Suhaas Kiran DG

sdg@umass.edu | LinkedIn | (413)-824-4724

EDUCATION

University of Massachusetts | Amherst, MA **September 2023 – May 2025**
Master of Science in Computer Science
Related Courses: Neural Networks, Advanced Algorithms, Software Engineering, Computer Vision, Deep Generative Models, 3D Visual Computing, Security and Privacy in Generative Models, Information Retrieval

PES University | Bangalore, India **July 2016 – July 2020**
Bachelor of Technology in Electronics and Communication Engineering
Related Courses: Data Science and Data Analysis, Artificial Neural Networks, Machine Learning, Digital Image Processing

WORK EXPERIENCE

STEALTH STARTUP **April 2025 - Present**
Founding AI Engineer Remote

- **Fine-tuned Mistral-7B-Instruct-v0.3** with **DPO** and **LoRA**, and **prompt engineering** for personalized summaries from user metrics.
- Designed score interpretation heuristics and summary logic with co-founders, incorporating iterative feedback to enhance output quality.
- Built RAG pipeline with **Pinecone + FAISS**; integrated **LangChain** agents for tool-calling.
- Conducted **A/B tests** showing 62% win rate for fine-tuned vs. zero-shot models in blind human evaluations, confirming better alignment.
- Deployed scalable LLM APIs using **Docker, vLLM**, and **CI/CD** (GitHub Actions); tracked experiments via **MLflow**.
- Integrated the summary generator into the user dashboard using React (TypeScript), Node.js, and Drizzle ORM.

GRAPHITE.IO **February 2025 - May 2025**
Student Researcher (Industry Mentorship) s Amherst, MA

- Developed a multi-agent based **long-form article generator** using **LangChain** and **RAG**-based expert simulations for content generation.
- Integrated FAISS for retrieval and implemented tool-calling pipelines to allow agents to access external documents and APIs.
- Designed **planner, writer and editor agents** with feedback loops using tool calling and retrieval orchestration.
- Developed a **custom LLM-as-a-judge evaluation framework**, achieving a 70% win rate over GPT-4o and a 5.04-point QA metric gain.

GOOGLE **August 2020 - August 2023**
Cloud Solutions Engineer Bangalore, India

- **Troubleshot and optimized customer ETL pipelines** and workflows, guiding integration of services like Dataflow (Beam), Composer (Airflow), BigQuery and Vertex AI through **hands-on debugging** and **PoC development**.
- Served as SME for Cloud Dataflow and Cloud Composer, supporting production deployments and scaling architectures on Google Cloud.
- Worked with product & sales teams for **root cause analysis**, to **enhance product reliability**, and **build internal docs & productivity tools**.
- Built a **machine-learning-based case sentiment analysis tool** with 76% accuracy to proactively flag low-sentiment cases, reducing SLA misses by 70% and improving resolution time by 27%.
- **Led strategic customer engagement** with Flipkart, ensuring high availability and proactive support during peak-scale events.

PROJECTS

GPU-based large-scale training of Graph Neural Networks (Research)

- Designed a **memory-state** and **dependency-graph** update mechanism for scalable **GNN training** on GPUs.
- Benchmarked NeutronStream and ETC in DGL and PyG, comparing link prediction accuracy and runtime on larger batches.

Zero-shot Visual Word Sense Disambiguation (Research)

- Built baseline models for VWSD using InternVL2 VLM, and context augmentation using LLMs and lexical databases.
- Developed a **novel re-ranking method** for **fine-grained image-text matching** using attribute decomposition and multi-step prompting.
- Achieved an **8.7 percent points increase** in top-1 accuracy for zero-shot matching and **0.8 points increase** in MRR over fine-tuned SOTA.

Movie Recommendation System

- Built **memory-based recommenders** using item-item similarity on Goodreads and MovieLens datasets, and implemented **latent factor** and **Bayesian Personalized Ranking models** for personalization.
- Handled **cold-start** scenarios using **Factorization Machines** with user/item metadata as side information.
- Assessed fairness across gender groups and achieved 12% increase in diversity metric by based on maximal margin relevance and reranking.

Keypoint Detection for Whale Images

- **Automated annotation** of 750+ whale drone images using YOLOv8, reducing manual labeling by 80%.
- Integrated **active learning** loop that improved keypoint detection accuracy by 35% while reducing labeled data requirements by 66%.

Adversarial Patch Defense

- Created targeted adversarial patches on ImageNet and implemented a **patch detection and segmentation** model to cover attacks.
- Improved Top-1 classification accuracy by 83.7% and Top-5 accuracy by 12.5% on adversarial ImageNet test sets.

TECHNICAL SKILLS

Languages: Python, C++, Java, R, SQL, Javascript
ML/AI: PyTorch, TensorFlow, Hugging Face, Scikit-learn, XGBoost, OpenCV, MLFlow, Langchain, CUDA, Deepspeed, vLLM
Data & Infra: GCP, AWS, Spark, Beam, Airflow, Kafka, Docker, Kubernetes, Git
Concepts: Data Preprocessing, Feature Engineering, Model Fine-tuning, Distributed Training, MLOps, Model Inference and Serving

PROFESSIONAL AWARDS & HONORS

- Award from Vice President, Google Cloud Customer Experience, India for successful execution of million-dollars sale event of Flipkart.
- Award for top case closures in Q1, Q2 2021 in EMEA.

PUBLICATIONS

"**Reinforcement Learning for Improving Coherence of Multi-turn Responses in Deep Learning-Based Chatbots**", Proceedings of International Conference on Communication, Circuits, and Systems, 2020