

Hemanth S

AI Engineer

9353575169 | hemanths01234@gmail.com | Bengaluru | [LinkedIn](#)

SUMMARY

AI Engineer with hands-on, end-to-end ownership of 20+ production-grade GenAI solutions across AWS, Azure, and GCP — from stakeholder discovery and requirements gathering through MVP delivery, deployment, and monitoring. Experience spans LLM-powered pipelines, agentic AI, and RAG architectures, including designing AI guardrails, system rules, and reusable prompt/workflow frameworks to standardize and improve LLM reliability. Delivered measurable business impact: a 40% reduction in LLM API costs and a 30% latency improvement through architecture and model selection trade-offs. Builds reliable, observable, and secure production systems with CI/CD (GitHub Actions, GitLab CI/CD), testing, and authentication/access-control best practices. AWS, GCP, and Databricks certified in Machine Learning and Generative AI.

CERTIFICATIONS

- AWS Certified Machine Learning – Associate
- GCP Generative AI Leader
- Databricks Certified Generative AI Engineer Associate

EXPERIENCE

Associate AI/ML Engineer | Rapyder Cloud Solution

Jan 2025 – Present

- Served as primary engineer driving end-to-end delivery on 20+ AI/ML initiatives across AWS, Azure, and GCP — owning stakeholder discovery and requirements gathering, and translating business needs into scalable technical solutions from MVP through production deployment
- Built GenAI solutions across diverse industries — Manufacturing, Healthcare, E-commerce, Gaming, and Fintech — mapping each client's workflow and constraints before selecting the right architecture (RAG, multi-agent orchestration, voice/chat interfaces) for the use case
- Reduced LLM API costs by 40% through model selection and right-sizing (routing simple queries to smaller/cheaper models), prompt optimization to cut token usage, and response caching for repeated queries
- Reduced system latency by 30% through parallelized data pipelines, optimized database read/write patterns, and caching frequently accessed data — improving throughput and turnaround time for end users

Data Science Intern | Accurate Info Solution

Aug 2023 – Nov 2023 · 4 mos

- Gained foundational experience in data extraction, cleaning, and basic visualization to support introductory data analysis tasks

KEY PROJECTS

EdTech Platform — AWS · EdTech

Built and deployed a production-grade EdTech platform on AWS serving 30,000+ end users/month, using Bedrock (Claude 4 Sonnet) to power personalized note generation, an AI job search assistant, AI voice-based interview assessments, multi-template resume generation, personalized learning roadmaps, and a course-specific RAG chatbot. Delivered a more personalized, end-to-end learning experience that improved student engagement and career readiness at scale.

Multilingual Post-Call Analytics — AWS · Banking/Fintech

Developed a production-grade multilingual post-call analytics solution on AWS for a major Indian bank, processing 6,000+ sales-agent calls/month with 90%+ analysis accuracy using Amazon Transcribe, Comprehend, Bedrock, and CloudFront to detect sentiment, fraud signals, and call-resolution status, backed by an analytics dashboard. Enabled the client to identify fraud risks and unresolved customer issues faster, reducing manual review effort and improving service quality oversight.

WhatsApp AI Agent for Retail — Azure · Retail Enterprise

Designed and deployed a production-grade WhatsApp-integrated AI chatbot on Azure for a retail enterprise, building an Azure AI Foundry agent that orchestrated RAG-based product/knowledge search, NL-to-SQL queries, user verification, and CRM-integrated ticket creation. Gave the client's customers 24/7 self-service support, reducing response times and easing load on their support team.

Personalized Recommendation & Content Moderation — AWS · Esports

Architected a production-grade recommendation and content moderation system on AWS for a leading Indian esports gaming platform, using Amazon Personalize to generate personalized game recommendations from user engagement and purchase data, and Amazon Rekognition to moderate live-stream gaming video — flagging and filtering abusive, NSFW, or spam content in real time. Helped the client boost user engagement through relevant recommendations while maintaining a safe, trusted platform environment.

RAG Chatbot for Retail Website — *GCP · Retail Enterprise*

Delivered a production-grade RAG chatbot on GCP for a retail platform's website, using Vertex AI to build a highly scalable, cost-efficient, low-latency conversational assistant. Enabled the client's end users to get instant, accurate answers on-site, reducing dependency on manual support channels and improving the overall customer experience.

SKILLS

Languages & Backend: Python · FastAPI · REST APIs · Microservices · Git

Testing, Observability & Auth: Unit Testing (Pytest) · Debugging · CI/CD (GitHub Actions, GitLab CI/CD) · Observability (LangSmith, OpenTelemetry, CloudWatch) · Authentication & Access Control (OAuth, JWT, IAM)

AI/GenAI: LangGraph · LangChain · LlamaIndex · HuggingFace · PyTorch · TensorFlow · GPT · LLaMA · Gemini · Claude (Anthropic) · RAG · Vector Databases (Pinecone, FAISS, pgvector) · Embeddings & Tokenization · Agentic AI · Multi-Agent Orchestration · AI Guardrails & System Rules · Reusable AI Skills & Prompt Frameworks · Function/Tool Calling · MCP (Model Context Protocol) · Prompt Engineering · Fine-tuning · NLP-to-SQL · MLOps · OpenSearch/Elasticsearch · Cursor · Claude Code

Cloud & DevOps: AWS · Azure · GCP · Docker · Kubernetes (working knowledge) · Terraform · CDK · Bedrock · Lambda · Rekognition · SageMaker · ECS · EC2 · Azure AI Foundry · Azure Functions · Azure AI Search · Azure OpenAI · GitLab CI/CD

Databases: PostgreSQL · MongoDB · Redis · Aurora pgvector · Pinecone · FAISS

EDUCATION

Bachelor of Engineering, Computer Science

ACS College of Engineering, Bengaluru | CGPA: 7.57 — 2024

Pre-University Course (Karnataka State Board)

Mandavya Excellence PU College, Mandya | 83.6% — 2020

Secondary School Examination (Karnataka State Board)

Sree Saraswathi Vidhya Mandir, Bengaluru | 74.84% — 2018