DUY NGUYEN

✉ email   🌐 website   github

Research Interests

My research focuses on post-training methods for (multimodal) LLMs, including RLHF and Reward Modeling. Additionally, I'm very interested in mechanistic interpretability and inference-time interventions for LLM safety alignment.

Education

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

Research Experience

VinAI Research Aug. 2022 – Aug. 2024
Research Resident *Hanoi, Vietnam*
- Advisor: Prof. Viet Anh Nguyen
- Research topics: LLM Truthfulness, Algorithmic Recourse, Recommender System

Publications

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

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.

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

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

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

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.

Preprints

Duy Nguyen[*], Archiki Prasad[*], Elias Stengel-Eskin, and Mohit Bansal. LASeR: Learning to Adaptively Select Reward Models with Multi-Arm Bandits. *Under Review.*

Bao Nguyen, Binh Nguyen, **Duy Nguyen**, and Viet Anh Nguyen. Semantically-Consistent, Distributional Intervention Policies for Language Models. *Under Review.*

Duy Nguyen, Bao Nguyen, and Viet Anh Nguyen. Cost Adaptive Recourse Recommendation by Adaptive Preference Elicitation. *Under Review.*

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), NeurIPS (2023, 2024), AISTATS (2022, 2023), UAI (2023), ACM FAccT (2023).