HARSHITH GUJJETI

Tampa, FL | +1 813-786-8772 | harshithgujjeti@icloud.com | linkedin.com/in/harshithgujjeti | github.com/Harshxth | harshxth.github.io

SUMMARY

M.S. AI & Business Analytics candidate at USF (4.0 GPA) building production GenAI and agentic systems. Shipped an on-prem LLM analytics assistant over 6M+ rows of enterprise ERP data; hands-on with LangGraph multi-agent orchestration, RAG, and evaluation (RAGAS, LangSmith). Prior AI research at ISRO.

EDUCATION

University of South Florida — M.S., Artificial Intelligence & Business Analytics (GPA: 4.0) *Aug 2025 – May 2027*

Jawaharlal Nehru Technological University — B.Tech, Computer Science (AI/ML Specialization) *Nov 2021 – Jul 2025*

SKILLS

GenAI & Agentic: LangGraph, LangChain, CrewAI, Google ADK, MCP, OpenAI API, Gemini, Groq, Ollama, RAG, RAGAS, LangSmith, prompt engineering, LLM fine-tuning

ML & Data: Python, PyTorch, TensorFlow, scikit-learn, XGBoost, SHAP, pandas, NumPy, SQL, Snowflake, Power BI

NLP, Speech & CV: Whisper, ElevenLabs, spaCy, sentence-transformers, YOLOv8, ByteTrack, CNN, transformers

Data & Infra: GCP BigQuery, Airflow, ETL, Docker, FastAPI, PostgreSQL, MongoDB, Supabase, ChromaDB, FAISS, HuggingFace Spaces, Vercel, Apache Spark, HDFS, Kubernetes, MLflow, CI/CD

EXPERIENCE

Reach Cooling Group | *Agentic AI Intern* *May 2026 – Aug 2026*

- Built and piloted (in Microsoft Teams) an on-prem GenAI analytics assistant over 6.2M+ rows of NetSuite ERP data (invoices, POs, forecasts, inventory), serving a local LLM (Ollama, qwen3:8b) with Groq cloud fallback so 98% of queries run fully on-prem; logged 1,000+ production queries with 0 runtime errors and a 0.6% true fallback rate.
- Engineered a safe alternative to NL-to-SQL: the LLM never writes SQL, classifying each question to one of 8 parameterized views (or read-only live NetSuite queries for current inventory), so query-injection and hallucinated-join risk are eliminated by construction, at ~3s median latency on an 8 GB GPU.
- Built a 52-case regression harness (100% passing across filters, dates, follow-ups, refusals, and adversarial inputs) plus an automated twice-daily review loop that replays real questions to catch routing regressions before users hit them.
- Integrated NetSuite via REST/OAuth (keyset-paged order-sync, quarantine-not-drop, nightly open-order refresh) and shipped a read-only Flask monitoring dashboard for data freshness, p50/p95 latency, and fallback health; led enablement sessions teaching directors and the supply-chain team to build agentic workflows.

Center for Urban Transportation Research (CUTR) | *Computer Vision Analyst* *Feb 2026 – Present*

- Built a production YOLOv8 + ByteTrack pipeline over 100+ hours of footage across 50+ camera sites, reaching ~0.90 mAP on a custom 9-class micromobility detector (bicycle / e-bike / seated scooter) for an FDOT-funded SS4A safety study.
- Designed a two-pass motion-ROI + human-review workflow with an OCR-timestamp hard-case mining loop into Label Studio; hand-labeled 5,000+ frames and shipped per-site CSV counts and a methodology report, sharply cutting manual annotation effort.

Muma College of Business, USF | *Data Engineering Research Assistant* *Jan 2026 – Present*

- Built large-scale entity-resolution pipelines (Python, pandas, SQL) linking companies across 325K+ records of messy, real-world research data: fuzzy name-matching across multiple algorithms reconciled in a consensus layer, verified with geographic, industry, and domain signals into tiered, explainable confidence scores for end-to-end data-quality auditing.

Indian Space Research Organisation (ISRO) | *AI Research Intern* *Jan 2025 – Jul 2025*

- Built an SRGAN + VGG19 perceptual-loss super-resolution pipeline (TensorFlow) achieving 4× upscaling at 31 dB PSNR / 0.86 SSIM across 5,000+ aligned satellite image pairs, under the Head of Analytics.
- Presented final methodology and results to 40+ senior scientists in ISRO's Advanced Analytics division.

Venkusa Technologies | *AI Developer Intern* *May 2024 – Nov 2024*

- Built an LLM code-completion tool (OpenAI API with custom prompt chaining and tuning) and benchmarked its suggestions against existing autocomplete tooling across 500+ developer interactions.

PROJECTS

CiteIQ — *Agentic RAG System · github.com/Harshxth/citeiq*

- Multi-agent RAG with dynamic LangGraph routing across 3 decision paths and ChromaDB + sentence-transformers retrieval; reached 90%+ RAGAS faithfulness (0.71 → 0.91 via a self-evaluation retry loop) with full LangSmith tracing and sub-2s latency, containerized with Docker and shipped to HuggingFace Spaces.

CareCall — *Agentic Voice AI for Post-Discharge Follow-Up* (Top 5, HackUSF 2026) · github.com/Harshxth/CareCall

- Built a multi-agent post-discharge patient-follow-up system (Google ADK + Gemini) operating across 70+ languages, with a 9-step care-protocol state machine, auto-escalation to nurses on urgent clinical flags, and a HIPAA-aligned Supabase backend.