Swapnil Dutta

swapnil@rycerz.es | LinkedIn | GitHub

Work Experience

Siemens Bengaluru, India
Research Engineer Intern *Oct 2025 – Present*

- Built a data analysis toolkit for electronic component datasets (18k+ rows) with semantic search, keyword search, and hybrid retrieval capabilities. Designed a chat interface with context-aware responses using DuckDB and LLM/embedding models.
- Evaluating 6D Object Pose Estimation models (SAM-6D, FoundationPose, GigaPose) for robotic manipulation applications in ROS2 Humble simulations.
- Exploring Vision-Language-Action (VLA) model architectures for integrating language-based task planning with robotic control in ROS2 environments.

Puch AI Remote
AI Engineering Intern *Sep 2025 – Oct 2025*

- Architected an asynchronous OpenAI wrapper for a WhatsApp chatbot, implementing a robust tool-calling framework supporting 30+ integrated tools with intelligent fallback mechanisms.
- Engineered a full observability suite using Sentry (error tracking), Langfuse (LLM traces), and PostHog (feature flags) to monitor conversation state and agent performance in real-time.
- Built automated test suites for async workflows to ensure reliability across multi-step agent conversations.

Cosdata Remote
Software Engineering Intern (Rust/Vector Search) *May 2025 – Sep 2025*

- Extended a Vector Database engine by implementing Hybrid Search (Dense + Sparse) and batch processing endpoints in Rust (Actix Web), significantly improving retrieval flexibility.
- Benchmarked retrieval performance using BEIR datasets against SOTA algorithms (DiskANN, HNSW), tracking QPS and CPU latency to guide core engine optimizations.
- Implemented memory-efficient chunked streaming (200k vectors/chunk) and remote embedding caching (vLLM), reducing indexing overhead and latency.

Projects & Open Source

Meta-PyTorch/OpenEnv | ***Open Source Contributor*** Dec 2025

- Authored and implemented architectural enhancements for concurrent environments, solving bottlenecks in RL training by designing multi-session support within single Docker containers.
- Standardized environment directory structures and dependency management using `uv` and `pyproject.toml`, streamlining CI/CD workflows for the open-source community.
- Migrated data models to Pydantic and added parameterized environment operations, improving validation, error handling, and API flexibility.

VaultAssist MCP Server | ***MCP, OAuth 2.1, Google APIs*** Aug 2025

- Built a secure Model Context Protocol (MCP) server exposing 50+ Google Workspace tools through OAuth 2.1, utilizing graph-database-backed memory for cross-service context management in personal AI assistants.

Agent Memory System | ***Neo4j, ChromaDB*** Jul 2025

- Designed a hybrid memory framework combining vector search (ChromaDB) and knowledge graphs (Neo4j) to manage semantic and episodic memories, enabling AI agents to retain long-term context.

Education

KIIT University Bhubaneswar, India
B. Tech. in Computer Science and Engineering *Graduation Date: Jun 2026*

Technical Skills

Languages: Python, Rust, Typescript, Go, SQL, C++, Bash
AI & Robotics: PyTorch, ROS 2 Humble, Hugging Face Diffusers, LangChain, LlamaIndex, vLLM
Systems & Cloud: AWS, Docker/Podman, Prometheus, Grafana, GitHub Actions, Sentry, PostHog
Data & Search: PostgreSQL, Neo4j, ChromaDB, Qdrant, DiskANN, HNSW, Xarray, Dask
Tools: Nix, uv, pixi, Actix Web, FastAPI, Next.js