PRASUN KUMAR

AI/ML TECHNICAL LEAD & ARCHITECT (GENAI, NLP, LLMOPS)

Pune, India 411046 ♦ +918793147065♦cehprasunsinha@gmail.com

♦ **LinkedIn:** https://www.linkedin.com/in/prasun-kumar-1708/ **WWW:** https://prasun0512.github.io/resume/

Summary

AI/ML Technical Lead & Architect with 8+ years of experience designing and delivering secure, scalable AI/GenAI solutions across enterprise and regulated workflows. Strong hands-on expertise in LLMs (fine-tuning, prompt engineering), retrieval-augmented generation (RAG), embeddings, multi-label classification, evaluation/monitoring (MLflow, Databricks), and production-grade NLP systems. Proven ability to lead teams, drive architecture decisions, collaborate with cross-functional stakeholders, and build repeatable delivery practices (templates, playbooks, dashboards, documentation). Experienced with privacy-aware processing (PHI/PII masking), compliance governance (HIPAA/SOX), and measurable outcomes. Quickly advanced from Python automation to ML/AI roles within one year of starting career, demonstrating rapid learning and leadership potential.

Core Competencies

- **GenAI Architecture:** retrieval-augmented generation design, vector search & hybrid retrieval, grounding strategies, prompt frameworks
- **LLMOps / MLOps:** model lifecycle management (MLflow tracking, registry), versioning, evaluation suites, drift/quality monitoring, CI/CD
- **NLP & Applied ML:** behavior classification, named entity recognition (NER), document intelligence, semantic search, toxicity detection
- **Secure AI Delivery:** PHI/PII data handling, compliance-aware pipelines (HIPAA, SOX), audit-ready documentation and reporting
- **Technical Leadership:** end-to-end solution design, architecture ownership, code reviews, mentoring, stakeholder alignment, Agile delivery
- **Product Enablement:** user demos and training, onboarding playbooks, reusable prompt and content libraries, feedback loop integration

Technical Skills

- **Languages:** Python
- **ML & NLP:** Transformers (Hugging Face), SentenceTransformers, embeddings, NER, text classification, model calibration, evaluation frameworks
- **GenAI:** retrieval-augmented generation (RAG), prompt engineering, few-shot prompting, tool-augmented LLM workflows, fine-tuning (LoRA/QLoRA, PEFT)
- **Frameworks/Libraries:** PyTorch, TensorFlow/Keras, scikit-learn, Hugging Face Transformers, LangChain, spaCy, NLTK, OpenCV
- **Platforms:** AWS, Azure (Azure AI/OpenAI), Databricks
- **MLOps & Data:** MLflow, Apache Airflow, Elasticsearch, SQL/NoSQL, Vector databases (FAISS, Pinecone)
- **Tools & Processes:** Docker, Git, CI/CD, Agile/Scrum methodology, technical documentation, stakeholder communication

Professional Experience

Harbinger Group — Pune, India
Associate Technical Specialist (AI/ML Technical Lead & Architect) (Jul 2024 – Present)

- Lead a team of ML engineers to develop scalable NLP and GenAI solutions for enterprise clients, owning architecture decisions, quality standards, and cross-functional delivery in an Agile environment.
- Designed and deployed multilingual NLP pipelines, improving sentiment analysis and NER accuracy by ~50% through advanced preprocessing and transformer fine-tuning.
- Built retrieval-augmented generation (RAG) solutions to improve domain-specific answer relevance for enterprise knowledge bases; incorporated stakeholder feedback loops to iteratively refine outputs and align with business needs.
- Developed an LLM evaluation and monitoring pipeline on Databricks with MLflow, tracking 10+ model versions. Enabled comprehensive comparison of model quality, drift, and performance through automated metrics logging and interactive dashboards for data-driven decision making.
- Engineered a memory-optimized LLM inference framework using adapter fusion and 4-bit quantization, reducing GPU memory usage by ~70% and enabling deployment of high-parameter models in production on limited hardware.
- Mentored junior engineers and promoted best practices through code reviews, technical workshops, and reusable templates, fostering a culture of excellence and continuous learning.

Harbinger Group — Pune, India
Senior Software Engineer (Apr 2021 – Jun 2024)

- Built an LLM-powered document analysis pipeline for medical insurance claims, integrating OCR and GPT-based NLP to extract PHI entities, classify case details, and generate summaries. Automated PHI masking and validation to ensure HIPAA compliance, boosting processing efficiency by ~80% and significantly reducing manual workload.
- Improved system performance and scalability by implementing multi-threading and asynchronous processing in Python, reducing key API response times by ~80% and enhancing user experience.
- Integrated AWS AI services (Amazon Comprehend, Kendra) and open-source NLP tools into client applications to enable intelligent search and document classification features, accelerating the adoption of machine learning capabilities in product roadmaps.
- Acted as interim tech lead on AI projects, guiding solution design and mentoring junior developers to deliver high-quality results.

Extentia Information Technology — Pune, India
Software Developer (Nov 2019 – Apr 2021)

- Built a semantic search chatbot using embeddings and similarity matching, improving domain Q&A accuracy by ~70%, and delivered custom NLP pipelines (rule-based + ML NER) to extract key entities and standardize downstream processing.

- Implemented a secure multilingual translation service using speech-to-text and translation models (e.g., Whisper) for real-time English–Hindi communication, and built stable, maintainable APIs with secure text handling and privacy-aware design.

Symantec Software India Pvt. Ltd. — Pune, India
Associate IT Applications Specialist (Jan 2017 – Oct 2019)

- Automated IT operations tasks (server patching, backups, and SOX audit reporting) using Python scripting and scheduling, improving process efficiency by ~70% and reducing manual compliance workload.
- Developed a Python/Flask dashboard to monitor ServiceNow tickets and system performance, improving SLA adherence through real-time alerts and facilitating proactive incident management.
- Built a strong foundation in Python automation and secure coding practices, facilitating a rapid transition into ML engineering roles.

Selected Projects

Behavior Scoring Engine – *Multi-label Email Classification (80+ behavior tags)*

- Designed and implemented a configurable classification pipeline on Azure Databricks and local environments, enabling controlled experimentation and seamless deployment of new behavior models.
- Built local ML baselines using SentenceTransformers embeddings (multi-qa-mpnet-base-dot-v1) with per-behavior classifiers (e.g., Random Forests), achieving high precision/recall on unseen data for the majority of behavior categories.
- Improved classification of edge-case behaviors by introducing a shallow neural network with class-weighted training, early stopping, and threshold calibration, boosting recall for under-represented classes.
- Benchmarked advanced prompt-based strategies using Azure OpenAI GPT-4 (few-shot examples, checklists, chain-of-thought reasoning) as a comparison and fallback for low-confidence predictions, combining LLM versatility with deterministic models.
- Delivered a comprehensive evaluation suite with train/validation/test splits and per-behavior metrics (precision, recall, F1), along with reproducible model artifacts and configuration files to ensure transparency and facilitate future enhancements.

Restaurant Chatbot – *Document Q&A System with RAG & Voice Integration*

- Developed a chatbot that answers questions based on proprietary recipe documents by extracting content via OCR and indexing knowledge into an Elasticsearch backend for semantic retrieval.
- Upgraded the system with a LangChain-based retrieval-augmented generation (RAG) layer using GPT-4, which improved answer accuracy, provided source-grounded responses, and handled follow-up queries more gracefully.
- Integrated Amazon Alexa as a voice interface for the chatbot, enabling hands-free access to recipe Q&A in kitchen environments and expanding the user reach through voice-enabled technology.

Medical Insurance Claims Automation – *Document Intelligence for Healthcare*

- Architected an end-to-end automation pipeline for US medical insurance claims processing, combining OCR, NLP, and RPA to handle high volumes of claim PDF documents with minimal human intervention.
- Implemented Protected Health Information (PHI) detection and masking in Python to sanitize data before LLM processing, supporting privacy compliance (HIPAA) in a machine-assisted review workflow.
- Leveraged GPT-4 for context-aware entity extraction (e.g., patient info, diagnosis codes) and fine-tuned GPT-3.5 on labeled case data to automatically classify claim documents into case type, subtype, and category, significantly reducing manual sorting and error rates.

Multi-Model LLM Evaluation Pipeline – *Model Performance Benchmarking & Monitoring*

- Developed an automated evaluation framework on Databricks, integrated with MLflow, to benchmark multiple large language model versions in parallel and track their performance over time.

- Built a Python-based automation + ETL pipeline to evaluate and monitor 10+ LLM variants, logging key metrics (quality, latency, cost) and tracking drift for production-like performance.
- Created dashboards for model comparison and deployment readiness, and supported continuous improvement using LLMOps practices like adapter merging and incremental updates.

Education

- **Master of Computer Applications (Artificial Intelligence)** – Pune University, 2017
- **Bachelor of Computer Applications** – Sikkim Manipal University, 2014

Certifications

- Executive Post Graduate Program in Machine Learning & AI – IIIT Bangalore, 2024
- Certified Information Security & Ethical Hacking (C|EH v8) – 2018
- Agile Project Management Certification – 2020

Awards & Recognition

- **Technical Star Award** (Harbinger Group) – Awarded for exceptional performance in multiple quarters (May 2023, Jul 2023, Jun 2024)
- **Superstar Award** (Harbinger Group) – Jun 2023
- **Team Player Award** (Harbinger Group) – Apr–May 2023