

Chandhan Saai Katuri

Applied AI Engineer — production LLM systems

San Francisco, CA · katurichandhansai@gmail.com · +1 (240) 370-1057
<https://chandhan.com> · github.com/ChandhanSaai · linkedin.com/in/chandhan-saai-katuri

EXPERIENCE

Verita AI — Software Engineer · San Francisco, CA **Apr 2026 – Present**

- Designed and shipped a production vision-LLM audit pipeline that analyzes contractor screen recordings against per-project task policies, classifies work hours (AI use, off-task, guideline violations), and produces sealed, priced, recomputable audit ledgers; benchmarked read models across providers via OpenRouter and achieved 80% precision and recall on a golden evaluation set, with 264 sessions sealed in production.
- Architected the core of an LLM data-annotation platform (Django REST Framework, React): five-role access control with a capability matrix, an audited workflow state machine with calibration routing, screening and qualification gates, an append-only audit log covering all 54 endpoints, and 10+ annotation widget types; expanded the test suite from 210 to 1,239 tests.
- Hardened the platform's image-ingestion pipeline with SSRF-safe fetching, presigned re-hosting to S3, and asynchronous batched SQS enqueueing that eliminated timeouts on 1,000-image imports; delivered 6,756 processed images to a client with geo-derived metadata backfill.
- Delivered the second-generation finance-interview rubric verifier for an AI interviewing product in six days, implementing LLM-judge claim validation, shadow-verifier gating, and hard execution time limits; instituted a mutation-testing standard requiring each test to demonstrably fail against a deliberate source defect.
- Led SOC 2 Type II readiness from 154 to 165 of 173 controls (95%) during an active audit window: executed a pgAudit and storage-encryption migration across 11 production PostgreSQL databases, deployed GuardDuty/CloudTrail/file-integrity monitoring, produced 40+ evidence artifacts, and remediated authenticated penetration-test findings.
- Developed a multi-scanner application-security pipeline (gitleaks, OpenGrep, Bandit, Trivy, CodeQL) with LLM-assisted triage, reducing 345 raw findings to 33 verified vulnerabilities including nine critical issues (remote code execution, SSRF, IDOR, exposed credentials).

Handshake AI — AI Data & Evaluation · Remote **Sep 2025 – May 2026**

- Evaluated frontier image-generation models against calibrated quality rubrics (rendering artifacts, anatomical consistency) to produce preference data for RLHF pipelines; reviewed Kaggle machine-learning codebases for algorithm selection and implementation quality.
- Selected as a Star Fellow for a program evaluating model reasoning against frontier research publications.

PROJECTS

GPT-2 Transformer Implementation and Training Pipeline (124M parameters) — *PyTorch, CUDA*

- Implemented the complete GPT-2 architecture and training pipeline in PyTorch with bfloat16 mixed precision, torch.compile, fused scaled-dot-product attention, gradient accumulation, cosine learning-rate scheduling with warmup, and BPE tokenization, achieving roughly 10x training throughput over the unoptimized baseline.

Adversarial LLM-Judge Evaluation Framework (Project D3) — *Python, Multi-Agent Systems, OpenAI API*

- Developed a structured adversarial evaluation framework in which advocate, judge, and jury agents deliberate over competing LLM responses; implemented single-round and iterative multi-round debate protocols with persona-diverse juror voting, per-evaluation token budgeting and cost tracking, and position-swap bias measurement; benchmarked on MT-Bench, AUTO-J, and AlignBench.

Metered API Billing Platform — *Django, PostgreSQL, React, TypeScript, Docker*

- Engineered a production-grade usage-based billing platform with idempotent event ingestion, tiered invoicing with precise monetary rounding, HMAC-signed webhooks with replay protection, and a database-enforced append-only audit log; shipped customer and operations dashboards as two React/TypeScript applications backed by 125+ tests against live PostgreSQL.

EDUCATION

University of Maryland, College Park — *Master of Engineering, Robotics* **Jul 2023 – May 2025**

SRM Institute of Science and Technology — *B.Tech, Mechatronics (Robotics specialization)* **Jul 2019 – Jun 2023**

SKILLS

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

LLM systems: VLM pipelines, RAG, embeddings & vector search (FAISS, Pinecone), LangChain/LangGraph, LLM evaluation (LLM-as-judge, golden datasets, mutation testing), OpenAI/Anthropic/OpenRouter APIs

Backend & web: Django/DRF, FastAPI, React, PostgreSQL, Celery, SQS, REST APIs

Infra & DevOps: AWS (S3, RDS, SQS, CloudWatch, GuardDuty, Elastic Beanstalk, ECR), Terraform, Docker, GitHub Actions, Sentry

ML / DL: PyTorch, Transformers, computer vision (YOLO, ViT), ONNX/TensorRT edge inference