Shreyansh Singh

+91-9654070844 shreyansh.pettswood@gmail.com shreyansh26 shreyansh26 Homepage

Education

IIT (BHU) Varanasi, India **Jul 2016 – May 2020**

Bachelor of Technology (B.Tech.) in Computer Science and Engineering *CGPA - 9.57/10.0*

Member of the Club of Programmers and founder of the Information Security Club

Experience

Level AI **Jan 2022 – Present**

Lead Machine Learning Engineer *(Remote - India) Mountain View, California*

- Leading the efforts on training a large-scale, in-house Instruction-Finetuned model (FLAN-T5-XXL (11B), Mistral-7B) on 8xA100 80GB GPUs, utilizing Fully Sharded Data Parallel (FSDP) and activation checkpointing for efficiency.
- Curated high-quality, diverse, multi-task instruction datasets leading to the model surpassing all internal task-specific NLP benchmarks and demonstrating robust zero-shot capabilities on new unseen tasks.
- Deployed the trained models efficiently using TensorRT-LLM and vLLM, optimizing for high performance and scalability. Currently experimenting with faster shared-prefix batch decoding techniques.
- Built the complete "Voice of the Customer" (VoC) product, using a generative model to summarize customer feedback from contact center conversations, followed by a three-level classification into product/service, issue type, and themes.
- Enhanced customer insight accuracy with the VoC solution, achieving 91% correctness and 87% quality in customer concern summarization, while maintaining a low latency of 300ms per conversation at peak load.
- Created AgentGPT, a tool for contact center agents to search for solutions to customer queries by simply asking in natural language. It sources answers from FAQs, articles, and historical conversations.
- Developed a pipeline to generate the resolution steps taken by an agent in a conversation. This allows the agent query to be searched against historical concerns and resolution steps to provide the relevant solutions with a sub-100ms latency.

Mastercard - AI Garage **Aug 2020 – Jan 2022**

Data Scientist *Gurugram, Haryana*

- Developed a graph-based representation learning algorithm for fraud detection in transactions, achieving a significant 6% increase in AUCPR and demonstrating a good tradeoff in training time vs. performance, compared to existing methods.
- Created a memory-efficient tabular GAN architecture, MeTGAN, which reduces memory usage by $\approx 80\%$ compared to the state-of-the-art model (at the time) specifically on datasets with high cardinality columns, without any drop in performance.

Samsung Research Institute - Bangalore **May 2019 – Jul 2019**

Research Intern *Bengaluru, Karnataka*

- Implemented and simulated the MAS5G architecture, a new 5G mobility scheme, published in IEEE FiCloud, 2019.
- Locally deployed and tested a proof-of-concept version of the architecture using Node.js, Cassandra and Kubernetes.

C3i Center, IIT Kanpur **Dec 2018 – Jan 2019**

Research Intern *Kanpur, Uttar Pradesh*

- Developed a system to classify Linux executables as malware or benign using static and dynamic analysis techniques.
- Achieved $\approx 96\%$ accuracy for the task and deployed the entire pipeline on their internal Malware Analysis system.

Innoplexus AG **May 2018 – Jul 2018**

Data Science Intern *Pune, Maharashtra*

- Developed a new OCR + NLP pipeline from scratch to extract and label segments of text from PDFs. Completely revamped the existing pipeline to make an 80% faster and more accurate ($\approx 92\%$) system.
- Experimented with image processing methods and Faster-RCNN model for detection and extraction of tables from PDFs.

Publications

- **MeTGAN: Memory Efficient Tabular GAN for High Cardinality Categorical Datasets** - ***Shreyansh Singh****, Kanishka Kayathwal, Hardik Wadhwa and Gaurav Dhama* at the 28th International Conference on Neural Information Processing (ICONIP), 2021
- **CuRL: Coupled Representation Learning of Cards and Merchants to Detect Transaction Frauds** - *Maitrey Gramopadhye*,* ***Shreyansh Singh****, Kushagra Agarwal, Nitish Srivasatava, Alok Singh, Siddhartha Asthana and Ankur Arora* at the 30th International Conference on Artificial Neural Networks (ICANN), 2021 (* ≡ Equal contribution)
- **IIT (BHU) Varanasi at MSR-SRST 2018: A Language Model Based Approach for Natural Language Generation** - ***Shreyansh Singh****, Avi Chawla, Ayush Sharma and A.K. Singh* in Proceedings of the 1st Workshop on Multilingual Surface Realisation at the 56th Association for Computational Linguistics (ACL), 2018

Projects

Accelerating LLM training and inference using Triton **Jan 2023 - Present**

- Implemented a high-performance linear layer (both forward and backward pass) with (optional) activation layer fusion using OpenAI's Triton.
- The use of the custom Triton-based linear layer demonstrated up to 1.6x speedup in training FlanT5-Base on the Samsum dataset and up to 3.5x speedup in inference.
- Automated the patching of PyTorch's nn.LinearLayer and associated activation layers to the new custom layers for inference using torch.fx for pattern matching and CUDA Graphs for reducing overheads.

Red-teaming Large Language Models **Oct 2023 - Dec 2023**

- Implemented some research papers and ideas around red-teaming large language models, including base, SFT and RLHF models like Llama-2-7B, Llama-2-7B-chat, Pythia-6.9B, GPT2-XL-1.5B and Phi-1.5B.
- Used techniques like red-teaming LLMs using the LLMs themselves to elicit toxic and offensive content generation and activation steering to steer and reduce the refusal nature of RLHF models.

Long-context Bi-Encoder **Aug 2023**

- Finetuned a custom BERT model pre-trained on 1024 token context length on document (sentence or paragraph) pairs/triplets using contrastive learning.
- Trained two versions of the model on 6.4 million and 64 million random pairs/triplets sampled from a total dataset size of 300GB. The dataset comprises data from Reddit, StackExchange, YahooAnswers, and many other popular datasets.
- Released the models publicly on HuggingFace Hub along with their evaluations on some common retrieval benchmarks.

Deep Learning Paper Implementations **Jun 2022 - Present**

- A collection of open-source reproducible implementations of some important and interesting concepts and research papers in the field of Deep Learning.
- Some examples include FlashAttention, Speculative Sampling, Lottery Ticket Hypothesis, Neural Tangent Kernels and various ML Optimizers.

Annotated ML Papers | Blog **Apr 2021 - Present**

- Regularly release annotated versions of research papers from the field of deep learning, natural language processing (NLP), ML systems and optimizations on GitHub to make reading research papers less daunting for newcomers.
- Authored multiple blog posts explaining research papers and ideas on a variety of deep learning topics in a concise manner.

Technical Skills

Languages: Python, C/C++, CUDA, Triton, Javascript, SQL, HTML/CSS, Bash
Technologies/Frameworks: PyTorch, JAX, vLLM, TensorRT-LLM, PySpark, Flask, Django, Docker, Kubernetes

Achievements/Extracurriculars

- Granted one provisional US patent for my work on Voice of the Customer and AgentGPT at Level AI.
- Granted two US patents for my work at Mastercard: One for leveraging reinforcement learning and NLP to suggest charities based on news articles, and the other for enhancing Mastercard's Threat Scan via a synthetic fraud transaction generation model using MeTGAN.
- Earned silver medal for ranking in the top 5% (115^{th} among 2426 teams) while participating solo in the Kaggle Shopee Price Match Guarantee competition, 2021.
- Ranked 55^{th} (top 10%) in the Multi-dataset Time Series Anomaly Detection challenge, KDD Cup - 2021.
- Ranked 15^{th} in CryptoHack CTF (as of May 2020), a modern-day cryptography-focused Capture the Flag event.
- Recipient of the student scholarship to attend Black Hat Asia 2019 in Singapore in which 100 students were selected from 82 countries.
- Ranked 8^{th} in AI Blitz#6 and 9^{th} in AI Blitz#7 competitions organized by AIcrowd.
- Event coordinator and problem setter for the Capture the Flag event of Technex'19, the technical fest of IIT (BHU) Varanasi and Codefest'19, the departmental fest of the CSE department.

Scholastic Achievements

- Secured all India rank of 576 in JEE Advanced 2016 among 0.2 million candidates and all India rank of 125 (99.99 percentile) in JEE Mains 2016.
- Secured all India rank of 116 in the Kishore Vaigyanik Protsahan Yojana (KVPY) examination 2015.
- Awarded NTSE scholarship through National Talent Search Examination (NTSE) in 2014 wherein 1000 meritorious students of class 10^{th} are selected at the national level.
- Top 1% ($\approx$ top 300) in India in each of the National Standard Examinations in Physics, Chemistry and Astronomy (NSEP, NSEC, NSEA) in 2015 and 2016.