Abulhair Saparov

Assistant Professor of Computer Science

475 Stadium Mall Drive, West Lafayette, Indiana, 47907, United States

asaparov {at} purdue.edu | asaparov.org | asaparov | YouTube

Research interests

- Applications of machine learning to natural language processing (NLP), semantic parsing, natural language understanding (NLU)
- Reasoning in large language models (LLMs), analysis of LLMs, mechanistic interpretability
- Symbolic and neuro-symbolic methods for reasoning, language understanding, worldbuilding
- Representations of meaning/knowledge, reasoning, especially in natural language understanding
- Statistical machine learning, interpretable machine learning, Bayesian nonparametrics, scalable inference
- Broadly interested in applications in e.g. epidemiology, linguistics, phylogenetics, biology, information security, etc.

Education

Carnegie Mellon University

Pittsburgh, PA

PhD in Machine Learning

2017 - 2022

- Advisor: Tom M. Mitchell
- Thesis: Towards General Natural Language Understanding with Probabilistic Worldbuilding
- Thesis committee: Tom M. Mitchell, William Cohen, Frank Pfenning, Vijay Saraswat

Carnegie Mellon University

Pittsburgh, PA

MS in Machine Learning

2013 - 2017

- Advisor: Tom M. Mitchell
- Degree requirements completed while in the PhD Program for Machine Learning.

Princeton University

Princeton, NJ

BSE in Computer Science

2009 - 2013

- Summa cum laude
- Certificate (minor) in Applied and Computational Mathematics
- Certificate (minor) in Neuroscience
- Thesis advisors: Ken A. Norman, David M. Blei

Publications

- Daking Rai, Yilun Zhou, Shi Feng, **Abulhair Saparov**, Ziyu Yao (2024). A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models. CoRR, abs/2407.02646. [**link**]
- Nitish Joshi, **Abulhair Saparov**, Yixin Wang, He He (2024). LLMs Are Prone to Fallacies in Causal Inference. CoRR, abs/2406.12158. [**link**]
- Usman Anwar, **Abulhair Saparov**, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramer, He He, Atoosa Kasirzadeh, Yejin Choi, David Krueger (2024). Foundational Challenges in Assuring Alignment and Safety of Large Language Models. CoRR, abs/2404.09932. [**link**]
- Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, **Abulhair Saparov**, Mrinmaya Sachan (2024). Do Language Models Exhibit the Same Cognitive Biases in Problem Solving as Human Learners? To appear in ICML. [**link**]
- Nitish Joshi*, Javier Rando*, **Abulhair Saparov**, Najoung Kim, He He (2023). Personas as a Way to Model Truthfulness in Language Models. CoRR, abs/2310.18168. [**link**]
  ***equal contribution*
- **Abulhair Saparov**, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, He He (2023). Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples. Advances in Neural Information Processing Systems 36, NeurIPS 2023. [**link**]

- Hongyi Zheng, **Abulhair Saparov** (2023). Noisy Exemplars Make Large Language Models More Robust: A Domain-Agnostic Behavioral Analysis. Empirical Methods in Natural Language Processing, EMNLP 2023. [**link**]
- Vaibhav Mavi, **Abulhair Saparov**, Chen Zhao (2023). Retrieval-Augmented Chain-of-Thought in Semi-structured Domains. Natural Legal Language Processing Workshop @ EMNLP 2023. [**link**]
- Andreas Opedal, Niklas Stoehr, **Abulhair Saparov**, and Mrinmaya Sachan (2023). World Models for Math Story Problems. Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics. [**link**]
- **Abulhair Saparov**, He He (2023). Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. International Conference on Learning Representations. [**link**]
- **Abulhair Saparov**, Tom M. Mitchell (2022). Towards General Natural Language Understanding with Probabilistic Worldbuilding. Transactions of the Association for Computational Linguistics (TACL), 10, 325–342. [**link**]
- **Ph.D. Thesis:** Towards General Natural Language Understanding with Probabilistic Worldbuilding [**link**]
- **Abulhair Saparov** (2022). A Probabilistic Generative Grammar for Semantic Parsing. CoRR, abs/1606.06361. [**link**]
- Emmanouil A. Platanios*, **Abulhair Saparov***, Tom M. Mitchell (2020). Jelly Bean World: A Testbed for Never-Ending Learning. International Conference on Learning Representations. [**link**]
  ***equal contribution*
- Tom M. Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bishan Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matther Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohammad, Ndapa Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, **Abulhair Saparov**, Malcolm Greaves, Joel Welling (2018). Never-Ending Learning. Communications of the ACM, 61(5), 103–115. [**link**]
- **Abulhair Saparov**, Vijay Saraswat, Tom M. Mitchell (2017). A Probabilistic Generative Grammar for Semantic Parsing. *Proceedings of the Twenty-First Conference on Computational Natural Language Learning*. [**link**]
- **Abulhair Saparov**, Michael A. Schwemmer (2015). Effects of passive dendritic tree properties on the firing dynamics of a leaky-integrate-and-fire neuron. *Mathematical Biosciences*, 269, 61-75.
- Tom M. Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohammad, Ndapa Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijay, Abhinav Gupta, Xinlei Chen, **Abulhair Saparov**, Malcolm Greaves, Joel Welling (2015). Never-ending Learning. AAAI. [**link**]
- Xiaobai Chen, **Abulhair Saparov**, Bill Pang, and Thomas Funkhouser (2012). Schelling Points on 3D Surface Meshes, *ACM Transactions on Graphics (Proc. SIGGRAPH)*. [**link**]

Honors & Awards

2023	**Best Reviewer**, Conference on Empirical Methods in Natural Language Processing	*Singapore*
2023	**Outstanding Area Chair**, 61st Meeting of the Association for Computational Linguistics	*Toronto, Canada*
2015	**Teaching Assistant Award**, Machine Learning Department, CMU	*Pittsburgh, PA*
2013	**Honorable Mention**, NSF Graduate Research Fellowship	
2013	**Best Paper**, Program in Applied and Computational Mathematics, Princeton University	*Princeton, NJ*

Teaching

10-601 Introduction to Machine Learning (graduate course) — *Pittsburgh, PA*

Teaching Assistant — *Jan. 2015 - May 2015*

- Recorded, edited, and uploaded lecture videos to YouTube.
- Led recitations, created and graded homework assignments/exams, and supervised student groups on final project work.
- Received the Teaching Assistant Award from the Machine Learning Department.

10-701/15-781 Introduction to Machine Learning (graduate course) — *Pittsburgh, PA*

Teaching Assistant — *Sep. 2014 - Dec. 2014*

- Recorded, edited, and uploaded lecture videos to YouTube.
- Led recitations, created and graded homework assignments/exams, and supervised student groups on final project work.

Experience

Purdue Department of Computer Science *West Lafayette, IN*

Assistant Professor *August 2024 -*

- Continuing research on utilizing symbolic methods to better study and understand NLU models, including LLMs.
- As well as research into improving the reasoning abilities of NLU models by incorporating symbolic methods.
- Teaching seminar and lecture courses in NLP.

NYU Center for Data Science *New York, NY*

Post-Doctoral Associate *June 2022 - July 2024*

- Worked on reasoning in NLP and NLU; deductive reasoning, causal reasoning, truthfulness in large language models.
- Advisor: He He, Research Group: ML2, CILVR

Rogo *New York, NY*

NLP Advisor *Nov. 2022 -*

- Consulting and advising the company on product research and development.
- Providing guidance and updates on cutting-edge NLP research relevant to the product.

IBM Thomas J. Watson Research Center *Yorktown Heights, NY*

Research Intern *Summer 2016*

- Developed grammar induction algorithms to train a semantic parser on a dataset of questions and corresponding logical forms.
- Advisor: Vijay Saraswat

The McGraw Center *Princeton, NJ*

Freshman Scholars Institute QuantLab Tutor *Summer 2012*

- Led group tutoring sessions for incoming Princeton University freshman students coming from schools without access to AP courses.

Princeton Department of Computer Science *Princeton, NJ*

Research Assistant *Feb. 2012 - May 2013*

- Developed and implemented novel inference algorithms to fit a probabilistic model of latent sources to fMRI brain activity data.
- Advisors: Sam Gershman, Ken Norman, David M. Blei

Princeton Program in Applied and Computational Mathematics *Princeton, NJ*

Research Assistant *Oct. 2011 - May 2013*

- Explored and analyzed novel dynamical behavior in the multi-compartment leaky integrate-and-fire model of a neuron.
- Advisors: Michael A. Schwemmer, Philip J. Holmes

Google *New York, NY*

Software Engineering Intern *Summer 2011*

- Designed and developed new features for the front-end of the DoubleClick for Publishers advertising platform, collaborating with many different software engineering and product management groups at every stage of the test-driven development process.

Re-verb Apparel Co.

Chief Web Designer *Nov. 2010 - 2012*

- Implemented and helped design the website, including online store, using HTML5 frontend technologies and a Wordpress backend.

Princeton Department of Computer Science *Princeton, NJ*

Research Assistant *Jun. 2010 - Jan. 2011*

- Designed a novel feature point detection algorithm for 3D surface meshes.
- Analyzed large data sets gathered from an online user study to find patterns in human-generated point sets on 3D meshes.
- Advisor: Thomas Funkhouser.

Other information

Languages native in English, conversational in Kazakh, intermediate in Spanish, beginner in Japanese