Chirag Bansal

Scarborough, ON, Canada — (+1) 437-431-3144 — chiragbansal254@gmail.com — LinkedIn

Professional Summary

Machine Learning Engineer with 3+ years of hands-on experience building **RAG systems, Agentic AI workflows, and conversational AI assistants**. Specialized in **LLM orchestration, vector databases, and production-ready AI pipelines** leveraging LangChain, RAG agents, and RASA. Proven ability to translate business problems into deployable AI systems—delivering measurable impact such as **76% retrieval accuracy in financial RAGs** and **38% faster decision workflows**. Experienced in **model deployment, observability** across GCP and Azure environments.

Work Experience

GenAI Analyst *Jul 2025 – Present*
Scotiabank, Toronto, ON, Canada

- Architected and deployed an **Agentic RAG framework** using **LangChain, Qdrant on GCP, and Vertex AI**, enabling retrieval and summarization of treasury reports and improving analysis speed by **27%**.
- Built an **LLM-powered assistant** integrated with BigQuery and secure vector embeddings to automate financial reconciliation workflows, reducing manual review effort by **43%**.
- Designed a **real-time observability dashboard** to track API latency, token usage, and system uptime.
- Partnered with engineering and product teams to identify new use cases for **RAG-based automation**, improving overall reporting accuracy and team productivity.

Business Analyst Intern – Fraud Optimization & AI Systems *Sep 2024 – Jul 2025*
Tangerine Bank, Toronto, ON, Canada

- Developed an **Agentic RAG workflow** using **LangChain and Vertex AI** to assist fraud analysts with claim investigation, cutting query turnaround time by roughly **28%**.
- Built a **context-aware RAG assistant** connected to Qdrant and BigQuery, allowing analysts to query fraud and claims data through natural language more efficiently.
- Implemented **multi-agent orchestration** to divide tasks such as SQL lookups, summarization, and case notes generation, improving operational efficiency by **30%**.
- Designed internal dashboards for tracking API performance and model reliability, leading to more consistent **LLM response quality**.

Software Application Support Engineer *Aug 2022 – Nov 2023*
Haryana Shahari Vikas Pradhikaran, Panchkula, HR, India

- Designed and deployed a **RASA-based citizen support chatbot** leveraging transformer-based NLP models and a MongoDB backend to automate user queries and service requests.
- Built a **Computer Vision GatePass System** using **YOLOv8 and OCR** to enhance on-premise security operations and vehicle entry validation.
- Integrated both solutions into a **Django + Docker** microservice pipeline with MLflow-based monitoring for model performance and versioning.
- Collaborated with cross-functional stakeholders to define AI solution requirements and ensure reliable deployment in production environments.

Technical Skills

- **Generative AI & LLMs**: OpenAI API, Groq API, Hugging Face Transformers, LangChain, RAG Pipelines, Prompt Engineering, Model Evaluation
- **Conversational AI**: RASA, LangGraph, Streamlit Chat Interfaces, Contextual Chatbots, Intent Classification
- **Machine Learning & Analytics**: scikit-learn, XGBoost, TensorFlow, Data Cleaning, Feature Engineering, Model Training, A/B Testing
- **Data & Querying**: SQL (BigQuery, MySQL), ETL Pipelines, Data Modeling, Advanced Excel, Reporting Automation, Dashboard Design
- **Visualization & BI Tools**: Google Looker, Power BI, Tableau — experienced in designing real-time dashboards for operational insights
- **Vector Databases & Retrieval**: Qdrant, FAISS, ChromaDB — semantic search, document indexing, and query optimization

- **Cloud & Deployment**: GCP (Vertex AI, BigQuery), Azure (Promptflow, Blob Storage), Docker, MLflow, API Integration
- **Programming & Dev Tools**: Python (Pandas, NumPy, FastAPI, Flask), Git, Jira, VS Code, CI/CD, Unit Testing (pytest)

Agentic AI & RAG Projects

Secure Voice IDE – Agentic RAG Hackathon Project *Sep 2025*
Google x AI Tinkerers Hackathon, Toronto, ON

- Built a **voice-first secure IDE** using **Agent Development Kit** with **Google Vertex AI, Llama Guard 4, and DLP De-identification**, enabling speech-to-code generation with privacy-safe prompts.
- Designed a **multi-agent observability dashboard** tracking API costs, latency, and token utilization, achieving **under 800ms average latency**.
- Recognized as **Runner-up among 30+ teams** for innovative design in **secure GenAI orchestration and cost transparency**.

Financial Document Intelligence System – Production RAG Application *Aug 2025*
Scotiabank, Toronto, ON, Canada

- Architected a production-grade **Retrieval-Augmented Generation (RAG)** system for Q1 2025 financial reports using **Groq API** for inference and **Jina-v3 embeddings** for semantic retrieval.
- Deployed a scalable **Qdrant vector database** achieving **76% retrieval accuracy** and sub-second response latency.
- Automated document parsing via **Docling** for PDFs containing financial tables, ensuring complete context retention across multi-format data.

Health Buddy – Medical AI Assistant (Google KaggleX Fellowship, Cohort 4) *Aug – Dec 2024*
Google KaggleX Fellowship Program

- Fine-tuned the **Gemma 2B LLM** with LoRA adapters for healthcare data, improving clinical query accuracy to **78%**.
- Deployed the system using **Gradio + HuggingFace Spaces**, handling **1000+ real-world queries** with monitored latency and quality metrics.
- Designed explainability visualizations and confidence scoring reports to ensure compliance with healthcare AI guidelines.

Intelligent Resume Assistant – Conversational RAG System *Oct 2024*
Independent Project

- Built a multi-turn conversational AI system using **LangChain + FAISS + OpenAI APIs** for natural language understanding and contextual resume improvement.
- Achieved **87% response accuracy** through hybrid retriever optimization and prompt tuning.
- Deployed using **Streamlit + A/B testing**, enabling continuous improvement via real-time feedback loops.

Education

Postgraduate Certificate in Artificial Intelligence and Machine Learning *Jan 2024 – Aug 2025*
Lambton College, Toronto, ON, Canada GPA: 3.1/4.0

- **President**, Google Developers Student Club - Led technical workshops on ML deployment, cloud computing, and best practices for 60+ students
- **LinkedIn Student Ambassador** - Conducted 3 optimization workshops on professional networking and career development, impacting 60+ students

Bachelor of Technology in Computer Science and Engineering *July 2018 – June 2022*
Punjabi University, Patiala, Punjab, India GPA: 8.1/10.0

- **CodeChef University Chapter President** - Organized 3 major ML/AI hackathons and technical workshops for 750+ participants, fostering coding culture
- **Community Growth Head**, CodeAsylums - Led off-campus expansion initiatives and mentored 12,000+ aspiring programmers with resources and guidance