

Laxman Kumar Busetty

San Francisco, CA | laxman.busetty@gmail.com | +1 909-499-1640 | [linkedin.com/in/laxmanbusetty](https://www.linkedin.com/in/laxmanbusetty) | github.com/LaxmanBusetty

SUMMARY

AI/ML Engineer with experience building production-grade AI agents, LLM systems, and scalable ML pipelines for customer-facing and enterprise use cases. Skilled in Python, FastAPI, LangChain, LlamaIndex, RAG, hybrid retrieval, LLM evaluation, safety guardrails, and cloud deployment. Hands-on experience designing, deploying, and improving reliable AI systems across document intelligence, conversational AI, and voice workflows, with a strong builder mindset and ability to ship in fast-moving environments.

EXPERIENCE

Machine Learning Engineer Intern | Vosyn Inc | United States (Remote) | Feb 2026 – Present

- Shipped a functional PoC for an Experience Engine in under one week, leading to investor demos and technical workshops
- Deployed the PoC to production via Vercel with continuous deployment from GitHub, enabling the co-founder to use it in investor demos and technical workshops supporting VosynVerse's funding and go-to-market efforts
- Cleaned and deduplicated a 6.69M-record multilingual translation dataset across 9 language pairs, removing 178,601 duplicate records (~2.7%) using a Pandas-based pipeline with zero corrupt records
- Fine-tuned a large language model for multilingual translation using LoRA (Low-Rank Adaptation), a parameter-efficient fine-tuning technique, combined with 4-bit quantization reducing GPU memory usage by 65-75% while maintaining BLEU scores of 15–25
- Built end-to-end supervised fine-tuning (SFT) pipeline on OPUS100 (1M+ sentence pairs) using Hugging Face and PEFT; improved training efficiency 3-5x via gradient checkpointing, bfloat16, and gradient accumulation

AI Engineer Intern (Part-Time) | Tilli | Palo Alto, CA (Remote) | Oct 2025 – Feb 2026

- Architected and built a production RAG coaching agent (Ask Tilli) for educators teaching social-emotional learning (SEL) emotional intelligence, self-awareness, and related skills to children under 10 years
- Reduced hallucination rate ~30% and response latency ~40% by grounding responses in curated, research-backed sources (UNICEF and CASEL) and architecting multi-LLM routing for optimized performance
- Architected and maintained a production knowledge base powering a RAG pipeline using curated UNICEF and CASEL content.
- Evaluated and monitored response quality using the RAGAS framework, tracking faithfulness, relevance, and context precision to reduce hallucinations
- Deployed a multi-layer moderation and guardrail framework combining LLM-based filtering with rule-based validation, improving agent safety for use in classroom and child-education contexts
- Containerized with Docker and deployed on GCP across web and WhatsApp supporting accessibility for educators in low-bandwidth or mobile-first regions

AI-Powered Document Insights Extern | Outamation (via Extern) | United States (Remote) | Oct 2025 – Dec 2025

- Built modular AI-powered pipelines processing 20+ page mortgage documents, combining OCR (Tesseract, PaddleOCR), PDF parsing (PyMuPDF), and RAG techniques for intelligent data extraction, classification, and search reducing manual review time by ~90%
- Developed a document retrieval system using LlamaIndex and RAG optimized for multi-document mortgage files, enhancing precision through chunk tuning, metadata filtering, and evaluation of open-source LLMs including Mistral and Phi-2
- Conducted end-to-end evaluation of the document intelligence system across 20+ page documents, benchmarking OCR accuracy, retrieval quality, and routing performance
- Delivered a technical report outlining model trade-offs, optimization strategies, and deployment recommendations; built a UI for stakeholder demo purposes

AI Research Assistant (Part time) Adversarial Machine Learning | CSUSB | San Bernardino, CA | Oct 2024 – Dec 2025

- Conducted AI research under **Prof. Nabeel Alzahrani** (Adversarial Machine Learning) and **Prof. Alireza Abdoli** (Computer Vision), first-authoring two comprehensive survey papers.
 - Published a survey accepted at **AIRC 2025**, analyzing adversarial attacks and defense mechanisms across computer vision, NLP, GNNs, and cybersecurity by reviewing **50+ papers** and synthesizing **20 high-impact studies**.
 - Reviewed **13 state-of-the-art object detection architectures** (including YOLOv7, Mask2Former, CSWin Transformer, and DETReg) across COCO, ImageNet, and nuScenes, identifying accuracy–efficiency trade-offs for scalable real-time vision systems.
 - Research supported by the **xREAL and OSR Summer 2025 Grant**, with findings presented to faculty and research lab members.
-

PROJECTS

FounderOS - AI-Powered Chief of Staff - *Next.js, TypeScript, InsForge, Multi-Agent Orchestration*

<https://i8u8yy2b.insforge.site>

- Solo-built a multi-agent system at the Agentic Dev Tools Hackathon (Entrepreneur First, San Francisco), orchestrating 4 specialized AI agents Planning, Market Intelligence, Opportunity Discovery, and Executive Briefing to take a founder from a single goal to a full execution roadmap
- Generated complete, structured roadmaps in under 10–15 seconds end-to-end; validated across 3+ real startup use cases during the hackathon
- Designed structured output validation between each agent layer, preventing silent error propagation and making the workflow trustworthy for non-technical stakeholders
- Received positive feedback from judges and fellow participants on the system's reliability and practical execution focus, despite building solo against team-based competitors

Wanderlust AI - AI Travel Planning Application - *Vertex AI (Gemini), GCP, Supabase, PostgreSQL*

<https://wanderlust-ai-330205915151.us-west1.run.app/>

- Built and deployed full-stack AI travel planner on GCP using Google Vertex AI (Gemini) with Supabase for user auth, database management, and persistent state
- End-to-end API pipeline from user input to personalized AI-generated itineraries with real-time accommodation and car rental recommendations; live in production

EDUCATION

Master of Science, Computer Science | California State University, San Bernardino | *San Bernardino, CA* | Aug 2023 – Dec 2025

Relevant Coursework: Artificial Intelligence, Machine Learning, Software Engineering, Theory of Algorithms

Bachelor of Technology, Electronics & Communication Engineering | Avanthi Institute of Engineering and Technology (AVIH), affiliated to JNTU Hyderabad, *India* | 2017 – 2021

- Fully funded by state government scholarship - awarded on academic merit

TECHNICAL SKILLS

AI & Machine Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, Hugging Face Transformers, PEFT, LoRA, Quantization, NLP, Computer Vision

LLM & Agentic AI: AI Agents, Multi-Agent Systems, LangChain, LlamaIndex, Retrieval-Augmented Generation (RAG), Hybrid Retrieval (BM25 + Vector Search), Prompt Engineering, Tool Calling, Context Engineering, LLM Evaluation (RAGAS), Safety Guardrails

Backend & APIs: FastAPI, REST APIs, Docker

Cloud & Deployment: Google Cloud Platform (GCP), Vertex AI, AWS, Vercel, CI/CD

Databases & Data: PostgreSQL, Pgvector, Qdrant, Vector Embeddings, Semantic Search, Pandas, NumPy

Developer Tools: Git, GitHub, Streamlit, Jupyter Notebook

LEADERSHIP & COMMUNITY

Vice President, Data Analytics Club, CSUSB | San Bernardino, CA | Feb 2025 – Aug 2025

- Served as Vice President for a 125 member club, co-organizing 3 workshops and speaker events to make data analytics approachable for students
- Mentored active members in Python, SQL, visualization, and machine learning through semester-long project work, culminating in members presenting their findings to the club
- Collaborated with the executive team and faculty to create networking and growth opportunities; helped shape club vision around teamwork, curiosity, and leadership