Allyson Ettinger

+1 (207) 975-6342 | allysone@allenai.org | aetting.github.io

Overview

Senior Research Scientist at Ai2 currently leading the AI for Science team under the Open Ecosystem group (the group that trains Olmo and Molmo models). I specalize in training data curation and methodologies for rigorous data ablation and quality verification. Previous work includes leading midtraining for Olmo 3, releasing state-of-the art open safety moderation for LLMs with the Ai2 Safety Toolkit, a large body of work pioneering methods for analysis, interpretability, and evaluation of LMs, as well as computational and wet lab research in cognitive neuroscience.

Academic Appointments

- **Allen Institute for AI** *2023 – Present*
  Research Scientist, Senior Research Scientist Seattle, WA
- **University of Chicago** *2019 – 2023*
  Assistant Professor, Department of Linguistics, Department of Computer Science Chicago, IL
- **Toyota Technological Institute at Chicago** *2018 – 2019*
  Research Assistant Professor Chicago, IL

Education

- **University of Maryland, College Park** *2013 – 2018*
  PhD in Computational Linguistics College Park, MD
  ◦ Dissertation: *Relating lexical and syntactic processes in language: Bridging research in humans and machines*
- **Brandeis University** *2010*
  BA in Linguistics and Psychology, Summa Cum Laude Waltham, MA

Publications (Selected)

Olmo Team, **Ettinger, A.**, Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., et al. (2025). **Olmo 3.** arXiv:2512.13961.

McGregor, S., **Ettinger, A.**, Judd, N., Albee, P., Jiang, L., Rao, K., Smith, W. H., Longpre, S., Ghosh, A., Fiorelli, C., et al. (2025). **To Err Is AI: A Case Study Informing LLM Flaw Reporting Practices.** In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Shi, W., Bhagia, A., Farhat, K., Muennighoff, N., Morrison, J., Walsh, E. P., Schwenk, D., Longpre, S., Poznanski, J., **Ettinger, A.**, et al. (2024). **FlexOLMo: Open Language Models for Flexible Data Use.** arXiv:2507.07024.

OLMo Team, Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., et al. (2024). **2 OLMo 2 Furious.** arXiv:2501.00656.

Jiang, L., Rao, K., Han, S., **Ettinger, A.**, Brahman, F., Kumar, S., Mireshghallah, N., Lu, X., Sap, M., Choi, Y., et al. (2024). **Wildteaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models.** arXiv:2406.18510.

Han, S., Rao, K., **Ettinger, A.**, Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., Dziri, N. (2024). **WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs.** arXiv:2406.18495.

West, P., Lu, X., Dziri, N., Brahman, F., Li, L., **Ettinger, A.**, Choi, Y. (2023). **The Generative AI Paradox: "What It Can Create, It May Not Understand."** In *International Conference on Learning Representations (ICLR)*.

Yang, C., **Ettinger, A.** (2023). **Can You Follow Me? Testing Situational Understanding for ChatGPT.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ettinger, A., Hwang, J. D., Pyatkin, V., Bhagavatula, C., Choi, Y. (2023). **Limits of LLMs as Analyzers of Abstract Meaning Representation.** In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Dziri, N., et al., **Ettinger, A.** (2023). **Faith and Fate: Limits of Transformers on Compositionality.** In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, J., Yu, L., **Ettinger, A.** (2023). **Counterfactual Reasoning in Language Models.** In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Misra, K., Taylor Rayz, J., **Ettinger, A.** (2023). **COMPS: Conceptual Minimal Pair Sentences.** In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Best Paper Award.

Li, J., **Ettinger, A.** (2023). **Heuristic Interpretation as Rational Inference.** *Cognition*.

Ettinger, A. (2020). **What BERT Is Not.** *Transactions of the Association for Computational Linguistics*.

Ettinger, A., Elgohary, A., Phillips, C., Resnik, P. (2018). **Assessing Composition in Sentence Vector Representations.** In *Proceedings of COLING 2018*.

Ettinger, A., Resnik, P., Carpuat, M. (2016). **Retrofitting Sense-Specific Word Vectors Using Parallel Text.** In *Proceedings of NAACL-HLT 2016*.

Grants

- **NSF EAGER Grant**
  World Modeling for Natural Language Understanding (Co-PI) 2019 – 2021
- **NSF Graduate Research Fellowship**
  2014 – 2017

Mentorship

- **PhD Advising**
  Primary Advisor / Committee Member
  - Chenghao Yang, PhD in Computer Science. *Beyond Surface Alignment: Grounding the Dynamics of Situational Understanding and Generative Control in LLMs*.
  - Lang Yu, PhD in Computer Science. *Analyzing and Improving Compositionality in Neural Language Models*.
  - Kanishka Misra (Purdue University), PhD in Natural Language Understanding. *On Semantic Cognition, Inductive Generalization, and Language Models*.
  - Sanghee Kim, PhD in Linguistics. *Encoding and retrieval of discourse structure during language comprehension.*;
  - Xiaomeng Ma (CUNY), PhD in Linguistics. *Evaluating Neural Networks as Cognitive Models for Learning Quasi-regularities in Languages*.
  - Karim Lasri (École normale supérieure), PhD in Language Science. *Linguistic Generalization in Transformer-Based Neural Language Models*.
  - Laura Aina (Universitat Pompeu Fabra), PhD in Translation and Language Sciences. *A Deep Learning Perspective on Linguistic Ambiguity*.
  - Daniel Edmiston, PhD in Linguistics. *Identifying Linguistic Structures in Continuous Spaces*.
- **Master's Thesis Advising**
  Advisor
  - Jiaxuan Li, Master's Program in Computational Social Science. *A Noisy Channel Model of N400 and P600 Effects in Sentence Comprehension*.
  - Wanitchaya (Mint) Poonpatanapricha, Master's Program in Computational Social Science. *Use of Emojis as Markers of Politeness*.
  - Ethan Liu, Master's Program in Statistics. *Causal Relations and Feature Extraction in Contextualized Sentence Representation*.
  - Zhou Xing, Master's Program in Computational Social Science. *Visually Grounded Multilingual Syntax Acquisition*.
  - Alexander (Sasha) Tyan, Master's Program in Computational Social Science. *Using Topic Modelling to Detect Bias Between News Sources*.
- **Undergraduate Thesis Advising**
  Advisor
  - Michael Hanna, BA in Linguistics and Computer Science. *Evaluating the Interpretability of Variational Autoencoders' Latent Space Representations of Sentences*.
  - Hanson Lu, BA in Linguistics and Computer Science. *Applying the Rational Speech Acts Model to Coreference Resolution*.

Honors and Awards

- **Best Paper Award, EACL**
  Awarded for paper: COMPS: Conceptual Minimal Pair Sentences 2023
  - Brief description of significance (e.g., top X% of submissions; competitive international conference).
  - Optional note (e.g., selection statistics, citation impact, media coverage).
- **Best Proposal Award, ACL Workshop**
  Awarded for workshop proposal on [Topic] 2016
  - Brief context about workshop scale or competitiveness.
- **NSF Graduate Research Fellowship**
  National Science Foundation 2014
  - Highly competitive national fellowship supporting graduate research in STEM.
  - Selection rate: XX% (optional).