CSCI 420 LLM Final Project Report

Jinyan Kuang

Exercise 1.1

Let D = max(d, dk, dq, dv). Then the complexities are given as follows:

- Time Complexity: $\Theta(L^2 \cdot dk)$
  - This was dominated by the computation of the attention scores matrix QK^T
- Space Complexity: $\Theta(L^2 + L \cdot D)$
  - This stores the attention weights matrix and the projected vectors Q, K, V, and output

These expressions reflect the dominant costs associated with the attention mechanism. The L^2 term comes from computing and storing the full attention score matrix between tokens, while the $L \cdot D$ term accounts for linear projections and output storage.

Exercise 1.2 Skip

Exercise 1.3

Task Description
In this project, I fine-tuned GPT-2 on a mental health conversation dataset from the HuggingFace Hub. The goal was to explore how a small, accessible model could respond to mental health-related queries. While more advanced chatbots today are trained using GPT-3 or GPT-4 with real counseling data or prompt engineering, this project demonstrates that even a simple prototype trained on open resources can show meaningful progress.

The task was to build a language model that could engage in empathetic conversations similar to those between counselors and clients. My motivation comes from my research with Professor Janice Zhang in the HCI lab, where we investigate how technology can support college students dealing with stress, anxiety, depression, and loneliness. We believe AI has strong potential in mental health due to its 24/7 availability and emerging capacity for empathy. Despite skepticism from many psychotherapists and the importance of human connection, this project explores a promising direction.

Model Description
The foundation model used is GPT-2 (124M parameters). This model is a small transformer-based language model. It includes multi-head attention, feedforward layers, and uses GELU activation and layer normalization. The model operates on 16-bit floating point precision and can run on typical GPUs and some CPUs without requiring BF16 or specialized hardware. I used top-k and top-p sampling for generation.

Data Description
The dataset includes a collection of questions and answers from online counseling platforms. I used the data loader to input the data. The questions span various mental health topics, and the responses are provided by licensed psychologists. This dataset was used to fine-tune GPT-2 to better simulate counselor-like responses. Although the dataset is relatively small, it provides a realistic structure for testing the model's response capabilities.

> *Example entry:*
> *Q: I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?*
> *A: If everyone thinks you're worthless, then maybe you need to find new people to hang out with.Seriously, the social context in which a person lives is a big influence in self-esteem.Otherwise, you can go round and round trying to understand why you're not worthless, then go back to the same crowd and be knocked down again.There are many inspirational messages you can find in social media. Maybe read some of the ones which state that no person is worthless, and that everyone has a good purpose to their life.Also, since our culture is so saturated with the belief that if someone doesn't feel good about themselves that this is somehow terrible.Bad feelings are part of living. They are the motivation to remove ourselves from situations and relationships which do us more harm than good.Bad feelings do feel terrible. Your feeling of worthlessness may be good in the sense of motivating you to find out that you are much better than your feelings today.*

Exercise 1.4: Obtaining a Baseline

To establish a baseline, I used the pretrained GPT-2 language model (124M parameters) from the HuggingFace Transformers library. This model is a general-purpose causal language model that has not been fine-tuned for mental health support. My goal was to observe how it performs on counseling-style prompts before any adaptation.

I evaluated the model in two ways:

1. **Qualitative Evaluation** — I selected a real training instance from my mental health dataset and prompted the model using the user query. I then compared the model's output to the reference response written by a licensed psychologist.
2. **Log-Likelihood Scoring** — I used the model to compute the per-token negative log-likelihood (NLL) of the reference response. A lower NLL indicates that the model finds the sentence more probable and fluent.

Baseline Evaluation

Prompt: I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?

GPT-2 Output: My first impulse was to just sit there and cry. I know I know the answers. I know I've been there and done it. I know I'm going to find a way to make this

Reference Response: If everyone thinks you're worthless, then maybe you need to find new people to hang out with.Seriously, the social context in which a person lives is a big influence in self-esteem.Otherwise, you can go round and round trying to understand why you're not worthless, then go back to the same crowd and be knocked down again.There are many inspirational messages you can find in social media. Maybe read some of the ones which state that no person is worthless, and that everyone has a good purpose to their life.Also, since our culture is so saturated with the belief that if someone doesn't feel good about themselves that this is somehow terrible.Bad feelings are part of living. They are the motivation to remove ourselves from situations and relationships which do us more harm than good.Bad feelings do feel terrible. Your feeling of worthlessness may be good in the sense of motivating you to find out that you are much better than your feelings today.

Negative Log-Likelihood (NLL): 3.0054

Analysis: The baseline GPT-2 response is fluent and grammatically correct, but it does not meaningfully address the user's question. Instead of offering guidance or empathetic support, the model echoes the tone of the prompt and continues the internal monologue, which results in a response that feels vague and somewhat incoherent. It fails to demonstrate the empathetic structure typically expected in counseling, and does not help the user reflect or move away from distorted thoughts. This highlights the limitations of a general-purpose language model when applied to emotionally sensitive, domain-specific tasks without fine-tuning.

Exercise 1.5: Fine-Tuning the Model

To improve the model's performance on this task, I fine-tuned GPT-2 using a dataset of real mental health Q&A pairs. Each entry consists of a user prompt and a psychologist-provided answer. I used the HuggingFace Trainer API and the DataCollatorForLanguageModeling to train the model with a standard causal LM objective.

Training Configuration:

- Base model: GPT-2 (124M)
- Epochs: 3
- Batch size: 4
- Max sequence length: 256 tokens
- File: train_data.csv

Exercise 1.6: Evaluation After Fine-Tuning

After training, I repeated the same evaluation on the fine-tuned model used the same prompt from the dataset:

Prompt: I'm going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here. I've never tried or contemplated suicide. I've always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?

Fine-Tuned GPT-2 Output: The best way to move from your "feelings" into positive self-talk and change yourself is to work on them. There is no wrong way to feel. Being able to change yourself as well as living in the present

Negative Log-Likelihood (NLL): 0.822

Analysis: The fine-tuned model produces a response that is more empathetic, structured, and aligned with the tone of professional mental health support. Unlike the baseline GPT-2 output, which was vague and disconnected from the user’s emotional needs, the fine-tuned model provides affirming language and begins to introduce constructive therapeutic ideas such as “positive self-talk” and “living in the present.” While the response is still somewhat incomplete, it clearly reflects a shift toward the language and tone used by real counselors. The significantly lower negative log-likelihood (0.822) indicates that the model has learned to assign higher

probability to this kind of structured, supportive response—evidence of successful domain adaptation.

Conclusion: By comparing the baseline GPT-2 model with the fine-tuned version, I observe a clear improvement in both linguistic quality and domain relevance. The fine-tuned model demonstrates more appropriate tone, therapeutic framing, and content tailored to mental health support. This is supported by both qualitative observations and quantitative evaluation: the negative log-likelihood dropped substantially from the baseline, indicating that the model has learned to prefer more human-aligned and supportive responses. These results highlight the potential of adapting large language models for specialized, sensitive applications such as mental health.