Dhruv Sandhu

Email: dhruvsandhu21@gmail.com
LinkedIn: linkedin.com/in/dhruvsandhu/
Portfolio: https://dhruvsandhu.vercel.app/
Kaggle: kaggle.com/dhruvsandhu1
Codechef: codechef.com/users/dhruv_sandhu22
Codeforces: codeforces.com/profile/Decoder_RAG

Education

Indian Institute of Technology, Bhubaneswar GPA: 7.59
B.Tech And M.Tech (Dual Degree) Mechanical Engineering *Nov. 2021 - present*

Amity International School,Ghaziabad
CBSE *April 2005 - March 2021*

Experience

Jazzee Technologies (June 2025 - July 2025)
ML Engineer

- Delivered an end-to-end AI interview platform that automated structured, voice-first screenings and live coding, orchestrated with LiveKit Agents (STT/TTS, VAD) and Python asyncio; reduced manual screening time and standardized evaluations.
- Generated role relevant questions from multiple sources like parsed resume, custom job description, or an interviewer curated database and a knowledge manager, with adaptive follow-ups to maintain interview flow.
- Implemented real time screen share analysis for a timed LeetCode task, detecting tab switching, off-platform activity, and cheating signals; triggered proactive interventions to protect assessment integrity.
- Produced audit ready artifacts full transcripts, screen analysis logs, and structured summaries persisted to Supabase for side-by-side comparison, replay, and decision support via a React + TypeScript admin UI.
- Built an extensible analytics layer enabling any interview analysis with custom metrics/KPIs (e.g., communication, problem solving, code quality, policy adherence), supporting pluggable scoring rubrics and auto generated reports; containerized services for scalable, repeatable deployments.

IIT Delhi (May 2024 - July 2024)
Summer Research Intern

- Worked upon finding correlation between different brain regions by analysing functional Magnetic Resonance Imaging(fMRI) which was labelled as Healthy Control(HC) or Major depressive disorder(MDD).
- The project was based on applying the AAC atlas in the multi site federated setting leveraging the power of GNN.
- Applied GCN and GAT together with various techniques such as MOE and adversarial domain adaption and calculated the correlation using Pearson's coefficient.

Coforge Ltd,Noida (May 2023 - July 2023)
NLP Research Intern

- Built an end-to-end pipeline that was required to extract feedback data from the client pdf document and perform sentiment analysis on the feedback extracted for customer retention, up-selling, and cross-selling.
- Worked on tweaking each component of the pipeline, such as using a different word embedding method, different ML models, and finally how to optimize it.
- Implemented transfer learning with pre-trained transformer architectures, including BERT and RoBERTa, to improve natural language processing model performance
- Explored NER using Bi-LSTM CRF.

Projects

Llama2 from scratch *NLP*

- Implemented a LLaMA-style Transformer in PyTorch featuring RMSNorm, rotary positional embeddings (RoPE), grouped-query attention (with separate $n_{\text{kv\_heads}}$), and per-token key-value caching.
- Built an inference engine supporting temperature scaling, top-p, top-k, greedy, random sampling, and beam search (with per beam cache and no repeat n-gram).
- Integrated SentencePiece for tokenization; added checkpoint and parameter loading, EOS/padding token handling, batching, and progress reporting.
- Optimized for half-precision (FP16) and CUDA by precomputing RoPE frequencies and minimizing memory copies.

GANs from scratch *GAN*

- Implemented a configurable DCGAN in PyTorch (latent_dim = 100) and trained on CelebA; automated sampling and visualization with `tqdm` and Matplotlib.
- Engineered latent-space arithmetic to transfer attributes: $\mathbf{v}_{\text{smiling},woman} - \mathbf{v}_{\text{neutral},woman} + \mathbf{v}_{\text{neutral},man} \rightarrow \mathbf{v}_{\text{smiling},man}$, scripted batch selection and reproducible z saving for analysis.
- Designed and trained an encoder to invert $G(z)$ (image $\rightarrow$ latent) using reconstruction loss, enabling reconstruction, interpolation, and attribute edits.

- Built inference notebooks for grid sampling, per-epoch visual checks, and artifact logging; exported tensors and grids for downstream analysis.
- Optimized training stability with BCEWithLogits, normalization, and evaluation/inference utilities; packaged configs via YAML for reproducibility.

Knowledge Distillation and Model Pruning *Model Compression*

- Implemented end-to-end LLM distillation using a custom `Trainer` (*KL divergence* + cross-entropy, temperature scaling, α-mixing), training a 135M student from a 360M teacher (SmolLM-360-base) on streaming datasets (1M+ samples).
- Added structured pruning (layer and width) to create smaller deployable variants ($\approx$ 81–95M params), enabling inference on smaller GPUs while preserving core capabilities.
- Optimized training with FlashAttention-2, bfloat16, 8-bit GaLore optimizer, gradient accumulation, custom warmup/hold/anneal LR schedule, W&B tracking, and HF Hub push/restore with checkpoint resume.

Key Skills

- **DSA,GenAI,RAG,Diffusion Models,GANs,VLMs,Transfer Learning,Orpo,Ppo**
- **Programming Languages:** Python,Java,C,C++
- **ML technologies and architecture:** CV,NLP,MLOps
- **Framework/Libraries:** Pytorch,Tensorflow,Keras
- **Competitive Programming:** Pupil(Max:1293)@codeforces, 3 star@codechef, Ranked 182 out of 4084 in Feburary Cook Off-2023,Codechef
- **Deployment:** Langchain,Streamlit,AWS,GCP,Dockers

Position of Responsibility

Data Science Head (July 2023 - May 2024)
Data Science Club, IIT Bhubaneswar

- Fostered a community of more than 100+ ML enthusiast
- Conducted various workshops explaining basic algorithms and giving insights on the topics

Mechanical ML lead (March 24 - April 24)
General Championship, IIT Bhubaneshwar

- Looked over a team of 15+ students and lead my branch to win bronze.

Research Paper collaborator (August 24 - September 24)
IIT Delhi