

Parth Gajanan Ghayal

+1 (857) 313-2404 | parthmghayal@gmail.com | [LinkedIn](#) | [Github](#) | [Tableau](#)

AI Engineer with 6 years across data and AI, the last 2.5 shipping production LLM systems and AI agents end to end. Build multi-agent orchestration, agentic RAG with MCP tools and pgvector, structured output pipelines, and LLM evaluation harnesses on OpenAI, Claude, and Bedrock, running as multi-tenant Go/Python microservices on Kubernetes and AWS EKS. Currently applying this in healthcare AI: ambient documentation, clinical NLP, and FHIR/HL7 integrations in a HIPAA-regulated environment. Own systems in production, including priority-tiered model routing, guardrails, observability, and instant rollback of every AI write path. MS in AI, Boston University.

EXPERIENCE

Senior AI Software Developer | Nephrolytics, Inc., United States

02/2025 - Present

- Architected the ambient AI scribe as a 3-pass LLM pipeline converting visit audio into signed, structured encounter notes: ICD-10 diagnoses, impression, plan, medications, orders and CPT/HCC-RAF coding; note accuracy 36% to 95%, documentation time down 90%.
- Embedded with 8+ nephrology practices as their engineering POC: weekly sessions with practice managers & practice directors, field feedback scoped into requirements, prioritized with the CTO, Product head, then implemented, shipped & demoed back to the clients.
- Built Smart Summary and Smart Recommendations: Go services fuse normalized labs, document summaries and prior encounters into a source-traceable, KDIGO-grounded brief clinicians open each visit with (pgvector RAG, MCP, KFRE); chart review down 95%.
- Designed the clinical document pipeline: EMR docs mapped into 15 categories, an LLM pass resolving unknown types, category-specific summarization, Textract OCR for lab/RPM records, & a lightweight model pre-screening the rest for extractable labs; 92% accuracy.
- Build nightly LOINC normalization enabling cross-vendor lab trending over 4M+ observations at ~99%, via pgvector search & LLM judge.
- Owned bidirectional, multi-tenant EHR integration on the Unity API & FHIR R4/HL7 v2 across 8+ practices, building the encounter-close reverse-ETL that returns clinical notes, ICD-10 diagnoses & lab/imaging/procedure/referral orders to the system of record.
- Pioneered overnight chart computation, building tomorrow's charts before clinic rather than fetching mid-visit: 1,000+ patients nightly across Veradigm, TouchWorks & eCW, pulling 14 data types via async workers & summarizing prior encounters into structured notes.

Senior AI and Data Engineer | SQOR.ai, New York, United States

08/2024 - 02/2025

- Built 7-agent NLP Agentic framework using RAG, LangChain, AWS, Azure, OpenAI, & Claude to deliver real-time insights on 50+ SaaS tools utilization. Engineered chatbots boosting enterprise SaaS tool insights by 80% query resolution & 88% response accuracy.
- Developed ETL pipelines with Snowflake, custom APIs & AWS, enabling real-time analytics & reducing data retrieval time by 30%.
- Built time-series predictive model with NLP & Temporal Fusion Transformer for SaaS KPIs with 94% accuracy, boosting decision-making.
- Tuned Causality Analysis model with Correlation, Granger Causality & Causal Autoencoders (CVAEs) for 400+ SAAS tools KPIs.

Artificial Intelligence Engineer / AI Product Manager | Radical AI, New York, United States

02/2024 - 07/2024

- Engineered & tested an AI-powered chatbot using GCP (Vertex AI), Python and Streamlit, resulting in a 19% surge in user engagement.
- Led 2 teams to develop AI career coach & question generator using PyTorch, GPT-4, GCP, RAG, Langchain, Data Loaders, & MongoDB.

System Specialist | Boston University, Boston, United States

09/2022 - 07/2024

- Managed AD, IAM, 1.5 PT data backups, Windows/Linux systems & servers, research support, inventory management(\$70,000 saving).

Data Scientist | Indian Internet Solution, Aurangabad, India

04/2022 - 08/2022

- Built & fine-tuned ML & deep learning models using Python & Pytorch for COVID-era sales prediction, achieving 10% surge in revenue.
- Architected ETL pipelines with Databricks & Snowflake yielding 45% speed surge & improved data quality via advanced Python/SQL.

Jr. Data Scientist | Lalnaa Retails, Aurangabad, India

01/2020 - 03/2022

- Resolved Sales Channel Inefficiencies leveraging SQL, ETL, Python, AWS, Adhoc & Tableau, boosting quarterly revenue by 12%.
- Programmed a hybrid LSTM, BERT and CNN models for sales prediction and customer segmentation with estimated 91% accuracy.

TECHNICAL SKILLS

Programming Languages: Python, R, SQL, C++, Java

AI/Gen AI: LLMs, Agentic AI, AI Agents, Multi-Agent Orchestration, RAG, Prompt Engineering, OCR, Vision Models

LLM/ML Frameworks: LangChain, LangGraph, MCP, OpenAI API, Claude API, Hugging Face, PyTorch, TensorFlow, scikit-learn

Libraries & Frameworks: MLOps, LLM Evaluation, LLM-as-a-Judge, Model Evaluation, Guardrails, Observability, MLflow, Docker, Kubernetes, Terraform, CI/CD, AWS (Bedrock, Textract, Lambda, S3), Azure, GCP

EDUCATION

Boston University - Master of Science: Artificial Intelligence

PICT - Bachelor of Engineering : Computer Engineering (Hons: Artificial Intelligence - Machine Learning)

PROJECTS

System Identification with Neural ODEs | Deep Learning, Time-series forecasting, Hyper-parameter tuning ([Research Paper Link](#))

- Leveraged Pytorch to create novel neural ODE models (baseline & Liquid Time Constant neural networks) for 2D trajectory prediction.
- Achieved 38% loss reduction & efficient transfer learning in 5 epochs for LTCN, matching N=100 trajectories to the baseline's N=829.