AI Model Comparisons

Model	Creator	Type	Key Features	Typical Use Cases
GPT-4o	OpenAI	Multimodal LLM	Processes text, images, and audio with real-time voice; improved reasoning and multilingual ability	AI assistants, coding, content creation, cloud copilots
GPT-3.5 Turbo	OpenAI	Text-based LLM	High speed and cost efficiency; optimized for large-scale text generation and transformation	Chatbots, summarization, translation
o1	OpenAI	Reasoning model	Advanced logical reasoning with step-by-step problem solving	Scientific research, math, law, coding
Grok 3	xAI	Multimodal reasoning AI	DeepSearch web retrieval, chain-of-thought reasoning, synthetic data training	Information analysis, social media data mining
Gemini 2.5 Pro	Google DeepMind	Multimodal LLM	1M token context window; reasoning across text, image, and coding tasks	Research, complex dialogues, multimodal analysis
Gemini 2.5 Flash	Google DeepMind	Fast reasoning LLM	Compact version of Gemini; prioritizes speed and low latency	Real-time conversation, embedded assistants
Claude 3.5	Anthropic	Constitutional AI model	Long context retention, ethical alignment, fact-grounding	Legal and enterprise workflows, document analysis
DeepSeek R1	DeepSeek	Compact high-reasoning LLM	Strong in reasoning benchmarks; efficient architecture	Coding, problem-solving, business automation
Llama 3.1 Ultra 253B	Meta (NVIDIA variant)	Open-weight reasoning LLM	128K context, strong open-source reasoning performance	Developer platforms, model fine-tuning
Sora	OpenAI	Generative video model	Creates realistic video sequences from text prompts	Animation, video generation, visual storytelling
Midjourney v7	Midjourney	Image generation model	High fidelity, stylistically diverse AI image synthesis	Creative industries, visual design, marketing
Recraft	RecraftAI	Vector-based AI design tool	Generates and edits scalable vector art	Digital marketing, UI/UX design, branding

GPT Model Comparison

Model	Release Year	Key Features	Architecture & Context	Performance Highlights	Ideal Use Cases
GPT-3.5 (Turbo)	2023	Fast and cost-efficient; strong at text-only tasks and dialogue	175B parameters; up to 16K token context window	Reliable for general text generation and summarization	Chatbots, writing, translation, customer support
GPT-4	2024	Major leap in reasoning, code interpretation, and factual precision	Estimated 1T parameters with Mixture of Experts design; up to 128K tokens	Stronger in logic, analysis, and contextual accuracy	Research, content creation, professional writing
GPT-4 Turbo	2024	Cheaper and faster than GPT-4; similar accuracy	Optimized for API and ChatGPT Plus plan	Improved efficiency for long sessions	Marketing, coding, and live conversation tools
GPT-4o (Omni)	2024	Native multimodal AI - unites text, audio, and vision in one neural network	End-to-end multimodal context handling; 128K tokens	89% accuracy, 87% precision; real-time voice and video understanding	Voice assistants, multimodal chat, real-time analysis
GPT-5	2025	Unified architecture merging GPT and o-series; adaptive reasoning	Incorporates o3 reasoning core; integrates tiered intelligence system	Auto-adjusts between speed and deep reasoning; strongest general AI to date	Enterprise AI, high-reliability reasoning, developers
GPT-5 Pro	2025	High-performing professional variant	Enhanced reinforcement learning and extended memory	Best accuracy in reasoning and multimodal synthesis	Advanced research, enterprise copilots
GPT-5 Mini	2025	Lightweight version for fast use cases	Condensed model for free-tier or overflow workloads	Prioritizes response speed under load	Lightweight chat tools, mobile AI
o1	2024	Dedicated reasoning model; chain-of-thought (CoT) logic	Trained to think before answering reinforcement learning refinement	Solves 83% of AIME questions; excels in math and coding	STEM, scientific, and logical problem solving
o1 Mini	2024	Faster, cheaper subset of o1	Reduced world knowledge but strong technical reasoning	Efficient for code and STEM tasks	Education, coding, analytics