AI Model Pricing and Capabilities (2025 Update)

Below is the updated pricing table for various advanced AI models, followed by a comprehensive summary of each model's capabilities, intended use cases, and key differences from similar models.

Updated Pricing Table

Model Name	Context Window	Price per Million Input Tokens	Price per Million Output Tokens
Claude 3.7 Sonnet	200,000 tokens	$3.00 ANTHROPIC.COM	$15.00 ANTHROPIC.COM
Claude 3.5 Sonnet	200,000 tokens	$3.00 ANTHROPIC.COM	$15.00 ANTHROPIC.COM
Claude 3.5 Haiku	200,000 tokens	$0.80 ANTHROPIC.COM	$4.00 ANTHROPIC.COM
Claude 3 Opus	200,000 tokens	$15.00 TECHCRUNCH.COM	$75.00 TECHCRUNCH.COM
DeepSeek Chat (V3)	64,000 tokens	$0.27 API-DOCS.DEEPSEEK.COM	$1.10 API-DOCS.DEEPSEEK.COM
DeepSeek Reasoner (R1)	64,000 tokens	$0.55 API-DOCS.DEEPSEEK.COM	$2.19 API-DOCS.DEEPSEEK.COM
Gemini 2.5 Pro (Experimental)	1,000,000 tokens	*TBA (not publicly announced)* BLOG.GOOGLE	*TBA (not publicly announced)* BLOG.GOOGLE
Gemini 2.0 Flash	1,000,000 tokens	$0.10 AI.GOOGLE.DEV	$0.40 AI.GOOGLE.DEV
Gemini 2.0 Flash-Lite	1,000,000 tokens	$0.075 AI.GOOGLE.DEV	$0.30 AI.GOOGLE.DEV
Gemini 1.5 Flash	1,000,000 tokens	$0.075 AI.GOOGLE.DEVAI.GOOGLE.DEV	$0.30 AI.GOOGLE.DEV
Gemini 1.5 Pro	2,000,000 tokens	$2.50 AI.GOOGLE.DEV	$10.00 AI.GOOGLE.DEV
Grok 2 (xAI)	131,072 tokens	$2.00 DOCS.X.AI	$10.00 DOCS.X.AI
GPT-4.5 Preview	128,000 tokens	$75.00 OPENAI.COM	$150.00 OPENAI.COM
GPT-4o	128,000 tokens	$2.50 OPENAI.COM	$10.00 OPENAI.COM
GPT-4o Mini	128,000 tokens	$0.15 OPENAI.COM	$0.60 OPENAI.COM
OpenAI o1	200,000 tokens	$15.00 OPENAI.COM	$60.00 OPENAI.COM
OpenAI o1 Pro	200,000 tokens	$150.00 HELICONE.AI	$600.00 HELICONE.AI
OpenAI o3 Mini	200,000 tokens	$1.10 OPENAI.COM	$4.40 OPENAI.COM
OpenAI o1 Mini	131,000 tokens	$3.00 NEWS.YCOMBINATOR.COM	$12.00 NEWS.YCOMBINATOR.COM

(Note: "Context Window" indicates the maximum tokens of context the model can handle. "TBA" denotes pricing to be announced.)

Model Summaries

Anthropic Claude Models

Claude 3.7 Sonnet – **Capabilities:** Claude 3.7 Sonnet is Anthropic's most advanced model, noted as the *"first hybrid reasoning model"* that combines chain-of-thought reasoning with powerful generation

ANTHROPIC.COM. It has state-of-the-art coding abilities, strong content creation skills, and can integrate tools or computer use within its responses ANTHROPIC.COM. With a **200K token context window**, it can ingest very large documents or datasets in a single prompt. Claude 3.7 excels at complex analysis and planning, often breaking down problems step-by-step and even self-correcting mistakes during its reasoning ANTHROPIC.COM. **Intended Use Cases:** Because of its high intelligence and massive context, Sonnet 3.7 is well-suited for tasks like large-scale data analysis, intricate coding assistance, long-form content generation, and strategic planning simulations. It's also effective in multi-turn conversations that require remembering extensive context or instructions provided earlier. **Key Differences:** Compared to earlier Claude models, 3.7 Sonnet delivers significant improvements in both speed and capability; it's roughly twice as fast as Claude 2 while being more intelligent ANTHROPIC.COM. Unlike the faster-but-smaller Claude Haiku model, Claude Sonnet prioritizes higher accuracy and reasoning depth over speed, making it more comparable to OpenAI's top models in quality (though at a higher cost). Its hybrid reasoning approach enables it to outperform peers on complex benchmarks, setting a new bar for Anthropic ANTHROPIC.COM.

Claude 3.5 Sonnet – **Capabilities:** Claude 3.5 Sonnet was the flagship model prior to 3.7, offering advanced reasoning and coding skills similar in scope to 3.7, but with slightly less refinement. It introduced many of the agentic capabilities now seen in 3.7 – for example, using tools or performing multi-step tasks autonomously

ANTHROPIC.COM. It also supports the full 200K context, enabling lengthy inputs/outputs. **Intended Use Cases:** Much like 3.7, the 3.5 Sonnet model is designed for complex applications: e.g. writing and debugging software, analyzing lengthy texts or logs, and acting as a general-purpose assistant in business or education settings. It's effective for developers needing high accuracy but who might not require the absolute latest model. **Key Differences:** Compared to Claude 3.7, the 3.5 Sonnet is slightly less capable on the toughest tasks (since 3.7 introduced further improvements in reasoning and coding) ANTHROPIC.COM. However, it shares the same pricing structure as 3.7 ANTHROPIC.COM, so users often choose 3.7 when available. Versus Claude 3.5 Haiku (its contemporary smaller model), 3.5 Sonnet is slower and more expensive but delivers much stronger performance on complex tasks and larger-scale problems.

Claude 3.5 Haiku – **Capabilities:** Claude 3.5 Haiku is Anthropic's **fastest** model in the Claude 3.5 lineup, optimized for speed and cost-efficiency while still providing advanced abilities. It offers strong coding help, improved tool use, and reliable reasoning – outperforming the previous generation's largest model (Claude 3 Opus) on many benchmarks despite its lower cost

ANTHROPIC.COM. Haiku 3.5 maintains a 200K token context, so it can handle very long inputs, but it is tuned to respond quickly. **Intended Use Cases:** This model is ideal for real-time or high-volume applications where quick turnaround is crucial – for example, powering interactive chatbots, streaming code autocompletion, or moderating content in real-time. It's well suited for user-facing products and "specialized sub-agent" tasks that may run concurrently in large numbersANTHROPIC.COMANTHROPIC.COM. **Key Differences:** Claude 3.5 Haiku trades a bit of raw power for speed. It is more budget-friendly than the Sonnet models (costing only $0.80/M input, $4/M output)ANTHROPIC.COM, and it achieves near state-of-the-art speed – e.g. reading a dense 10K-token document in under 3 secondsANTHROPIC.COM. However, it is slightly less capable in complex reasoning or creative tasks compared to Claude Sonnet. Anthropic positions Haiku as the entry point to the Claude 3.5 family – **faster and cheaper, but not quite as *"intelligent"*** as Sonnet or Opus modelsANTHROPIC.COM.

Claude 3 Opus – **Capabilities:** Claude 3 Opus is the **largest and most powerful model** of the Claude 3 family (preceding the 3.5 update). It was Anthropic's state-of-the-art model in early 2024, delivering near-human performance on many knowledge and reasoning benchmarks

ANTHROPIC.COM. Opus exhibits exceptional comprehension, fluency, and expertise across domains – from undergraduate-level knowledge tests to graduate-level reasoning and mathematicsANTHROPIC.COM. It also supports a 200K token context window for very large inputs. **Intended Use Cases:** Opus is designed for the most demanding tasks where quality is paramount: in-depth research analysis, complex problem solving (such as legal reasoning or scientific research assistance), and any scenario where the highest level of accuracy and understanding is required. Enterprises may use Claude Opus for tasks like analyzing extensive reports or databases, given its large context and top-notch reasoning ability. **Key Differences:** Compared to later models like Claude 3.5/3.7 Sonnet, Opus 3 is *slower and far more expensive* (at $15/M input and $75/M output)TECHCRUNCH.COM. It was noted that Opus runs at roughly the same speed as the older Claude 2 models (i.e. not as fast as the new Haiku/Sonnet models)ANTHROPIC.COM. In exchange, it offers superior intelligence – it was the model that led Anthropic's benchmarks before the Claude 3.5 series arrived. Today, Claude 3 Opus remains an option when maximum intelligence is needed and cost is less of a concern, but many users have migrated to Claude 3.7 Sonnet for better efficiency.

DeepSeek Models

DeepSeek Chat (V3) – **Capabilities:** DeepSeek Chat V3 is the third-generation model from DeepSeek, designed as a general-purpose conversational AI with an emphasis on **high throughput and cost-effectiveness**. Under the hood it uses a massive Mixture-of-Experts architecture (671 billion parameters, with 37B active per token) to achieve strong performance

API-DOCS.DEEPSEEK.COM, but it remains accessible via an open API. This model supports a 64K token context, enabling fairly lengthy conversations or documents to be processed. It's capable of fluent dialogue,

answering questions, and basic reasoning. Notably, V3 improved its processing speed ~3× over V2 (up to ~60 tokens/second)[API-DOCS.DEEPSEEK.COM]. **Intended Use Cases:** Given its low cost and decent intelligence, DeepSeek Chat V3 is well-suited for applications requiring **scalable chat or assistant functionality** – for example, customer support chatbots, bulk text processing, or interactive agents that need to handle lots of queries in parallel. Its design balances competence with speed, making it useful for real-time services. **Key Differences:** Unlike "inference-oriented" models that explicitly generate reasoning traces, DeepSeek Chat focuses on straight response generation. It does, however, integrate **context caching** to reduce costs on repeated inputs (only $0.07/M for cached tokens)[API-DOCS.DEEPSEEK.COM]. Compared to DeepSeek's Reasoner model, Chat V3 is simpler – it doesn't output an explicit chain-of-thought. It is also cheaper and faster than the Reasoner R1, but may not perform as well on extremely complex, multi-step problems that benefit from seeing intermediate reasoning. Another differentiator is DeepSeek's open-source ethos: the model weights and research for V3 are publicly available [API-DOCS.DEEPSEEK.COM], which is relatively uncommon among large models (OpenAI and Anthropic do not open-source theirs).

DeepSeek Reasoner (R1) – **Capabilities:** DeepSeek's Reasoner R1 is a specialized model geared towards **deliberate reasoning via chain-of-thought**. It has the same 64K context as the Chat model, but when asked complex questions, it can produce a step-by-step reasoning process (not directly shown to the end-user unless requested) before giving a final answer. In fact, the API will return both the reasoning tokens and the final answer tokens as output, and both are billed equally [API-DOCS.DEEPSEEK.COM]. This approach often improves accuracy on complicated tasks like math word problems or logical puzzles, because the model "thinks out loud." R1's architecture is slightly different to facilitate this, and it uses more computational resources per query (hence the higher cost). **Intended Use Cases:** DeepSeek R1 is intended for scenarios where **transparency and strong reasoning** are needed. Developers might use it for AI agents that need to justify their answers or for complex planning tasks where seeing the intermediate steps is useful. It's also valuable in evaluation settings – e.g., to produce rationales for answers that can be audited. **Key Differences:** Compared to DeepSeek Chat V3, the Reasoner is both slower and costlier (roughly double the price per token)[API-DOCS.DEEPSEEK.COM]. In return, it tends to be better at multi-step problems and provides insight into its decision-making by outputting a chain-of-thought. This is similar in spirit to OpenAI's "reasoning" models (like o1), but DeepSeek's implementation allows the reasoning content to be retrieved via the API. In summary, choose DeepSeek Chat V3 for speed and interactive chat, but choose DeepSeek R1 when you need the model to *"show its work"* and tackle harder problems with a systematic approach.

Google DeepMind Gemini Models

Gemini 2.5 Pro (Experimental) – **Capabilities:** Gemini 2.5 Pro Experimental is Google DeepMind's latest **"thinking" model**, representing their most advanced AI to date. It is a **multimodal** reasoning model that can process text and other inputs (e.g. images or possibly code/context) and is built to "think" step-by-step before responding

MEDIUM.COMBLOG.GOOGLE. With an enormous 1M token context window, Gemini 2.5 Pro can absorb truly massive contexts (roughly 5× the context of GPT-4) – for example, entire books or codebases – and still perform coherent reasoning over them. In internal evaluations, it has achieved top-tier results: it debuted at #1 on the LMArena benchmark (which measures human preference comparisons)BLOG.GOOGLEBLOG.GOOGLE, and has demonstrated superior performance in coding and math tasks compared to most rival models. For instance, Google reports it scored 68.6% on a code-editing benchmark (Aider Polyglot), beating OpenAI's o3-mini and Anthropic's Claude 3.7 on that testGIGAZINE.NET. **Intended Use Cases:** Currently in a limited preview, Gemini 2.5 Pro is aimed at complex tasks that benefit from robust reasoning and multimodal understanding. This includes generating or debugging code in large projects, creating **agentic applications** (AI agents that plan and act, since Gemini can internally simulate thoughts), and even building visual web applications from a descriptionMEDIUM.COM. Its multimodal ability suggests it can analyze images or possibly diagrams combined with text, which could be useful for things like data analysis with charts, or assisting in design and creative tasks. As it's labeled "Experimental," it's mainly available for research and testing purposes through Google's Gemini Advanced plan (a $20/mo subscription)MEDIUM.COM. **Key Differences:** Compared to earlier Gemini models, 2.5 Pro Experimental is a significant leap in reasoning capability – Google explicitly calls it *"our most intelligent model to date"*MEDIUM.COM. It is intended to incorporate reasoning as a standard, always-on feature (whereas previous models might only do straightforward completion). In contrast to OpenAI's top reasoning model (o1-Pro), Gemini 2.5 Pro is expected to be more *multimodal* and possibly more efficient (OpenAI's o1-Pro is extremely costly and currently text-only). However, pricing for Gemini 2.5 Pro is **not yet publicly announced**BLOG.GOOGLE – it's likely to be introduced once the model moves out of experimental phase. In summary, Gemini 2.5 Pro (Experimental) is at the cutting edge, designed for the most demanding tasks, but it's still in preview with limited access and pending cost details.

Gemini 2.0 Flash – **Capabilities:** Gemini 2.0 Flash is the **flagship general model** of the Gemini 2.0 series (launched late 2024). It is a multimodal model capable of handling text, images, and even video and audio inputs, unified under a single model interface

AI.GOOGLE.DEV. It also features a 1M token context window, which was revolutionary at its release – enabling it to process vast amounts of context in one go. Gemini 2.0 Flash was built *"for the era of Agents,"* meaning it's optimized for tasks involving tool use and autonomous task execution in addition to standard QA or text generationAI.GOOGLE.DEV. It provides **great all-around performance** across diverse tasks (creative writing, coding, reasoning, etc.), though without the explicit step-by-step "thinking" mechanism of the 2.5

series. **Intended Use Cases:** Because of its broad skills and multimodal nature, Flash 2.0 is used for everything from powering conversational assistants (like an upgraded version of Google's Bard) to performing complex tasks in business settings (summarizing long documents, analyzing videos, etc.). Developers choose Gemini Flash when they need a versatile model that can handle **various input types** and maintain coherence over long interactions. It's also appropriate for building agent-like systems that carry out multi-step procedures (since it was designed with those in mind). **Key Differences:** Gemini 2.0 Flash sits between the lightweight Flash-Lite and the reasoning-focused 2.5 Pro. Compared to **Flash-Lite**, it offers much stronger performance and the ability to handle rich media (images/videos) – but at a moderately higher cost and slightly higher latency. Compared to **Gemini 1.5** models, 2.0 Flash has improved quality and supports audio input/output (with separate pricing for audio tokens)AI.GOOGLE.DEV, reflecting a jump in multimodal integration. It's also notably **33% cheaper per token than earlier pricing** – Google simplified the pricing to $0.10/M input for text/image and $0.40/M output, making huge context usage more affordableDEVELOPERS.GOOGLEBLOG.COMTHREADS.NET. While not as explicitly "reasoning-heavy" as the later 2.5 Pro, Gemini 2.0 Flash is a workhorse model for general AI tasks, analogous to OpenAI's GPT-4 (but with a larger context and modality support when it launched).

Gemini 2.0 Flash-Lite – **Capabilities:** Flash-Lite 2.0 is the **smaller, cost-optimized sibling** of Gemini 2.0 Flash. It provides the same 1M token context window and core multimodal capabilities, but on a much leaner model that trades off some accuracy and sophistication to massively reduce cost. At only $0.075 per million input tokens and $0.30 per million output tokens

AI.GOOGLE.DEV, Flash-Lite is **even cheaper than many legacy models**, which is remarkable given it still understands images and long context. It's described by Google as *"our smallest and most cost effective model, built for at-scale usage."*AI.GOOGLE.DEV This suggests it likely has a significantly reduced parameter count (and possibly was akin to a Gemini 1.5 "8B" model updated with the new context). **Intended Use Cases:** Flash-Lite is ideal for applications that need to **deploy AI at massive scale or on edge devices**. For example, if a service needs to handle millions of requests with only straightforward queries, Flash-Lite can drastically cut costs. It's suitable for lightweight chatbot duties, text classification, or simple summarization tasks, especially when the user is cost-sensitive. It can also serve as a fallback model for less critical requests in a system that dynamically chooses models based on needed sophistication. **Key Differences:** The main difference is in performance and cost. Flash-Lite offers about the same raw *features* as Flash (long context, multimodal input), but its outputs are less advanced – think of it as the "budget tier" model. It will generally lag behind full Flash on complex linguistic tasks or understanding very nuanced instructions. However, its pricing is on par with or lower than even OpenAI's smallest GPT-4o Mini model, making it extremely attractive where high volume matters. In the Gemini family, Flash-Lite stands out by enabling **cost-effective scaling**: you can use it for repetitive or simpler tasks while reserving the more powerful (and expensive) models for truly hard problems.

Gemini 1.5 Flash – **Capabilities:** Gemini 1.5 Flash was the **fastest model of the Gemini 1.5 series**, introducing the idea of a 1M token context window to Google's lineup and focusing on quick, repeatable tasks. It's a multimodal model as well, and it significantly improved speed over the earlier Gemini 1.0. Google touted 1.5 Flash as having *"great performance for diverse, repetitive tasks"* while maintaining the long context

AI.GOOGLE.DEV. It has relatively strong conversational ability, can handle coding tasks (to a degree), and is efficient at scanning or generating large texts rapidly. **Intended Use Cases:** This model found its niche in scenarios such as real-time chatbot interactions (where latency matters), large-scale document parsing (where it could ingest big inputs quickly), and high-frequency API calls where throughput is important. For example, a platform that personalizes content for thousands of users could use 1.5 Flash to generate each user's content swiftly. **Key Differences:** Being a generation behind the 2.0 models, Gemini 1.5 Flash lacks some of the latest fine-tuning and might not be as accurate on complex prompts. It is, however, extremely cheap for its capabilities (pricing was $0.075/M input, $0.30–$0.60/M output depending on context length)AI.GOOGLE.DEV. One key difference in the 1.5 era was the introduction of **Gemini Flash-8B**, a further distilled version with only ~8 billion parameters. In fact, "Flash-Lite" as a concept really began with 1.5 – an 8B model that had the same pricing of $0.0375/M input and $0.15/M output for smaller promptsSIMONWILLISON.NETSIMONWILLISON.NET. So, Gemini 1.5 Flash (the full model) sits between that 8B lite model and the larger 1.5 Pro. It provides a balance of speed and competence, but for challenging tasks (like intricate reasoning or writing long code), it would be outperformed by the Pro version or newer models. In summary, 1.5 Flash was a milestone for enabling **million-token contexts at low cost**, paving the way for the improvements seen in Gemini 2.0.

Gemini 1.5 Pro – **Capabilities:** Gemini 1.5 Pro was the **high-end model** of the 1.5 series, tuned for maximum intelligence and context size. Notably, it came with a *"breakthrough 2 million token context window,"* meaning it could handle extremely long inputs (the first widely known model to go that high)

AI.GOOGLE.DEV. It offered strong performance on reasoning, coding, and complex knowledge tasks — Google positioned it for advanced enterprise use. For instance, 1.5 Pro could potentially take in a couple thousand pages of text and answer nuanced questions about them. It is also multimodal (like other Gemini models, it can process images along with text). **Intended Use Cases:** This model is aimed at tasks that require both **high intelligence and huge context**. Examples include: analyzing entire books or multi-document dossiers, performing long dialogs without losing track of early details, or handling data-extensive tasks like reviewing codebases or scientific literature in one go. Its multimodal capability also makes it useful for processing long transcripts with embedded images or for generating detailed reports that mix text and graphics. **Key Differences:** Compared to Gemini 1.5 Flash, the Pro model is far more powerful but also much slower and pricier. At the paid tier, its cost was $2.50/M input and $10/M

output for large prompts[AI.GOOGLE.DEV] – about **10× the cost of Flash**. This model was essentially Google's answer to models like GPT-4 (which had a 32K context at that time) and Anthropic's Claude 2 (100K context then). It significantly surpassed them in context length, though quality-wise it was in a similar league for tough tasks. One key difference from OpenAI's models: even 1.5 Pro's **vision capabilities** were integrated, whereas OpenAI did not add vision to GPT-4 until later. However, by late 2024, OpenAI's introduction of the o1 model (with heavy reasoning ability) meant the landscape shifted – 1.5 Pro is extremely competent, but newer "inference" models like o1 or Gemini 2.5 have since raised the bar on certain reasoning tasks. In short, Gemini 1.5 Pro filled the gap of ultra-long-context AI with strong general performance, and while it's now surpassed by the 2.x series, it remains relevant for its unique context length and still-high capability.

xAI Model

Grok 2 – **Capabilities:** Grok 2 is xAI's second-generation large language model, released in late 2024 as an improvement over their initial Grok (Beta). It features a **131K token context window** (approximately 128k tokens) and focuses on robust language understanding, coding, and factual question-answering. Elon Musk's xAI designed Grok to have a bit of a personality – reportedly it has a style inspired by the Hitchhiker's Guide to the Galaxy (witty and bold) – but underneath that style, it's a serious model geared toward high-end tasks. Grok 2 can handle complex reasoning and has some degree of *real-time knowledge access*, integrating up-to-date information via tools or retrieval (the specifics of which were evolving). According to an xAI announcement, Grok's training included a large swath of public data and it was taught to be a **"rebellious" chatbot with internet access**, aiming to answer queries with cutting-edge information and humor. **Intended Use Cases:** Grok 2 is positioned for users who need an AI similar to OpenAI's GPT-4 but possibly with fewer restrictions and more up-to-date data. It can be used for coding help, general Q&A, drafting content, and analysis of documents (within that 131k token limit). Enterprises interested in diversity of AI providers might integrate Grok 2 for tasks like report generation or research assistance, especially if they value the style or the possibility of more recent info via xAI's system. **Key Differences:** In terms of pricing and performance, Grok 2 sits roughly in the same class as OpenAI's GPT-4o model – its cost is $2 per million input and $10 per million output, which is on par with GPT-4o's $2.50/$10

[BLOG.PROMPTLAYER.COMOPENAI.COM]. One distinguishing factor is xAI's philosophy: Grok is intended to be a bit more unfiltered (within safe bounds) and *"answer almost anything"*, in contrast to the sometimes cautious refusals of models like ChatGPT. Technically, Grok 2 doesn't yet claim to surpass the very top OpenAI/Anthropic models on benchmarks, but it offers an alternative approach (and possibly advantages in real-time knowledge). It also had a smaller cousin *Grok Vision-Beta* for image tasks, but Grok 2 itself is primarily a text model (with 131k context for text). In summary, Grok 2 provides a **fresh entrant** in the advanced LLM field, combining a large context and solid reasoning abilities with a unique style – making it different from the more "corporate" AIs by OpenAI, while still being robust enough for enterprise use[BLOG.PROMPTLAYER.COM].

OpenAI GPT-4 and "o-Series" Models

GPT-4.5 Preview – **Capabilities:** GPT-4.5 is OpenAI's **largest GPT-series model** currently in a research preview. It's described as *"the largest GPT model designed for creative tasks and agentic planning"* OPENAI.COM. Essentially, GPT-4.5 sits between GPT-4 and a potential future GPT-5, incorporating new advances to handle extremely complex or open-ended tasks. It can produce highly elaborate and coherent content, engage in long-form dialogues planning actions or storylines, and likely has enhanced problem-solving skills. Like GPT-4, it supports a 128k token context, so it can take in very long inputs. This model is expected to excel at things like writing lengthy creative pieces (stories, scripts), devising strategies or plans (hence "agentic planning"), and possibly powering autonomous AI "agents" that need to make decisions. **Intended Use Cases:** Because it's in preview, GPT-4.5 is mainly used in research and by select developers to test its abilities. Ideal use cases include **high-end creative work** (e.g. drafting a novel or designing a marketing campaign with minimal human input) and complex decision support (e.g. analyzing vast data and suggesting actionable plans). It's also likely a testbed for future tools – one could use GPT-4.5 to orchestrate other models or services, given its planning focus. **Key Differences:** The most notable difference is that GPT-4.5 Preview is *extremely expensive* (roughly double the cost of base GPT-4 models) at $75/M input and $150/M output OPENAI.COM, which limits its practical use. OpenAI's pricing signals that 4.5 is computationally very heavy. It's also not generally available yet, indicating OpenAI is still refining it. Compared to **OpenAI o1-Pro** (another very large model focused on reasoning), GPT-4.5 might be more geared toward creativity and less toward pure logical reasoning. It doesn't explicitly advertise the tool use or vision that some other models have, so its strength lies in freeform generation and planning. In short, GPT-4.5 is an experimental powerhouse for when GPT-4 isn't enough, but it comes with steep costs and is only in limited use during this preview phase HELICONE.AI.

GPT-4o – **Capabilities:** GPT-4o is OpenAI's high-intelligence model for general use, essentially an upgraded version of GPT-4 optimized for both performance and cost. It retains GPT-4's strengths in understanding complex inputs and producing detailed, contextually accurate outputs, and it also has multimodal abilities (it can process images in addition to text). Importantly, GPT-4o operates with a 128k context window, allowing very long conversations or documents. It's noted as *"faster and cheaper than GPT-4 Turbo with stronger vision capabilities."* AZURE.MICROSOFT.COM. This means GPT-4o can analyze images or visual data more effectively than earlier models. **Intended Use Cases:** GPT-4o is a versatile model intended for a broad range of tasks – from answering tough research questions, writing high-quality content, to interpreting images (e.g., explaining what's in a photo or chart). Given its balance of power and price, many developers use GPT-4o as the workhorse model for complex tasks that require intelligence but need to be cost-effective at scale. For instance, a knowledge assistant ingesting and answering questions about lengthy documents

would benefit from GPT-4o's long context and accuracy. **Key Differences:** Compared to **GPT-4 (original)**, GPT-4o offers a larger context and lower pricing, making advanced capabilities more accessible (GPT-4 32k context was more limited and expensive). Compared to **GPT-4.5**, GPT-4o is smaller and less "creative"; however, it's vastly more economical ($2.50/M vs $75/M input)[OPENAI.COM]. Within OpenAI's lineup, GPT-4o is the top model for *general-purpose complex tasks* that don't specifically require the new reasoning paradigm of the o1 series. It's also positioned as a direct competitor to Anthropic's Claude 3.5/3.7 and Google's Gemini Flash in both capability and price. Its introduction of **vision** sets it apart from many older text-only models, enabling use cases that overlap with computer vision tasks. In essence, GPT-4o is the go-to model when one needs something smarter than GPT-3.5 but wants to avoid the extreme cost of something like o1-Pro or GPT-4.5[AZURE.MICROSOFT.COM].

GPT-4o Mini – **Capabilities:** GPT-4o Mini is the **small, cost-efficient variant** of GPT-4o. OpenAI released it as a replacement for the earlier GPT-3.5 Turbo series, offering significant improvements in understanding and context length at a low price. It's described as *"OpenAI's most cost-efficient small model, intended to replace GPT-3.5 Turbo"*

[BLOG.PROMPTLAYER.COM]. Despite being "mini," it still supports the full 128k context window of the GPT-4o family, which means it can handle very long conversations or inputs that GPT-3.5 could never manage. Its capabilities include decent reasoning, fast response generation, and effective handling of everyday tasks such as summarization, casual creative writing, and customer support dialogues. **Intended Use Cases:** GPT-4o Mini shines in high-volume, real-time scenarios. For example, if an application needs to process thousands of user chats or do on-the-fly personalization, this model provides the speed and affordability to do so. It's suitable for building chatbots that require some level of understanding beyond GPT-3.5's ability, performing batch document processing (like summarizing a bunch of articles), or acting as an assistant in apps where response time and cost are critical. **Key Differences:** With an input cost of only $0.15 per million tokens[OPENAI.COM], GPT-4o Mini is **20× cheaper on inputs** than full GPT-4o – a massive difference – while output tokens cost $0.60/M[OPENAI.COM]. This makes it comparable in price to smaller open-source models, yet GPT-4o Mini benefits from OpenAI's training (meaning it generally has far better quality than models like GPT-3.5 at similar price). The trade-off is raw power: it won't match GPT-4o on very complex queries or highly nuanced understanding. In benchmarks, it performs lower, as expected; however, it *"offers a cost-effective alternative with decent performance"* and is particularly suited for chaining or parallel calls in workflows[BLOG.PROMPTLAYER.COMBLOG.PROMPTLAYER.COM]. Essentially, GPT-4o Mini is the choice when you need **scale and speed** over absolute best accuracy – it brings some of the GPT-4 family strengths into a lightweight package.

OpenAI o1 – **Capabilities:** OpenAI o1 is a **frontier reasoning model** introduced as part of a new series in late 2024. It is specialized for complex, multi-step problem solving and supports tools, structured outputs, and even vision input

OPENAI.COM. The model has a 200k token context, indicating it can digest very large inputs. What sets o1 apart is that it was designed to *"think"* more deeply: it uses an internal chain-of-thought mechanism to reason through problems, leading to more accurate and coherent solutions on challenging tasks. For example, o1 is excellent at mathematical reasoning, logic puzzles, and debugging code, often significantly outperforming models that just do direct completion. It can also integrate tools/plug-ins (structured outputs) and analyze images as part of its input, making it a very flexible advanced model. **Intended Use Cases:** o1 is ideal for **STEM applications and any use case requiring deep reasoning**. This includes complex calculations, long-form logical deductions, solving engineering or scientific problems, and understanding visual data (like interpreting the content of an image or graph). It's also used where an answer needs to cite a reasoning process or use tools – for instance, an AI agent that can use a calculator or call external APIs might rely on o1. Essentially, if the task is something like "figure out a detailed solution step-by-step" or "analyze this diagram and give conclusions," o1 is a top choice. **Key Differences:** The o1 model inaugurated a shift in OpenAI's approach – compared to GPT-4, it is far more **reasoning-oriented**, albeit at a high cost and latency. Its pricing of $15/M input, $60/M output is significantly higher than GPT-4o's pricing OPENAI.COM. OpenAI specifically notes that the o1 series is suited for *"complex, multi-step tasks"*, recommending it for tough problems and STEM domains OPENAI.COM. Unlike GPT-4.5 (which is geared towards creative generation), o1 focuses on getting difficult answers right. Another difference is that o1 can produce **structured outputs** (e.g., in JSON) more reliably, which is useful for applications needing formatted answers. In comparison to Google's models, o1 was one of the first *inference-focused* models on the market (Google followed with Gemini Flash Thinking). Overall, OpenAI o1's introduction set a new standard for reliability on tasks that require the model to effectively *solve* problems, not just regurgitate information.

OpenAI o1 Pro – **Capabilities:** o1 Pro is an enhanced version of the o1 reasoning model, pushing the limits of reasoning even further by allocating more computational effort to each query. Internally, it can be instructed to use a "high reasoning effort" mode

HELICONE.AI, meaning it will think longer and consider more possibilities before finalizing an answer. The result is improved reliability and accuracy on extremely complex tasks – it *"thinks harder"* and often produces more refined solutions HELICONE.AI. Like o1, it supports tools and vision, and maintains the 200k token context. In essence, o1-Pro is designed for scenarios where you want the absolute best reasoning quality and are willing to incur heavy costs for it. **Intended Use Cases:** Given its extraordinary cost, o1-Pro is used sparingly, typically in high-stakes or exceptionally challenging problems. Use cases might

include advanced scientific research analysis, legal reasoning on very lengthy case documents, or solving difficult mathematical proofs. It could also be used in AI safety research to evaluate tricky prompts, since it's less likely to make reasoning errors. Some early users joked that o1-Pro is like having an *"AI PhD on call"* – you wouldn't ask it to do trivial tasks, but for something really hard, it might succeed where others fail. **Key Differences: Cost** is the glaring difference: at $150 per million input tokens and $600 per million output[HELICONE.AI], o1-Pro is the most expensive model on the market by a wide margin. This cost is 10× higher than base o1 and even surpasses GPT-4.5's pricing. The high price corresponds to using more computational steps ("effort") per token. Also, o1-Pro is the first OpenAI model that is *only* available through a newer API (the "Responses API") – not the standard chat completion endpoint – reflecting its special status[HELICONE.AI]. In practical terms, o1-Pro may only be worth using when simpler models (like o1 or GPT-4o) are not giving correct results, but the task is critical. It's significantly slower due to the extra computation. In benchmark terms, some reports indicate mixed results: o1-Pro can shine on certain problems (like very tricky coding or math tasks) but offers diminishing returns on others[HELICONE.AI]. It's a testament to how far one can push LLM reasoning – and also an experiment in what kinds of tasks justify its use. For most applications, o1-Pro is overkill; but it represents the cutting edge for maximum reasoning capability in 2025.

OpenAI o3 Mini – **Capabilities:** OpenAI o3-mini is a **small, cost-efficient reasoning model** introduced alongside o1. Despite the "mini" name, it's part of the reasoning-optimized family, focusing on coding, math, and science tasks with higher throughput. It has a 200k context length like the larger models [OPENAI.COM]. Essentially, o3-mini is designed to give some of the benefits of the advanced reasoning models but at a fraction of the cost, making it practical for frequent use. It's optimized to handle structured outputs and tool usage as well. For instance, o3-mini would do well in writing code snippets, solving medium-hard math problems, or answering science questions where factual accuracy is needed but the absolute top model isn't necessary. **Intended Use Cases:** Because of its low cost, o3-mini is suitable for **production use at scale** where moderate reasoning is required. Developers might use it to power features like code autocompletion, data cleaning scripts, or QA bots in technical domains. It's also a good default for multi-step reasoning in scenarios where using o1 would be too expensive. For example, an ed-tech app that helps students with math homework could use o3-mini to show steps and explanations without incurring enormous costs. **Key Differences:** OpenAI's pricing for o3-mini is $1.10/M input and $4.40/M output[OPENAI.COM], which is dramatically lower than o1 (by ~15× on output). This makes o3-mini roughly **60% cheaper than GPT-4o** for inputs, though GPT-4o still has an edge on certain general tasks[NEWS.YCOMBINATOR.COM]. The "o3" naming suggests it's a tier below "o2" (which presumably was skipped or internal) and o1 in capability. In practice, o3-mini often outperforms legacy models like GPT-3.5 on reasoning-heavy tasks, but will be outperformed by the larger o1 on very complex cases. Another difference: o3-mini is specifically *optimized for coding and STEM*, so it might lack broad world knowledge that models like GPT-4o have. This is by design – focusing on STEM lets it be smaller while

still effective in that niche[OPENAI.COMOPENAI.COM]. In sum, o3-mini fills an important gap: it brings down the cost of advanced reasoning, making techniques like chain-of-thought affordable for everyday use, at some tradeoff in raw power.

OpenAI o1 Mini – Capabilities: OpenAI o1-mini is a scaled-down version of the o1 model, created to offer **cost-efficient reasoning** while maintaining much of o1's core strengths. OpenAI noted that *"o1-mini achieves comparable performance on many useful reasoning tasks, while being significantly more cost efficient."*

[OPENAI.COM]. It is heavily optimized for STEM reasoning (math and coding in particular), nearly matching the full o1 on those benchmarks[OPENAI.COM]. Where it falls short is on tasks requiring broad general or factual knowledge outside of STEM, due to its smaller size and targeted training[OPENAI.COM]. It uses a 128k–131k token context (the precise context may be slightly lower than o1's 200k, as Azure documentation suggests 128k for o1-mini[AZURE.MICROSOFT.COM]). **Intended Use Cases:** o1-mini is intended for **developers and applications that need serious reasoning abilities but can't afford o1's cost** for every request. Use cases include math problem solvers, automated code debugging tools, and logical reasoning tutors – essentially, any scenario where the process of reasoning through a solution is needed, but a slightly lower accuracy than o1 is acceptable. It's also great for running multiple reasoning tasks in parallel (like solving many small problems at once) which would be prohibitively expensive with o1. **Key Differences:** The o1-mini model is *80% cheaper* than the original o1-preview was[OPENAI.COM]. In concrete pricing, this translates to about **$3.00 per million input and $12.00 per million output tokens** [NEWS.YCOMBINATOR.COM], versus o1's $15/$60. This huge drop in cost comes with only a modest drop in performance on many benchmarks – a trade-off favoring o1-mini for most practical purposes. However, o1-mini is not as capable when it comes to open-domain tasks (e.g., answering questions about history or literature) because it doesn't prioritize broad knowledge the way GPT-4o or full o1 might[OPENAI.COM]. It's very much a specialist. In terms of placement, o1-mini can be seen as bridging the gap between standard models (like GPT-4o) and the heavy-duty o1; it gives some of o1's reasoning prowess at a cost closer to GPT-4o's range. For many developers, this makes it the *go-to model for complex reasoning* by late 2024/early 2025, as it offers one of the best price-to-performance ratios for that category of problem[NEWS.YCOMBINATOR.COM].

Sources: Official pricing and model info from Anthropic [ANTHROPIC.COMANTHROPIC.COM], DeepSeek[API-DOCS.DEEPSEEK.COMAPI-DOCS.DEEPSEEK.COM], Google DeepMind[AI.GOOGLE.DEVBLOG.GOOGLE], xAI[DOCS.X.AI], and OpenAI[OPENAI.COMOPENAI.COM]. Additional details on capabilities and use cases are drawn from developer documentation and press releases[OPENAI.COMBLOG.PROMPTLAYER.COMHELICONE.AI]. Each model summary incorporates the

latest known information (as of 2025) to highlight where each model excels and how they differ from one another.

 Sources