SIREESH GURURAJA

sireesh.gururaja@gmail.com ⋄ https://siree.sh

EDUCATION

Carnegie Mellon University, Pittsburgh, PA *Expected: May 2027*
PhD in Language and Information Technologies (GPA: 4.0)

Carnegie Mellon University, Pittsburgh, PA *May 2023*
M.S. of Language Technologies (GPA 3.99)
Coursework: Introduction to Machine Learning (PhD), Advanced Natural Language Processing, Human-centered NLP, Prototyping Algorithmic Experiences

Columbia University, New York, NY *May 2015*
B.A. in Computer Science, Concentration in Physics (GPA 3.80)
Honors: *Cum Laude*, Dean's List (all semesters). Coursework: Natural Language Processing, Programming Languages and Translators, Advanced Sanskrit

PUBLICATIONS

Tom Corringham, Nupoor Gandhi, Bryan Flores, Emma Strubell, **Sireesh Gururaja**, Jacob Dunafon, Tristan Romanov. Extracting Structured Policy Information from Climate Action Plans. **Dataset proposal at the Tacking Climate Change with Machine Learning Workshop at NeurIPS 2025.**

Lucy Suchman, **Sireesh Gururaja**, David Gray Widder. AI interdisciplinarity as critical technical practice. **Cambridge Forum on AI: Culture and Society**, vol. 1, p. e2, 2025. doi:10.1017/cfc.2025.10004

Sireesh Gururaja, Nupoor Gandhi, Jeremiah Milbauer, Emma Strubell. Beyond Text: Domain Expert Needs in Document Research. **Findings of ACL 2025.**

Sireesh Gururaja*, Yueheng Zhang*, Guannan Tang, Tianhao Zhang, Kevin Murphy, Yu-Tsen Yi, Junwon Seo, Anthony Rollett, Emma Strubell. Collage: Decomposable Rapid Prototyping for Information Extraction on Scientific PDFs. **Scholarly Document Processing Workshop at ACL 2025.**

Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, **Sireesh Gururaja**, Tzu-Sheng Kuo, Jenny T. Liang, Ryan Liu, Ihita Mandal, Jeremiah Milbauer, Xiaolin Ni, Namrata Padmanabhan, Subhashini Ramkumar, Alexis Sudjianto, Jordan Taylor, Ying-Jui Tseng, Patricia Vaidos, Zhijin Wu, Wei Wu, Chenyang Yang. LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs **CHI 2025 Case Study.**

Sireesh Gururaja*, Amanda Bertsch*, Clara Na*, David Gray Widder, and Emma Strubell. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing. **EMNLP 2023.**

Sireesh Gururaja, Ritam Dutt, Tinglong Liao, and Carolyn Rosé. Linguistic representations for fewer-shot relation extraction across domains.. **ACL 2023.**

Srijan Bansal, Suraj Tripathi, Sumit Agarwal, **Sireesh Gururaja**, Aditya Srikanth Veerubhotla, Ritam Dutt, Teruko Mitamura, and Eric Nyberg. 2022. R3 : Refined Retriever-Reader pipeline for Multidoc2dial. **DialDoc Workshop at ACL 2022.**

WORKING PAPERS

David Gray Widder, **Sireesh Gururaja**, Lucy Suchman. Basic Research, Lethal Effects: Military AI Research Funding as Enlistment. Under Review.

Sireesh Gururaja, Junwon Seo, Hung-Yi Lin, Jeremiah Milbauer, Anthony Rollett, Emma Strubell. Data-driven Design as a High-Impact, Ecologically Valid Benchmark for Document Understanding. **Dataset proposal at AI4Science Workshop at Neurips 2025, non-archivally presented at the NAACL Workshop on AI and Scientific Discovery**. Working Paper.

PATENTS

Szanto, A., **Gururaja, S.**, & Puzio, D. (2020). *Dynamically updated text classifier* U.S. Patent No. 11,586,987.

Ramsey, M., Parlato-Altay, G., Szanto, A., Liu, A., **Gururaja, S.**, Hsu, B., & Puzio, D. (2022). *Named entity recognition and disambiguation engine* U.S. Patent No. 11,366,966. Washington, DC: U.S. Patent and Trademark Office.

INVITED TALKS

Basic Research, Lethal Effects: Military AI Research Funding as Enlistment. Talk with David Widder at the Emergent Nonfiction Lab at the University of Salford. March 25, 2025.

PROFESSIONAL EXPERIENCE

Ikigai Labs Cambridge, MA
ML/UX Intern *May 2025 - August 2025*

- Designed and implemented user-controllable smart alerts for supply chain management, allowing users to iteratively set and refine thresholds to manage notification overload.

Kensho Technologies Cambridge, MA
Team Lead, ML Ops and Internal Tools *May 2020 - May 2021*

- Managed an infrastructure team of six, overseeing the **ML model lifecycle** and development tooling for a company of120 people.
- Architected and built ML Ops tooling to catalog datasets, ensure **reproducible experimentation and training**, and **monitor and retrain** models using a combination of open-source and vendor technology, simplifying ML Ops and saving weeks of ML developer time per quarter, and increasing artifact reliability.

Machine Learning Engineer *January 2018 - May 2020*

- Led a team of four engineers, researching and improving Kensho's **named entity disambiguation** models with **graph-based features** to improve out-of-domain generalization and improving F1 scores 4 points on the client benchmarks.
- Built robust pipelines for **rapidly creating text classifiers**, using knowledge graph-based sampling approaches and active learning to accelerate model creation.
- Proposed and **ran an audit** of Kensho's transcription offering, Scribe, focusing on gender and accent bias originating in deep imbalance in the training data.

IBM Watson New York, NY
Software Engineer/Staff Software Engineer *August 2015 - December 2017*

- Led a machine learning squad on a corporate tax project, building **text classification pipelines** backed by knowledge graph verification.
- Designed and implemented a **continuous integration** approach for those machine learning pipelines, allowing for transparent tracking of results and accountability in model and feature choices.

The Mathworks, Inc. Natick, MA
Software Engineering Intern *May 2014 - August 2014*

- Designed and built a prototype for passing model-level information to a backend verification engine in Simulink.
- Collaborated with the MATLAB Graphics team to create an algorithm for smart text placement.

RESEARCH EXPERIENCE

CMU Language Technologies Institute Pittsburgh, PA
August 2021-Present *Graduate Research Assistant*

- Leading a collaboration between researchers at the LTI and the CMU Materials Science and Engineering department to study multimodal information extraction for data-driven design (DDD) for novel carbide materials. Informed DDD work with interviews of experts in multiple fields.
- Studying the comparative effectiveness of methods for on-device information extraction, and how to effectively evaluate these systems with minimal annotated data.

- Building a prototype system to enable user schema-driven information extraction on PDF documents in the browser.
- Studied the conception of AI in the U.S. Department of Defense through an analysis of over 7,000 grant solicitations. Maintained technical infrastructure supporting the project, including PDF parsing, indexing and search.
- Studied trends in NLP, focusing on both incentives and funding using both quantitative and qualitative methods.
- Evaluated the extent to which using structured objectives during pretraining affects language models' robustness.
- Investigated the feasibility of using linguistic representations to improve few-shot transfer of relation extraction in procedural text.

Columbia Astrophysics Laboratory New York, NY
Volunteer Research Assistant *September 2019 - August 2021*

- Developed a resnet approach to sex tsetse fly pupae in order to accelerate sterile insect technique-based population control that achieved an AUC of 0.97. Additionally verified the learned features and results by studying LIME explanations over a held-out corpus.

Columbia University MESAAS Department New York, NY
Research Assistant *January 2014 - September 2014*

- Reviewed and formatted the digitization of Sanskrit texts, and annotated structural divisions.
- Used regular expressions and Python to automate pre-processing tasks on Sanskrit texts.

TEACHING EXPERIENCE

Carnegie Mellon Language Technologies Institue Pittsburgh, PA
Teaching Assistant - Introduction to Language Technologies Research *August 2025 - December 2025*

- Participated in course design.
- Managed course logistics.

Carnegie Mellon Language Technologies Institue Pittsburgh, PA
Teaching Assistant - Natural Language Processing *January 2024 - May 2024*

- Participated in assignment and exam design.
- Designed and delivered guest lecture: How do we think of progress in NLP.
- Held office hours once weekly.

Columbia University Computer Science Department New York, NY
Teaching Assistant - Introductory Computer Science *January 2014 - December 2014*

- Held office hours and review sessions to help students with basic concepts in computer science and Java.
- Graded students' problem sets and provided feedback for a class of over 300 students.

SERVICE

Academic	
Reviewing	ACL ARR (2024, 2025, all sessions), TMLR, COLM (2024, 2025)
Comittee Assignments	CMU SCS Doctoral Advising Committee (Fall 2024 - Present), LTI Education Committee (AY 2022-2023), LTI Student Admissions Reviewer (AY 2022-2023), LTI SysAdmin Job Search (Summer 2022), LTI Branding Committee (Fall 2022)
Mentorship	LTI Peer Mentoring Program (2022-2024), CMU Undergraduate AI Mentoring Program (AY 2023-2024)
Professional	
DEI	Kensho D&I Committee (3/18 - 6/20):
CSR	Kensho CSR Outreach program (3/20 - 5/21)
Mentorship	Industry mentor, Yonkers Partners in Education (12/20 - 5/21)
Other	
Volunteering	Code for Boston (7/19 -4/20), Garfield Communty Farm (6/21 - present)

SKILLS

Programming Languages	Python, Java(Intermediate), Typescript(Intermediate)
Tools	pyTorch, NumPy, pandas, React, Docker, k8s, Git, Bash, LaTeX, SQL
Protocols	GRPC, REST, GraphQL, AMQP, jsonnet
Language	Intermediate French; Conversational Hindi, Kannada, and Sanskrit
Hobbies	Wok Seasoning, Fly Fishing, Knitting, and Coffee