Lesson 06: Benchmarking and Evaluating LLM Capabilities

Overview:

In this activity, you will engage in a hands-on demonstration using LangChain and ChromaDB to create a MultiPDF retriever. By following the provided instructions, you will gain practical experience with these tools, enhancing your proficiency in leveraging LangChain and ChromaDB.

Instructions:

1. Use **Demo 02: MultiPDF QA Retriever with ChromaDB and LangChain.ipynb** for this activity.
2. Tweak **Step 6: Passing the Query** to retrieve answers from multiple PDF files for the questions given in the task below.
3. Check your answers with the key answers provided at the end of the activity.

Tasks:

Task 1: Retrieve answers for the questions provided below:

1. What is the concept of Attention with Linear Biases (ALiBi) introduced in the paper?
2. How does ALiBi compare to the original sinusoidal position method in terms of efficiency and extrapolation abilities?
3. What is the role of training sequence length in transformer-based language models?
4. How does ALiBi facilitate efficient extrapolation in transformer models?
5. What is the impact of ALiBi on the performance of transformer language models on the WikiText-103 benchmark?

Additional Task (Optional)

If time permits, perform the below task:

1. Review all the PDF documents used in this task and extract the answers formulated from those files.

Answer Key

Step 6: Passing the Query

query = " What is the concept of Attention with Linear Biases (ALiBi) introduced in the paper?"
docs = vectordb.similarity_search(query)
print(docs[0].page_content)

query = " How does ALiBi compare to the original sinusoidal position method in terms of efficiency and extrapolation abilities?"
docs = vectordb.similarity_search(query)
print(docs[0].page_content)

query = " What is the role of training sequence length in transformer-based language models?"
docs = vectordb.similarity_search(query)
print(docs[0].page_content)

query = " How does ALiBi facilitate efficient extrapolation in transformer models?"
docs = vectordb.similarity_search(query)
print(docs[0].page_content)

query = " What is the impact of ALiBi on the performance of transformer language models on the WikiText-103 benchmark?"
docs = vectordb.similarity_search(query)
print(docs[0].page_content)