Cloudflare AI Gateway

Pricing, Cost Control & Value

Cloudflare Solutions Engineering
cloudflare.com/ai-gateway

OVERVIEW

What Is AI Gateway?

One Line of Code. Full Visibility & Control.

WITHOUT AI GATEWAY

- No visibility into token spend
- No cache — identical requests billed again
- Runaway cost if usage spikes
- Vendor lock-in — hard to switch models

WITH AI GATEWAY

- Real-time analytics — tokens, cost, latency
- Response caching for repeated queries
- Rate limiting to cap spend
- Model fallback — switch providers instantly

Supports **OpenAI, Anthropic, Google Gemini, Workers AI, Replicate** and more — all through a single proxy endpoint

AI Gateway Pricing

FREE — ALL PLANS

- Dashboard analytics
- Response caching
- Rate limiting
- DLP scanning (2 profiles)
- Model fallback & retry

PERSISTENT LOGS

Workers Free

100,000 logs

across all gateways

Workers Paid

10,000,000 logs

per gateway

PAID ADD-ONS

Guardrails

Billed as Workers AI tokens

(Llama Guard 3 8B inference)

Logpush (Paid plan only)

10M req/mo free

+$0.05 per million after

AI Gateway does not charge per token or per request for core features. You pay the model provider — not Cloudflare.

Workers AI — On-Platform Model Pricing

BILLING UNIT

Neurons = GPU compute units. Billed at **$0.011 / 1,000 Neurons**.

Free allocation: **10,000 Neurons/day** on all plans.

KEY LLM MODELS (PER M TOKENS)

Model	Input	Output
Llama 3.2 1B	$0.027	$0.201
Llama 3.1 8B fp8-fast	$0.045	$0.384
Llama 3.1 70B fp8-fast	**$0.293**	**$2.253**
DeepSeek R1 32B distill	$0.497	$4.881
Qwen3 30B (fp8)	$0.051	$0.335

OTHER MODEL TYPES

Embeddings

BGE-M3 from **$0.012/M tokens**

Image Generation

Flux-1-Schnell from **$0.0000528** per 512×512 tile

Speech / Audio

Whisper from **$0.0005/audio minute**

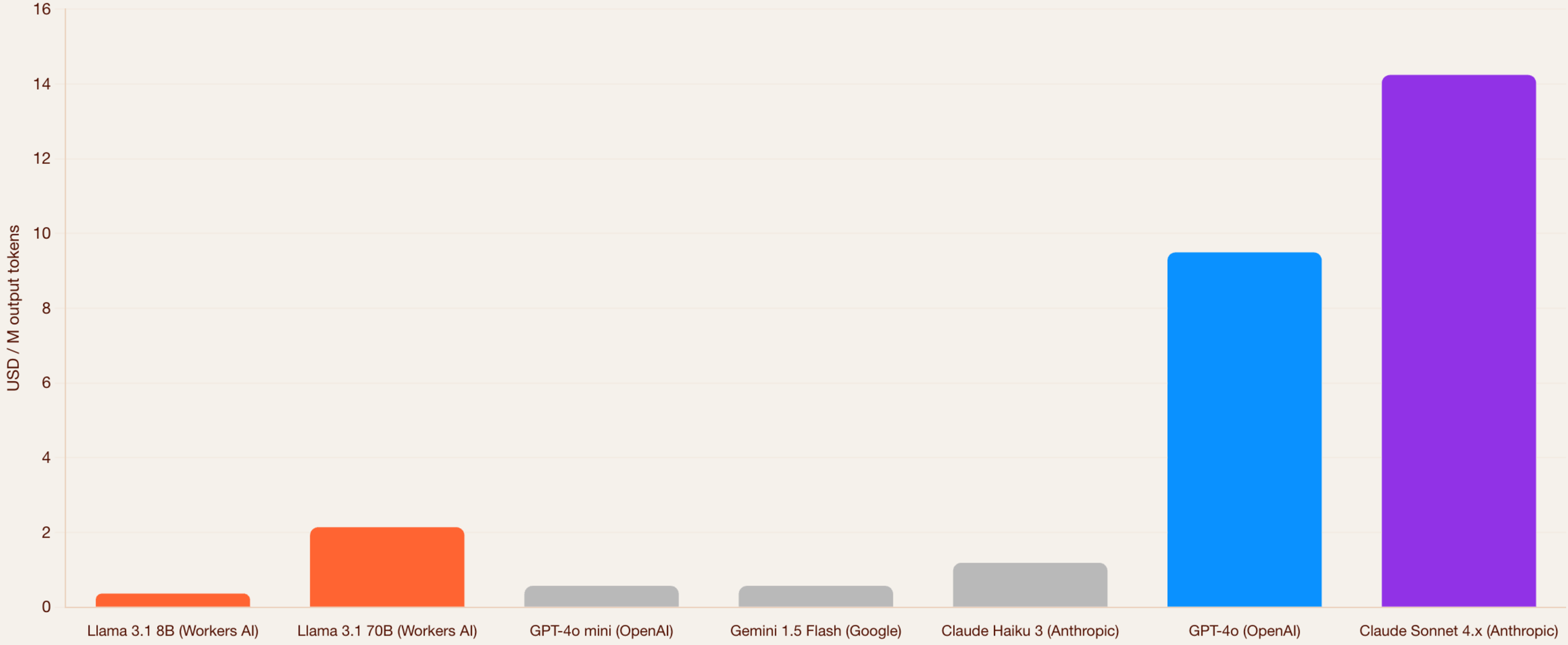
All Workers AI models run on **serverless GPUs** — no reservation or minimum commitment required.

COST COMPARISON

Workers AI vs. Frontier LLMs

Output Token Cost: Workers AI vs. Frontier Models

USD per million output tokens — the dominant cost driver in most AI workloads



Source: developers.cloudflare.com/workers-ai/platform/pricing — verified May 2026. Frontier prices from respective provider pages.

AI Gateway: Built-In Cost Controls

Response Caching

Cache identical responses from model providers. Cached hits cost **$0** — you're served from Cloudflare's edge.

Best for: FAQ bots, repeated template queries, demos

Model Fallback

Automatically route to a cheaper or alternative model if the primary fails. Route low-stakes requests to smaller models to reduce cost.

Rate Limiting

Set hard caps on requests per minute/hour per gateway or per consumer. Prevents runaway cost from unexpected traffic spikes.

Analytics & Logging

Full per-request visibility: model used, token count, cost, latency. Identify expensive prompts and optimise before they scale.

Up to **10M logs/gateway/month** on Workers Paid

Note: Current caching is **exact-match** on identical requests. **Semantic caching** (similar queries) is on the AI Gateway roadmap.

VALUE PROPOSITION

The Real ROI

Why AI Gateway Changes the Economics

$0

Gateway cost for
core features

4–17×

Cheaper than frontier
models via Workers AI

1 line

of code to add
analytics, caching & control

Security Included

DLP scanning, Guardrails (Llama Guard), and Data Loss Prevention — no separate security stack needed for AI endpoints.

Provider Agnostic

OpenAI, Anthropic, Gemini, Workers AI, Replicate and more. Move between providers in seconds with no code changes — full pricing optionality.

Get started free: developers.cloudflare.com/ai-gateway/get-started