Anish Ghule

(+91) 90210 49314 • anishghule12@gmail.com • linkedin.com/in/anishghule
github.com/anishghule • anishghule.github.io

Founding engineer building production AI deployment systems for high-stakes workflows. Designed and shipped RISA's entire AI intelligence layer from scratch - multi-agent harness, MCP/A2A server infrastructure, and quantitative eval frameworks across OpenAI, Claude, Gemini, and TogetherAI, running 14B annual RPA actions in production healthcare. Strong in technical scoping, rapid prototyping, cost/performance tradeoffs, and shipping reliable systems with auditability and interpretability.

Selected Impact (AI Deployment)

- Led RISA's end-to-end AI deployment lifecycle (0→1), from technical scoping and architecture to production rollout of agentic systems in healthcare workflows.
- Delivered >30% zero-touch prior-auth workflows and 10× order-volume scaling without linear headcount growth.
- Avoided an estimated $420K/year by pivoting from in-house LLM hosting to API-based architecture and vendor-agnostic model routing.
- Reduced order ingestion from 75 to 7 minutes, saving 500-1,000+ hours/year in fulfillment capacity.
- Partnered directly with community oncology CEOs to diagnose workflow inefficiencies, taking ownership of end-to-end deployment lifecycle, from scoping ambiguous problem statements to shipping agentic features.

Skills

AI / Agents: Multi-Agent Systems, MCP, A2A Protocol, Agent Harness Design, Prompt Engineering, LLM Evaluation & Benchmarking, OpenAI / Claude / Gemini APIs

Languages: Python, Kotlin, Java, C, C++, SQL

ML / Research: PyTorch, TensorFlow, Keras, Vision Transformers, MAE, DPR, BERT fine-tuning

Infra / Systems: GCP, Kubernetes, Docker, FastAPI, BigQuery, Firebase, MongoDB, Elasticsearch, Grafana, Kibana

Tools: LaTeX, Git, Selenium, Playwright, Postman

Experience

Founding Engineer, Software AI, RISA, Palo Alto, US — Jul'24 – Present

- Designed and shipped a **multi-agent orchestration stack** using A2A with asyncio-based async message passing; built router-to-specialist delegation for PDF parsing and FHIR transformation, enabling parallel pipelines and structured handoffs.
- Built an **MCP server** with JSON-RPC handshake and permission-scoped tool discovery so agents only receive task-relevant schemas, eliminating static prompt stuffing and reducing token overhead.
- Architected a **vendor-agnostic AI Agent Factory** with runtime switching across GPT, Claude, Gemini, and TogetherAI; normalized provider I/O via adapters and enabled hot-swap failover and cost-aware model selection.
- Created **quantitative evaluation and benchmarking framework** across latency, accuracy / zero-touch rate, latency, and cost per million tokens; failure-time screenshots and DOM dumps to cut debug cycles by 60%.
- Prototyped and deployed an **LLM-powered Workflow Builder** with Magic Selectors and NoVNC remote-browser capture, enabling adoption among non-engineers for generating workflow configs and off-loading engineering dependency for 10-12 configs/month.
- Implemented **HITL, audit-trails, and observability** for automated decisions through masked BigQuery logs, customer-facing dashboards, and permission-scoped tool access, ensuring human oversight and auditability at scale.
- Stabilized 9-microservice internal platform via HPAs, zombie-process RCA, and cleanup schedules; improved fulfillment throughput and reduced ingestion from 75 to 7 minutes.

Stack: Python, OpenAI / Claude / Gemini APIs, MCP, A2A, FastAPI, Kubernetes, GCP, Pydantic, Selenium, Playwright

Software Developer (SDE), Sprinklr, Gurugram, India — Jan'24 – Jun'24

- Implemented resiliency checks for CPU under/over-utilisation in Kubernetes deployments with Teams webhook alerts; identified issues in 60%+ of deployments, impacting 40% in 95th-percentile replica-count clusters.
- Developed jobs to detect Elasticsearch shard mapping discrepancies and fixed missing-replica bugs; added coordinating-node logging for search requests across ES clusters.
- Diagnosed and fixed 15+ infrastructure issues using jstack, heap dumps (Eclipse MAT), Grafana, Kibana, and KEDA HPA analysis.

Stack: Java, Python, Spring Boot, Kubernetes, Elasticsearch, Grafana, Kibana

Research Intern, Indian Institute of Science, Bangalore, India Mar'22 – Dec'23

- Designed a novel ViT + MAE architecture for micro-expression emotion recognition; curated a multi-angle facial dataset and outperformed SOTA by +11% on SAMM and CASME2.
- Conducted comparative analysis of CNN baselines (ResNet, VGG-16) vs Vision Transformers for classification accuracy.

Stack: Python, PyTorch, TensorFlow, Keras, Vision Transformers, MAE
Supervisor: Dr. Punit Rathore

Research and Development Intern, Soroco, Bangalore, India Aug'22 – Apr'23

- Built a rule engine to summarize 4-10 action events into single representative events, deflating data size to <25% and generating synthetic training data for behavior analysis using heuristics and GPT-3.
- Anonymized sensitive data by mapping words to domain categories, supporting public dataset creation and safer downstream LLM training pipelines.

Software Engineering Intern, Sprinklr, Gurugram, India May'23 – Jul'23

- Deployed storage-analysis APIs to identify unused Elasticsearch fields, reducing storage costs 30%; built Azure Cost Analyzer with 12 drill-down views for anomaly detection and forecasting.

Research Intern, CEERI, Chennai, India May'22 – Jul'22

- Implemented Eulerian Video Magnification for remote plethysmography (rPPG) with spatio-temporal decomposition, 0.8–1.0 Hz bandpass, and 20× colour / motion magnification on UBFC-RPPG and CEERI proprietary datasets.

Supervisor: Dr. Madan Lakshmanan

Projects

Multi-Retriever Closed-Domain QA, BITS Pilani, India Aug'22 – Dec'23

- Authored a closed-domain QA pipeline (HayStack) with dual sparse/dense DPR retriever, FARMReader, and Re-Ranker; fine-tuned BERT on CORD-19, improving F1 by +9.6% (72.0% → 81.6%) on CovidQA benchmark.

Supervisor: Dr. Kamlesh Tiwari

Text-to-Image Generation, BITS Pilani, India Jan'23 – Apr'23

- Designed a GAN-based pipeline for high-fidelity Indian facial image generation from text prompts; curated 1,500+ annotated images to fine-tune StyleGAN, TediGAN, and CLIP.

Supervisor: Dr. Kamlesh Tiwari

Conversational AI & User Experience, BITS Pilani, India Aug'23 – Dec'23

- Conducted a field study on conversational AI agents; comparative analysis of Regression, SVM, and XGBoost for entity identification in CFPB consumer complaints to study LLM-human interaction quality.

Supervisor: Dr. Virendra Singh Nirban

Education

B.E. Computer Science, Birla Institute of Technology and Science, Pilani, India **2020 – 2024**
CGPA: 9.3/10.0 Relevant Courses: Deep Learning, Probability & Statistics, Data Structures, Operating Systems, Database Systems, Computer Networks, Compilers, Theory of Computation

Maharashtra State Board, Maharashtra, India **2018 – 2020**
Percentage: 96.15%