Questions and Answers - Chapter 10: Augmented Large Language Models (RAG)

True/False:

Q1. Prompt quality in RAG does not impact coherence.

Answer: False

Q2. RAG is slower than pure generation models.

Answer: True

Q3. RAG index remains static.

Answer: False

Q4. RAGAs framework only evaluates generation.

Answer: False

Q5. Recursive reasoning reuses problem-solving steps.

Answer: True

Multiple Choice:

Q6. Goal of Augmented LLMs:

(a) Smaller models (b) Enhance with external knowledge (c) Faster decoding (d) Train from scratch

Answer: (b) Enhance with external knowledge

Q7. Prompting in RAG helps:

(a) Improve logits (b) Create context from retrieval (c) Prune memory (d) Tokenize search

Answer: (b) Create context from retrieval

Q8. Purpose of sharding in RAG:

(a) Reduce attention heads (b) Manage large data efficiently (c) Compress model (d) Distill embeddings

Answer: (b) Manage large data efficiently

Q9. Metric not used for retrieval evaluation:

(a) BLEU (b) Precision@k (c) Recall@k (d) MRR

Answer: (a) BLEU

Q10. LLM agents differ by:

(a) More parameters (b) Access to APIs and data (c) Faster decoding (d) Limited vocabulary

Answer: (b) Access to APIs and data

Prepared for the User

Compiled by ChatGPT