AAMOD KHATIWADA

📍 Mountain View, California ✉ khatiwada.a@northeastern.edu 🔗 https://aamodkh.github.io
in https://www.linkedin.com/in/aamod-khatiwada G https://scholar.google.com/citations?user=OnkSeCsAAAAJ

RESEARCH INTERESTS

LLM Optimization, Information Retrieval, Data Discovery and Integration, Tabular Representation Learning, Knowledge Graphs, Data lake Management, Search and Ads

EDUCATION

Northeastern University, Khoury College of Computer Sciences, Boston, MA, USA *Sep 2020 - Dec 2024*
PhD in Computer Science
Advised by: Dr. Renée J. Miller
Thesis Title: *Table Discovery and Integration in Data Lakes*
Research summary: *Develop novel tools, algorithms and foundation models to add semantics to the open data tables, and use such semantics for data management, table cleansing, table discovery and table integration.*

Northeastern University, Khoury College of Computer Sciences, Boston, MA, USA *Sep 2020 - Apr 2023*
Masters in Computer Science

Tribhuvan University, Institute of Engineering (IOE), Kathmandu, Nepal *Nov 2014 - Nov 2018*
Bachelor's Degree in Electronics and Communication Engineering
Distinction, Department Topper

SELECTED EXPERIENCE

Microsoft AI, Silicon Valley Campus, CA, USA *Feb 2025 - Present*
Data Scientist II (Microsoft Ads Team)

- Develop novel algorithms and leverage language models to enhance the effectiveness of recommending web search results to the users in Bing Search and Co-pilot.

Data lab, Northeastern University, Boston, MA, USA *Sept 2020 - Dec 2024*
Graduate Research Assistant

- Developed a principled way of using LLMs to generate benchmarks for tabular tasks.
- Led a group project entitled *SANTOS* on finding the unionable open data tables by detecting their semantic types.
- Developed novel techniques and algorithms to integrate open data tables and web tables in a principled way.
- Created open data benchmarks for data discovery tasks.

PATENTS

- **A. Khatiwada**, H. Kokel, G. Rossiello, U. Khurana, O. Hassanzadeh, D. Bhattacharjya, K. Srinivas, A.M. Gliozzo, and H. Samulowitz, *"Sensible Join Discovery Framework for Tabular Data Integration in Data Lakes"* (US Patent filed)
- O. Hassanzadeh, **A.Khatiwada** and S. Shirai, *"Link Prediction Using an Ensemble of Representations and Rules"* (under revision)

SELECTED PUBLICATIONS

1. **A. Khatiwada**, R. Shraga and R. J. Miller, *"Diverse Unionable Tuple Search"*, in EDBT 2026. (to appear)
2. **A. Khatiwada**, R. Shraga and R. J. Miller, *"Fuzzy Integration of Data Lake Tables "*, in EDBT 2026. (to appear)
3. H. Kokel, **A. Khatiwada**, T. Pedapati, H. Ananthakrishnan, O. Hassanzadeh, H. Samulowitz, K. Srinivas, *"TOPJoin: A Context-Aware Multi-Criteria Approach for Joinable Column Search"*, in TaDA@VLDB 2025. `https://arxiv.org/abs/2507.11505`
4. **A. Khatiwada**, Harsha Kokel, I. Abdelaziz, S. Chaudhury, J. Dolby, O. Hassanzadeh, Z. Huang, T. Pedapati, H. Samulowitz, K. Srinivas, *"TabSketchFM: Sketch-based Tabular Representation Learning for Data Discovery over Data Lakes"*, in ICDE 2025. `https://doi.ieeecomputersociety.org/10.1109/ICDE65448.2025.00118`

5. **A. Khatiwada**, Harsha Kokel, I. Abdelaziz, S. Chaudhury, J. Dolby, O. Hassanzadeh, Z. Huang, T. Pedapati, H. Samulowitz, K. Srinivas, *"TabSketchFM: Sketch-based Tabular Representation Learning for Data Discovery over Data Lakes"*, in TRL@NeurIPS 2024. `https://arxiv.org/abs/2407.01619` (Best Paper Runner-up Award)
6. D. C. Fox, **A. Khatiwada** and R. Shraga, *"Generative Benchmark Creation Framework for Detecting Common Data Table Versions"*, in CIKM 2024. `https://doi.org/10.1145/3627673.3679157`
7. K. Pal, **A. Khatiwada**, R. Shraga and R.J. Miller, *"ALT-GEN: Benchmarking Table Union Search using Large Language Models"*, in TaDA@VLDB 2024. `https://arxiv.org/abs/2308.03883` (Best Paper Award)
8. **A. Khatiwada**, R. Shraga and R. J. Miller, *"DIALITE: Discover, Align and Integrate Open Data Tables"*, in SIGMOD-Companion, 2023. `https://doi.org/10.1145/3555041.3589732`
9. **A. Khatiwada**, G. Fan, R. Shraga, Z. Chen, W. Gatterbauer, R. J. Miller and M. Riedewald, *"SANTOS: Relationship-based Semantic Table Union Search"*, in Proc. ACM Manag. Data (PACMMOD) 1, 1, Article 9 (May 2023), 2023. `https://doi.org/10.1145/3588689`
10. **A. Khatiwada**, R. Shraga, W. Gatterbauer and R.J. Miller, *"Integrating Data Lake Tables"*, in PVLDB, 16(4):932-945, 2022. `https://doi.org/10.14778/3574245.3574274`
11. **A.Khatiwada**, S. Shirai, K. Srinivas and O. Hassanzadeh, *"Knowledge Graph Embeddings for Causal Relation Prediction"*, in Deep Learning for Knowledge Graphs Workshop (DL4KG@ISWC), 2022. `https://ceur-ws.org/Vol-3342/paper-8.pdf`
12. S. Shirai, **A.Khatiwada**, D. Bhattacharjya and O. Hassanzadeh, *"Rule-Based Link Prediction over Event-Related Causal Knowledge in Wikidata"*, in 3rd Wikidata Workshop (Wikidata@ISWC), 2022. `https://ceur-ws.org/Vol-3262/paper14.pdf`
13. O. Hassanzadeh, P. Awasthy, K. Barker, O. Bhardwaj, D. Bhattacharjya, M. Feblowitz, **A. Khatiwada**, L. Martie, S. F. Mbouadeu, J. Ni, A. Saha, S. Shirai, K. Srinivas and L. Yip, *"Knowledge-Based News Event Analysis Toolkit"*, in ISWC, 2022. `https://ceur-ws.org/Vol-3254/paper399.pdf`
14. **A. Khatiwada**, P. Kadariya, S. Agrahari and R. Dhakal, *"Big Data and Deep Learning Based Sentiment Analysis System for Sales Prediction"*, in IEEE International Conference on Innovating Technology for Humanity, Pune, 2019. pp. 1-6. `https://doi.org/10.1109/PuneCon46936.2019.9105719`

PRIOR EXPERIENCES

Microsoft AI, Redmond, WA, USA *June 2024 - Aug 2024*
Data Science Intern (Microsoft Ads Team)
Mentors: Deepak Saini, Qiang Lou and Jian Jao

- Developed an optimization algorithm that reduces LLM inference latency by over 29 % with negligible accuracy drop in relevant Ads recommendation task.
- Trained Language Model to understand the semantics of online web search queries and recommend relevant Ads to the users.

IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA *May 2023 - Aug 2023*
AI Research Scientist Intern
Mentors: Dr. Udayan Khurrana, Dr. Kavitha Srinivas and Dr. Oktie Hassanzadeh

- Led a project on building a large language model for data management tasks.
- Developed systematic ways of creating benchmarks to fine tune large language models for data discovery tasks.
- Developed TOPJoin algorithm for searching joinable tables from the data lake which is a part of IBM's Watsonx.data product (`https://www.ibm.com/products/watsonx-data`).

Northeastern University, Boston, MA, USA *Sept 2022 - Dec 2022*
Teaching Assistant (Khoury College of Computer Sciences)

- Contributed to the development of course materials and course projects for a graduate-level course *CS7290: Special Topics in Data Science* (`https://northeastern-datalab.github.io/cs7290.f22/`)
- Taught three lectures covering state-of-the-art systems for data discovery and management.

IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA *May 2022 - Aug 2022*
Exploratory Science Research Intern
Mentors: Dr. Oktie Hassanzadeh and Dr. Dharmashankar Subramanian

- Led a project on detecting the causal relations in Knowledge graphs using embeddings and GNN-based models.

Tribhuvan University, Kathmandu, Nepal *Nov 2018 - Jan 2020*
Teaching Assistant (Department of Electronics and Computer Engineering)

- Carried out lab classes on C programming, Digital Logic and Big Data Analytics.
- Assisted final year students to design and debug their major projects.

PRE-PRINT

1. K. Srinivas, J. Dolby, I. Abdelaziz, O. Hassanzadeh, H. Kokel, **A. Khatiwada**, T. Pedapati, S. Chaudhury, H. Samulowitz, *"LakeBench: Benchmarks for Data Discovery over Data Lakes"*. 2023. `https://arxiv.org/abs/2307.04217`

SELECTED TALKS

- **A. Khatiwada**, *"Session on Dataset Discovery in the era of Large Language Models"*, in Deeplearning.AI Lecture Series. Mar 2025.
- **A. Khatiwada**, *"Guest Lecture on Table Discovery and Integration in Data Lakes"*, in Boston University, Boston, MA, USA. Mar 2024. `https://bu-disc.github.io/CS561/slides/CAS-CS561-Class14.pdf`
- **A. Khatiwada**, R. Shraga, W. Gatterbauer and R.J. Miller, *"Invited Video Showcase of Integrating Data Lake Tables"*, in TaDA@VLDB 2023. `https://www.youtube.com/watch?v=4c6SYCwQ7uc`
- **A.Khatiwada** and G. Fan, *"Table Discovery and Integration in Data Lakes: Challenges and Solutions "*, in Northeast Database Day. Mar 2023. `https://northeastern-datalab.github.io/nedbday/2023/`

AWARDS AND EXTRA CURRICULAR

- Best Paper Runner-up Award in Tabular Representation Learning@NeurIPS 2024
- Best Paper Award in TaDA@VLDB 2024
- SIGMOD Student Travel Award from ACM SIGMOD. (Apr 2023)
- PhD Startup Fund from Khoury College of Computer Sciences, Northeastern University, MA, USA. (Sept 2020)
- Leadership And Mentorship Program (LAMP) Scholarship from the Dean of the Whitacre College of Engineering, Texas Tech University, TX, USA. (Jan 2020)
- Awarded as **the best undergraduate student** at Electronics and Computer Department, Tribhuvan University, Nepal for three consecutive years (Sophomore, Junior and Senior). (2016, 2017, 2018)

PROFESSIONAL SERVICE

- **PC Member:** KDD Research Track 2025, KDD ADS Track 2025, TaDA@VLDB 2025 , HILDA@SIGMOD 2025, HILDA@SIGMOD 2024, TaDA@VLDB 2024 , SIGMOD Availability and Reproducibility 2023, SemTab Challenge@ISWC 2023
- **Reviewer:** SIGIR Conference 2024, SIGIR Conference 2023
- **Student Volunteer**: SIGMOD Conference 2023
- Member at Computer Science PhD Admission Committee, Northeastern University (2021 and 2024)

SKILLS

- **Programming:** Python, C, C++, PHP, SQL, SPARQL, JavaScript, Java, Solidity, Assembly Programming
- **Tools and frameworks:** Pytorch, Knowledge Graphs, Large language models, Hadoop MapReduce, Spark, Laravel, jQuery, D3, Tensorflow, LaTeX

GRADUATE COURSEWORK

Information Visualization (project: *https://aamodkh.github.io/theta-join-visualization*), Distributed Systems (project: *https://github.com/aamodkh/distributed-datalake-tapestry*), Principle of Scalable Database Management, Advanced Algorithm, Large Scale and Parallel Data Processing, Special Topics in Database Management