

Yash Rawal

+91-9399159685 | yashrawal987@gmail.com | linkedin.com/in/rawal-yash | github.com/YashRL | Surat, Gujarat, India

Professional Summary

Applied AI & Machine Learning Engineer serving a **1M+ user scale** by architecting enterprise agentic platforms and ERP AI infrastructure across 120+ organizations. Specialized in core model training, embedding optimization, and low-latency agent execution frameworks. Active contributor to Spring AI, bridging machine learning models with high-throughput production systems.

Technical Skills

Languages & Backend: Python, Java, SQL, Spring Boot, FastAPI, PostgreSQL, SAP HANA, Data Engineering

AI & ML Engineering: Machine Learning, Model Training & SFT, Agentic AI, MCP, Vector Embeddings

AI & ML Frameworks: PyTorch, Spring AI, LangGraph, LangChain, LlamaIndex, CrewAI, AutoGen, vLLM, SAP Build, SAP Joule

Platforms & Models: Azure OpenAI, SAP AI Core, Amazon Bedrock, Google Vertex AI, Claude, Gemini, Qwen, Meta Llama & Open Weights

Infrastructure & Security: Docker, Kubernetes, Cloud Foundry, Linux, CI/CD, LLM-as-Judge, Red Teaming, PII Sanitization (BIO NER)

Experience

TalenTeam Ltd., UK (via Bigscal Technologies)

AI/ML Engineer

Oct 2024 - Present

Remote

- Deployed a production-scale agentic platform and multi-agent systems serving **1M+ active ERP users** across 120+ enterprise tenants.
- Designed a capability contract layer to execute nested profiles inside a single ReAct loop, reducing execution overhead by 3x and token consumption by 70%.
- Trained domain-specific **Qwen2.5-7B** models on A100 GPUs and fine-tuned E5-large-v2 embeddings using PyTorch to optimize query intent mapping.
- Developed the Prompt Lab evaluation control plane to collect reasoning trajectories and human preference annotations, establishing a scalable **Human Feedback (RLHF) Pipeline**.
- Implemented a hierarchical RAG pipeline indexing 10,000+ technical documents using PostgreSQL/PGVector, achieving a 90%+ Recall@10 retrieval rate.
- Secured agentic workflows against OWASP LLM vulnerabilities via runtime tool monitoring and trained token-classification models for automated PII redaction.

Healthray (Bigscal Technologies)

AI/ML Engineer Intern

Apr 2024 - Sep 2024

Surat, India

- Engineered clinical summarization pipelines for 116K+ medical records with 4-bit quantized Llama 3, and designed real-time WebSocket audio streaming for Indic ASR (Whisper, Vosk, NeMo).
- Researched audio digital signal processing (STFT, Mel-spectrograms, MFCC feature extraction via librosa/torchaudio) and CNN audio architectures (VGGish, ResNet), compiling custom Indic LJSpeech-format datasets.

Open Source & Projects

Spring AI (spring-projects/spring-ai) | Open Source

Merged PR #5711 to propagate reasoningContent from AWS Bedrock and OpenAI-compatible models, and resolved Converse API prompt-caching TTL latency consistency.

Java, Spring Boot

Modular Agent Harness & Memory System ("Brain") | Open Source

Principal Creator & Architect of an open-source plug-and-play agentic framework designed as modular, composable building blocks for flexible enterprise workflows, implementing a formal Java-based ReAct loop (Thought-Action-Observation) with step budgets, tactical pgvector tool routing, and dynamic MCP settings caching.

Java, Spring Boot, PostgreSQL, WebSockets

AI Agent Security Red Teaming - OpenAI | Security Researcher

Designed adversarial prompt-injection pipelines targeting guardrails to evaluate security boundaries across 2,000+ agent test cases.

Python, Kaggle

Education

Master of Computer Applications (Artificial Intelligence & Machine Learning) - Parul University

2022 - 2024

Certifications

Microsoft Certified: Azure AI Fundamentals (AI-900)

Microsoft

Microsoft Certified: Azure Fundamentals (AZ-900)

Microsoft