

KAUSHAL KUMAR PRAJAPATI

Staff AI/ML Engineer | AI Platforms, Agent runtime, Applied AI | 8+ Years

+91 9862949885 | kaushalcse07@gmail.com | linkedin.com/in/kforcode | github.com/kforcodeai

PROFESSIONAL SUMMARY

Staff AI/ML Engineer who builds the system layer that makes AI products actually work - **runtime**, **harness** and **headless developer platforms** at HighLevel, **workflow**, **post-training personalization** and **eval gates** for compliance agents at Zania, a **unified Document AI extraction stack** at Intellect. 8+ years turning research-grade models and ideas into production platforms, with publications at **EMNLP 2024** and **Elsevier ASOC**.

TECHNICAL SKILLS

AI Platforms: AI agents, agent harnesses, MCP, model routing, tool/function calling, hybrid RAG, memory, long-context systems, Voice AI, copilots

Model Quality & Serving: Evals, prompt-optimization, post-training, LoRA/QLoRA, PEFT, vLLM serving, Document AI, OCR

Engineering: Python, TypeScript, PyTorch, Hugging Face, FastAPI, Node.js, SDK/CLI design, distributed systems, observability/SLOs, reliability

Cloud & Data: AWS, GCP, Azure, Docker, Kubernetes, PostgreSQL, Redis, OpenSearch, vector/hybrid search, CI/CD, LLMops

EXPERIENCE

HighLevel

Dallas, TX | Remote

Staff Software Engineer, AI Platform

Jul 2025 - Present

- Set platform direction and architected the **unified AI runtime and agent harness**, the shared execution layer for **6+ AI product surfaces** including Voice AI, Conversation AI, Agent Studio, Ask AI, Ads Copilot, and Email Agent. Standardized model routing, tool authorization, long-context state, memory, and provider fallback.
- Scaled production AI execution to **15M+ conversations** and **800B+ tokens per trailing quarter**, serving **250K+ active locations** at **99.9% success rate** and supporting **70% YoY growth** in AI adoption.
- Ideated, built, and own HighLevel's **headless developer platform** - `ghl-mcp`, `ghl-agent-sdk`, `ghl-cli`, `ghl-plugins` - exposing **600+ public operations over MCP** so customers connect via Claude, ChatGPT, Codex, Cursor, or any standard agent through OAuth and **automate end-to-end with no UI**.
- Standardized how internal teams build AI agents by turning runtime primitives into **reusable SDKs**, plugin contracts, evaluation gates, and safety patterns adopted across product surfaces.
- Drove **4.7M+ production voice conversations** at **99.9%+ success** under **real-time call constraints** by designing the Retell-based LLM execution path with OpenAI STT/TTS, custom-LLM backends, in-call tool execution, latency-aware routing, and provider fallback.

Zania AI

San Francisco, CA | Remote

Applied AI Engineer, Compliance Agents

Nov 2024 - Jun 2025

- **Retained the largest enterprise account** by shipping **Auto Compliance Agents** for SOC 2, NIST, HIPAA, and third-party risk assessment (TPRA) workflows with auditor-ready citations.
- Improved evidence-retrieval **F1@10 from 86% to 97%** despite sparse, control-specific evidence by combining hybrid retrieval, reranking, control-aware chunking, and citation-span ranking to increase auditor trust in generated responses.
- Built **post-training and evaluation release gates** using synthetic control-evidence pairs, hard negatives, rejection sampling, and held-out audit QA to measure faithfulness, citation accuracy, hallucination rate, and control coverage.
- Fine-tuned Qwen2.5-7B, Llama-3.1-8B, Phi-3-mini, and Gemma-2-9B with Hugging Face, LoRA/QLoRA, PEFT, and 4-bit quantization; deployed **cost-aware routing** between open SLMs and frontier LLM APIs.

Intellect AI, Intellect Design Arena

Bengaluru, India | Remote

Lead Machine Learning Engineer

Sep 2021 - Nov 2024

- Led commercialization of **Underwriting Copilot**, a multi-agent Document AI system for intake, loss-run analysis, policy comparison, and underwriter decision support, driving **4x sales growth** and **6 enterprise client onboardings in one month**.
- Replaced **23+ document-specific models and services** with one **layout-aware extraction architecture** using OCR tokens, bounding boxes, schema-aware decoding, and task-specific losses, reducing maintenance burden for new document types.
- Cut AWS Textract spend by **70%** while scaling layout-based chunking and OCR microservices for **6M+ PDFs** across underwriting, claims, and compliance workflows.
- Improved zero-shot extraction **F1 from 73% to 79%** and cut retraining iterations by **60%** through instruction tuning, PEFT/LoRA adapters, synthetic document generation, and schema-constrained decoding.

Eagle Labs, First American India

Bengaluru, India | Remote

Research & Development Engineer

Jul 2018 - Sep 2021

- Reduced title-insurance claims by **17%** with **parcel-level risk models** using GIS layers, claims APIs, Google Maps, and property attributes, improving risk scoring through ensemble geospatial learning.
- Built **Plymouth**, a reusable document-classification framework for **2M+ documents** across **1,400+ classes**, standardizing training, inference, and monitoring pipelines across teams.

SELECTED PUBLICATIONS

- **EMNLP 2024** - De-Identification of Sensitive Personal Data in IIT-CDIP-Derived Datasets: layout-aware QA ensemble and OCR-redaction pipeline. aclanthology.org/2024.emnlp-main.1198
- **Elsevier ASOC** - Parameter-Efficient Fine-Tuning for Hospital Discharge Summarization: PEFT fine-tuning of 3B/7B LLMs. sciencedirect.com/.../S1568494624003053
- **Elsevier ASOC** - DoRA+: Enhancing Weight-Decomposed Low-Rank Adaptation for Hospital Discharge Paper Summarization. sciencedirect.com/science/article/pii/S1568494626013098

EDUCATION

National Institute of Technology, Arunachal Pradesh

Arunachal Pradesh, India

B.Tech, Computer Science & Engineering | CGPA 8.68

2014 - 2018

Summer Research Fellow, SRTF at IIT Guwahati (2017): biomedical text-mining for drug entity recognition, **improving F1 from 60% to 78%**.