## 5. Generation Layer — MCQs

### A. OpenAI GPT Models

1. Which of the following OpenAI models is widely used in RAG pipelines for high-quality generation?
   a) GPT-2
   b) GPT-3.5
   c) GPT-4
   d) Both b & c
   **Answer:** d) Both b & c ✅
2. GPT-4 is known for:
   a) Better reasoning capabilities
   b) Handling longer contexts
   c) Better integration with retrieval frameworks
   d) All of the above
   **Answer:** d) All of the above ✅
3. In RAG, GPT models primarily serve as:
   a) Index builders
   b) Generators
   c) Embedding creators
   d) Document parsers
   **Answer:** b) Generators ✅
4. Which GPT variant is cost-effective for RAG when low-latency is important?
   a) GPT-4
   b) GPT-3.5
   c) GPT-2
   d) Codex
   **Answer:** b) GPT-3.5 ✅

---

### B. Anthropic Claude

5. Claude 3 and Claude 2 are designed by:
   a) OpenAI
   b) Anthropic

c) Meta
d) Cohere
**Answer:** b) Anthropic ✅

6. A unique strength of Claude compared to GPT is:
   a) Training on larger multimodal datasets
   b) Safety-first and explainability-oriented responses
   c) Faster embedding generation
   d) Indexing speed
   **Answer:** b) Safety-first and explainability-oriented responses ✅
7. Claude models are widely used in RAG pipelines when:
   a) High interpretability is needed
   b) Lower cost than GPT-4 is preferred
   c) Large context window requirements exist
   d) All of the above
   **Answer:** d) All of the above ✅

---

## C. Mistral / Mixtral

8. **Mistral** models are popular in RAG setups because they are:
   a) Open-source
   b) Small, efficient, and high-performing
   c) Optimized for fine-tuning
   d) All of the above
   **Answer:** d) All of the above ✅
9. **Mixtral** is different from Mistral because:
   a) Mixtral uses a Mixture of Experts (MoE) architecture
   b) Mixtral is proprietary
   c) Mixtral is slower
   d) Mixtral doesn't support open-source fine-tuning
   **Answer:** a) Mixtral uses a Mixture of Experts (MoE) architecture ✅
10. A key advantage of Mistral in RAG workflows is:
    a) Higher accuracy than GPT-4
    b) Cost-efficiency and better on-device deployment
    c) Better multimodal understanding

d) Integration with proprietary APIs only
**Answer:** b) Cost-efficiency and better on-device deployment ✅

---

## D. LLaMA 2

11. **LLaMA 2** models are developed by:
    a) OpenAI
    b) Meta
    c) Google DeepMind
    d) Cohere
    **Answer:** b) Meta ✅
12. LLaMA 2 is considered ideal for RAG when:
    a) Open-source preference exists
    b) Deployment on private servers is required
    c) Full control over fine-tuning is needed
    d) All of the above
    **Answer:** d) All of the above ✅
13. A unique benefit of LLaMA 2 in RAG is:
    a) Closed-source high-cost API
    b) Supports multilingual queries effectively
    c) Limited to small documents
    d) Low compatibility with orchestration tools
    **Answer:** b) Supports multilingual queries effectively ✅

---

## E. Cohere Command R

14. **Cohere Command R** models are optimized for:
    a) Vector indexing
    b) RAG-specific generation
    c) Embedding creation only
    d) Dataset labeling
    **Answer:** b) RAG-specific generation ✅
15. A key reason why **Cohere Command R** performs well in RAG is:
    a) Trained on retrieval-enhanced corpora

b) Optimized for structured document generation
c) Proprietary embeddings
d) Task-specific search APIs
**Answer:** a) Trained on retrieval-enhanced corpora ✅

---

## F. LangChain & LlamaIndex

16. LangChain and LlamaIndex act as:
    a) Generators only
    b) Orchestration frameworks
    c) Embedding creators
    d) Evaluation metrics
    **Answer:** b) Orchestration frameworks ✅
17. The primary role of orchestration in RAG is:
    a) To combine retriever + generator seamlessly
    b) To generate embeddings
    c) To create index tables
    d) To replace embeddings
    **Answer:** a) To combine retriever + generator seamlessly ✅
18. In RAG, LlamaIndex is specifically strong in:
    a) Chunking and document structuring
    b) LLM fine-tuning
    c) Embedding compression
    d) Token optimization
    **Answer:** a) Chunking and document structuring ✅

---

## G. Mixed Concepts

19. Which combination is most optimal for an **open-source RAG pipeline**?
    a) LangChain + GPT-4
    b) LlamaIndex + Mistral
    c) Haystack + Claude 3
    d) Cohere Command R + LangChain
    **Answer:** b) LlamaIndex + Mistral ✅

20. If you want a **fully private** RAG system with **on-premise deployment**, the best approach is:
    a) GPT-4 + LangChain
    b) LLaMA 2 + LlamaIndex
    c) Claude 3 + Cohere Command R
    d) OpenAI Embeddings + Haystack
    **Answer:** b) LLaMA 2 + LlamaIndex ✅