Vageesha Datta Ganapaneni

vageeshadattag@gmail.com | linkedin | Github | +1(469) 805-1906 | Dublin,California

Education

- **Master of Science in Computer Science** Aug 2022 - May 2025
  The University of Texas at Dallas *CGPA: 3.6/4.0*
  Relevant Coursework: Machine Learning, Artificial Intelligence, Computer Vision, Natural Language Processing, Statistics in AI and ML, Operating Systems, Design and Analysis of Algorithms

Experience

- **Allen Institute for AI (AI2)** Dallas
  AI Researcher *Jan 2025 - May 2025*
  - Implemented Optimization by Prompting (OPRO) to enhance few-shot performance in LLMs, improving hypothesis generation accuracy by 12% across active reasoning benchmarks.
  - Created belief-tracking and uncertainty quantification modules using entropy and KL divergence to monitor confidence dynamics during multi-step LLM inference.
  - Developed a D3.js visualization platform to analyze the evolution of model beliefs over reasoning chains, streamlining debugging and fostering transparent model behavior.
  - Integrated AutoGen-based reasoning loops to simulate iterative self-verification in LLMs, enabling more consistent and explainable model responses under dynamic inputs.
- **Rocktop Technologies** Dallas
  Software Engineer Intern *Sep 2023 - July 2024*
  - Developed and deployed microservices using PyTorch and Flask to serve fine-tuned LLMs that automated fixed-income data ingestion, reducing analysis time by over 30%.
  - Built scalable Retrieval-Augmented Generation (RAG) pipelines using FAISS and LangChain to enable grounded, domain-specific responses in high-stakes financial workflows.
  - Designed NLP-based natural language query systems with under 100ms response latency, allowing non-technical users to retrieve structured insights from 10+ financial datasets.
  - Prototyped Dockerized, quantized LLM agents with simulated edge deployment and built real-time feedback integration to explore on-device intelligence in enterprise contexts.
- **Computer Vision and Multimodal Computing (CVMC) Lab** Dallas
  Graduate Researcher- UT Dallas *Nov 2022 - Aug 2023*
  - Optimized CUDA kernels and multi-GPU training loops for T2AV, a text-to-audio transformer, achieving a 27% reduction in inference time on real-world audio synthesis tasks.
  - Engineered real-time diagnostic tools for attention heatmaps, spectrograms, and latent vectors to evaluate and explain multimodal alignment and system behavior.
  - Co-developed T2AV-Bench, a distributed contrastive benchmarking framework with GPU fault-tolerance to evaluate cross-modal consistency in generative audio models.
  - Built analysis pipelines to track embedding drift and modality collapse during training, enabling architecture-level tuning through ablation-informed evaluation loops.

Projects

- **Persona Weaver (In Progress)**:
  - Designing a full-stack system using Python and FastAPI to create and chat with customizable AI personas, with multi-trait conditioning and Gemini Pro-powered dialogue generation.
  - Actively building dynamic prompt logic and memory-based multi-turn flow to support persistent, context-aware conversations aligned with user-defined identity, tone, and behavior.
- **InsightBridge:An LLM-powered document analysis tool**:
  - Designed and implemented a recursive text chunking pipeline with LangChain's RecursiveCharacterTextSplitter, enabling efficient vectorization and semantic retrieval from long-form documents.
  - Integrated FAISS-based vector store for low-latency dense retrieval and constructed a Retrieval-Augmented Generation (RAG) chain with ChatOpenAI to produce grounded, context-aware responses.
- **MediQuery:An Instruction-Tuned Healthcare Chatbot**:
  - Developed a modular React frontend with real-time chat interface, integrating complex state management and optimized GPU batch inference for sub-second response times.
  - Implemented backend Flask APIs using FAISS and SentenceTransformers for dense retrieval, enhancing contextual accuracy by 19% via RAG-based instruction tuning.

Technical Skills

- **Languages:** Python, JavaScript, TypeScript, C++, SQL, Bash, HTML, CSS
- **Frameworks:**PyTorch, TensorFlow, Flask, FastAPI, React, LangChain, AutoGen, D3.js, FAISS, Material UI
- **Tools:** Docker, AWS, Git, Jenkins, Prometheus, Grafana, REST APIs, Unix/Linux, Terraform, Kafka, CockroachDB