

Himanshu Kumar

AI/ML Engineer | Agentic AI | RAG | Cloud-Native AI Platforms

+91 89358 28600 | himanshu.kumar0012@gmail.com | LinkedIn | GitHub

SUMMARY

AI Engineer with experience building agentic AI systems, RAG workflows, document intelligence pipelines, and cloud-native AI platforms for BFSI and enterprise compliance use cases. Skilled in agent orchestration, Kubernetes deployments, LLM inference infrastructure, multi-tenant SaaS and on-premise implementations, CI/CD automation, and client-facing solution delivery across AWS environments.

EDUCATION

Indian Institute of Technology, Kharagpur 2020 – 2025
M.Sc. Economics, 7.9 / 10.0 CGPA Kharagpur, India

TECHNICAL SKILLS

Languages: Python, SQL, C
AI / ML: LLMs, RAG, Agentic AI, LangGraph, LangChain, LlamaIndex, Transformers, SentenceTransformers, Prompt Engineering, Document Intelligence, OCR, Re-ranking, NLP
Frameworks: FastAPI, Pydantic, PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, LightGBM, Pandas, NumPy, Matplotlib, Seaborn, NLTK, SpaCy, BeautifulSoup
Cloud / Infra: Kubernetes, Docker, AWS EC2, AWS S3, AWS Bedrock, GCP, Azure, Cloudflare, Linux, GitHub Actions, Argo CD, CI/CD
Data Systems: MongoDB, Redis, Qdrant, Pinecone, ChromaDB, MySQL, RabbitMQ, Trino, Bifrost
Tools: Git, VS Code, Jupyter, Anaconda, Streamlit, Power BI, Excel

EXPERIENCE

AI Engineer Jun 2025 – Present
OnFinance AI Bengaluru, India

- Architected and led Agent Studio, a LangGraph-based agentic workflow engine featuring a connected file-system, dynamic tool registry, and skill-prompt routing, enabling configurable, scalable AI pipeline orchestration across multiple product verticals; comparable in design to Codex or ChatGPT Agent Builders.
- Engineered high-volume LLM, RAG, and document intelligence pipelines processing 100M+ tokens/month using LiteLLM, AWS Bedrock (Claude 4.5/4.6 Sonnet, Haiku 4.5), OCR, PyPDF, FusionQuery search, and re-ranking for large-scale financial and regulatory document ingestion.
- Owned AWS Kubernetes infrastructure for 20+ enterprise clients, managing shared clusters with dedicated node pools for databases, application services, and LLM inference via Bifrost; operated production data systems across MongoDB, Redis, Qdrant, RabbitMQ, and Trino with CI/CD via GitHub Actions and Argo CD.
- Drove end-to-end client delivery across 20+ enterprise accounts, including NSE, BSE, Axis Finance, BharatPe, PayU Finance, and FedFina etc., covering solution finalization, tenant provisioning, integration planning, infra sizing, network provisioning, and sign-off across SaaS and on-premise deployments on AWS.
- Secured \$260K+ in infrastructure credits — \$10K in AWS credits for POC deployments and \$250K via Cloudflare Startup Program — engaging directly with Cloudflare engineering stakeholders on organizational-level integration planning and implementation.
- Contributed to securing \$4.2M Pre-Series A from PeakXV Partners through technical demos and executive presentations of the product; deeply involved in early product development including LangGraph agentic workflows over SEBI, RBI, and IRDAI regulatory data and clause-level RAG pipelines over compliance circulars.

LLM Trainer May 2025 – Jun 2025
Outlier Remote

- Performed RLHF preference-ranking and pairwise comparison of LLM-generated responses across coding and reasoning tasks, providing evidence-based justifications for quality ratings.
- Authored structured feedback and failure-mode annotations used to inform reward-model signals for downstream reinforcement learning.
- Reviewed code-generation outputs for correctness, efficiency, and adherence to task specifications as part of domain-specific model evaluation.

LLM Trainer Oct 2024 – Dec 2024
Soul AI Remote

- Evaluated and benchmarked multi-model LLM outputs across STEM and non-STEM domains for factual accuracy, reasoning correctness, and response quality as part of human-feedback data pipelines.
- Audited AI agent tool-calling behavior, validating function-call schemas, parameter correctness, and API integration reliability in code-execution traces.
- Engineered and optimized prompts using zero-shot, few-shot, and chain-of-thought techniques to improve model reasoning consistency and output quality.

AI Intern May 2024 – Jul 2024
Schneider Electric Bengaluru, India

- Built a GenAI-powered internal knowledge bot for developers, applying retrieval-augmented generation over internal documentation and codebases to surface contextual, source-grounded answers and code references.
- Prototyped GenAI solutions across knowledge management, automated testing, and code-assistance workflows, evaluating LLM integration patterns and embedding-based retrieval for internal developer tooling.
- Designed AI-driven knowledge management workflows, including semantic indexing and vector search over internal documentation, to improve discoverability of engineering knowledge org-wide.

PROJECTS

Hierarchical Document Parsing & Agentic Traversal Engine Sep 2024 – Nov 2024
Document Intelligence

- Parsed documents into a hierarchical tree (pages → sections → bounding boxes) using Claude Haiku 4.5 vision calls to infer section boundaries and generate per-node natural-language descriptions across diverse layouts.
- Built an agentic traversal engine with tree-navigation tools (parent/child/sibling traversal, section-label search, bounding-box filtering) that cut Vision LLM API calls by 60%+ via cached tree state, re-querying only ambiguous nodes.