

Prompt Caching: Redução de 80% no Custo de Contexto Recorrente

Impacto econômico do reaproveitamento de tokens de contexto em assistentes com manuais e código.

🎯

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

Respostas 4x mais rápidas em conversas longas e bases de suporte

-80% de Custo em Tokens de Entrada

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: Anthropic API Performance Insights 2026

oquevem.ai · Intelligence Hub

CUSTO VS VELOCIDADE

Elite

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

US\$ 3.00 – US\$ 4.50 / 1M tokens

Sweet Spot

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

US\$ 0.10 – US\$ 0.55 / 1M tokens

Zero Risk

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

Custo Fixo de GPU Interna

< 150ms

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

Sub-150ms TTFT

Respostas 4x mais rápidas em conversas longas e bases de suporte

-80% de Custo em Tokens de Entrada

TI

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

CANAL TELEGRAM
t.me/oquevemais