

Yash Pal

📍 Mumbai ✉ 2004yashpal@gmail.com ☎ 9326259927 🌐 yashpal.me in LinkedIn GitHub

Profile

Recent AI/ML graduate with a published research paper and hands-on experience building production-grade LLM systems. Designed MITRA-AI — a multi-agent platform integrating 5 LLM providers with NLP-based routing and async API architecture — and completed a QA automation internship at Xemi. Strong foundation in Python, FastAPI, and generative AI development. Eager to contribute to AI product teams building real-world applications.

Skills

Agentic AI: Multi-Agent System Design, LLM Reasoning & Planning Workflows, Tool-Use Patterns, Agentic Orchestration, MCP (Model Context Protocol), Prompt Engineering, Context-Aware NLP Routing

Generative AI & LLMs: OpenAI API, Claude API (Anthropic), Google Gemini API, DeepSeek API, LLM Inferencing Pipelines, RAG Pipelines (actively building), Embedding Models & Semantic Search (conceptual)

Cloud & Azure: Azure AI Foundry (actively upskilling), Azure OpenAI (actively upskilling), AWS (deployed projects), Firebase Suite, Google AI Studio

Backend & APIs: Python, JavaScript, FastAPI, REST API Design & Integration, Microservices Architecture, Async Patterns, Webhooks

Vector Databases: pgVector / OpenSearch / Milvus (conceptual & actively integrating), Cloud Firestore

Frameworks & Tools: LangChain (actively upskilling), LangGraph (actively upskilling), TensorFlow, PyTorch, Flutter, Git/GitHub, Linux, Docker (basic)

Practices: Responsible AI, Observability & Logging, Secure API Design, Technical Documentation, Agile Development

Education

University of Mumbai

Nov 2022 – May 2026

Bachelor of Engineering in Computer Science (Artificial Intelligence & Machine Learning)

- GPA: 7.5/10.0
- **Coursework:** Artificial Intelligence, Deep Learning, Machine Learning, Reinforcement Learning, Data Structures, Project Management.

Experience

Automation Engineering Intern

March 2026 – April 2026

Xemi Private Limited

- Developed Python scripts to automate testing workflows, reducing manual regression effort and improving system reliability across application modules.
- Performed systematic bug identification, root-cause analysis, and diagnostic reporting — applying analytical problem-solving to complex software issues.
- Monitored application performance and contributed logging analysis and observability insights to the QA pipeline.
- Collaborated cross-functionally and maintained technical documentation aligned with responsible engineering practices.

Projects

MITRA-AI — Multi-Agent Agentic AI Platform (Python / FastAPI)

[MITRA-AI](#) 

- Designed and built a production-grade agentic AI platform integrating 5 LLM providers (Claude, GPT-4, Gemini, DeepSeek, Kimi K2) with intelligent NLP-based routing implementing core agentic patterns: reasoning, planning, tool-use, and autonomous workflow execution.
- Architected a scalable Python/FastAPI backend with multi-agent orchestration, async request handling, and enterprise-style REST API design — analogous to Azure AI Foundry's agent service architecture.
- Implemented MCP (Model Context Protocol) awareness and prompt engineering strategies to optimize context-aware, multi-turn LLM interactions across providers.
- Deployed on AWS with Git-based version control; designed for responsible AI output with logging and performance monitoring across agent pipelines.
- Tools Used: Python, FastAPI, Flutter/Dart, OpenAI API, Claude API, Gemini API, DeepSeek API, REST APIs, MCP, AWS, Git.

Generative Adversarial Networks (GANs) — Autonomous Training Pipeline

[GAN](#) 

- Designed an autonomous adversarial training pipeline in Python/TensorFlow where Generator and Discriminator agents iteratively reason and adapt — a foundational pattern shared with modern agentic AI feedback loops.
- Implemented custom training loops with performance monitoring and iterative model optimization.
- Tools Used: TensorFlow, Python, Jupyter Notebook

SRCNN Super-Resolution — End-to-End AI Inference Pipeline

[SRCNN](#) 

- Built a complete AI inference pipeline — data ingestion, preprocessing, embedding-style feature extraction, model inferencing, and output evaluation — structurally analogous to a RAG pipeline for enterprise knowledge retrieval.
- Evaluated pipeline performance using PSNR metrics, demonstrating a systematic approach to AI output quality monitoring.
- Tools Used: TensorFlow, Keras, OpenCV, Python

Book-My-Turf — Cloud-Native SaaS with Enterprise API Integration

[BookMyTurf](#) 

- Developed a cloud-backed SaaS platform integrating Firebase (managed cloud DB), REST APIs, and real-time authentication — demonstrating enterprise API and cloud service integration patterns.
- Designed scalable data architecture for multi-entity systems (users, bookings, turf owners) with Cloud Firestore as a managed metadata store.
- Tools Used: Flutter, Firebase, Dart, REST APIs, Cloud Firestore

Certifications & Achievements

Research Paper Published — Multimodal Conversational AI System

[Google](#) 

AI Fluency: Framework and Foundations

[Anthropic](#) 

Machine Learning Specialization – DeepLearning.AI

[Coursera](#) 

Generative AI Studio

[Google](#) 

Oracle Cloud Infrastructure AI Foundations

[Oracle](#) 

Career Essentials in Generative AI by Microsoft and LinkedIn

[LinkedIn](#) 

Career Essentials in GitHub Professional Certificate

[LinkedIn](#) 

Microsoft Copilot: The Art of Prompt Writing

[LinkedIn](#) 

AI For All by Skill India

[Skillindia](#) 