UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

CS 498: Introduction to Generative AI
Part II Exam

04/01/2026

Name: ____

NetID: ____

This exam contains 4 pages (including this cover page) and 6 questions, The point total is 18.

All questions in this exam are single- or multiple-choice questions. Please clearly circle your answer. If you change your response, erase your previous markings cleanly. Unless otherwise specified, each question has one correct answer. If we cannot find/cannot read your solution, we reserve the right to give that question/subquestion a 0.

Distribution of Marks

Question:	1	2
Points:	2	2
Score:		
Question:	3	4
Points:	2	4
Score:		
Question:	5	6
Points:	4	4
Score:		
Question:		Total
Points:		18
Score:		

Part 1: True/False

Please clearly circle your answer. Each question is worth 2 points. If you change your answer, please erase your previous marking cleanly.

1. (2 points) In cross-attention in an encoder–decoder Transformer, queries come from encoder states, while keys and values come from decoder states.

   A. True

   B. False

   Solution. B

2. (2 points) During training of a decoder-only Transformer, next-token predictions for all positions can be computed in parallel even though generation at test time proceeds one token at a time.

   A. True

   B. False

   Solution. A

3. (2 points) In multi-head cross-attention, the queries Q come from decoder states, while keys K and values V come from encoder states.

   A. True

   B. False

   Solution. A

Part 2: Multiple Choice

Please clearly circle your answer. If you change your answer, please erase your previous marking cleanly. Each choice is graded separately and is worth 1 point.

4. (4 points) Multi-head attention computes H attention outputs and then combines them. Which of the following statements is correct? **Select all that apply.**

   A. Different heads can use different learned projection matrices, which can lead to different attention patterns.

   B. Concatenating the head outputs typically increases the feature dimension before a final linear projection maps back to d_{model}.

   C. Using more heads makes causal masking unnecessary in a decoder.

   D. Multi-head attention can be viewed as computing several weighted sums in parallel and then mixing them.

   Solution. A,B,D

5. (4 points) (*PyTorch-level concept*) A common implementation pattern for attention is:

$$\texttt{scores} \leftarrow \frac{QK^\top}{\sqrt{d_k}}, \qquad \texttt{scores} \leftarrow \texttt{scores} + \texttt{mask}, \qquad \texttt{attn} \leftarrow \texttt{softmax}(\texttt{scores}).$$

   Which of the following statements is correct? **Select all that apply.**

   A. A causal mask is used in the encoder so that the encoder attends only to previous source tokens.

   B. Adding `mask` before softmax is a way to enforce that some (i, j) pairs receive near-zero attention weight.

   C. If `mask` uses large negative values for disallowed pairs, those pairs receive negligible probability after softmax.

   D. The main purpose of the mask is to change the value vectors V by setting some entries of V to zero.

   Solution. B, D

6. (4 points) Consider a decoder-only Transformer trained for next-token prediction on sequences $(y_1, \ldots, y_L)$. Which of the following statements is correct? **Select all that apply.**

   A. During training, we can compute a loss for each position t using the predicted distribution for y_{t+1} and the known target token.

   B. During inference, cross-attention to an encoder is required for any Transformer model.

   C. During inference, tokens are generated sequentially because the next token is unknown until it is produced.

   D. During training, causal masking is unnecessary because the targets are known.

   Solution. A,C

7. (4 points) Consider a decoder–encoder model with decoder states $S_{\text{dec}} \in \mathbb{R}^{L_{\text{dec}} \times d_{\text{model}}}$ and encoder states $H_{\text{enc}} \in \mathbb{R}^{L_{\text{enc}} \times d_{\text{model}}}$. Which statements correctly describe cross-attention? **Select all that apply.**

   A. Cross-attention typically needs a causal mask to prevent attending to future decoder positions.

   B. The attention weight matrix has size $L_{\text{dec}} \times L_{\text{enc}}$ for each head.

   C. For each head h, $Q^{(h)} = S_{\text{dec}} W^{Q,h}$ and $K^{(h)} = H_{\text{enc}} W^{K,h}$.

   D. In cross-attention, both $K^{(h)}$ and $V^{(h)}$ are computed from H_{enc}.

   Solution. B,C,D

8. (4 points) Which of the following statements about training versus inference in a decoder-only Transformer is correct? **Select all that apply.**

   A. In training with teacher forcing, all positions can be computed in parallel within a layer because the target tokens are known.

   B. In inference, generation is sequential because token y_{t+1} is unknown until the model produces it.

   C. Even in training, causal masking is used so that position t does not use information from target positions $> t$ when predicting y_{t+1}.

   D. In inference, causal masking is unnecessary because the model already knows future tokens.

   Solution. A,B,C

9. (4 points) Consider the following pseudocode for masked attention logits before softmax:

$$\text{scores} = QK^\top/\sqrt{d_k}, \qquad \text{scores}[\,\text{mask} = 0\,] = -\infty, \qquad A = \text{softmax}(\text{scores}).$$

   Which statement is correct?

   A. This masking is primarily to make attention computations faster by reducing the asymptotic complexity from $O(L^2)$ to $O(L)$.

   B. A padding mask is used to prevent attention to PAD tokens that were added only for batching.

   C. Setting masked logits to a very negative value makes the corresponding softmax weights close to zero.

   D. A causal mask is used to prevent a decoder position t from attending to positions $> t$ during next-token prediction.

   Solution. C

This page is intentionally left blank to accommodate work that wouldn’t fit elsewhere and/or scratch work.