

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](#)
in linkedin.com/in/vaidehi-patil-262916151

RESEARCH INTERESTS

AI Safety & Security, Privacy, Agent Safety & Evaluation, Adversarial Robustness, Post-Training, Machine Un-learning

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
🏆 Recipient of the Google North America PhD Fellowship 2025 in Machine Learning & ML Foundations
🏆 Recipient of the AdvML Rising Star Award 2026

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 Talukdar, Google Research India

WORK EXPERIENCE

Salesforce AI Research, Palo Alto, USA *May - Aug, 2026*
AI Research Intern Advisors: Shafiq Joty, Zixuan Ke
Working on the evaluation of self-evolving tool-use agents in an evolving harness setup

Google DeepMind, Cambridge, USA *May - Dec, 2025*
Student Researcher | Advisors: Gintare Karolina Dziugaite, Matthew Jagielski, Eleni Triantafillou

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 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

SELECTED PUBLICATIONS

* - equal contribution

12. **EvoHarnessBench: Can Your Agents Keep Pace with an Evolving Harness?**
Zixuan Ke*, Vaidehi Patil*, Haizhou Shi*, Yang Li, Ye Liu, Sarath Shekizhar, Anurag Koul, Jiayu Wang, Xuan Phi Nguyen, Semih Yavuz, Mohit Bansal, Shafiq Joty
Under Review at ICLR 2027
11. **Playing Along: Learning a Double-Agent Defender for Belief Steering via Theory of Mind.**
Hanqi Xiao*, Vaidehi Patil*, Zaid Khan, Hyunji Lee, Elias Stengel-Eskin, Mohit Bansal
COLM 2026

10. **The Sum Leaks More Than Its Parts: Compositional Privacy Risks and Mitigations in Multi-Agent Collaboration** .
Vaidehi Patil, Elias Stengel-Eskin, Mohit Bansal
Under Review at NAACL 2027
9. **UPCORE: Utility-Preserving Coreset Design for Balanced Unlearning.**
Vaidehi Patil, Elias Stengel-Eskin, Mohit Bansal
Under Review at TMLR
8. **LEMUS: Latent Concept Erasure for Robust Unlearning in Diffusion Transformers** .
 Fengli Wu*, Vaidehi Patil*, Han Lin, Jaehong Yoon, Yue Zhang, Mohit Bansal
Under Review at WACV 2026
7. **Safeguarding LLMs via Model-Agnostic Latent Safety Signals from Dark Knowledge** .
 Wonjun Lee, Kyungsik Yang, Gaeun Ji, **Vaidehi Patil**, Park, Ham, Mohit Bansal, Suhyun Kim
NeurIPS 2026
6. **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
5. **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
4. **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)
3. **Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks.**
Vaidehi Patil*, Peter Hase* and Mohit Bansal
ICLR 2024 as Spotlight
2. **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)
1. **Detecting Document Versions and Their Ordering In a Collection.**
 Natwar Modani, Anurag Maurya, Gaurav Verma, Inderjeet Nair, **Vaidehi Patil**, Anirudh Kanfode.
International Conference on Web Information Systems Engineering (WISE) 2021
🏆 Best Paper Runner-Up Award.

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 Kanfode.
US patent application | Adobe Inc.

AWARDS

- Recipient of the **Google North America PhD Fellowship in Machine Learning & ML Foundations** [’25]
- Selected for AdvML, MLCommons **Rising Stars 2026**, the 13th Heidelberg Laureate Forum [’26]
- 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]

PROFESSIONAL RESPONSIBILITIES

- *Lead Workshop Organizer* - **Machine Unlearning for Generative AI**, ICML 2025
- *Undergraduate TA* - Center for Machine Intelligence and Data Science, IIT Bombay