Swayam Gupta

2024 Graduate · Material Science & Engineering
Indian Institute of Technology, Kanpur

gupta.swayam123@gmail.com · (+91) 8957370095
swayamg20 · guptaswayam20
swayamgupta20.medium.com · swayamg20.github.io

Education

Year	Degree	Institution	Score
2024	B.Tech, MSE	Indian Institute of Technology, Kanpur	6.32/10.0
2019	CBSE – XII	Pt. Deen Dayal Vidyalaya, Kanpur (CBSE)	89%
2017	CBSE – X	Pt. Deen Dayal Vidyalaya, Kanpur (CBSE)	9.4/10.0

Achievements

- 2nd place in ixigo hackweek'25, built GenAI travel product *2025*
- Won ixigo Premio Award for top contributor Engineers *2025*
- Silver Medal in PS by ISRO hosted by Inter IIT Tech Meet *2022*

Work Experience

ixigo *Aug'24 - present*

Software Engineering 2 - AI Products

- Architected a real-time ASR-LLM-TTS conversational pipeline, solving Voice AI UX challenges like **turn detection and barge-in** using open source **vad** like **TEN/ Silero**
- Built an internal WebRTC and event-driven streaming framework, achieving sub-500ms latency for real-time voice AI
- Did POCs with **ElevenLabs**, **conversational agent pipelines**, and **LiveKit WebRTC**, validating the shift from custom WebRTC streams to a scalable managed real-time media platform
- Implemented custom FFT/PCM audio analysis adding accurate speech boundary detection, **silence trimming**, and **queuing**
- Designed multi agent orchestration layer supporting dynamic tool registration, workflows, proactive triggers, and LLM observability
- Built the Conversational AI **Client SDK** from scratch, a cross platform, event-driven integration layer enabling **realtime multimodal omnipresent AI** across products
- Developed a **low-latency contextual memory system** using Redis and Zep, achieving millisecond-level retrieval for long and short-term context
- Optimized voice AI infra by reducing overall latency 50% and compute cost 40% through semantic caching and streaming optimizations, TTS caching, LLM function tool call optimizations
- Built LLM based post-call analytics, including summarization, structured intents, call scoring, sentiment analysis, and NPS
- Developed ixigo's **public facing MCP server** end-to-end with proxied OAuth 2.0, persistent memory, and production ready reliability, enabling ticket booking, hotel queries, and complex agentic workflows via AI agents

Overlayy AI *Apr'24 - Aug'24*

Fouding Engineer

- Built scalable production-ready LLM backend pipeline from scratch using advanced RAG techniques, improving search latency while managing complete startup technology infrastructure
- Managed comprehensive AWS DevOps infrastructure for multiple microservices, implementing CI/CD automation and integrating AI-enhanced Plausible analytics using Elixir/Phoenix framework
- Developed secure real-time WebSocket messaging system with robust API authorization, architecting custom analytics solution using ClickHouse data warehouse for optimal performance

Previous Internship Experience

Fischer Jordan (offered PPO) *Jun'23 - Feb'24*

Built React.js client management portal with Redux state management, RESTful APIs, and WebSocket real-time features

Llama Mindfulness *Apr'23-Jun'23*

Developed Node.js backend with Slack Bolt integration, configured Moodle LMS plugins, created JavaScript admin portal

Vaticinari Solutions *Jul'23-Sep'23*

Performed cybersecurity threat analysis using Neo4j graph databases, Bloodhound AD mapping, and CVE data processing

C3i Hub, IIT Kanpur *Jul'21-Jul'22*

Integrated Drupal CMS with PHP Composer, built React.js Threatmap frontend using React.js and Express APIs

Key Projects

Reel2Trip *March'25*

ixigo Hackweek'25 - *Live on Production*

- Built **FastAPI** backend integrating **Google Gemini 2.0** and OpenAI GPT-4 for video-to-itinerary conversion system
- Developed Instagram DM method using Facebook **Graph API** with real-time **webhook** processing, currently live production
- Implemented intelligent caching layer reducing AI processing costs by 70% through reel deduplication and optimization
- Created context aware chat system with **ZEP memory** management enabling iterative itinerary editing and personalization

MCP Framework (Auth based) *Jul'25 - present*

Self Project

- Architected a production ready **OAuth 2.0** framework for **MCP servers**, reducing implementation complexity by **90%** for DX
- Built enterprise authentication system with PKCE flow, AES-256 encryption, automated token refresh, and multi-provider support
- Published open-source npm package with comprehensive documentation, TypeScript definitions, and developer-friendly APIs following modular architecture
- Led e2e development of first-to-market OAuth solution for MCP ecosystem, establishing production standards and compatibility

ISRO Web Based X-RAY Burst Automation *Feb'22 - Apr'22*

Inter IIT Tech Meet 10.0

- Developed **web automated system** for identifying **solar** bursts, utilising the **ReactJS framework** with data segregation features
- Organised curve data by saving specific details in **react hooks** in **JSON** to ensure smooth management and API calls to the server
- Led end-to-end **product** documentation and orchestrated the development of a user journey and research, ensuring project success
- Won a **Silver Medal** in the **mid-prep event**, with IIT Kanpur's contingent securing the **2nd position** in the **overall Meet**

Journal Scrapper for Data Mining *Jul'22 - Dec'22*

Supervised By- Proff. Shikhar Krishan Jha

- **Automated** system for fetching **.ris/.bib** files, extracting author details programmatically, and optimizing data retrieval processes
- Proficiently conducted **NER (Named Entity Recognition)** to extract author entities and **DOI URLs** from scholarly articles
- Used Scopus **Cited By API** to implement citation counts of **input journals**, enhancing our tool's functionality & data accuracy
- Seamlessly connected **Flask app** and frontend was established through **HTTP** or other **REST-based API** calls via backend

Skills

Languages: TypeScript, JavaScript, Python, Elixir, C++
Backend & Realtime Systems: Node.js, FastAPI, WebRTC, WebSockets, SSE, RPC
AI / ML: Agent Orchestration, Multi-Agent Systems, RAG, Semantic Caching, LLM Observability, Memory Layer, Vector DBs
Databases: PostgreSQL, MongoDB, SQLite, Neo4j
Infra: Docker, CI/CD, AWS (EC2, RDS, S3), nginx

Positions of Responsibility

Head Web & App, Techkriti'23 *Jul'22 - May'23*

- Led a 400+ cross-functional team across engineering, finance, design, and operations, ensuring alignment and smooth execution
- Re-architected server infrastructure and built 10+ unified portals, improving scalability, reliability, and user experience
- Managed 350K+ monthly traffic and 30K+ daily users, maintaining a 15K+ user database with consistent performance