# Rubric for HW6

## Q1 SFT (30 points)

### Q1.1 Task 1 (6 points)

- Correct (6 points)
- Explanation states that we ignore question token predictions because we want the model to learn to generate responses, not predict the input/question (3 points)
- Explanation states that masking with -100 ensures the loss function only considers answer tokens (2 points)
- Response is 1-2 paragraphs in length (1 point)

### Q1.2 Task 2 (24 points)

- Correct (24 points)
- All 7 hyperparameters are specified (batch_size, gradient_accumulation_steps, learning_rate, max_epochs, patience, lora_r, lora_drop) (3 points)
- Model achieves format rate greater than 90% on test set (3 points)
- Model achieves accuracy greater than 30% on test set (among properly formatted responses) (3 points)
- Response clearly states whether overfitting was observed (yes or no) (2 points)
- Response describes specific hyperparameter changes made or that would be made to address overfitting (e.g., adjusting learning rate, dropout, early stopping patience, or LoRA rank) (3 points)
- Response clearly states whether slow training or memory issues were encountered (yes or no) (2 points)
- Response describes specific hyperparameter changes made or that would be made to address these issues (e.g., adjusting batch_size, gradient_accumulation_steps, or learning_rate) (3 points)
- Response identifies which task is more challenging: formatting instructions OR mathematical reasoning (2 points)

- Response provides reasoning that explains why one task is more challenging than the other (3 points)

# Q2 RAG (50 points)

## Q2.1 Task 1 (8 points)

- Correct (8 points)
- Basic prompting code generates answers using only the question as the prompt (2 points)
- Instruction prompting code generates answers using the format “Instruct: {question}\nOutput:” (2 points)
- Explanation describes qualitative differences between basic and instruction prompting (e.g., instruction format produces more focused/concise responses, basic prompting generates additional questions) (2 points)
- Explanation mentions that SFT trains the model to follow the instruction format (1 point)
- Response is 1-2 paragraphs in length (1 point)

## Q2.2 Task 2 (6 points)

- Correct (6 points)
- Code runs evaluate_model with mode=“no_context” and max_examples=100 (2 points)
- Code runs evaluate_model with mode=“gold_context” and max_examples=100 (2 points)
- Code prints both accuracy and average time for each evaluation (2 points)

## Q2.3 Task 3 (20 points)

- Correct (20 points)
- compute_embeddings function: tokenizes input texts with padding=True, truncation=True, and max_length=512 (3 points)
- compute_embeddings function: uses mean pooling across the sequence dimension (2 points)

- Convert final embeddings to numpy arrays on CPU (2 points)
- Returns stacked numpy array (np.vstack) of embeddings (2 points)
- retrieve_top_context function: computes query embedding using compute_embeddings (2 points)
- retrieve_top_context function: computes cosine similarity using vectorized numpy operations (3 points)
- retrieve_top_context function: uses np.argmax to find index of highest similarity (2 points)
- retrieve_top_context function: returns the context string at the top-ranked index (2 points)
- Retrieval should achieve at least 50% accuracy (1 point)
- Retrieval should average less than 100ms per query (1 point)

### Q2.4 Task 4 (16 points)

- Correct (16 points)
- Total time should average less than 1 second per query (3 points)
- RAG system achieves at least 50% accuracy on question answering (tested on 100 examples) (3 points)
- Response quantifies the improvement in accuracy that RAG provides over no context (2 points)
- Response discusses the runtime overhead of RAG (1 point)
- Response describes at least one circumstance when RAG would be beneficial (e.g., frequent updates, large data, factual grounding) (3 points)
- Response is 2-3 sentences in length (1 point)
- Response mentions at least one advantage AND one disadvantage of RAG compared to fine-tuning (2 points)
- Response is 2-3 sentences in length (1 point)

## Q3 LLM Risk (20 points)

### Q3.1 Task 1 (7 points)

- Correct (7 points)

- Identifies at least two distinct societal challenges or risks from the article (3 points)
- Explains each challenge/risk in the student's own words (2 points)
- Response is 1-2 paragraphs in length (1 point)
- References concepts from the article (challenges must be actually discussed in the paper) (1 point)

## Q3.2 Task 2 (7 points)

- Correct (7 points)
- Summarizes at least two specific suggestions from Section 7 of the article (3 points)
- Takes a clear position on whether these suggestions can mitigate Task 1 risks (2 points)
- Connects the paths forward to the specific risks identified in Task 1 (1 point)
- Response is 1-2 paragraphs in length (1 point)

## Q3.3 Task 3 (6 points)

- Correct (6 points)
- Takes a clear position (agree or disagree with the claim) (2 points)
- Provides at least two reasons or pieces of reasoning to support the position (2 points)
- Response is 1-2 paragraphs in length (1 point)
- Reasoning demonstrates engagement with the topic (not generic or superficial) (1 point)