Marathalli, Bangalore,
Karnataka,India, 560037

PANKAJ DEB ROY

(+91)8967405776 | pankajdebroy2000@gmail.com
LinkedIn | GitHub | Portfolio

SKILLS

- Python | C++ | Keras | PyTorch | Scikit-learn | Pandas | Matplotlib | Seaborn | SQL | Numpy | Docker
- Machine Learning | Deep Learning | Linear Regression | Logistic Regression | XGBoost | SVM | CNN | Generative AI (GenAI)
- Natural Language Processing | Transformers | Large Language Model | LLAMA | T5 | Bart | Mamba | GPT | LoRA
- Quantization | Pruning | Knowledge Distillation
- RHLH | DPO
- Training : DDP | FSDP | DeepSpeed

PAPER & PATENT

- **[Patent]**
  - **A method for unlocking One-For-All Large Language Models on Edge.**
  - SIPMS No : **WI-202504-022-1-IN0**
  - Grade : **A1**

CERTIFICATION

- Convolutional Neural Networks [link]
- Neural Networks and Deep Learning [link]
- 2022 Complete Python Bootcamp From Zero to Hero in Python

AWARDS & ACCOMPLISHMENT

- Cleared **Samsung Professional Exam**
- Samsung Excellence Award [For integrating GenAI Feature on Tab S10+]
- Spot Award for LLM Model deployment in S24
- Led two business trips to **South Korea** for the successful integration of **GenAI** (LLM 3B) models on **Samsung S24** and **Tab S10** devices.
- Best Team Award (From TCS)

EMPLOYMENT

Samsung Research Institute, Bangalore

Machine Learning Engineer specializing in LLM fine-tuning, model quantization. Designed and optimise LLM inference pipelines.

Senior Machine Learning Engineer — **October 2022 - Present**

- **Developed Text Summarisation** model using **T5 Base, achieving 0.46 ROUGE** scores on **Article Summarisation.** Implemented model **deployed** on **S25** device reducing **inference time** by **60%.**
- **Implemented parameter-efficient fine-tuning [PEFT]** techniques, **LoRA, QLoRA,** to adapt a pre-trained **LLAMA:2B** model for samsung specific **health** use-case and next line prediction use-case for **samsung keyboard**, achieving a **96%** accuracy rate.
- Instruct Fine-tuned **Qwen-32B** model on large context size **[8K]** for **Code Generation** task with **Reward Architectures** [**RLHF,DPO]**. Trained the model with multi-gpu parallalization [**DeepSpeed]**.
- **Automated Human Annotation** of **Code Dataset** by generating **Reward Model** with **LLMs.** Which improves **data-labelling** by **90%.**
- **Fine-tuned Llama 3B** model on a proprietary dataset of **150,000** samsung specific transcripts to create a specialised chatbot. Resulted in a **20% increase** in **user satisfaction scores.**
- Performed **8-bit quantization** for **MTK NPU** chipsets, achieving **98.41% quantization accuracy** and **reducing peak memory [RAM] by 84%**. Experimented with various tokenizers and samplers to further enhance accuracy.
- **Implemented Fixed Shape KV-Cache** management, significantly **improving memory** efficiency for quantised model on **Qualcomm NPU.**
- Implemented **GitHub Actions** for automated model testing and Rogue score validation, improving deployment reliability.
- Worked on fine-tuning Noise Removal model for custom use-case and quantization of the model for **Exinos NPU** of edge devices like **S24** and **A54** which improves **10% of Model Inference timing** and successfully achieved **~95% accuracy over base model.**

TATA Consultancy Services LTD

Assistant System Engineer Trainee — **August 2021 - October 2022**

As a Manual Tester, I worked on functionality testing of L&D tool of our client, ensuring that it meets functionality and performance as per requirements. We have developed and executed application specific teset-cases to ensure they met functional and performance requirements.

PROJECT

Paper Implementation

- Implemented Computer Vision models from papers.
- Models : **GoogleNet**, **ResNet**, **VGGNet**, NeuralStyleTransfer using **Unet**
- Skills : Pytorch, Matplotlib, Machine Learning, Deep Learning, Neural Network, Convolutional Neural Network.
- Project Link: **https://github.com/yotaAI/ConvNets**

CarO

- Implemented a mobile application which will allow to book vehicles (Toto/Auto) inside a very large complex / society. Which will solve the complexity of finding small vehicles at any time whenever people want to travel inside the premises.
- Made with: **HTML, CSS, JavaScript, React, Firebase**
- Project Link: **https://github.com/yotaAI/car-o**

EDUCATION

RCC Institute of Information Technology — 2017-2021
B.Tech in **Electrical Engineering** — 8.14