

Large Language Models: Oral Knowledge Questions

Explain each answer in your own words and justify the underlying mechanism.

1. **Self-attention and contextual representations.** In the sentence “The animal didn’t cross the street because it was too tired,” how can self-attention help the representation of “it” incorporate information from “animal”? Explain the roles of queries, keys, attention weights, and values.

2. **Multiple attention heads and the feed-forward network.** Why does a Transformer use multiple attention heads? Explain how their learned projections allow different attention patterns, how their outputs are combined, and how the subsequent feed-forward network performs a different operation.

3. **Position and token order.** Consider a Transformer encoder with unrestricted self-attention and no positional encodings. It receives “dog bites man” and “man bites dog.” Why would the representation computed for “dog” be the same in both cases? Explain how adding positional encodings to token embeddings allows the model to distinguish the two orders.

4. **Encoder-decoder versus decoder-only Transformers.** Compare the original encoder-decoder Transformer with a GPT-style decoder-only model. Explain the roles of encoder self-attention, masked decoder self-attention, and encoder-decoder cross-attention. Which of these attention mechanisms does a standard GPT-style model use?

5. **From a hidden representation to generated text.** How does the final decoder representation become a prediction over vocabulary tokens? Explain the linear layer, logits, softmax probabilities, and token selection. How is the selected token used to continue generation, and how does the next-token cross-entropy objective train this process?