LUCAS DE SOUZA SANTOS

AI Engineer | Sistemas Multiagente · RAG · Document AI

Curitiba, PR · (41) 99668-6405 · lss.ctba@gmail.com · linkedin.com/in/lucas-de-souza-santos

RESUMO PROFISSIONAL

AI Engineer com trajetória de automação de processos desde 2019, evoluindo da modernização de sistemas legados para IA generativa em produção. Construção de agentes em GPT-3.5 com tool calling manual já no início de 2023, antes do suporte nativo de function calling. Atuação atual em orquestração de agentes, RAG e Document AI: arquiteturas multiagente, extração documental com saída validada por schema, prompt engineering anti-alucinação e observabilidade ponta a ponta. Impacto mensurável: redução de mais de 50% no custo com LLMs em produção via gateway de inferência.

COMPETÊNCIAS TÉCNICAS

Sistemas Multiagente & Orquestração: Semantic Kernel, LangGraph, LangChain; planning, memory, tool calling; separação arquitetural entre extração e decisão

LLMs & Inferência: modelos comerciais (GPT-4/5, Gemini) e open-weights locais (Qwen3, DeepSeek-R1) via Ollama / vLLM; roteamento por complexidade com primário + fallback; saída estruturada validada por Pydantic; prompt engineering anti-alucinação

RAG & Busca Vetorial: recuperação vetorial, embeddings, reranking, chunking ciente de página (OpenSearch, PGVector, Azure AI Search)

Document AI: OCR, extração de layout, processamento em lotes (Azure AI Content Understanding, Docling, pypdf, python-docx)

Avaliação (Evals): acurácia e cobertura contra gabarito, avaliação contínua em produção (Azure AI Evaluation, Azure AI Foundry)

Backend & APIs: Python (FastAPI, asyncio), C# (.NET, ASP.NET Core), Go, Node.js; PostgreSQL, SQL Server, REST

Engenharia de Dados: Kafka, Kafka Connect, Debezium (CDC), arquitetura event-driven, PySpark, OpenSearch / Elasticsearch

Cloud & Observabilidade: OpenTelemetry, telemetria ponta a ponta; Azure (Identity, Blob, Cosmos DB, Monitor / App Insights), AWS (EC2, S3, IAM)

DevOps: Docker, CI/CD (GitHub Actions, Azure DevOps), Terraform, pytest / pytest-asyncio, Git, Linux

Metodologias: Scrum, Kanban, Code Review, TDD, Clean Architecture, Event-Driven Architecture

Idiomas: Português (Nativo) · Inglês (Intermediário, Leitura Técnica Avançada)

Domínio: Govtech / legaltech (Lei 14.133/2021), COMEX / DUIMP, avaliação de redações (ENEM)

EXPERIÊNCIA PROFISSIONAL

Prover | AI Engineer

Desde Abr 2026

Engenharia de IA generativa para a BigBrain, em projetos de avaliação documental e textual para os setores público e educacional.

- Pipeline multiagente (Semantic Kernel) para análise de habilitação em licitações (Lei 14.133/2021), combinando parsing determinístico e extração por LLM com saída validada por schema.
- Sistema multiagente de avaliação de redações (modelo ENEM) para o programa Redação Paulista (Seduc-SP): análise de tangência ao tema, proposta de intervenção e elementos textuais, com geração de feedback pedagógico, em produção em larga escala na rede estadual.
- Camada de Document AI com chunking ciente de página e processamento em lotes para documentos extensos (Azure AI Content Understanding).
- Camada de confiabilidade nas chamadas de LLM: modelo primário + fallback e concorrência controlada.

RESULTADOS OBTIDOS

- Substituição de revisão manual que consumia meio período de um analista por edital.
- Decisão auditável e sem alucinação de veredito, com separação estrita entre extração e decisão (nenhum LLM define o resultado final) e validação antes da persistência.
- Trilha de auditoria ponta a ponta, com rastreabilidade do dado extraído até a página de origem (OpenTelemetry, Azure Monitor).

CCM do Brasil | AI Engineer

Abr 2025 a Abr 2026

Arquiteto e referência técnica solo da plataforma de IA corporativa.

- Plataforma GenAI (AYA Search): RAG end-to-end com orquestração de LLMs, plugins e recuperação semântica; ingestão de documentos (Docling), chunking, reranking e prompt engineering avançado.
- Gateway de inferência agnóstico: roteamento entre LLMs comerciais (GPT-4) e locais (Qwen3, DeepSeek-R1) por complexidade.
- Orquestração de agentes autônomos (plugins, planning, memory) com execução de tarefas em sistemas corporativos e gestão de contexto e estado.
- Guardrails role-based e plataforma de dados em tempo real: arquitetura event-driven com Debezium (CDC), Kafka Connect e microsserviços agregadores em Go; modelagem RAW → FCT.

RESULTADOS OBTIDOS

- Redução de mais de 50% no custo com APIs comerciais de LLM.
- Liberação de 5 a 10 especialistas de trabalho operacional repetitivo.
- AYA Search adotada por 10 a 50 usuários ativos, com evals em produção sobre uso real.
- Deploy de LLMs locais em produção (infra híbrida on-premise + cloud GPU com vLLM).

OVD Importadora | Analista de Sustentação e Desenvolvedor, TI

Mai 2023 a Mar 2025

- Integração COMEX com LLMs para descrições técnicas em conformidade DUIMP (Receita Federal).
- Reconciliação (batimento) de preços entre sistema legado COBOL e ambiente de produção em Elasticsearch (B2B), com detecção de divergências.
- Modernização de sistemas legados (COBOL para Elasticsearch + Python), pipelines ETL multi-fonte e sustentação de e-commerce B2B.

RESULTADOS OBTIDOS

- 100% de conformidade nas descrições técnicas DUIMP.
- Divergências de preço entre legado e produção detectadas e direcionadas à correção de causa raiz em microsserviço Kafka.

OVD Importadora | Analista de Dados e Desenvolvedor, Eng. de Produto

Abr 2019 a Abr 2023

- Agente autônomo de e-mail (abr/2023): fluxo agêntico completo em GPT-3.5 com tool calling manual via parsing de intenção e roteamento por JSON estruturado, anterior ao suporte nativo de function calling da API.
- Pipeline de análise de dados do setor de engenharia, integrando sistema legado (COBOL) e Oracle EBS para apoio à decisão.
- Sistema de logística reversa com direcionamento autônomo, integrado a Telecontrol e Service Telecontrol (assistência técnica).
- Suíte de automação para engenharia de produto: geração de certificados de qualidade, orçamentos com consulta de preços e documentação de pós-venda, além de classificador fiscal de NCM integrado ao SISCOMEX.

RESULTADOS OBTIDOS

- Agente de e-mail operando 100% sem intervenção humana.
- Redução de tempo e ganho de acurácia em decisões a partir da pipeline integrada (COBOL + Oracle EBS).
- Reduções de 80% a 90% no trabalho manual nos processos automatizados.

FORMAÇÃO ACADÊMICA

Pós-graduação: Arquitetura de Software · Universidade Tuiuti do Paraná · 2026

Bacharelado: Engenharia Elétrica · Universidade Tuiuti do Paraná · 2024