Ghazal Khalighinejad

gk126@duke.edu | Google Scholar | GitHub | ghazalkhalighinejad.github.io

Education

Duke University September 2021 - Present
Ph.D. Student in Computer Science; GPA: 3.92/4
Advisors: Bhuwan Dhingra, Sam Wiseman

Sharif University of Technology September 2017 - May 2021
Bachelor of Science in Computer Science

Research Interest

My research focuses on Natural Language Processing and Multimodal Large Language Models, with an emphasis on improving embedding methods for multimodal retrieval and developing post-training techniques that enhance scientific reasoning, particularly in areas such as AI for code and scientific discovery.

Positions

Google Research *Student Researcher* May 2025 – November 2025
Developed a paper-to-code retrieval system by fine-tuning a code-embedding model. Created a benchmark for evaluating ML concept–to–code mapping.

The Simons Foundation & Polymathic AI *Guest Researcher* March 2025 – May 2025
Worked on a multimodal foundation model for astronomical-observation retrieval and designed post-training methods to improve late-interaction image-to-image retrieval.

Adobe Research *Research Intern* May 2024 – August 2024
Built a text-to-video retrieval system that identifies relevant shots and assembles them into coherent sequences using a dynamic programming algorithm to select and order shots.

Publications

1. MatViX: Multimodal Information Extraction from Visually Rich Articles. Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, Bhuwan Dhingra. ***NAACL* 2025,** Oral Presentation.

2. It's LIT! LLMs with Interpretable Tool Calling. Ruixin Zhang, Jon Donnelly, Zhicheng Guo, Ghazal Khalighinejad, Haiyang Huang, Alina Jade Barnett, Cynthia Rudin. ***MTI-LLM @ NeurIPS* 2025**.

3. Training Neural Networks as Recognizers of Formal Languages. Alexandra Butoi, Ghazal Khalighinejad, Anej Svete, Josef Valvoda, Ryan Cotterell, Brian DuSell. ***ICLR* 2025.**

4. IsoBench: Benchmarking Multimodal Foundation Models on Isomorphic Representations. Deqing Fu*, Ruohao Guo*, Ghazal Khalighinejad*, Ollie Liu*, Bhuwan Dhingra, Dani Yogatama, Robin Jia, Willie Neiswanger (*Equal contribution). ***COLM* 2024.**

5. Extracting Polymer Nanocomposite Samples from Full-Length Documents. Ghazal Khalighinejad, Defne Circi, L.C. Brinson, Bhuwan Dhingra. ***Findings of ACL* 2024**.

6. How Well Do Large Language Models Understand Tables in Materials Science? Defne Circi, Ghazal Khalighinejad, Anlan Chen, Bhuwan Dhingra, L.C. Brinson. *Integrating Materials and Manufacturing Innovation (**IMMI 2024**).*

7. Galloping in FastGrowth Natural Merge Sorts. Elahe Ghasemi, Vincent Jugé, Ghazal Khalighinezhad, Helia Yazdanyar. ***Algorithmica* 2024** and ***ICALP* 2022**.

8. Approximating CKY with Transformers. Ghazal Khalighinejad, Ollie Liu, Sam Wiseman. ***Findings of EMNLP* 2023**.

9. Exploring the Effect of Frequency Resolution in FNet. Gregory Szumel, Ghazal Khalighinejad, Rickard Stureborg and Sam Wiseman. ***SustaiNLP @ ACL* 2023**.

10. Retrieval of Synthesis Parameters of Polymer Nanocomposites using LLMs. Defne Circi, Ghazal Khalighinejad, Shruti Badhwar, Bhuwan Dhingra, L. Brinson. ***AI4MAT @ NeurIPS* 2023**.

Preprints

1. Which View Works Best? Evaluating Representations for Scientific Document Retrieval. Ghazal Khalighinejad, Raghuveer Thirukovalluru, Bhuwan Dhingra. *In submission*, 2025.

Research Experience

Multimodal Large Language Models for Sciences Current

- Created a large dataset and evaluation method for extracting structured data from visually rich scientific articles using vision-language models.
- Proposed novel evaluation metrics for assessing model performance on tasks involving multimodal data extraction, with a focus on curve similarity and hierarchical structure alignment.

Algorithmic Reasoning in Transformers 2022-2023

- Trained transformers to approximate CKY parsing, replacing CKY in modern constituency parsers without accuracy loss and improving runtime from cubic to quadratic dependence on sentence length.

Sorting Algorithms 2021-2022

- Proved that several merge sorting algorithms are as efficient as TimSort when employing its galloping sub-routine.

Awards & Achievements

Oral Presentation NAACL 2025

LLMs for Materials Hackathon, Anthropic Award Won Anthropic API credits.

aiM National Science Foundation Fellow Full-tuition scholarship and funding for research in AI + Materials.

ACM-W Research Conference Scholarship Travel scholarship to attend NeurIPS 2022.

CRA-WP Scholarship Travel scholarship to attend CRA-WP Grad Cohort for Women Workshop.

Teaching Assistance Experience

Natural Language Processing (Graduate Course), Duke University, Fall 2022
Algorithms (Graduate Course), Duke University, Spring 2021
Game Theory (Graduate Course), Sharif University of Technology, Spring 2020
Advanced Programming (Undergraduate Course), Sharif University of Technology, Fall 2019

Doctoral Courseworks

Neurosymbolic Machine Learning, Natural Language Processing, Deep Learning, Causality and Interpretability, Probability and Statistics, Algorithms

Services

Reviewer CVPR 2026, COLM 2025, ACL Rolling Review 2024, 2025

Skills

Programming: Python, Java, C/C++

Libraries: PyTorch, JAX, TensorFlow