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)
Focusing on standardizing retrieval-augmented generation with LLMs [11] [12] [7]. Worked on multilingual information retrieval [4] [13] [1], data and model efficiency [6] and reproducibility [5] [3]. Recently focusing on retrieval-augmented generation with 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 (SELECTED)

[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.
To appear in 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.
To appear in SIGIR 2024 - Resource Track.

[3] Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard.
Ehsan Kamalloo, **Nandan Thakur**, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, Jimmy Lin.
To appear in SIGIR 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)
TACL 2023 & WSDM Cup 2023 Competition.

[5] 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.

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

[7] 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.

[8] GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval.
Kexin Wang, **Nandan Thakur**, Nils Reimers, Iryna Gurevych.
NAACL-HLT 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 2021 - Datasets and Benchmark Track.

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

PREPRINTS (SELECTED)

[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.
Under review at ACL 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.

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)
Worked on topic modeling for clustering millions of candidate resumes. Extracted keywords using FlashText and automatically clustered candidates using GuidedLDA, a semi-supervised LDA algorithm.

HONOR AND AWARDS

University of Waterloo (UW) Graduate Scholarship 2021 - Present
BEIR benchmark: CS224U teaching material at Stanford University. 2021
Created 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

Finalists in Technology Premier League (TPL) held by CIO & Leader, IT Next. 2019
Received a fully-funded ML fellowship in EMBL Heidelberg 2018

Teaching Experience

Teaching Assistant, University of Waterloo 2021 - Present

- CS 116 (Introduction to Computer Science 2) - Winter 2024
- CS 370 (Numerical Computation) - Fall 2023, Summer 2024 (Upcoming)
- 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

Competition Lead Organizer: WSDM Cup 2023.
Shared-task Lead Organizer: (Upcoming) TREC RAG 2024.
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

IIIT Delhi: IR Benchmarking in Domains and Languages India, Jan 2024
IIT Delhi: IR Benchmarking in Domains and Languages India, Jan 2024
Koç University: Advanced Information Retrieval (Tutorial) Virtual, Jun 2023
Stanford University: Heterogenous Benchmarking in IR Research USA, Nov 2022
OpenNLP Meetup: BEIR, An Open-Source Benchmark for IR Systems Virtual, Jun 2021

Coursework

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

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

Press and Media

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

Competences

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