Sai Pranav H R

Portfolio: saipranav.com — contact@saipranav.com — +91 9740448593 — Bangalore
GitHub:SAP2911 — LinkedIn: Sai Pranav H R

Education

- **JSS Academy of Technical Education, Bangalore** 2021 – 2025
  Bachelor of Engineering in Information Science and Engineering **CGPA: 8.96**

Professional Experience

- **Yellow.ai** – Forward Deployed Engineer **(On Site)** Sep 2026 – Present
  - Building **voice bots** using **NLP**, **speech technologies**, and **LLM-based conversational AI** to automate real-time customer interactions at enterprise scale.
  - Working on the conversational AI stack behind voice agents, including intent understanding, dialogue management, and low-latency response generation.
- **Hitachi Digital Services** – Gen AI Intern **(On Site)** Dec 2025 – Jul 2026
  - Architected a production **multi-agent platform (OFE)** with **150+ specialized AI agents** across **12 operational domains** (Incident Response, SRE, Cloud, FinOps, Security), each with **decorator-based registry**, **confidence-gated output**, and **circuit-breaker-protected LLM calls**; deployed on **Azure** for enterprise IT operations.
  - Designed **MCDA-based cost-aware LLM routing** scoring **4+ models** (**GPT-4o**, **GPT-4o-mini**, **GPT-4.1**, **Claude Sonnet**) across **5 weighted criteria** (cost, quality, speed, context window, availability) with **min-max normalization**, routing trivial prompts to cheaper models while preserving quality for complex reasoning.
  - Built an end-to-end **RAG ingestion pipeline** with **adaptive chunking**, **multi-format parsing** (PDF, DOCX, HTML), **OpenAI embeddings**, **post-ingest chunk validation**, **document version-chain tracking**, and **semantic conflict detection** via centroid-based embedding comparison.
- **Thinkleap** – Junior AI Engineer **(On Site)** June 2025 – Nov 2025
  - Architected end-to-end **Retrieval-Augmented Generation (RAG)** pipelines, optimizing retrieval accuracy using **LangChain**.
  - Orchestrated AI workflows with **LangGraph** to automate document intelligence.
  - Implemented intelligent document processing using **LayoutLM**, **Transformers**, and **PaddleOCR**, enabling accurate structured data extraction.
- **Dom Ventas** – Technical Intern **(On Site)** Sep 2024 – Jan 2025
  - Developed and maintained backend features using **Flask** and frontend components using **Vue.js**.
  - Built web scrapers for automated multi-source data extraction, reducing manual effort.
  - Designed, implemented, and tested RESTful API endpoints for backend services.

Technical Expertise

- **Generative AI & Agents:** LangChain, LangGraph, RAG, Prompt Engineering, OpenAI/Anthropic APIs, HuggingFace, Multi-modal LLMs, Tool-Calling, Structured Output Parsing, Voice Bots / Conversational AI.
- **Vector Computing & Search:** FAISS, ChromaDB, Semantic Search, Hybrid Search (Keyword + Vector), Embedding Models, Data Indexing Strategies.
- **Machine Learning & NLP:** Scikit-learn, SpaCy, pandas, numpy, NLTK, Tesseract OCR, LayoutLM, Transformer Architectures (Attention Mechanisms), NER, Sentiment Analysis, Text Classification.
- **Full-Stack & Backend:** Python (FastAPI/Flask), JavaScript/TypeScript (React/Next.js/Node.js), RESTful API Design, Microservices.
- **Blockchain & Security:** Solidity, Hardhat, Ethers.js, Proxy/Factory Patterns, OpenZeppelin, ed25519 Cryptography.
- **Infrastructure & Databases:** Docker, Git/GitHub Actions, Linux/Bash, MySQL, PostgreSQL, MongoDB.

Key Technical Projects

- **Trendly Support Assistant: Agentic E-commerce Support Agent** GitHub Link
  - Built an agentic customer-support assistant for a D2C fashion retailer that looks up orders, answers policy questions, decides return/exchange eligibility, and escalates to a human, using **Next.js**, **TypeScript**, and **13 Zod-typed tools**. Deployed live on **Vercel**.
  - Separated reasoning from truth: the LLM decides *what to do*, while a **deterministic policy engine** decides eligibility, business-day math, and refund timelines. **Input/output guards** block card numbers, prompt injection, and ungrounded amounts or dates.
  - Added **circuit-breaker failover** across Groq and Gemini, session-gated tool access to prevent identity forgery, and a cassette-based offline eval harness. **315 tests** with **99.7% coverage** on core logic.
- **Mumzworld PDP Generator: Bilingual AI Product Description Pipeline** GitHub Link
  - Built a **4-agent AI pipeline** that takes a single product image and generates publish-ready descriptions in both **English and Gulf Arabic** in under 10 seconds - replacing a manual copywriting step for an e-commerce platform.
  - Agents handle distinct roles: vision extraction, English copy, Arabic copy, and self-critique. Low-confidence fields are flagged for human review instead of being guessed, achieving **0% hallucination rate** with **100% schema compliance** across 15 test cases.
- **Yellow: AI-Powered Bank Customer Support Chatbot** GitHub Link
  - Built a conversational AI assistant for a bank that answers customer questions by searching through official bank documents - a smart FAQ bot that actually understands natural language, not just keywords.
  - Implemented **Hybrid Retrieval** (BM25 + FAISS) so the system catches both exact keyword matches and meaning-based matches; top results are passed to an LLM to generate grounded, cited answers. Includes a **human escalation flow** and adversarial prompt detection to block jailbreak attempts.
  - Built with a **FastAPI** backend, session-based multi-turn conversation history, and a clean chat UI served statically.
- **KeyLinx: USB Hardware Password Manager with Blockchain Auth** GitHub Link
  - Built a physical password manager using a **Raspberry Pi Pico W** that acts like a USB security key - it proves identity using **ed25519 digital signatures**, so authentication requires the physical device and cannot be spoofed remotely.
  - Designed an authentication gateway on the **Sepolia blockchain** using **Solidity smart contracts** (upgradeable proxy pattern) so device verification is decentralized and tamper-proof.
  - Wrote the Solidity contract and deployed it on the Sepolia testnet.