Duy Nguyen

duykng@cs.unc.edu duykhuongnguyen.github.io github.com/duykhuongnguyen

Research Interests

My research focuses on mechanistic interpretability and inference-time steering methods for interpreting and monitoring the behaviors of (multimodal) LLMs. Additionally, I am interested in post-training methods for LLMs, including Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning with Verifiable Rewards (RLVR).

Education

The University of North Carolina at Chapel Hill **Aug. 2024 – Aug. 2029 (expected)**
Ph.D. in Computer Science *Chapel Hill, NC, US*

- Advisor: Prof. Mohit Bansal

Hanoi University of Science and Technology **Sep. 2018 – Sep. 2022**
B.S. in Computer Science *Hanoi, Vietnam*

- GPA: 3.65/4.00, graduated with Excellent Degree

Publications

Duy Nguyen[*], Archiki Prasad[*], Elias Stengel-Eskin, and Mohit Bansal. LASeR: Learning to Adaptively Select Reward Models with Multi-Arm Bandits. In *Neural Information Processing Systems (NeurIPS)*, 2025. [pdf]

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-Attribute Steering of Language Models via Targeted Intervention. In *Association for Computational Linguistics (ACL)*, 2025. [pdf]

Bao Nguyen, Binh Nguyen, **Duy Nguyen**, and Viet Anh Nguyen. Risk-Aware Distributional Intervention Policies for Language Models. In *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)*, 2025. [pdf]

Ngoc Bui, **Duy Nguyen**, Man-Chung Yue, and Viet Anh Nguyen. Coverage-Validity-Aware Algorithmic Recourse. In *Operations Research*, 2024. [pdf]

Hieu Nguyen, **Duy Nguyen**, Khoa Doan, and Viet Anh Nguyen. Cold-start Recommendation by Personalized Embedding Region Elicitation. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024. [pdf]

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Distributionally Robust Recourse Action. In *International Conference on Learning Representations (ICLR)*, 2023. [pdf]

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Feasible Recourse Plan via Diverse Interpolation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. [pdf]

Ngoc Bui, **Duy Nguyen**, and Viet Anh Nguyen. Counterfactual Plans under Distributional Ambiguity. In *International Conference on Learning Representations (ICLR)*, 2022. [pdf]

Tuan-Duy H. Nguyen, Ngoc Bui, **Duy Nguyen**, Man-Chung Yue, and Viet Anh Nguyen. Robust Bayesian Recourse. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022. [pdf]

(*) denotes equal contribution

Preprints

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs. *Under Review.* [pdf]

Duy Nguyen, Bao Nguyen, and Viet Anh Nguyen. Cost Adaptive Recourse Recommendation by Adaptive Preference Elicitation. *Forever Preprint.* [pdf]

Experience

Amazon Science **May 2025 – Aug. 2025**
Applied Scientist Intern *Seattle, WA, USA*

- Research topics: LLM reasoning for tabular data

VinAI Research **Aug. 2022 – Aug. 2024**
Research Resident *Hanoi, Vietnam*

- Advisor: Prof. Viet Anh Nguyen
- Research topics: LLM safety, interpretability and explainability

Honors and Awards

Honorable Mention - Undergraduate Operations Research Prize **Oct. 2022**
INFORMS

Best thesis presentation award **Aug. 2022**
School of Information and Communication Technology, HUST

Excellence Scholarship for the academic year **Sep. 2019**
School of Information and Communication Technology, HUST

Professional Academic Services

Reviewer at ICLR (2025, 2026), ICML (2025), NeurIPS (2023–2025), ACL Rolling Review (2025).