shinydocs

Data Intelligence Solution

Best Open-Source LLM's by Use Case

Unlock the Potential of Open-Source LLMs for Diverse Applications

In the rapidly evolving landscape of artificial intelligence, open-source large language models (LLMs) have emerged as powerful tools for a wide range of applications. From general-purpose tasks to specialized use cases, these models offer flexibility, efficiency, and cost-effectiveness. This guide highlights the best open-source LLMs tailored to various needs, helping you choose the right model for your specific requirements. Whether you're looking for robust all-purpose models, enterprise-grade solutions, specialized applications, or cost-efficient options, we've got you covered.

Best All-Purpose LLMs

Mistral & Mixtral (Mistral AI)

- **Sizes:** Mistral 7B, Mixtral (MoE, 8x7b, 8x22b)
- **Strengths:** Efficient generalist models with robust reasoning and performance. Mixtral uses Mixture-of-Experts for high-quality, cost-effective outputs.
- **Enterprise Use:** Cost-effective structured enrichment, quick Q&A workflows.

Llama 3.1 (Meta)

- **Sizes:** 8B, 70B, 405b
- **Strengths:** Excellent instruction-following, versatile for Q&A, document summarization, structured enrichment, and general enterprise tasks. Extended context with special variants available.
- **Enterprise Use:** High-performance RAG, multi-purpose document analysis.

Gemma 3 (Google DeepMind)

- **Sizes:** 2B, 7B
- **Strengths:** Versatile for general instruction-following, secure and robust, excels in responsible, reliable general-purpose workflows.
- **Enterprise Use:** Broad applicability, from structured enrichment and doc Q&A to security-conscious general use.

Best for Enterprise AI & Secure On-Prem Use

Gemma 3 (Google DeepMind) (also listed as All-Purpose)

- **Sizes:** 2B, 7B
- **Strengths:** Specifically built for secure, responsible deployment. Compact, secure architecture with excellent performance for sensitive environments.
- **Enterprise Use:** Secure, compliant document analytics, sensitive data enrichment, on-premises Q&A solutions.

IBM Granite 3.x (IBM Research)

- **Sizes:** Dense (2B, 8B), Vision variant
- **Strengths:** Exceptional long-context capabilities (128K+ tokens), multimodal support, embedding generation, tool-use proficiency.
- **Enterprise Use:** Comprehensive RAG, visual/textual document enrichment, advanced metadata extraction.

Best for Structured Enrichment & Coding

CodeQwen (Qwen 2.5-Coder, Alibaba)

- **Sizes:** 0.5B–32B
- **Strengths:** Specialized in structured data extraction, logic-driven tasks, coding efficiency, structured output formatting (JSON, CSV, etc.).
- **Enterprise Use:** Structured enrichment workflows, backend automation, automated logic, and code-driven enrichment tasks.

CodeLlama (Meta)

- **Sizes:** 7B, 13B, 34B
- **Strengths:** Optimized for structured document processing and code reasoning. Ideal for structured extraction and enrichment pipelines.

Best for Cost Efficiency & Low Compute

Phi-3 (Microsoft)

- **Sizes:** 3.8B
- **Strengths:** Lightweight, efficient, yet capable of structured enrichment and rapid reasoning tasks, optimized for minimal resource environments.
- **Enterprise Use:** Efficient doc Q&A, structured enrichment in resource-limited setups.

SmolLM2

- **Sizes:** 135M, 360M, 1.7B
- **Strengths:** Compact models optimized for on-device or low-resource environments. Despite size, the 1.7B variant delivers strong instruction-following, summarization, function-calling, mathematical reasoning, and knowledge
- **Enterprise Use:** Perfect for lightweight backend services, edge deployment, or cost-sensitive pipelines requiring basic document Q&A, structured extraction, or summarization.

To explore how these open-source LLMs compare on cost and performance, visit our comprehensive AI cost comparison chart at **Shinydocs.com/AI**

linkedin.com/company/Shinydocs

info@Shinydocs.com