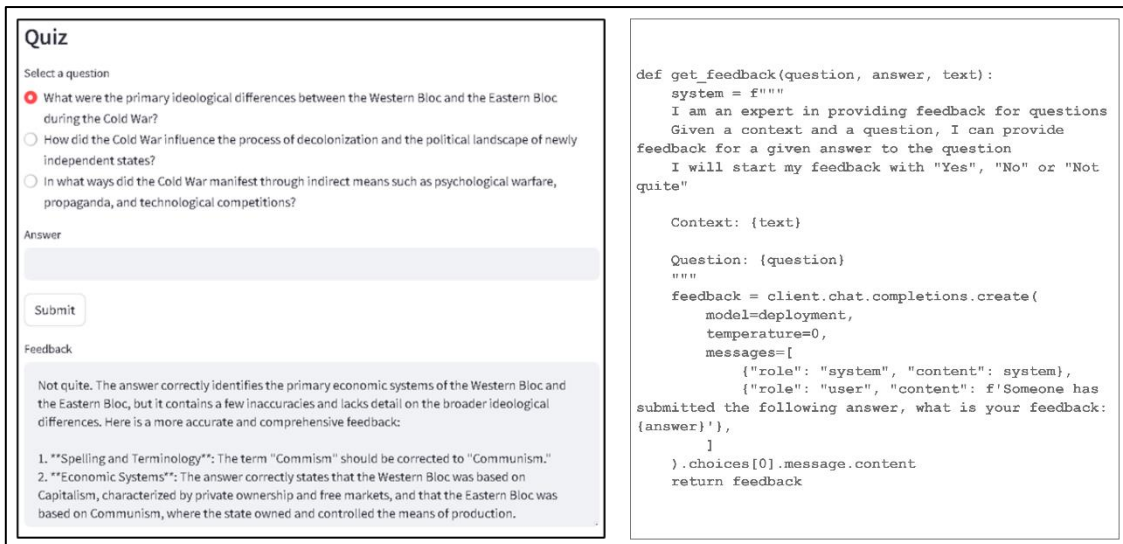


Figure 3 - Quiz and feedback

Figure 4 presents the chatbot interface in addition to code snippets for obtaining ratings. You can ask the chatbot questions about any of your uploaded documents. A 1-5 rating is provided, giving an idea of how accurate the response is based on its given context. The next section will discuss how the ratings for a prompt vary with temperature.



Figure 4 - Chatbot and Rating

Evaluation

The chatbot uses context given from the 3 uploaded files, as well as prior chat messages, to answer queries. A QA evaluator from Hugging Face is used to provide 1-5 ratings on the quality of the chatbot’s responses based on the context for 3 different temperatures. Figures 5 and 6 are examples of how responses may vary for the same prompt based on the temperature of the language model. In this case, Figure 5 includes a concluding sentence, while Figure 6 doesn’t.

User: In what ways did the Cold War manifest through indirect means such as psychological warfare, propaganda, and technological competitions?

Assistant: The Cold War manifested through various indirect means, including:

- Psychological Warfare:** Both superpowers engaged in efforts to influence public opinion and morale. This included spreading disinformation, conducting covert operations to undermine the other side’s political stability, and using media to shape perceptions.
- Propaganda Campaigns:** The United States and the Soviet Union used propaganda to promote their ideologies and discredit the other. This included broadcasting radio programs, distributing pamphlets, and using films and literature to spread their messages.
- Espionage:** Both sides invested heavily in intelligence operations to gather information and conduct covert activities. Agencies like the CIA and KGB played crucial roles in spying on each other and influencing political events in various countries.
- Embargoes:** Economic measures were used to exert pressure. For example, the U.S. imposed trade embargoes on the Soviet Union and its allies to weaken their economies.
- Sports Diplomacy:** International sports competitions, such as the Olympics, became arenas for demonstrating national superiority. The superpowers sought to showcase their athletic prowess as a reflection of their societal and ideological strengths.
- Technological Competitions:** The Space Race is a prime example, where the U.S. and the Soviet Union competed to achieve significant milestones in space exploration. This included the launch of satellites, manned space missions, and ultimately, the U.S. landing on the moon in 1969.

These indirect methods were crucial in the Cold War, as they allowed both superpowers to exert influence and compete for global dominance without engaging in direct military conflict.

Rating: 4.426

Figure 5

User: In what ways did the Cold War manifest through indirect means such as psychological warfare, propaganda, and technological competitions?

Assistant: The Cold War manifested through various indirect means, including:

- Psychological Warfare:** Both superpowers engaged in efforts to influence public opinion and morale. This included spreading disinformation, conducting covert operations to undermine the other side’s political stability, and using media to shape perceptions.
- Propaganda Campaigns:** The United States and the Soviet Union used propaganda to promote their ideologies and discredit the other. This included broadcasting radio programs, distributing pamphlets, and using films and literature to spread their messages.
- Espionage:** Both sides invested heavily in intelligence operations to gather information and conduct covert activities. Agencies like the CIA and KGB played crucial roles in spying on each other and influencing political events in various countries.
- Embargoes:** Economic measures were used to exert pressure. For example, the U.S. imposed trade embargoes on the Soviet Union and its allies to weaken their economies.
- Sports Diplomacy:** International sports competitions, such as the Olympics, became arenas for demonstrating national superiority. The superpowers sought to showcase their athletic prowess as a reflection of their societal and ideological strengths.
- Technological Competitions:** The Space Race is a prime example, where the U.S. and the Soviet Union competed to achieve significant milestones in space exploration. This included the launch of satellites, manned space missions, and ultimately, the U.S. landing on the moon in 1969.

Rating: 4.4828

Figure 6

For temperatures 0 and 0.5, the language model produced the output in Figure 5 with rating 4.42. For temperature 1, the language model produced the output in Figure 6 with a rating of 4.4828. This procedure is repeated 9 times for different prompts to produce Table 1:

Table 1 Ratings for Cold War Document

Prompt	Temperature		
	0	0.5	1
In what ways did the Cold War manifest through indirect means such as psychological warfare, propaganda, and technological competitions?	4.4260	4.4260	4.4828
How did the Cold War influence the process of decolonisation and the political landscape of newly independent states?	4.2910	4.4100	4.4771
What were the primary ideological differences between the Western and Eastern Bloc during the Cold War?	4.6540	4.6540	4.6540
What was the Cold War, and when did it take place?	4.9905	4.9905	4.9905
Why is the term "Cold War" used to describe the period of tension between the United States and the Soviet Union?	4.9708	4.9708	4.9757
What were some of the indirect methods used by the superpowers to express their struggle for dominance during the Cold War?	4.9908	4.9873	4.9817
What marked the beginning and the end of the Cold War?	4.9896	4.9900	4.9900
How were the Western Bloc and Eastern Bloc different in terms of political and economic systems?	4.5468	4.5804	4.6276
How did the Cold War affect former colonial states during the period of decolonisation?	4.5006	4.2704	4.5844
What role did the United States and the Soviet Union play in regime changes around the world during the Cold War?	3.6067	4.3294	4.4782
Mean	4.5967	4.6609	4.7242
Standard deviation	0.4375	0.2994	0.2321

Table 2 contains some more ratings taken from various documents and prompts.

Table 2 Ratings for various documents

Prompt number	Temperature		
	0	0.5	1
1	4.9745	3.8795	4.1768
2	4.9685	4.4691	4.3629
3	4.9812	4.4096	4.9532
4	4.9799	4.3437	4.9644
5	4.7315	4.1668	4.8749
6	4.9833	4.5392	4.2969
7	4.5771	3.7909	4.9736
8	4.4380	4.9767	4.5128
9	4.9545	4.9867	4.9855
10	4.3970	4.9848	4.6054
Mean	4.8986	4.4547	4.6706
Standard deviation	0.2061	0.4488	0.3138

Discussion

This report set out to assess the quality of the chatbot's responses based on the context given in the uploaded documents. It's worth noting that the QA evaluator is assumed to be accurate.

For the Cold War document, we can see a robust set of ratings, with a mean rating above 4.5 for each of the temperatures. There was, however, an unacceptable rating of 3.6067 given for one of the responses, likely because the question asked wasn't as closely related to the context as the other questions. It can be observed in Table 1 that the ratings tend to be higher and more consistent with higher temperatures. It can be concluded that, with a temperature of 1, the Research Companion App will produce reliable outputs for the Cold War document. Using a temperature of 1 should prevent unacceptable ratings from arising. However, using a temperature of 1 may not be optimal for all documents.

In Table 2, we can again observe an array of strong ratings, with a mean rating of 4.6746. The standard deviation is also low, on average 0.3. This suggests that the output of the chatbot tends to be of high quality for most documents, prompts and temperatures. Nevertheless, it can be observed that there are 2 ratings below 4, which is not ideal. This is likely due to a drawback of using RAG: the reliance on the quality of the retrieval process. The output produced won't be high quality if inaccurate or irrelevant information is retrieved from the databases. From the table, we can see that this problem happens least for a temperature of 0, so fixing the temperature to 0 seems to be a temporary solution. Further measures for improving the application will be discussed below.

Next Steps

To improve the application, steps can be taken to produce short summaries reliably. Currently, the summaries given are too long and trying to truncate them leads to summaries having incomplete sentences at the end. If there had been more time, I could have crafted more examples for the large language model to improve the quality of outputs. In addition, more exploration can be done on Hugging Face to find a higher quality QA evaluator to get more accurate evaluation results.

References

1. Streamlit Homepage: <https://streamlit.io/>. Last accessed 03-September-2024
2. Hugging Face Homepage: <https://huggingface.co/>. Last accessed 03-September-2024
3. Microsoft Azure OpenAI Service page: <https://learn.microsoft.com/en-us/azure/ai-services/openai/overview>. Last accessed 03-September-2024

Acknowledgements

I want to acknowledge my supervisors, Audrey Ekuban and Dr Venetia Brown, and the KMi team at the Open University, who assisted and provided the credits for Microsoft Azure OpenAI's service. The KMi Scholarship for Black Students supported this work.