Shashank Yadav AI/ML Engineer

Denver, CO | yadav.shashank1201@gmail.com | shashankyadav.dev | GitHub | LinkedIn

SUMMARY

AI/ML engineer with 3+ years building production software. At Ally Financial I put an LLM into a real-time fraud risk pipeline under banking audit rules. On my own I build retrieval and ML systems and measure each one against a baseline before I trust it.

EXPERIENCE

Software Developer, Ally Financial — Jul 2025 – Present | Remote, USA

- Added an LLM to a real-time transaction risk scoring pipeline to write anomaly narratives for fraud analysts, with prompt guardrails and validation built with the AI and compliance teams to meet banking audit rules.
- Owned the core scoring services (Python, streaming, caching) of a new real-time fraud detection platform, which cut fraud losses 18% in its first quarter.
- Cut risk evaluation and alerting API latency from 340 ms to 265 ms by reworking query plans and pooling connections.
- Built a React fraud analytics dashboard that replaced spreadsheet reports and cut investigation time by about 21%.
- Owned service auth (OAuth 2.0, JWT) and CI/CD. Wrote 85+ tests that run on every PR and cover over 90% of critical paths.

Graduate Teaching Assistant, University of Massachusetts Boston — Sep 2024 – May 2025 | Boston, MA

Ran 8 lab cohorts (90+ students) in circuit analysis and MATLAB, and gave feedback on 70+ lab reports within 24 hours.

Software Engineer, Tata Consultancy Services — Aug 2022 – Aug 2023 | Gurgaon, India

- Led a zero-downtime Oracle to PostgreSQL migration of a 20 GB production database, refactoring 15,000 lines of SQL and code. Saved $45K a year in licensing and cut page loads by 4 seconds.
- Shipped seven features to a core banking platform (Java, Spring Boot) handling 50K+ transactions a day. Wrote 120+ JUnit tests and owned the Jenkins pipeline. Defects fell 40% from the prior release.
- Managed a 6-person Tier-2 support team for banking clients and raised 24-hour SLA compliance from about 65% to 90%.

Software Development Engineer Intern, NTT Data — Jan 2021 – Jul 2022 | Remote, India

- Built Angular UI and Spring Boot APIs for an enterprise app, including a Kafka order service handling 2K+ transactions a day.
- Added role-based access control (Spring Security, JWT) and CI test gates that moved releases from biweekly to weekly.

PROJECTS

PitchPerfect | TypeScript, Next.js, Express, Fastify, MongoDB, OpenAI API, Playwright

- Sales practice and hiring site where students pitch an LLM buyer with hidden objections and get a 7-part scorecard.
- Each buyer sentence is moderated and leak-checked before it streams, and every scorecard quote is matched to the transcript. Tested with 31 evals (9 adversarial) on a live model, plus 3,860 unit and 49 Playwright tests in CI.

Ask the Filings | Python, PostgreSQL, pgvector, Gemini API, cross-encoder reranker

- Answers questions on SEC 10-K filings with source passages and refuses when they lack the answer: 111 of 117 answerable right and zero wrong on a 144-question eval. Hybrid search and a reranker lifted recall at 8 from 0.60 to 0.89.

pytest-triage | Python, Gemini API, BM25, scikit-learn, GitHub Actions

- Suggests labels and duplicates for new pytest issues daily in shadow mode, scored against the maintainers' own labels: macro-F1 0.58 vs 0.46 for the best method without an LLM. None of 104 prompt-injection attack runs got through.

AccuScan | PyTorch, CTC, ONNX Runtime Web, TypeScript

- In-browser handwriting reader: 20.1% character error on unseen words vs 51.4% for zero-shot TrOCR-small, at 1/50 its size.

GestureSign | PyTorch, MediaPipe, ONNX Runtime Web, React

- In-browser ASL recognition for 250 signs from MediaPipe landmarks: 72.1% top-1 and 88.1% top-5 on signers it never saw.

SKILLS

LLM and ML: RAG, hybrid search, reranking, evals, guardrails, prompt design, Gemini/OpenAI APIs, PyTorch, Transformers, NLP
Backend and data: Python (FastAPI), Java (Spring Boot), Node.js (Express), C++, gRPC, Kafka, PostgreSQL, MongoDB, Redis, Oracle
Cloud and DevOps: AWS (Lambda, Fargate, EKS, S3, SQS), Azure, Docker, Kubernetes, Helm, ArgoCD, GitHub Actions, Jenkins
Testing and monitoring: pytest, JUnit, Mockito, Testcontainers, Vitest, Playwright, Cypress, Prometheus, Grafana
Frontend: TypeScript, React, Next.js, Tailwind CSS, Angular, ONNX Runtime Web, Web Speech API

EDUCATION

MS in Computer Science, University of Massachusetts Boston | GPA 3.81/4.0 — Sep 2023 – May 2025

B.Tech in Computer Science and Engineering, Graphic Era Hill University, India | GPA 8.59/10 — Aug 2018 – May 2022