Vaidehi Patil

Department of Computer Science
University of North Carolina at Chapel Hill
vaidehi99.github.io
github.com/Vaidehi99

vaidehi@cs.unc.edu
Google Scholar
linkedin.com/in/vaidehi-patil-262916151

Research Interests

Language Models, AI safety, Privacy, Security, Multimodality, Multi-Agent privacy, Interpretability, Model editing

Education

University of North Carolina at Chapel Hill (UNC Chapel Hill) *Aug 2022 - May 2027 (Expected)*
PhD student in Computer Science
Graduate Research Assistant, Advisor: Prof. Mohit Bansal

Indian Institute of Technology Bombay (IIT Bombay) (**CPI: 9.52/10**) *2017 - 2022*
Interdisciplinary Dual Degree: B. Tech. in Electrical Engineering, M. Tech. in AI and Data Science
Minor in Computer Science and Engineering
🏆 **Undergraduate Research Award**, **Best IDDDP Dissertation Award**
Thesis: Multilingual Representations for Closely Related Languages
Advisors: Prof. Sunita Sarawagi, IIT Bombay; Dr. Partha Talukar, Google Research India

Work Experience

Apple, Cupertino, USA *May - Aug, 2024*
ML Research Scientist Intern
Advisors: Siddharth Patwardhan
Improving faithfulness in RAG by reducing conflict between parametric and contextual knowledge

Amazon Alexa team, Seattle, USA *May - October, 2023*
Applied Scientist Intern
Advisors: Markus Dreyer, Mengwen Liu & Leonardo Ribeiro
Self-refinement of multimodal LLMs to create a dataset for multimodal summarization

Adobe Research India *May - August, 2021*
Research Intern (🏆Pre-placement offer as Software Development Engineer)
Advisors: Dr. Vishwa Vinay & Dr. Kuldeep Kulkarni
Worked on scene expansion via scene graph using graph generative models

Adobe Research India *May - August, 2020*
Research Intern (🏆Re-internship offer as research intern for Summer 2021.)
Advisor: Natwar Modani
Introduced a novel problem of detecting versions of documents and created a dataset for it

AWL Japan *February - April, 2022*
Computer Vision R&D Intern

Publications

* - equal contribution

10. **UPCORE: Utility-Preserving Coreset Design for Balanced Unlearning**.
Vaidehi Patil, Elias Stengel-Eskin, Mohit Bansal
Under Review at ICML 2025

9. **Safree: Training-Free and Adaptive Guard for Safe Text-to-Image And Video Generation** .
Jaehong Yoon,* Shoubin Yu*, **Vaidehi Patil**, Huaxiu Yao, Mohit Bansal
ICLR 2025

8. **Unlearning Sensitive Information in Multimodal LLMs: Benchmark and Attack-Defense Evaluation**.
Vaidehi Patil, Yi-Lin Sung, Peter Hase, Jie Peng, Tianlong Chen, Mohit Bansal
TMLR 2024

7. **RefineSumm: Self-Refining MLLM for Generating a Multimodal Summarization Dataset** .
Vaidehi Patil, Leonardo F. R. Ribeiro, Mengwen Liu, Mohit Bansal and Markus Dreyer.
ACL Main Conference 2024 (🏆 Amazon Conference Travel Grant)

6. **Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks**.
Vaidehi Patil*, Peter Hase* and Mohit Bansal
ICLR 2024 as **Spotlight**

5. **Debiasing Multimodal Models via Causal Information Minimization**.
Vaidehi Patil, Adyasha Maharana and Mohit Bansal
EMNLP 2023 Findings
NeurIPS 2023 workshop on Causal Representation Learning

4. **GEMS: Scene Expansion using Generative Models of Graphs**.
Rishi Agarwal*, Tirupati Saketh Chandra*,**Vaidehi Patil***, Aniruddha Mahapatra*, Kuldeep Kulkarni and Vishwa Vinay.
WACV 2023

3. **Overlap-based Vocabulary Generation Improves Cross-lingual Transfer Among Related Languages**. [Oral Presentation]
Vaidehi Patil, Partha Talukdar and Sunita Sarawagi.
ACL Main Conference 2022 (🏆*Google Conference Travel Grant*)

2. **Detecting Document Versions and Their Ordering In a Collection**.
Natwar Modani, Anurag Maurya, Gaurav Verma, Inderjeet Nair, **Vaidehi Patil**, Anirudh Kanfade.
International Conference on Web Information Systems Engineering (WISE) 2021
🏆 **Best Paper Runner-Up Award**.

1. **Exploiting Language Relatedness for Low Web-Resource Language Model Adaptation: An Indic Languages Study**.
Yash Khemchandani*, Sarvesh Mehtani*, **Vaidehi Patil**, Abhijeet Awasthi, Partha Talukdar and Sunita Sarawagi.
ACL Main Conference 2021

Patents

2. **Expanding a scene graph using proposals from a generative model of scene graphs**.
Vishwa Vinay, Tirupati Saketh Chandra, Rishi Agarwal, Kuldeep Kulkarni, Hiransh Gupta, Aniruddha Mahapatra, **Vaidehi Patil**
US patent application | Adobe Inc.

1. **Systems for generating indications of relationships between electronic documents**.
Natwar Modani, **Vaidehi Patil**, Inderjeet Nair, Gaurav Verma, Anurag Maurya, Anirudh Kanfade.
US patent application | Adobe Inc.

Professional Responsibilities

- *Lead Workshop Organizer* - **Machine Unlearning for Generative AI**, ICML 2025
- *Reviewer* - TMLR 2025, ICLR 2025, ACL 2025, NAACL 2025, TMLR 2024, ACL ARR 2024, CVPR 2022
- *Undergraduate TA* - Center for Machine Intelligence and Data Science, IIT Bombay

Major Academic Achievements

- Awarded **Carolina Computing Fellowship** by the Computer Science Dept, UNC Chapel Hill ['22]
- **2nd rank** in AI and Data Science batch of Centre for Machine Intelligence and Data Science ['22]
- Granted Advanced Performers grade (awarded to **top 1%**) grade for excellent performance ['18]
- Awarded an option of branch change due to exceptional academic performance in first year ['18]
- Secured All India Rank in **top 0.25%** in JEE Mains 2017 and **top 2%** in JEE Advanced 2017 ['17]

Invited Talks

- **Deep Learning: Classics and Trends**
Can Sensitive Information Be Deleted from LLMs? [Sep '24]

- **Ploutos**
UPCORE: Utility-Preserving Coreset Selection for Balanced Unlearning [Apr '25]