https://llm-price.com/
https://openai.com/pricing
https://www.anthropic.com/pricing
https://cohere.ai/pricing

Understanding LLM Pricing and Cost Comparison

Background

Many organizations are successfully integrating LLMs (like GPT or Claude) into production for chatbots, document summaries, data parsing, etc.

While LLMs are easy to start with and seem cheap initially, costs can grow quickly as usage scales (more tokens processed, more requests).

It’s crucial to understand how LLMs are priced and plan for the real costs as your applications scale.

How Are LLMs Priced?

Two Main Types of LLMs:

Managed LLM Providers (e.g., OpenAI, Anthropic)

Pricing is based on the number of tokens processed (input + output tokens).

Tokens are units of text (roughly 1 token ≈ 4 characters or 0.75 words).

Costs vary depending on model size, capability, and context window (max tokens per request).

Managed providers charge per million tokens processed.

Open-Source LLMs (e.g., Llama2, Mistral, Bert)

No per-token charge because you self-host.

Costs come from infrastructure (GPUs, memory, compute instances) needed to run the model.

Larger models require more expensive hardware and more resources.

Managed LLM Pricing Examples (OpenAI)

Model	Input Cost (per 1M tokens)	Output Cost (per 1M tokens)	Context Window (max tokens)
gpt-4o	$5	$15	128K
gpt-4-turbo	$10	$30	128K
gpt-4	$30	$60	32K
gpt-3.5-turbo-0125	$0.50	$1.50	16K

Anthropic’s Claude models have similar tiered pricing but generally vary in cost based on model size and capabilities.

Open-Source LLM Model Examples

Model	Parameters	Memory Required	GPUs Required

Mistral-7b-v0.3	7B	24GB	1 GPU
Mistral-8x7B-Instruct	56B	64GB	3 GPUs
Llama3-8b	8B	20GB	1 GPU
Llama3-70B	70B	160GB	8 GPUs
Bert	110M	8GB	1 GPU or CPU

Larger open-source models require significant hardware and can cost thousands of dollars per month on cloud providers.

Smaller models like Bert can be run for a few hundred dollars per month.

You'll need to scale horizontally (more instances) to handle higher traffic, increasing costs linearly.

Estimating Real-World Costs: Example Use Case (Sentiment Analysis)
Assume 30 requests/minute

Average input prompt size: 150 tokens

Average output size: 45 tokens

Using OpenAI GPT-4-Turbo pricing:

Input tokens:
30 req/min × 150 tokens × 60 min = 270,000 tokens/hour
270,000 × 24 hours × $10/1M = $64.80/day

Output tokens:
30 req/min × 45 tokens × 60 min = 81,000 tokens/hour
81,000 × 24 hours × $30/1M = $58.32/day

Total daily cost: $64.80 + $58.32 = $123.12/day (~$3,693/month)

Open-Source Deployment Cost Example (Llama3-8b on AWS)
Using AWS g5.2xLarge instance (about $1.212/hour)

Single instance cost: $29.08/day or ~$872/month

For safety and scalability, using 2 replicas: ~$1,744/month

Costs can reduce if using spot instances (cheaper, but less reliable)

No token cost, but you pay for infrastructure 24/7

Which Option Should You Choose?
Managed LLMs are easier to scale, maintain, and update, but can be expensive at scale.

Open-source LLMs can be cheaper at high volumes but require managing infrastructure, scaling, and maintenance.

Decision depends on:

Security and data privacy requirements

Traffic volume and latency needs

Model accuracy and complexity

Your team’s infrastructure capabilities

Cost constraints and expected growth

Additional Costs to Consider
RAG (Retrieval Augmented Generation) Pipelines: Storing and retrieving external context/data for the LLM adds costs (vector databases, caching, APIs).

Supporting Infrastructure: Autoscaling, monitoring, model routing, security layers, and additional ML components add complexity and cost.

These overheads should be included in your total cost estimate.

Conclusion
LLM pricing is complex and varies based on many factors.

Carefully track tokens, latency, accuracy, and infrastructure costs.

Revisit your cost estimates regularly as models and pricing evolve.

Balance between managed services and self-hosting based on your unique needs.