Nandan Thakur

thakur-nandan.github.io

Email: nandan.thakur@uwaterloo.ca
[Google Scholar] [Twitter] [GitHub] [LinkedIn]

Education

Univesity of Waterloo, Waterloo, ON, Canada — Sep 2021- Present
Ph.D. Student, David R. Cheriton of Computer Science
Supervisor: *Prof. Jimmy Lin*

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

Research Experience

Univesity of Waterloo — Sep 2021 - Present, Canada
Ph.D. Student – Supervisor: Prof. Jimmy Lin
Working on multilingual information retrieval [4] [13] [1], data and model efficiency [6], and reproducibility [5] [3]. Recently, I have been interested in improving retrieval-augmented generation (RAG) evaluation with large language models (LLMs) [11] [12] [7].

Vectara — Feb 2024 - Present, Virtual
Research Internship – Mentor: Amin Ahmad
Working on reducing LLM hallucinations present in multilingual retrieval-augmented generation (RAG) settings [11] by constructing large-scale multilingual instruction and DPO training datasets.

Google Research — Sep 2022 - May 2023, USA
Student Researcher – Mentors: Daniel Cer, Jianmo Ni
Worked on improving existing multilingual retrieval systems using PaLM 2 generated synthetic data, without expensive human-labeled training data for 18 languages [1].

UKP Lab, Technical University of Darmstadt — Nov 2019 - Aug 2021, Germany
Research Assistant – Supervisors: Prof. Iryna Gurevych, Nils Reimers
Developed a zero-shot benchmark to evaluate out-of-domain (OOD) evaluation of retrieval systems [9] and data-augmentation to generate synthetic data for domain adaptation in pairwise sentence [10] and retrieval tasks [8].

Publications

[1] 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.
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024.

[2] 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.
ACM SIGIR Conference on Research and Development in Information Retrieval, 2024 (Resource Track).

[3] Resources for Brewing BEIR: Reproducible Reference Models and Statistical Analyses.
Ehsan Kamalloo, **Nandan Thakur**, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin.
ACM SIGIR Conference on Research and Development in Information Retrieval, 2024 (Resource Track).

[4] 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)
Transactions of the Association for Computational Linguistics (TACL), 2023.

[5] SPRINT: A Unified Toolkit for Evaluating and Demystifying Zero-shot Neural Sparse Retrieval.
Nandan Thakur, Kexin Wang, Iryna Gurevych, Jimmy Lin.
ACM SIGIR Conference on Research and Development in Information Retrieval, 2023 (Resource Track).

[6] Injecting Domain Adaptation with Learning-to-hash for Effective and Efficient Zero-shot Dense Retrieval.
Nandan Thakur, Nils Reimers, Jimmy Lin.
Workshop on Reaching Efficiency in Neural Information Retrieval, 2023 (ReNeuIR: Oral Presentation).

[7] Evaluating Embedding APIs for Information Retrieval.
Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, **Nandan Thakur**, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Jimmy Lin.
The Annual Conference of the Association for Computational Linguistics (ACL), 2023 (Industry Track).

[8] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval.
Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych.
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022.

[9] BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych.
NeurIPS (Datasets and Benchmarks Round 2), 2021.

[10] Augmented SBERT: Data Augmentation for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks.
Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych.
Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2021.

PREPRINTS

[11] 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.
Currently under review, 2024.

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

[13] 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**, Jheng-Hong Yang, Xinyu Zhang.
Arxiv Preprint, 2023.

SELECTED AWARDS & GRANTS

David R. Cheriton Graduate Scholarship	2024
Snowflake AI Research & University of Waterloo Collaborative Grant	2024
BEIR benchmark in CS224U Teaching Material at Stanford University	2021
Developed both the ELLIS NLP 2021 and SustaiNLP 2021 workshop websites.	2021
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

INDUSTRY EXPERIENCE

KNOLSKAPE Sep 2018 - Oct 2019, India
Data Scientist (Manager: Chaithanya Yambari)
Worked on developing Krawler.ai, an enterprise product for effectively searching a company's large messy content libraries (pdf, xlsx, docx, etc.) with multimodal search. Implemented search functionality using Elasticsearch and backend data ingestion using Flask, Apache-Airflow and MongoDB.

Belong.co Jul 2017 - Dec 2017, India
Data Science Intern (Manager: Vinodh K. Ravindranath)
I worked on topic modeling for clustering millions of candidate resumes. Extracted keywords using FlashText and developed an automatic clustering algorithm using GuidedLDA, a semi-supervised algorithm.

Teaching Experience

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

- CS 116 Introduction to Computer Science 2 — Winter 2024
- CS 370 Numerical Computation — Fall 2023, Summer 2024
- CS 479/679 Introduction to Artificial Intelligence — Winter 2023
- CS 136 Elementary Algorithm Design — Spring 2023, Winter 2022
- CS 241 Foundations of Sequential Programs — Spring 2022
- CS 135 Designing Functional Programs — Fall 2021

Services

Shared-Task Lead Organizer on Retrieval Augmented Generation (RAG) Track at TREC 2024.
Competition Lead Organizer on MIRACL at WSDM Cup 2023.
Reviewer (*CL/NLP conferences): ACL Rolling Review: Oct-Nov (2021), Jan-Apr (2022)
Reviewer (ML conferences): NeurIPS 2023.
Reviewer (IR conferences): SIGIR 2023, ECIR 2024, NAACL 2024.

Invited Talks

Heterogenous IR Benchmarking across Domains and Languages, *IIIT Delhi & IIT Delhi* — India, 2024
Advanced Information Retrieval (Tutorial), *Koç University* — Virtual, 2023
Heterogenous Benchmarking in IR Research, *Stanford University* — USA, 2022
BEIR, An Open-Source Benchmark for IR Systems, *OpenNLP Meetup, Deepset.ai* — Virtual, 2021

Coursework

University of Waterloo: (Fall, Winter and Spring)

- CS 680 Introduction to Machine Learning — Fall 2023
- CS 889 Data Sources for Emerging Tech — Spring 2023
- CS 886 Graph Neural Networks — Winter 2023
- CS 886 Robustness of Machine Learning — Spring 2022
- CS 848 Information Retrieval & CS 679 Neural Networks — Winter 2022
- CS 854 Experimental Performance Evaluation & CS 649 Human-Computer Interaction — Fall 2021

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

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 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*

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
Dr. Daniel Cer, Senior Research Scientist, Google Research
Dr. Nils Reimers, Director of Machine Learning, Cohere.ai