

AMBUJ KUMAR TRIPATHI

GenAI Engineer | Conversational AI | Agentic RAG Systems | LLM Fine-Tuning (QLoRA) | LLMOps

Available Immediately | Remote and Hybrid | Open to relocation

Email: kumarambuj8@gmail.com | Mobile: +91 9431 801363 | Location: Gorakhpur, UP, India

Portfolio: ambuj-portfolio-v2.netlify.app | GitHub: github.com/Ambuj123-lab

HuggingFace: huggingface.co/invincibleambuj | Docs: ambuj-rag-docs.netlify.app

PROFESSIONAL SUMMARY

GenAI Engineer with 5+ years of enterprise technology experience building production-ready AI applications. Specialized in designing **Adaptive ReAct Agentic RAG systems** featuring intelligent retrieval, Human-in-the-Loop (HITL), confidence-aware routing, live web search, and multi-channel AI delivery across Web and WhatsApp using **Meta WhatsApp Business Cloud API** with webhook-driven event architecture. Built scalable conversational AI platforms capable of **multi-domain intent routing** across Legal, Financial, Technical Documentation, and Programming assistance, processing **31,500+ chunks** extracted from unstructured legal and financial documents, including text, tables, images, and complex layouts, using intelligent document analysis and hallucination mitigation. Active open-source contributor with three published fine-tuned Indian Legal LLMs on Hugging Face, collectively achieving **5,500+ global downloads**.

KEY HIGHLIGHTS

- Built and deployed **3 enterprise-grade RAG systems** - all live and publicly accessible
- Indexed **31,500+ chunks** and **28,000+ precision vectors** across 20+ legal acts and financial frameworks
- Designed a **9-Node LangGraph Agent** implementing a **ReAct reasoning pattern** with dual-path web search activation - confidence-gated HITL fallback (<45% threshold) and direct Out-of-Scope routing - with LLM-as-judge Hallucination Guard and grounded answer generation.
- Extended the production-grade Agentic RAG platform to an **omnichannel conversational AI system** by integrating **Meta WhatsApp Business Cloud API** with webhook-driven event processing, enabling shared FastAPI/LangGraph orchestration across Web and WhatsApp interfaces.
- Implemented **Jina v3 MRL (Matryoshka Representation Learning)** embeddings - compressed 28K+ vectors from 1024 to 256 dimensions achieving **75% storage savings**
- Implemented **Langfuse observability**, **Redis semantic caching (sub-100ms)**, pybreaker circuit breakers, and system health monitoring
- Built **GDPR-compliant AI pipelines** using Microsoft Presidio + spaCy NLP, OAuth 2.0, JWT, MongoDB TTL, and **7-layer enterprise upload security**
- Published **3 Hugging Face GGUF models (1B, 3B, 8B)** using QLoRA, achieving **5,500+ downloads**
- Maintaining **4 live deployment systems** with **99.75% uptime** and **sub-300ms response times** via automated health monitoring

Core Strengths: RAG Architecture | LLMOps | ReAct Agents | Web Search Fallback | Neural Reranking | LangGraph State Machines | Multi-Turn Context | Circuit Breakers | Vector Databases (Pinecone, ChromaDB, Qdrant) | NLP | GDPR Privacy

OPEN SOURCE CONTRIBUTIONS

Open Source Projects & Published Models | Remote

October 2025 - Present

- Published production-grade Agentic RAG architecture with the LangChain developer community (**51K+ views**)
- GitHub:** Advanced RAG repository - **65 stars, 22 forks**
- HuggingFace Hub:** 3 published GGUF models - **5,600+ downloads**
- Fine-tuned Llama 3.x models (**1B, 3B, 8B**) using **QLoRA (4-bit)** on **14,543** Indian Legal QA pairs
- Published GGUF Q4_K_M models and LoRA adapters on HuggingFace - CPU-first inference via **LM Studio & Ollama**
- Built local inference workflows using LM Studio and HuggingFace ecosystem
- Conducted comparative evaluation of **Fine-Tuning vs RAG** architectures
- Built public demos and evaluation workflows for experimentation and learning

Models: 1B GGUF | 3B GGUF | 8B GGUF

FEATURED PROJECTS (ALL LIVE AND ACCESSIBLE)

Adaptive ReAct Omnichannel RAG Platform

Tech: FastAPI, React, LangGraph (9-Node), Pinecone, Qwen 3.6 Plus, Jina v3 MRL, Cohere Reranker, LlamaParse

- Built a **9-Node LangGraph Agent** with dual-path web search (<45% confidence HITL fallback + direct Out-of-Scope routing) and Dual-Chunking (**LlamaParse Vision LLM**).
- Hybrid semantic retrieval** using Pinecone vector search with **Cohere Neural Reranking**.
- AI Workflow Orchestration:** Implemented **adaptive retrieval** with **confidence-aware autonomous fallback** from enterprise knowledge base to **live web search** via LangGraph conditional edges.
- Designed **Multi-stage Hallucination Guard** - Confidence Gating -> LLM-as-Judge -> Adaptive ReAct Fallback
- Designed **stateless FastAPI services** enabling horizontal scaling with Redis-backed shared state.
- Implemented **async SSE streaming** responses for improved perceived latency.
- Integrated **Meta WhatsApp Business Cloud API** using webhook-driven event architecture, exposing the same Agentic RAG backend through conversational Web and WhatsApp interfaces without duplicating business logic.
- Implemented **asynchronous webhook handlers** with Redis-backed conversational session management, enabling scalable multi-turn AI interactions across multiple communication channels.

Live: <https://agentic-rag-financial-parser.onrender.com/>

Multi-Tenant Legal RAG Platform

Tech: FastAPI, React, Qdrant Cloud, Qwen 3 235B, Supabase (Postgres + Storage), Google OAuth 2.0

- Processed **31,500+ chunks** across 20+ legal acts with **Parent-Child Chunking** and deep metadata enrichment for GDPR-compliant multi-tenant vector search
- Built **SHA-256 sync engine** - auto-handles new files, changed files, orphan cleanup, and unchanged file skip
- Deployed containerized production platform with secure OAuth 2.0 authentication and multi-tenant access control
- Performed **retrieval quality evaluation** and iterative prompt optimization using real-world legal queries.

Live: <https://indian-legal-ai-expert.onrender.com>

Citizen Safety - Adaptive Chat Routing & Emergency Fallback

Tech: FastAPI, React, LangChain.js, ChromaDB, Llama 3.3 70B

- Deployed **spaCy-driven trilingual routing** (English, Hindi, Hinglish) with dual-indexing and intelligent 112/100 emergency fallback signaling
- Resolved production retrieval and indexing issues through systematic debugging and deployment optimization

Live: <https://citizen-safety-ai-assistant.vercel.app>

CORE COMPETENCIES AND TECHNICAL SKILLS

RAG & AI Architecture

Omnichannel AI Platforms, Conversational AI, Webhook Architecture, Event-Driven APIs, WhatsApp Business Cloud API, Channel-Agnostic AI Systems, LangGraph StateGraphs, LangChain, Agentic RAG, AI Agents, QLoRA, LoRA, Model Fine-Tuning, GGUF Quantization, llama.cpp, Cohere Neural Reranking, Parent-Child Chunking, MRL Embeddings, Hybrid Search, Hallucination Control, Confidence Gating, LLM-as-Judge, NLP, Multi-Turn Dialog, Prompt Engineering, Few-Shot/Zero-Shot Learning, Metadata Enrichment, Retrieval Augmented Generation, Multi-Agent Systems, Adversarial Testing, Red Teaming, Model Validation.

Vector Databases

Pinecone Serverless (namespaces, metadata filtering), Qdrant Cloud (payload indexes, upsert patterns), ChromaDB (persistent client, metadata-filtered deletion).

LLMs & APIs

Qwen 3.6 Plus, Qwen 2.5 72B, Qwen 3 235B, Llama 3.3 70B, Llama 3.2 1B, Llama 3.2 3B, Llama 3.1 8B, DeepSeek R1, GPT-4, OpenAI, Claude, Gemini - via OpenRouter, Direct APIs, Token Optimization, Hugging Face Transformers, Google Generative AI SDK, Gemini Vision API.

LLMOps & Observability

Langfuse distributed tracing, Redis semantic caching (SHA-256), pybreaker circuit breakers, SlowAPI rate limiting, UptimeRobot, Multi-method health endpoints, MongoDB feedback collection, Cost Controls, Real-Time Analytics.

Security & Compliance

Microsoft Presidio + spaCy NLP, custom Indian PII recognizers, Google OAuth 2.0 (Authlib), JWT (HS256), GDPR pipelines, 7-layer upload security, SHA-256 file fingerprinting, Magic Bytes verification, PDF Bomb Protection, CORS, ProxyHeaders Middleware, CSRF Protection, Input Validation.

Backend & Deployment

Meta WhatsApp Business Cloud API, Webhook Integrations, Event-Driven Architecture, FastAPI (async/await), Render, Vercel, Koyeb, GCP (Vertex AI, Cloud Run), IBM WatsonX, Google Colab, MongoDB Atlas, Supabase, Upstash Redis, React 18, Vite, LM Studio, Ollama, Tailwind CSS, SSE streaming, Express.js, Flask, Streamlit Cloud, Netlify, Hugging Face Spaces, Swagger UI, REST APIs.

Other Tools & Languages

Python, JavaScript ES6+, Node.js, HTML5, CSS3, Unsloth, LlamaParse (VLM), PyMuPDF, Jina v3, Bruno API Testing, Puppeteer, Git, Jupyter Notebooks, VS Code, NotebookLM, Gradio, Streamlit.

PROFESSIONAL EXPERIENCE

AI Prompt Engineer

Hogarth Worldwide (WPP Marketing) | Remote

September 2025 - October 2025

- Engineered constraint-based prompts (72-77 tokens) for commercial AI image generation (Flux, SDXL) ensuring brand compliance
- Developed 'Smart AI Prompt Builder' tool with keyword classification across 8 business use-cases
- Executed adversarial red-teaming to identify hallucination edges and prompt drift behaviors

Associate Engineer - Applied AI and Workflow Automation

British Telecom Global Services Pvt. Ltd. | Gurugram, India

January 2022 - August 2024

- Developed internal AI-assisted automation prototypes using **OpenAI APIs** and **Gemini Vision** for network documentation workflows
- Built internal NLP query interface with **Gradio UI** for real-time database deployments
- Drove FTTP optimization on GIS frameworks delivering **15% improvement** in network delivery
- Architected Python data pipelines reducing manual efforts by **60%**

Operations and Maintenance Engineer

Tata Communications Transformation Services Ltd. | Lucknow, India

November 2021 - January 2022

- Optimized ISP infrastructure with Orion monitoring - reduced downtime by **10%**

Optical Fiber Execution Engineer

Annu Infra Construct India Pvt. Ltd. | Kolkata, India

December 2017 - June 2018

- Deployed classified optical fiber across 3 high-security state zones under Ministry of Defense compliance

System Administrator

Tata Consultancy Services (TCS) | Noida, India

October 2013 - November 2014

- Maintained **99% uptime** on cloud production instances through proactive monitoring

INDUSTRY CERTIFICATIONS AND CREDENTIALS (ALL VERIFIED)

AI/ML and GenAI: NVIDIA: Building RAG Agents, AI on Jetson Nano | Google Cloud: 6 Skill Badges (Vertex AI, Gemini API, TensorFlow, GenAI Apps) | IBM AI: Fundamentals, Action, Deep Learning | Anthropic: MCP, Claude | Linux Foundation: Ethical AI

Job Simulations: BCG X: AI Financial Chatbot | Big 4: AWS SA, PwC Digital Assurance, Deloitte Analytics, Tata iQ GenAI, Siemens

All certifications verifiable through official platforms. Verification links available in online portfolio.

EDUCATION

Post Graduate Diploma in Power Transmission and Distribution

National Power Training Institute, Ministry of Power | 2015 - 2016 | New Delhi, India

Bachelor of Technology - Electrical and Electronics Engineering

UPTU - Uttar Pradesh Technical University | 2009 - 2013 | Lucknow, India