📌 PERSONAL INFORMATION

Abhishek Singh
📍 Pune, Maharashtra – 411014, India (Open to Relocation)
📞 +91-7223933444
✉️ maiabhishek1012@gmail.com
🌐 Website
💼 LinkedIn | GitHub

🎯 PROFESSIONAL SUMMARY

Strategic and hands-on **AI Engineer** with 3.3 years of work experience and 5+ years of domain experience driving next-gen innovation in **Generative AI, NLP, and Agentic RAG systems**.

I specialize in **bridging cutting-edge research with real-world implementation**, blending deep expertise in **LLMs, NLP, RAG systems**, and intelligent automation with strong product thinking and engineering execution. From crafting scalable AI tools for **pricing, forecasting, contract intelligence,** and **customer experience** to building **agentic systems** that learn, adapt, and act — I thrive where complex problems meet elegant solutions.

I'm the architect behind **RAGentX**, a next-gen agentic RAG framework that fuses memory, planning, and tool use for autonomous, multimodal task execution. My stack spans **Python, FastAPI, LangChain, LangGraph, Hugging Face, Streamlit, MLOps,** and **React** — but beyond the tools, it's the impact that drives me.

Outside the codebase, I'm a **design tinkerer**, lifelong learner, and advocate for **ethical, responsible AI**. I thrive in **zero-to-one environments**, move with urgency, and don't wait for direction — I anticipate it.

Let's build AI that doesn't just work — it **thinks, evolves, and delivers**.

🧠 CORE COMPETENCIES

🔮 Generative AI & NLP

- **LLMs & APIs**: OpenAI (GPT-4/4o), Cohere, Mistral, Ollama
- **Frameworks**: LangChain, LlamaIndex, LangGraph, Haystack
- **RAG Systems**: FAISS, Chroma, Pinecone, Semantic Search
- **Prompt Engineering & Optimization**
- **Transformers & Models**: Hugging Face, BERT, GPT, T5, CLIP
- **LLM Fine-tuning**: LoRA, QLoRA, Adapters, Quantization (GGUF, AWQ)
- **NLP Tasks**: Text Generation, Summarization, NER, Classification
- **Monitoring**: LangSmith, Evaluation Pipelines

🤖 Machine Learning Engineering

- **Libraries**: Scikit-learn, XGBoost, LightGBM, CatBoost, PyTorch Lightning, TensorFlow
- **Time Series**: ARIMA, Prophet, DeepAR
- **Model Optimization**: Optuna, ONNX
- **Explainability**: SHAP, LIME
- **Reinforcement Learning**, **Computer Vision (YOLO, OpenCV)**
- **Distributed Training**: DDP, FSDP
- **Federated Learning** (Basics)

🧩 Agentic AI Systems

- **Autonomous Agents**: CrewAI, AutoGen, LangGraph
- **Memory, Planning, Tool Use Integration**
- **LangServe, FastAPI for Agent Deployment**

🛠️ Data Engineering

- **Languages**: Python, SQL, PySpark
- **ETL/ELT & Orchestration**: Apache Spark, Airflow, Dagster
- **Storage**: PostgreSQL, MongoDB, Delta Lake, Iceberg
- **Warehousing**: Snowflake, BigQuery
- **Streaming**: Kafka, Flink
- **Data Quality**: Great Expectations
- **IaC**: Terraform
- **Cloud Data Platforms**: AWS, GCP, Azure

🚀 MLOps & DevOps

- **Deployment**: FastAPI, LangServe, Hugging Face Spaces, Docker, Flask
- **CI/CD**: GitHub Actions, Bitbucket Pipelines
- **Monitoring**: MLflow, Grafana, Prometheus
- **Containerization & Orchestration**: Docker, Kubernetes, Helm
- **Model Lifecycle**: DVC, MLflow, Kubeflow
- **Auth & Access**: Firebase, Clerk, Auth0 (optional use)

🌐 Full Stack & Frontend (Optional/Light)

- **Frontend Tools**: Streamlit, React, Next.js, Tailwind
- **Visualization**: Power BI, Plotly, D3.js, Dash

☁️ Cloud Platforms

- **AWS**: SageMaker, Lambda, S3, ECS
- **Azure**: Azure ML, App Service, Databricks
- **GCP**: Vertex AI, BigQuery, Cloud Run

🛠️ Project & Team Tools

- **Version Control & Collaboration**: Git, Confluence, Jira, Bitbucket
- **Productivity**: Notion, Trello (for personal planning)

💼 PROFESSIONAL EXPERIENCE

AI Engineer & Data Scientist
Vcreatek (India) — Jan 2023 – Present

AI-Driven Business Solutions | GenAI Prototyping | ML System Development

🧠 Project: EPRMate – Agentic Packaging Data Enrichment for EPR Compliance

- Designed and deployed an **agentic AI pipeline** to auto-enrich missing packaging attributes across millions of SKUs, enabling accurate Extended Producer Responsibility (EPR) fee calculations.
- Integrated **multi-agent architecture** combining rule-based logic, ML classification (XGBoost, LightGBM), external registry lookups, and GenAI-based inferencing (OpenAI GPT-4) for intelligent enrichment.
- Achieved a **90% reduction in incomplete records** and **70% cut in manual remediation effort**, significantly improving audit readiness and regulatory compliance.
- Built traceable enrichment flows with **confidence scoring**, reasoning logs, and human-in-the-loop fallbacks to ensure high data quality and explainability.
- Orchestrated workflows using **PySpark**, Airflow, Delta Lake, and OpenAI APIs, delivering scalable enrichment with continuous learning via human feedback loops.

🧠 TEXTSPECTRUM – No-Code NLP App

- Built a no-code NLP platform enabling business users to analyze unstructured text via classical ML and OpenAI models (GPT-3.5).
- Integrated modular capabilities like sentiment analysis, topic modeling, classification, and a ChatGPT-style assistant.
- Empowered non-technical teams to generate insights without writing code, significantly reducing reliance on data science teams.

📈 POS Forecasting Optimizer

- Developed a retail sales prediction engine using ARIMA and Prophet, achieving >92% forecasting accuracy across SKUs.
- Deployed scalable pipelines using Azure Data Factory with integrated triggers, logs, and health monitoring.
- Enabled automated forecasting workflows that eliminated manual intervention and improved decision-making agility.

📑 Claim Recovery Engine

- Engineered a GenAI-driven contract validation system using NER and clause parsing to auto-verify deduction claims.
- Identified discrepancies between claims and contract clauses, unlocking $500K+ in potential recoveries.
- Automated a previously manual audit process, enhancing compliance, transparency, and policy enforcement.

🛒 TDP – Space-Aware Assortment Optimizer

- Designed a hybrid ML + heuristic system to optimize product placement and maximize shelf-space efficiency.
- Reduced shelf-space wastage by 18% while improving assortment strategies through predictive modeling.
- Delivered actionable recommendations to merchandisers for real-world planogram improvement and SKU rationalization.

🚀 FLAGSHIP PROJECT

RAGentX – Agentic Research & Execution Assistant
Personal Open-Source Project — 2024 – Present

An intelligent GenAI assistant with RAG, long-term memory, and autonomous task planning.

- Built from scratch using **LangChain**, **LLMs**, **Vector DBs**, and **Toolformer-style agentic planning**.
- Integrates **multi-source retrieval** (docs, web, APIs) + intelligent memory context injection.
- Features Trello-style task orchestration, conversational memory, and customizable agent workflows.
- Packaged for web with **FastAPI + Streamlit + Tailwind UI**, optimized for both interactivity and performance.

🧪 PROJECTS HIGHLIGHTS

- **E-Commerce Review Analytics** – Built a GenAI dashboard for auto-insight generation from 100K+ reviews.
- **Walmart Pricing Analytics** – Developed elasticity models using regression + uplift modeling.
- **ResumeGPT (Upcoming)** – Smart GPT-powered resume & JD matcher for job seekers (under dev).

🎓 EDUCATION

📌 PGDM: Research and Technical Business Analytics -> 2021-23

Prestige Institute of Global Management, Indore, MP

📌 Bachelor of Commerce -> 2017-2020

Renaissance College of Commerce & Management, Indore, MP

🏆 CERTIFICATIONS & TRAININGS

- LangChain Agent Frameworks (Advanced) – DeepLearning.AI
- Prompt Engineering for Developers – OpenAI x DeepLearning.AI
- Hugging Face Transformers – Coursera
- MLOps Fundamentals – Google Cloud

🌍 COMMUNITY & OPEN SOURCE

- Contributor: LangChain templates + LangServe integrations
- Mentor: Junior AI devs via LinkedIn + GitHub Discussions
- Open Source: RAGentX (Under Release)

📢 REFERENCES

Available upon request.