

SAKETH KANCHI

sakethkanchi3@gmail.com | 201-552-8414

linkedin.com/in/saketh-kanchi | github.com/SakethKanchi | saketh-kanchi.vercel.app

SUMMARY

Full Stack AI Engineer and recent CS MS grad (Stevens, Dec 2025) who ships full-stack product work and production **AI – React/Node** features, hybrid-retrieval **RAG**, multi-provider LLMs, nightly evals, **~80%** host-cost cut, and a **10,000+** install VS Code extension.

EXPERIENCE

Fund Flow OS

Dec 2025 – Present

Full Stack AI Engineer

Jersey City, NJ

- Owned customer-facing features end to end in **React**, **Node.js**, and REST APIs; integrated **Apollo.io** so the product auto-enriches and stores **500+ lead** records
- Built the production **RAG** stack: hybrid retrieval (**pgvector** + BM25, Reciprocal Rank Fusion), cross-encoder reranking, and multi-provider LLM routing (**Gemini 2.5**, **GPT-4o**, **Claude 3.5**) via **LiteLLM**, with a golden-dataset eval harness running nightly in CI
- Cut monthly hosting costs by **~80%** by migrating long-running workloads from **Vercel** to **Fly.io** and redesigning the deployment architecture for production

SideQuest

Jun 2025 – Dec 2025

Software Engineer, Automation

Jersey City, NJ

- Led automation work for a small eng team: shipped internal tools (Instagram data scraper, **Google Apps Script** jobs) that removed recurring manual data entry
- Built an AI quest-generation pipeline (**Foursquare Places API**, **Google Gemini**, Unsplash) with batch processing and retries to fetch, enrich, and categorize points of interest
- Automated **Mapbox** tileset generation from **PostgreSQL/GeoJSON**, powering dynamic location-based quest maps in the product

PROJECTS

Kitty VS Code Theme | *Node.js, VS Code Extension API*

- Shipped a production **VS Code** theme with **10,000+ Marketplace downloads** and a near 5-star rating
- Owns the full release loop: vsce publishing, semantic versioning, accessibility fixes, and support across thousands of installs

Parley | *Node.js, Python, discord.js, SQLite, FastAPI · live demo*

- Open-sourced a self-hosted **Discord** meeting bot that records voice channels, transcribes locally (Python faster-whisper sidecar), and writes AI notes via pluggable LLM backends (Gemini, OpenAI, Ollama)
- Designed per-speaker capture and a Node.js orchestration pipeline with SQLite + FTS5 for searchable history and concurrent multi-channel recording

OPEN SOURCE CONTRIBUTIONS

adi1090x/widgets (800+ stars): Added new color schemes to a popular Linux desktop customization framework

nyxxbit/discord-quest-completer (260+ stars): Ported the relay from TypeScript to **Python**, extending the tool to a new language ecosystem

TECHNICAL SKILLS

Languages & Frameworks: JavaScript/TypeScript, React.js, Node.js, Python, SQL, HTML/CSS

AI / Data: RAG (hybrid retrieval, reranking), LLM Evaluation, LiteLLM, Vector Embeddings, Pandas, Scikit-learn

Databases & APIs: PostgreSQL, pgvector, SQLite, RESTful APIs, GraphQL, Firebase

Cloud & DevOps: Fly.io, Vercel, Cloudflare Workers, Docker, Linux, GitHub Actions, CI/CD

Tools: Postman, VS Code Extension APIs, Playwright, Mapbox, Git

EDUCATION

Stevens Institute of Technology — M.S. Computer Science (GPA: 3.83), Hoboken, NJ, December 2025

Certification: Oracle Cloud Infrastructure 2025 AI Foundations Associate

Gandhi Institute of Technology and Management — B.S. Computer Science (GPA: 8.29/10.0), Hyderabad, TG, June 2023