Parth Shastri

shastriparth28@gmail.com +91 7588330340
linkedin.com/in/parthshastri github.com/parth-shastri

Summary

Versatile AI Engineer and Data Scientist with 3+ years of cross-domain experience spanning fintech, enterprise automation, healthcare, and voice AI, translating cutting-edge research into production systems that drive measurable business outcomes. Deep expertise across the full AI stack: retrieval-augmented generation, LLM fine-tuning, agentic systems, real-time inference, and multi-cloud MLOps. Driven by a core belief in **bridging the gap between AI research and real-world execution** by deeply understanding the pain points of inter-domain AI adoption and building systems that make advanced AI accessible, reliable, and impactful across industries.

Skills Summary

- **LLM & GenAI**: Prompt Engineering, RAG / Retrieval-Augmented Generation, LLM Fine-tuning (SFT, LoRA/QLoRA), LLM Evaluation, Agentic Systems (ReAct, LangGraph), Guardrails, vLLM, LlamaIndex, LangChain, HuggingFace (transformers, unsloth, datasets, PEFT), Unsloth
- **ML & Data Science**: Vector Embeddings, Semantic Search, BM25, Dense Retrieval, Pytorch, Tensorflow, XGBoost, Scikit-learn, Spacy, OpenCV, Pandas, SQL
- **Infrastructure & Tools**: FastAPI, Docker, Kubernetes, RabbitMQ, Celery, PostgreSQL, DuckDB, Redis, Qdrant, WebRTC, GIT, Jupyter Notebooks
- **Cloud**: GCP (Vertex AI, Model Garden, BigQuery, Cloud SQL, Document AI, Cloud Build, Dataflow, Cloud Storage, Compute Engine, Cloud Functions, Composer, GKE), Azure (AI Foundry, Azure Database for PostgreSQL Flexible Server, AKS, Azure Container Registry, Azure Monitor, STT/TTS), Multi-cloud Architecture

Professional Experience

- **Fulcrum Digital (Remote)** Aug 2025 – Present
  Lead AI Scientist
  - **AI Studio Platform: Multi-vertical RAG Knowledge Intelligence**:
    - Led architecture and delivery of a **multi-tenant RAG platform** enabling businesses in insurance, e-commerce, and enterprise to query proprietary knowledge (policies, contracts, SOPs, product catalogs) through natural language, acting as a **"second brain"** for organizational intelligence.
    - Designed a **hybrid retrieval layer** combining dense vector embeddings (via **Qdrant**) + BM25 sparse retrieval with **Reciprocal Rank Fusion (RRF)**, and a **query-routing layer** that classifies intent at inference time to direct queries to the correct path: RAG index, structured SQL store, or external tool-use.
    - Built **agentic multi-step workflows with LangGraph** (ReAct loops) enabling decomposition of complex business queries across multiple knowledge sources, producing **grounded, citation-backed responses**.
    - Architected the platform for **multi-cloud deployment** (GCP, Azure, AWS) to meet diverse client infrastructure and compliance requirements; with environment-agnostic containerization via Docker and Kubernetes.
    - Led a team of **4 AI engineers**; owned technical design, sprint planning, and pilot deployments for insurance and e-commerce verticals.
  - **Voice AI for Enterprise Invoice Processing**:
    - Built an end-to-end invoice processing and querying platform handling **10,000+** documents with **95%+** structured extraction accuracy via LLM-based validation and completeness scoring.
    - Developed a **ReAct agent** with RAG (semantic retrieval via **Qdrant** + fuzzy matching over PostgreSQL), delivering real-time streaming responses with **<100ms** time-to-first-token.
    - Integrated **RingCentral SIP** + Pipecat WebSocket audio streaming with Azure STT/TTS and Silero VAD to achieve **sub-500ms** voice-to-voice latency and **99%+** uptime, including live-agent transfer.
    - Deployed the full stack on **Azure Kubernetes Service (AKS)**, leveraging **Azure AI Foundry** for model lifecycle management and deployment, **Azure Database for PostgreSQL Flexible Server** for scalable relational storage, Azure Container Registry, and AKS autoscaling for production-grade reliability and cost efficiency.
  - **Multimodal Insurance Sales Coach**:
    - Built a real-time **WebRTC voice agent** with parallel pipeline architecture for concurrent LLM response generation, AI-powered sales nudges, and audio recording.
    - Developed **intent classification**-driven contextual nudge generation aggregating conversation context every 30 frames, delivering real-time coaching via WebRTC data channels, driving improved BDM adoption and conversion rates.
    - Designed a scalable persistence layer (**PostgreSQL** + cloud blob storage) supporting multiple concurrent sessions with comprehensive BDM performance analytics; used **Qdrant** for RAG-powered context retrieval for policy data and internal documents.

    - Deployed microservice architecture on **GCP Kubernetes Engine (GKE)**, using **Vertex AI Model Garden** for model serving and management and **Cloud SQL (PostgreSQL)** for managed relational storage, enabling independent scaling of voice processing, nudge generation, and analytics services with **99%+** uptime.

- **Bridgeweave Ltd., Hyderabad** Sep 2023 – July 2025
  Data Scientist
  - **AI-powered Equity Baskets**:
    - Developed deep learning models using a novel combination of **evolutionary algorithms and computer vision** (chart pattern recognition) to curate stock-baskets with **over 50% annual returns**, garnering an **AUM of over 100 Cr.**
    - Designed a low-risk stock-basket with **intelligent model switching** based on stock-level risk metrics, achieving a cumulative return of **40%**, **beating NIFTY 500 by 27%.**
    - Streamlined model evaluation and production workflows, **increasing team efficiency by 80%** (reduced model iteration cycle from 5 days to 1 day).
    - Backtested **1000+ models** using in-house Python tooling to deliver critical performance KPIs to stakeholders.
  - **LLM Agent for Platform**:
    - Developed and fine-tuned an **LLM-based agentic assistant** for stock analysis and investor education using **RAG** (Qdrant for vector retrieval), **prompt engineering, and tool-use**, increasing platform user engagement by **25%**.
    - Reduced LLM hallucinations by **60%** through guardrails, query-analysis, and structured **retrieval-augmented generation** with grounded citations.

- **Searce Cosourcing Services Pvt. Ltd., Pune** Jun 2022 – Aug 2023
  Machine Learning Engineer
  - **Plant Power Prediction** (Client: Tata Power): Trained and deployed an XGBoost regression model on Vertex AI achieving **95% prediction accuracy** for solar/wind plant power production.
  - **Entity Extraction via GCP Document AI** (Client: Apollo 24/7): Deployed fine-tuned document parsers to extract structured patient data from lab-test reports, significantly reducing manual review effort.
  - **Document Intelligence Service** (Client: Rockwell Automation): Built an end-to-end document processing service increasing downstream efficiency by **70%**; designed a novel tabular-data extraction pipeline using fine-tuned **CascadeTabNet** and image processing at **98% accuracy**; deployed on GCP via Kubernetes and Cloud Build, reducing uptime costs by **20%**.

- **Samsung Research Institute, Bangalore (Remote)** Aug 2021 – Mar 2022
  Project Intern *Prism Program*
  - Designed a novel **GAN-based data augmentation** method (Mel-spectrogram synthesis) to combat data scarcity in multilingual speech for Spoken Language Identification, published at **IEEE IConSIP 2022** and as an arXiv pre-print (arXiv:2205.15747).
  - Built a **CNN-LSTM** baseline LID model for benchmarking; augmentation improved classification accuracy on code-switched speech.

Publications and Blogs

- **Adversarial Synthesis-based Data Augmentation for Speech Classification** Aug 2022
  IEEE IConSIP 2022
  - DOI: 10.1109/ICoNSIP49665.2022.10007491
- **Adversarial Synthesis-based Data Augmentation for Code-Switched Spoken LID** May 2022
  Samsung Research Institute / arXiv
  - Pre-print: arXiv:2205.15747
- **Demystifying Diffusion Models** Dec 2022
  Searce Engineering Blog (Medium)
  - blog.searce.com/demystifying-diffusion-models
- **Phi2-ning: Fine-tuning a Small Language Model** May 2024
  Medium
  - medium.com/@shastriparth28/phi-2ning

Certifications

- **Google Cloud**: Certified **Professional Machine Learning Engineer** | Certified **Associate Cloud Engineer**

Education

- **College of Engineering, Pune** | B.Tech in Electronics & Telecom. Engg. | CGPA: 7.93 Aug 2018 – May 2022