SATYAM KALRA

AI ENGINEER

United States | skalra@gwu.edu | +1-(571)-622-8689 | https://github.com/satkalra | https://www.linkedin.com/in/satkalra/

AI Engineer specializing in Large Language Models (LLMs), retrieval-augmented generation (RAG), and agentic architectures. Experienced in designing, developing, and deploying scalable AI systems that transform user intent into intelligent automation from RAG bots and analytical agents to auto-dashboard generation systems. Proven ability to lead cross-functional teams, build modular AI products and prompt orchestration.

RELEVANT WORK EXPERIENCE

Strategy (Formerly MicroStrategy)   **Tysons, VA**
AI Engineer   **July 2024 - Present**

- Led development of flagship AI products "Agents" and "Auto Dashboards" from concept to release, collaborating across product, delivery, and engineering teams.
- Designed the Agent system, enabling users to create custom AI agents from structured and unstructured data to answer analytical questions with **~90% accuracy**.
- Architected a **multi-provider AI layer**, supporting seamless integration and switching between **OpenAI, Gemini, and Claude** for dynamic inference routing.
- Architected a **semi-orchestrated deterministic flow** for the Auto Dashboard system — transforming natural-language intent into fully formatted dashboards within seconds, cutting dashboard creation time by **over 95%.**
- Led a cross-functional team to develop a central prompt library, standardizing and optimizing all prompt templates used across enterprise AI applications.

Data Scientist (Internship)   **July 2023 – May 2024**

- Developed Auto Expert, a publicly available **RAG-based AI assistant** for Strategy Software that helps customers find product information, troubleshoot issues, and access community resources, leading to **a 20% reduction** in customer support cases through automated case deflection.
- Implemented machine learning guardrails and hallucination reduction algorithms before guardrail frameworks were mature, achieving a **10% improvement** in factual accuracy.
- Created a user-friendly tester UI using React and Node, leading to a 30% boost in testing efficiency among project testers while smoothly integrating and optimizing data flow into a separate database, reducing data access time by 40% during testing.

Rupeek Fintech   **Bengaluru, India**
Software Development Engineer   **January 2022 - August 2022**

- Designed a **multi-app platform** for the autonomous lending of gold loans using Node.js, React, Electron.js, and TypeScript.
- Created and managed app development by implementing scalable backend and frontend for customer-facing products.
- Impacted the loan dispersal process, **resulting in $122K disbursed to over 1000 customers**, opening new revenue sources.
- Collaborated across teams to identify business requirements and execute scalable solutions, improving client-server interactions by reducing **API latency by 1 second.**

TECHNICAL PROJECTS

Team Leader, MCP Integration for Agents   **July 2025 - August 2025**

- Led the integration of Model Context Protocol (MCP) into the Agents product, allowing ingestion of any data source from MCP servers.
- Designed Agent–MCP interoperability for seamless contextual augmentation and multi-source reasoning.

Team Leader, NER Agent   **November 2024 - December 2024**

- Designed a mechanism to extract keywords from user queries and map them to relevant tags using Named Entity Recognition.
- Matched extracted tags with database table columns, leveraging Redis for efficient lookup and similarity scoring.

Team Leader, Omni Counsel - Research Paper Chatbot with Language Model   **August 2023 - December 2023**

- Engineered a highly efficient **chatbot pipeline for ML research papers**, utilizing OpenAI's LLM. Curated and processed a dataset of 100 papers and implemented advanced text-cleaning and token-based segmentation for rapid answer retrieval in Flask.
- Pioneered a dynamic question-to-answer pipeline with context extraction using cosine similarity and **cluster prediction with an accuracy of 89%**, optimizing prompt generation and temperature adjustment.

TECHNICAL SKILLS

- Languages and Frameworks: Python, SQL, Typescript, JavaScript, React.js, Django, Node.js, Fast API, Pydantic, Flask
- Data & AI Stack: Spark, Kafka, Redshift, Airflow, Hadoop, AWS, Azure, PostgreSQL, MongoDB, Scala, LLM, Redis.
- ML/LLM: RAG, Prompt Engineering, MCP, Generative AI, Semantic Retrieval, orchestration, Langchain, LangGraph.
- Software and Systems: Tableau, Docker, Linux, Mac, MATLAB, Unix, Data Modeling, MapReduce, Docker, Kubernetes.

EDUCATION

The George Washington University, School of Engineering & Applied Science   **Washington, DC**
Master of Science in Computer Science   **May 2024**
Relevant Courses: Computer Architecture, Big Data and AI, Machine Learning, Cloud Computing. Database

Thapar Institute of Engineering and Technology   **Patiala, India**
Bachelor of Engineering   **June 2021**
Relevant Courses: Data Structures and Algorithms, Artificial Intelligence, Digital Image Processing