

Mohammad Yaser Hussain

mohammadyaserhussain@gmail.com · [linkedin.com/in/yaser0004](https://www.linkedin.com/in/yaser0004) · github.com/yaser0004 · mohammadyaserhussain.com

EDUCATION

Bachelor of Technology | CGPA: 7.62/10 | May 2026

Jayaprakash Narayan College of Engineering (*Affiliated to JNTU, Hyderabad*)

Major: Artificial Intelligence and Machine Learning

PROJECT EXPERIENCE

QuantLM: Native Android LLM Inference Engine | [\[GitHub\]](#)

- Engineered a privacy-first on-device AI engine capable of running state-of-the-art LLMs (Llama, Phi, Qwen) and Vision-Language Models entirely offline on low-resource hardware (2-3 GB RAM).
- Architected a multi-engine inference pipeline, integrating a custom C++17 JNI bridge for llama.cpp alongside Google MediaPipe and TensorFlow Lite to handle low-latency text and multimodal processing.
- Optimized asynchronous model loading and memory management via Kotlin Coroutines, ensuring stable inference and a fluid, non-blocking UI during heavy computation.
- Built a robust local infrastructure utilizing Clean Architecture, Dagger Hilt, and a WorkManager-based system for seamless background downloading and hot-swapping of multi-gigabyte model files.

F.R.I.D.A.Y: On-Device Multimodal AI with Autonomous System Control | [\[GitHub\]](#)

- Architected a fully offline, privacy-first Windows desktop AI assistant integrating local LLMs (Qwen2.5), multimodal vision, STT, and TTS with zero cloud dependencies.
- Engineered a hybrid ML/DL intent pipeline combining LinearSVC (TF-IDF) for near-instant classification of simple intents with DistilBERT (ONNX) reserved for complex queries, dynamically routing via complexity scoring with adaptive CPU/GPU scheduling.
- Built a concurrent PyQt6 application decoupling DL inference, wake-word detection (openWakeWord → faster-whisper → TTS), and OS interactions across isolated worker threads with a structured JSONL telemetry and tracing system for diagnostics.
- Implemented semantic memory via ChromaDB + SQLite with spaCy NLP pipelines for entity extraction and multi-turn context, alongside real-time diagnostic UI with live audio waveform and hardware telemetry panels.

SEC Insider Trading Analysis: Multi-Agent System | [\[GitHub\]](#)

- Architected a multi-agent orchestration system using CrewAI Flow (@router, @listen) to coordinate autonomous agents across data fetching, analysis, and report generation stages.
- Developed custom tools to ingest real-time SEC EDGAR filings (Form 4/8-K) via ATOM feeds, leveraging Pydantic for type-safe configuration and BeautifulSoup for robust XML parsing.
- Implemented data validation guardrails to enforce data integrity and integrated Google Gemini 2.0 Flash to automatically verify and correct generated insight reports against raw filing data.

PUBLICATIONS & RESEARCH

Enlightening Paths: Python's Vision into the Electric Vehicle Market | [\[PAPER\]](#)

Hussain, M. Y., Abrar, M. A., Aiman, M. I., Srinivas, T. A. S. (2024). Journal of Communication Engineering & VLSI Design, 2(2), 28–35. DOI:10.48001/JoCEVD.2024.222835

TECHNICAL SKILLS

Languages & Tools: Python, Kotlin, C++, SQL, CMake, Bash/PowerShell, Git/GitHub, Pydantic, BeautifulSoup, Google Gemini API

AI/ML & NLP: PyTorch, TensorFlow Lite, Transformers (Hugging Face), ONNX, Scikit-learn, spaCy, CrewAI, Model Quantization (GGUF, ONNX)

Frameworks & Runtimes: Jetpack Compose, JNI, Dagger Hilt, MediaPipe, PyQt6, llama.cpp, faster-whisper, ChromaDB

CERTIFICATIONS & TRAINING

- AI Internship, Edunet/IBM Skillsbuild (Nov 2024–Dec 2024)** | [\[CERTIFICATE\]](#)