Gaurav Chintakunta

Chicago, IL | +1-872-258-2126 | gaurav.pvt25@gmail.com | LinkedIn | GitHub | Portfolio

Education

University of Illinois at Chicago | Master of Science in Computer Science | GPA - 3.89/4.0 *May 2026*
Coursework: Responsible AI Engineering, Neural Networks, Quantum Computing, Deep Learning with CV

JNTU Hyderabad | Bachelor of Technology in Computer Science *May 2024*
Coursework: Machine Learning, Design and Analysis of Algorithms, Data Science, Big Data

Technical Skills

- **Languages:** Python, Go, JavaScript, TypeScript, Scala, SQL, C
- **Frameworks & Tools:** FastAPI, React, Next.js, MCP, Kubernetes, Docker, CrewAI, Git, GitHub Actions, pytest, Linux, Bash
- **ML & AI:** PyTorch, TensorFlow, LangChain, Pinecone, RAG, LLMs, Transformers, PEFT, DSPy, Scikit-learn, PydanticAI
- **Data & Cloud:** PostgreSQL, Pandas, NumPy, Spark, OpenCV, Tableau, AWS, Azure, Prometheus, Streamlit
- **CS Fundamentals:** Data Structures & Algorithms, System Design, Object-Oriented Design, Distributed Systems

Experience

- **Graduate Researcher** *Jan 2025 - Jan 2026*
  University of Illinois at Chicago, Computer Science Chicago, IL
  - Improved spatial prediction accuracy by 30% by designing a Transformer-based GAN pipeline (Pix2Path) processing 2TB+ of spatialomics data on UIC's HPC cluster with optimized Docker workflows.
  - Reduced model training time per epoch by 40% by engineering parallelized patch extraction and memory-efficient data ingestion pipelines using Python and Spark.
  - Improved experiment reproducibility across the research team by building automated TensorBoard tracking and custom checkpointing systems for multi-GPU training runs.
  - Accelerated iteration across 5+ GAN architecture variants by designing modular, reusable OOP pipeline components in Python, eliminating codebase restructuring between experiments.
- **Machine Learning Intern** *May 2023 - Sep 2023*
  AIML Analytics Solutions Hyderabad, Telangana
  - Improved business metric forecasting accuracy by 15% by building and validating Scikit-learn models against real-time SQL data streams.
  - Built RESTful data validation services integrated with SQL backends, reducing data inconsistencies by 25% across model training and evaluation pipelines.
  - Integrated ML model outputs into Tableau dashboards used by 3 cross-functional teams, cutting the decision-making cycle by 20%.

Open Source Contributions

- **Contributor, OpenSRE**
  AI-powered SRE agent framework for automated incident response (1,900+ GitHub stars) View Contribution
  - Redesigned the root-cause analysis prompt to reduce LLM hallucination and diagnostic bias by introducing a structured multi-step reasoning sequence with hypothesis generation, evidence validation, and elimination before classification.
  - Added a structured alternative-hypothesis output section, reducing false-positive root-cause classifications by 30% and updating the response parser for backward-compatible extraction across mixed-signal incidents.

Projects

- **Winnow: AI Email Triage Platform | Python, FastAPI, Next.js, TypeScript, PostgreSQL, Docker** GitHub
  - Built a full-stack email triage platform with a Next.js/TypeScript dashboard and FastAPI backend, featuring tiered ML classification that retrains on user corrections via a nightly learning loop.
  - Deployed to production on Railway and Vercel with Docker Compose orchestration, Alembic migrations, OAuth integration, and a live demo serving concurrent sessions at zero LLM cost.
- **Autonomous Agentic SRE Pipeline | Go, Python, MCP, Kubernetes, CrewAI, Gemini** GitHub
  - Engineered an autonomous SRE pipeline with a Go-based watchdog that polls Prometheus metrics to trigger self-healing workflows in Kubernetes clusters.
  - Built a closed-loop remediation pipeline using CrewAI that analyzes pod logs and executes strategic merge patches to resolve OOM errors without human intervention.
- **ToolGen: Multi-Agent Data Generator | Python, Pydantic, NetworkX, LLM APIs** GitHub
  - Architected a multi-agent pipeline for generating synthetic tool-use conversation data with structured JSON output, constraint sampling, and LLM-as-Judge quality validation.
  - Designed typed Pydantic data models with automatic repair logic, achieving consistent structured output across diverse tool schemas with full test coverage.
- **AI Research Assistant | Python, LangChain, FastAPI, Pinecone, RAG** GitHub
  - Built a production-ready RAG API enabling researchers to query academic documents in natural language, sustaining sub-200ms latency under concurrent load with 95% retrieval consistency.
  - Optimized Pinecone vector indexing and metadata filtering to improve semantic retrieval accuracy across large document corpora, with automated regression testing via GitHub Actions.