

Fine-Tuning & LoRA vs RAG & Prompt Engineering

Análise de custo total de propriedade (TCO) para especialização de modelos em dados proprietários.

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

Fine-tuning reservado apenas para tom de voz ultrasensível e vocabulários técnicos raros

RAG Híbrido vence em 78% dos casos corporativos

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: Pinecone & Cohere Enterprise Survey 2026

oquevem.ai · Intelligence Hub

Fine-Tuning & LoRA vs RAG & Prompt Engineering

Análise de custo total de propriedade (TCO) para especialização de modelos em dados proprietários.

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

Fine-tuning reservado apenas para tom de voz ultrasensível e vocabulários técnicos raros

RAG Híbrido vence em 78% dos casos corporativos

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: Pinecone & Cohere Enterprise Survey 2026

oquem.ai - Intelligence Hub