ANMOL GAUTAM

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

EXPERIENCE

Lead Applied Scientist - AI/ML

8bit.ai

10/2024 - Present Bengaluru, India

- **Neutrino -** leading the development of Multi agentic **workflow automation system** combining Information Retrieval using APIs as tools, text to sql and RAG workflows.
- Building ETL pipelines for data migration and normalisation of structured data.
- **Serverless Inference -** research and development for building serverless inference for 8bit GPU cloud. Set up vLLM inference engine using Ray for autoscaling.
- Set up model quantization and model pruning pipelines for faster throughput and inference speed.

Applied Scientist - AI/ML

SuperAGI

11/2023 - 10/2024 Bengaluru, India

- Build T**ext to SQL and RAG** based Conversational Multi Agentic System for **SuperSales.**
- Built **SuperCoder2.0 - achieved 33% on SWE-Bench-Lite,** multiagentic autonomous code navigation and code issue resolving system. Using custom RAG solution and code generation.
- **Fully Autonomous Multi Agent System**, a conversational **GenAI** powered platform built using **Multi Agent System (MAS).** Used different **Open-Source LLM** (**Llama series, Mistral and Mixtral**) and Closed Source LLMs (**OpenAI - GPT4, Claude**) for **Task Mining**, **Task Execution** (**ReAct Agent**) from scratch.
- Applied various **Fine-Tuning** and **Alignment** strategies (**SFT, DPO, ORPO, model merging**) to get the desired results. Developed custom **evaluation techniques** using Multi agent system to automate the testing process for evaluation of fine-tuned LLMs.
- **Complete ownership of the project from Researching the PoC**, experiments and taking the final design to production through collaboration with CEO, CTO and Engineering team to deploy the MAS in production on AWS using Agile development.
- **Development of SAM -7B (Instruct Fine-Tuned Mistral -7B) - Model Card (3200+ Downloads on HuggingFace), Dataset, Blog-** Build SAM by fine-tuning Mistral 7B base model using **LoRA** and **PEFT** on **dataset generated by only Open-source models**. Created custom dataset by **generating Explanation Traces** using other open-source models – **Yi-34B, Falcon-40B and Mixtral.** Explanation Traces followed **Chain of thought** schema to allow **knowledge distillation** of reasoning capacities of larger LLMs to be incorporated into smaller LLMs. SAM achieved performance **comparable to GPT 3.5 and outperformed Orca on GSM 8k and ARC challenge despite being trained on 97% smaller dataset.**

Associate Consultant

Oracle

08/2022 - 10/2023 Bengaluru, India

- **Automated Invoice Processing -** Prepared dataset using OCI data labelling services. Performed **NER**, Key Value Extraction using OCI Document Understanding services. **Fine-Tuned EasyOCR** to **improve performance by 7% on NER and Key-Value extraction**.
- **Closed Domain Question Answering system and Document Search for large Corpus using RAG -** Built a custom solution using **Falcon-7B, Llama-13B** and sentence transformers (**BERT**, **MPNET**)to create vector embeddings. Utilised **Chroma DB** as vector store, used **Cohere LLM API** to perform **semantic chunking**.
- **Custom Face Recognition system -** Build and deployed a complete Face Recognition system using Open-Source Models and **Dlib**. Built isolated VM using Anaconda environments to perform inference in real-time. Used **keras-vggface** (**ResNet-50)** and **MTCNN** to detect faces. Used **Transfer Learning in TensorFlow** by fine-tuning the model on custom data to **improve performance by 37%.**

Research Intern

NVIDIA

05/2021 - 04/2022 Bengaluru, India

- Used **Nvidia Nemo (Natural Language Processing)** framework to **develop English to-Hindi Machine Translation application** using Hugging Face pre-trained models for machine translation tasks. Worked on different Computer Vision tasks including Object detection, Image segmentation for different use cases using transfer learning, data augmentation and test time augmentation. Also worked on ML problems using scikit-learn, pandas and numpy.

Machine Learning Intern

Gahan AI

01/2022 - 05/2022 Bengaluru, India

- Managed a project **to classify teaching and non-teaching video** for the e-learning platform. Created a video classification dataset from scratch. Developed a novel
3D CNN – RNN based model to classify videos. Model Quantization and deployment of the model using Flask and REST API. **Won Best Paper Award at CVMI 2022.**

EDUCATION

M.Tech. Computer Science and Engineering

National Institute of Technology (NIT), Meghalaya

CGPA 10 / 10

2020 - 2022 Shillong, Meghalaya, India

- **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

Region of Interest Segmentation in Biomedical Images

2021 - 2022

MTech Thesis in Collaboration with Nvidia

- Dataset Creation, Biomedical Data Processing, Data Augmentation, Data Visualization, CNN Models and implementation, and Performance analysis, **achieved SOTA result published in IEEE.** Improved UNet and outperformed Deep Lab Family, FPN Net.
- **Received Institute Best Master's Thesis Award**

SKILLS

Generative AI | Deep Learning | NLP | LLM | Pytorch | Python | Machine Learning | MultiModal LLM | HuggingFace | BERT | Transformers | AI Engineer | Conversational AI