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 (Remote), *Vijil AI*, California, US. **Feb 2024 - Present**

- Built vijil-fuzzer, an LLM red-teaming framework designed to mutate a set of seed prompts towards aggressiveness and jailbreak.
- Maintained vijil_objects, a repository for ingesting red-teaming data from the web to be used by the evaluation engine.

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

- Spearheaded data synthesis initiatives to improve autonomous agent capabilities in open-source Large Language Models (LLMs), utilizing tools like AgentBench for benchmarking. Built Synchaev, an end-to-end system for agent data synthesis.
- Developed Mandrill, a framework for fine-tuning LLMs and assessing them on prevalent agent benchmarks. This project also included overseeing the mentorship of 10 research interns from Disruption Lab, University of Illinois at Urbana-Champaign (UIUC).
- Co-developed Cacti, a web3 transaction chatbot, in collaboration with a seasoned industry expert and former VP of Quora.
- Engineered AutoEval, an evaluation framework to facilitate autonomous end-to-end 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.
- Enhanced the SADGA-GaP framework by implementing pre-processing and post-processing steps, resulting in a notable 10% increase in accuracy.
- Implemented a value copying mechanism in the model, addressing previous limitations such as missing table names and row values in the generated queries.
- Authored a research paper on this work.

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

- Developed the core architecture for Content Studio, a tool for generating and analyzing content, enhancing blog composition, product marketing, and various other creative writing tasks.
- Implemented a comprehensive system covering Text Detoxification, Customer Intent Categorization, Two-Level Clustering (used initially for recommendation purposes), and Creation of Outlines & Titles (supported with Pegasus and GPT-3), alongside a Question-Answering module (also supported with GPT-3).
- Curated datasets from SemRush and Common Crawl to establish the foundational database.

Publications

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

Devise a set of *reverse preference attacks* which illustrate how LLMs can be made harmful by providing adversarial reward during RLHF. We find that these attacks uniquely expose a critical safety gap of safety-aligned LLMs in RL settings: they easily explore harmful actions resulting in learning harmful text generation policies. We explore mitigation strategies from defences against harmful SFT and develop a taxonomy that unifies current defence methods and connects them with a framework from Safe RL: Constrained MDPs.

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

Demonstrate that the robustness of a linear predictor on downstream tasks can be constrained by the robustness of its underlying representation, regardless of the protocol used for pretraining.

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

Developed semantic consistency measures for evaluating the performance of pretrained language models (PLMs) in text generation, demonstrating enhanced alignment with human evaluations and surpassing traditional lexical metrics in assessing consistency across paraphrased prompts.

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

Introduce and evaluate heuristic-based input transformations and adversarial attacks to assess the robustness of BERT-based biomedical concept normalization models, revealing significant vulnerabilities in model performance and proposing mitigation strategies to enhance model robustness against adversarial perturbations.

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

Presented application of pre-trained Vision-Language Models (VLMs) in Vision-and-Language Navigation (VLN), significantly enhancing performance by aligning visual features with linguistic context and reducing reliance on extensive data synthesis. This approach yielded improvements in benchmark metrics across multiple datasets (R2R, RxR, REVERIE) and validated the efficacy of VLMs in diverse VLN agent frameworks.

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

Developed a novel metric for evaluating semantic consistency in Large Language Models (LLMs), introducing the Ask-to-Choose (A2C) prompting strategy, which significantly enhances the accuracy and consistency in open-ended text generation and closed-book question answering, as validated by the TruthfulQA benchmark.

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

The GANDALF architecture represents an efficient approach in deep learning for tabular data, distinguished by its novel Gated Feature Learning Unit (GFLU), demonstrating superior or comparable performance to existing state-of-the-art methods on various recognized benchmarks.

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

Developed a semantic parsing system that transforms natural language queries into SQL, enhancing database accessibility for business users. This innovation involved refining SoTA models with strategic pre and post-processing techniques, culminating in a production-ready tool complemented by an automated visualization framework for query results.

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

Proposed a multi-level product categorization model, demonstrating an 86% accuracy rate on a 20,000-product dataset from Flipkart, significantly enhancing e-commerce product search and categorization efficiency."

Projects

ConsistencyBench **(LLM Evaluation)(under development)**:

- A library to evaluate self-consistency in Large Language Models (LLMs). To be released under vijil.

Repo-Level Prompt Engineering-Solidity **(Prompt Engineering) (Oct '22, Yield Protocol)**:

- Modified the original repository to support the blockchain-industry's widely used programming language Solidity.
- Integrated the Abstract Syntax Tree (AST) of Solidity with modifications to align with the existing framework.

Antibody-Antigen Binding Classifier **(Computational Biology, Mar '22)**:

- Developed a model to predict the binding affinity between antibodies and antigens. The model leverages a dual-modality strategy, combining sequential and structural data of the molecules using SE3 Equivariant Transformer for structural properties and a Protein Transformer for sequential data.
- Attained a 78% prediction accuracy on the SAbDab dataset (after filtering out small sequence molecules) by considering the structural attributes of the molecules only.

Honors and Awards

Vision and Language Navigation: Secured **3rd** place in the success-to-length ratio on the R2R benchmark, a leading Vision and Language Navigation measure. Our model uses significantly less compute than the top performers. Link to standings. ID: MLR_Lab_DTU.

CodeForces Raif ML Round 1: Achieved **Global Rank 7** and stood **1st** in the country during the Raif ML Round 1 competition organized by Raiffeisen Bank International AG in June 2021. Link to standings. ID: harsh777111raj.

ML Safety Workshop, NeurIPS'22: Won the Best Paper Award for the paper Measuring Reliability of Large Language Models through Semantic Consistency with a cash price of $5000.