Simran Singh Sandral

+91-9502732909 | simransandral05@gmail.com | Simran32909 | simraann | Simran's Portfolio

Education

GITAM University 2022 – 2026
B.Tech in Computer Science and Engineering (CGPA: 7.8/10.0) *Hyderabad, India*

Experience

Center for Visual Information Technology | IIIT-H March 2025 - Present
Research Intern - Under Prof. Ravi Kiran *Hyderabad*

- Developing an advanced HTR system for low-resource historical scripts with irregular layouts, from real data collection to large-scale synthetic data generation.
- Architecting and training a novel attention-based Vision Transformer model as part of a generalized pipeline designed to handle recognition tasks for any low-resource script.

Language Technologies Research Centre | IIIT-H August 2024 - March 2025
Research Intern - Under Prof. Vineet Chaitanya *Hyderabad*

- Engineered a segmentation system for Sanskrit using the ByT5 model, boosting model performance by 30% by implementing a more efficient decoding algorithm.
- Achieved 90% alignment with expert linguistic standards in Sanskrit word sense disambiguation by evaluating and integrating various LLMs (Llama-3.1-8B, Gemma-2).

Projects

ExplainIt-Phi: An End-to-End Fine-Tuned LLM [|]
Phi-2, PyTorch Lightning, Hydra, Transformers, llama.cpp, GGUF

- Engineered a specialized ELI5 model by fine-tuning Microsoft's 2.7B parameter Phi-2, using the memory-efficient QLoRA technique to achieve a 20% improvement in text simplicity based on Flesch-Kincaid readability scores.
- Deployed the fine-tuned model in GGUF format for llama.cpp compatibility through a custom quantization pipeline, achieving over **250+ community downloads** on Hugging Face within the first week of release and enabling efficient local inference capabilities.

Cloud-Based IDE Platform [ ]
Python, Flask, Docker, JWT, SQLAlchemy, Git

- Engineered a full-stack cloud IDE featuring a containerized execution environment using Docker to ensure security and dependency isolation.
- Built a scalable, multi-tenant file system supporting real-time code execution in multiple languages with strict resource limitations.

CodeLLaMA [ ]
LLaMA, Transformers, PEFT, LoRA, Flask

- Reduced model training time by 60% by fine-tuning a LLaMA 3.2-1B model for code generation using Low-Rank Adaptation (LoRA).
- Developed a Flask-based web interface with an optimized preprocessing pipeline to handle complex coding prompts and generate accurate code.

Technical Skills

Languages & Web: Python, JavaScript, C++, Java, Springboot, TypeScript, Go, HTML/CSS, React, Flask, SQL, REST APIs, JAX, Ruby on Rails, Next.js, Node.js, Express.js, Three.js
ML & AI: PyTorch, TensorFlow, TensorBoard, OpenCV, Albumentations, Fastai, PyTorch Lightning, DL4J, DeepSpeed, Hydra, TensorRT, Augraphy, Wandb
Tools & Libraries: Git, Docker, NumPy, Pandas, Scikit-learn, CLIP, FAISS, Matplotlib, Seaborn, ONNX, Hugging Face, tmux, Bash, Streamlit, Linux, IntelliJ IDEA, llama.cpp, GGUF
Research Skills: Data Analysis, Statistical Modeling, Research Methodology, Technical Writing