VINAY CHANAMALLU

GENERATIVE AI & LLM ENGINEER | RAG & AGENTIC AI | MACHINE LEARNING & MLOPS

New York, NY | (240) 879-6884 | chanamalluvinay@gmail.com | linkedin.com/in/chanamalluvinay | github.com/VinChan2001 | vinaychanamallu.com

PROFESSIONAL SUMMARY

Generative AI and Machine Learning Engineer with 3+ years building production LLM, RAG, NLP, and forecasting systems. Own retrieval, LangGraph orchestration, and evaluation for a multi-tenant GenAI platform on AWS Bedrock serving three enterprise clients.

TECHNICAL SKILLS

- **Generative AI & LLMs:** Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), agentic workflows, multi-agent systems, LangGraph, LangChain, AWS Bedrock (Anthropic Claude, Titan Embeddings), OpenAI API, Ollama, tool calling, function calling, vector databases, semantic search, LLM evaluation, prompt engineering, Hugging Face Transformers, fine-tuning
- **Machine Learning:** Python, PyTorch, Scikit-learn, LightGBM, deep learning, classification, regression, predictive modeling, time series forecasting, ensemble methods, feature engineering, hyperparameter tuning, model evaluation, SHAP explainability
- **MLOps & Cloud:** AWS (SageMaker, ECS, S3, SQS, CloudWatch), Google Cloud (Vertex AI, BigQuery, Cloud Run), Docker, Terraform, Git, GitHub Actions, CI/CD, MLflow, ONNX Runtime, Evidently, model deployment, model monitoring, drift detection
- **Data, Backend & Analytics:** SQL, Pandas, NumPy, Spark (PySpark), Airflow, ETL and data pipelines, event-driven architecture, FastAPI, REST APIs, Pydantic, PostgreSQL, Redis, Pinecone, OpenSearch, FAISS, A/B testing, statistical analysis, Tableau, Power BI

PROFESSIONAL EXPERIENCE

GOML | AI Engineer (Full-Time, Contract) | New York, NY ***May 2026 – Present***

Project: *Multi-tenant enterprise GenAI platform for 3 client teams: 850+ queries/day across 65K documents.*

- Designed hybrid search for the RAG assistant (BM25 + Pinecone vectors, Cohere reranking, structure-aware chunking) and the evaluation pipeline (with LLM-as-judge) that measured it: Recall@5 rose from 72% to 84% and hallucinated answers fell 20%.
- Built a LangGraph-orchestrated workflow that extracts, cross-checks, and summarizes onboarding documents for analyst approval, reducing analyst handling time per case 30%.
- Cut inference cost per query 20% and p95 latency 16% with tenant-scoped Redis caching and leaner context.
- Enforced tenant isolation and document-level access control outside the LLM, with Bedrock Guardrails and output validation; deployed services on AWS ECS Fargate using Terraform and GitHub Actions CI/CD with gated releases.

GOML | AI Engineer Intern | New York, NY ***Sep 2025 – Feb 2026***

Project: *Document Q&A proof of concept (RAG) for a professional-services client across 4,800 contracts, SOWs, and process manuals.*

- Implemented the document ingestion and retrieval pipeline and raised answer quality from 3.4 to 3.9 out of 5 in human review by tuning chunking, retrieval, and prompts; earned a return offer after the proof of concept went to production.

BRILLIO | AI/ML Engineer | India ***Aug 2023 – Aug 2024***

Project: *Production NLP ticket-routing service and GenAI document summarization and Q&A pilots for a financial-services client.*

- Fine-tuned DistilBERT to route service tickets into 12 categories, raising macro F1 from 0.73 to 0.80 over a TF-IDF baseline.
- Served the model through FastAPI on AWS ECS (5,000+ requests/day); ONNX Runtime and inference optimizations cut p95 latency 43% with comparable classification performance.
- Automated training, evaluation, and deployment (SageMaker, MLflow, GitHub Actions), shortening release cycles from 10 to 7 business days; drift and performance monitoring caught an F1 drop after a new product launch, and retraining recovered it.
- Prototyped BART summarization and an early RAG Q&A tool (FAISS, FLAN-T5); in an analyst pilot, median review time fell 25%.

WIPRO | Data Scientist | India ***Oct 2021 – Jul 2023***

Project: *Store-SKU demand forecasting, stockout-risk scoring, and KPI reporting for a retail client with ~100 stores.*

- Developed XGBoost models that forecast weekly store-SKU demand up to four weeks ahead from 8.4M sales line items, lowering forecast error (WAPE) 10% vs. a seasonal-naive baseline.
- Analyzed an 80-store randomized pilot of forecast-driven replenishment; stores using the forecasts had a 13% lower stockout rate, supporting a staged rollout.
- Wrote PySpark and Airflow jobs feeding Tableau/Power BI dashboards, cutting weekly manual reporting by more than half.

PROJECTS

- **Agentic Research Assistant (LangGraph, MCP, OpenAI/Claude):** multi-agent planner-executor workflow with tool use, memory, and human review.
- **Telco Churn ML Platform (GCP Vertex AI, BigQuery, Cloud Run, SHAP, Docker):** event-driven BigQuery ingestion, Vertex AI champion/challenger training, Cloud Run batch scoring, and SHAP explanations; 0.86 ROC AUC.

EDUCATION, CERTIFICATION & PUBLICATION

UNIVERSITY OF MARYLAND, COLLEGE PARK | M.S., Business Analytics (GPA 3.6) ***Aug 2024 – May 2026***

JNTU KAKINADA | B.Tech., Computer Science (CGPA 8.57/10), India ***May 2018 – Jun 2022***

- **AWS Certified Machine Learning – Specialty (2026)** | **IEEE CICT 2025 (first author):** *Forecasting Hourly Electricity Consumption Using Ensemble Gradient Boosting and SHAP Insights*; 16% lower RMSE than a tuned LightGBM baseline.