Vineet Telang

Microsoft Certified AI Engineer
linkedin.com/in/vineet-telang

Email : telangvineet@gmail.com
Mobile : +91-9987559734
github.com/Vineet314

Education

College of Engineering, Pune Pune, India
Bachelor of Technology in Mechanical Engineering; CGPA: 7.73/10 *[Nov 2020 – Jun 2024]*

DAV Public School (CBSE) Navi Mumbai, India
PCM with Computer Science in Python; Grade: 94% *[July 2020]*

Experience

Ednius AI Vancouver, Canada (Remote)
AI Engineer | Generative AI & Cloud Computing *[Mar 2025 - Present]*

- Refining an **Agentic AI** approach to automate examination grading, leveraging **Microsoft Azure.**
- Developed robust end-to-end **APIs** using **Azure Functions**, establishing a seamless pipeline for grading PDFs .
- Implemented **asynchronous programming** to mitigate network I/O bottlenecks, substantially improving efficiency.
- Deployed custom-trained **YOLO-based models** for diagram detection (93.6% accuracy) and smart-cropping, and integrated a **CNN-based model** for false-positive filtering (96.5% accuracy).
- Implemented **Agglomerative Clustering** and **text embeddings** to accurately group answers, achieving a **silhouette score of 0.66**. Visualized English sentences from embeddings, **employing PCA and t-SNE**.
- Containerized applications with **Docker** and deployed them to production environments via **CI/CD pipelines**.

COEP Technological University Pune, India (Hybrid)
DL Research Associate | Distributed Deep Learning & High Performance Computing *[Feb 2025 - Present]*

- Conducting research on **distributed DL** techniques to enhance the scalability and efficiency of LLMs.
- Applied cutting-edge advancements *viz.* **Flash Attention, MHLA, GRPO** to improve the performance of LLMs.
- Developed a custom small-scale LLM and trained it on a local HPC cluster **- Nvidia DGX** with **4x V100 GPUs**
- **Mentored and guided** junior researchers/interns in the implementation and training of LLMs.

Reliance Industries Navi Mumbai, India (On-Site)
Graduate Engineering Trainee | Mechanical Engineering *[Aug 2024 - Feb 2025]*

- Drew insights by **analyzing data**, performing **Root Cause Analysis** for performance improvements.
- Acquired practical and industrial skills, utilized SAP for **Enterprise Resource Planning** (ERP) systems.

Projects

Distributed Model training *[Mar 2025 - Present]*

- Working on training LLMs, *viz.*,**GPT2** on single-GPU & advancing towards **OpenR1**, on multi-GPU.
- Implemented **data/model sharding**, **hybrid parallelism** for scaling LLM training on singe-node systems.
- Adapting and exploring various **PEFT techniques** to fine-tune LLMs comprising several billion parameters.

Retrieval-Augmented Generation (RAG) with Low-Level Tools *[May 2025 - Present]*

- RAG across four distinct approache. *EasyRAG* - Fast and accurate results employing Gemini Files API, Gradio.
- *MedRAG* - Customization using **LangChain**, *HardRAG* - **ChromaDB**, OpenAI for retrieval, vectors & generation.
- *'RAG from Scratch'*- No tools except Numpy for (dense) retrieval, HuggingFace for embeddings & generation.

Smart Boiler Modelling *[Nov 2023 - May 2024]*

- Developed a mathematical model incorporating **Support Vector Regression** (SVR), predicting efficiency.
- **Enhanced** model performance by integrating physics-based insights, achieving a prediction error rate of just 5%.

Skills

Python: DeepSpeed, `torch.distributed`, PyTorch, TensorFlow, HuggingFace, LangChain, OpenAI SDK, Sklearn, `asyncio`, FastAPI, Flask, WandB, MLFlow, Seaborn, Matplotlib, Pandas, Numpy

Technologies: Azure, GCP, Bash, Git, GitHub/Actions, Docker, LaTeX, SLURM, PowerBI, MySQL

Certifications

Microsoft: Microsoft Certified Azure AI Engineer

Coursera: ML/DL Specializations, Advanced ML on GCP, Microsoft AIML, MS PowerBI Specialization

Awards: Innovation and excellence UG Project Competition award