Shreyansh Singh

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

Experience

Level AI **Jan 2022 – Present**

Principal Machine Learning Engineer *Remote (India), HQ: Mountain View, CA*

- Building LLM-powered agentic “AI Workers” for contact-center workflows - coaching, deep transcript search, complex analytical insights, deep-research-style report generation, combining multi-step reasoning with parallel tool calling and efficient context handling, along with rigorous automated evaluations to solve complex analyst workflows end-to-end.
- Implemented inference optimizations for high-QPS in-house Llama-style fine-tuned models, AutoQA and Summarization using prefix caching, chunked prefilling, `torch.compile` (with CUDA Graphs) and fp8 quantization, achieving up to 3x faster throughput with comprehensive benchmarking on `vLLM` and `SGLang`, ensuring no quality loss.
- Deployed Medusa-based speculative decoding for 1.4x lower latency for low-QPS models, and accelerated bi-encoder and cross-encoder inference using `torch.compile`, yielding 1.5x faster inference for our intent classification model.
- Led the org-wide adoption of this optimized LLM inference stack across 8 production NLP services (mix of high and low-QPS), improving throughput and latency for all major AI features.
- Developed a post-training pipeline (SFT followed by DPO) on Llama 8B, with FlashAttention-2, FSDP2, `torch.compile`, BFD packing, async distributed checkpointing and activation checkpointing on 8xH100 GPUs to train a general-purpose model for contact center tasks which surpassed internal NLP benchmarks and exhibited strong zero-shot performance.
- Built the “Voice of the Customer” solution to automatically discover a three-level hierarchy of customer concerns and classify contact center conversations by product/service, issue type, and theme, achieving 75%+ accuracy across customers and <200 ms per conversation latency.

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

Data Scientist *Gurugram, India*

- 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) on datasets with high cardinality columns, without any drop in performance.

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 (* $\equiv$ 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

CUDA and Triton Kernels for Multi-LoRA LLM Inference **Oct 2025**

- Implemented CUDA and Triton kernels for Batched Gather Matrix-Vector Multiplication (BGMV) and Segmented Gather Matrix-Vector Multiplication (SGMV) algorithms for efficient Multi-LoRA inference in LLMs.
- These operations enable heterogeneous batching of inference requests where each request may need a different LoRA adapter over a shared backbone model.

Qwen3 Wordle Solver using GRPO **Sept 2025**

- Built an end-to-end RL pipeline that trains Qwen3-4B to solve Wordle using SFT on synthetic multi-turn teacher traces (with ⟨think⟩ / ⟨guess⟩ XML formatting) and GRPO (+5% accuracy) in a verifiable multi-turn Wordle environment.
- Implemented a scalable training & inference stack with FlashAttention-2, FSDP2 (with options for tensor and context parallelism) and a `vLLM` rollout server + verifiers-based GRPO trainer and OpenAI-API-compatible batch evaluation.

Low-level ML Systems and Deep Learning Paper Implementations **Jun 2022 – Present**

- A collection of open-source reproducible implementations of key interesting concepts and research papers in the field of ML Systems and Deep Learning.
- Some examples include Multi-head Latent Attention (MLA), FlashAttention (in Triton and Pytorch), various LLM Sampling techniques (from scratch), KV cache compression techniques, Byte-Pair Encoding, Lottery Ticket Hypothesis.

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.Linear` layer 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.

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, OpenAI-Agents SDK, 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% (≈ top 300) in India in each of the National Standard Examinations in Physics, Chemistry and Astronomy (NSEP, NSEC, NSEA) in 2015 and 2016.