Mihir Upadhyay

mu2253@nyu.edu | LinkedIn | +1 (551)-362-5762 | Google Scholar | GitHub | Website

Academic History

New York University — New York, US
Master of Science in Data Science — *Sep 2024 – Present*
GPA: 3.93/4.0

Indian Institute of Technology BHU — Varanasi, India
Bachelor of Technology, First Class with Honours — *Jul 2017 – May 2021*

Academic Publications

[3] **Mihir Upadhyay***, Bernardo Vilegas*, Siobhan Mackenzie Hall, Katherine M. Collins, Umang Bhatt. Characterizing Higher-order Externalities of Human-AI Interaction on Human-Human Interaction. *Working paper*.

[2] Elisabeth Kirsten, Jost Grosse Perdekamp, Qinyuan Wu, **Mihir Upadhyay**, Krishna P. Gummadi, Muhammad Bilal Zafar. Characterizing Web Search in the Age of Generative AI. The Fourteenth International Conference on Learning Representations *(ICLR'26), AIWILD*.

[1] Umang Bhatt*, Sanyam Kapoor*, **Mihir Upadhyay**, Katherine M. Collins, Ilia Sucholutsky, Francesco Quinzan, Adrian Weller, Andrew Gordon Wilson, Muhammad Bilal Zafar. When Should We Orchestrate Multiple Agents? *(arXiv'25)*.

Research Experience

TRACE Lab, New York University — New York, US
Research Assistant — *Sep 2025 — Present*

- Developed taxonomy of AI-Human workflows across 5+ domains, mapping 4 oversight levels and governance structures.
- Implemented multi-agent orchestration platform using LangChain & CrewAI supporting 10+ workflow configurations.
- Designed sandbox evaluation testbed using synthetic datasets and simulated human agents to study AI orchestration on real-world tasks from O*NET, analyzing cost and accuracy trade-offs.

IBM Thomas J. Watson Research Center — New York, US
Student Researcher — *Sep 2025 – Dec 2025*

- Curated a dataset of 140+ scenarios evaluating agents on compositional instructions and multi-agent orchestration tasks.
- Created industrial workflows with agent dependency graphs, capturing temporal, logical, and data-dependent relationships.
- Proposed the DRFR evaluation framework, revealing consistent task decomposition (ROUGE ≈ 0.25) but significant variance in parameter prediction across 10 open-source LLMs.

RC Trust, Ruhr University — Bochum, Germany
Research Intern — *May 2025 – Aug 2025*

- Developed evaluation framework for generative search agents across 6+ domains, capturing stability & LLM stochasticity.
- Observed 27% decision flips and low source overlap (Jaccard ≈ 0.18–0.45) across repeated executions in generative search.
- Applied LLOOM-based concept induction and AST-based parsing to quantify topical coverage and latent reference structures in generated outputs, achieving 25% improvement over LDA.

Industry Experience

Flixstock — Gurugram, India
Technical Product Manager — *Oct 2022 – Sep 2023*

- Led development of a multi-stage generative imaging pipeline (diffusion, BLIP, segmentation) for creating AI-generated human models with dynamic apparel rendering.
- Added human-in-the-loop validation and QCs, increasing accuracy by 20% and reducing manual intervention time by 30%.

Standard Chartered Bank — Bengaluru, India
AI/ML Analyst — *Jul 2021 – Sep 2022*

- Led adoption of trustworthy ML practices by integrating Truera-based bias and drift monitoring for deployed ML models.
- Integrated Dockerized workflows into CI/CD pipelines using Jenkins, improving automated failure detection by 25%.

Data Science Intern — *May 2020 – Jul 2020*

- Developed and optimized fraud detection classifiers in card transactions using bagging, boosting, and ensemble methods, improving precision by 15%.

Projects

Benchmarking Human-AI Collaboration in Multi-Agent Systems *May 2025 — Aug 2025*

- Developed game-based multi-agent testbed to study human-AI teams of 3 across competitive and cooperative settings.
- Achieved +21.4% gains in team payoff in cooperative sequential tasks and +18.7% higher response diversity with memory in social deduction games (100+ runs).

Derivative-Free Optimization of In-Context Learning for LLMs *Jan 2026 — May 2026*

- Framed prompt optimization for LLMs as a derivative-free optimization problem and implemented a GEPA-based evolutionary framework under strict evaluation budgets ($\leq$500 queries).
- Designed a unified evaluation pipeline across fact verification (HoVer) and program synthesis (HumanEval), incorporating sandboxed execution, multi-objective Pareto archiving, and cost-efficient surrogate screening.
- Achieved performance gains of +20% Pass@1 (GPT-3.5) and +26% (Qwen-3-8B) on HumanEval, and +6.4% on HoVer.

Skills

Programming Languages: Python, C, C++, Java, JavaScript, Rust, SQL, Bash
Tools and Platforms: PyTorch, TensorFlow, Hugging Face Transformers, LangChain, LangGraph, CrewAI, vLLM, LangEval, Docker, AWS, Kubernetes, Jenkins, Supabase, Git, Apache, Spark, Hadoop

Awards & Leadership

Recipient of **Swami Vivekanand Scholarship for Academic Excellence** toward MS funding 2024–26
Co-organizer NYU Computational Intelligence, Vision, and Robotics (CILVR) Seminar Series 2024–26
Teaching Assistant NYU Courant Institute of Mathematical Sciences (Algebra, Trigonometry, and Functions) Fall'25
Teaching Assistant NYU Center for Data Science (Responsible Data Science) Spring'25
Member NYU Graduate Community Building Group (GCBG) 2024–25
Recipient of **IIT BHU Merit Scholarship** for the academic years 2017–21
Gold Medal IIT BHU Inter-Hostel Meet 2019 (Football)
Silver Medal Inter-IIT Cultural Meet 2019 (Music track)
Member IIT BHU Cultural Council 2019–20
Secretary IIT BHU Western Music Club 2019–20