email
linkedin
website
github
gscholar

Harsh Raj

Education

B.Tech in Engineering Physics, *Delhi Technological University*, New Delhi, India. **2019–2023**
Grade: 8.35/10.0

Skills Summary

Languages: Python, Java, C++, C, SQL, Unix scripting

Frameworks & Tools: Kubernetes, Docker, GIT, Matlab, Tensorflow, Pytorch, FastAPI

Experience

Applied Scientist (full-time) (remote), *Vijil AI*, California, US. **Feb 2024 - Present**

- Developed vijil-fuzzer, an LLM red-teaming framework that mutates seed prompts for aggressive behavior and jailbreak scenarios.
- Integrated over 30 open-source benchmarks into the **vijil-evaluation** service.
- Contributed to **vijil-dome** and **guardrails**, integrating guardrails into tools like Langchain with minimal latency impact.

ML Engineer (full-time) (remote), *Yield Protocol*, Chicago, US. **Sept 2022 - Jan 2024**

- Led data synthesis to improve LLMs as agents and built Synchaev, a complete system for agent data generation.
- Developed Mandrill, a framework for fine-tuning LLMs and validation on agent benchmarks. Supervised 10 research interns from the Disruption Lab at UIUC as part of the project.
- Co-created Cacti, a web3 transaction chatbot, in collaboration with a former VP of Quora.
- Built AutoEval, a framework for automating the evaluation of the Cacti bot.

Applied Researcher (intern) (remote), *Thoucentric*, Bangalore, India. **May 2022 - Aug 2022**

- Developed a comprehensive NL2SQL conversion system tailored for data analysis purposes. Incorporated automatic visualization feature with tools like DeepEye.
- Improved the SADGA-GaP framework by introducing pre-processing and post-processing steps, resulting in a 10% increase in accuracy. Additionally, implemented a value-copying mechanism in the model, resolving issues such as missing table names and row values in generated SQL queries.
- Authored a white paper

Data Scientist (intern) (on-site), *Attryb Tech*, Bangalore, India. **June 2021 - Dec 2021**

- Developed the complete pipeline for Content Studio, a tool for generating and analyzing content for blog composition, product marketing, and various other creative writing tasks.
- Finetuned Pegasus and GPT-2 for content summarization and title generation.
- Data mined SemRush and Common Crawl to create the content database.

Publications

Defences against Reverse Preference Attacks **(under review, SaTML'24)**: Domenic Rosati, Giles Edkins, **Harsh Raj**, David Atanasov, Kai Williams, Subhabrata Majumdar, Janarthanan Rajendran, Frank Rudzicz, Hassan Sajjad

On Transfer of Adversarial Robustness from Pretraining to Downstream Tasks **(NeurIPS'23)**: Laura Fee Nern, **Harsh Raj**, Maurice Georgi, Yash Sharma

Measuring Reliability of Large Language Models through Semantic Consistency **(Best Paper Award, ML Safety, NeurIPS'22)**: **Harsh Raj**, Domenic Rosati, Subhabrata Majumdar

Evaluating the Robustness of Biomedical Concept Normalization **(Transfer Learning, NeurIPS'22)**: Sinchani Chakraborty, **Harsh Raj**, Srishti Gureja, Tanmay Jain, Atif Hassan, Sayantan Basu

Decoding Percepts in Vision Language Navigation: Is it about better features or more data? (under review, ACM TIST): **Harsh Raj**, Ashutosh Pandey, Shaurya Kumar, Kavinder Singh, Nihal Kumar, Anil Singh Parihar

Improving Consistency in Large Language Models through Chain of Guidance **(under review, TMLR)**: **Harsh Raj**, Vipul Gupta, Domenic Rosati, Subhabrata Majumdar

GANDALF: Gated Adaptive Network for Deep Automated Learning of Features **(under review, TMLR)**: Manu Joseph, **Harsh Raj**

AskYourDB: An end-to-end system for querying and visualizing relational databases using natural language: Manu Joseph, **Harsh Raj**, Abhinav Yadav, Aaryamann Sharma

Extract It! Product Category Extraction by Transfer Learning **(CICT'22)**: **Harsh Raj**, Aakansha Gupta, Rahul Katarya

Projects

Synchaev **(Data Synthesis)**:

- A framework for generating conversational data between LLM agents and environments. The environments are taken from AgentBench.

Repo-Level Prompt Engineering-Solidity **(Prompt Engineering)**:

- An implementation of the paper Repo-Level Prompt Generation to support SOLIDITY code, a widely used programming language in the Blockchain industry.

Antibody-Antigen Binding Classifier **(Computational Biology)**:

- A graph neural network-based architecture to predict the binding affinity between antibodies and antigens, utilizing the SAbDab dataset for affinity prediction.

Honors and Awards

Vision and Language Navigation: Secured **3rd** place in the SPL metric on the R2R benchmark, a widely recognized Vision and Language Navigation measure. Link to standings. ID: *MLR_Lab_DTU*.

ML Safety Challenge: Won NeurIPS'22 Best Paper Award with a cash prize of $5000.

Codeforces ML Round: Achieved **Global Rank 7** and ranked **1st** in the country in the Raif ML Round 1, organized by Raiffeisen Bank International AG. Link to standings. ID: *harsh777111raj*.