

Anuj Sadani

Principal Engineer · AI/ML Systems · Technical Leadership

E-mail : anuj.k.sadani@gmail.com
Phone : +91 9766880844
Website : anujsadani.in
LinkedIn : linkedin.com/in/anujsadani
Github : github.com/asadani
Location : Pune, India

SUMMARY

Engineering leader with 17 years of software development experience, including the last 4 years focused on applied machine learning and generative AI systems. Leads technical direction and people development while staying hands-on with production AI architecture, from model routing and fine-tuning to agent tooling and evaluation. Experience spans document intelligence, distributed systems, cloud platforms, and engineering standards.

CORE EXPERTISE

Technical Leadership

Technical roadmaps, architecture ownership, hiring, mentoring, performance development, team structure, cross-functional collaboration, engineering standards

Applied AI Systems

Generative AI, Retrieval-Augmented Generation (RAG), agentic systems, Model Context Protocol (MCP), LLMs, SLMs, Vision-Language Models (VLMs), model evaluation, context and prompt engineering

Agent and LLM Frameworks

LangChain, LangGraph, LlamaIndex, CrewAI, PydanticAI, DSPy

Models, Serving and Routing

AWS Bedrock, Amazon SageMaker, vLLM, Ollama, LiteLLM, OpenRouter, multi-provider model routing and fallback

Machine Learning and Data

Fine-tuning (LoRA, QLoRA), quantization, synthetic data generation, YOLO, OCR, spaCy, Databricks

Engineering and Platform

Python, FastAPI, Flask, microservices, AWS, Docker, Kubernetes, PostgreSQL (RDBMS), MongoDB (NoSQL), vector databases

EXPERIENCE

Infrdd

Bengaluru, India · Jan 2021 – Present

Principal Software Development Engineer

- Led technical direction and people development for 12+ engineers across three teams, including hiring, mentoring, promotions and team restructuring.
- Own the generative AI and machine learning architecture for a document extraction platform processing approximately 1M pages per day, at 1–4¢ per page depending on document class.
- Designed and led implementation of a hybrid machine learning and LLM architecture that routes routine workloads through machine learning models and reserves LLM inference for harder cases, reducing operational cost by 30–40% against an LLM only approach.
- Designed a Model Context Protocol based agent tooling layer for mortgage document processing, working with underwriting subject matter experts to turn their domain rules into tools that agents call in production.
- Fine-tuned small open models for entity extraction from unstructured documents and integrated model based processing into production workflows.
- Own and evolve organization wide Python engineering practices, covering security standards, application patterns and code quality across every team that ships Python.

- Architected core generative AI and machine learning capabilities behind Infrd's recognition by Gartner, Everest Group and Deep Analysis.

Delivery Hero

Berlin, Germany · Feb 2020 – Dec 2020

Software Engineer II

- Built and scaled Python and Flask microservices supporting real time order processing within a distributed logistics platform.
- Worked with Product Owners and engineering leadership to translate business objectives into technical objectives and key results, roadmaps and delivery plans.
- Partnered with data science and business intelligence teams to integrate analytical workflows into production services.
- Contributed to operational reliability and incident management using Terraform, Helm, Grafana and New Relic.

NVIDIA

Pune, India · Aug 2009 – Jan 2020

Senior Tools Development Engineer

- Led a team of 8+ engineers building embedded automation frameworks and internal engineering tools supporting graphics processor and system on chip development across multiple product lines.
- Owned engineering tools across the full lifecycle, from requirements and architecture through implementation, deployment and long term maintenance.
- Drove root cause analysis and corrective action for production issues, and improved engineering productivity through automation first practices.
- Introduced a crowdsourcing approach that let engineers outside the core team contribute to internal tooling.

PUBLICATIONS

Tool Attention Is All You Need

Dynamic tool gating and lazy schema loading for scalable agentic workflows.

arxiv.org/abs/2604.21816 · 2026

Locale-Conditioned Few-Shot Prompting for On-Device PII Substitution

Mitigating demonstration regurgitation with small language models.

arxiv.org/abs/2605.13538 · 2026

The Great Recalibration

The 2026 pivot from generalist wrappers to application layers and industrialized AI services.

dx.doi.org/10.2139/ssrn.6071412 · 2026

PATENTS

Three pending patent applications with the United States Patent and Trademark Office.

- [19/375,298](#) System and Method for Optimizing Queries for Extracting Entities from Documents
- [19/375,285](#) System and Method for Optimizing Queries for Extracting Entities and Entity Values from Documents
- [19/432,186](#) System and Method for Extraction of Entities from Documents

EDUCATION AND CERTIFICATIONS

Bachelor of Engineering, Information Technology

2005 – 2009

Government College of Engineering, Amravati, India

Certified ScrumMaster, Scrum Alliance Credential ID 001022708

IBM Enterprise Design Thinking, Co-Creator