Build LLM from Scratch Research Assignment

Building a Task-Specific Small Language Model from Scratch

Overview

In our previous sessions, we explored the core components of transformer-based language models and how they can be trained from scratch. Your next assignment will involve building your own small language model using a domain of your choice, based on the Colab notebook shared below.

You will receive access to:

- A Google Colab notebook that walks through building a GPT-style language model from scratch:
  Colab: Build a GPT Language Model from Scratch
- The lecture recording and documentation (accessible on the course dashboard)

Task Description

Your assignment consists of the following steps:

Step 1: **Choose a field or domain** of interest (examples: Legal contracts, Medical records, Mythology texts, Movie subtitles, Recipe instructions, etc.)

Step 2: **Assemble a dataset** containing task-specific or domain-specific text data. Preprocess it into a suitable format for tokenization and training.

Step 3: **Use the provided Colab notebook** to train your own language model on this dataset. You may modify model architecture, tokenizer, context length, or training loop to suit your domain.

Step 4: **Evaluate your model** by generating meaningful completions or outputs specific to your dataset.

Research Questions

Alongside your implementation, respond to the following questions in your submission:

Q1. What challenges did you face in creating a domain-specific dataset? How large was your final training corpus?

Q2. How did your model perform in generating fluent, domain-specific text? Provide qualitative examples.

Q3. Did you make any changes to the model architecture or training pipeline? If so, why?

Q4. How would you improve your model if you had access to more compute or data?

Collaboration and Submission

- You may work in teams of up to **three students**.
- **Final Submission Deadline: 2nd June**
- Please submit:
  - A short report (PDF) addressing the research questions
  - Link to your modified Colab notebook
  - Sample outputs from your model

Future Opportunity

Selected submissions may be showcased on Vizuara's Youtube or LinkedIn channel!

All the best!