CS 696 Applied Large Language Models
Spring Semester, 2026
San Diego State University -- This page last updated on 3/17/26

Midterm Exam
Part 1
Due Mar 20

1. In a RAG pipeline, why might a vector-based "Semantic Search" fail to retrieve the correct document for the query "Project X-15 Status Report," and how would "Hybrid Search" resolve this?

2. How do Query's transformations and reranking chunks improve RAG?

3. How does an LLM "know" which MCP tool to call?

4. If a student provides 20 retrieved documents to an LLM, but the answer is located in the 10th document, the model often fails to find it. Explain this in the context of the Attention Mechanism and positional biases.

5. Compare the advantages and disadvantages of fine-tuning a pretrained LLM versus continuous pretraining. Provide examples of how each approach is more beneficial.

6. Explain LoRA (Low-Rank Adaptation) and QLoRA. How do these techniques enable efficient fine-tuning of large models on consumer-grade GPUs? What advantage does QLoRA have over LoRA? What advantage does LoRA have over QLoRA?

7. Define catastrophic forgetting in the context of LLMs. Propose and evaluate strategies to mitigate this issue when fine-tuning an LLM on a domain-specific dataset.

Part 2
Due Mar 29

You likely will need to use the GPU cluster for the following. See lecture document 13 for instructions on accessing the cluster.

8. Fine-tune the microsoft/Phi-4-mini-instruct model on the Magicoder-OSS-Instruct-75K dataset, https://huggingface.co/datasets/ise-uiuc/Magicoder-OSS-Instruct-75K. Fine-tune it two ways: modify all parameters and use LoRA. For each method, report and comment on the following:
    1. Time required to train the model
    2. Vram memory used to train each model
    3. Model size after training
    4. Performance on the MBPP (Mostly Basic Python Problems) data (https://github.com/google-research/google-research/tree/master/mbpp) on tasks 20 - 70.

Your report and comment should be at least multiple paragraphs. You need to explain what the numbers mean, not just report them.

9. Use bitsandbytes to quantize the models you produced in #7. Compare the memory (CPU and GPU) and run time required by a FP16 quantized model, an 8-bit quantized model, and a 4-bit model. Is there any difference between the fully trained and the LoRA-trained model in their quantized versions? How does the 4-bit model perform on the MBPP task 20-70, compared to the LoRA model from #7?
10. Given the number of parameters in the model from #7 compute the savings in the model by the 4-bit model. How does your computed value compare to the actual size reduction of the model?
11. Can you detect any catastrophic forgetting in the model from #7 trained by modifying all the parameters?
12. (Extra Credit) Train the Qwen3-8B model on the Magicoder-OSS-Instruct-75K data using Unsloth and QLoRA. Compare the performance to the best model from #7.

What to turn in.

If any answer is determined to be a direct output of an LLM, or otherwise copied from the internet, it will result in a 50-point deduction per question. Answers to questions 1 through 6 can be turned in as a PDF or a Word Document. Turn it in on canvas.

Answers to questions 8+ are to be done in a Jupyter notebook in the style of assignment 1. Put your notebook(s) into a zip file and turn it in on Canvas. The zip file name must contain your name.

Grading

Problems	Percent of Grade
1 - 7	5 points each
8	40 points
9	20 points
10	10 points
11	5 points
12 (Extra Credit)	15 points

Late Policy

An exam part turned in late will lose 20% of the total value per day late.