

Tanish P D

AI/ML Engineer

+91-6364203536 | pdtanish1@gmail.com | Bangalore, India
linkedin.com/in/tanish-p-d | github.com/Tanishpd | tanishpd.github.io/portfolio

SUMMARY

Builds production LLM systems — LangGraph multi-agent orchestration, retrieval-augmented generation (RAG), LLM fine-tuning, and conversational & voice AI. **Best Paper** co-author (IEEE ICDSBS 2026). Independent projects span LLM security guardrails, cost-aware model routing, and real-time ML infrastructure.

EXPERIENCE

Python AI Developer Intern

Dec 2025 – Present

Desible.ai (Basal Analytics Private Limited)

Bangalore, India

- Built BFSI conversational AI with LangGraph multi-agent orchestration, shipping **200+ production pull requests** to automate customer interactions
- Built multilingual AI voice agents powering 24/7 automated customer service
- Deployed production RAG and context-management systems; fine-tuned domain-specific LLMs, cutting inference cost and latency

PUBLICATIONS & AWARDS

Best Paper Award, IEEE ICDSBS 2026 — “Career-Copilot: An AI-Powered Recruitment and Career Preparation Platform,” 3rd Intl. Conf. on Data Science & Business Systems, SRM IST Chennai (IEEE Madras Section); co-author

TECHNICAL PROJECTS

Career Copilot | *Python, Flask, Google Gemini, Whisper, Sentence-BERT*

- Achieved **91.8%** interview-assessment and **89.6%** resume-analysis accuracy with a Gemini AI interviewer using adaptive follow-ups and Whisper ASR scoring
- Cut candidate screening time by **68%**, replacing keyword matching with Sentence-BERT semantic resume-JD scoring

Secure Air-Gapped LLM Platform | *FastAPI, ONNX Runtime, Qdrant, Ollama*

- Detected prompt injection and PII at **1.0 recall** (0.854 precision) inside a network-isolated LLM enclave — Presidio pseudonymization and an ONNX DeBERTaV3 classifier fronting a local Ollama model, with RAG over a Qdrant vector database (280 tests passing)
- Enforced isolation with a dual-layer in-process egress guard (socket monkeypatch + PEP 578 audit hook) that raises EgressViolation on any non-loopback connection

Real-Time Streaming Feature Store | *PyFlink, Kafka, Redis, Feast*

- Designed a real-time Kafka → Flink → Redis pipeline (windowed aggregates + MiniLM embeddings in Feast’s native encoding): **2.17 ms p99** online reads at **8,500 events/s** sustained
- Hardened feature serving with 89 automated tests and 62 fault-injection drills

LLM Cost Optimizer | *FastAPI, Redis, NumPy, Ollama*

- Shipped an OpenAI-compatible router (913 tests): a 5-tier cascade (caches → 0.5B → 3B local → frontier) that escalates only on measured token log-prob uncertainty, with per-gate logistic calibrators
- Added a Redis + NumPy semantic cache over 768-d embeddings with a **47 ms p50** hit latency

Multi-Region Failover Proxy | *Starlette, httpx, Redis, Lua*

- Survived a simulated region outage with **0 client errors** across 801 requests via priority-ordered httpx failover spanning 3 regions, **4.8 ms p50** fault detection, and atomic Redis token-bucket rate limiting (Lua)
- Proved resilience with an open-loop chaos harness (kills regions under load) and 58 unit tests

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript, TypeScript, Java, C/C++

Machine Learning: PyTorch, TensorFlow, scikit-learn, LangChain, LangGraph, Sentence-BERT, Whisper, Ollama

Backend & Infra: FastAPI, Redis, Kafka, Flink, Qdrant, Docker, ONNX Runtime, pytest, GitHub Actions (CI/CD)

Focus Areas: Multi-Agent Systems, LLM Fine-Tuning, RAG, NLP, MLOps, Model Deployment, Prompt Engineering

EDUCATION

Dayananda Sagar University

Aug 2022 – May 2026

B.Tech in Computer Science & Engineering; CGPA: 8.64/10.0

Bangalore, India