Sri Charan Sripadi

AI Engineer | Machine Learning Engineer

Liverpool, UK | +44 7508 594911 | sricharansripadi618@gmail.com | LinkedIn | GitHub | Portfolio | Hugging Face

Available to start now; full time from January 2027 on the UK Graduate visa, no sponsorship needed

SUMMARY

AI engineer who builds LLM agents, fine-tunes small models to evaluate them, and ships them with Docker, CI/CD and monitoring. My fine-tuned 150M-parameter judge beats a 27B LLM judge (88.0% vs 81.2%), and my agent-evaluation scorer was merged into Weights & Biases' open-source toolkit. MSc Advanced AI & Data Science, University of Liverpool; BTech GPA 9.4/10.

SKILLS

Languages: Python, SQL, Java, JavaScript, C, C++
AI & ML: PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, pandas, NumPy, NLP, computer vision
LLMs & Agents: LLM agents, multi-agent systems, Google ADK, tool calling, RAG, LLM evaluation, fine-tuning
MLOps: FastAPI, Flask, REST APIs, Docker, GitHub Actions CI/CD, MLflow, ONNX, Pytest, Git, Linux
Cloud & Data: AWS (EC2, S3), Google Cloud, Azure, SQL databases, FAISS, Chroma; A/B testing, statistics

EXPERIENCE

Independent AI Engineer, Open-Source Agent Tooling and ML Systems | Remote 2025 – Present

- Built and open-sourced three ML systems end to end (TinyJudge, haan, SkillMiner), each answering an open request in Google ADK, AWS Strands Agents or Pipecat, from data and fine-tuning to Docker and CI.
- Shared benchmarks and code with maintainers on public issues (e.g. pipecat-ai/smart-turn #42) with written methods.
- Wrote Pytest suites (unit, integration, offline) so pipelines run in CI without API keys or model downloads.

Open Source Contributor | Weights & Biases, Hugging Face 2025 – Present

- **W&B rai-toolkit:** wrote the TaskCompletionScorer (PR #113, merged), which checks an AI agent's real final state against task success criteria rather than its reply; 93 tests.
- **Hugging Face Transformers:** PR #46777, replaced wildcard imports with explicit ones in TimeSformer.

PROJECTS

TinyJudge: judge model for AI agents | GitHub | *PyTorch, ModernBERT, MLflow, FastAPI, Docker* 2026

- Fine-tuned a 150M-parameter model to catch policy breaches and wrong tool use by AI agents; on 208 real held-out conversations it scored **88.0%**, beating a **27B** LLM judge (81.2%).
- Calibrated its confidence (error 0.20 → 0.10) and shipped it as a FastAPI service with Prometheus metrics and drift signals; CI blocks any release whose accuracy drops below 85%.

haan: turn-taking model for phone voice agents | GitHub | Model | *PyTorch, Whisper, ONNX* 2026

- Voice agents trained on studio audio misjudge when a phone caller has finished speaking; fine-tuned a fix that cut wrongly cut-off callers from **11.09% to 8.64%**.
- Tested on 31,527 clips in 23 languages with paired bootstrap 95% confidence intervals; better in all 23.

SkillMiner: multi-agent system on Google ADK | GitHub | *Python, LiteLLM, scikit-learn, Pytest* 2026

- Multi-agent pipeline that finds workflows an AI agent repeats and writes them as reusable skills; its LLM intent grouping scored ARI 0.93 against 0.48 for a TF-IDF baseline.
- Added PII redaction and an A/B validator; fixed a retry bug that ran refund tools twice; 23 tests.

Speech Emotion Recognition | *Python, TensorFlow, Docker, AWS EC2* 2024

- Classified 7 emotions at 88% accuracy on 10,000 samples, under 300 ms inference; deployed on AWS EC2.

Driver Drowsiness Detection | *Python, OpenCV, Dlib, Flask* 2024

- Real-time computer vision web app served through a Flask API at 24 FPS; 95% accuracy over 200+ sessions.

EDUCATION

University of Liverpool, MSc Advanced Artificial Intelligence and Data Science 2026 – 2027
KL University, India, BTech AI and Data Science; GPA 9.4/10, First Class with Distinction 2021 – 2025

CERTIFICATIONS

Google Cloud Associate Cloud Engineer; TensorFlow Developer Certificate (Google) 2024