AKHIL CHINTALAPATI

AI Software Engineer

Durham, NC | +1 (445) 256-3786 | akhil.chintalapati01@gmail.com | linkedin.com/in/akhil-c-/ | github.com/AkhilByteWrangler

EDUCATION

Duke University ***Aug 2024 - May 2026***

Master's in Artificial Intelligence, 3.97/4.00

Teaching Assistant - Intelligent Agents (Spring 2026)

VIT University, India ***Aug 2020 - May 2024***

Bachelor's in Computer Science (Data Science), 3.70/4.00

TECHNICAL SKILLS

- **Programming:** Python, C++ (Performance-Critical), Rust, TypeScript, SQL (Query Optimization)
- **AI/ML:** PyTorch, JAX, Fine-tuning (LoRA, RLHF, DPO), Distributed Training (DeepSpeed, FSDP), Custom Embedding Models, Systematic Prompt Engineering
- **GenAI & Retrieval:** Vector Search (FAISS, HNSW), Pinecone, Weaviate, Hybrid Search, Cross-Encoder Re-ranking, Context Compression, Latency-Optimized Streaming
- **On-Device AI:** Core ML, MLX, ONNX Runtime, INT4/INT8 Quantization, Knowledge Distillation, Mobile Inference (Sub-100ms)
- **Backend:** FastAPI, Django, Node.js, gRPC, GraphQL, Redis (Caching & Rate Limiting), PostgreSQL (Optimized Queries)
- **Infrastructure:** Docker, Kubernetes (Auto-scaling), AWS (SageMaker, Bedrock, Lambda, S3), Terraform, CI/CD (Blue-Green Deployments), Airflow
- **Frontend & Mobile:** React, Next.js (SSR/SSG), Swift/SwiftUI, iOS SDK, WebSockets, Tailwind CSS

WORK EXPERIENCE

Tesla Motors - Data Algorithmic Engineer Intern, GenAI & Supply Chain Analytics, *Fremont, CA* ***May 2025 - Dec 2025***

- Engineered self-improving agentic AI system with feedback loops that autonomously refines predictive models and generates interactive insights, enabling natural language-driven supply chain analytics across highly unstructured data sources.
- Built custom agent tools with safety guardrails to query and manipulate 1M+ database tables at scale, implementing secure execution, validation layers, and optimization strategies to handle production data complexity while preventing corruption.
- Developed full-stack analytics platform with Django REST APIs, React frontend, Redis caching, and async streaming, delivering real-time KPIs and intelligent visualizations with sub-500ms latency for 30+ cross-functional stakeholders.
- Worked as data scientist and ML engineer on 10+ supply chain optimization problems including predictive modeling, anomaly detection, and forecasting, orchestrating ETL pipelines with Airflow across multi-million row datasets.

Samsung Research - PRISM AI Intern, *Vellore, India* ***Dec 2023 - Apr 2024***

- Engineered safety testing framework for LoRA fine-tuning workflows, identifying critical vulnerabilities where adapter layers bypassed base model alignment and safety guardrails in open-source LLMs.
- Built automated evaluation pipeline using sentence-transformers to measure safety degradation across LoRA configurations, implementing metrics for jailbreak detection, harmful output classification, and alignment drift measurement.
- Developed best practices and mitigation strategies for production LoRA deployments, documenting secure fine-tuning protocols that preserve model safety while maintaining task-specific performance gains.

Hindustan Petroleum Corporation Limited - Data Science Intern, *Visakhapatnam, India* ***Mar 2023 - Jul 2023***

- Engineered predictive maintenance pipeline processing 100K+ daily sensor readings from refinery equipment, building forecasting models to predict component failures 24-48 hours in advance and prevent unplanned downtime through proactive scheduling.
- Developed real-time monitoring dashboard with automated alerting system for 20+ critical assets, enabling operators to intervene before failures and reducing emergency maintenance response time from 45 minutes to 15 minutes.
- Automated ETL workflows using Python to ingest sensor streams and SQL to integrate legacy system data, implementing self-updating model retraining cycles that reduced manual data processing time by 50%.

RESEARCH PUBLICATIONS

- **SOAR-ML - Synthetic Optimization for Oral Cancer Prediction (Springer CCIS):** Addressed critical data scarcity in early oral cancer detection by pioneering a GAN and VAE-based synthetic data augmentation framework, generating high-fidelity samples of rare positive OSCC cases. Combined synthetic augmentation with ensemble methods (CatBoost, XGBoost, LightGBM) to achieve 93.5% detection accuracy, significantly outperforming conventional diagnostic approaches limited by small datasets. https://link.springer.com/chapter/10.1007/978-3-031-79086-7_16.
- **Sentiment Analysis on Reddit Trading Data (IEEE Xplore):** Developed a multi-modal sentiment analysis framework combining VADER, Fourier Transforms, and LSTM networks to analyze 'r/wallstreetbets' discussions, transforming unstructured social trading sentiment into quantitative market signals. Achieved 87% accuracy in predicting short-term stock movements, bridging the gap between speculative retail trading discourse and systematic, data-driven investment strategies. https://ieeexplore.ieee.org/document/10493402.

PROJECTS

Reducing False Positives in Automated Code Review with Reinforcement Learning

- Addressed critical problem where automated code review tools flag 70-80% false positives, wasting developer time on irrelevant suggestions, by building RLHF system that learns from human approval/rejection feedback.
- Engineered dual-model architecture using Proximal Policy Optimization (PPO) - reward model trained on preference pairs distinguishes valuable suggestions from noise, while policy model optimizes review generation with KL-divergence constraints to prevent policy collapse.
- Implemented observability with LangSmith for monitoring policy gradient updates, reward predictions, and KL-penalty dynamics, achieving measurable reduction in false positives as agent learns developer preferences through continuous feedback loops.

AI Wardrobe Search & Outfit Recommendations

- Built multimodal search system enabling natural language wardrobe queries by combining CLIP embeddings with hybrid retrieval (BM25) and metadata filtering, implementing HNSW indexing and cross-modal re-ranking for sub-50ms outfit discovery while automatically scraping fashion trends from online and runway data to enhance recommendations.
- Engineered context-aware engine integrating real-time weather APIs, calendar sentiment analysis (detecting "interview" vs "casual coffee"), and color theory algorithms to deliver seasonally appropriate suggestions with intelligent memory management tracking user acceptance patterns.