Prajwal Ganugula

+1 480-942-3759 | pganugul@asu.edu | linkedin.com/in/prajwal-ganugula | Tempe, Arizona

Education

Arizona State University — Tempe, AZ, US
M.S. in Computer Science ***(GPA: 4.0/4.0)*** — *Aug 2024 – May 2026*

Indian Institute of Technology, Hyderabad — Hyderabad, India
B.Tech in Computer Science and Engineering ***(GPA: 8.5/10)*** — *Jul 2018 – May 2022*

Technical Skills

Languages & Libraries: Python , C/C++, Java, SQL, PyTorch, TensorFlow, NumPy, scikit-learn, pandas
GenAI: LLMs, RAG, LangChain/LangGraph/LangSmith, MCP/A2A agents, multi-agent orchestration, Transformers
Cloud: AWS, Azure, DevOps, CI/CD, infrastructure, monitoring, APIs, microservices, Docker, Kubernetes
Systems: arch., Agentic AI, Responsible AI, message queues(RabbitMQ, Kafka), cross-system,

Work Experience

AI Engineer (Data Infrastructure & Platforms) — Oct 2025 – Present
Everest Global Insurance — *Remote, US*

- Built a production RAG-based claims chatbot, orchestrating Kafka/Airflow pipelines on Azure to ingest, monitor, and preprocess multi-source claims data for **1,200+** underwriters globally.
- Designed LLM multi-agent workflows to detect and repair missing or inconsistent claim fields, improving claims data quality and reducing human-in-the-loop underwriting review time by **25%**.

Data Scientist — Dec 2024 – Sep 2025
ASU Decision Theater — *Tempe, AZ, US*

- Built and scaled a RAG-based LLM chatbot over **100K+ research articles**, improving answer groundedness and relevance through hybrid search (dense + sparse), adaptive chunking, metadata filtering, and prompt engineering.
- Designed automated evaluation pipelines for retrieval and generation quality, measuring **faithfulness, answer relevance, context relevance, and completeness** using LLM-as-a-judge, without gold labels.

Software Engineer - Deep Learning Research & Innovation — June 2022 – Aug 2024
OPLUS (OPPO-OnePlus) Research and Development — *Hyderabad, India*

- Optimized LLMs for Edge SoCs using FlashAttention, domain-specific QLoRA, and INT8 post-training quantization, enabling **sub-100ms** on-device chat under strict memory limits.
- Designed a mobile **AI eraser** pipeline with Edge-SAM for object selection and a ControlNet Stable Diffusion inpainter for background inpainting (demo).
- Compressed the Stable Diffusion UNet backbone with soft/hard distillation, post-training quantization, and block pruning, cutting model size by **50%** and speeding inpainting by **8x** on phones.

Projects

Structured Image Captioning with Distributed Training (Multi-GPU) — Mar 2025 – May 2025

- Designed a multimodal pipeline (BLIP-2 + ViT-G) fine-tuned with QLoRA; optimized training efficiency on limited compute using memory-efficient attention techniques.
- Implemented distributed training strategies using DeepSpeed ZeRO-3 and FSDP, leveraging NCCL for efficient inter-GPU communication to scale model performance.

Multi-Agent RAG Copilot for Cloud Operations - Personal Project — Aug 2024 – Nov 2024

- Built an agentic multi-agent RAG copilot in Python/FastAPI for SRE and DevOps on-call engineers, using LangChain, LangGraph and MCP-style tools so incidents can be diagnosed via a single chat interface.
- Orchestrated A2A agents with vector database of runbooks, tickets and postmortems plus live AWS/Azure logs and metrics, deploying with Docker and CI/CD to replace manual dashboard-hopping with guided incident workflows.

Technical Writing & Publications

MOSAIC: Multi-Object Segmented Arbitrary Stylization Using CLIP — **Prajwal Ganugula**, et al.
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023

Application for Produced Crop Price Forecasting Through Deep Learning — **Prajwal Ganugula**, et al.
International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), Nov 2023