Understanding Generative AI - Quiz Questions

The following 15 multiple-choice questions are designed to reinforce core learning objectives from my Book Chapter on **Transformer-based generative models for text generation**, focusing on GPT-style architectures, training methods, decoding strategies, and limitations. The questions cover recall, application, and deeper analysis to ensure a comprehensive understanding of generative language modeling.

SECTION 1 — Recall Questions (5 Questions)

Question 1 — Definition of a Transformer
Which component is responsible for enabling a Transformer to handle long-range dependencies in text?

A) Convolution layers
B) Recurrent loops
C) Self-attention mechanism
D) Pooling layers
E) Batch normalization

Correct Answer: C

Explanation:
Self-attention allows the model to compare each token with every other token, capturing long-range dependencies that RNNs struggle with. Options A and D are used in CNNs; B is used in RNNs; E is not responsible for dependency modeling.

Reference: Section 2.1 "Transformer Architecture Basics," p. 5.

Question 2 — Tokenization
GPT-style models primarily rely on which tokenization method?

A) Word-level tokenization
B) Byte-Pair Encoding (BPE)
C) Character-level encoding
D) One-hot tokenization
E) Lemmatization-based encoding

Correct Answer: B

Explanation:
GPT models use BPE to efficiently represent rare words while keeping vocabulary size manageable. Other methods either inflate vocabulary (A) or lose semantic structure (C).

Reference: Section 2.2 “Tokenization and Embeddings,” p. 7.

Question 3 — Training Objective

What is the core training objective of a GPT-style language model?

A) Predict masked tokens
B) Predict the next token in a sequence
C) Classify sentiment
D) Minimize reconstruction loss
E) Select a response from predefined choices

Correct Answer: B

Explanation:
GPT uses causal language modeling (next-token prediction). Option A describes BERT; D describes autoencoders; C and E are downstream tasks.

Reference: Section 3.1 “Causal Language Modeling,” p. 10.

Question 4 — Position Encoding

Why do Transformer models need positional encodings?

A) To reduce training time
B) To add sequential order information
C) To decrease overfitting
D) To improve loss scaling
E) To reduce model parameters

Correct Answer: B

Explanation:
Transformers process all tokens in parallel, so positional encodings provide ordering information. Other options don't reflect their function.

Reference: Section 2.3 "Positional Encodings," p. 6.

Question 5 — Decoding Strategy

Which decoding method selects the top-k most probable next tokens before sampling?

A) Greedy
B) Beam search
C) Top-k sampling
D) Top-p sampling
E) Random sampling

Correct Answer: C

Explanation:
Top-k sampling restricts sampling to the top-k candidates, controlling randomness. Top-p samples from a dynamic set; greedy picks only the max-probability token.

Reference: Section 4.1 "Decoding Methods," p. 12.

SECTION 2 — Application Questions (5 Questions)

Question 6 — Choosing a Decoding Strategy

You want the model to generate creative marketing slogans while avoiding incoherent text. Which decoding setting is most appropriate?

A) Greedy decoding
B) Temperature = 0.1
C) Temperature = 1.2 + Top-k sampling
D) Random sampling with no constraints
E) Beam search

Correct Answer: C

Explanation:
Temperature > 1.0 increases creativity, while top-k restricts chaos. Greedy or beam search produces repetitive text; random sampling is too uncontrolled.

Reference: Section 4.2 “Balancing Creativity and Coherence,” p. 13.

Question 7 — Handling Long Prompts

A user provides a very long prompt and the model truncates part of it. Which configuration should you adjust?

A) Learning rate
B) Max position embeddings
C) Batch size
D) Tokenizer vocabulary size
E) Dropout

Correct Answer: B

Explanation:
The max position embedding length determines how many tokens the model can attend to. Other parameters do not influence context window size.

Reference: Section 2.4 “Context Window Limitations,” p. 8.

Question 8 — Detecting Overfitting

During fine-tuning on a small dataset, the training loss decreases, but text generations become repetitive. What does this indicate?

A) Underfitting
B) Overfitting
C) Poor tokenization
D) Incorrect optimizer
E) Low temperature

Correct Answer: B

Explanation:
Repetition in generations often reflects overfitting to narrow patterns. Low temperature could worsen repetition but doesn't explain decreasing training loss alone.

Reference: Section 5.1 "Limitations and Failure Modes," p. 16.

Question 9 — Improving Model Diversity

A model keeps generating identical responses. Which modification helps increase output diversity?

A) Lowering top-p
B) Lowering temperature
C) Increasing number of training epochs
D) Increasing temperature
E) Switching to greedy decoding

Correct Answer: D

Explanation:
Higher temperature increases randomness, improving variety. Lowering top-p or using greedy decoding reduces diversity.

Reference: Section 4.1 "Temperature Scaling," p. 12.

Question 10 — Domain Adaptation

You fine-tune GPT-2 on legal documents, and it begins generating structured legal language. Which concept explains this shift?

A) Zero-shot learning
B) Catastrophic forgetting
C) Domain adaptation
D) Data augmentation
E) Reinforcement learning

Correct Answer: C

Explanation:
Domain adaptation occurs when the model adjusts to patterns in a new domain. Catastrophic forgetting would imply loss of earlier ability, not a stylistic shift.

Reference: Section 3.3 “Fine-Tuning and Domain Adaptation,” p. 11.

SECTION 3 — Analysis & Evaluation Questions (5 Questions)

Question 11 — Comparing GPT and RNN Performance

Why do GPT-style Transformers outperform LSTM-based models on long-form text generation?

A) LSTMs store more information in their hidden states
B) Transformers use recurrence and attention jointly
C) Transformers capture long-range dependencies efficiently through self-attention
D) LSTMs train faster than Transformers
E) Transformers require no training data

Correct Answer: C

Explanation:
Self-attention lets the model access any token at any position, enabling superior handling of long context. LSTMs struggle with vanishing gradients over long distances.

Reference: Section 2.1 “Transformer Advantages Over RNNs,” p. 5.

Question 12 — Ethical Risk Assessment

A model trained on biased social media text generates toxic outputs. What is the root cause?

A) Incorrect decoding strategy
B) Too many layers
C) Bias in training data
D) Low-quality tokenizer
E) Insufficient GPU memory

Correct Answer: C

Explanation:
Generative models reproduce biases present in their datasets. Decoding strategies or GPU memory do not introduce social biases.

Reference: Section 5.2 “Bias, Ethics, and Safety,” p. 17.

Question 13 — Perplexity Interpretation

Model A has lower perplexity than Model B on the same dataset. Which interpretation is correct?

A) Model A memorized the dataset
B) Model A assigns higher probability to true next tokens
C) Model B is overfitting
D) Model B must use a different tokenizer
E) Model A has fewer parameters

Correct Answer: B

Explanation:
Lower perplexity means the model more accurately predicts next tokens. It does not necessarily imply overfitting or model size differences.

Reference: Section 3.4 “Evaluation Metrics,” p. 12.

Question 14 — Impact of Training Data Structure

Why does GPT-2 fine-tuned on Wikitext-2 generate more formal text than GPT-2 fine-tuned on Reddit posts?

A) GPT-2 cannot generate informal language
B) Model size determines writing style
C) Transformers ignore training data style
D) Generative models internalize statistical patterns from their datasets
E) Reddit uses a different tokenizer

Correct Answer: D

Explanation:
Generative models learn distributional patterns, including tone and structure. Fine-tuning shifts the model toward the domain of the new dataset.

Reference: Section 3.3 “Dataset Influence on Output Style,” p. 11.

Question 15 — Evaluating Decoding Trade-offs

A company wants controllable, non-chaotic text generation while still keeping slight creativity. Which approach offers the best trade-off?

A) Pure greedy decoding
B) Beam search with no sampling
C) Top-k or top-p sampling with moderate temperature
D) Fully random sampling
E) Temperature set to 0

Correct Answer: C

Explanation:
Top-k or top-p restrict randomness to meaningful tokens, while moderate temperature adds controlled creativity. Greedy and beam search are too rigid; random sampling is too chaotic.

Reference: Section 4.2 “Controlling Generation Quality,” p. 14.