

Provider	Model Name	Model Size	Model Type	Notes
OpenAI	GPT-4o	Not specified	Chat/Multimodal	Processes and generates text, images, and audio
	GPT-4o-mini	Not specified	Chat/Multimodal	A smaller and cheaper version of GPT-4o, replacing GPT-3.5 Turbo. Cost-effective
	GPT-4	Not specified	Chat/Multimodal	Capable of processing text and images
	GPT-4o with Scheduled Task	Not specified	Agent	An extension of GPT-4o designed to autonomously plan and execute multi-step tasks, providing comprehensive reports with citations.
	Operator	Not specified	Agent	An AI agent capable of carrying out tasks by using virtual browsers, e.g. booking tables at restaurants. Currently available only in a preview for ChatGPT Pro subscribers: 200 dollars per month in the US.
	o1	Not specified	Reasoning	Designed to solve complex problems by spending more time "thinking" before responding; effective in science, coding, and reasoning tasks.
	o3-mini	Not specified	Reasoning	A lighter and faster version of the o3 model, offering improved reasoning capabilities with reduced computational requirements.
	o3-mini-high	Not specified	Reasoning	A premium variant of o3-mini, providing even higher-quality responses, especially in coding tasks.

Provider	Model Name	Model Size	Model Type	Notes
	o1-pro	Not specified	Reasoning	An upgraded version of the o1 model, utilizing more computational resources to provide better answers; available to ChatGPT Pro subscribers (200 dollars per month.)
Microsoft	Copilot	Not specified	Chat/Multimodal/Reasoning	An AI-powered assistant."Think Deeper" feature provides reasoning based on OpenAI's o1 reasoning model.
Google Gemini	Gemini 1.5 Flash	8B	Chat/Multimodal	Supports reasoning tasks with a context window of 1,048,576 tokens; ideal for high-efficiency applications.
	Gemini 2.0 Flash	Not specified	Chat/Multimodal	
	Gemini 1.5 Pro	> 200B	Chat/Multimodal	Multimodal processing. Optimized for a wide range of reasoning tasks; can process large amounts of data, including extensive audio and video inputs.
	Gemini 2.0 Pro	Not specified	Chat/Multimodal	Capable of handling complex instructions with a context window of 2,000,000 tokens; integrates tools like Google Search and code execution.
	Gemini 2.0 Flash Thinking Experimental	Not specified	Reasoning	Best for multi-step reasoning
	Gemini 2.0 Flash Thinking Experimental with apps	Not specified	Reasoning	reasoning across YouTube, Maps & Search
	Gemini Advanced	Not specified	Reasoning	Offers features like Deep Research, large context window (e.g., 1 million tokens for Gemini 2.0 Flash), with the ability to create custom versions, called "Gems." Requires "Google One AI Premium"

Provider	Model Name	Model Size	Model Type	Notes
				subscription (19.99 dollars per month)
<b>Anthropic</b>	Claude 3 Haiku	Not specified	Chat	Part of the Claude 3 model family, offering advanced conversational abilities with a focus on safety.
	Claude 3 Sonnet	Not specified	Chat	An enhanced version in the Claude 3 series.
	Claude 3 Opus	Not specified	Chat	The most advanced model in the Claude 3 family, strong with coding and creative writing
<b>Liquid AI</b>	LFM-7B	7B	Chat	Alternative architecture (Synthesis of Tailored Architecture) instead of "Transformer" architecture, Fast and efficient inference
<b>Perplexity</b>	pplx-7b-online	7B	Search	Perplexity allows user to choose models from other providers; Online LLMs fine-tuned on Mistral-7B, Website excerpts are provided as inputs
	pplx-70b-online	70B	Search	Perplexity allows user to choose models from other providers; Online LLMs fine-tuned on Mistral-7B, Website excerpts are provided as inputs
<b>DeepSeek</b>	DeepSeek-R1 (Open-source)	671B	Reasoning	A reasoning model that has matched capabilities of recent models from OpenAI, Anthropic, and Meta, developed at a significantly lower cost.
	DeepSeek-V3 (Open-source)	671B	Chat	A chat model that outperforms Meta's Llama 3.1 and Alibaba's Qwen 2.5, while matching OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet.

Provider	Model Name	Model Size	Model Type	Notes
<b>Allen Institute for AI (Ai2)</b>	Tulu 3 (True Open-source)	7B / 8B / 70B	Reasoning	Strong instruction-following model with reasoning capabilities.
<b>Meta</b>	Llama 3 (Open-source)	8B / 70B	Chat	Rivals or exceeds models like Gemini Pro 1.5 and Claude Sonnet
	Llama 2 (Open-source)	7B / 13B / 70B	Chat	Instruction-tuned and optimized for dialogue use cases
<b>Mistral AI</b>	Mistral 7B (Open-source)	7B	Chat	Outperforms larger models like Llama 2 13B, highly-efficient
	Mistral 8x7B (Open-source)	47B	Chat	Mixture-of Experts (MoE) model
	Mistral Large	Significantly larger than 47B	Chat (?)	Mistral's most powerful model
	Mistral Medium	Between 7B and >47B	Chat	Offers a balance of performance and cost-effectiveness
	Mistral Small	Smaller than Mistral Medium	Chat	Designed for efficiency and cost-effectiveness