**LARGE LANGUAGE MODELS: RENDEZVOUS 2023**

***August 2025, Presented by B Das***

**Overview**

**This document summarizes key large language models (LLMs) from 2023 to 2025, their features, use cases, and licensing. It includes enhancements for a GenAI Architect interview, emphasizing hands-on LLM experience, fine-tuning, deployment, and ethical considerations.**

**Evolution of LLMs (August 2024)**

**The LLM landscape has evolved rapidly, focusing on:**

* **Open Models: Increased accessibility through open-weight and open-source licenses.**
* **Extended Context: Larger token limits for long-document processing.**
* **Enterprise Readiness: Scalable, cost-effective solutions for business applications.**

**Key Focus Areas for ML Architects (2024)**

**For GenAI Architects, the following are critical:**

1. **Cost Efficiency: Optimize resource allocation using open-source models and efficient inference techniques (e.g., quantization, pruning).**
2. **Open-Source Fine-Tuning: Leverage tools like Hugging Face Transformers, LoRA, and QLoRA for cost-effective model adaptation.**
3. **Retrieval-Augmented Generation (RAG): Combine retrieval and generation for contextually accurate outputs, using frameworks like LangChain or LlamaIndex.**
4. **GPU-Friendly Deployment: Optimize models for GPU inference using ONNX, TensorRT, or vLLM for high throughput and low latency.**

**Hands-On Tip: Architects should be proficient in fine-tuning LLMs using frameworks like PyTorch and Hugging Face, deploying models on cloud platforms (e.g., AWS SageMaker, Azure ML), and implementing RAG pipelines with vector databases (e.g., Pinecone, Weaviate).**

**Leading Proprietary LLMs (2023-2025)**

**GPT-4o: The Multimodal Virtuoso**

* **Developer: OpenAI**
* **Launch: May 2024**
* **License: Proprietary**
* **Token Limit: 128K**
* **Key Features: Multimodal (text, vision, code), 50% cheaper API costs than GPT-4.**
* **Use Cases: Creative tasks, document summarization, code generation.**
* **Hands-On Insight: Fine-tune GPT-4o via OpenAI’s API for custom enterprise applications, leveraging its multimodal capabilities for image-to-text workflows.**

**Claude 3.7 Sonnet: The Developer’s Companion**

* **Developer: Anthropic**
* **Launch: October 2024**
* **License: Proprietary**
* **Token Limit: 200K**
* **Key Features: Excels in code generation, debugging, and complex reasoning; analyzes diagrams.**
* **Use Cases: Software development, technical analysis, ethical Q&A.**
* **Hands-On Insight: Use Claude 3.7 Sonnet for structured reasoning tasks, integrating with AWS Bedrock for scalable deployment.**

**Gemini 2.0 Pro: The Google Powerhouse**

* **Developer: Google DeepMind**
* **Launch: December 2024**
* **License: Proprietary**
* **Token Limit: 128K**
* **Key Features: Advanced reasoning, seamless integration with Google Cloud.**
* **Use Cases: Enterprise applications, research, multilingual tasks.**
* **Hands-On Insight: Deploy Gemini 2.0 Pro on Google Cloud Vertex AI, optimizing for low-latency inference in production.**

**Leading Open-Source LLMs (2023-2025)**

**LLaMA 4 (Scout & Maverick)**

* **Developer: Meta**
* **Launch: March 2025**
* **License: Open-weight**
* **Token Limit: 32K–128K**
* **Key Features: Mixture-of-Experts (MoE) architecture, multimodal, multilingual.**
* **Use Cases: Custom deployments, multilingual tasks, RAG pipelines.**
* **Hands-On Insight: Fine-tune LLaMA 4 using Hugging Face’s PEFT library, deploy on-premises with vLLM for cost efficiency.**

**Gemma 3**

* **Developer: Google DeepMind**
* **Launch: February 2024**
* **License: Open-source**
* **Token Limit: 8K–32K**
* **Key Features: 1B to 27B parameters, optimized for single GPU deployment.**
* **Use Cases: On-device applications, research, custom pipelines.**
* **Hands-On Insight: Use Gemma 3 for edge computing, leveraging JAX or PyTorch for efficient inference.**

**Mistral 8x22B**

* **Developer: Mistral AI**
* **Launch: January 2025**
* **License: Open-source (Apache 2.0)**
* **Token Limit: 32K**
* **Key Features: MoE with 8 experts, high performance, resource-efficient.**
* **Use Cases: High-throughput applications, cost-effective solutions.**
* **Hands-On Insight: Optimize Mistral 8x22B with quantization (e.g., 4-bit) for low-GPU deployments.**

**Qwen 2.5-Max**

* **Developer: Alibaba**
* **Launch: November 2024**
* **License: Open-source**
* **Token Limit: 64K**
* **Key Features: Up to 110B parameters, strong multilingual performance.**
* **Use Cases: Multilingual applications, open-source AI.**
* **Hands-On Insight: Fine-tune Qwen 2.5-Max for domain-specific tasks using Alibaba Cloud’s PAI.**

**Earlier LLMs (2023-2024)**

**Claude 2**

* **Developer: Anthropic**
* **Launch: July 2023**
* **License: Commercial**
* **Token Limit: 100K**
* **Key Features: Constitutional AI, emphasizing helpfulness and ethics.**
* **Use Cases: Ethical AI, RAG, long-form Q&A.**
* **Hands-On Insight: Integrate Claude 2 with AWS Bedrock for secure, enterprise-grade deployments.**

**GPT-3.5 Turbo**

* **Developer: OpenAI**
* **Launch: March 2023**
* **License: Commercial**
* **Token Limit: 16K**
* **Key Features: Cost-efficient, high-speed version of ChatGPT.**
* **Use Cases: Internal copilots, chatbots, enterprise integration.**
* **Hands-On Insight: Use GPT-3.5 Turbo for rapid prototyping via OpenAI’s API.**

**LLaMA 2**

* **Developer: Meta**
* **Launch: July 2023**
* **License: Open-weight**
* **Token Limit: 4K–32K**
* **Key Features: Variants (7B, 13B, 70B), ideal for custom fine-tuning.**
* **Use Cases: RAG, custom deployments, enterprise pilots.**
* **Hands-On Insight: Fine-tune LLaMA 2 with LoRA for domain-specific tasks, deploy using Hugging Face Inference Endpoints.**

**Falcon 40B**

* **Developer: Technology Innovation Institute (TII)**
* **Launch: June 2023**
* **License: Apache 2.0**
* **Token Limit: 2K–4K**
* **Key Features: Cost-effective, fine-tunable for public sector.**
* **Use Cases: Public sector LLMs, compliant internal AI.**
* **Hands-On Insight: Deploy Falcon 40B on Kubernetes for scalable enterprise solutions.**

**MPT-7B**

* **Developer: MosaicML (acquired by Databricks)**
* **Launch: June 2023**
* **License: Apache 2.0**
* **Token Limit: 65K (StoryWriter variant)**
* **Key Features: Extensible, cost-effective, long-context variant.**
* **Use Cases: Document summarization, training from scratch.**
* **Hands-On Insight: Use MPT-7B with Databricks for scalable training and inference.**

**Claude 2.1**

* **Developer: Anthropic**
* **Launch: November 2023**
* **License: Commercial**
* **Token Limit: 200K**
* **Key Features: Best-in-class for long-context Q&A in 2024.**
* **Use Cases: Legal documents, internal audits, ethical Q&A.**
* **Hands-On Insight: Implement Claude 2.1 for long-document processing with AWS Bedrock.**

**Mistral 7B & Mixtral**

* **Developer: Mistral AI**
* **Launch: September–December 2023**
* **License: Apache 2.0**
* **Token Limit: 8K**
* **Key Features: Lightweight, MoE architecture (Mixtral), fast inference.**
* **Use Cases: Efficient inference, low-GPU-cost applications.**
* **Hands-On Insight: Optimize Mistral 7B with 4-bit quantization for edge deployments.**

**Command R & R+**

* **Developer: Cohere**
* **Launch: April–June 2024**
* **License: Commercial/Open-weight (R+)**
* **Token Limit: 128K**
* **Key Features: RAG-first design, low-latency, multilingual.**
* **Use Cases: Enterprise Q&A, knowledge retrieval.**
* **Hands-On Insight: Integrate Command R+ with Cohere’s API for RAG pipelines.**

**Phi-2**

* **Developer: Microsoft Research**
* **Launch: December 2023**
* **License: MIT (non-commercial)**
* **Token Limit: 2K**
* **Key Features: Lightweight, efficient for edge computing.**
* **Use Cases: Tiny chatbots, edge LLMs.**
* **Hands-On Insight: Deploy Phi-2 on resource-constrained devices using ONNX.**

**OpenChat (OpenHermes)**

* **Developer: Community-driven**
* **Launch: Ongoing (2024)**
* **License: Permissive**
* **Key Features: Built on LLaMA/Mistral, excels in instruction-following.**
* **Use Cases: Prototyping, local experiments.**
* **Hands-On Insight: Use OpenChat for rapid prototyping with Direct Preference Optimization (DPO).**

**LLM Cheat Sheet (August 2024)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Company** | **License** | **Context Window** | **Open Weights?** | **Key Use Case** |
| **GPT-4 Turbo** | **OpenAI** | **Commercial** | **128K** | **No** | **Document summarization, coding** |
| **Claude 2.1** | **Anthropic** | **Commercial** | **200K** | **No** | **Long-document Q&A, ethical AI** |
| **LLaMA 2** | **Meta** | **Open-weight** | **4K–32K** | **Yes** | **Custom fine-tuning, RAG** |
| **Falcon 40B** | **Technology Innovation Institute** | **Apache 2.0** | **2K–4K** | **Yes** | **Public sector, cost-effective** |
| **MPT-7B** | **MosaicML (Databricks)** | **Apache 2.0** | **65K** | **Yes** | **Document summarization** |
| **Mistral 7B** | **Mistral AI** | **Apache 2.0** | **8K** | **Yes** | **Fast inference, low GPU cost** |
| **Command R+** | **Cohere** | **Open-weight** | **128K** | **Yes** | **RAG, enterprise Q&A** |
| **Phi-2** | **Microsoft Research** | **MIT** | **2K** | **Yes** | **Edge computing, tiny chatbots** |
| **Gemma** | **Google DeepMind** | **Commercial/Research** | **8K** | **Yes** | **Custom pipelines, on-device** |

**Performance Metrics (2025)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **AI Quality (Accuracy)** | **Cost (Input/Output per 1M Tokens)** | **Latency (Index)** | **Throughput (Tokens/Sec)** |
| **GPT-4o** | **0.75** | **$2.5/$10** | **1.39** | **378.89** |
| **Mistral-Large** | **0.74** | **$2/$6** | **0.92** | **182.31** |
| **LLaMA-3.2-11B** | **0.43** | **$0.37/$0.37** | **0.78** | **384.16** |
| **Phi-3-medium** | **0.42** | **$0.17/$0.68** | **0.88** | **193.3** |
| **GPT-3.5-Turbo** | **0.40** | **$0.5/$1.5** | **0.89** | **363.71** |

**Hands-On Insight: Evaluate models based on trade-offs between accuracy, cost, and latency. For high-throughput applications, prioritize models like LLaMA-3.2-11B; for cost-sensitive projects, use Phi-3-medium.**

**Future Trends in LLMs**

1. **Increased Accessibility: Open-source models and APIs democratize LLM usage.**
2. **Multimodal Processing: Integration of text, images, and audio for richer applications.**
3. **Ethical Considerations: Frameworks like Anthropic’s Constitutional AI ensure responsible use.**
4. **Efficient Resource Use: Techniques like MoE and quantization reduce computational costs.**
5. **Real-World Applications: LLMs in healthcare (e.g., medical record summarization), education (e.g., personalized tutoring), and customer service (e.g., chatbots).**
6. **Ecosystem Integration: Seamless integration with platforms like AWS, Azure, and Google Cloud.**

**Hands-On Insight: Stay updated with emerging frameworks (e.g., LangChain, LlamaIndex) and experiment with multimodal LLMs for cross-domain applications.**

**GenAI Architect Interview Preparation**

**For hands-on LLM experience, focus on:**

* **Fine-Tuning: Use LoRA/QLoRA with Hugging Face for efficient model adaptation.**
* **Deployment: Deploy modelslandır**

**System: using vLLM, TensorRT, or cloud platforms (e.g., AWS, Azure).**

* **RAG Pipelines: Build RAG systems with vector databases and frameworks like LangChain.**
* **Optimization: Apply quantization, pruning, and MoE architectures for performance.**
* **Ethics: Implement guardrails (e.g., content filtering) and evaluate models for bias using tools like Fairlearn.**
* **Monitoring: Use tools like Prometheus and Grafana to monitor model performance in production.**

**Sample Interview Question: "How would you fine-tune LLaMA 4 for a domain-specific RAG pipeline?"**

* **Answer: Use Hugging Face’s Transformers with LoRA to fine-tune LLaMA 4 on domain-specific data. Integrate with Pinecone for vector storage and LangChain for RAG. Optimize inference with vLLM and monitor performance with Prometheus.**

**Embedding Models: 2023–2025**

*An Overview for GenAI Architects*

**Introduction**

Embedding models convert text, images, or other data into dense vector representations, enabling tasks like semantic search, clustering, and RAG. This document covers key embedding models from 2023 to 2025, categorized by open-source and commercial offerings, with relevance for GenAI applications.

**Key Trends in Embedding Models**

* **Multimodal Embeddings**: Models supporting text, images, and audio for versatile applications.
* **Efficiency**: Lightweight models optimized for low-latency inference and edge deployment.
* **Open-Source Growth**: Increased availability of high-quality open-source embeddings.
* **Scalability**: Models designed for large-scale enterprise search and retrieval.

**Commercial Embedding Models**

**OpenAI text-embedding-3-large**

* **Developer**: OpenAI
* **Launch**: January 2024
* **License**: Commercial
* **Dimension**: 3072
* **Key Features**: High-dimensional embeddings, optimized for semantic search and clustering.
* **Use Cases**: Enterprise search, RAG, text classification.
* **Relevance**: High accuracy for English-centric tasks; accessible via OpenAI’s API.
* **Hands-On Insight**: Integrate with OpenAI’s API for RAG pipelines, using vector databases like Pinecone.

**Cohere Embed v3**

* **Developer**: Cohere
* **Launch**: November 2023
* **License**: Commercial
* **Dimension**: 1024
* **Key Features**: Multilingual support, optimized for low-latency retrieval.
* **Use Cases**: Knowledge retrieval, multilingual search, classification.
* **Relevance**: Ideal for enterprise applications requiring fast, multilingual embeddings.
* **Hands-On Insight**: Use Cohere’s API with Weaviate for scalable RAG systems.

**Google Vertex AI Embeddings**

* **Developer**: Google
* **Launch**: March 2023
* **License**: Commercial
* **Dimension**: 768
* **Key Features**: Integrated with Google Cloud, supports text and multimodal embeddings.
* **Use Cases**: Enterprise search, recommendation systems, multimodal RAG.
* **Relevance**: Seamless integration with Google’s ecosystem for large-scale deployments.
* **Hands-On Insight**: Deploy on Vertex AI, combining with BigQuery for data-driven applications.

**Open-Source Embedding Models**

**Sentence-BERT (SBERT)**

* **Developer**: UKP Lab (Hugging Face Community)
* **Launch**: Ongoing updates (2023–2025)
* **License**: Apache 2.0
* **Dimension**: 384–768
* **Key Features**: Lightweight, fine-tunable, supports multiple languages.
* **Use Cases**: Semantic search, text similarity, clustering.
* **Relevance**: Highly customizable for domain-specific tasks; widely adopted in research.
* **Hands-On Insight**: Fine-tune SBERT with Hugging Face’s sentence-transformers library for niche applications.

**BGE (BGE-large-en-v1.5)**

* **Developer**: BAAI (Beijing Academy of AI)
* **Launch**: September 2023
* **License**: Apache 2.0
* **Dimension**: 1024
* **Key Features**: State-of-the-art for English text embeddings, optimized for retrieval.
* **Use Cases**: RAG, semantic search, question answering.
* **Relevance**: Competitive with commercial models, ideal for open-source RAG pipelines.
* **Hands-On Insight**: Deploy BGE with Hugging Face’s Transformers, integrating with Faiss for vector search.

**E5 (intfloat/e5-large-v2)**

* **Developer**: Microsoft (Community-driven)
* **Launch**: December 2023
* **License**: MIT
* **Dimension**: 1024
* **Key Features**: High performance for text and code embeddings, multilingual support.
* **Use Cases**: Code search, document retrieval, multilingual RAG.
* **Relevance**: Versatile for cross-domain applications, especially in technical domains.
* **Hands-On Insight**: Use E5 with ONNX for efficient inference in production.

**Mixtral Embed**

* **Developer**: Mistral AI
* **Launch**: February 2024
* **License**: Apache 2.0
* **Dimension**: 768
* **Key Features**: Lightweight, MoE-based embeddings, optimized for low-GPU usage.
* **Use Cases**: Edge applications, cost-effective retrieval.
* **Relevance**: Ideal for resource-constrained environments.
* **Hands-On Insight**: Deploy Mixtral Embed with vLLM for low-latency inference.

**CLIP-ViT-L-336px**

* **Developer**: OpenAI
* **Launch**: June 2023
* **License**: MIT
* **Dimension**: 512
* **Key Features**: Multimodal (text and image) embeddings, high accuracy for vision tasks.
* **Use Cases**: Image-text retrieval, multimodal search, content moderation.
* **Relevance**: Pioneering multimodal embeddings for cross-modal applications.
* **Hands-On Insight**: Use CLIP with PyTorch for multimodal RAG pipelines.

**Embedding Model Cheat Sheet (2025)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Company** | **License** | **Dimension** | **Open-Source?** | **Key Use Case** |
| text-embedding-3-large | OpenAI | Commercial | 3072 | No | Enterprise search, RAG |
| Cohere Embed v3 | Cohere | Commercial | 1024 | No | Multilingual retrieval, Q&A |
| Google Vertex AI | Google | Commercial | 768 | No | Multimodal RAG, enterprise search |
| SBERT | UKP Lab (Hugging Face) | Apache 2.0 | 384–768 | Yes | Semantic search, clustering |
| BGE-large-en-v1.5 | BAAI | Apache 2.0 | 1024 | Yes | RAG, question answering |
| E5-large-v2 | Microsoft | MIT | 1024 | Yes | Code search, multilingual RAG |
| Mixtral Embed | Mistral AI | Apache 2.0 | 768 | Yes | Edge applications, retrieval |
| CLIP-ViT-L-336px | OpenAI | MIT | 512 | Yes | Multimodal search, image retrieval |

**Relevance for GenAI Architects**

* **Commercial Models**: Offer high accuracy and enterprise support but require API costs and vendor lock-in. Ideal for rapid deployment in production.
* **Open-Source Models**: Provide flexibility for fine-tuning and cost-effective deployment, especially for research or custom applications.
* **Hands-On Skills**:
  + Fine-tune SBERT or BGE for domain-specific embeddings using Hugging Face.
  + Build RAG pipelines with open-source embeddings (e.g., BGE) and vector databases (e.g., Faiss, Pinecone).
  + Optimize multimodal embeddings (e.g., CLIP) for cross-modal search using PyTorch.
  + Monitor embedding performance with metrics like cosine similarity and recall@K.

**Future Trends in Embedding Models**

1. **Multimodal Integration**: Unified embeddings for text, images, and audio.
2. **Efficiency**: Lightweight models for edge and on-device applications.
3. **Open-Source Adoption**: Growing community contributions to models like BGE and E5.
4. **Scalability**: Optimized embeddings for large-scale retrieval in enterprise settings.
5. **Ethical Considerations**: Addressing biases in embeddings for fair representations.

**Hands-On Insight**: Experiment with multimodal embeddings (e.g., CLIP) for innovative applications like visual question answering, and use tools like Faiss for scalable vector search.