

Frontier Models 2026: Custo vs Inteligência

Mapeamento de Score de Raciocínio (MMLU-Pro / LMSYS) vs Custo de Inferência por 1M de Tokens.

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

-94% no custo em 24 meses

De \$60.00/1M tokens para \$0.10 – \$0.55/1M tokens em modelos de alta capacidade

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: LMSYS Arena, Artificial Analysis, OpenAI, Anthropic, DeepSeek (2026)

oquevem.ai · Intelligence Hub

Frontier Models 2026: Custo vs Inteligência

Mapeamento de Score de Raciocínio (MMLU-Pro / LMSYS) vs Custo de Inferência por 1M de Tokens.

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

-94% no custo em 24 meses

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: LMSYS Arena, Artificial Analysis, OpenAI, Anthropic, DeepSeek (2026)

oquem.ai - Intelligence Hub