

YASH SRIVASTAVA

+91-7599732598 · officialyashsrivastava1504@gmail.com · [LinkedIn](#) · [GitHub](#) · [LeetCode](#) · Meerut, UP, India

Summary

AI Engineer building production-grade GenAI systems including Agentic RAG, LLM orchestration, and vision-model pipelines, backed by scalable FastAPI/backend engineering. Experienced in shipping AI products end-to-end using modern agentic development tooling, with a focus on reliable, secure, and production-oriented AI applications.

Experience

AI Engineer – Independent Projects

June 2026 – Present

Self-directed

Remote

- Designing and shipping independent, production-style AI applications end-to-end, combining full-stack engineering with LLM systems and Applied AI Engineering to solve real-world data and workflow problems.

Gen AI Engineer

Jan 2026 – May 2026

OneHeal Technologies

Remote

- Built a RAG-based clinical decision-support system for NCCN oncology guidelines – vision-LLM extraction turns flowchart PDFs into a structured decision graph (PostgreSQL + pgvector), resolving cross-page/cross-document links automatically.
- Designed a hand-rolled ReAct agent that traverses the decision graph against patient inputs, halting to request missing data instead of guessing, with page-ID citations on every output.
- Built a bilingual (Hindi/English) vision-based OMR pipeline processing 100+ medical questionnaire templates – homography-aligned bubble detection with agentic confidence-gated escalation to a vision LLM for ambiguous marks, validated against a hand-labeled eval set with Langfuse observability across the full pipeline.

Technical Intern

Dec 2025 – Dec 2025

OneHeal Technologies

Remote

- Gained foundational experience in scalable backend system design using FastAPI and RESTful APIs on a health-care project.

AI Intern

Sept 2025 – Dec 2025

AD&AR Technologies

Onsite, Noida, India

- Built an agentic voice AI assistant on n8n – STT/TTS pipeline with dynamic tool-calling and intent-based branching, backed by RAG retrieval for real-time business conversations.
- Fine-tuned an open-source LLM on a custom domain dataset using QLoRA (parameter-efficient fine-tuning) for low-resource adaptation.

Projects

DataSage (Magic Excel) – AI Spreadsheet Automation

[GitHub \(FE\)](#) | [GitHub \(BE\)](#) | [Live](#)

React, TypeScript, FastAPI, Python, PostgreSQL, Pandas, LLMs

- Architected an LLM-powered spreadsheet automation pipeline that converts natural-language requests into structured, typed transformation operations, enabling users to query and manipulate Excel data without formulas or SQL.
- Designed a two-stage LLM orchestration system that routes simple aggregate queries directly while translating complex requests into deterministic transformation pipelines for reliable execution.
- Built an AST-based expression sandbox and Pandas execution engine to safely evaluate LLM-generated conditional logic and row-level operations while preventing arbitrary code execution.
- Engineered a stateless, preview-first data processing architecture that dynamically executes transformations and generates non-destructive diffs for modifications, deletions, and affected spreadsheet regions.

Manah.ai – LLM-Powered Conversational AI Platform

[GitHub \(BE\)](#) | [GitHub \(FE\)](#) | [Live](#)

FastAPI, Python, React, TypeScript, PostgreSQL, OpenRouter, Gemini, SSE

- Engineered an LLM-powered conversational system integrating Gemini through OpenRouter with sliding-window context management and real-time SSE streaming for low-latency AI interactions.
- Built an asynchronous FastAPI backend with secure API proxying, JWT authentication, PostgreSQL persistence/RLS, request validation, and IP-based rate limiting.
- Implemented fault-tolerant LLM infrastructure using AsyncIO, retries, exponential backoff, and request timeouts.

to handle rate limits and transient API failures reliably.

Technical Skills

- **Languages:** Python, SQL
- **Generative AI:** LLMs, Generative AI, RAG, Retrieval-Augmented Generation, Agentic AI, AI Agents, LLM Integration, LLM Orchestration, Prompt Engineering, Context Engineering, Structured Outputs, Function Calling, Tool Calling, Conversational AI
- **LLM Engineering:** RAG Pipelines, Embeddings, Vector Search, Context Management, Token Optimization, LLM Streaming, Prompt Evaluation, LLM Evaluation, AI Safety, Guardrails, Hallucination Mitigation
- **AI Systems & MLOps:** Evaluation Pipelines, Observability, LLM Tracing, Logging, Performance Monitoring, Latency Optimization, Token/Cost Monitoring, AI Application Monitoring
- **Backend & APIs:** FastAPI, REST APIs, AsyncIO, Server-Sent Events (SSE), API Authentication, Rate Limiting, Retry & Exponential Backoff
- **Data & Databases:** PostgreSQL, Supabase, Pandas, Vector Databases
- **Security:** JWT Authentication, PostgreSQL Row-Level Security (RLS), API Security, Input Validation, AST-based Sandboxing

Education

Meerut Institute of Engineering and Technology (MIET)

2022 – 2026

Bachelor of Technology in Computer Science and Engineering

Meerut, UP, India

Certifications

- **AI Fluency: Framework & Foundations** – Anthropic — **Claude 101** – Anthropic
- **Claude Code 101** – Anthropic — **Journey to Cloud: Envisioning Your Solution** – IBM