

Sistemas Críticos com Custo Marginal Próximo de Zero

Decisões de arquitetura, benchmarks reais em produção e computação determinística.

Rodrigo Almeida | Desenvolvedor de Software · Technical Contractor

orodrigoalme.com
rodrigo@orodrigoalme.com

Atuação direta: Projeto, construo e opero em produção sistemas críticos, atuando da arquitetura à segurança de borda, do pipeline de dados ao produto final. Atendo software houses, agências e empresas de tecnologia que precisam absorver gargalos de infraestrutura, desacoplamento, segurança ou inteligência artificial aplicada, com autonomia sênior e sem camadas de intermediação.

-95%

CUSTO DE INFERÊNCIA

Filtros de sinal na borda antes de qualquer chamada a VLM

> 99%

ECONOMIA EM EMBEDDINGS

Zero custo de tokens de API com FastEmbed local na CPU

< 2,5%

BOUNCE RATE DE DADOS

Validação estrita em fontes públicas oficiais e registros DNS/MX

Visão Sintética dos Três Sistemas em Produção

SISTEMA	GARGALO RESOLVIDO	BENCHMARK / RESULTADO EM PRODUÇÃO
Trueoria Engine Visão, borda & auditoria	Integridade fotográfica e documental para sinistros: detecção de fraudes, edições e imagens geradas por IA sem vazamento de PII's.	-95% em custo de inferência Resposta em ~2,1s Custódia de integridade ISO/IEC 27037
Memória L4 + Graph RAG MCP local contínuo	Impede a amnésia de agentes de software entre sessões, eliminando custos de embeddings e sprawl de bancos.	>99% de economia em embeddings Latência de 22 a 35 ms na tool Zero vazamento de código de clientes
Motor de Inteligência B2B Prospecção determinística	Descoberta de sócios e decisores reais na Receita Federal em 15 s, sem pagar plataformas de data brokers.	Pesquisa de 40 min para ~15 s Custo de dados < R\$ 25/mês Acurácia de sócios > 88%

COMPARATIVO DE EFICIÊNCIA: MERCADO VS ARQUITETURA DETERMINÍSTICA

Base: 50.000 requisições/mês em infraestrutura de produção

Custo Mensal de APIs e Nuvem

-96,5% de Custo

Mercado (Stack Típica Nuvem)

R\$ 9.800 / mês

Trueoria (VPS Dedicada)

R\$ 350 / mês

Latência Média de Inferência por Requisição

7,1x mais Rápido

Mercado (VLMs em Nuvem)

15.000 ms (~15 s)

Trueoria (Física + XAI Local)

2.100 ms (~2,1 s)

Diretriz de Arquitetura: Onde a maioria das implementações recorre a APIs proprietárias de custo recorrente e comportamento probabilístico, construo pipelines determinísticos em código de máquina, reservando a IA generativa estritamente para as etapas onde ela gera diferencial de produto.

Trueoria Engine: Detecção de Fraudes e Edições em Imagens

Plataforma determinística de alta vazão para detecção em tempo real de adulteração fotográfica e documental em sinistros de seguros, cobrindo edições manuais e imagens geradas por IA.

BENCHMARK: Resposta em ~2,1 s com custo marginal fixo vs R\$ 0,15 a 0,80 por imagem em VLMs de terceiros.

ARQUITETURA DO PIPELINE DE INTEGRIDADE: BORDA LOCAL VS IA Desacoplamento Go + Python + Custódia Criptográfica

01. INGESTÃO & BORDA

Gateway Go Concorrente

- mTLS com CA privada
- Rate limiting com tarpitting
- Hashing SHA-256 imediato

Zero-Trust Edge

02. SINAL ÓPTICO & IA

Motor Heurístico & ML

- Ruído de sensor PRNU / ELA
- Detecção de GenAI e edições
- Coerência fotométrica local

Determinístico < 50ms

03. TRIAGEM DE CUSTO

Filtro de Invocação IA

- 82% resolvidos sem custos de API
- 18% demandam chamada VLM
- Explicação vetorial XAI

Economia de Tokens

04. RELATÓRIO & CUSTÓDIA

ISO/IEC 27037 Compliant

- Assinatura digital do lote
- Relatório técnico PDF/JSON
- Zero vazamento para WAN

Cadeia de Custódia

Decisões Técnicas & Trade-offs Críticos

Motor Próprio & ML vs VLM em Nuvem

Análise de física óptica, ruído de sensor (PRNU) e redes neurais locais para detecção de manipulação manual e imagens geradas por IA (Midjourney, DALL-E, etc.). Alternativa descartada: VLMs comerciais por custo imprevisível, falta de determinismo e risco de jailbreak.

Gateway Go + Worker Python

Go absorve concorrência massiva de I/O na borda; Python isola o pipeline matemático. Alternativa descartada: monólito Python com gargalo de GIL em I/O.

Borda Zero-Trust com mTLS

Validação criptográfica mTLS em nível de zona com CA privada. Conexões diretas ao IP morrem no handshake. Sessões em Redis com tarpitting progressivo.

Bilhetagem Atômica por Processo

Débito transacional único (10 fotos = 1 crédito). Severidade máxima consolida o relatório: uma imagem adulterada sinaliza o processo inteiro.

Benchmark Técnico Comparativo

DIMENSÃO TÉCNICA	VLMs COMERCIAIS EM NUVEM	TRUEORIA ENGINE (IMPLEMENTAÇÃO PRÓPRIA)
Custo por Inferência	R\$ 0,15 a 0,80 por imagem	Marginal próximo de zero (infra própria em VPS)
Latência Média	12 a 25 segundos	~2,1 segundos
Auditabilidade Legal	Caixa-preta probabilística	Sinal determinístico + XAI explicável
Proteção de Borda	Exposta a bypass de IP	mTLS zone-level com CA privada

Stack Técnica: Go · C++ / Python Nativo · OpenCV · Node.js / BuLLMQ · Redis · PostgreSQL / pgvector · Cloudflare · Docker Multi-stage

Memória L4 + Graph RAG via MCP

Camada de contexto contínuo que impede a amnésia de agentes de software entre sessões, com >99% de redução em custos de embeddings.

BENCHMARK: >99% de economia em embeddings (zero taxas de API), recuperação híbrida (vetor + grafo) em 22 a 35 ms, zero código enviado a terceiros.



Decisões Técnicas & Trade-offs Críticos

FastEmbed Local (ONNX/CPU)

Geração de embeddings locais na CPU em 8 a 14 ms por chunk. Custo recorrente zero e isolamento estrito de código confidencial do cliente.

Graph RAG em SQL Puro

CTE recursiva no PostgreSQL/pgvector com corte de ciclos. Elimina a necessidade de gerenciar instâncias dedicadas de Neo4j ou Pinecone.

MCP Stateless HTTP/SSE

Middleware ASGI resolvendo dessincronizações de sessão atrás de Cloudflare e proxies reversos. Autenticação por token protegida contra timing attacks.

Triade Contextual Estruturada

Obrigatório armazenamento em formato Gatilho → Causa Raiz → Solução, impedindo que logs crus poluam o índice vetorial.

Benchmark Técnico Comparativo

MÉTRICA DE AVALIAÇÃO	STACK TÍPICA EM NUVEM (BROKERS / APIS)	IMPLEMENTAÇÃO L4 / POSTGRESQL LOCAL
Custo de Embeddings	Cobrança contínua por token	>99% de economia / R\$ 0 em tokens (FastEmbed CPU)
Busca Vetorial	60 a 180 ms (API externa)	7 a 12 ms (pgvector HNSW local)
Consulta em Grafo	40 a 120 ms (Neo4j gerenciado)	2 a 5 ms (CTE SQL Recursivo)
Latência Total da Tool MCP	200 a 500 ms	22 a 35 ms
Complexidade de Infraestrutura	3 serviços gerenciados em nuvem	1 VPS containerizada + 1 PostgreSQL

Stack Técnica: Python / FastMCP · FastEmbed / ONNX · PostgreSQL / pgvector · Docker · Cloudflare

Motor de Inteligência B2B

Descoberta e validação de sócios e decisores reais na Receita Federal em 15 segundos, sem data brokers caros.

BENCHMARK: Redução de 40 min para ~15 s por lead com custo de dados < R\$ 25/mês vs R\$ 3.000 a 6.500 de plataformas de broker.



Decisões Técnicas & Trade-offs Críticos

Sourcing Híbrido Oficial

Combinação de Google Places, DDGS, RDAP Registro.br e Quadro de Sócios e Administradores (QSA) da Receita Federal. Decisor real com poder de assinatura.

Validador Léxico Anti-Alucinação

Similaridade léxica determinística antes de associar qualquer CNPJ ao lead. Código determinístico respondendo em menos de 2 ms onde LLMs alucinariam.

Pipeline em Duas Fases

Fase 01: radar leve de descoberta; Fase 02: enriquecimento profundo somente nos alvos aprovados. Economia de 70% em processamento computacional.

IA Exclusiva para Síntese

LLM atua unicamente no copywriting consultivo. Scraping, validação sintática e verificação DNS/MX são 100% código determinístico.

Benchmark Técnico Comparativo

MÉTRICA DE AVALIAÇÃO	PROSPECÇÃO CONVENCIONAL / BROKERS	MOTOR DE INTELIGÊNCIA PROPRIETÁRIO
Custo de Dados por Mês	R\$ 3.000 a R\$ 6.500 / mês	< R\$ 25 / mês
Tempo de Pesquisa por Lead	25 a 40 minutos (manual)	~15 segundos
Bounce Rate de E-mails	12% a 18%	< 2,5% (sem risco de spam em ping SMTP)
Acurácia de Decisores	~45% (redes sociais defasadas)	> 88% (QSA oficial + RDAP)

Stack Técnica: Python / FastAPI • React 19 / TypeScript • SQLite WAL • RDAP • DNS / MX Validator • Gemini Flash

Modelo de Contratação & Padrão de Entrega

Desenvolvimento autônomo para squads de tecnologia e soluções sob medida para operações.

