

PARAMVEER SINGH

Full stack developer and agentic ai engineer with a strong backend focus and experience building scalable web applications

☎ +917488287651 ✉ sparamveer1001@gmail.com 💻 paramveer.me 🌐 GitHub 📍 Durgapur, India

WORK EXPERIENCES

Open Source Maintainer / Project Admin

Oct 2024 - Nov 2024

Remote • Remote

GSSOC'24

- Led development of an **open source web platform** with **50+ contributors**, ensuring code quality and scalability.
- Reviewed and merged **50+** pull requests, maintaining clean architecture and stable releases.
- Implemented **CI/CD pipelines** using **GitHub Actions**, achieving consistent deployments and zero downtime.
- Mentored contributors in **React, Node.js, and Git workflows**, improving onboarding efficiency.

Technologies / Skills Used : Node.js, GitHub Actions

PROJECTS

Distributed Multi-VM Inference System on AWS | 🌐 Github

May 2026 - May 2026

- Architected and implemented a distributed model inference pipeline spanning multiple AWS EC2 instances.
- Set up private subnet networking for secure internal communications.
- Developed and integrated an RPC-based framework for worker nodes to communicate and coordinate inference tasks.
- Automated resource provisioning (compute instances, networks, etc.) using Terraform, ensuring replicability and easy environment setup.
- Documented system design, operational debugging strategies, and engineering trade-offs to aid future maintainers.
- Troubleshoot latency bottlenecks imposed by t3.micro compute constraints, clearly articulating trade-offs and performance characteristics.

Technologies / Tools Used : Python, HCL (Terraform), TypeScript, AWS, RPC, Distributed Systems, Systemd, iii

Junior Doctor — AI Clinical Document Processing Agent | 🌐 Github

- Built an AI-powered multi-stage agent pipeline to convert scanned patient records into structured clinical discharge summaries, addressing hallucination risks of naive PDF-to-LLM approaches.
- Implemented OCR fallback (Tesseract) for scanned PDFs alongside PyMuPDF native text extraction, enabling reliable ingestion of real-world medical documents.
- Engineered deterministic reconciliation and conflict detection modules in Python, flagging medication discrepancies and contradictory evidence without allowing the LLM to override clinical data.
- Designed full provenance tracking — every extracted clinical fact stores source document, page number, and source text — ensuring auditability and traceable outputs.
- Maintained explainability via chronological execution trace exports and structured review flag files for clinician oversight.

Technologies / Tools Used : Python, GeminiAPI, PyMuPDF, Tesseract OCR, Pydantic, Agent Architecture

ATLAS RAG API — RAPTOR-Indexed Knowledge Agent | 🌐 Github

- Architected a production-grade RAG API service supporting multi-format document ingestion (PDFs, images, URLs, code) with a fully containerized microservices architecture.
- Implemented RAPTOR (Recursive Abstractive Processing Tree Of Retrieval) hierarchical indexing using UMAP + HDBSCAN clustering with LLM summarization, enabling multi-level semantic search across large knowledge bases.
- Designed a multi-stage retrieval pipeline: vector similarity search (pgvector HNSW) + BM25 full-text search fused via Reciprocal Rank Fusion (RRF) → Jina reranker → Gemini 2.0 Flash answer generation with source citations.
- Integrated IBM Docling as a dedicated Python microservice for high-fidelity PDF parsing, alongside Jina AI v3 embeddings with late chunking for cross-chunk attention context.
- Added full observability stack — Prometheus metrics, Pino structured logging, Grafana dashboards — and Docker Compose orchestration for one-command local and production deployment.

Technologies / Tools Used : Fastify, pgvector, prisma, JinaAI, RAPTOR, UMAP, HDBSCAN, Prometheus

smartnotes-ai | 🌐 Github 🌐 Website

- Developed a robust note-taking application ("smartnotes") primarily using TypeScript (98%+ codebase), ensuring type safety, scalability, and maintainability.
- Architected and implemented core features such as rich text editing, note organization, and search functionality.
- Integrated modern web technologies and best practices for code structure, resulting in a clean, modular, and testable codebase.
- Optimized application performance by identifying and resolving bottlenecks, leading to an enhanced user experience.

Technologies / Tools Used : Next.js, Tailwind CSS, Postgre SQL, Supabase, Google Gemini API, Vercel, TypeScript

Evolve — Resume Analyzer (ATS Scoring) | 🌐 Github 🌐 Website

- Developed an **intelligent resume analysis tool** evaluating resumes based on **8+ ATS criteria**, including keyword density, formatting, and section completeness.
- Extracted structured data and provided **5+ actionable feedback points per resume**; designed a **0-100 scoring system** to simulate recruiter filtering.

Technologies / Tools Used : Next.js, GoogleGenAI

SKILLS

Databases :	MongoDB, PostgreSQL, pgvector, SQL
Frameworks & Libraries :	Node.js, Express.js, Django, React.js, Next.js, Tailwind CSS, Fastify
Programming Languages :	C, C++, Java, Python, JavaScript, TypeScript, SQL, Rust, Solidity
Tools & Platforms :	Docker, Git, GitHub Actions, CI/CD, Linux, AWS, Terraform, Systemd, Prometheus, Grafana
AI / ML :	RAG, RAPTOR, pgvector, Jina AI, Gemini API, LLM Integration, Pydantic, UMAP, HDBSCAN, Numpy, ScikitLearn, Pandas, pytorch

EDUCATIONS

Bengal College of Engineering and Technology | Durgapur

Aug 2023

B.Tech in Computer Science

CGPA : 7.3 / 10

AWARDS & ACHIEVEMENTS

- Solved **500+** DSA problems across competitive programming platforms.
- **LeetCode Max Rating:** 1459
- **CodeChef Rating:** 1300