Mourad Heddaya

mourad@uchicago.edu | mheddaya.com | 425-753-5800

Experience building evaluation and alignment systems for LLMs. Built and deployed AI systems with automated pipelines for detecting hallucination, bias, and unsafe model behavior. Background in reasoning compression, long-context and multi-turn evaluation, and large-scale data processing.

Education

Ph.D. in Computer Science, University of Chicago, 2021–2026 (expected May). *Advisor: Chenhao Tan*

B.S. in Informatics, University of Washington, 2015–2019. *Advisors: Noah Smith & Mari Ostendorf*

Selected Publications

[1] **M. Heddaya**, R. Wadhawan, M. Roberts, C. Tan. When Internalization Fails: Finding Better Targets for Reasoning Compression. Under review, *ACL 2026*.

[2] **M. Heddaya**, K. MacMillan, A. Malani, H. Mei, C. Tan. CaseSumm: A Large-Scale Dataset for Long-Context Summarization. *Findings of NAACL 2025*. *Accepted with talk @ ALEA 2024*.

[3] **M. Heddaya**, C. Tan, R. Voigt, Q. Zeng, A. Zentefis. A Century of Inflation Narratives. *SSRN 2025*.

[4] **M. Heddaya**, Q. Zeng, C. Tan, R. Voigt, A. Zentefis. Causal Micro-Narratives. *WNU @ EMNLP 2024*.

[5] **M. Heddaya**, S. Dworkin, C. Tan, R. Voigt, A. Zentefis. Language of Bargaining. *ACL 2023*.

Experience

Doctoral Researcher, University of Chicago CS — 2021–2026

- [ongoing] Developing **AI to support critical thinking and civic deliberation** through CivicChats.org. Built an automated evaluation pipeline to detect hallucinations, bias, and stereotyping through **iterative prompt optimization over collected multi-turn user data**. Follow-up work on **large-scale data curation and synthetic data generation for post-training** to elicit coherent perspectives in models to support more productive user interactions. Wrote about this effort here.
- Built a 25.6K **long-context understanding benchmark** from Supreme Court opinions [2]. Showed fine-tuning improves automatic metrics while increasing hallucination rate. Found that a) **frontier models hallucinate less but more subtly, creating critical safety challenges in mitigating their risk,** and b) neither standard metrics nor LLM-as-judge reliably correlate with human judgement.
- Developed a method for **at-scale automatic extraction of causal narratives and applied it to 4.2M sentences** from a century of U.S. inflation coverage [4]. Found that causal narratives predict consumer expectations and spread via contagion and differentiation [3].
- Showed that language reduces price variance by 3x, through experimental design and controlled experiment (∼400 participants). Found that successful buyers use empathetic framing while pushing prices down whereas sellers succeed by asking user needs early in the conversation [5].

ML Research Scientist Intern, Abridge AI — Summer 2025

- Demonstrated that chain-of-thought internalization fails on complex reasoning [1]. Proposed **post-think reasoning distillation**, a simple yet effective compression target that preserves the teacher's solution path and achieves the **best accuracy–efficiency trade-off**. Motived by the need for faster inference that tightens the human oversight loop in **safety critical healthcare settings**.

Applied Scientist Intern, Amazon AWS AI Labs (Bedrock) — Summer 2023

- Developed a method for **efficient LLM safety alignment at scale without human annotation**, where the LLM scores its own diverse outputs and uses self-feedback as a regularizer, drastically reducing training speed and complexity by eliminating need to serve reference and reward models.

Research Engineer, University of Washington 2020–2021

- Industry collaboration (T-Mobile) building an unsupervised information extraction system for **massive real-world conversational speech data**. Identified distinct conversation paths corresponding to **successful vs. unsuccessful customer interactions** using learned HMM topology, delivering methods and analyses to industry partner. (Advisors: Noah Smith & Mari Ostendorf). [paper]

Other Papers & Research Mentorship

- J. Fu, **M. Heddaya**, C. Tan. Automatically Generating Hard Math Problems from Hypothesis-Driven Error Analysis. *LLM Reasoning @ ICLR 2026*.
- C. Shah, A. Agarwal, K. Garg, **M. Heddaya**. LLM Rationalis? Measuring Bargaining Capabilities of AI Negotiators. *MTI @ NeurIPS 2025*.

Invited Talks

- **Freestone Grove Partners Investment Firm**, April 2025 — *Causal Micro-Narratives*.
- **Max Planck Institute for Research on Collective Goods**, February 2025 — *NLP in Law & Economics*.
- **University of Chicago LEAP Workshop**, January 2023 — *Language of Bargaining*.

Service

- ARR Reviewer: June 2024, November 2024, October 2025, January 2026.
- IC2S2 Reviewer: 2022, 2023, 2024, 2025, 2026.
- Program Committee, Symposium on Human + AI, Fall 2022.

Teaching

University of Chicago
CMSC 25400 – Machine Learning, Winter 2023
CMSC 25300 / 35300 – Mathematical Foundation of Machine Learning, Fall 2022
CMSC 35100 – Natural Language Processing, Winter 2022