Shreyansh Singh

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

Experience

Level AI
Jan 2022 – Present

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

- Developing LLM-powered AI agents for contact center workflows (e.g. sales coaching, competitive analysis), integrating on-the-fly data retrieval, filter detection, and multi-step reasoning based dynamic response generation.
- Implemented inference optimizations for high-QPS models, AutoQA and Summarization - using prefix caching, chunked prefilling, and fp8 quantization, achieving up to 3x faster throughput on AutoQA and 1.5x faster throughput on summarization, with comprehensive benchmarking on vLLM and SGLang, ensuring no quality loss.
- For low-QPS models, deployed Medusa-based speculative decoding to achieve 1.3-1.4x faster latency, and for our intent solution, accelerated bi-encoder and cross-encoder inference via torch compile, yielding 1.4x faster performance.
- Developed an instruction fine-tuning + DPO pipeline on Llama 8B, with FSDP and activation checkpointing on 8xH100 80GB GPUs to train a general purpose contact center model.
- Curated diverse multi-task instruction datasets that enabled the model to surpass internal NLP benchmarks and exhibit robust zero-shot performance, leveraging synthetic data from a distilled Llama 70B–Deepseek R1 model.
- Built the "Voice of the Customer" solution to classify contact center conversations by product/service, issue type, and theme, achieving 91% accuracy and less than 200ms per conversation 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.

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 InfoSec Club

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

Sparse Matrix Computations in CUDA **Sept 2024**

- Developed CUDA and C++ kernels for SpMV (sparse matrix-vector) and SpMM (sparse matrix-matrix) multiplication across various sparse formats (COO, CSR, CSC, BSR, BSC, ELL) to boost performance on CUDA-enabled GPUs.
- Implemented sparse matrix conversion routines and integrated libtorch for kernel validation, supported by a robust, component-specific build system.

Accelerating LLM training and inference using Triton **Jan 2024 – Feb 2024**

- 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 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.

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 (in Triton and Pytorch), various LLM Sampling techniques from scratch, Lottery Ticket Hypothesis, Neural Tangent Kernels and different 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, vLLM, SGLang, TensorRT-LLM, MLC-LLM, JAX, 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% (115th among 2426 teams) while participating solo in the Kaggle Shopee Price Match Guarantee competition, 2021.
- Ranked 55th (top 10%) in the Multi-dataset Time Series Anomaly Detection challenge, KDD Cup - 2021.
- Ranked 15th 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 8th in AI Blitz#6 and 9th 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 10th 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.