DAVID MOLIZANE

Engenheiro de Software & IA

david.molizane@icloud.com | github.com/molizanee | linkedin.com/in/david-molizane

RESUMO PROFISSIONAL

Engenheiro de Software & IA com **+2 anos** de experiência na construção de aplicações full-stack e com IA. Experiência comprovada na arquitetura de sistemas de produção **RAG (Retrieval-Augmented Generation)** com **bancos de dados vetoriais (Qdrant, pgvector)**, **integração de LLMs (Langchain, Pydantic AI)**, **ingestão de documentos, chunking, geração de embeddings e recuperação semântica**. Arquitetou **fluxos de trabalho multi-agente** com gerenciamento de estado via **checkpoints do LangGraph** e roteamento determinístico através de **camadas de validação semântica**, alcançando um **aumento de 40% na conversão** e **redução de 60% no suporte**. Proficiência na construção de sistemas de backend com **Python (FastAPI, SQLAlchemy)**, **Node.js & Bun.js (Express.js, Next.js, React.js, Vitest)**, **AWS (S3, EC2, Bedrock)**, **GCP (Bucket, Vertex AI, Cloud Run)**, e **contêineres Docker** e **pipelines de CI/CD (GitHub Actions)**. Arquitetou soluções que processam grandes quantidades de dados, aumentaram a produtividade da empresa e reduziram custos. Participou da primeira Equipe de IA no **Grupo Guanabara** (o maior conglomerado de transporte de ônibus do Brasil, com mais de 20.000 funcionários), liderando com sucesso a implementação de uma cultura de IA saudável, reduzindo os custos de IA em **30%** com a otimização do proxy LiteLLM e eliminando **+1 dia de trabalho manual** através da automação com IA.

HABILIDADES TÉCNICAS

IA & Dados: Sistemas RAG (melhora de 40% na precisão de recuperação), Bancos de Dados Vetoriais (Qdrant, pgvector), Integração de LLM (Langchain, Pydantic AI), Embeddings, Chunking de Documentos, Busca Semântica, Engenharia de Prompt, Dados Sintéticos, LiteLLM (redução de 30% nos custos), LangFuse

Linguagens & Frameworks: Python, FastAPI, SQLAlchemy, JavaScript, TypeScript, Node.js, Bun.js, Express.js, GoLang, Go-chi, React, Next.js, Prisma ORM, Drizzle ORM

Dados & Infraestrutura: PostgreSQL, MongoDB, Redis, Microsoft SQL Server, MySQL, APIs REST, WebSockets

Cloud & DevOps: AWS (EC2, S3, Route 53, Bedrock), GCP (VM, Cloud Run, Bucket, Vertex AI), Docker, Kubernetes, Terraform, CI/CD (GitHub Actions), Observabilidade (Grafana)

Testes & Práticas: Jest, Vitest, React Testing Library, Agile/Scrum, Microsserviços

Segurança & Guardrails de IA: Moderação de conteúdo (AWS Bedrock Guardrail), Redação de PII (Microsoft Presidio), Design de resposta determinística, Mitigação de alucinações

EXPERIÊNCIA PROFISSIONAL

Molizane SYS (Autônomo) *Fevereiro 2025 - Presente*

Engenheiro de IA

- Desenvolvendo o **FestaPro AI Agent**, um sistema de agentes construído em **Python com FastAPI**. Apresenta um **Agente React para a API do WhatsApp Business** que retorna informações sobre locais de eventos e festas, responde a perguntas contratuais, agenda visitas a locais, envia fotos de locais e fornece links sociais. Usa **Tortoise ORM** para gerenciamento de esquema, **LangChain e LangGraph** para estado de conversação e roteamento de fluxo, com checkpoints salvos em **PostgreSQL**. Suporta uploads de arquivos (PDF, TXT, DOCX) para **QdrantDB** para contexto semântico.

Grupo Guanabara *Setembro 2025 - Presente*

Engenheiro de IA

- Construiu o **Commercial AI**, um Agente de IA Text-to-SQL que recebe perguntas de negócios e retorna respostas usando dados reais da empresa. Arquitetado com **sistema multi-workflow e multi-agente n8n**, **QdrantDB** para validação de busca semântica, **LiteLLM Proxy** para consumo de modelo, **camada semântica CubeJS** para geração de payload SQL com acesso ao data lake, e **Langfuse** para traces e evals. **Impacto:** Executivos obtêm insights de negócios rápidos sem depender da equipe de BI, reduzindo

drasticamente o tempo de resposta para perguntas de negócios. Baixo custo alcançado através da seleção estratégica de LLM: roteamento específico de tarefas entre Claude Sonnet 4.5, GPT-5, Nvidia Nemotron, Llama 4/3.1 e Qwen 3 com base em benchmarking de custo-qualidade.

- Projetou o **Guanabara Chat**, uma pilha de IA completa para toda a empresa integrando **Open WebUI** para interface de funcionário, **LiteLLM Proxy** para gerenciamento de modelo e custo por equipe, **AWS Bedrock** e **Google Vertex AI** para modelos, **Langfuse** para traces e feedback, **Microsoft Presidio** para mascaramento de dados PII, e **AWS Bedrock Guardrail** para moderação de conteúdo. **Impacto:** Acesso seguro à IA em toda a empresa, mantendo dados sensíveis na infraestrutura local.
- Implantou o **PCO**, um assistente de IA da **API do WhatsApp Business Meta** que ajuda os motoristas de ônibus com incidentes e suporte. Usa **QdrantDB** para embeddings de incidentes, garantindo respostas determinísticas e minimizando alucinações em situações críticas, **Redis DB** para gerenciamento de histórico de conversas e **Langfuse** para traces e evals. **Impacto:** Motoristas ganham eficiência e precisão sobre como proceder em situações críticas.

Grupo Guanabara *Fevereiro 2025 - Agosto 2025*

Desenvolvedor de Automação

- Construiu o **Punctuality**, um pipeline de dados de alerta de atraso de veículos em tempo real. Consome de um **WebSocket** que traz dados de GPS (serviço construído em **Golang**), processa dados em **n8n** com arquitetura assíncrona, executa consultas no **AWS Athena** e realiza cálculos de geolocalização e tempo para alertas de atraso em tempo real. **Processa mais de 864.000 pontos de dados de veículos por dia** com baixo custo e alta eficiência. **Impacto:** Visão estratégica de negócios com geração massiva de dados relevantes (rotas, veículos e horários com mais atrasos), abrindo possibilidades para ML prever atrasos.
- Criou o **Orders Track**, uma aplicação full-stack (frontend **React.js/Next.js**, backend **Express.js/Bun.js**) que fornece um painel interativo em tempo real para pedidos de peças de veículos atrasadas com alertas automáticos por e-mail. Usa **autenticação Supabase** para gerenciamento de usuários e **PostgreSQL** para armazenamento de dados. **Impacto:** Equipe de logística **40% mais eficiente** devido à análise unificada otimizada das responsabilidades de pedidos atrasados.

Bravion *Janeiro 2024 - Fevereiro 2025*

Engenheiro Full-Stack Júnior

- Projetou e construiu o microsserviço principal para o **GenX**, uma plataforma de genética clínica que processa grandes volumes de dados genéticos de pacientes e gera **relatórios em PDF em menos de um minuto**. Stack: **Node.js, TypeScript, Prisma ORM, PostgreSQL, react-pdf, Google Cloud Platform**.
- Desenvolveu funcionalidades essenciais para o **Cmorq**, uma plataforma de criptomoedas com IA e **conformidade com a LGPD**, construindo uma aplicação **Next.js** segura com **APIs WalletConnect** para proteção de dados financeiros do usuário.

Bravion *Setembro 2023 - Dezembro 2023*

Estagiário de Engenharia Full-Stack

- Projetou um processo robusto de migração de dados da plataforma legada para **PostgreSQL**, incluindo um **web crawler baseado em Python** capaz de se adaptar a múltiplos layouts de páginas legadas.
- Construiu e manteve a aplicação full-stack **Motorbooks** com backend **FastAPI** e frontend **React.js/Next.js**, entregando uma interface moderna que melhorou a usabilidade e a manutenibilidade.

EDUCAÇÃO & IDIOMAS

Bacharelado em Engenharia de Computação | Fundação Herminio Ometto | 2024 - 2030 (Previsto, cursando o 3º ano)

Idiomas: Inglês (Fluente - Proficiência Profissional), Português (Nativo)