NARUTATSU RI

nr3764@princeton.edu | narutatsu-ri | narutatsuri | narutatsuri.github.io

EDUCATION

Princeton University — Princeton, New Jersey
Ph.D. in Computer Science — Sep 2025 – Present
Advisor: Sanjeev Arora

Columbia University — New York, NY
M.S. in Computer Science, MS-GRA (Full-Ride Graduate Research Assistant) — Sep 2024 – May 2025
Advisor: Kathleen McKeown

Columbia University — New York, NY
B.S. in Computer Science, Egleston Scholar — Sep 2020 – May 2024
Advisors: Kathleen McKeown, Daniel Hsu, Nakul Verma

HONORS

Gordon Wu Fellowship	2025
Theodore R. Bashkow Research Award	2024, 2025
Dean's List	2020 – 2024
Upsilon Pi Epsilon Candidate	2023
Tau Beta Pi Candidate Junior Cohort	2022
Egleston Scholar	2020
Ezoe Memorial Foundation Academic Scholarship	2019

SELECTED PUBLICATIONS

[1] Reranking-based Generation for Unbiased Perspective Summarization
Narutatsu Ri, Nicholas Deas, Kathleen McKeown
ACL 2025 Findings

[2] Speak Easy: Eliciting Harmful Instructions from LLMs with Simple Interactions
Yik Siu Chan*, **Narutatsu Ri***, Yuxin Xiao*, Marzyeh Ghassemi
ICML 2025

[3] Latent Space Interpretation for Stylistic Analysis and Explainable Authorship Attribution
Milad Alshomary, **Narutatsu Ri**, Marianna Apidianaki, Ajay Patel, Smaranda Muresan, Kathleen McKeown
COLING 2025

[4] The Effect of Model Capacity on the Emergence of In-Context Learning in Transformers
Berkan Ottlik*, **Narutatsu Ri***, Daniel Hsu, Clayton Sanford
ICLR 2024 ME-FoMo Workshop

[5] Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations
Yanda Chen, Ruiqi Zhong, **Narutatsu Ri**, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, Kathleen McKeown
ICML 2024 (Spotlight)

[6] Enhancing Few-shot Text-to-SQL Capabilities of Large Language Models: A Study on Prompt Design Strategies
Linyong Nan, Yilun Zhao, Weijin Zou, **Narutatsu Ri**, Jaesung Tae, Ellen Zhang, Arman Cohan, Dragomir Radev
EMNLP 2023 Findings

[7] Contrastive Loss is All You Need to Recover Analogies as Parallel Lines
Narutatsu Ri, Fei-Tzin Lee, Nakul Verma
ACL 2023 RepL4NLP Workshop

GRANT AND RESEARCH EXPERIENCE

IARPA HIATUS Program, Columbia University — May 2024 – Aug 2024, Jan 2025 – May 2025
Graduate Research Assistant, with Kathleen McKeown

- Developed a text style-transfer–based authorship obfuscation system for submission to IARPA's evaluation. Improved on previous model performance in English and Russian.
- Developed an explainable authorship-attribution method mapping the embedding space of attribution models to informative and interpretable natural language features. Achieved improved performance compared to multiple baseline stylistic explanation methods.

Knight First Amendment Institute, Columbia University — Sep 2024 – Dec 2024
Graduate Research Assistant, with Kathleen McKeown

◦ Worked on mitigating input bias (e.g., length, position, stance, etc.) in large language models (LLMs) for multi-document perspective summarization. Developed new metrics for measuring summary coverage and faithfulness and new method for generating unbiased perspective summaries.

DARPA CCU Program, Columbia University Sep 2023 – May 2024
Undergraduate Research Assistant, with Kathleen McKeown

◦ Developed changepoint detection methods for multi-turn conversations. Implemented a translate-train approach and data augmentation techniques for multilingual datasets. Resulted in ∼20% performance improvements in Chinese and Spanish and established new strong baseline for Russian.

◦ Investigated the precision of explanations generated by large language models. Developed a novel metric, counterfactual simulatability, to assess the accuracy of LLM-generated explanations.

Data Science Institute, Columbia University May 2023 – Jan 2024
Undergraduate Research Assistant, with Daniel Hsu

◦ Investigated the emergence of in-context learning in transformer models within a statistical fixed design regression setting. Demonstrated how limiting model capacity and training data diversity encourages transformers to shift from a memorization (Bayesian) estimator to a generalizing (James-Stein) estimator for out-of-distribution data.

LILY Lab, Yale University Jan 2023 – May 2023
Undergraduate Research Assistant, with Dragomir Radev

◦ Worked on implementing and benchmarking in-context learning capabilities for Text-to-SQL tasks using large language models. Studied LLMs by optimizing prompt design strategies and investigating demonstration selection methods. Achieved state-of-the-art performance on Spider dataset.

International Research Center for Neurointelligence, The University of Tokyo May 2022 – Sep 2022
Visiting Researcher, with Mingbo Cai

◦ Analyzed manifold structure in contextualized BERT embeddings, identifying and characterizing horseshoe effect within embeddings. Designed and developed a novel framework for decoding syntactic information from raw fMRI brain activity for visual movie scene descriptions.

Department of Computer Science, Columbia University Jan 2021 – May 2024
Undergraduate Research Assistant, with Nakul Verma

◦ Conducted systematic analysis of word embedding models, and proved intrinsic relationship between the formation of analogies as parallel structures (analogy parallelism) and word co-occurrence statistics. Developed Contrastive Word Model (CWM), word embedding model employing a simple contrastive learning objective. Demonstrated analogy recovery performance on par with existing word embedding models with ∼50 times shorter training time.

TEACHING

Columbia University, Department of Computer Science Sep 2024 – Dec 2024
Teaching Assistant, Unsupervised Learning (COMS 4774)

◦ Served as Head TA; oversaw assignments, office hours, and final projects.

◦ Course covers the theoretical and algorithmic foundations of unsupervised machine learning (clustering and its guarantees; linear and non-linear dimensionality reduction; density estimation; latent variable models; manifold and topological methods; metric embeddings).

Columbia University, Department of Computer Science Jul 2022 – May 2024
Teaching Assistant, Machine Learning (COMS 4771)

◦ Served three terms as Head TA and two as regular TA; oversaw creating and grading assignments, exam preparation, and office hours.

◦ Course covers the core theory and practice of machine learning (MLE; Bayesian, generative and discriminative classifiers; kernel methods and SVMs; regularized regression; PAC/VC theory; clustering EM; PCA manifold learning; graphical models and HMMs).

SERVICE

Conference Reviewer

ICLR 2024
ACL, RepL4NLP Workshop 2024