Satyapriya Krishna

Webpage: satyapriyakrishna.com
Email: skrishna@g.harvard.edu,
krishna.satyapriya123@gmail.com
Google Scholar : satyapriya-krishna
GitHub: github.com/y12uc231

Education

Harvard University — Cambridge, MA
Doctor of Philosophy in Computer Science — Jan 2025

- Thesis: *From Understanding to Improving Artificial Intelligence: New Frontiers in Machine Learning Explanations*
- Advisor: Prof. Himabindu Lakkaraju and Prof. Finale Doshi-Velez

Carnegie Mellon University — Pittsburgh, PA
Master of Science, School of Computer Science, Advisor: Prof. Anatole Gershman — 2018

- Thesis: "Conversational Agents based on Deep Reinforcement Learning"
- GPA : 3.91/4.00

LNM Institute of Information Technology — Jaipur, India
Bachelors in Computer Science & Engineering — 2016

- GPA: 9.29/10.00
- Rank: 3/350

Publications

[1] **Krishna, Satyapriya**, N. Mehrabi, A. Mohanty, M. Memelli, V. Ponzo, P. Motwani, and R. Gupta, "Evaluating the critical risks of amazon's nova premier under the frontier model safety framework", *Amazon Technical Reports*, 2025.

[2] **Krishna, Satyapriya**, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, U. Shyam, and M. Faruqui, "Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation", *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[3] **Krishna, Satyapriya** and other authors, "Ailuminate: Introducing v1. 0 of the ai risk and reliability benchmark from mlcommons", *MLCommons Benchmarks*, 2025.

[4] **Krishna, Satyapriya** and other authors, "Humanity's last exam", *ArXiv*, 2025.

[5] **Krishna, Satyapriya**, C. Agarwal, and H. Lakkaraju, "Understanding the effects of iterative prompting on truthfulness", *The Forty-first International Conference on Machine Learning (ICML)*, 2024.

[6] A. J. Li, **Krishna, Satyapriya**, and H. Lakkaraju, "More rlhf, more trust? on the impact of human preference alignment on language model trustworthiness", *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

[7] B. Peng, D. Goldstein, Q. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, Ferdinan, **Krishna, Satyapriya**, *et al.*, "Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence", *Conference on Language Model (COLM)*, 2024.

[8] S. Casper, C. Ezell, C. Siegmann, N. Kolt, T. L. Curtis, B. Bucknall, A. Haupt, K. Wei, J. Scheurer, M. Hobbhahn, **Krishna, Satyapriya**, *et al.*, "Black-box access is insufficient for rigorous ai audits", *The ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2024.

[9] **Krishna, Satyapriya**, J. Ma, D. Slack, A. Ghandeharioun, S. Singh, and H. Lakkaraju, "Post hoc explanations of language models can improve language models", *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[10] **Krishna, Satyapriya**, "On the intersection of self-correction and trust in language models", 2023.

[11] N. Kroeger, D. Ley, **Krishna, Satyapriya**, C. Agarwal, and H. Lakkaraju, "Are large language models post hoc explainers?", *(Under Review) The Twelfth International Conference on Learning Representations*, 2023.

[12] **Krishna, Satyapriya**, J. Ma, and H. Lakkaraju, "Towards bridging the gaps between the right to explanation and the right to be forgotten", *The Fortieth International Conference on Machine Learning (ICML)*, 2023.

[13] C. Agarwal, **Krishna, Satyapriya**, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju, "Openxai: Towards a transparent evaluation of model explanations", *36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[14] **Krishna, Satyapriya**, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective", *Interpretable Machine Learning in Healthcare (IMLH) at Thirty-ninth International Conference on Machine Learning (ICML)*, 2022.

[15] **Krishna, Satyapriya**, C. Agarwal, and H. Lakkaraju, "On the impact of adversarially robust models on algorithmic recourse", *7th AAAI Conference on AI, Ethics, and Society*, 2024.

[16] D. Slack, **Krishna, Satyapriya**, H. Lakkaraju, and S. Singh, "Explaining machine learning models with interactive natural language conversations using talktomodel", *Nature Machine Intelligence*, pp. 1–11, 2023.

[17] C. Agarwal, N. Johnson, M. Pawelczyk, **Krishna, Satyapriya**, E. Saxena, M. Zitnik, and H. Lakkaraju, "Rethinking stability for attribution-based explanations", in *ICLR 2022 Workshop on PAIR*, 2022.

[18] **Krishna, Satyapriya**, R. Gupta, A. Verma, J. Dhamala, Y. Pruksachatkun, and K.-W. Chang, "Measuring fairness of text classifiers via prediction sensitivity", *The Joint Conference of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 2022.

[19] U. Gupta, J. Dhamala, V. Kumar, A. Verma, Y. Pruksachatkun, **Krishna, Satyapriya**, R. Gupta, K.-W. Chang, G. V. Steeg, and A. Galstyan, "Equitable text generation with distilled language models via counterfactual role reversal", *The Joint Conference of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022 - Findings)*, 2022.

[20] **Krishna, Satyapriya**, R. Gupta, and C. Dupuy, "Adept: Auto-encoder based differentially private text transformation", *The 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[21] J. Dhamala, T. Sun, V. Kumar, **Krishna, Satyapriya**, Y. Pruksachatkun, R. Gupta, and K.-W. Chang, "Bold: Dataset and metrics for measuring biases in open-ended language generation", *The ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2021.

[22] Y. Pruksachatkun, **Krishna, Satyapriya**, A. Ramakrishna, J. Dhamala, and R. Gupta, "Trustnlp: First workshop on trustworthy natural language processing", *The North American Chapter of the Association for Computational Linguistics*, 2021.

[23] A. Patel, R. Gupta, M. Harakere, **Krishna, Satyapriya**, A. Alok, and P. Liu, "Towards classification parity across cohorts", *ML-IRL Workshop, International Conference on Learning Representations (ICLR)*, 2020.

[24] Y. Tao, S. Gupta, **Krishna, Satyapriya**, X. Zhou, O. Majumder, and V. Khare, "Finetext: Text classification via attention-based language model fine-tuning", *Amazon Machine Learning Conference (AMLC)*, 2019.

[25] **Krishna, Satyapriya**, Q. Hu, Y. Zhang, B. Yin, and H. Rangwala, "Document expansion with keyword extraction from massive product catalog", *Amazon Machine Learning Conference (AMLC)*, 2020.

[26] **Krishna, Satyapriya**, M. de Jong, and A. Agarwal, "Grounding complex navigational instructions using scene graphs", *arXiv preprint arXiv:2106.01607*, 2018.

Work Experience

Amazon AGI (Frontier AI Safety) Boston, MA
Sr. Researcher Feb 2025 - still working

- Released first frontier AI high-risk safety evaluations for `Nova Premier` (Amazon's most capable AI model).

Google Gemini/Deepmind New York City, NY
Research Intern May 2024 –Aug 2024

- Built a search agent evaluation benchmark, named `FRAMES`, being used to evalaute frontier search agents. We presented this work at NAACL 2025.

Meta Platforms, Inc. Menlo Park, CA
Research Scientist Intern May 2022 –Aug 2022

- Worked with Responsible AI team to build a novel explanation method that not only generate attribution scores but also provide the reliability estimates associated with explanations.

Amazon Alexa AI Boston, MA
Sr. Researcher (Applied Scientist - 2) Dec. 2019 –Aug 2021

- Led Trustworthy Natural Language Understanding (NLU) team to build robust, fair and interpretable language processing models for Amazon Alexa
- Launched the first fairness/bias measurement and mitigation pipeline for Alexa

Amazon Search (A9.com) Palo Alto, CA
Applied Scientist - 1 Oct. 2018 –Nov. 2019

- Launched the first set of document summarization models on amazon retail worldwide (15 countries)
- Launched the first version of Amazon Product Graph which improved search recall by 27%

Amazon Web Services Seattle, WA
Machine Learning Engineer Jul 2018 –Sep 2018

- Launched universal state-of-the-art text-classification models on AWS Sagemaker

Carnegie Mellon University Pittsburgh, PA
Teaching Assistant Nov 2016 –May 2018

- Teaching assistant in Math for Machine Learning(10-600) to Prof. Geoffrey J. Gordon
- Teaching assistant in Practical Data Science(15-688) to Prof. Zico Kolter
- Teaching assistant in PhD-Intro to ML(10-701) to Prof. Manuela M. Veloso and Prof. Pradeep Ravikumar

Amazon Research Seattle, WA
Research Intern May 2017 –Aug 2017

- Developed software to generate alerts for potential physical-virtual geolocation mismatches, when the physical location of a particular logistic (trailer/package) is not consistent with the system
- Applied clustering algorithms(EM, K-Means) to investigate factors behind mismatches and developed a real-time geo-location tracker which tracked logistic movement for Amazon fulfillment centers worldwide

Indian Institute of Technology (IIT) Chennai, India
Research Intern May 2015 –Sep 2015

– Developed state-of-the-art face recognition algorithms under the supervision of Prof. Sukendu Das from Visualization and Perception Lab

National Institute of Technology (NIT) Rourkela, India
Research Intern May 2014 –Sep 2014

– Developed statistical models to detect and remove shadow regions from low-resolution videos from security video sensors to maximize feature extraction in Computer Vision Lab, supervised by Prof. Banshidhar Majhi

Scholarships and Awards

- Awarded Nicole A. Chen and Karina A. Chen Graduate Student Research Fellowship 2021–2022
- Awarded Student Volunteer Award at the 60th Annual Meeting of the Association for Computational Linguistics 2022
- Awarded Director's Scholarship four times for being in the top 3 (out of 400) performers, LNM Institute of Information Technology (India) - ₹80000 2012–2016

Coursework

- **Undergraduate:** Design & Analysis of Algorithms, Data Structures, Genetic Programming, Software Engineering, Graph Theory, Probability & Statistics, Database Management , Machine Learning
- **Masters:** Bayesian Statistics, Multivariate Analysis, Causal Inference, Deep Reinforcement Learning, Deep Learning, Conversational agents, Machine Learning, Big Data Analytics, Robotics & Machine Learning, Natural Language Processing, Applied Machine Learning, Math for Machine Learning, Independent Research : Machine Learning for Social Good
- **PhD:** COMPSCI 226R: Topics in Theory for Society: The Theory of Algorithmic Fairness, COMPSCI 288: AI for Social Impact, and COMPSCI 242: Computing at Scale.

Notable Projects

- Grounding Complex Navigational Instructions Using Scene Graphs — Ranked 2nd out of 40 projects
  Tools: Python, Tensorflow
  - Trained agents to navigate in gaming environment(ViZDoom) using natural language instructions
  - Achieved state-of-the-art accuracy in reaching target objects with 20% faster travel time

- **SynSem** : Semi- Goal Oriented Conversational Agent — Masters Research Project
  Tools: Python, CuDNN, C++, PyTorch
  - Developed chat agents with better context-transitions and initiations, improving conversation time by 5 times compared against other end-to-end chatbots, along with 70% improvement in end-to-end training speed
  - Experimented with Asynchronous Actor-Critic (A3C) and Proximal Policy Optimization (PPO) learning algorithms, and improved inference time with C++/cuDNN programmed modules

- Universal Semantic Parser — Capstone Project
  Tools: Pytorch
  - Developed seq-2-seq based semantic parser trained for multiple formalisms (ATIS, GeoQuery, Overnight) to achieve faster parsing across different formalisms by maximizing parameter-sharing
  - Reduced parameters by 30% against state-of-the-art baseline with comparable performance

Skills

- **Programming:** Python • Java • C++ • C • R
- **Frameworks:** PyTorch • Keras • OpenGym • CUDA, cuDNN • MXNET • Tensorflow • Pandas • Hadoop • Spark • MongoDB • Apache Hive • AJAX • jQuery • Kubernetes • Docker
- **Tools & Utilities:** Bash • AWS Services • Spring • Apache Pig • Node.js • SQL • Apache Hive • Vim • Conda • Jupyter • Rstudio • tmux • Git • Slurm Workload Manager • Microsoft Office.
- **Operating Systems :** Unix (MacOS) • Linux (CentOS and Ubuntu) • Windows 10

Service to the Community, Organization, and Profession

Editorial Committees Boston, MA

Area Chair/Program Committee/Organizing Committee/Meta-Reviewer/ Reviewer

- **Area Chair** : AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI
- **Program Committee** : FAccT (2023), AAAI (2022, 2023), IEEE International Conference on Big Data (2023)
- **Reviewer** for AAAI (2021,2022,2023), NeurIPS (2021, 2022, 2023), ICML(2023), ACL (2022, 2023), EMNLP (2022), ICLR(2024) and AACL(2022)
- **Organizing Committee** : TrustNLP@NAACL-2021
- Reviewed a submission for Nature Machine Intelligence Journal in 2022
- Reviewed two submissions for Artificial Intelligence for the Earth Systems (AIES) Journal

Machine Fairness - Amazon Research Seattle, WA

Panelist & Tutor Jul 2018 –current

- Participated in panel discussions on social bias in machine learning at Amazon Research. Also organized hackathons, poster presentations, and tutorials to motivate research in the area of Fairness, Accountability, Transparency and Ethics (FATE)

Speech & Language Research - Amazon Research Seattle, WA

Mentor Dec 2018 –current

- Mentoring research scientists on conducting high-quality & effective research at Amazon through weekly discussions

LNM Institute Of Information Technology Jaipur, India

Mentor May 2020 –current

- Mentoring students from my undergrad university on various topics related to career path, graduate applications, and academic research.

ConvAI - Carnegie Mellon University Pittsburgh, PA

Organizer & Speaker 2016 - 2018

- Organized reading group to invite discussions on research problems related to conversational AI. This group was joined by students and researchers from various departments in the university, which also resulted in several research publications