Ngoc Bui

New Haven, CT, US · (+1) 203-809-4258
ngocbh.pt@gmail.com · ngoc.bui@yale.edu · ngocbh.github.io · scholar: LVklD_cAAAAJ

RESEARCH INTERESTS

My recent research focuses on large (vision-)language models (LLMs and VLMs), with an emphasis on long-context, long-horizon, and open-ended generation, where computation and memory efficiency are especially critical. These capabilities are essential for long-form reasoning, persistent conversational agents, and streaming applications. I am also excited about building personalized assistants, generating synthetic data, and designing memory-efficient architectures that support continual knowledge updates for streaming and real-time applications.

EDUCATION

- **Ph.D. in Computer Science** *2023 – 2028 [expected]*
  Yale University
  - Advisor: Prof. Rex Ying
- **M.S. in Data Science** *2021 – 2023*
  Hanoi University of Science and Technology (HUST)
  - GPA: 3.9/4.0, Major GPA: 4.0/4.0.
  - Thesis: Evolutionary and Deep Reinforcement Learning Algorithms for Optimizing the Lifetime of Wireless Sensor Networks.
  - Advisors: Prof. Thuan Do Phan and Dr. Phi Le Nguyen
- **Engineer in Computer Science** *2016 – 2021*
  Hanoi University of Science and Technology (HUST)
  - GPA: 3.67/4.0, Major GPA: 3.88/4.0.
  - Honors: Excellent (≈ 1st class honors).

UNDER REVIEW (MOST RECENT)

- **Ngoc Bui**, Shubham Sharma, Simran Lamba, Saumitra Mishra, Rex Ying. "Cache What Lasts: Token Retention for Memory-Bounded KV Cache in LLMs". *Under review*. [Openreview].
  Keywords: LLMs, forgetting mechanism, KV cache eviction, long-context, long-reasoning.

PUBLICATIONS

- Neil He, Jiahong Liu, Buze Zhang, **Ngoc Bui**, Ali Maatouk, Menglin Yang, Irwin King, Melanie Weber, Rex Ying. "Position: Beyond Euclidean - Foundation Models Should Embrace Non-Euclidean Geometries". **Oral Presentation** at *Learning on Graphs Conference (LOG)*, 2025. [arxiv].
  Keywords: geometric deep learning, foundation models.

- **Ngoc Bui**, Menglin Yang, Runjin Chen, Leonardo Neves, Mingxuan Ju, Rex Ying, Neil Shah, Tong Zhao. "Learning Along the Arrow of Time: Hyperbolic Geometry for Backward-Compatible Representation Learning". The *International Conference on Machine Learning (ICML)*, 2025. [arxiv].
  Keywords: geometric deep learning, image retrievals, backward-compatible training.

- **Ngoc Bui**, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, Rex Ying. "Mixture-of-Personas Language Models for Population Simulation". The *Findings of the Association for Computational Linguistics (ACL Finding)*, 2025. [arxiv].
  Keywords: LLMs, synthetic data generation, personalized agents.

- **Ngoc Bui**, Hieu Trung Nguyen, Viet Anh Nguyen, and Rex Ying. "Explaining Graph Neural Networks via Structure-aware Interaction Index". The *International Conference on Machine Learning (ICML)*, 2024. [arxiv].
  Keywords: interpretability, Shapley values, game theory, graph neural network.

- Jiasheng Zhang, Jialin Chen, Ali Maatouk, **Ngoc Bui**, Qianqian Xie, Leandros Tassiulas, Jie Shao, Hua Xu, Rex Ying. "LitFM: A Retrieval Augmented Structure-aware Foundation Model For Citation Graphs". *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. [arxiv].
  Keywords: citation graphs, retrieval augmented generation.
- **Ngoc Bui**, Duy Nguyen, Man-Chung Yue, and Viet-Anh Nguyen. "Coverage-Validity-Aware Algorithmic Recourse". The *Operations Research (OPRE) Journal*, 2024. [arxiv].
  Keywords: interpretability, distributionally robust optimization.
- Duy Nguyen, **Ngoc Bui**, Viet-Anh Nguyen. "Feasible Recourse Plan via Diverse Interpolation". The *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023. [arxiv].
  Keywords: interpretability, distributionally robust optimization.
- Duy Nguyen, **Ngoc Bui**, Viet-Anh Nguyen. "Distributionally Robust Recourse Action". The *International Conference on Learning Representations (ICLR)*, 2023. [arxiv].
  Keywords: interpretability, distributionally robust optimization.
- **Ngoc Bui**, Duy Nguyen, Viet-Anh Nguyen. "Counterfactual Plans under Distributional Ambiguity". The *International Conference on Learning Representations (ICLR)*, 2022. [arxiv].
  Keywords: interpretability, distributionally robust optimization.
- Tuan-Duy Hien Nguyen, **Ngoc Bui**, Duy Nguyen, Man-Chung Yue, Viet Anh Nguyen. "Robust Bayesian Recourse". The *Association for Uncertainty in Artificial Intelligence (UAI)*, 2022. [arxiv].
  Keywords: interpretability, distributionally robust optimization.
- **Ngoc Bui**, Phi Le Nguyen, Viet Anh Nguyen, Phan Thuan Do. "A Deep Reinforcement Learning-based Adaptive Charging Policy for WRSNs". IEEE *International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, 2022.
- **Ngoc Bui**, Tam Nguyen, Binh Huynh Thi Thanh, and Trong Vinh Le. "A phenotype-based multi-objective evolutionary algorithm for maximizing lifetime in wireless sensor networks with bounded hop". The *Journal of Soft Computing*, 2023.
- **Ngoc Bui** and Viet-Trung Tran. "A Novel Conditional Random Fields Aided Fuzzy Matching in Vietnamese Address Standardization". The *International Symposium on Information and Communication Technology (SoICT)*, 2019.

EXPERIENCES

- **JPMorgan Chase & Co.** *June 2025 - August 2025*
  Data & Analytics Research Associate
  - Developing an efficient inference algorithm to improve the memory and computation bottleneck of long-context and long-generation in LLM inference, providing $\times 2$ inference speed and preserving (sometimes outperforming) the performance with only $\sim 3\%$ memory budget compared to vanilla inference. This is 58%-200% gain compared to SOTA baselines.
  - Advisor: Dr. Shubham Sharma
- **Snap Inc.** *June 2024 - November 2024*
  Research Intern
  - Developing backward-compatible training techniques based on hyperbolic geometry to improve performance and efficiency of large retrieval systems, particularly in the context of continual updates of the embedding models.
  - Advisor: Dr. Tong Zhao
- **VinAI Research (now acquired by Qualcomm)** *August 2021 - July 2023*
  Research Resident
  - Focusing on robust & trustworthy ML, studying different paradigms of explanation methods for machine learning models and their robustness.
  - Advisor: Prof. Viet Anh Nguyen
  - Applied Rotation Project: Interactive Tool for 3D Point Cloud Segmentation.

- **IBM** *July 2019 - October 2019*
  AI Research Intern
  - Applying PowerAI Vision to visual inspection problems in the car manufacturing process to detect dirt, and dust defects in the car body after painting.
  - Advisor: Dr. Tam Le Nhan

AWARDS & HONORS

- Honorable Mention in INFORMS Undergraduate Operations Research Prize. *2022*
- Best Thesis Presentation Award. *2021*
- Problem Winner in ASEAN-India Hackathon. *2021*
- Third prize in ACM/ICPC Asia - Ho Chi Minh Regional. *2017*
- Third prize in Vietnam Olympiad in Informatics. *2016*

TEACHING ASSISTANT

- Trustworthy Deep Learning, *Yale* *Spring 2024/Fall 2025*
- Applied Algorithms classes, *HUST* *2019 - 2021*

PROFESSIONAL SERVICES

- Reviewer at AISTATS 2022/2023, FaCCT 2023, UAI 2023, ICML 2025, NeuRIPS 2023/2024/2025 (Top Reviewer), ICLR 2024/2025.
- Reviewer at ACM Transactions on Sensor Networks (TOSN).
- Organizer at Non-Euclidean Foundation Models and Geometric Learning (NEGEL) Workshop at NeurIPS 2025.