|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LLM** | | | | | |
| **Model Family** | **Developer** | **Key Variants** | **Parameter Sizes (where known)** | **Key Features & Uses** | **Access**  **/Availability** |
| GPT | OpenAI | GPT-5, GPT-4o, GPT-4o mini, GPT-3.5 | Up to trillions (rumored for GPT-5) | Multimodal (text, image, audio, video), advanced reasoning, coding, agentic workflows; powers ChatGPT. | OpenAI API, paid subscriptions. |
| Claude | Anthropic | Claude 4 (Opus 4, Sonnet 4, Haiku), Claude 3.5 | Not publicly disclosed | Constitutional AI for safety, extended thinking mode, multimodal (image/text), coding, enterprise tasks; 1M token context in beta. | Anthropic API, AWS Bedrock, Google Vertex AI. |
| Gemini | Google | Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.0 Flash | Up to billions (e.g., Pro is largest) | Multimodal (text, image, audio, video), tool use, reasoning, coding; up to 2M token context. | Google AI Studio, Vertex AI, free tiers available. |
| Gemma | Google | Gemma 3 (1B, 4B, 12B, 27B), Gemma 2 (9B, 27B) | 1B to 27B | Open-source, lightweight, runs locally; text generation, multilingual support. | Hugging Face, Google Vertex AI. |
| LLaMA | Meta | LLaMA 4 (Scout, Maverick), LLaMA 3.2 (1B-90B), LLaMA 3.1 (8B-405B), LLaMA 3.3 (70B) | 1B to 405B | Open-source, multimodal (image/text), reasoning, coding, multilingual; up to 10M token context. | Hugging Face, Meta downloads. |
| Grok | xAI | Grok 4 (Heavy), Grok 3, Grok 2, Grok 1.5 | Up to 314B (Grok-1) | Multimodal (image/text), real-time search, coding, reasoning, AI agents; 128K token context. | xAI API, Grok platform. |
| Mistral | Mistral AI | Mistral Medium 3, Mixtral 8x22B, Magistral Medium | 39B active (Mixtral), up to 141B total | Mixture-of-Experts (MoE), multilingual, coding/math, low-latency; open-source options. | Mistral API, Hugging Face. |
| Command | Cohere | Command A, Command R+, Command R7B | 7B to 104B | Enterprise-focused, multilingual, RAG with citations, tool use; 256K token context. | Cohere API, AWS Bedrock. |
| DeepSeek | DeepSeek | DeepSeek V3.1, DeepSeek-R1 | Up to 671B | MoE, reasoning, math, coding; open-source (MIT), 128K token context. | DeepSeek API, Hugging Face. |
| Qwen | Alibaba | Qwen3 (235B-A22B), Qwen2.5, Qwen1.5 | 0.5B to 235B | MoE, multilingual (12+ languages), coding/vision/audio; open-source. | Alibaba Cloud, Hugging Face. |
| Ernie | Baidu | Ernie 4.5, Ernie X1 | Billions (not specified) | MoE, powers Ernie chatbot, open-sourced series in 2025. | Baidu API. |
| Falcon | Technology Innovation Institute | Falcon 3, Falcon 2, Falcon Mamba 7B | 1B to 180B | Open-source, multilingual, multimodal (vision/text), efficient for long sequences. | Hugging Face. |
| Granite | IBM | Granite 3.2, Granite Vision, Granite Code | Up to 34B | Open-source (Apache 2.0), enterprise-ready, RAG, code tasks, risk detection. | watsonx.ai, Hugging Face. |
| BLOOM | BigScience/  Hugging Face | BLOOM | 176B | Multilingual (46 languages + 13 programming), open access for research. | Hugging Face. |
| GPT-NeoX | EleutherAI | GPT-NeoX | 20B | Language understanding, math, few-shot tasks; open-source. | Hugging Face. |
| Vicuna | LMSYS | Vicuna-13B | 13B | Chatbot, multi-turn conversations; open-source. | Hugging Face. |
| BERT | Google | BERT family | 342M | Transformer-based, sequence-to-sequence, query understanding; foundational for NLP. | Hugging Face. |

|  |  |
| --- | --- |
| **Model Family** | **When to Use** |
| GPT | Ideal for general-purpose AI applications requiring high accuracy in reasoning, coding, creative writing, or multimodal tasks (e.g., analyzing images/videos alongside text). Use GPT-5 or GPT-4o for complex agentic workflows, enterprise chatbots, or when safety alignments and tool integration are critical. Opt for GPT-4o mini or GPT-3.5 for cost-sensitive, high-volume tasks like quick translations or simple Q&A. Best in production environments needing reliability and scalability via OpenAI's ecosystem. |
| Claude | Choose for safety-focused enterprise applications, such as legal document review, ethical AI agents, or tasks needing extended context (e.g., 1M tokens for analyzing long reports). Claude 4 Opus is suited for advanced reasoning in research or coding with "thinking" modes; Sonnet/Haiku for faster, lighter deployments like customer support bots. Great when constitutional AI (bias mitigation) is a priority, or integration with AWS/Google clouds. |
| Gemini | Use for multimodal projects involving text, images, audio, or video, such as content moderation, educational tools, or app development with long contexts (2M tokens). Gemini 2.5 Pro excels in reasoning-heavy tasks like math puzzles or data analysis; Flash variants for low-latency mobile apps. Prefer when free tiers suffice or for Google ecosystem integration (e.g., Vertex AI for ML pipelines). |
| Gemma | Opt for open-source, lightweight deployments on edge devices or local hardware, like mobile apps for text generation or multilingual chat. Smaller variants (1B-4B) for resource-constrained environments; larger (27B) for fine-tuning on custom datasets in research. Useful when privacy is key (no API calls) or for prototyping in Hugging Face. |
| LLaMA | Select for open-source flexibility in research, fine-tuning, or custom AI models, especially multimodal ones (e.g., vision-text). LLaMA 4/3.3 for high-performance reasoning/coding; smaller 1B-8B for on-device inference. Ideal for long-context tasks (10M tokens) like summarizing books or building RAG systems. Use when cost-free scaling or community extensions (via Hugging Face) are needed. |
| Grok | Choose for real-time, dynamic applications like search-augmented chat, AI agents, or coding with humor/uncensored outputs. Grok 4 Heavy for complex multimodal reasoning; lighter versions for quick queries. Best in scenarios requiring integration with X (formerly Twitter) data or when xAI's API offers unique tools for analysis. |
| Mistral | Use for efficient, low-latency tasks in multilingual environments or MoE-based scaling (e.g., handling variable workloads). Mixtral for math/coding in enterprise; open-source variants for custom fine-tuning. Suited for cost-effective APIs or when blending with Hugging Face for hybrid models. |
| Command | Prefer for enterprise RAG setups with citations, multilingual support, or tool-use in business apps (e.g., CRM integration). Command R+ for high-context (256K) document processing; smaller for chatbots. Ideal via Cohere/AWS when compliance and scalability are priorities. |
| DeepSeek | Opt for open-source MoE models in math, coding, or reasoning-heavy research (e.g., algorithm development). DeepSeek V3.1 for large-scale inference; suitable for MIT-licensed projects needing 128K context without proprietary dependencies. |
| Qwen | Use in multilingual (12+ languages) or multimodal (vision/audio) applications, like global e-commerce bots or content creation. Qwen3 for MoE efficiency in cloud setups; open-source for fine-tuning on Alibaba infrastructure. |
| Ernie | Choose for Chinese-language dominance or MoE in Asian markets, such as search engines or chatbots. Ernie 4.5 for integrated Baidu ecosystems; open-sourced variants for research in 2025. |
| Falcon | Select for open-source multilingual/multimodal tasks with efficiency (e.g., long-sequence processing). Falcon 3 for vision-text in apps; Mamba for state-space efficiency in real-time systems via Hugging Face. |
| Granite | Use in enterprise settings for open-source RAG, code generation, or risk detection (e.g., compliance tools). Granite 3.2 for IBM watsonx integration; Vision/Code for specialized multimodal/coding workflows. |
| BLOOM | Opt for research in multilingual NLP (46+ languages/programming) when open access is needed for benchmarking or ethical AI studies via Hugging Face. |
| GPT-NeoX | Choose for open-source language understanding or few-shot tasks in non-commercial research, especially math, via Hugging Face. |
| Vicuna | Use for chatbot development focusing on multi-turn conversations; open-source for quick prototyping in conversational AI. |
| BERT | Fundamental for NLP tasks like sequence classification, query understanding, or embeddings in pipelines; use when transformer basics are sufficient via Hugging Face. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EMBEDDINGS** | | | | |
| **Model** | **Developer** | **Dimensions** | **Key Performance & Pricing (if applicable)** | **Use Cases & Strengths** |
| text-embedding-3-large/small | OpenAI | 3072/1536 (adjustable) | High on English tasks; $0.13-$0.02 per million tokens. | Reliable for enterprise search, semantic similarity; Matryoshka-trained. |
| voyage-3-large/lite/code-3 | Voyage AI | 2048/512 (adjustable) | Top relevance in retrieval; $0.18-$0.02 per million tokens, free trial. | Cutting-edge for long contexts (32K tokens), multilingual, domain-specific (e.g., code). |
| Embed v3 (multilingual-v3.0) | Cohere | 1024/384 | Strong multilingual (100+ languages); $0.12 per million tokens. | Multilingual RAG, faster light versions. |
| text-embedding-004 (Gemini) | Google | 768 (adjustable) | Modest but free with limits; multilingual. | Cost-effective search, low-latency. |
| Jina Embeddings v3 | Jina AI | 1024 (adjustable to 32) | Task-optimized, long text (8K tokens); free tier + paid. | Long documents, multilingual (89+ languages). |
| Stella 400M/1.5B v5 | Dun Zhang | 1024 | Top open model on MTEB; open-source (MIT). | Fine-tuning, multilingual retrieval. |
| ModernBERT Embed Base/Large | Nomic AI/Answer.AI/LightOn AI | 768/1024 | Speed/accuracy focus; open-source (Apache 2). | English-only, efficient inference. |
| nomic-embed-text / mxbai-embed-large / all-minilm | Ollama | 1024/384 | Local/privacy-focused; free (hardware cost). | Offline semantic search, privacy-sensitive apps. |
| Amazon Titan Embeddings | AWS | 1536 | Scalable, enterprise security; pay-per-use. | Cloud-based RAG, integration with AWS services. |
| Azure Embeddings | Microsoft | Varies | Integrated with Azure; pay-per-use. | Enterprise NLP, hybrid cloud setups. |
| Hugging Face Models (e.g., Sentence-BERT, intfloat/e5-large-v2) | Hugging Face Community | 768+ | Zero-shot, open-source; free. | Universal sentence embeddings, RAG customization. |
| Grok Embeddings | xAI | Varies | Real-time, multimodal; API pricing. | Research, data analysis with Grok integration. |

|  |  |
| --- | --- |
| **Model** | **When to Use** |
| text-embedding-3-large/small | For high-accuracy English semantic search or RAG in production (e.g., chatbots retrieving docs). Large for precision; small for speed/cost; adjustable dims for storage efficiency. |
| voyage-3-large/lite/code-3 | Ideal for domain-specific retrieval like code search or long-context (32K) multilingual RAG. Use code-3 for programming tasks; free trials for testing. |
| Embed v3 (multilingual-v3.0) | Choose for global apps needing 100+ languages in RAG or search; light versions for faster inference in enterprise via Cohere. |
| text-embedding-004 (Gemini) | Opt for cost-effective, multilingual search in low-latency apps; free limits make it great for prototypes or Google-integrated pipelines. |
| Jina Embeddings v3 | Use for long-document (8K) embeddings in multilingual research or apps; adjustable dims for fine-tuning. |
| Stella 400M/1.5B v5 | Select as top open-source for MTEB benchmarks in multilingual retrieval; MIT license for custom fine-tuning. |
| ModernBERT Embed Base/Large | For English-focused speed/accuracy in inference; open-source Apache for privacy-sensitive setups. |
| nomic-embed-text / mxbai-embed-large / all-minilm | Choose for local/offline semantic search (e.g., personal data apps); free hardware-based for privacy. |
| Amazon Titan Embeddings | Use in scalable AWS cloud for enterprise RAG with security features. |
| Azure Embeddings | Ideal for Microsoft-integrated hybrid setups in enterprise NLP. |
| Hugging Face Models (e.g., Sentence-BERT, intfloat/e5-large-v2) | For zero-shot, customizable sentence embeddings in RAG; free community models for experimentation. |
| Grok Embeddings | Opt for real-time multimodal embeddings in research or xAI-integrated analysis. |