

Modelos Compactos (SLMs) na Borda vs Modelos de Fronteira

Casos de uso e viabilidade econômica de modelos de 1B a 8B parâmetros em dispositivos locais.

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

Privacidade local total em smartphones, notebooks e veículos inteligentes

Zero Custo de API por Usuário

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: Apple Intelligence Whitepaper, Qualcomm AI Research 2026

oquevem.ai · Intelligence Hub

Modelos Compactos (SLMs) na Borda vs Modelos de Fronteira

Casos de uso e viabilidade econômica de modelos de 1B a 8B parâmetros em dispositivos locais.

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

Privacidade local total em smartphones, notebooks e veículos inteligentes

Zero Custo de API por Usuário

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: Apple Intelligence Whitepaper, Qualcomm AI Research 2026

oquevem.ai - Intelligence Hub