AMBUJ KUMAR TRIPATHI

AI Prompt Engineer | LLM Integration Specialist | Conversational AI Developer

Available Immediately | Remote/Hybrid | Open to Relocation

Email: kumarambuj8@gmail.com | Mobile: +91 9431 801363 | Location: Gorakhpur, Uttar Pradesh, India
Portfolio: https://ambuj-portfolio-v2.netlify.app/ | GitHub: https://github.com/Ambuj123-lab

PROFESSIONAL SUMMARY

AI Prompt Engineer and Automation Architect specialized in commercial-grade LLM optimization, RAG pipelines, and Conversational AI. Leveraging **5+ years of technical experience** in Telecom and Tech sectors to build cost-efficient, production-ready AI solutions. Successfully transitioned from legacy automation frameworks to GenAI-driven architectures, managing publicly deployed AI agents processing **525,000+ tokens** in live environments with **Langfuse observability** and **Redis** for real-time user analytics.

KEY TECHNICAL ACHIEVEMENTS

- **Production-Scale GenAI Deployment:** Deployed **5+ scalable AI agents** on Render and Streamlit Cloud, successfully processing **525,000+ tokens** and **10,000+ conversation turns** tracked via Langfuse traces and OpenRouter API routing
- **Production Observability and Monitoring:** Integrated **Langfuse** to monitor production metrics including request traces, response latency, and token usage patterns, providing visibility into model performance and API costs
- **Real-Time User Analytics:** Implemented **Redis** to track active user sessions and concurrency in real-time, providing instant visibility into system load and usage spikes without database latency
- **Enterprise Data Privacy:** Architected **Microsoft Presidio + SpaCy PII masking layer** to detect and redact sensitive information (Names, Emails, Phone Numbers) before LLM processing, ensuring **GDPR-compliant inference pipelines**
- **User Engagement Analytics:** Integrated **Google Analytics 4 (GA4)** on the portfolio UI to monitor user sessions, tracking click-through rates (CTR) on AI demos and project dwell time to analyze real-world engagement patterns
- **Systematic Prompt Optimization:** Built reusable prompt libraries and comprehensive documentation for diverse creative requirements, reducing prompt engineering time by **40%** through systematic optimization methodology and team feedback-driven iteration
- **Commercial AI Production Experience:** Engineered prompts for commercial-grade generative AI models at **Hogarth Worldwide (WPP)** with strict technical constraints

Core Strengths: Prompt Engineering | RAG Pipelines | LLM API Integration | Conversational AI Development | Model Validation | Red Teaming | Python/JavaScript/Node.js | Cloud Platforms | Production AI Workflows | Data Analysis | Automation | Quality Assurance

GENERATIVE AI PROJECTS (ALL LIVE & ACCESSIBLE)

Secure Enterprise RAG Assistant with Auto-Recovery Architecture | January 2026

Core RAG Pipeline & Auto-Recovery:

- Architected production-grade RAG system using **LangChain and Llama 3.3 70B**, featuring **auto-recovery mechanism** that automatically rebuilds vector database (ChromaDB) from source upon corruption, ensuring **98% system reliability**
- Developed **Custom Confidence Scoring Algorithm** using Vector Distance Normalization to quantify retrieval quality and flag potential hallucinations in real-time, enhancing user trust through transparency
- Optimized indexing latency by **98% (2 minutes to under 2 seconds)** through incremental document processing, utilizing **RecursiveCharacterTextSplitter (1000/200)** for semantic context preservation

Session Management & Data Persistence:

- Implemented **Multi-Turn Conversational Memory** using sliding window (last 3 messages) with **MongoDB Atlas** for session persistence, enabling context-aware responses across email-based authenticated sessions
- Configured **MongoDB TTL Index for automatic 30-day data expiration**, ensuring GDPR-compliant data retention and preventing storage overflow in production environments
- Integrated real-time analytics using **Upstash Redis** to track user engagement and session metrics securely

Enterprise Security & Compliance:

- Integrated **Enterprise Privacy Layer** using **Microsoft Presidio and SpaCy**, detecting and masking PII (Names, Phones, Emails) before LLM processing to ensure GDPR compliance
- Developed **Content Moderation System** with regex-based abusive language filter and **zero fabrication policy** through strict system prompts, ensuring safe and reliable user interactions

Monitoring & Deployment:

- Integrated **LangFuse Observability Platform** to monitor production-grade metrics including Traces, Latency, and Token Cost, providing full visibility into model reasoning processes and enabling data-driven optimization
- Optimized User Experience with **Interactive Dashboard** featuring real-time latency tracking (Streamlit metrics) and Source Viewer component, following enterprise-grade UI standards for production AI systems
- Engineered **OS-agnostic deployment pipelines**, resolving cross-platform path inconsistencies (Windows/Linux) and enabling seamless cloud deployment on Streamlit Cloud
- **Live Demo:** https://ambuj-rag-chatbot.streamlit.app/

  Technologies: *Python, LangChain, LangFuse, Llama 3.3 70B, ChromaDB, Streamlit, Microsoft Presidio, SpaCy, MongoDB Atlas, Upstash Redis, RecursiveCharacterTextSplitter*

Geo AI - Geospatial Narrative & Image Analysis Tool | November 2025

- Developed AI-powered geospatial analysis tool using **Gemini 1.5 Flash** to generate location-based narratives from map images
- Implemented interactive map rendering with **Leaflet.js**, allowing users to capture specific regions or upload custom images for AI analysis
- Integrated secure image handling and API token management for seamless AI analysis requests
- Deployed on **Render** with automated CI/CD pipelines from GitHub
- **Live Demo:** https://geo-narrator-ai.onrender.com/

  Technologies: *Node.js, Express, Gemini API, Leaflet.js, HTML5, CSS3, Render*

LLM-Integrated Resume Chatbot | Meta Llama 3.3 70B | August 2025

- **Stateless Backend Architecture:** Developed a lightweight Flask REST API where conversation history is managed **client-side** and passed via JSON payloads, eliminating the need for server-side databases (Redis/SQL) and reducing latency
- **Context-Aware Prompt Engineering:** Implemented **Dynamic Context Injection** by loading structured resume data (JSON) directly into the LLM's System Prompt, ensuring factual accuracy without complex RAG pipelines
- **AI Guardrails and Moderation:** Coded custom Python logic (**is_abusive function**) to filter toxic keywords before API calls, ensuring interaction safety and professional output constraints
- **Production Deployment:** Configured **Gunicorn (WSGI)** as the application server on Render to handle concurrent requests efficiently, utilizing python-dotenv for secure API key management
- **Error Resilience:** Built robust exception handling for OpenRouter API to gracefully manage timeouts and upstream errors, ensuring continuous service availability
- **Live Demo:** https://ambuj-resume-bot.onrender.com

__Technologies:__ Meta Llama 3.3 70B, OpenRouter API, Python, Flask, Gunicorn, JSON Context Injection

Professional Task Management System | Privacy-First PWA | September 2025

- **Privacy-First Architecture:** Architected a **Zero-Backend design** using LocalStorage for **100% client-side data persistence** with JSON Export/Import capabilities (via Blob API), ensuring complete data sovereignty
- **PWA and Offline Capability:** Implemented an installable **Progressive Web App (PWA)** structure with Service Worker registration, enabling Add-to-Home-Screen functionality and offline resource caching
- **Dual Audio Integration:** Integrated Web Audio API for real-time synthesized feedback sounds (Oscillators) and Web Speech API for text-to-speech welcome messages, enhancing UI accessibility
- **Smart Session Tracking:** Developed an **Activity-Aware Session Timer** using Page Visibility API and Event Listeners (Mouse/Keyboard) to track accurate focus time, automatically pausing during user inactivity
- **Advanced Search Logic:** Built a multi-field search algorithm filtering tasks by ID, Category, Priority, and Content simultaneously, improving data retrieval efficiency

__Technologies:__ React 18, Tailwind CSS, PWA, Service Workers, Web Audio API, Web Speech API, LocalStorage

Smart AI Prompt Builder - Intelligent Prompt Generation Tool | August 2025

- Developed intelligent web application helping users create effective AI prompts and recommending optimal AI platforms
- Built advanced keyword matching algorithm automatically detecting **8 problem categories**
- Implemented smart AI tool recommendation engine suggesting best platforms based on user requirements
- Created bilingual interface (**Hindi/English**) with **6 customizable tone styles**

__Technologies:__ HTML5, CSS3, JavaScript ES6+, Client-Side Processing, Netlify

Resume-as-Code - Automated Resume Builder Pipeline | July 2025

- Built automated pipeline generating ATS-friendly PDF resumes from single HTML/CSS source file
- Implemented Node.js backend with **Puppeteer** for programmatic HTML rendering and PDF export
- Achieved **100% formatting consistency** reducing update time by **90%**

__Technologies:__ Node.js, Puppeteer, HTML5, CSS3, Headless Browser Automation

Additional AI Projects

- **Rule Based Chatbot (IBM Watson):** Engineered a structured NLU chatbot achieving **95% intent recognition accuracy**. Deployed the frontend on **Netlify CDN** backed by IBM Cloud architecture, ensuring **99.9% uptime** through decoupled static hosting

CORE COMPETENCIES & TECHNICAL SKILLS

AI & Prompt Engineering

Prompt Engineering, Large Language Models (GPT-4, Claude, Gemini, Meta Llama 3.3 70B, IBM WatsonX), RAG Pipelines (Retrieval-Augmented Generation), Vector Databases (ChromaDB, FAISS), Semantic Search and Embeddings, LangChain Orchestration, LangFuse Observability, Multi-Turn Dialog Systems, Sliding Window Memory, Hallucination Control, Context-Aware Systems, LLM API Integration, Conversational AI Development, Chatbot Development, Multi-Modal AI, Adversarial Testing, Red Teaming, Model Validation, Model Robustness Testing, Few-Shot Learning, Zero-Shot Prompting, Token Optimization, Content Evaluation, Content Moderation Systems, Data Quality Assessment, Fact-Checking and Verification, Responsible AI, AI Ethics

Programming Languages & Frameworks

Python, JavaScript (ES6+), Node.js, Express.js, HTML5, CSS3, React 18, Flask, Gradio, Streamlit, Tailwind CSS, Babel, RESTful API Development, Async/Await, Promise Handling

AI Tools & Platforms

MongoDB Atlas, Microsoft Presidio (PII Masking), SpaCy (NLP), LangChain, LangFuse, ChromaDB, Upstash Redis (Analytics), Hugging Face Transformers, OpenRouter API, Gemini API, Google Generative AI SDK, Gemini Vision API, Vertex AI, IBM WatsonX, Jupyter Notebooks, VS Code, NotebookLM, REST APIs, Puppeteer, Git, Conversation State Management, Multi-Turn Dialog Systems

Cloud Platforms

Google Cloud Platform (Vertex AI, Cloud Functions, Cloud Run, GKE), IBM (WatsonX Services, Cloud Foundry, Speech to Text, Text to Speech), Streamlit Cloud, Netlify, Render.com, Hugging Face Spaces

Specialized Skills

Enterprise Data Privacy (GDPR Compliance), Incremental Indexing Logic, Progressive Web Apps (PWA), Service Workers, Offline Functionality, Data Processing and Visualization (Papa Parse, SheetJS, Plotly.js), Web Speech API, Web Audio API, CORS, Rate Limiting, Environment Secrets Management, Security Best Practices, Input Validation, API Rate Limiting, Session Management, PDF.js, html2canvas, jsPDF, Conversational Memory Architectures, Regex-based Content Moderation, Content Moderation Systems, Pattern-Based Guardrails

Additional Technical Skills

Data Analysis, Automation, Network Optimization, GIS Tools, Fiber Optic Network Design, Excel Advanced Functions, Business Process Documentation, Performance Monitoring, Quality Assurance, Cross-functional Collaboration, Technical Documentation

PROFESSIONAL EXPERIENCE

AI Prompt Engineer

Hogarth Worldwide (WPP Marketing Communications India Pvt. Ltd. via TeamLease) | Remote, India

September 2025 – October 2025 | **6-Month Contract (Concluded due to mandatory office relocation requirement - health reasons)**

- Engineered and optimized prompts for commercial-grade generative AI models (**Flux, SDXL**) within strict technical constraints (**72-77 token limits** across Base, Positive, Negative prompts) for brand-compliant visual outputs
- Conducted systematic model validation and adversarial testing, identifying critical limitations including instruction adherence failures, hallucination patterns, layout inconsistencies, and edge-case behaviors across multiple parameters
- Developed **Smart AI Prompt Builder tool** to standardize prompt creation workflows across team, featuring intelligent keyword matching algorithm detecting **8 problem categories** and recommending optimal AI platforms based on task requirements
- Collaborated with cross-functional teams to define AI integration protocols and quality assurance standards for production pipelines
- Developed prompt refinement methodology translating subjective business feedback into quantifiable technical adjustments

  Key Achievement: *Created systematic prompt optimization framework reducing iteration cycles and improving output quality consistency across diverse creative requirements*

  Technologies: *Flux, SDXL, Prompt Engineering, Adversarial Testing, Model Validation, QA Protocols*

Associate Engineer - Fibre & Network Delivery (Applied AI & Workflow Automation)

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

January 2022 – August 2024

- Led FTTP network planning using **GIS tools**, reducing deployment time by **15%** through systematic optimization
- Streamlined network estimation workflows by automating data extraction from **Piper database** to structured documentation including duct diagrams, cable diagrams, and SOC costing sheets, reducing manual effort by **60%**

- Collaborated on transition from manual processes to **Autotron-based automation pipeline**, standardizing documentation workflows across engineering teams
- Developed Excel-based automation solutions and Python scripts for repetitive data transformation tasks, improving data accuracy and processing speed
- Explored AI-powered document intelligence tools (May-August 2024) using **Gemini 1.5 Flash API** for automated extraction of technical parameters from network specifications and site survey reports
- Prototyped conversational AI interface using **Gradio** enabling stakeholders to query network deployment metrics through natural language, demonstrating potential for cross-team accessibility improvements
- Experimented with multimodal AI capabilities for automated interpretation of network design maps and infrastructure blueprints, achieving promising results in pilot testing phase
- Collaborated with MIS/Power BI teams to explore AI-driven analytics solutions and automation pipelines, receiving recognition for innovative problem-solving approaches
- Coordinated cross-functional teams achieving **98% quality assurance compliance** with streamlined processes
- Awarded **"Top Performer"** (September 2022) for innovative problem-solving and process improvement initiatives

  Innovation Achievement: *Successfully automated critical network planning workflows and pioneered exploration of AI technologies for telecom operations, establishing foundation for intelligent automation initiatives*

  Technologies: *GIS Tools, Network Optimization, Piper Database, Autotron, Excel Advanced Automation, Python Scripting, Gemini 1.5 Flash API, Gradio*

Operations & Maintenance Engineer

Lobo Staffing (Client: Tata Communications Transformation Services Ltd.) | Lucknow, India
November 2021 – January 2022

- Optimized network performance and monitoring systems, reducing downtime by **10%** through comprehensive data analysis
- Managed billing systems for OSP/ISP services with **100% accuracy** using systematic quality control processes
- Improved Field Engineer SLA compliance, achieving **98% TAT** through workflow optimization

  Technologies: *ServiceNow, Orion, Network Monitoring Systems, Excel Advanced Functions*

Optical Fiber Execution Engineer

Annu Infra Construct India Pvt. Ltd. | Kolkata, India
December 2017 – June 2018

- Executed fiber optic deployment projects for **Ministry of Defense** across **3 states** ensuring complete security compliance
- Ensured **100% compliance** with security protocols and technical specifications through rigorous quality assurance
- Reduced project delays by **20%** through effective vendor coordination and strategic resource optimization

System Administrator

Teamware Solutions (Client: Tata Consultancy Services) | Noida, India
October 2013 – November 2014

- Managed revenue systems for **Delhi Jal Board**, ensuring **99% uptime** through proactive monitoring
- Trained **20+ portal users** on system optimization and best practices
- Supported **eMigrate portal (Ministry of Overseas Affairs)** with document verification processes

INDUSTRY CERTIFICATIONS & CREDENTIALS (ALL VERIFIED)

AI/ML & Generative AI Certifications

- **NVIDIA:** Building RAG Agents with LLMs (LangChain, FAISS, FastAPI), AI on Jetson Nano (Edge AI, CNNs, ResNet-18)
- **Google Cloud:** 6 Skill Badges — Vertex AI Prompt Design, Gemini API, TensorFlow Image Classification, GenAI Apps with Streamlit, Cloud Run & GKE Deployment
- **IBM AI:** AI Fundamentals, Generative AI in Action, Deep Learning Essentials, Chatbot Development, Python for Data Science
- **Azure AI:** Language Models on Databricks, Responsible AI Principles
- **Anthropic Academy:** Model Context Protocol (MCP), Claude with Google Vertex AI
- **Linux Foundation:** LFS118 - Ethical Principles for Conversational AI

Enterprise Job Simulations (Forage Platform)

- **BCG X:** AI-Powered Financial Chatbot Development (NLP, Data Extraction, Financial Analysis)
- **Big 4 & Enterprise:** AWS Solutions Architecture, PwD Digital Assurance, Deloitte Australia Analytics, Tata iQ GenAI, Siemens

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

EDUCATION

Post Graduate Diploma in Power Transmission & Distribution

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

Bachelor of Technology - Electrical & Electronics Engineering

Uttar Pradesh Technical University | 2009 – 2013 | Lucknow, India