SHAYINI SIDDARDHA

+91 86883 57687 | siddardhashayini3@gmail.com | LinkedIn | GitHub | Portfolio | Hyderabad, Telangana

PROFESSIONAL SUMMARY

AI/ML Engineer focused on resource-efficient AI systems, autonomous agents, and cognitive memory architectures. Creator of OlaiEdge, TAEL-Agent, and HERM4AI, with experience in local LLM optimization, model fine-tuning, agent runtime engineering, and AI research. Open-source developer with research publications in IEEE Xplore and Zenodo.

TECHNICAL SKILLS

Languages: Python, C, C++, Java, JavaScript, TypeScript, SQL, HTML/CSS
AI/ML: LLMs, Agentic AI, Cognitive Architectures, RAG, Knowledge Graphs, Temporal Reasoning, LLM Fine-Tuning (QLoRA/LoRA), NLP, Computer Vision, scikit-learn, TensorFlow, YOLOv12n
LLM & Agent Systems: Ollama, llama.cpp, GGUF, KV-Cache Quantization, Prompt Caching, Grammar-Constrained Decoding, Tool Calling, Persistent Agent Memory
Frameworks & Tools: FastAPI, LangChain, LanceDB, SQLite, Git, GitHub, VS Code Extension Development, REST APIs

PROJECTS

HERM4AI *(Research Published on Zenodo)* | GitHub | Publication — Sep 2026 – Nov 2026

- Designed and implemented a **model-free cognitive memory architecture** addressing **temporal retrieval errors, entity representation collapse, and stale-context hallucinations** in RAG systems.
- Integrated **valid-time reasoning, associative knowledge graphs, 7-slot working memory, CLS-inspired consolidation, and ACT-R-inspired arbitration**; achieved **100% accuracy across 7 benchmark suites**, with **0.065 ms query latency** and **less than 6 MB RAM**.

OlaiEdge *(Open Source • Research Published)* | GitHub | Publication — Jun 2026 – Present

- Built a **headless, embeddable local LLM runtime** with SDKs and a background daemon, integrating **LCP slot caching, 8-bit KV-cache quantization, GBNF-constrained decoding, and RAM-budget enforcement**.
- Reduced steady-state prompt-evaluation latency from **~3,500 ms to ~73 ms** on a **CPU-only, 8 GB RAM system**, achieving **zero OOM failures** in long tool loops; implemented **process watchdogs, multi-tab caching, offline signed licensing, and 26 tests**.

TAEL-Agent *(Local Autonomous SWE Agent)* | GitHub | Hugging Face — Sep 2026 – Present

- Built a **privacy-preserving SWE agent** for **CPU-only, 8 GB RAM systems** by fine-tuning **IBM Granite 3.1 2B Instruct** using **QLoRA** on a **29,900-sample curriculum**, deploying a **4-bit GGUF model**.
- Engineered an **autonomous runtime** with **9-category routing, 10 sandboxed tools, four-pillar memory, and context-budget management**; developed a **VS Code extension** and achieved **8/8 benchmark passes**.

WildGuardAI *(IEEE Published)* | GitHub | Publication — Aug 2025 – Jun 2026

- Developed a real-time wildlife conflict monitoring system using **YOLOv12n, BoT-SORT tracking, Kalman filtering, Flask, and SQLite** to detect animals and generate proximity-based alerts.
- Evaluated the detection model at **91.2% mAP@0.5** and published the research in IEEE Xplore.

Open Source Contributor — sklearn-genetic-opt | GitHub PR #178 — Sep 2025 – Mar 2026

- Documented parallel processing with **n_jobs** in **GASearchCV**; **PR #178** merged, resolving **issue #152**.

WORK EXPERIENCE

AI Engineer Intern | Infosys Springboard — Feb 2025 – Apr 2025

- Built an electricity demand & price forecasting system with Facebook Prophet & a Flask real-time forecast interface.

AI Engineer Intern | YBI Foundation — Dec 2024 – Jun 2025

- Built an Air Quality Prediction System for Indian cities using Random Forest with multi-city data pipelines.

EDUCATION

Bachelor of Technology – Computer Science & Engineering (AI & ML) — 2022 – 2026
Vignan Institute of Technology and Science, Hyderabad | CGPA: 7.72/10.00

CERTIFICATIONS & ASSESSMENTS

- **Oracle:** Agentic AI Certified Foundations Associate (Jul 2026), Oracle Certified Foundations Associate (Oct 2025)
- **TCS iON NQT (IT):** Qualified – Dec 2025 | Strong performance in Python & problem-solving
- **NPTEL:** Programming with Generative AI – Nov 2025
- **Coursera:** Cloud Machine Learning Engineering & MLOps – Dec 2024
- **Infosys Springboard:** DSA, AI Primer, Principles of Generative AI, Generative Models for Developers