Huy Nghiem

nghiemh@umd.edu | @hnghiem_ai | huy-nghiem | Homepage | Scholar

Education

2023 – 2027 **Ph.D. in Computer Science, University of Maryland, College Park**
Dean's Fellowship recipient.
Advisor: Professor Hal Daumé III.
GPA: 3.96/4.0.
Dissertation focus: Safety, Robustness and Trustworthiness in AI Systems.

2018 – 2021 **M.S. in Computer Science, University of Southern California, Los Angeles**
Concentration: Data Science.
GPA: 3.83/4.0.

2013 **B.S. in Applied Mathematics & Public Health, University of California, Irvine**
GPA: 3.5/4.0.

Research Experience

2023 – 2027 **Doctoral Student & Research Assistant, University of Maryland, College Park**
- **AI Oversight and Governance** Design oversight methods to detect and mitigate misaligned or deceptive behaviors in deployed AI systems.
- **LLM Safety & Alignment** Develop data-efficient, self-correcting LLM frameworks using preference optimization to improve robustness in content moderation.
- **Bias in AI Systems** Mitigate social and systemic biases in LLMs for employment, education, and healthcare applications using causal inference.
- **Causal Fairness** Develop counterfactual fairness methods with respect to protected attributes, enhancing transparency in AI decision-making.
- **Explainable AI** Leverage causal inference and in-context learning to reduce stereotypes in generative explanations from large language models.

Winter, 2026 **CBAI Research Fellow, Boston**
- **Mentor:** Dylan Hadfield-Menell
- **AI Technical Governance** Developing Model Spec Sheet as a blueprint to map components of training pipeline to downstream behaviors.

Spring, 2026 **MATS 9.0 & 9.1 Fellow, Berkeley**
- **Mentor:** Noam Kolt
- **AI + Legal Alignment** Developing audits methods to evaluate, predict, and orient LLM and AI tools to legal requirements.

2019 – 2023 **Researcher, Information Sciences Institute, Marina del Rey**
- **Cryptocurrency Fraud Detection** Develop deep CNN & RNN models achieving <6% error margin in detecting pump-and-dump frauds using social and financial data.
- **Content Moderation** Create crowdsourced annotation frameworks and improved anti-Asian hate speech detection by 8% F1-score using BERT & RoBERTa models.
- **Meta-Learning for NLP** Design Prototypical-MAML classifiers achieving >70% accuracy in cross-dataset offensive speech detection with only 100 labels.

Research Experience (continued)

2017 – 2023 **Data Scientist & Team Lead, Children's Data Network, Los Angeles**

- **Predictive Modeling for Social Good** Lead development of Random Forest & XGBoost models to optimize California's child welfare system, ensuring model transparency through SHAP and LIME interpretability tools.
- **Large-Scale Data Mining** Analyze historical administrative databases to build predictive classifiers for child abuse prevention, directly informing state and federal policy through dashboards and legislative reports.
- **Healthcare Analytics** Conduct mental health analyses using ICD-9/10 codings to improve Medicaid-insured children's services.

Industry Experience

2026 **Research Intern, AI Applied Science, New York City, NY**

- Design novel QUR evaluation framework for RAG artifacts to categorize risk profiles for downstream performance, leading to insights for optimization and system selection.
- Adapt QUR to agentic RAG systems to augment reliable contextual engineering.

2025 **GenAI Research Intern, Oracle Cloud Infrastructure (OCI), Burlington, MA**

- Build an iterative alignment framework for conversational medical assistants using preference optimization methods.
- Achieve up to 42% improvement in harmful query detection on the CARES-18K benchmark across multiple LLMs.

2024 **Graduate Intern, Capital One, McLean, VA**

- Implement alignment techniques (SFT, preference optimization, KTO) to strengthen LLM safety guardrails, boosting defense against adversarial attacks by 250%.
- Deploy LLM prototypes (Llama 2/3, Mixtral), reducing inference latency by 75%.
- Manager-nominated Leadership Award for initiating impactful research collaborations.

2015 – 2017 **Data Programmer, Edwards LifeSciences, Irvine, CA**

- Maintain ORACLE-based clinical database systems for FDA-compliant clinical trials.
- Create machine learning models to automate health record analysis and reporting.

Honors & Awards

2026 **MATS 9.0 & 9.1 scholar** – Acceptance rate below 7%

2025 **Outstanding Reviewer Award** – IJCNLP-ACL 2025

Best Paper Award – ML4H Symposium

Outstanding Paper Award – AAAI PDLM Workshop

2024 **Leadership Award** – Capital One (Manager-nominated)

2023 – 2027 **Dean's Fellowship** – University of Maryland, College Park

2021 **Finalist** – Data Science for Social Good Fellowship, Carnegie Mellon University

Awardee – UCSD Summer Training Academy for Research Success

2020 **Head of Programs** – USC Graduate and Rising in Information and Data Science

2019 **Finalist** – Center for Knowledge-Powered and Interdisciplinary Data Science

2018 **2nd Place** – USC–Boeing Inaugural Data Hackathon

Research Publications

Conference Proceedings

1. **H. Nghiem**, S.-T. Ho, and H. D. III, "Cait: Mitigating and certifying counterfactual sensitivity via activation editing," in *Under ICLR submission*, 2026.
2. **H. Nghiem**, S.-T. Ho, S. Wiegreffe, and H. D. III, "Trait-space monitoring for emergent misalignment during supervised finetuning," in *COLM Actionable Interpretability Worshop*, 2026. URL: https://arxiv.org/abs/2606.07631.
3. **H. Nghiem**, M. Kostylew, and N. Kolt, "HireWatch: Evaluating LLM Compliance with U.S. Employment Law Under Contextual Pressure," in *NeurIPS Evaluation & Dataset Track*, 2026.
4. **H. Nghiem**, P. A. Nguyen-Le, S.-T. Ho, and H. D. III, "Bias in the tails: How name-conditioned evaluative framing in resume summaries destabilizes LLM-based hiring," in *EMNLP*, 2026. URL: https://arxiv.org/abs/2604.19984.
5. **H. Nghiem**, A. Sachdeva, and H. Daumé III, "SMARTER: A data-efficient framework to improve toxicity detection with explanation via self-augmenting large language models," in *ACL*, 2026. URL: https://arxiv.org/abs/2509.15174.
6. S.-T. Ho, M. Liu, **H. Nghiem**, and F. Huang, "Soundnessbench: Can llms judge scientific soundness?" In *Under submission*, 2026. URL: https://arxiv.org/abs/2605.30329.
7. M. Sharma, C. B. C. Zhang, **H. Nghiem**, *et al.*, "Researchrubrics: A benchmark of prompts and rubrics for deep research agents," in *International Conference on Learning Representations*, 2026. URL: https://arxiv.org/abs/2511.07685.
8. **H. Nghiem**, P.-A. Nguyen-Le, J. Prindle, R. Rudinger, and H. Daumé III, "'Rich Dad, Poor Lad': How do large language models contextualize socioeconomic factors in college admission?" In *EMNLP*, 2025. URL: https://arxiv.org/abs/2509.16400.
9. **H. Nghiem**, S. Panda, D. Khatwani, H. V. Nguyen, K. Kenthapadi, and H. D. III, "Balancing safety and helpfulness in healthcare ai assistants through iterative preference alignment," in *ML4H 2025 Symposium Proceedings (Best Paper)*, 2025.
10. Z. Li, X. Wu, H. Du, F. Liu, **H. Nghiem**, and G. Shi, "A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE / CVF, 2025. DOI: 10.1109/CVPRW67362.2025.00147.
11. M. K. Rad*[1], **H. Nghiem***, S. Wadhwa, A. Luo, and M. S. Sorower, "Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment," in *AAAI Workshop on Preventing and Detecting LLM Misinformation (PDLM)*, 2025.
12. **H. Nghiem**, J. Prindle, J. Zhao, and H. Daumé III, "'You gotta be a doctor, Lin': An investigation of name-based bias of large language models in employment recommendations," in *EMNLP*, 2024, pp. 7268–7287. URL: https://aclanthology.org/2024.emnlp-main.413.
13. **Huy Nghiem** and H. Daumé III, "HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models," in *EMNLP Findings*, 2024, pp. 5938–5956.

Journal Articles

1. T. Nguyen, J. Merchant, X. Yue, *et al.*, "A decade of tweets: Visualizing racial sentiments towards minoritized groups in the united states between 2011 and 2021," *Epidemiology*, vol. 35, no. 1, pp. 51–59, 2024.
2. E. Putnam-Hornstein, R. Foust, S. Cuccaro-Alamin and J. Prindle and **H. Nghiem** and E. Ahn and L. Palmer, "A population-based study of mental health diagnoses and child protection system involvement among medicaid-insured children," *The Journal of Pediatrics*, vol. 252, pp. 117–123, 2023.

[1] *:equal contribution

3. S. Cuccaro-Alamin, A. L. Eastman, R. Foust, J. McCroskey, **H. Nghiem**, and E. Putnam-Hornstein, "Strategies for constructing household and family units with linked administrative records," *Children and Youth Services Review*, vol. 120, p. 105 706, 2021.
4. **H. Nghiem**, G. Muric, F. Morstatter, and E. Ferrera, "Detecting cryptocurrency pump-and-dump frauds using market and social signals," *Expert Systems with Applications*, vol. 182, p. 115 284, 2021.
5. R. Foust, **H. Nghiem**, J. Prindle, J. Hoonhout, J. McCroskey, and E. Putnam-Hornstein, "Child protection involvement among homeless families," *Journal of Public Child Welfare*, vol. 14, no. 5, pp. 518–530, 2020.

Technical Skills

- **Programming:** Python, SQL, SAS, JavaScript, MATLAB
- **ML/AI Frameworks:** PyTorch, TensorFlow, Keras, Hugging Face Transformers
- **LLM Techniques:** Fine-tuning (SFT, DPO, GRPO, KTO), alignment, RLHF, vLLM serving
- **ML Methods:** Deep Learning (CNN, LSTM, RNN, GAN), Causal Inference, Meta-Learning
- **Data & Analytics:** Spark, AWS, Kubernetes, Git, Tableau, Statistical Analysis
- **Specialized:** SHAP/LIME Interpretability, Crowdsourcing (AMT), Clinical Data Systems

Teaching Experience

2025 – 2026	**Research Assistant** Personalization of LLM-generated content
Spring 2024	**Teaching Assistant**, CSMC 116: You, Me and Generative AI
Fall 2024	**Teaching Assistant**, CMSC 839E: Uncertainty Communication for Decision-making
Fall 2023	**Teaching Assistant**, CMSC 320: Introduction to Data Science

Service

Committee

2025	**Junior Chair** Bias and Fairness Area, ML4H Symposium

Reviewing

- **Reviewer:** NeurIPS, ICLR
- **ACL Rolling Review (ARR):** ACL, EMNLP, AACL
- **Trustworthy NLP Workshop**
- **ACM Transactions on Intelligent Systems and Technology**
- **Journal of Medical Internet Research**
- **Machine Learning 4 Health (ML4H)**
- **Expert Systems with Applications**

Invited Talks

Winter 2026	**HireWatch: Measuring LLM compliance to US Labor Law**, MATS 9.0 Symposium
Spring 2025	**'You gotta be a Doctor Lin': The Risk of Bias in LLM**, UMD INST 414
Fall 2024	**A tutorial on Trustworthiness and Bias in LLMs** UMD Values-centered AI Institute
Summer 2023	**Summer Education Program for High Schools in STEM**