

A Deflação Histórica do Custo de Tokens (2024–2026)

Evolução do custo de inferência por 1 milhão de tokens de entrada e saída.

MATRIZ DE ALOCAÇÃO ESTRATÉGICA DE RECURSOS

CUSTO VS VELOCIDADE

 **Raciocínio de Alta Precisão** Elite

Auditorias, código crítico, decisões de crédito e planejamento estratégico.

US\$ 3.00 – US\$ 4.50 / 1M tokens

 **Soberania & On-Premise** Zero Risk

Dados sigilosos, defesa, saúde privada e conformidade estrita com LGPD.

Custo Fixo de GPU Interna

 **Automação em Massa** Sweet Spot

Triagem de dados, suporte ao cliente, extrações estruturadas e e-mails.

US\$ 0.10 – US\$ 0.55 / 1M tokens

 **Real-Time & Voz Interativa** < 150ms

Agentes de voz bidirecionais, busca instantânea e automações de tela.

Sub-150ms TTFT

DESTAQUE DE MERCADO

Queda de 94.2% em 2 anos

A inteligência mais rápida e barata da história da computação

TI

DIRETRIZ DE ENGENHARIA

Adoção de Model Routers inteligentes reduz a fatura de computação em até 70% sem sacrificar acurácia.

CANAL TELEGRAM

t.me/oquevemai

Fontes: Artificial Analysis Price Index 2026

oquevem.ai · Intelligence Hub

A Deflação Histórica do Custo de Tokens (2024–2026)

Evolução do custo de inferência por 1 milhão de tokens de entrada e saída.

🎯

MATRIZ DE ALOCAÇÃO ESTRATÉGICA DE RECURSOS

CUSTO

VS

VELOCIDADE

🧠

Raciocínio de Alta Precisão

Elite

Auditorias, código crítico, decisões de crédito e planejamento estratégico.

US\$ 3.00 – US\$ 4.50 / 1M tokens

⚡

Automação em Massa

Sweet Spot

Triagem de dados, suporte ao cliente, extrações estruturadas e e-mails.

US\$ 0.10 – US\$ 0.55 / 1M tokens

🛡️

Soberania & On-Premise

Zero Risk

Dados sigilosos, defesa, saúde privada e conformidade estrita com LGPD.

Custo Fixo de GPU Interna

🚀

Real-Time & Voz Interativa

< 150ms

Agentes de voz bidirecionais, busca instantânea e automações de tela.

Sub-150ms TTFT

DESTAQUE DE MERCADO

A inteligência mais rápida e barata da história da computação

Queda de 94.2% em 2 anos

TI

DIRETRIZ DE ENGENHARIA

Adoção de Model Routers inteligentes reduz a fatura de computação em até 70% sem sacrificar acurácia.

CANAL TELEGRAM

t.me/oquevemai

Fontes: Artificial Analysis Price Index 2026

oquem.ai - Intelligence Hub