# Rajarshi Haldar

PhD Candidate · Computer Science · University of Illinois Urbana-Champaign

Champaign, IL • rajarshihal@gmail.com • Google Scholar • rajarshihaldar • rhaldar • rajarshihaldar.com

## RESEARCH INTERESTS

Evaluating Long-Form Text · Interpreting Evaluators · Code Intelligence · Structure & Knowledge in NLP

## EDUCATION

**University of Illinois Urbana-Champaign**, Urbana, IL — Aug 2018 — Present

*Ph.D. in Computer Science* · Advisor: Julia Hockenmaier · Natural Language Processing

**Indian Institute of Technology Kharagpur**, Kharagpur, India — Jul 2014 — May 2018

*B.Tech. in Computer Science and Engineering*

- Undergraduate research on knowledge graphs over the ACL Anthology (NAACL-HLT 2018 demo) and on venue recommendation for event-based social networks (IEEE TKDE).

## RESEARCH EXPERIENCE

**University of Illinois Urbana-Champaign**, Urbana, IL — Aug 2018 — Present

*PhD Candidate & Graduate Research Assistant* · Advised by Prof. Julia Hockenmaier

- Used sparse autoencoders on LLM-judge activations; probes beat the judges' own scores on 7 of 8 criteria (in review).
- Isolated the few dozen SAE features that track expert-annotated summary quality in Gemma 3 and Qwen 3.5 — under 1% of each dictionary — and showed ordinal probes on them recover positive agreement with experts on fluency, where the judges' own decoded scores fall below chance.
- Showed LLM judges rescore identical inputs inconsistently, with Krippendorff's alpha as low as 0.27 (EMNLP 2025).
- Traced LLM code-summarization scores to token overlap in function names, not code structure (LREC-COLING 2024).
- Introduced a multi-perspective code-search architecture, lifting MRR 28% over a code-token baseline (ACL 2020).
- Measured intra-rater reliability across three reruns on SummaC, SummEval, and MT-Bench: Llama 3.1 reached a Krippendorff's alpha of 0.265 and DeepSeek-R1 0.507, with Qwen 3 the most consistent but still short of standard reliability thresholds.
- Built the dataset, training, and evaluation pipelines in PyTorch and Hugging Face Transformers.
- Designed a system to auto-grade code-reading assignments, collaborating with four researchers to devise a validated scoring rubric and to produce a dataset of code snippets, student responses, and quality scores (SIGCSE 2020).

**Schlumberger**, Menlo Park, CA — Summer 2022

*Digital Intern*

- Built a knowledge-graph embedding model that learns from limited labeled data via labeling functions and Snorkel AI.
- Shipped a query-answering model on those embeddings that replaced the team's system at lower latency and overhead.

**IBM Research**, Yorktown Heights, NY — Summers 2019 & 2020

*Graduate Research Intern*

- Led an automated tech-support agent that resolves IT queries by extracting events from unstructured incident text.
- Induced an event schema over 802K technical documents, raising retrieval MRR from 0.468 to 0.538 over a BERT baseline.
- Derived the schema with AMR parses, variational graph auto-encoders, and clustering under minimal supervision; the resulting retrieval system also beat parse-based and AMR-based baselines, reaching Recall@1 of 0.425 against 0.343 for BERT.

## TEACHING

**University of Illinois Urbana-Champaign**, Urbana, IL — 2022 — 2025

*Teaching Assistant, CS 447 — Natural Language Processing*

- Designed a Google Colab coding assignment on machine translation and wrote its auto-grader on Gradescope.
- Held office hours twice a week and answered student questions on Campuswire.

**University of Illinois Urbana-Champaign**, Urbana, IL — Summers 2024 & 2025

*Teaching Assistant, CS 416 — Data Visualization*

- Graded a midterm and a final project in which students built dashboards in Tableau and D3.js.
- Held weekly office hours and answered student questions on Campuswire.

## PUBLICATIONS

[7] **Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks** — Findings of EMNLP 2025
**Rajarshi Haldar**, Julia Hockenmaier

[6] **Analyzing the Performance of Large Language Models on Code Summarization** — LREC-COLING 2024
**Rajarshi Haldar**, Julia Hockenmaier

[5] **A Multi-Perspective Architecture for Semantic Code Search** — ACL 2020
**Rajarshi Haldar**, Lingfei Wu, JinJun Xiong, Julia Hockenmaier

[4] **Deep Learning Driven Venue Recommender for Event-Based Social Networks** — IEEE TKDE 2020
Soumajit Pramanik, **Rajarshi Haldar**, Anand Kumar, Sayan Pathak, Bivas Mitra

[3] **A Validated Scoring Rubric for Explain-in-Plain-English Questions** — SIGCSE 2020
Binglin Chen, Sushmita Azad, **Rajarshi Haldar**, Matthew West, Craig Zilles

[2] **Cluster Aware Mobility Encounter Dataset Enlargement** — IWCMC 2019
**Rajarshi Haldar**, Salih Safa Bacanlı, Moayad Aloqaily, Adel Ben Mnaouer, Damla Turgut

[1] **CL Scholar: The ACL Anthology Knowledge Graph Miner** — NAACL-HLT Demo 2018
Mayank Singh, Pradeep Dogga, Sohan Patro, Dhiraj Barnwal, Ritam Dutt, **Rajarshi Haldar**, Pawan Goyal, Animesh Mukherjee

## PATENTS

**Unsupervised Event Extraction** — US 12,001,896 B2
**Rajarshi Haldar**, Yu Deng, Lingfei Wu, Ruchi Mahindru, Shu Tao · IBM · Granted Jun 2024

**Graph-Based Event Schema Induction for Information Retrieval** — US 11,615,152 B2
**Rajarshi Haldar**, Yu Deng, Lingfei Wu, Ruchi Mahindru, Julia Hockenmaier, Sinem Guven Kaya · IBM · Granted Mar 2023

**Multi-Perspective, Multi-Task Neural Network Model for Matching Text to Program Code** US 11,132,512 B2

Lingfei Wu, JinJun Xiong, Julia Hockenmaier, **Rajarshi Haldar** · IBM · Granted Sep 2021

## SKILLS

**Languages & Tools:** Python, SQL, C, Git, Linux, LaTeX, Typst
**ML & NLP:** PyTorch, Hugging Face Transformers, TensorFlow, scikit-learn
**Research:** Large Language Models, LLM Evaluation, Interpretability, Code Intelligence, Information Retrieval

## SERVICE

- Reviewer, Annual Meeting of the Association for Computational Linguistics (ACL)
- Reviewer, Conference on Empirical Methods in Natural Language Processing (EMNLP)
- Reviewer, North American Chapter of the ACL (NAACL)