Final Project: Large Language Models

Ella Bryant

**Exercise 1.1.** What are the space and time complexities for the (non-sparse) scaled dot product attention mechanism applied to a single input as a function of

* L, the number of tokens being processed,
* d, the dimension of the input and output, and
* , the embedding dimensions for the keys, values and quereies?

The scaled dot product attention mechanism is the primary method for computing context vectors in the current iteration of transformer models. For this to happen, several steps must be completed: compute the attention scores, compute attention weights, and compute context vectors. Before starting, the space complexities of the input query, key, values matrices is as follows: Θ(Ldq + Ldk + Ldv). Next, the attention scores were computed using the following formula 𝑎𝑙𝑝ℎ𝑎𝑖𝑗 = 𝑞𝑖𝑘𝑗 √𝑑𝑘 where j = 1,…,L. Therefore, the space complexity is that of the output matrix Θ(𝐿 2 ). The time complexity is represented by the matrix multiplication of query and key matrices such that Θ(𝐿 2𝑑𝑘). Then, the computation of the attention weight from the function is as follows 𝑤𝑗 = 𝑠𝑜𝑓𝑡𝑚𝑎𝑥(𝑎𝑙𝑝ℎ𝑎𝑖 ). Since the complexity of the softmax function is linear, the space/time complexity is the same as the input: Θ(𝐿 2 ). Finally, the context vector is computed based on ℎ𝑖 = ∑ 𝑤𝑖𝑗𝑣𝑗 𝐿 𝑗=1 . Therefore, the time complexity is Θ(𝐿 2𝑑𝑣 ), and the space complexity is that of the output matrix: Θ(𝐿𝑑𝑣 ). Overall, the space complexity is Θ(𝐿𝑑𝑞 + 𝐿𝑑𝑘 + 𝐿𝑑𝑣 + 𝐿 2 + 𝐿𝑑𝑣 ) = Θ (L(dq + dk + 𝑑𝑣 + 𝐿)), and the time complexity is Θ(L 2𝑑𝑘 + 𝐿 2 + 𝐿 2𝑑𝑣 ) = Θ(L 2 (𝑑𝑘 + 𝑑𝑣 )).

**The source code and fine-tuned model for the following project can be found at** [**ViolettGee/Retrieval-Augmented-Generation-Framework**](https://github.com/ViolettGee/Retrieval-Augmented-Generation-Framework)**.**

**Exercise 1.3.** You may choose any task of your choice for this project and you may browse through HuggingFace Hub or the interwebs for ideas. You should pick a dataset that has at least 1,000 examples in it.

Your report should contain the following sections.

* A task description. Describe the task that you would like to use LLMs to solve. What was your motivation for choosing this task? What kinds of capabilities are required for an LLM to do well at this task?
* A model description. State the foundation model and your reasons for choosing it. How many parameters are in the model? What is its structure? Does it rely on binary16 or bfloat16 instructions that are commonly found in CPUs? What computational resources does it take to run?
* A data description. Describe the data used for fine-tuning the model on this task and for assessing performance. How many examples are contained in the train, evaluation, and test splits? Show at least one example of the data for your taks.

**Task Description.** The project explored the usage of a retrieval augmented generation framework on specializing a language model to a certain type of information. This is essentially mapping each user query to reference data to specialize the knowledge of a language model. Right now, the most prevalent problem with language models is hallucinations. A hallucination is when a language model produces a response that looks factual but is fabricated. The phenomena appear most when asking a model about data it does no have information on or when it references a different information context. Thus, the goal is to give the language model the background data or context on a query using the retrieval augmented generation framework. In the context of this project, the specialized topic is the impacts of advancements in AI on mental health.

**Model Description.** The model used in this project is a retrieval augmented generation model which couples both an embedding model and large language model in the following framework. Specifically, the database of the specialized data is embedding using an embedding model. This embedding is saved as a vector database to be referred to later. When a user queries the model, that query is embedded using the same embedding model. The embedded user query is then compared to the data examples in the vector database finding the top k matches. In the context of our model, the matches were determined based on cosine similarity. Then, the text of the top matches is compiled with the original user query to form an optimum prompt. This prompt is then passed to the language model to retrieve an output. For this project, the embedding model was fine-tuned from the “jinaai/jina-embeddings-v2-base-en” model. This model is based on a BERT architecture with about 137 million parameters. The model required usage of the sentence transformers library. For this project, the language model was from the “gpt-3.5-turbo” model. The model uses a transformer architecture with about 174 billion parameters and uses bfloat16 precision. The model required usage of the OpenAI library.

**Data Description.** The data used within this model was compiled from free-text articles found from searching “Machine Learning/AI and Mental Health: in PubMed. The references to these articles can be found in the “References” document on the GitHub repository. The training data was compiled in the format: “User Query”, “Text”, “Title” as shown below. There were 1014 training examples. The user query was found by prompting OpenAI to generate a question about the given text for each example.

A screenshot of a white and orange text

AI-generated content may be incorrect.

The database data was compiled in the format: “Title”, “Text” as shown below. There were 1017 database examples.

A screenshot of a computer

AI-generated content may be incorrect.

No standard evaluation dataset was used when training the model. The testing data was a combination of the database data shown above and 3 user queries to evaluate responses to shown below.

Question 1: "What role can AI play in predicting mental health crises before they occur, and how can we validate such predictions ethically?"

Question 2: “How can multimodal AI systems (text, voice, facial expression) be integrated to provide more accurate mental health assessments?"

Question 3: "How can AI contribute to large-scale mental health research without violating individual consent or autonomy?"

**Exercise 1.4.** Obtain a baseline on the performance of the foundation model on your task. Your report should contain the following section.

* Obtaining a baseline. How are you evaluating the baseline model. You should present one training instance and the model’s corresponding output.

**Obtain a baseline.** The baseline model can have two definitions in the scope of this project: retrieval augmented generation using pre-trained embedding model and language model without retrieval augmented generation framework. Thus, each model—retrieval augmented generation using fine-tuned embedding model, retrieval augmented generation using pre-trained embedding model and stand-alone language model—was used to generate responses for the 3 user queries described above. The responses of each of the models are described in Exercise 1.6.

**Exercise 1.5.** Fine-tune the model for your task. You are free to choose whether you tune the full model or use parameter-efficient tuning method such as adapter or a LoRA variant.

Your report should contain the following selection.

* Method for finetuning. Describe your approach to fine-tuning. You shou also include a discussion of any major choices you had to make, such as the number of training epochs, checkpointing, validation during training etc.

**Method for Fine-tuning.** Due to the nature of the retrieval augmented framework, fine-tuning was completed on a pre-trained embedding model such that the training example pairs would be embedded more similarly. The inputs were a pair between the user query and text with labels of 1.0 to indicate a match. The loss function was selected as cosine similarity for comparison. Since the training data size was relatively small, an epoch value of 5 was selected with a corresponding learning rate of 0.005

**Exercise 1.6.** Discuss the results of your fine-tuning. Your report should contain the following section.

* Compare and contrast the performance of your model before and after fine-tuning. Present the model’s response on the baseline examples you gave.

**Compare and Contrast.**

Question 1: "What role can AI play in predicting mental health crises before they occur, and how can we validate such predictions ethically?"

|  |  |
| --- | --- |
| Base Language Model Output | The response effectively answers both questions in a robust way. The model’s answers are effectively more descriptive than both other models. However, the same limitations as before still remain that there is no clear way of determining where the information came from. |
| Base RAG Output | The response answers the first portion of the question explaining the role AI can play in mental health. However, the response was incredibly generalized, and it had a logical leap from analyzing data to immediately predictive capabilities. The response attempted to answer more concerning ethical implications despite not pulling references to text containing commentary on the ethics involved. Thus, the responses to ethical commentary were not very descriptive in nature and it made some logical leaps and claims that it was not able to back up. Still, none of the statements were necessarily incorrect. |
| Fine-Tuned RAG Output | The response answers the first portion much to the same caliber as the Base RAG Output. The topic is addressed, however, not entirely descriptive. The major difference is this response effectively tackles the ethical implications explaining the need for informed consent among all parties involved and explaining the justification of model effectiveness to those parties as well. The response overall seems better put together and more descriptive. However, this could be due to confirmation bias. |

Question 2: How can multimodal AI systems (text, voice, facial expression) be integrated to provide more accurate mental health assessments?”

|  |  |
| --- | --- |
| Base Language Model Output | The response effectively answers the question robustly. It gives a complete list of different modalities and the information they can obtain. However, the response does not outline specifically how they are combined, and the same limitations as before remain that there is no clear way of determining where the information came from. |
| Base RAG Output | The response answers the question lacking technical detail about how modalities are integrated with one another. The response lacks depth for what each modality is measuring. Overall, the response is incredibly general and does not give a descriptive answer. This is likely due to the references having minimal relation to the question with 3 out of the 5 only mentioning the effects of standard Generative AI on mental health. |
| Fine-Tuned RAG Output | The response effectively answers the question robustly. It gives a complete list of different modalities and the information they obtain. The response is not significantly different from the language model’s response with the only difference being that its slightly more descriptive about how these multimodal systems can better help with mental health. Still, the response does not add descriptions on how they get combined. However, there is no clear way of determining where the information came from as the references appear to not heavily relate to the question. |

Question 3: “How can AI contribute to large-scale mental health research without violating individual consent or autonomy?”

|  |  |
| --- | --- |
| Base Language Model Output | The response answers the question detailing several examples. Moreover, the response does not take into account that just because data is publicly available does not mean the individuals effectively consented to its usage such as copyrighted material which is a fact conflicting hallucination. Still, the response has the same limitations as before where there is no way of determining where the information came from. |
| Base RAG Output | The response effectively iterates how to go about getting protecting individuals through both informed consent and giving participants-controlled autonomy. However, the details are not iterated in exactly how AI can contribute to mental health research, not fully answering the question. The lack of robust answer is interesting when considering that the references were relevant to the question in this instance. |
| Fine-Tuned RAG Output | The response effectively iterates several examples for how to increase individual consent or autonomy in AI research. These are more detailed than either of the other models. Still, there is a lack of answer for AI contributing to large-scale mental health research besides a singular ambiguous line. The lack of complete answer is interesting when considering that the references were relevant to the question in this instance. The response overall seems better, but this could be due to confirmation bias. |

The above questions show that the Base RAG model is predominantly the least effective model. However, due to the limited number, there is a high likelihood that sampling bias may be at play. In these examples, the Base Language Model and Fine-Tuned RAG Model do not consistently have significant differences in responses. However, this is an improvement as we can be more confident about where the information is coming from and the RAG model is less likely to produce hallucinations for this reason. We see this phenomenon at play in the Base Language Model response to question 3. Still, there is a need to acknowledge the chance of confirmation bias in the analysis of the Fine-Tuned RAG Model. The reference cosine similarity measures pulled as the top matches in the Fine-Tuned RAG Model are an increase of about 0.1 when compared to the Base RAG Model.