Report: Part 3 Pinecone VectorDB and RAG

Dataset

The dataset we chose contains news stories and summaries for each story (likely used to train a model), for our purposes we used the stories as context for the RAG pipeline specifically the column `text`.

Link to the dataset

Rationale

As mentioned in the lecture and the tutorial, LLMs suffer from hallucinations, general understanding of topics and static knowledge (up to a certain cutoff), thus this dataset helps us mitigate these obstactles:

1. **Rich Contextual Information:**

   The dataset covers a broad spectrum of news topics, ensuring that the RAG model can handle a wide range of queries and generate relevant responses across different domains. more over, the full text of news stories offers a wealth of contextual information that is crucial for generating accurate and detailed responses in the RAG pipeline.

2. **Hallucinations:**

   Providing the relevant contexts (using the vector DB) will help us to minimize hallucinations as much as possible.

3. **Static Knowledge:**

   The specfic LLM we used has a knowledge cutoff up to January 2024, however, the dataset features stories that are after that date, and the dataset may feature some data that the LLM wasn't trained on. Thus using this dataset will improve the performance and the reliability of the LLM.

4. **Enhancing Response Quality:**

   As we mentioned before LLMs are great with general knowledge of topics, yet struggle with deep understanding of these topics, for example they might struggle citing a figure like new-worth so they tend to answer with a random number. Using the dataset especially with the high level of detail in the stories will allow the LLM to provide accurate, conscise and (hopefully) correct figures and information

Anecdotal Examples

To evaluate the effictiveness of the RAG pipeline we defined a function that prompts the LLM once with the query without any additional context and another time with the context provided from the vector database.

```python
def check_effictiveness(query):
    results = []
    results.append(rag.prompt(query, add_context=False))
    results.append(rag.prompt(query, add_context=True))
    df = pd.DataFrame(results, columns=['text','query',
'source_knowledge'])
    df.index = ['no', 'yes']
    df.index.name = 'context?'
    display(HTML(df.to_html()))
```

Our process involved choosing a target context (1 story from the dataset), formulating a question about it, and then evaluating the model

Example 1

Query: *On what date will facebook discontinue Moments no need to mention year?*

Target context: *Facebook will discontinue its standalone private photo and video-sharing app 'Moments', launched in 2015, on February 25 later this year. Users can retrieve the content stored on the app by either storing it on Facebook or downloading it to their device via a Facebook link. Facebook attributed the discontinuation to lesser people using the app but didn't share user numbers.*

LLM w no context: Facebook discontinued Moments on June 25.

LLM w context: February 25 **context**:

- Facebook will discontinue its standalone private photo and video-sharing app 'Moments', launched in 2015, on February 25 later this year. Users can retrieve the content stored on the app by either storing it on Facebook or downloading it to their device via a Facebook link. Facebook attributed the discontinuation to lesser people using the app but didn't share user numbers.

- Facebook is reportedly testing solar-powered internet drones, to beam internet connectivity from the Earth's stratosphere, in Australia with aeronautics company Airbus. It was in talks with Airbus to conduct test flights, scheduled for November and December 2018, with Airbus' Zephyr drone, the report added. In June 2018, Facebook had closed its solar-powered aircraft-building facility in the UK.

- Facebook is testing 'LOL' feature, a dedicated feed consisting of meme videos and other viral content. The feature, with content categorised by topics like 'For You', 'Animals', and 'Fails' is currently in private beta with around 100 high school students. Facebook said it is still finalising if the feature will become part of the main app or a standalone app.

Analysis

First of, we can see that the LLM without the context didn't provide the correct answer and instead returned a random date likely due to hallucenations. After providing the LLM the correct context we can wee that it in fact returned the correct answer. Important thing to note is that the question purposefully asks the LLM to omit the year, otherwise the LLM with context returned an incorrect year, which is to be expected since its answering based on the context provided to it. Finally, looking at the results returned by the vector database, we can see that most importantly we retrived the target context needed to answer the query, interestingly, the database returned other contexts that discuss additional projects by facebook just like "Moments", however, these contexts were not needed to answer the query.

Example 2

Query: *Which video was Huwaei accused to copying for an ad?*

Target context: *Chinese telecoms giant Huawei has been accused of copying a music video in an ad to promote its tablet 'MediaPad M3 Lite' by New Zealand composer Nigel Stanford. Huawei's ad video, allegedly similar to Stanford's video 'Cymatics: Science Vs. Music' released in 2014, was flagged by Stanford himself. "We're looking into this matter and will share further updates," Huawei said*

LLM w no context: In 2018, Huawei was accused of copying a video created by a Dutch filmmaker, Wouter Bouvijn, for one of their smartphone ads. The original video, titled "The Story of a Peculiar Motion," was allegedly imitated in Huawei's "Beauty Has No Borders" ad campaign.

LLM w context: Huawei was accused of copying Nigel Stanford's video "Cymatics: Science Vs. Music" for its ad promoting the 'MediaPad M3 Lite' tablet. **context**:

- Chinese telecoms giant Huawei has been accused of copying a music video in an ad to promote its tablet 'MediaPad M3 Lite' by New Zealand composer Nigel Stanford. Huawei's ad video, allegedly similar to Stanford's video 'Cymatics: Science Vs. Music' released in 2014, was flagged by Stanford himself. "We're looking into this matter and will share further updates," Huawei said.
- US President Donald Trump defended high school students who appeared to confront a Native American man in a viral video. The "students were treated unfairly with early judgements proving out to be false - smeared by media", Trump tweeted. Other videos showed a group of black protestors hurling slurs at the students, before their encounter with the Native American man.
- As part of the viral #10YearChallenge, Anil Kapoor shared a collage of his videos from various years [1989, 1999, 2009 and 2019] and wrote, "Forget the #10YearChallenge, take the #AKChallenge!" The video includes songs from Anil's films 'Ram Lakhan', 'Taal', a scene from 'Slumdog Millionaire' and the latest song from his upcoming film 'Ek Ladki Ko Dekha Toh Aisa Laga'.

Analysis

In this example, the LLM without the context provided an incorrect answer, citing a different incident involving Huawei and a Dutch filmmaker, which is not related to the query about the specific video accusation. This demonstrates the LLM's tendency to generate plausible but incorrect responses when not supplied with the precise context.

After providing the LLM with the correct context, it accurately identified that Huawei was accused of copying Nigel Stanford's video "Cymatics: Science Vs. Music" for its ad promoting the 'MediaPad M3 Lite' tablet. This highlights the importance of context in guiding the LLM to generate accurate and relevant responses.

Additionally, the retrieved contexts included other unrelated information about various viral videos and media incidents. Although these contexts were not necessary for answering the specific query about Huawei, they demonstrate the broader scope of content available in the database, which can sometimes yield extraneous information. The key takeaway is that providing targeted context significantly enhances the LLM's ability to produce precise and relevant answers.

Example 3

Query: *What was the name of the student that was killed in Australia at La Trobe University? and who were they?*

Target context: *An Israeli student was killed by an unknown assailant in Australia while she was on video call with her sister. Aiia Maasarwe had been in Australia for about six months on a study abroad program at La Trobe University. A detective said police believed it was a random attack, however, they hadn't ruled out possibility Maasarwe had been sexually assaulted.*

LLM w no context: Zhongjun (Zack) Qiao, a 22-year-old Chinese national, was a student at La Trobe University in Australia. He was killed in a tragic incident on the university's Melbourne campus in 2022.

LLM w context: Aiia Maasarwe was the name of the student who was killed in Australia while on a study abroad program at La Trobe University. She was an Israeli national. **context**:

- An Israeli student was killed by an unknown assailant in Australia while she was on video call with her sister. Aiia Maasarwe had been in Australia for about six months on a study abroad program at La Trobe University. A detective said police believed it was a random attack, however, they hadn't ruled out possibility Maasarwe had been sexually assaulted.
- A 21-year-old student who was pursuing LLB from Delhi University was found hanging in the room of her paying guest accommodation in Noida on Sunday night. She allegedly committed suicide after a man whom she had befriended online didn't pick her calls following an argument, the police said. Her parents haven't filed a complaint, the police added.
- Archaeologists have found the remains of British explorer Captain Matthew Flinders, who is credited with naming Australia, near a railway station in London. Captain Flinders led the first circumnavigation of Australia. The discovery of his burial site was made as archaeologists were preparing the site where a railway station will be built.

Analysis

In this example, the LLM without context provided an incorrect answer, naming a different student and incident unrelated to the query about the specific student killed at La Trobe University. This response highlights the LLM's tendency to generate erroneous information in the absence of relevant context.

When provided with the correct context, the LLM accurately identified Aiia Maasarwe as the student who was killed in Australia while on a study abroad program at La Trobe University. It correctly stated that she was an Israeli national. This underscores the importance of supplying the LLM with precise context to guide it towards generating accurate and pertinent answers.

The retrieved contexts included additional unrelated information about other tragic incidents and discoveries, such as a student's suicide in Delhi and the discovery of Captain Matthew Flinders' remains. While these contexts were not necessary for answering the specific query about Aiia Maasarwe, they demonstrate the variety of information available in the database. The key takeaway is that providing the targeted context significantly improves the LLM's ability to deliver correct and relevant responses.

Retrival System & Prompts

Retrival System

Our retrival system is built on top of Pinecone Vector database, given a query we find the top 3 most relevant documents in the database according to cosine similarity (which was declared when creating the index).

The choice of top 3 closest was made since the topics discussed in the data are relatively sparse and not related to each other, so in most cases you need at most 1-2 documents to answer a given query.

Prompts

We noticed that the LLM provides a lengthy and detailed answer, however, we wanted a shorter more precise and conscise answer, so after each query we added the following prompt:

```
[query] + be consciense and get straight the point, maximum of 3-4 lines,
if you don't know say you don't know
```

Finally when we want to add context to our query we used the following:

```
Using the contexts below, answer the query.
Contexts:
[context1, context2, context3]
If the answer is not included in the source knowledge - say that you don't
know.
Query: [query]
```

Insights

- In order to try and reduce hallucinations we added `if you don't know say you don't know`, we tried several combinations of this sentence the queries we presented as examples and other topics, however, we found out that the one we added provided the best (although not optimal) performance, since according to the examples we can see that it did in fact hallucinate.
- Finding the examples we provided above was not as easy as we thought it would be, since the LLM has a relatively recent knowledge cutoff (~Jan 2024) and has dynamic knowledge fetching, making it harder for us to find examples that the LLM won't succeed in answering correctly, however, as we can see above we managed to find examples.
- Our strategy to find relevant examples was to ask the LLM a question with very specific details about the topic, which provided us with the best results.