ANMOL GAUTAM

+91-8447063045 anmolgautam2428@gmail.com www.linkedin.com/in/anmolgautam28 https://github.com/anmolgautam Bengaluru

SUMMARY

Lead Applied Scientist with **4+ years** of experience architecting and shipping production-grade AI systems from **multi-agent platforms** and **Text-to-SQL engines** to e**nterprise RAG and real-time multimodal solutions. Gold Medalist** (NIT Meghalaya, 10.0 CGPA) with **6 published papers** (IEEE, Springer, Arxiv). Proven ability to take complex AI projects from research prototype to scalable production.

EXPERIENCE

Lead Applied Scientist - AI/ML

8bit.ai

10/2024 - Present Bengaluru, India

- Architected Neutrino, a **multi-agent AI platform** powering **enterprise search, Text-to-SQL, and workflow automation,** built on FastAPI/SSE with **human-in-the-loop execution**, **observability**, and multi-LLM orchestration; deployed across 5 major ISV partners.
- **Fine-tuned domain-specific** LLMs using **LoRA**, **DoRA**, PEFT, and alignment techniques (**DPO**, **GRPO**) for partner use cases, improving task accuracy and response quality while reducing training compute and enabling rapid model customization across deployments.
- Built a **multi-schema Text-to-SQL** engine using agentic **ReAct** workflows across **PostgreSQL** and **Trino**, enabling natural-language querying over partner-specific enterprise data.
- Designed **end-to-end enterprise search spanning RAG ingestion, hybrid retrieval, knowledge-graph augmentation**, PII tagging, and **multi-tenant data discovery,** adapted and delivered for each partner's unique data landscape.
- Optimized inference latency and cost through benchmarking, pruning, and **quantization**, used **vLLM** and **SGLang**, delivered real-time multimodal solutions including voice and sign-language AI for partner deployments.

Applied Scientist - AI/ML

SuperAGI

11/2023 - 10/2024 Bengaluru, India

- Built Text-to-SQL and RAG-based conversational multi-agent systems for SuperSales.
- **Developed SuperCoder2.0**, a **multi-agent** autonomous code navigation and issue-resolution system, **achieving 33% on SWE-Bench-Lite** using custom RAG and code generation.
- Architected a fully autonomous multi-agent platform using open-source and closed-source LLMs for task mining and **ReAct**-style task execution, and **Planner** and **Orchestration** Pattern for task decomposition, finally, taking projects from PoC to AWS production.
- **Developed SAM-7B,** an instruction-tuned Mistral-7B model with custom datasets and evaluation pipelines; achieved GPT-3.5-comparable performance and outperformed Orca on GSM8K and ARC despite training on a much smaller dataset.

Associate Consultant

Oracle

08/2022 - 10/2023 Bengaluru, India

- Built document AI and information extraction pipelines using OCI Document Understanding and EasyOCR, improving NER and key-value extraction by 7%.
- Developed RAG-based question answering and computer vision systems using Falcon, Llama, ChromaDB, TensorFlow, and transfer learning, including a face recognition pipeline that improved performance by 37%.

Research Intern

NVIDIA

05/2021 - 04/2022 Bengaluru, India

- Worked on **NLP and computer vision systems** using **NVIDIA NeMo**, **HuggingFace**, and transfer learning, including **English-to-Hindi machine translation**, object detection, and image segmentation.

Machine Learning Intern

Gahan AI

01/2022 - 05/2022 Bengaluru, India

- Built a **teaching vs non-teaching video classification system** for e-learning using a custom dataset and a **3D CNN-RNN** architecture; deployed via Flask/REST and received **Best Paper Award at CVMI 2022**.

EDUCATION

M.Tech. Computer Science and Engineering

National Institute of Technology (NIT), Meghalaya

2020 - 2022 Shillong, Meghalaya, India

CGPA 10 / 10

- **Gold Medalist in Academics**
- **Institute Best Masters Thesis Award** - Region of Interest Segmentation in Biomedical Images

PUBLICATIONS

SuperCoder2.0: Technical Report on Exploring the feasibility of LLMs as Autonomous Programmer

Arxiv

https://arxiv.org/abs/2409.11190

Veagle: Advancements in Multimodal Representation Learning

Arxiv

https://arxiv.org/abs/2403.08773

SAU-NET: Scale Aware Polyp Segmentation using Encoder-Decoder Network

IEEE

https://ieeexplore.ieee.org/abstract/document/9864338

ED-NET: Educational Teaching Video Classification Network

Springer

https://link.springer.com/chapter/10.1007/978-981-19-7867-8_12

Batch Image Encryption and Compression using Chaotic Map Infused Autoencoder Network

IEEE

https://ieeexplore.ieee.org/document/9986385

Li-SegPNet: Encoder-Decoder Mode Lightweight Segmentation Network for Colorectal Polyps Analysis

IEEE

https://ieeexplore.ieee.org/document/9926143

PROJECTS

Dendrux - Open Source Runtime

https://github.com/dendrux/dendrux

Dendrux: Building an open-source runtime **for real-world agents with tool calling, persistence**, **observability**, FastAPI/SSE hosting, and a **client-tool bridge** for pause/resume execution.

Region of Interest Segmentation in Biomedical Images

2021 - 2022

MTech Thesis in Collaboration with Nvidia

- A**chieved SOTA result published in IEEE.** Improved UNet and outperformed Deep Lab Family, FPN Net. **Received Institute Best Master's Thesis Award**

SKILLS

AI / LLM Systems

Agentic AI, Multi-Agent Systems, RAG, Text-to-SQL, Fine-tuning, Evaluation

ML / DL

PyTorch, Hugging Face, Transformers, NLP, Computer Vision

Backend

Python, FastAPI, Go, Node.js, REST APIs, SSE, WebSockets

Data / Retrieval

PostgreSQL, MongoDB, Trino, ChromaDB, Milvus, pgvector

Infra

Docker, AWS, Azure, MLflow, vLLM, SGLang, Quantization