

Janelas de Contexto: 128k vs 1M vs 2M Tokens

Precisão de recuperação (Needle in a Haystack) e custo de processamento em documentos longos.

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

Elimina a necessidade de segmentação excessiva de dados em RAG simples

99.8% Precisão em 1M Tokens (Gemini/Claude)

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 Context Benchmark, Google DeepMind Research 2026

oquevem.ai · Intelligence Hub

Janelas de Contexto: 128k vs 1M vs 2M Tokens

Precisão de recuperação (Needle in a Haystack) e custo de processamento em documentos longos.

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

Elimina a necessidade de segmentação excessiva de dados em RAG simples

99.8% Precisão em 1M Tokens (Gemini/Claude)

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 Context Benchmark, Google DeepMind Research 2026

oquem.ai - Intelligence Hub