Mohammad Taufeeque

taufeeque9.github.io · 9taufeeque9@gmail.com · github/taufeeque9 · mtaufeeque

Education

Indian Institute of Technology Bombay 2018–2022
B.Tech (with Honors) in Computer Science and Engineering Mumbai, India

- GPA - **9.24**/10.0
- Bachelor's Thesis - Fianchetto: Speed, Belief, Guile, Caution to Win at Reconnaissance Blind Chess

Selected Publications

[1] A. Tamkin, **M. Taufeeque**, and N. Goodman, "Codebook Features: Sparse and Discrete Interpretability for Neural Networks", in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 21–27 Jul 2024.

[2] **M. Taufeeque**, P. Quirke, M. Li, C. Cundy, A. D. Tucker, A. Gleave, and A. Garriga-Alonso, "Planning in a recurrent neural network that plays Sokoban", 2024. arXiv: 2407.15421 [cs.LG].

[3] K. Pelrine*, **M. Taufeeque***, M. Zając, E. McLean, and A. Gleave, "Exploiting Novel GPT-4 APIs", 2023. arXiv: 2312.14302 [cs.CR].

[4] A. Gleave*, **M. Taufeeque***, J. Rocamonde*, E. Jenner, S. H. Wang, S. Toyer, M. Ernestus, N. Belrose, S. Emmons, and S. Russell, "imitation: Clean Imitation Learning Implementations", 2022. arXiv: 2211.11972 [cs.LG].

[5] G. Perrotta, R. W. Gardner, C. Lowman, **M. Taufeeque**, N. Tongia, S. Kalyanakrishnan, G. Clark, K. Wang, E. Rothberg, B. P. Garrison, P. Dasgupta, C. Canavan, and L. McCabe, "The Second NeurIPS Tournament of Reconnaissance Blind Chess", in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, ser. Proceedings of Machine Learning Research, vol. 176, 2022, pp. 53–65.

[6] **M. Taufeeque***, S. Koita*, N. Spicher, and T. M. Deserno, "Multi-camera, multi-person, and real-time fall detection using long short term memory", in *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11601, SPIE, 2021, pp. 35–42.

Industry & Research Experience

FAR.AI - Research Engineer Berkeley, USA
Supervisor: Dr. Adam Gleave Aug '22 - Present

- Reverse-engineered the **planning algorithm** learned by an RNN to play Sokoban through model-free RL
- **Codebook Features**: Developed an interpretability tool that discretizes NN activations into features
- Red-teaming: Found novel ways to exploit the **GPT-4** Fine-tuning and Assistants API
- Benchmarked and improved the performance of the open-source library ***imitation***

Microsoft Research - Research Intern Bangalore, India
Guide: Prof. Sunita Sarawagi (IIT Bombay) & Dr. Sriram Rajmani (Microsoft Research) Dec '21 - Aug '22

- Developed an online algorithm against OOD data to integrate noisy feedback rules to a trained ML text classifier
- Deployed the algorithm to **Microsoft's Ads** system & improved the compliance of the system by over **55%**

Goldman Sachs - Summer Analyst Bangalore, India
Guide: Kesavan Mukunthan Summer 2021

- Created a webapp module to compute and display performance exposures to different factors for each stock in a portfolio of mutual funds that helps portfolio managers to analyse the drivers of performance of every fund

Technische Universität Braunschweig - Research Intern Braunschweig, Germany
Guide: Prof. Thomas Deserno April '20 - July '20
GitHub: taufeeque9/HumanFallDetection

- Developed an application that **detects falls in real-time** using human pose keypoints from multiple cameras
- Maintainer of the open-source project on GitHub with **270+ stars & 60+ forks**

AI Agent for Reconnaissance Blind Chess IIT Bombay
Bachelor's Thesis | *Guide: Prof. Shivaram Kalyanakrishnan* | Won NeurIPS competition (RBC) Aug '21 - Dec '21

- Developed an AI Agent for RBC, a variant of Chess where only a 3x3 region can be sensed before making a move
- Won the **NeurIPS 2021** competition on RBC with **91.3%** win rate & **100 Elo pts** margin from the runner-up

Randomized Planning Algorithms for POMDPs IIT Bombay
Guide: Prof. Shivaram Kalyanakrishnan Spring 2021

- Designed planning algorithms for **POMDPs** that achieved **20%** higher rewards than the SoTA algorithms
- Combined random subsets of nodes from Finite State Controllers of weak policies to obtain a strong policy

SAFE App Vulnerabilities IIT Bombay
Guide: Prof. Bhaskaran Raman Aug '20 - Apr '21

- Found **severe vulnerabilities** in the SAFE App IITB, used by many institutions to conduct remote exams
- Reported **data-leak**, **APK signature verification** and **timing-based** vulnerabilities in the Android app

Scholastic Achievements

Attended Google Research Week	Selected among 50 undergraduates nationwide	2022
Best AI Agent out of 18 Bots	RBC chess competition in NeurIPS 2021	2021
AP (Advanced Performer) Grade	Best performance in Machine Learning course (GNR 638)	2020
All India Rank 303	JEE Advanced (230,000 aspirants)	2018
All India Rank 330	JEE Mains (1.2 Million aspirants)	2018
Merit-cum-Means (MCM) Scholarship	IIT Bombay	2018
National top 1%	Indian National Physics Olympiad (INPhO)	2017
National top 1%	Indian National Chemistry Olympiad (INChO)	2017
KVPY Science Fellowship	Government of India	2016

Technical Skills

Programming	Python, C++, C, Java, Bash, Bazel, Racket, Prolog, MIPS, PostgreSQL
ML Libraries	PyTorch, JAX, Transformers, TransformerLens, Scikit-learn, Pandas, Numpy

Mentorship & Teaching

Research Manager at Impact Academy Apr '24 - Present

- I provide technical guidance & manage research projects at Impact Academy, an AI Safety field-building org

Teaching Assistantships

- Intro to ML Safety (Summer 2022 & Spring 2023)
- Artificial Intelligence & Machine Learning (Autumn 2021)
- Medical Image Computing (Spring 2022)
- Calculus II (Autumn 2021)

English Language Tutor at ELIT IIT Bombay Autumn 2020

- Designed an English language curriculum and held classes for 50 undergraduates registered for the program

Member in Developer's Community (DevCom) IIT Bombay Jan '19 - Aug '20

- Maintained, developed & updated features of **InstiApp**, the institute app with over **10,000+** downloads

Mentor of SoC project - Intrusion Detection System | Github: taufeeque9/IDS Summer 2020

- Mentored a team of **9 developers** in building a real-time system to monitor networks for malicious activity