

Matheus Levi da Silva Barbosa

Engenheiro de IA/ML | 4+ anos em IA | Infraestrutura de LLMs, Avaliação de Modelos & Data Engineering

mathlevi20.github.io

linkedin.com/in/matheus-levi-621125150/

github.com/MathLevi20

matheuslevi2012@gmail.com

+55 86 99904-1794

RESUMO

Engenheiro de Machine Learning com 5+ anos em desenvolvimento de software e 4+ anos focados em IA, incluindo LLMs, IA Generativa e engenharia de dados. Atua na construção de datasets e pipelines de treinamento/avaliação de modelos em português, com base sólida em backend (Python/FastAPI), full stack e processamento de dados em escala, combinando atuação CLT e consultoria PJ em projetos de P&D.

EXPERIÊNCIA DE TRABALHO

Engenheiro de Machine Learning Pleno, SoberanIA — Novembro/2025 – Atual • São Paulo, Brasil • Remoto

Projeto de IA soberana brasileira, com foco em LLMs, engenharia de dados e infraestrutura de inferência.

- Projetei e implementei pipelines de agentes de IA para curadoria e geração de dados, com observabilidade e mecanismos de contingência para garantir a continuidade do sistema.
- Desenvolvi plataforma de avaliação e segurança de modelos de linguagem, incluindo documentação técnica e dashboard de resultados.
- Criei um leaderboard público de avaliação colaborativa de modelos de linguagem em português, com autenticação institucional.
- Construí sistemas de ingestão de dados multimodais (texto, áudio, imagem) e pipelines de coleta de dados em larga escala.
- Implementei observabilidade de infraestrutura e pipelines de CI/CD para validação e deploy contínuo dos serviços.

Engenheiro de Software (PJ / Consultor), AWTech — Setembro/2024 – Maio/2026 • Piauí, Brasil • Híbrido

Consultor PJ em projetos de P&D em IA, gestão pública e processamento de dados, alguns sob NDA, incluindo análise documental, o sistema AWControle (prefeituras) e o SIRVA (reconstrução 3D em parceria com a SSP) — em paralelo a outras atuações CLT do período.

- Projetei soluções de análise documental com LLMs, RAG e LangChain para textos médicos, jurídicos, técnicos e administrativos, com backend em Python, FastAPI e PostgreSQL, usando Docker e Docker Compose para deploy.
- Desenvolvi o AWControle, sistema de controle interno e organização administrativa para prefeituras, com backend em Java e frontend em Angular (TypeScript, HTML, CSS, Tailwind CSS).
- Participei do desenvolvimento do SIRVA, projeto em parceria com a SSP (Secretaria de Segurança Pública), implementando reconstrução de malhas 3D a partir de imagens e nuvens de pontos (LiDAR) com técnicas de Structure from Motion e reconstrução de superfície (Poisson).
- Estruturei os pipelines de persistência de dados do SIRVA (FastAPI, SQLAlchemy, Alembic, PostgreSQL), as visualizações 3D interativas em Three.js e as aplicações mobile em Next.js, Flutter e Swift.
- Automatizei testes com Pytest e pipelines de CI/CD (GitHub Actions) para validação contínua e maior confiabilidade.

Engenheiro de Software Full Stack, Alva — Fevereiro/2024 – Abril/2025 • São Paulo, Brasil • Remoto

Empresa de Ciência de Dados e desenvolvimento Full Stack, com foco em aplicações Django.

- Atuei de ponta a ponta no desenvolvimento Full Stack, da modelagem do banco de dados à entrega de frontend e backend.
- Construí aplicações com Django, HTML, CSS, JavaScript e Bootstrap, conduzindo a migração do stack para Next.js, Tailwind CSS e Node.js, com pipelines de CI/CD (GitHub Actions) para deploy contínuo.
- Modelei análises de dados com Python, Pandas, NumPy e Scikit-learn, transformando dados brutos em visualizações para decisões de negócio.
- Desenvolvi pipeline mensal de reconciliação fiscal cruzando mais de 172 mil XMLs de NFe contra registros SPED (ICMS), usando distância de Levenshtein para casar divergências e recalculando a base de ICMS-ST, processando cerca de 29.741 notas fiscais e 48.047 registros de apuração por mês.
- Apliquei IA Generativa para acelerar a geração de insights e conteúdos visuais a partir de grandes volumes de dados.

Desenvolvedor Mobile (Meio Período), IFCE — Julho/2023 – Dezembro/2023 • Remoto

Atuação em meio período no desenvolvimento de aplicações mobile para o setor de Inovação e Tecnologia, em paralelo à graduação.

- Desenvolvi, do zero, um app iOS de adoção de animais com Swift, SwiftUI e SwiftKit, cobrindo toda a jornada do usuário.
- Integrei o app a um banco de dados NoSQL (Firebase) e otimizei a performance para reduzir o tempo de carregamento.
- Projetei interfaces intuitivas, responsivas e funcionais para o público-alvo do app.

Desenvolvedor Python Back-end, Comunicare Solutions LTDA — Janeiro/2022 – Fevereiro/2024 • Piauí, Brasil

Empresa de IA para tratamento fonoaudiológico, com foco em processamento e análise de dados de áudio.

- Projetei APIs de ASR, migrando de Flask para FastAPI e integrando Whisper e Wav2Vec para Speech-to-Text, com foco em precisão e desempenho.
- Implementei o NVIDIA Triton Inference Server, processando até 200 áudios simultâneos por segundo em uma única GPU T4, com dynamic batching e autoscaling.
- Implementei observabilidade com Prometheus e Grafana, monitorando latência, recursos e disponibilidade em produção.

- Conduzi P&D em Knowledge Distillation, reduzindo a complexidade computacional dos modelos sem perder precisão em ambientes com recursos limitados.
- Gerenciei deployment e infraestrutura em AWS (EC2, S3, EBS) e GCP, com CI/CD garantindo disponibilidade e escalabilidade.
- Implementei testes unitários (Pytest) e de carga (k6), validando robustez e baixa latência em produção.
- Desenvolvi API de captura e análise estatística de áudio, com relatórios automatizados via Pandas, NumPy e Matplotlib.

PROJETOS PARALELOS

Equitas: Plataforma de automação com web scraping e análise financeira, aplicando scripts e métodos acadêmicos desenvolvidos pela Universidade Federal do Piauí (UFPI) para avaliação financeira e cálculos específicos.

EDUCAÇÃO E CERTIFICADOS

- 2021 - 2025 Graduação em Análise e Desenvolvimento de Sistemas, Instituto Federal do Piauí.
- 2024 DeepLearning.AI TensorFlow Developer Specialization, DeepLearning.AI.
- 2024 Machine Learning with Python - Carga horária 20h, Coursera.
- 2024 Databases and SQL for Data Science with Python - Carga horária 20h, Coursera.
- 2023 Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning - Carga horária 18h, Coursera.
- 2023 Version Control - Carga horária 13h, Coursera.
- 2023 Programming in Python - Carga horária 44h, Coursera.
- 2023 Advanced React - Carga horária 26h, Coursera.
- 2022 React Basics - Carga horária 26h, Coursera.
- 2022 Linux II - Carga horária 4h, Alura.
- 2022 Linux I - Carga horária 8h, Alura.
- 2022 Shell Scripting II - Carga horária 8h, Alura.
- 2022 Shell Scripting I - Carga horária 8h, Alura.

HABILIDADES TÉCNICAS

ML & Dados: LLMs, IA Generativa, RAG, LangChain, LangGraph, vLLM, Langfuse, Inspect AI, TensorFlow, PyTorch, Scikit-learn, Keras, NumPy, Pandas, Matplotlib, análise estatística, Knowledge Distillation, ASR, Whisper, Faster-Whisper, Docling, Wav2Vec, NVIDIA Triton, pgvector, engenharia e curadoria de dados

Linguagens: Python, JavaScript, TypeScript, Java, Swift

Backend: FastAPI, Flask, Django, NestJS, APIs REST, PostgreSQL, SQLAlchemy, Alembic, ClickHouse, Prisma, MongoDB, Keycloak

Frontend: React, Next.js, Angular, Vite, Tailwind CSS, HTML, CSS, Firebase

Mobile: Swift, SwiftUI, SwiftKit, Flutter, desenvolvimento iOS

Cloud & DevOps: AWS (EC2, S3, EBS, Lambda, SageMaker), GCP, Docker, Docker Compose, CI/CD (GitHub Actions), Pytest, k6, Prometheus, Grafana, Metabase, Git

Web Scraping & Automação: Scrapy, Playwright, Selenium, gestão de proxies (Tor, proxies residenciais), bypass de proteções anti-bot

Outros: Three.js, LiDAR e processamento de nuvens de pontos, MinIO, Redis, Scrum

IDIOMAS

Inglês — Intermediário (B2) | Português — Nativo