Siddharth Singh

AI Backend Engineer · Python · Async Distributed Systems · LLM Inference Pipelines · Kubernetes · Model Serving
Noida, Uttar Pradesh, India · +91-82679-45097
siddharthsingh8418@gmail.com · github.com/Siddharthsinghkumar · linkedin.com/in/siddharth-singh

PROFILE

AI Backend Engineer with 3+ years of hands-on experience building Python backend systems, distributed data pipelines, and production AI services across independent projects, open-source work, and professional roles. Strong in async programming, LLM inference pipelines, model serving, Docker/Kubernetes deployments, and fault-tolerant microservices. Built production systems for multi-provider LLM routing, streaming inference, OCR-to-LLM pipelines, and vector search with a focus on reliability, observability, and fast iteration in AI environments.

TECHNICAL SKILLS

- **AI Backend:** Python, FastAPI, asyncio, LLM inference pipelines, model serving, SSE streaming, async request orchestration, LangGraph, RAG, RAGAS, OCR-to-LLM pipelines, multi-provider LLM systems
- **Distributed Systems & Data:** Microservices, distributed systems, system design, job queues, Kafka, PostgreSQL, Redis, MongoDB, SQLite, vector databases (Pinecone, Chroma), Sentence-Transformers
- **Infra & Reliability:** Docker, Docker Compose, Kubernetes (k3s), GitHub Actions, CI/CD pipelines, Prometheus, Grafana, Sentry, AWS, GCP, Linux, Git, pytest

PROFESSIONAL EXPERIENCE

AI Backend Engineer — LLM Travel Planner Jan 2026–Present
Python, FastAPI, asyncio, LangGraph, Ollama, OpenAI, GCP/Gemini, Docker, Kubernetes, Prometheus/Grafana github.com/ai-t

- Improved reliability of a distributed AI backend, as measured by eliminating cascading provider failures during API degradation, by building async circuit breakers, provider health checks, and automatic failover across local and cloud LLM inference pipelines.
- Increased production resilience for model serving, as measured by zero-downtime credential rotation and rate-limit recovery, by implementing an async routing layer with atomic locking, cooldown windows, and fallback orchestration.
- Strengthened deployment scalability and operational readiness, as measured by successful rollout on a multi-node k3s Kubernetes cluster and 13+ production runbooks, by deploying containerized FastAPI microservices with Docker, reverse proxying, health checks, and Prometheus/Grafana observability.
- Improved user-facing response performance, as measured by real-time token-by-token SSE streaming and uninterrupted cross-provider fallback, by architecting async Python pipelines for request orchestration, streaming delivery, and structured decision generation in a fast-paced AI product environment.

Backend Engineer — Sindhey Pathology Apr 2026–Present
Python, FastAPI, PostgreSQL, GitHub Actions, CI/CD, Sentry, Next.js sindheypathology.com

- Delivered production backend systems for a zero-to-one healthcare platform, as measured by live patient booking, admin operations, and report-delivery workflows, by designing APIs, data models, and service logic for day-to-day clinic operations.
- Improved workflow reliability and operational correctness, as measured by a stateful 5-step booking flow with real-time price aggregation across lab visits and home collections, by implementing backend orchestration, validation, and audit-aware handling of patient and order data.
- Increased release speed and deployment consistency, as measured by production launch readiness and repeatable releases, by setting up GitHub Actions CI/CD pipelines, Sentry monitoring, and environment-aware deployment workflows.
- Increased platform extensibility for future automation, as measured by support for staged payments, WhatsApp/OTP flows, and report-delivery workflows, by building backend-first service architecture suited for a fast-moving startup environment.

KEY PROJECTS

Smart Job Scanner v2 — Distributed OCR + LLM Inference Pipeline 2025–Present
Python, asyncio, OCR, SQLite, multiprocessing, structured logging github.com/smart-job-scanner-v2

- Built a production OCR + LLM inference pipeline, as measured by processing 15–20 GB of newspaper PDFs daily for 75+ days unattended, by engineering an 11-stage Python workflow with checkpointing, structured JSONL logging, and automated alert delivery.
- Improved throughput 3.2x and reduced false positives by 90%, by optimizing parallel worker pools, GPU-aware multiprocessing, VRAM-constrained scheduling, and iterative pipeline refinement across multiple production versions.

Open-Source & Independent AI Backend Work 2023–Present
Python, FastAPI, LLM systems, OCR pipelines, backend services

- Built and maintained a public portfolio of AI backend systems across 2023–2026, including LLM applications, OCR-to-LLM pipelines, streaming APIs, and production-oriented Python services.

Auto Pay Reminder System — Workflow Automation Service 2025
Google Apps Script, Google Sheets, Telegram Bot API github.com/auto-pay-reminder-system

- Built a lightweight automation service for payment reminders and workflow escalation, as measured by zero-fee transaction support and automated follow-up handling, by integrating Telegram bot workflows, command parsing, and dynamic payment link generation.

EDUCATION

B.Tech 2020–2024