Nandan Thakur

https://thakur-nandan.github.io

Email: *nandant@gmail.com*

◇ Google Scholar ◇ Semantic Scholar ◇ Twitter ◇ GitHub ◇ LinkedIn

EDUCATION

Univesity of Waterloo — September '21 - Present
MS / PhD in Computer Science — Ontario, Canada
(advised by *Prof. Jimmy Lin*)
Topic: *Heterogeneous Benchmarking of IR Systems Across Domains and Languages*

Birla Institute of Technology & Science (BITS) Pilani — July '14 - July '18
B.E. (Hons.) in Electronics & Instrumentation, Minor in Finance — Goa, India

EMPLOYMENT & INTERNSHIPS

Databricks & Mosaic Research — August '24 - Present
Research Intern under Omar Khattab and Prof. Michael Carbin — San Francisco, CA
◇ Automatic framework for RAG benchmark construction.

Vectara — February '24 - July '24
Research Intern under Amin Ahmad — Palo Alto, CA (remote)
◇ Multilingual RAG benchmarking [12] & LLM hallucinations [1].

Google Research — September '22 - May '23
Student Researcher under Daniel Cer and Jianmo Ni — Mountain View, CA
◇ Synthetic construction of multilingual retrieval datasets using LLMs [2].

UKP Lab, Technical University of Darmstadt — November '19 - August '21
Research Assistant under Nils Reimers and Prof. Iryna Gurevych — Darmstadt, Germany
◇ Zero-shot IR benchmarking [10] & data augmentation [11] [9].

KNOLSKAPE — September '18 - October '19
Data Scientist — Bangalore, India
◇ Constructed Krawler.ai, an enterprise multimodal search product.

(EMBL) European Molecular Biology Laboratory — June '18 - August '18
Research Trainee under Manjeet Kumar and Prof. Toby Gibson — Heidelberg, Germany
◇ ML Prediction toolkit to predict phosphorylation sites within protein sequences.

Belong.co — July '17 - December '17
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 Research at EMBL Heidelberg	2018

INVITED TALKS

IIT Delhi & IIIT Delhi — January '24
(*Heterogenous IR Benchmarking across Domains and Languages*) — Delhi, India

Koç University — July '23
(*A Tutorial on Advanced Information Retrieval*) — Virtual

Stanford University — November '22
(*Heterogenous Benchmarking in IR Research*) — Palo Alto, CA

OpenNLP Meetup, Deepset.ai — August '21
(*BEIR, An Open-Source Benchmark for IR Systems*) — Virtual

Academic Service

Lead Organizer of Retrieval Augmented Generation (RAG) Track at TREC 2024 (program)
Competition Lead Organizer on multilingual retrieval at WSDM Cup in February '23 (program)
Reviewer (NLP conferences): ARR Oct-Nov (2021), Jan-Apr (2022), 2023 & 2024 (all cycles)
Reviewer (ML & IR conferences): NeurIPS 2023, SIGIR '23, ECIR '24

Publications

Check my Google Scholar profile for all of my publications. NeurIPS, ACL, EMNLP, NAACL, TACL, SIGIR are top-tier peer-reviewed conferences in ML / NLP / IR with an acceptance rate between 20-30%.

[1] Knowing When You Don't Know for Robust Multilingual Retrieval-Augmented Generation
Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, Jimmy Lin.
EMNLP 2024 (Findings)

[2] Leveraging LLMs for Synthesizing Training Data Across Many Languages in Multilingual Dense Retrieval
Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Frederick Wieting, Jimmy Lin, Daniel Cer.
NAACL 2024

[3] Systematic Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR
Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, Jimmy Lin.
SIGIR 2024 (Resource Track) – Oral

[4] Resources for Brewing BEIR: Reproducible Reference Models and Statistical Analyses
Ehsan Kamalloo, *Nandan Thakur*, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin.
SIGIR 2024 (Resource Track) – Oral

[5] MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages
Xinyu Zhang*, *Nandan Thakur**, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, Jimmy Lin. (∗ denotes equal contribution)
TACL 2023

[6] SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval
Nandan Thakur, Kexin Wang, Iryna Gurevych, Jimmy Lin.
SIGIR 2023 (Resource Track)

[7] Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval
Nandan Thakur, Nils Reimers, Jimmy Lin.
ReNeuIR Workshop @ SIGIR 2023 – Oral

[8] Evaluating Embedding APIs for Information Retrieval
Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, *Nandan Thakur*, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Jimmy Lin.
ACL 2023 (Industry Track)

[9] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval
Kexin Wang, *Nandan Thakur*, Nils Reimers, Iryna Gurevych.
NAACL 2022

[10] BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych.
NeurIPS 2021 (Datasets and Benchmarks Track)

[11] Augmented SBERT: Data Augmentation for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks
Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych.
NAACL 2021

PREPRINTS

[12] MIRAGE-Bench: Automatic Multilingual Benchmark Arena for Retrieval-Augmented Generation Systems
Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, Amin Ahmad.
Arxiv 2024 (under review)

[13] Ragnarök: A Reusable Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track
Ronak Pradeep*, *Nandan Thakur**, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, Jimmy Lin. (∗ denotes equal contribution)
Arxiv 2024 (under review)

[14] UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor
Shivani Upadhyay, Ronak Pradeep, *Nandan Thakur*, Nick Craswell, Jimmy Lin.
Arxiv 2024 (under review)

[15] A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution
Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, *Nandan Thakur*, Jimmy Lin.
Arxiv 2023

[16] Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval
Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, *Nandan Thakur*, Xinyu Zhang.
Arxiv 2023

TEACHING EXPERIENCE

Head TA at University of Waterloo (Fall, Winter and Spring)

- CS 116 Introduction to Computer Science 2 — Winter '24
- CS 370 Numerical Computation — Fall '23, Summer '24
- CS 479/679 Introduction to Artificial Intelligence — Winter '23
- CS 136 Elementary Algorithm Design — Spring '23, Winter '22
- CS 241 Foundations of Sequential Programs — Spring '22
- CS 135 Designing Functional Programs — Fall '21

PRESS & MEDIA

Moving Beyond BEIR: Snowflake AI Research Joins Forces with the University of Waterloo, *Snowflake AI*
Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages, *WSDM Cup 2023*
Domain Adaptation with Generative Pseudo-Labeling (GPL), *Pinecone.ai*
Extending Neural Retrieval Models to New Domains and Languages, *Zeta Alpha*
BEIR benchmark in CS224U Teaching Material Stanford University, *Stanford University*
BEIR benchmark as a helpful ML library, *ML News by Yannic Kilcher*
Making the Most of Data: Augmentation with BERT, *Pinecone.ai*
Advance BERT model via transferring knowledge from Cross-Encoders to Bi-Encoders, *Towards Data Science*

COURSEWORK

University of Waterloo: (Fall, Winter and Spring)

- CS 680 Introduction to Machine Learning — Fall '23
- CS 889 Data Sources for Emerging Tech — Spring '23
- CS 886 Graph Neural Networks — Winter '23
- CS 886 Robustness of Machine Learning — Spring '22

- CS 848 Information Retrieval & CS 679 Neural Networks — Winter '22
- CS 854 Experimental Performance Evaluation & CS 649 Human-Computer Interaction — Fall '21

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

COMPETENCIES

Languages: Bengali (*native*), English (*fluent*, TOEFL 110), Hindi (*fluent*), German (*elementary*, A2)
Programming: Python, JavaScript, ReactJS, R, C++, HTML, CSS, Excel, MATLAB, Racket, LaTeX.
Libraries and Services: Pytorch, JAX, Tensorflow, Flask, Django, SQL, MongoDB, Docker, Elasticsearch, Redis, RabbitMq, Apache-Airflow, Postman.

REFEREES

Prof. Jimmy Lin, Full Professor, University of Waterloo
Prof. Iryna Gurevych, Full Professor, TU Darmstadt; Adjunct Professor, MBZUAI
Nils Reimers, Director of Machine Learning, Cohere.ai
Omar Khattab, Post-Doctoral Researcher, Databricks; Incoming Assistant Professor, MIT
Daniel Cer, Senior Research Scientist, Google Research

CO-CURRICULAR

Mime Club Coordinator, BITS Goa 2016-2017
Led a team of 30 student performers in BITS Goa. Involved in acting, sound mixing, designing slides and creating stories for more than 10 shows over a span of 4 years, check it out on YouTube.