Name: _____ **Roll No.:** _____

Exam: Set 2

Course: DS:246 — Generative and Agentic AI in Practice

Instructions: This exam consists of **20 MCQs** (1 mark each) and **10 MSQs** (2 marks each). Indicate the correct option(s) clearly. For MSQs, multiple answers may be correct. Write the correct option(s) inside the box provided at the right side of each question.

Set 2 — 30 questions

(Each question lists full text and options. Type: MCQ = single correct answer, MSQ = multiple correct answers.)

1. (MCQ) Suppose a model has a 200M parameter dense layer stored in FP16 format. If we instead quantize it to 4-bit integers (ignoring extra storage for scaling constants), what is the approximate memory savings?

A. 2× smaller

B. 4× smaller

C. 8× smaller

D. No savings, same size

2. (MCQ) A Transformer layer contains two linear projections: $W_Q \in R^{768\times768}$ and $W_V \in R^{768\times768}$. If we fine-tune both using LoRA with rank $r = 16$, how many parameters are trainable in LoRA compared to fine-tuning both layers fully?

A. 24,576 vs 1.18M

B. 49,152 vs 1.18M

C. 196,608 vs 1.18M

D. 393,216 vs 2.36M

3. (MCQ) You build a RAG pipeline on a corpus of 1,000,000 documents. Each document averages 500 tokens. You chunk documents into windows of 250 tokens (no overlap), embed each chunk into a 512-dimensional FP32 vector, and store them in a vector DB. Assume each FP32 number = 4 bytes and metadata per vector = 128 bytes. Approximately how many gigabytes of storage are required?

A. 4.0 GB

B. 4.3 GB

C. 4.6 GB

D. 5.0 GB

4. (MCQ) Which of the following best describes how LoRA works?

A. LoRA continues the original pre-training objective on new data to update the weights of the original model.

B. LoRA freezes all weights in the original model layers and introduces new components which are trained on new data.

C. LoRA trains a smaller, distilled version of the pre-trained LLM to reduce model size.

D. LoRA decomposes the original weight matrix into two smaller rank matrices and trains those instead of the full model weights.

5. (MCQ) What is the main limitation of using a fixed-size context vector in encoder–decoder sequence models?

A. It improves translation accuracy.

B. It increases training stability.

C. It can create a bottleneck for long input sequences.

D. It allows direct alignment between source and target tokens.

6. (MCQ) In the Transformer architecture, positional encoding is used to:

A. Normalise the hidden representations.

B. Provide information about the order of tokens in a sequence.

C. Reduce overfitting during training.

D. Increase the model's parameter count.

7. (MCQ) Which of the following best describes "cross-attention" in a Transformer decoder?

A. The decoder attending to its own outputs.

B. The encoder attending to the decoder outputs.

C. The decoder attending to the encoder outputs.

D. Tokens attending to themselves within the encoder.

8. (MCQ) Consider a sequence of five tokens

$$X = [\text{token}_1,\ \text{token}_2,\ \text{token}_3,\ \text{token}_4,\ \text{token}_5].$$

The (unmasked) row-wise attention score matrix (dot-product scores) for this sequence is

$$\text{Scores} = \begin{pmatrix} \text{token}_1\cdot\text{token}_1 & \text{token}_1\cdot\text{token}_2 & \text{token}_1\cdot\text{token}_3 & \text{token}_1\cdot\text{token}_4 & \text{token}_1\cdot\text{token}_5 \\ \text{token}_2\cdot\text{token}_1 & \text{token}_2\cdot\text{token}_2 & \text{token}_2\cdot\text{token}_3 & \text{token}_2\cdot\text{token}_4 & \text{token}_2\cdot\text{token}_5 \\ \text{token}_3\cdot\text{token}_1 & \text{token}_3\cdot\text{token}_2 & \text{token}_3\cdot\text{token}_3 & \text{token}_3\cdot\text{token}_4 & \text{token}_3\cdot\text{token}_5 \\ \text{token}_4\cdot\text{token}_1 & \text{token}_4\cdot\text{token}_2 & \text{token}_4\cdot\text{token}_3 & \text{token}_4\cdot\text{token}_4 & \text{token}_4\cdot\text{token}_5 \\ \text{token}_5\cdot\text{token}_1 & \text{token}_5\cdot\text{token}_2 & \text{token}_5\cdot\text{token}_3 & \text{token}_5\cdot\text{token}_4 & \text{token}_5\cdot\text{token}_5 \end{pmatrix}.$$

Question: After applying causal masking and then the row-wise softmax, how many entries of the resulting attention-weight matrix are *exactly zero*?

A. 0

B. 5

C. 11

D. 10

9. (MCQ) In language model pre-training, the "compute budget" refers to:

A. The maximum allowable depth and width of the model.

B. The maximum allowable training time and FLOPs.

C. The maximum allowable size of the training dataset.

D. The maximum allowable vocabulary size used in training.

10. (MSQ) What are the main advantages of parameter-efficient fine-tuning compared to full fine-tuning?

A. Reduced storage and memory footprint

B. Faster adaptation to downstream tasks

C. Ability to completely retrain the base model from scratch

D. Reusability of task-specific modules without duplicating the base model

11. (MCQ) In QLoRA, quantization is applied to reduce memory usage while still enabling efficient fine-tuning of large models. Which of the following statements most accurately captures the key design principle of QLoRA?

A. QLoRA quantizes both the base model weights and the LoRA adapter parameters to 4-bit, ensuring that all trainable parameters remain low-precision.

B. QLoRA uses 4-bit quantization of the frozen base model weights, but keeps LoRA adapter parameters in higher precision (e.g., 16-bit), ensuring stable gradient updates while retaining low memory usage.

C. QLoRA trains the full base model in 4-bit precision directly, thereby eliminating the need for LoRA adapters and achieving better fine-tuning accuracy.

D. QLoRA avoids quantization during training but compresses the model into 4-bit weights only at inference time, reducing memory costs after fine-tuning is complete.

12. (MCQ) In language model decoding, the temperature parameter τ rescales logits before applying softmax:

$$p_i = \frac{\exp\left(\frac{z_i}{\tau}\right)}{\sum_j \exp\left(\frac{z_j}{\tau}\right)}.$$

Suppose a model produces logits [2, 1, 0] for tokens $[A, B, C]$. Which of the following best describes the effect of changing the temperature?

A. At $\tau \to 0^+$, the distribution converges to uniform over all tokens, since scaling makes differences negligible.

B. At $\tau = 1$, the distribution is a normal softmax over the logits, and *increasing* τ above 1 makes the distribution more peaked.

C. At $\tau \to \infty$, the distribution approaches a delta on the largest logit (token A), because extreme scaling amplifies the maximum.

D. Lowering τ increases the relative probability gap between higher and lower logits, while raising τ flattens the distribution towards uniform.

13. (MSQ) Which of the following statements correctly describe the relationship between PEFT and catastrophic forgetting? (Select all that apply)

A. PEFT mitigates catastrophic forgetting by freezing most pretrained parameters and only training small task-specific modules.

B. Full fine-tuning is more prone to catastrophic forgetting because all parameters are updated for each new task.

C. PEFT entirely eliminates catastrophic forgetting, making continual learning trivial.

D. PEFT allows sharing the same base model across tasks while reducing interference between them.

14. (MCQ) How many LLMs are required for the (Direct Preference Optimization) DPO Method in Alignment Tuning ?

A. 1

B. 2

C. 3

D. 4

15. (MCQ) A language model at one decoding step produces the following probability distribution over eight candidate next tokens (probabilities in percent):

Token	A	B	C	D	E	F	G	H
Probability (%)	28	18	12	10	9	8	7	8

Using **top-p (nucleus) sampling** with $p = 0.60$, which set of tokens form the nucleus?

A. $\{A, B, C\}$

B. $\{A, B, C, D\}$

C. $\{A, B\}$

D. $\{A, B, C, D, E\}$

16. (MCQ) A system that can retrieve images based on text queries and retrieve text based on image queries is an example of:

A. Multi-modal classification

B. Single-modal generation

C. Cross-modal retrieval

D. Unimodal embedding

17. (MCQ) In Denoising Diffusion Probabilistic Models (DDPM), what happens during the forward process?

A. Clean images are gradually corrupted by adding noise

B. The model learns to generate images from scratch

C. Text is converted into visual representations

D. Audio signals are processed into spectrograms

18. (MCQ) What is the main training objective used in CLIP (Contrastive Language-Image Pre-training)?

A. Minimize reconstruction loss between images and text

B. Train separate classifiers for images and text independently

C. Maximize similarity between correct image-text pairs while minimizing similarity between incorrect pairs

D. Use autoencoder loss to compress both modalities

19. (MSQ) ReAct prompting integrates reasoning traces with external actions (e.g., API calls, search). Which of the following correctly describe its advantages and limitations? (Select all that apply)

A. It improves factuality by grounding reasoning in retrieved evidence

B. It completely eliminates hallucination by constraining the model to deterministic reasoning steps

C. It can be more resource-intensive due to multiple external calls during inference

D. It requires careful orchestration to avoid infinite loops or redundant reasoning steps

20. (MSQ) Multi-agent alignment approaches including self-play and debate models offer unique advantages but also face specific challenges. Which of the following are accurate characterizations? (Select all that apply)

A. Self-play creates adversarial training loops that can improve model robustness

B. Debate models help expose weaknesses in reasoning through structured argumentation

C. Multi-agent systems always converge to optimal solutions given sufficient time

D. The quality of debate-based alignment depends critically on judge reliability

E. These approaches can scale oversight by leveraging multiple perspectives on complex problems

21. (MSQ) A major challenge in scaling RLHF is "reward hacking". Which emerging techniques target this issue? (Select all that apply)

A. Direct Preference Optimization (DPO), by eliminating the explicit reward model altogether

B. Active Learning, by selectively querying the most uncertain samples to reduce labeling cost

C. Debate models, where multiple AI agents critique each other's reasoning

D. Supervised Fine-Tuning (SFT), by training on a larger dataset of demonstrations

22. (MCQ) Fill in the blanks: _____ involves using many prompt-completion examples as the labeled training dataset to continue training the model by updating its weights. This is different from _____ where you provide prompt-completion examples during inference.

A. Pre-training, Instruction fine-tuning

B. In-context learning, Instruction fine-tuning

C. Prompt engineering, Pre-training

D. Instruction fine-tuning, In-context learning

23. (MSQ) Consider encoder-only, decoder-only, and encoder-decoder architectures. Which of the following correctly describe their trade-offs? (Select all that apply)

A. Encoder-only models are best suited for bidirectional context understanding, making them effective for classification

B. Decoder-only models excel at generative tasks since they model left-to-right dependencies

C. Encoder-decoder models are computationally lighter because they separate encoding and decoding roles

D. Encoder-decoder models allow cross-attention, which is essential for tasks like translation

24. (MSQ) Multi-Head Attention and its masked variant are crucial in transformer architectures. Which of the following are true? (Select all that apply)

A. Multi-Head Attention enables the model to capture relationships at multiple representation subspaces simultaneously

B. Masked Multi-Head Attention prevents information leakage from future tokens during autoregressive generation

C. Self-Attention and Cross-Attention differ in that the query comes from the decoder in cross-attention

D. Multi-Head Attention reduces computational cost compared to single-head self-attention

25. (MSQ) Residual connections and normalization layers play key roles in transformer training stability. Which statement is correct?

A. Residual connections mitigate vanishing gradients by preserving gradient flow across layers

B. Layer Normalization is applied across the feature dimension, while Batch Normalization normalizes across the batch

C. Pre-Norm architectures apply normalization before residual addition, improving training stability in deep transformers

D. Batch Normalization is universally preferred over Layer Normalization in NLP models

26. (MSQ) Self-play reinforcement learning in LLMs draws inspiration from game AI. Which of the following capture unique challenges for LLM self-play? (Select all that apply)

A. Language models use discrete action spaces while games use continuous spaces

B. Language models cannot generate their own training data effectively

C. Self-play in LLMs requires human supervision while game AI is fully automated

D. Unlike games with clear win/loss conditions, LLMs face the challenge of defining meaningful adversarial tasks and evaluation metrics in the open-ended language domain

27. (MCQ) What is the main training objective of a decoder-only Transformer model, such as GPT?

A. To predict the next token in a sequence given the preceding tokens.

B. To reconstruct the original input sequence after adding noise.

C. To learn a shared embedding space for text and images.

D. To generate a complete sequence from a single starting token.

28. (MCQ) In the context of large language models, what does "temperature" control during text generation?

A. The number of tokens in the generated output.

B. The randomness of the generated text.

C. The length of the input sequence.

D. The size of the model's vocabulary.

29. (MCQ) What is the core innovation of the original Transformer architecture that addresses the limitations of Recurrent Neural Networks (RNNs) in sequence processing?

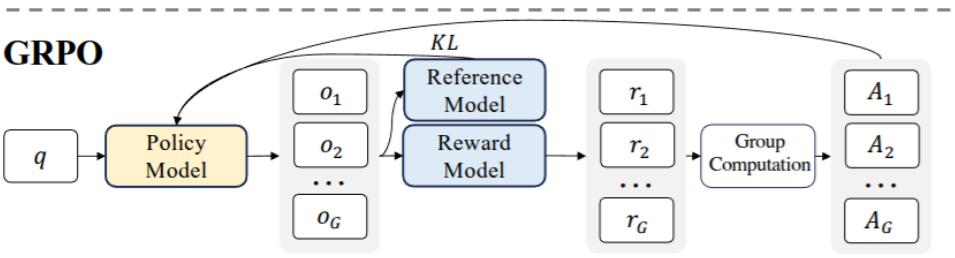
A. The use of a simple feedforward network, eliminating the need for complex architectures.

B. The introduction of attention mechanisms, which allows the model to weigh the importance of different parts of the input sequence regardless of their distance.

C. The ability to process only one token at a time, ensuring precise sequential processing.

D. The use of a single large convolutional layer to capture all dependencies at once.

30. (MSQ) In the GRPO framework shown below, which of the following best explains why GRPO eliminates the need for a separate value model compared to PPO?



A. GRPO directly computes group-level advantages by normalizing rewards across multiple sampled outputs, removing the need for state-value estimation.

B. GRPO uses the relative ranking of responses within a group as an implicit baseline, making a learned critic unnecessary.

C. GRPO assumes that KL divergence alone provides sufficient feedback for optimization, making the value model redundant.

D. GRPO trains the reward model and the policy jointly, so a separate value model is not necessary.

Rough Sheet