CSE 576 Quiz 7

Transformer Pretraining Quiz – Answer Key

1. **(D)** By enabling different heads to attend to diverse relational patterns across positions.
   Multi-head attention lets each head focus on different types of relationships (syntax, long-range dependencies, local patterns, etc.), leading to richer contextual representations than a single head.

2. **(A)** Decoder layers with masked self-attention.
   GPT is a decoder-only Transformer that uses causal (masked) self-attention, making it naturally suited for left-to-right sequence generation.

3. **(B)** Unidirectional left-to-right flow; limits full context understanding.
   Because each token can only attend to previous tokens, GPT cannot use future tokens as context during pretraining, which restricts its ability to build fully bidirectional representations.

4. **(C)** Jointly optimizing masked token reconstruction and inter-sentence coherence discrimination.
   BERT combines the masked language modeling (MLM) objective with next sentence prediction (NSP), allowing it to capture both deep token-level context and relationships between sentences.

5. *Sample answer:*
   The pretrain–fine-tune pipeline can suffer from:

   - **Domain and task mismatch**: representations learned on large generic corpora may not align well with a specialized downstream domain.
   - **Catastrophic forgetting or instability**: fine-tuning on small task datasets can overwrite or destabilize useful pretrained knowledge.
   - **Bias propagation**: biases present in the pretraining data are carried into many tasks.
   - **Inefficiency and rigidity**: changing tasks or domains often requires retraining or carefully repeating the fine-tuning process.

6. *Sample answer:*
   GPT's autoregressive objective optimizes next-token prediction using only left context. This:

   - Encourages strong local fluency but does not directly optimize global sequence-level objectives like holistic classification, ranking, or discourse structure.
   - Prevents the model from using future context during training, which can be important for tasks that require understanding the entire span before making a decision.
   - Can lead to shallow, locally coherent generations that miss deeper global constraints or logical consistency.

7. *Sample answer:*
   The main challenge of MLM is the **pretraining–inference mismatch**: during pretraining the model sees artificial `[MASK]` tokens and learns to predict them, but at fine-tuning and inference time those tokens never appear. This distribution shift can:

- Make the learned representations somewhat biased toward masked, corrupted inputs rather than natural text.
- Limit performance on generation tasks and require additional adaptation.
- Reduce sample efficiency because only a subset of tokens are directly supervised at each step.

8. *Sample answer:*
Unidirectional (left-to-right) processing is ideal for **text generation**: the model can generate tokens sequentially while conditioning on all previous outputs. However:

- For tasks like text classification or sequence labeling, having only left context at each position is suboptimal compared to fully bidirectional models like BERT.
- GPT often needs architectural tricks (e.g., adding a special classification token at the end and conditioning on the whole prefix) to approximate holistic understanding, whereas bidirectional models can natively attend to both past and future tokens.