Wallace Souza

Senior AI/Machine Learning Engineer

wallacesouza1014@gmail.com +55 91 98540-7361 Tatajuba, Capitão Poço - PA, Brazil

Wallace Souza

Profile

Senior AI Engineer with 8 years of experience leading the development and deployment of scalable machine learning and generative AI solutions. Expertise in fine-tuning large language models (LLMs), prompt engineering, and cloud-native AI systems on GCP, AWS, and Azure. Skilled in MLOps, API architecture, vector search, and implementing ethical, privacy-focused AI practices.
I am passionate about learning cutting-edge technologies and delivering impactful AI products across various industries.

Professional Experience

Senior AI Engineer 11/2022 – 05/2025
AI21 Labs (Remote) New York, USA

- Led the development and deployment of large language models (LLMs) for NLP and RAG applications, focusing on building next-gen AI solutions that push the boundaries of conversational AI.
- Designed and optimized LLM architectures, improving text generation accuracy by 20% and reducing inference times by 25% using cutting-edge techniques like model distillation and pruning.
- Collaborated with cross-functional teams to integrate AI solutions into commercial products, enhancing user engagement by 30% through intelligent, context-aware interactions.
- Pioneered AI-driven content generation tools, increasing content production efficiency for clients by 30% while maintaining high-quality output.
- Integrated AI workloads into GCP, reducing cloud costs by 15% through optimized resource management and automation.
- Implemented an automated model retraining pipeline with MLOps, reducing the time for model updates by 50% and increasing system reliability.

Machine Learning Developer 01/2020 – 10/2022
Reonomy (Remote) New York, USA

- Developed a predictive property value model, reducing forecasting errors by 18%, improving sales and marketing strategies.
- Built a natural language processing model for real estate market trend analysis, enhancing decision-making by 30%.
- Streamlined machine learning workflows, cutting model training time by 40%.
- Led integration of AI models with backend systems, reducing query latency by 20%.
- Created a custom model monitoring system, improving model performance tracking and reducing issues by 15%.

Data Scientist Intern 09/2018 – 01/2020
Stitch Fix San Francisco, USA

- Contributed to a recommendation system optimization, enhancing personalized product recommendations and improving customer conversion rates by 12%.
- Worked on improving predictive models for demand forecasting, leading to a 10% improvement in inventory management efficiency.
- Developed a demand forecasting model that improved inventory management efficiency by 10%.
- Contributed to customer segmentation using unsupervised learning, aiding targeted marketing efforts.

Education

Bachelor's Degree in Computer Science 04/2016 – 09/2018
University of San Francisco (USF) San Francisco, USA

Skills

Programming Languages
Python, C++, Java, Javascript, Rust

AI Agents
SmolAgents, LlamaIndex, LangGraph, n8n

Cloud Platforms
AWS, Google Cloud Platform(Vertex, BigQuery), Microsoft Azure

APIs & Integrations
REST, GraphQL, OAuth2, JWT, LLM API integration(OpenAI, Cohere)

Leadership & Collaboration
Team leadership, mentoring, cross-functional teamwork, communication skills

Machine Learning Frameworks
TensorFlow, PyTorch, Hugging Face, Transformers, Scikit-learn

Generative AI & LLM
GPT-4, LLaMA2, Claude 2, BERT, generative models(GANs, diffusion), Langchain

DevOps & MLOps
Terraform, Pulumi, MLflow, Docker, Kubernetes, CI/CD

Databases
PostgreSQL, MongoDB, VectorDB(Pinecone, Chroma)

Edge AI
Tensorflow Lite, ONNX, model optimization