PARTHO DAS
email: parthodas6176@gmail.com
daspartho.github.io
GitHub: @daspartho LinkedIn: @daspartho

WORK EXPERIENCE

AI Safety @ UCLA, *Collaborator* Feb. 2023 — Present

- Collaborating with AI Safety @ UCLA's alignment research team on mechanistic anomaly detection project and exploring variants of the causal scrubbing hypothesis.
- Worked on training the poisoned models, evaluating the impact of poisoning, creating activation dataset, training a binary classifier for anomaly detection, and evaluating it.

UC Berkeley SPAR, *Student researcher* Feb. 2023 — May. 2023

- Worked on RLLF (RL from language feedback) project under the mentorship of Juan Rocamonde for UC Berkeley's Supervised Program for Alignment Research.
- Designed and implemented a high-performance photorealistic image augmentor using ControlNet models for observation images of various gym environments.
- Experimented with different ControlNet pre-processing algorithms and prompt engineering for improving photorealistic image generation. Evaluated the impact of the augmentor on the overall training process.

CodeDay Lucknow, *Co-organizer* Mar. 2023 — Jun. 2023

- Led the workshop team, doing outreach, managing, and helping participants design and run their own workshops.
- Hosted my own workshop on “reverse engineering neural networks” with attendees finding circuits for various natural language tasks in real-world language models.

PERSONAL PROJECTS

Pronoun Prediction, *github / doc* May. 2023

- Investigating the circuit responsible for correctly predicting gendered pronouns given a subject name implemented by GPT-2 Small model.
- Made a preliminary report summarizing key findings around localizing the model’s computation for the task with activation patching result visualizations.

Prompt Extend, *github / demo* Nov. 2022

- Text-Generation model to help with prompt engineering by generating suitable style cues for Stable Diffusion prompts, resulting in better image generations.
- Processed the diffusiondb dataset and trained a new tokenizer and a GPT-2 model on the dataset of stable diffusion prompts for generating style cues.
- The model has 30k downloads on HuggingFace Hub. Received 2x $1000 grant from algovera.ai for the project.

MagicMix, *github / demo* Dec. 2022

- Implementation of MagicMix: Semantic Mixing with Diffusion Models paper. This technique allows for mixing two different concepts in a semantic manner to create a new concept using Diffusion Models.
- Implemented the paper in PyTorch using components from the diffusers library and successfully reproduced results from the paper. Added the implementation as a community pipeline to the diffusers library.

Predict Subreddit, *github / demo* Oct. 2022

- Multi-class text classification model to predict the subreddit of a post based on its title.
- Wrote python scripts to scrape posts from top subreddits, cleaned and processed the collected data for training.
- Fine-Tuned DistilBERT model on the collected dataset of post title pairs from the top 250 subreddits using huggingface transformers library.

OPEN SOURCE CONTRIBUTIONS

- huggingface/transformers
- huggingface/diffusers

COURSES

Intro to ML Safety Feb. 2023 — Apr. 2023

Participated in the Intro to ML Safety course for spring 2023. Learned about various technical topics and research areas in AI safety. Participated in weekly discussion sessions facilitated by Richard Moulange.

FastAI DL Foundations Oct. 2022 — Mar. 2023

Received scholarship for participating in the live cohort of FastAI's 2022 deep learning foundations course. Learned how various key components in modern deep learning work under the hood and implemented them from scratch.

SKILLS

Languages: Python, HTML, CSS, JavaScript, SQL.
Libraries: PyTorch, HuggingFace, FastAI, TransformerLens, NumPy, Pandas, Matplotlib, Ploty, Gradio, Streamlit, Flask, Selenium.
Tools: Git, GitHub, Jupyter, VS Code, Bash, Linux, AWS, TensorBoard, CI/CD.

EDUCATION

APS Nehru Road, Lucknow Graduation Year: 2022

Achieved third place in a national-level coding competition and secured the runner-up position in a state-level coding competition representing my high school.