

ASHWIN RACHHA

ashwin.rachha@gmail.com | [linkedin.com/in/ashwinrachha](https://www.linkedin.com/in/ashwinrachha) | +1-540-824-8748 | github.com/ashwinrachhavt | Website

Summary

Founding and applied AI engineer building production agentic systems, financial infrastructure, and end-to-end product workflows. Experienced across LLM applications, backend and platform engineering, product ownership, and high-trust operational domains including fintech and mortgage technology. Passionate about AI, Philosophy, Chess, Travel and Art.

Experience

Loan Labs

Applied AI Engineer

Manhattan Beach, CA

Jan 2026 – Oct 2026

- Built **Lois**, an agentic platform for business-purpose mortgage-loan processing, internal operations, and pilot workflows; re-architected one-off Ruby LLM calls into a LangGraph system deployed on Amazon Bedrock AgentCore.
- Automated document classification, lender-specific file renaming, and policy validation for unstructured loan-document workflows; exposed real-time agent capabilities through borrower email intake and an in-product conversational interface in LoanOS.
- Designed a **fail-closed authorization layer** for Composio-connected SaaS tools, replacing a coarse email toggle with independently managed write, send, merge, and archive permissions for high-trust mortgage workflows.
- Integrated customer-connected workflows across Google Drive, Box, Salesforce, HubSpot, Pipedrive, OneDrive, and SharePoint through fine-grained, least-privilege controls, enabling Lois to retrieve context and take approved actions across customer systems.
- Introduced an AI-assisted **Software Factory** that connected Linear product work, Notion specifications, company knowledge, AI coding tools, and senior-engineer review into a shared delivery workflow; enabled the CEO and Product Designer to participate directly in scoped, reviewable implementation work.
- Established Linear and Notion as the company knowledge system, consolidating product requirements, technical specifications, meeting decisions, and linked implementation context into structured, backlink-driven documentation for both people and AI agents.
- Designed the factory around Marr's three levels of analysis—computational goals, representations/algorithms, and implementation—and introduced an iterative *scope, plan, assess* loop to improve task clarity, constrain agent autonomy, and create self-correcting implementation feedback.
- Scaffolded a Next.js marketing-site codebase and collaborated with the CEO and Product Designer to ship Loan Labs' marketing website, establishing a practical template for cross-functional, AI-assisted product development.
- Moved LoanOS toward agent-friendly client-server API boundaries while retaining Rails as the system of record, giving LangGraph agents structured access to loan context and approved actions without exposing direct MVC-layer access.

Finally

AI Product Engineer / Tech Lead

Boca Raton, FL

Feb 2024 – Jan 2026

- Joined as Finally's first AI Product Engineer and led a three-engineer team that took **Classify AI** from prototype to production: an LLM-assisted bookkeeping workflow processing **50,000+ transactions daily** and reducing manual categorization by approximately **80%**.
- Designed a retrieval-augmented transaction-classification system using LangChain, Pinecone, Elasticsearch, Redis, Celery, Django, and Weights & Biases; combined semantic transaction history, merchant enrichment, and custom charts of accounts for bookkeeping recommendations.
- Evolved Classify AI from CSV upload to automated bank-data ingestion with Plaid/Teller, OCR-supported statement processing, reconciliation checks, and QuickBooks push; helped reduce first-month close time from 4+ months to approximately 2 weeks.
- Architected reusable banking infrastructure for account linking, token lifecycle management, encrypted access-token storage, webhooks, transaction synchronization, statement retrieval, and provider-normalized account data.
- Built cash-based underwriting for Finally's corporate-card product using 90-day bank data, daily-balance reconstruction, Z-score logic, weekly recalculation, audit history, notifications, and manual-override controls.

UNAR Labs

Machine Learning Engineering Intern

Portland, ME

Jun 2023 – Aug 2023

- Developed accessibility-focused backend systems and data pipelines for visually impaired users using computer vision and NLP technologies.
- Architected preprocessing, model-training, and inference pipelines using OpenCV, PyTorch, Transformers, and FastAPI.
- Deployed scalable ML solutions on GCP with Docker containerization and Hugging Face model integration.

Outreach

Machine Learning Platform Engineering Intern

Seattle, WA

May 2022 – Aug 2022

- Built reusable NLP inference and deployment infrastructure using PySpark, MLflow, ONNX, NVIDIA Triton, Docker, Go/Python microservices, CI/CD, and GKE.
- Developed text-processing pipelines and production model-serving workflows for topic detection and sentiment-analysis use cases.

Mindbrowser Inc.

Software Engineer, Machine Learning

Pune, India

Dec 2019 – Feb 2021

- Developed a facial-expression-recognition system for CRM meeting analysis using VGG-19 transfer learning, achieving 73% accuracy.
- Built Flask services with RBAC/JWT authentication, PostgreSQL user and session metadata, and MongoDB GridFS storage for raw video and per-frame inference results.
- Implemented video ingestion, frame sampling and preprocessing with OpenCV, inference workflows, and REST APIs for job status, paginated results, and CSV/JSON exports.

Education

Virginia Tech

M.S. in Computer Science (Thesis) — GPA: 4.0/4.0

Blacksburg, VA

Aug 2021 – Dec 2023

Pune Institute of Computer Technology (PICT)

B.E. in Computer Science — GPA: 8.70/10.00

Pune, India

Jul 2016 – Nov 2020

Skills

Languages: Python, Go, Ruby, TypeScript, SQL

Docker, Kubernetes, GKE

AI / Agents: LangGraph, LangChain, Amazon Bedrock Agent-Core, RAG, MCP, Composio, Pinecone, Elasticsearch, W&B

ML / Platform: PyTorch, Hugging Face, ONNX, NVIDIA Triton, MLflow, PySpark, OpenCV

Backend / Data: Rails, Django, FastAPI, PostgreSQL, Redis, Celery, Snowflake SQL, Plaid, Teller

Engineering: REST APIs, OAuth, webhooks, CI/CD, ArgoCD, Datadog, Sentry

Web / Cloud: React, Next.js, Tailwind CSS, AWS, GCP,

Selected Projects

Gurukul LMS Platform — *Django, Next.js, TypeScript, Tailwind CSS, Prisma, Supabase, Docker*

[LIVE](#) | [CODE](#)

- Master's thesis project and research platform: an LLM-enhanced learning environment using RAG and guardrails to promote Socratic reasoning in data-structures and algorithms education.
- Built a coding platform with a custom editor, GPT-4 tutoring integration, structured problem tables, and guided learning workflows.

Neuralflow — **AI-Powered Productivity App** — *Next.js, TypeScript, Clerk, Supabase*

[LIVE](#) | [CODE](#)

- Built a full-stack productivity application combining a Pomodoro focus timer and Kanban task management with Clerk SSO and responsive interfaces.
- Developed foundations for LLM-assisted task categorization and workflow structuring from unstructured inputs.

Publications

LLM-Enhanced Learning Environments for CS: Exploring DSA with Gurukul — *IEEE Frontiers in Education (FIE)* [PAPER](#) | N

- Published an evaluation of Gurukul, a specialized LMS incorporating LLMs, RAG, and guardrails for computer-science education; conducted a user expert review with more than 40 students.

Explainable AI in Education — *IEEE SouthEastCon*

[PAPER](#) | Apr 2023

- Published research on explainable-AI approaches for education, including an algorithmic and design-choice framework that informed Gurukul's tutoring system.

Community & Writing

Technical Writing & Open-Source Community

- Kaggle Expert, ranked in the top 1% of Notebooks; contributor of data-science tutorials and competition solutions since 2020. kaggle.com/ashwinrachha1
- Technical writer covering LLMs, engineering, distributed systems, and ML inference. medium.com/@ashwin_rachha