Ankit Kumar Singh

Senior Data Scientist

+91-8660868651 | ankitsingh540@gmail.com | LinkedIn

PROFILE SUMMARY

Senior Machine Learning Engineer with 5 years of experience specializing in architecting and deploying large-scale, enterprise-grade Generative AI systems. Proven track record of delivering end-to-end agentic solutions, from fine-tuning LLMs for specialized tasks to deploying multi-agent RAG platforms for over 124,000 users. Expertise in designing robust, scalable workflows using frameworks on multi-cloud to solve critical business problems in BI and enterprise search.

TECHNICAL SKILLS

Core Competencies	Machine Learning, Deep Learning, Natural Language Processing (NLP), Generative AI, Large Language Models (LLMs), Agentic Frameworks(CrewAI, LangGraph), RAG, LLM application Frameworks(Langchain, LlamaIndex), Vector Databases, Model Fine-tuning (LoRA, QLoRA), Quantization (GPTQ, AWQ), Deployment Frameworks (TGI, vLLM), Evaluation and Monitoring (OpenTelemetry, Arize AI, MLFLow)
Programming & Tools	Python, TensorFlow, PyTorch, SQL, Unix
Cloud Technologies	Azure(Azure OpenAI, Storages, VMs), AWS (Textract, Bedrock, EC2, S3), GCP (Vertex AI)

PROFESSIONAL EXPERIENCE

- **Enterprise Contract Intelligence and Search Platform**
  - Architected and deployed an AI-powered search platform for a 45,000-document legal repository, successfully adopted by procurement teams to reduce contract analysis and retrieval time by over 75%.
  - Engineered an intelligent query router using a fine-tuned LLM classifier to interpret user intent. This router dynamically triages requests, either generating SQL from natural language (NL2SQL) for metadata-based queries or initiating the semantic search workflow for content-level analysis.
  - Designed a high-performance, multi-step RAG pipeline that first queries a Redis semantic cache for previously computed answers. On a cache miss, it executes a hybrid search (BM25 + Dense Vector) across an OpenSearch cluster, feeding the retrieved context to an LLM for precise answer generation.
  - Instituted a rigorous custom evaluation framework to measure the accuracy of the core decision-making logic. Achieved **95%+ precision in classifying user** queries for NL2SQL vs. Semantic Search and a **99% hit rate for cacheable queries**, validated against a golden dataset and human-in-the-loop reviews.
  - Developed a scalable offline data processing pipeline using AWS Lambda to perform Named Entity Recognition (NER) and clause extraction. This structured metadata, stored in PostgreSQL, forms the backbone of the high-accuracy NL2SQL capability and faceted filtering.

- **Conversational BI Agent**
  - Developed an advanced conversational AI agent for ad-hoc business intelligence, enabling users to query complex, structured insurance data through natural language dialogue.
  - Engineered a cyclical, stateful agentic workflow using LangGraph to manage multi-turn conversations and orchestrate complex tasks, including query planning, code generation, tool execution, and self-correction.
  - Fine-tuned a Llama-3 8B Instruct model for a high-accuracy Text-to-SQL generation task, employing Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA). This boosted query

success rates by over 15% on hard and extra-hard query complexities compared to the base model.

- Implemented a robust "reflection" and error-handling mechanism within the graph, allowing the agent to analyse failed SQL executions, correct its own code, and re-query the database, significantly improving the reliability of ad-hoc analytics.
- Spearheaded the evaluation framework for the agent, benchmarking the fine-tuned model against the industry-standard BIRD-SQL dataset. The superior accuracy was instrumental in a **successful Proof-of-Concept delivery**, leading to positive stakeholder reception and paving the way for wider tool adoption.

- **Chat with your own data (CWYD)**
  - Architected and deployed a company-wide, GenAI-powered productivity platform for **a user base of 124,000 employees**, achieving high adoption with over **12% daily active users (15,000+) and an 85% accuracy rate on core tasks.**
  - Engineered a sophisticated multi-agent system featuring a central GPT-4o-powered orchestrator that intelligently routes requests to specialized agents for document comparison (Gemini Pro), summarization, and Retrieval-Augmented Generation (GPT-4o-mini), optimizing for cost and task accuracy.
  - Designed a robust, multi-format ingestion pipeline on AWS, processing diverse unstructured data (PDF, DOCX, PPTX) and semantic chunking before generating embeddings and indexing in a Qdrant vector database.
  - Implemented a scalable and resilient backend leveraging AWS Lambda and DynamoDB for critical state management, including user session history, prompt tracking, and data lineage, ensuring observability and reliability at scale.
  - Pioneered a multi-cloud and multi-model AI strategy, integrating models from both Azure OpenAI and Google to prevent vendor lock-in, enhance system resilience, and select the most cost-effective model for each specific sub-task (e.g., routing, comparison, generation).

ONLINE PROFILES

- **GitHub**: https://github.com/ankit201
- **Kaggle**: https://www.kaggle.com/ankitsingh1996

COURSES UNDERTAKEN

- AI Agents in LangGraph (Deeplearning.ai)
- Ai Agentic Design Patten with AutoGen (Deeplearning.ai)
- Generative AI for Software Development (Deeplearning.ai)
- Large Language Models: Foundation Models from the Ground Up (LLM102x Databricks)
- Deep Learning Specialization (Deeplearning.ai)
- Nano degree in Machine Learning (Udacity)

WORK HISTORY

Employer	Title	Dates of employment
Fractal Analytics	Senior Data Scientist	March 2023 - Present
Implesys India	Machine Learning Engineer	October 2021 – March 2023
Lowe's	Analyst	September 2020 - October 2021

EDUCATION

Bachelor of Engineering (Information Science and Engineering) from Visvesvaraya Technological University in the year 2020, India.