Vedant Andhale

Applied AI Engineer | Agentic Systems, LLM Evals & Backend Infrastructure
Nashik, Maharashtra, India | vedant@vedant.me | vedant.me
linkedin.com/in/VedantAndhale | github.com/VedantAndhale

Summary

Applied AI Engineer specializing in 0-to-1 multi-agent architectures, LLM evaluation pipelines, and high-reliability backend microservices. Production experience shipping decision-support systems with strict validation boundaries, client-side WebGPU inference, and deterministic fallback logic.

Experience

AI Engineer, agntworks.ai (Shinka Labs LLP) – Jaipur, Rajasthan, India — March 2026 - present

- Co-architected 0-to-1 multi-agent decision systems with the founder, automating pricing and inventory optimization workflows over enterprise catalogs.
- Built asynchronous FastAPI microservices on GCP with strict Pydantic v2 validation boundaries, structured outputs, and rate-limited LLM orchestration.
- Designed automated evaluation pipelines using deterministic assertions and LLM-as-a-judge harnesses, tracking drift and decision reliability across model iterations.

AI and Cloud Intern, Edunet Foundation – Remote — February 2024 - March 2024

- Developed predictive ML classification models on clinical datasets in Python (scikit-learn), identifying disease risk patterns across patient cohorts.
- Engineered end-to-end data preprocessing pipelines: automated cleaning, exploratory analysis (EDA), and feature engineering; deployed models on IBM Cloud.

Projects

Engram - Local-First Second Brain with On-Device AI
Tools Used: React 19, TypeScript, WebGPU, LiteRT-LM, MCP SDK, IndexedDB | GitHub | Website

- Engineered an offline-first PWA with bidirectional knowledge graphs, IndexedDB caching, and zero-conflict sync to private GitHub repositories.
- Integrated client-side on-device LLM inference via WebGPU and LiteRT-LM (Gemma 4 E2B), enabling zero-cost, private document Q&A with exact passage citations.
- Implemented a hosted Model Context Protocol (MCP) server exposing semantic search, highlight retrieval, and async note suggestions to external coding agents.

FreightSense - Inspectable AI Decision-Support Engine
Tools Used: Python, FastAPI, LLMs, Decision Support, Docker | GitHub | Website

- Built a hybrid decision-support engine decoupling deterministic financial/delay routing formulas from LLM natural language intervention reasoning.
- Structured an immutable human-override telemetry log to systematically benchmark agent precision and establish ground-truth datasets for model refinement.
- Implemented modular FastAPI microservices with structured JSON outputs, validation boundaries, and containerized Docker deployment.

Crop Cure - Plant Pathology Classifier
Tools Used: PyTorch, Computer Vision, FastAPI, WhatsApp API | GitHub | Website

- Trained a deep learning computer vision model in PyTorch to detect and classify grape leaf diseases from field photography.
- Connected the classifier to a multilingual WhatsApp messaging workflow via FastAPI for low-latency automated farmer triage.

loadfile - Unified Tabular Ingestion Library
Tools Used: Python, Pandas, Cloud Storage, Open Source | GitHub | Website

- Developed an open-source Python library unifying tabular data loading across local files and cloud storage (S3/GCS) into Pandas.
- Automated format sniffing, cloud credential resolution, and streaming parsers to eliminate repetitive ingestion boilerplate.

Technical Skills

AI & LLM systems: Multi-agent workflows, LLM evaluation (evals), prompt caching, structured outputs (Pydantic v2), MCP, PyTorch
Backend & cloud: Python (asyncio), FastAPI, REST APIs, Docker, Google Cloud Platform (GCP), WebGPU, Linux, Git, CI/CD
Databases & ingestion: SQL (PostgreSQL, MySQL), pandas, NumPy, data pipelines, schema validation, exploratory data analysis (EDA)
Developer tools: VS Code, LiteRT-LM, Jupyter Notebook, Typst, RenderCV, Jira, GitHub Actions

Education

Savitribai Phule Pune University (SITRC) — 2021 - 2025
Bachelor of Engineering in Artificial Intelligence and Data Science (CGPA: 7.75) — Nashik, Maharashtra, India

Certifications

- AWS Certified Cloud Practitioner | Credential
- Astronomer Certification for Apache Airflow 3 Fundamentals | Credential
- GitHub Certified: GitHub Foundations | Credential
- Microsoft Certified: Azure AI Fundamentals | Credential
- Google Data Analytics Certificate | Credential