CS 696 Applied Large Language Models
Spring Semester, 2025
Assignment 2v2
San Diego State University -- This page last updated on 2/25/25

Assignment 2
LLMs Generating Project Ideas for LLM Project

Due March 4, 10 pm

1. Run one model from assignment one on the GPU cluster.

2. Compute the memory used by each of the models.

Select one of the models for the rest of the assignment

3. Decrease the number of hidden layers and attention heads by 1/2 each. How does this affect the a) output of the model, b) the memory size, and c) model runtime? If the amount of memory used does not decrease, the number of layers and heads will not decrease.

Create a notebook that contains the results. The notebook should be a report, not just a listing of code and output. Explain what you are doing and comment on the results.

What to turn in.

Download the .ipynb file and turn it in.

You will turn in the assignment in Canvas.

Grading

	Percent of Grade
Each Problem	25%
Report	25%

Late Policy

An assignment turned in 1-7 days late, will lose 10% of the total value of the assignment per day late. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.