Nandan Thakur

200 University Ave W, Waterloo, ON, N2L 3G1

nandan.thakur@uwaterloo.ca

thakur-nandan.github.io

Research Interests

I develop real-world evaluation benchmarks and curated training datasets to advance search and retrieval-augmented generation (RAG) systems. My research focuses on enabling retrieval models to accurately retrieve and reason over information in niche domains, complex problem settings, and diverse languages.

Education

Univesity of Waterloo — September 2021 – Present
Ph.D. in Computer Science — Ontario, Canada
Advised by *Prof. Jimmy Lin*
Heterogeneous Benchmarking of IR Systems Across Domains and Languages

Birla Institute of Technology & Science (BITS) Pilani — July 2014 – July 2018
B.E. (Hons.) in Electronics & Instrumentation, Minor in Finance — Goa, India

Past Employment

Databricks & Mosaic Research — August – December 2024
Research Intern under Omar Khattab and Prof. Michael Carbin — San Francisco, CA
Automatic framework for RAG benchmark construction [1].

Vectara — February – July 2024
Research Intern under Amin Ahmad — Palo Alto, CA
Multilingual RAG benchmarking [7] & LLM hallucinations [9].

Google Research — September 2022 – May 2023
Student Researcher under Daniel Cer and Jianmo Ni — Mountain View, CA
Synthetic construction of multilingual retrieval datasets using LLMs [10].

UKP Lab, Technical University of Darmstadt — November 2019 – August 2021
Research Assistant under Nils Reimers and Prof. Iryna Gurevych — Darmstadt, Germany
Zero-shot IR benchmarking [17] & data augmentation [18] [16].

KNOLSKAPE — September 2018 – October 2019
Data Scientist — Bangalore, India
Constructed Krawler.ai, an enterprise multimodal search product.

(EMBL) European Molecular Biology Laboratory — June – August 2018
Research Trainee under Manjeet Kumar and Prof. Toby Gibson — Heidelberg, Germany
ML Prediction toolkit to predict phosphorylation sites within protein sequences.

Belong.co — July – December 2017
Data Science Intern under Vinodh K. Ravindranath — Bangalore, India
Semi-supervised topic modeling and keyword extraction with GuidedLDA.

Selected Awards & Grants

David R. Cheriton Graduate Scholarship of $20,000 for two academic years	2024
Snowflake AI Research & University of Waterloo Collaborative Grant	2024
Huawei Technologies & University of Waterloo Collaborative Grant	2022
Got Selected as a Speaker for PyCon Italia in 2020 (Cancelled due to Covid-19)	2020
Received a fully-funded ML fellowship to work at EMBL Heidelberg	2018

Invited Talks

IISc Bangalore *Beyond Models: Rethinking Benchmarks, Data, and Evaluation for RAG*	October 2025 Bangalore, India
MILA, Hamel Husain's Course & Weaviate Podcast *Modern IR Evaluation for RAG* (Over 4000+ signups)	September 2025 Virtual
Microsoft Research India *Accelerating Multilingual RAG Systems*	January 2025 Virtual
IIT Delhi & IIIT Delhi *Heterogenous IR Benchmarking across Domains and Languages*	January 2024 Delhi, India
Koç University *A Tutorial on Advanced Information Retrieval*	July 2023 Virtual
Stanford University *Heterogenous Benchmarking in IR Research*	November 2022 Palo Alto, CA
OpenNLP Meetup, Deepset.ai *BEIR, An Open-Source Benchmark for IR Systems*	August 2021 Virtual

Academic Service

- Lead Organizer of the TREC's inaugural RAG track: TREC-RAG 2024 & 2025.
- Lead Organizer on multilingual IR competition task (MIRACL) at WSDM Cup 2023.
- **Reviewer (NLP conferences):** ACL ARR Oct-Nov (2021), Jan-Apr (2022), 2024–Present
- **Reviewer (ML & IR conferences):** NeurIPS 2023, SIGIR 2023 & 2025, ECIR 2024 & 2026, ICLR 2026.

Peer-Reviewed Conference Publications

Check my Google Scholar (https://scholar.google.com/citations?user=CE9GJoMAAAAJ) for all of my publications. NeurIPS, ACL, EMNLP, NAACL, TACL, SIGIR etc. are top-tier peer-reviewed conferences in NLP/IR with an acceptance rate between 20-30%. (* *denotes equal contributions.*)

[1] FreshStack: Building Realistic Benchmarks for Evaluating Retrieval on Technical Documents. **N. Thakur**, J. Lin, S. Havens, M. Carbin, O. Khattab, A. Drozdov.
Conference on Neural Information Processing Systems: Datasets and Benchmarks Track (NeurIPS D&B Track). 2025.

[2] Hard Negatives, Hard Lessons: Revisiting Training Data Quality for Robust Information Retrieval with LLMs. **N. Thakur***, X. Zhang*, X. Ma, J. Lin.
Findings of the Association for Computational Linguistics (EMNLP Findings). 2025.

[3] Assessing Support for the TREC 2024 RAG Track: A Large-Scale Comparative Study of LLM and Human Evaluations. **N. Thakur**, R. Pradeep, S. Upadhyay, D. Campos, N. Craswell, J. Lin.
Conference on Research and Development in Information Retrieval (SIGIR). 2025.

[4] The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. R. Pradeep, **N. Thakur**, S. Upadhyay, D. Campos, N. Craswell, J. Lin.
Conference on Research and Development in Information Retrieval (SIGIR). 2025.

[5] A Large-Scale Study of Relevance Assessments with Large Language Models Using UMBRELA. S. Upadhyay, R. Pradeep, **N. Thakur**, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, J. Lin.
Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). 2025.

[6] Ragnarök: A Reusable Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. R. Pradeep*, **N. Thakur***, S. Sharifymoghaddam, E. Zhang, R. Nguyen, D. Campos, N. Craswell, J. Lin.
Findings of the European Conference on Information Retrieval (*ECIR Findings*). *2025.*

[7] MIRAGE-Bench: Automatic Multilingual Benchmark Arena for Retrieval-Augmented Generation Systems. **N. Thakur**, S. Kazi, G. Luo, J. Lin, A. Ahmad.
Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (*NAACL*). *2025.*

[8] MMTEB: Massive Multilingual Text Embedding Benchmark.
K. Enevoldsen, I. Chung, ..., **N. Thakur**, ... (80 additional authors).
International Conference on Learning Representations (*ICLR*). *2025.*

[9] "Knowing When You Don't Know": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation. **N. Thakur**, L. Bonifacio, X. Zhang, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, B. Chen, M. Rezagholizadeh, J. Lin.
Findings of the Association for Computational Linguistics (*EMNLP Findings*). *2024.*

[10] Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval. **N. Thakur**, J. Ni, G. Hernández Ábrego, J. F. Wieting, J. Lin, D. Cer.
Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (*NAACL*). *2024.*

[11] Systematic Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR. **N. Thakur**, L. Bonifacio, M. Fröbe, A. Bondarenko, E. Kamalloo, M. Potthast, M. Hagen, J. Lin.
Conference on Research and Development in Information Retrieval (*SIGIR Resource & Reproducibility Track*). *2024.* Oral Presentation.

[12] Resources for Brewing BEIR: Reproducible Reference Models and Statistical Analyses.
E. Kamalloo, **N. Thakur**, C. Lassance, X. Ma, J. H. Yang, J. Lin.
Conference on Research and Development in Information Retrieval (*SIGIR Resource & Reproducibility Track*). *2024.* Oral Presentation.

[13] MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages.
X. Zhang*, **N. Thakur***, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, J. Lin.
Transactions of the Association for Computational Linguistics (*TACL*). *2023*

[14] SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval. **N. Thakur**, K. Wang, I. Gurevych, J. Lin.
Conference on Research and Development in Information Retrieval (*SIGIR Resource & Reproducibility Track*). *2023.*

[15] Evaluating Embedding APIs for Information Retrieval
E. Kamalloo, X. Zhang, O. Ogundepo, **N. Thakur**, D. Alfonso-Hermelo, M. Rezagholizadeh, J. Lin.
Annual Conference of the Association for Computational Linguistics (*ACL Industry Track*). *2023.*

[16] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval K. Wang, **N. Thakur**, N. Reimers, I. Gurevych.
Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (*NAACL*). *2022.*

[17] BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. **N. Thakur**, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych.
Conference on Neural Information Processing Systems: Datasets and Benchmarks Track (*NeurIPS D&B Track*). *2021.* Widely used benchmark by OpenAI, Google, IBM, Microsoft, Elastic Search, Cohere and many others. Over 2K GitHub stars & 1.5K citations.

[18] Augmented SBERT: Data Augmentation for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks. **N. Thakur**, N. Reimers, J. Daxenberger, I. Gurevych.
Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (*NAACL*). *2021.*

Workshop Publications

[19] BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. Z. Chen, X. Ma, ..., **N. Thakur**, ..., J. Lin.
Workshop on Multi-Turn Interactions in Large Language Models (*MTI-LLM*) *@ NeurIPS 2025.* Spotlight Presentation.

[20] Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval. **N. Thakur**, N. Reimers, J. Lin.
Workshop on Reaching Efficiency in Neural Information Retrieval (*ReNeuIR*) *@ SIGIR 2023.* Oral Presentation.

ArXiv Preprints

[21] Chatbot Arena Meets Nuggets: Towards Explanations & Diagnostics in the Evaluation of LLM Responses S. Sharifymoghaddam*, S. Upadhyay*, **N. Thakur***, R. Pradeep, J. Lin. 2025 (under review at ICLR 2026)

[22] UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor S. Upadhyay, R. Pradeep, **N. Thakur**, N. Craswell, J. Lin. 2024.

[23] A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution. E. Kamalloo, A. Jafari, X. Zhang, **N. Thakur**, J. Lin. 2023.

[24] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval. J. Lin, D. Alfonso-Hermelo, V. Jeronymo, E. Kamalloo, C. Lassance, R. Nogueira, O. Ogundepo, M. Rezagholizadeh, **N. Thakur**, X. Zhang. 2023.

Teaching Experience

Head TA at University of Waterloo

• CS 370 Numerical Computation	Winter, Spring & Fall 2025, Fall 2023, Spring 2024
• CS 486/686 Introduction to Artificial Intelligence	Winter 2023
• CS 116 Introduction to Computer Science 2	Winter 2024
• CS 136 Elementary Algorithm Design	Spring 2023, Winter 2022
• CS 241 Foundations of Sequential Programs	Spring 2022
• CS 135 Designing Functional Programs	Fall 2021

Advising & Mentoring

Current M.S. Students:
- Linsen Gao (open source reproduction of [10])

Current Undergraduate Students:
- Suraj Subrahmanyan (extension of [1])
- Nathan Kuissi (extension of [1])
- Jonathan Zhao (extension of [16])

Press & Media

2025: FreshStack has been included in the new RTEB benchmark — *Hugging Face*
2025: Modern RAG evaluation is part of the Hamel Husain's RAG mini book — *Maven*
2025: Invited Guest Speaker (#124) at Weaviate Podcast — *Weaviate*
2025: TREC RAG research has been incorporated in Vectara's product — *Vectara*
2023: Domain Adaptation with Generative Pseudo-Labeling (GPL) — *Pinecone.ai*
2022: Extending Neural Retrieval Models to New Domains and Languages — *Zeta Alpha*
2022: BEIR benchmark in Stanford's CS224U Teaching Material — *Stanford University*
2022: BEIR benchmark as a helpful ML library — *ML News by Yannic Kilcher*
2021: Making the Most of Data: Augmentation with BERT — *Pinecone.ai*

Coursework

University of Waterloo (2021–Present): CS 680: Introduction to Machine Learning, CS 889: Data Sources for Emerging Technology, CS 886: Graph Neural Networks, CS 886: Robustness of Machine Learning, CS 848: Information Retrieval, CS 679: Neural Networks, CS 854: Experimental Performance Evaluation, CS 649: Human-Computer Interaction.

BITS Pilani (2014–2018): Machine Learning, Neural Networks & Fuzzy Logic, Data Structures & Algorithms, Probability & Statistics, Linear Algebra, Econometric Methods, Discrete Mathematics.

Co-Curricular

Mime Club Coordinator, BITS Pilani — 2016–2017
Led a team of 30 student performers. Involved in acting, sound mixing, designing slides and creating stories for more than 10 shows over a span of 4 years. (YouTube)

Referees

Prof. Jimmy Lin, Full Professor, University of Waterloo
Prof. Iryna Gurevych, Full Professor, TU Darmstadt; Adjunct Professor, MBZUAI
Omar Khattab, Assistant Professor, MIT
Daniel Cer, Senior Research Scientist, Google Research
Nick Craswell, Principal Architect, Microsoft