Pricing for Top Large Language Models (LLMs) (2025)

Below we compile current pricing data for **50+ LLMs**, organized by model provider and service type. Pricing is given per 1,000 tokens for **input** (prompt) and **output** (completion) tokens, with context window sizes noted where available. We also note available pricing tiers (e.g. batch discounts, enterprise plans) and any special discount triggers. All data is sourced from official provider documentation as of early 2025.

1. Commercial API Models

These models are accessed via the providers' own APIs and are typically closed-source. Pricing is pay-as-you-go unless otherwise noted. **Batch** or bulk-processing discounts and enterprise volume plans are highlighted where applicable.

OpenAI – GPT Series Models

OpenAI offers multiple GPT series models with different capabilities and context lengths. Pricing is usage-based, and **Batch API** usage can save **50%** on token costs.

Model (Context)	Input $/1K tokens	Output $/1K tokens	Notable Tiers/Discounts	Source
GPT-4 (8k)	$0.03	$0.06	N/A (standard rate)	
GPT-4 (32k)	$0.06	$0.12	N/A	
GPT-4.5 (128k)	$0.075	$0.150	*Research preview* (higher capability)	
GPT-4 Turbo (128k)	$0.01	$0.03	Vision-enabled variant available at same rate	
GPT-4o (128k)	$0.005	$0.015	High-performance optimized GPT-4 (faster responses)	

GPT-4o Mini (128k)	$0.00015	$0.00060	Small, cost-efficient model for everyday tasks
GPT-3.5 Turbo (4k)	$0.0015	$0.002	N/A (base 4k context model)
GPT-3.5 Turbo (16k)	$0.0005	$0.0015	N/A (extended 16k context)

Notes: "Cached input" prices (not shown above) are 50% of input rates for repeat prompts. OpenAI's **Batch** API for asynchronous jobs offers **50% off** both input and output token costs. Enterprise plans can negotiate volume discounts or higher rate limits via sales.

Anthropic – Claude Models

Anthropic's Claude series includes models optimized for chat and assistant tasks with high context windows (up to 100K tokens on Claude 2). Pricing is pay-as-you-go; Amazon Bedrock offers batch and caching discounts for Claude via its service (noted below).

Model	Input $/1K	Output $/1K	Special Tiers/Discounts	Source
Claude Instant (≈9k)	$0.0008	$0.0024	Fast, lightweight model (lower latency)	
Claude 2 (100k)	$0.008	$0.024	100k token context assistant	
Claude 2.1 (100k)	$0.008	$0.024	Same pricing as Claude 2 (model iteration)	
Claude 3.5 Sonnet (128k)*	$0.003	$0.015	Newer Claude 3.5 model ("Sonnet" version, ~128k context)	
Claude 3 Opus (>=128k)*	$0.015	$0.075	High-end Claude 3 model ("Opus", largest context, higher quality)	

Claude 3.x models (Haiku, Sonnet, Opus) are available via partners (e.g. AWS Bedrock) – "Haiku" has a smaller context/window at much lower cost, "Sonnet" medium, and "Opus" the largest and most costly..

Notes: Through Amazon Bedrock's **On-Demand** API, **batch mode** calls for Claude models are **50% off** input/output costs. Bedrock also offers caching: repeated prompts ("cache read") cost a fraction of input price. Enterprise volume pricing is available through Anthropic's sales team (not publicly listed).

Cohere – Command Models

Cohere offers the Command family of models for text generation and chat. Pricing is per million tokens (converted to per-1K below) and varies by model size. Fine-tuning and an **Enterprise** tier are available for custom needs.

Model	Input $/1K	Output $/1K	Other Pricing Tiers	Source
Command *R+* (128k)	$0.003	$0.015	Largest model (higher quality, longer context)	
Command *R* (128k)	$0.0005	$0.0015	Standard model for production use	
Command *R (fine-tuned)*	$0.002	$0.004	Fine-tuned model cost (plus $0.008/1K for training tokens)	
Command A *(Alpha)*	$0.0025	$0.010	"Command A" efficient model for enterprises (shorter context)	

Notes: All Cohere usage is pay-go with monthly billing. The above rates are **per 1M tokens** broken down per 1K (e.g. Command R $0.50/M = $0.0005/1K). An **Enterprise tier** offers dedicated instances and support; pricing is custom-negotiated. Free trial API keys are available with limited rate limits.

AI21 Labs – Jurassic-2 / Jamba Models

AI21 provides the **Jurassic-2** family and newer **Jamba** models via the AI21 Studio API. Pricing is per million tokens (converted to per-1K below). A free trial with $10 credit is offered, and enterprise plans with volume discounts are available.

Model	Input $/1K	Output $/1K	Context Window	Source

Jamba 1.5 Mini	$0.0002	$0.0004	*Up to 4k tokens*
Jamba 1.5 Large	$0.002	$0.008	*Long context (≥8k)*
Jurassic-2 Mid	$0.0125	$0.0125	8192 tokens
Jurassic-2 Ultra	$0.0188	$0.0188	8192 tokens

Notes: AI21's pricing **per token** is about 30% more text per token compared to some others (due to larger token definition), effectively providing cost savings. All plans are pay-as-you-go by default. **Volume discounts** and private hosting are offered under custom enterprise plans. Amazon Bedrock also hosts Jurassic-2 models with identical pricing and offers 50% batch discounts.

IBM watsonx – Granite Models

IBM's **Granite** family (open-source foundation models optimized for enterprise) is offered on the IBM watsonx.ai platform. Pricing can be pay-per-token or hourly, depending on deployment mode. Below are pay-per-token rates for IBM-hosted models:

Model	Input/Output $/1K	Context Window	Source
Granite-3-2B Instruct (v3.1)	~$0.00010*	128k tokens	
Granite-3-8B Instruct (v3.1)	~$0.00020*	128k tokens	

*IBM pricing is listed per 1M tokens: e.g. Granite-3-2B Instruct is **$0.10 per 1M** ($0.00010 per 1K). Granite models with "lab" or older versions (e.g. Granite-7B Lab) are charged higher rates ($5.22 per 1M) due to legacy hosting costs.*

Notes: IBM offers **tiered plans**: a Free Trial (50K tokens/month free) and paid tiers (Essentials, Standard) that include monthly token allowances. The token prices above are for excess usage beyond included amounts. **Custom model hosting** on IBM Cloud or other clouds is also available, charged hourly instead of per token. Enterprise customers can negotiate bespoke pricing and larger deployments.

2. Hosted Open-Source Models

This category covers open-source LLMs provided as managed services. The models are typically freely available to self-host, but here we include pricing for using them through cloud providers or API platforms (which charge for infrastructure or usage).

Meta Llama 2 via AWS Bedrock

Meta's **Llama 2** Chat models (13B and 70B parameters) are offered on Amazon Bedrock (fully-managed). Pricing is per 1,000 tokens on input and output. Fine-tuning (customization) and provisioned throughput options are available for these models.

Model	Input $/1K	Output $/1K	Notes	Source
Llama 2 Chat 13B	$0.00075	$0.00100	On-Demand usage (pay-go)	
Llama 2 Chat 70B	$0.00195	$0.00256	On-Demand usage (pay-go)	
Llama 2 (Fine-tuned)	$0.00149	N/A	$0.00149/1K training tokens (for custom)	

Notes: These rates are for on-demand inference. **Batch processing** on Bedrock (where available) would be 50% of these rates. After fine-tuning a Llama 2 model on Bedrock, usage is charged per-hour of model unit (rather than per token) for the custom model. Provisioned throughput (reserved capacity) pricing for Llama 2 is available at ~$21.18/hour per model unit (1-month commit) or $13.08/hour (6-month commit) for 13B/70B.

Hugging Face Inference Endpoints (Open Models)

Hugging Face allows hosting *any* open-source model as a dedicated endpoint. Pricing is not per token but per **instance-hour** of the hardware used. For example, a basic CPU instance costs **$0.033/hour**, while a single T4 GPU starts at **$0.40/hour**, and larger multi-GPU setups scale up to ~$8.30/hour. There are no token limits; cost is purely time-based. This option suits high-volume or custom deployments of open models (like Falcon, GPT-J, MPT, etc.). **Autoscaling** and **enterprise plans** (with custom SLAs) are available.

Notes: Since pricing is hourly, the effective per-1K-token cost varies by model speed and load. Hugging Face no longer offers a simple pay-per-request Inference API for new customers (it was sunset in favor of endpoints). However, *prepaid* tiers (e.g. $9/month for certain throughput) were previously available for low-volume use – existing users may still have access.

Other Hosted Open-Source Services

- **Azure AI Model Catalog:** Azure's Model Catalog can deploy open models (like Llama 2, Mistral 7B, etc.) on Azure infrastructure. Pricing is typically similar to Hugging Face (based on VM/hour rates) rather than per token. (As of 2025, Azure has partnered with Meta to host Llama 2 and Code Llama for Azure users – pricing is case-by-case via Azure ML or endpoint billing).

- **Databricks / MosaicML:** Databricks (which acquired MosaicML) offers the MPT series and other open models as part of its platform. Pricing is integrated into Databricks compute costs (per cluster GPU hour) rather than explicit per-token fees. MosaicML's MPT-30B, etc., can also be self-hosted or deployed on their platform under custom contracts.

---

3. Pay-as-You-Go Cloud Models

This section covers LLM services from cloud providers (and third-party models offered through cloud platforms) where pricing is purely usage-based with no long-term commitment. These often overlap with the above categories but are listed here by platform for clarity, including any unique models (like Amazon's Titan or Google's PaLM/Gemini). Volume discounts or **provisioned throughput** (commitment contracts) are noted where applicable.

Microsoft Azure OpenAI Service (Pay-go)

Azure OpenAI Service provides OpenAI models on Azure, with identical token pricing to OpenAI's standard rates. Azure charges by consumption on your Azure bill (no separate OpenAI account needed). All models are pay-as-you-go; **no upfront fees** (but no explicit volume discount published).

Model (via Azure)	Input $/1K	Output $/1K	Context	Source
GPT-3.5 Turbo (4k)	$0.0015	$0.0020	4k	
GPT-3.5 Turbo (16k)	$0.0005	$0.0015	16k	
GPT-4 (8k)	$0.03	$0.06	8k	
GPT-4 (32k)	$0.06	$0.12	32k	
GPT-4 Turbo (128k)	$0.01	$0.03	128k	

GPT-4 Turbo w/ Vision (128k)	$0.01	$0.03	128k
GPT-4o (128k)	$0.005	$0.015	128k

Notes: Azure OpenAI also offers older base models (e.g., Davinci-002) and embeddings, charged per 1K tokens at rates similar to OpenAI's legacy prices (Davinci ~$0.02/1K). Fine-tuning on Azure is charged per 1K training tokens plus an hourly hosting fee (e.g. ~$0.0004/1K for training Davinci-002 and $1.70/hour hosting). Azure does not currently offer an official batch discount program for OpenAI models (all requests are on-demand), but enterprise agreements may provide negotiated discounts for large Azure spend.

Amazon Bedrock (On-Demand)

Amazon Bedrock is AWS's managed service hosting various foundation models from AWS and partners. **On-Demand** pricing (pay-go) is per token for text models. **Batch** jobs cost 50% of on-demand rates for supported models. Below are Bedrock's own models and unique offerings:

- **Amazon Titan** – AWS's proprietary LLMs:
  - **Titan Text Lite** (GPT-like, 4k context): **$0.00015** per 1K input tokens, **$0.00020** per 1K output tokens. *(Very low-cost basic model.)*
  - **Titan Text Express** (8k context, more powerful): **$0.0008** per 1K input, **$0.0016** per 1K output . *(Higher capability than Lite, still cheap.)*
  - **Titan Embeddings**: $0.0001–$0.00002 per 1K tokens (input only) for text embeddings models.
  - **Provisioned throughput:** Titan models can be reserved: e.g., Text Lite at ~$5.10/hour (6-mo commit) and Text Express at ~$14.80/hour (6-mo) for guaranteed capacity.
- **AI21 Jurassic-2** – (See **AI21** in Commercial section for pricing; Bedrock matches those rates. Batch = 50% off.)
- **Anthropic Claude** – (See **Anthropic** section; Bedrock offers Claude Instant, 2, 3 etc. at the same token prices. Batch = 50% off, caching discounts available.)
- **Cohere Command** – (Bedrock hosts Cohere models; pricing not explicitly listed on public docs, but would align with Cohere's pricing per million.)
- **Stability AI** – Bedrock hosts **Stable Diffusion** image generation; priced per image (e.g. $0.04 per 1024×1024 image for standard quality). *[Not token-based, so omitted from tables.]*

Notes: Bedrock's unique **"Prompt Caching"** feature lets users cache prompts/outputs. **Cache reads** are very cheap (e.g. $0.0003 per 1K for Claude 3.5 Sonnet) to incentivize reuse. **Provisioned throughput**

pricing allows buying dedicated capacity by model unit (with 1-mo or 6-mo commitments) – suitable for consistent high-volume usage. For example, Claude Instant is ~$22/hour per unit on a 6-month term. Enterprise customers can also get custom pricing via AWS sales.

Google Cloud Vertex AI / PaLM & Gemini

Google's Vertex AI offers Generative AI models on a pay-as-you-go basis, charged by **characters** instead of tokens. (Approx. 1 token ≈ 4 chars.) Below are key model prices, converted to per-1K-token estimates for comparison:

Model (Google)	Input	Output	Context	Source
PaLM 2 Text-Bison	~$0.004/1K tokens	~$0.004/1K tokens	4k	
PaLM 2 Chat-Bison	~$0.002/1K tokens	~$0.002/1K tokens	4k	
Gemini 1.0 Pro	$0.0005/1K	$0.0015/1K	128k	
Gemini 1.5 Pro	$0.0035/1K	$0.0105/1K	128k+	
Gemini 2.0 Flash (1M)	$0.0001/1K	$0.0004/1K	1,000k	
Gemini 2.0 Flash-Lite (1M)	$0.000075/1K	$0.0003/1K	1,000k	

Notes: *PaLM 2* models are billed at $0.0010 per 1,000 **characters** for Text-Bison, and $0.0005 per 1,000 chars for Chat-Bison (roughly $0.004 and $0.002 per 1K tokens respectively). Google's next-gen *Gemini* models (available via the separate Google AI Studio/Gemini API) have extremely low pricing to encourage adoption. For instance, **Gemini 2.0 Flash** (1M context) costs only $0.10 per 1M input tokens (i.e. $0.0001 per 1K) . During previews, Google often provides promotional discounts (Vertex AI Gen AI was 100% discounted during its preview phase). Enterprise pricing tiers can include committed-use discounts or monthly spend discounts through Google Cloud marketplace deals.

References: All pricing data above is from official documentation or portals: OpenAI, Azure, Anthropic/AWS, Cohere, AI21, IBM, AWS Bedrock, and Google Cloud/Google AI . Please consult the linked documentation for the most up-to-date information and any region-specific pricing variations.