Pulkit Jai

Email: hey@pulxit.com
LinkedIn: pulkitjai
GitHub: limahotel44

Summary

AI engineer focused on LLM evaluation, red teaming, and reliable deployment. Experience building reproducible evaluation tooling, multi-agent systems, and production ML pipelines across startup environments.

Research & Selected Projects

- **ARTEMIS: Advanced AI Reasoning Threat Evaluation Intelligence System**
  OpenAI gpt-oss-20b Red-Teaming Challenge | Python, LLM Evaluation, Red Teaming Aug 2025
  - **Honorable Mention** in OpenAI's gpt-oss-20b red-teaming challenge for work on Chain-of-Thought reasoning manipulation, reproducing **12/12 successful bypasses** across 4 attack vectors and 3 harmful content categories in controlled testing.
  - Built ARTEMIS, a reproducible CLI/notebook evaluation framework with dual-LLM scoring for harmfulness and guardrail-bypass detection; executed 88 automated tests across the reproduction pipeline.
  - Published the methodology, fresh-conversation constraint, and validation setup through a public Kaggle write-up, GitHub repository, reproduction notebook, and structured JSON findings.
- **VoiceVision: Accessible OCR-to-Speech Device for the Visually Impaired**
  B.Tech. Final Year Thesis | Python, Raspberry Pi, TrOCR, pyttsx3, LLM API Jan 2024 – Jul 2024
  - Built a standalone assistive device that captures a scene, extracts text, and reads it aloud, targeting offline accessibility for visually impaired users.
  - Fine-tuned a quantized TrOCR model (ONNX) for local inference on Raspberry Pi and implemented a confidence-gated fallback to a GPT API for low-confidence or handwritten text.

Experience

- **Hivel.ai** Remote
  Applied AI Engineer (Promoted from Intern) Apr 2025 – Sep 2025
  - Built a production 6-agent code-review system in Python/MCP with asynchronous orchestration and shared agent state for structured review generation.
  - Designed the workflow end to end, including task decomposition, inter-agent communication, conflict handling, and deterministic output formatting.
  - Implemented authentication/authorization, audit logging, PII scrubbing, and deployment guardrails aligned with SOC 2, HIPAA, and GDPR requirements; set up CI/CD and observability with PostHog and Sentry.
  - Piloted the system with enterprise customers and incorporated structured feedback to meet integration requirements.
- **ReWorked.ai** Bengaluru, IN
  AI/ML Engineer (Intern → Full-Time) Jun 2024 – Mar 2025
  - Built and shipped PyTorch + scikit-learn classification models for solar, roofing, and mortgage lead scoring, achieving F1 > 0.8 across all three verticals.
  - Designed parallel inference pipelines processing 5,000+ records/day and deployed models via FastAPI on AWS with monitoring, alerting, and rollback procedures.
  - Owned model versioning, retraining schedules, and production monitoring end to end after conversion from intern to full-time engineer.
- **EssentiallySports** Remote
  F1 Journalist and Social Media Manager May 2023 – Apr 2024
  - Published 427 Formula 1 articles for a major sports media platform while completing a full B.Tech., demonstrating sustained writing throughput and deadline discipline.
  - Developed technical writing and science communication skills by translating motorsport engineering concepts for general audiences.

Education

- **Manipal University Jaipur** Jaipur, IN
  B.Tech., Electrical & Electronics Engineering; Minor in Computer Science & Machine Learning Jul 2020 – Jul 2024
  - GPA: 7.6/10 overall; final-semester GPA: **9.0/10**.
  - Coursework: Machine Learning, Neural Networks, Computer Vision, Data Structures & Algorithms, Database Systems.
  - Electronics Club — Research & Engagement Head: organized technical programs and inter-university events.

Skills

- **Evaluation / Safety**: adversarial testing, jailbreak analysis, failure analysis, eval harness design, classifier-based detection
- **ML / LLMs**: PyTorch, JAX, scikit-learn, ONNX, transformer fine-tuning (LoRA), retrieval systems, multi-agent workflows, evaluation pipelines
- **Systems**: FastAPI, AWS, CI/CD, observability, PostHog, Sentry
- **Programming**: Python, TypeScript