XINYU ZHANG (Crystina)

xinyucrystina.zhang@uwaterloo.ca
http://crystina-z.github.io

EDUCATION

University of Waterloo — Waterloo, Canada
Ph.D. in Computer Science | Advisor: Prof. Jimmy Lin — Sept. 2020 – Present

Hong Kong University of Science and Technology — Hong Kong SAR China
B.E in Computer Science | Thesis Advisor: Prof. Lei Chen — Sept. 2015 - Jan. 2020

University of California, Los Angeles — Los Angeles, United States
International Exchange Program — Jun. 2017 – Aug. 2017

EMPLOYMENTS

Google DeepMind | New York City, United States — Nov. 2023 – Aug. 2024
Student Researcher | Host: Dr. Jing Lu, Dr. Donald Metzler

Cohere | Toronto, Canada — May 2023 – Oct. 2023
Research Intern | Host: Dr. Sebastian Hofstätter, Dr. Patrick Lewis

Max Planck Institute for Informatics | Saarbrücken, Germany — Feb. 2020 – Aug. 2020
Visiting Scholar Fellowship | Host: Prof. Andrew Yates

NAVER Clova AI Lab | Seongnam, South Korea — Sept. 2018 – Dec. 2018
AI Research Intern | Supervisor: Prof. Lei Chen, Prof. Sung Kim

UROP | Hong Kong SAR, China — Sept. 2017 – Aug. 2018
Undergraduate Student Researcher | Supervisor: Prof. Samuel Chow, Prof. Gary Shueng Han Chan

PUBLICATIONS

[1] FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture
Wenyan Li, **Xinyu Zhang**, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, Desmond Elliott
In the 2024 Conference on Empirical Methods in Natural Language Processing

[2] Words Worth a Thousand Pictures: Measuring and Understanding Perceptual Variability in Text-to-Image Generation
Raphael Tang, **Xinyu Zhang**, Lixinyu Xu, Yao Lu, Wenyan Li, Pontus Stenetorp, Jimmy Lin, Ferhan Ture
In the 2024 Conference on Empirical Methods in Natural Language Processing

[3] NoMIRACL: 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
In the 2024 Conference on Empirical Methods in Natural Language Processing

[4] CELI: Simple yet Effective Approach to Enhance Out-of-Domain Generalization of Cross-Encoders.
Xinyu Zhang*, Minghan Li*, Jimmy Lin
In the 2024 Conference of the North American Chapter of the Association for Computational Linguistics Jun. 2024

[5] Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models
Raphael Tang*, **Xinyu Zhang***, Xueguang Ma, Jimmy Lin and Ferhan Ture
In the 2024 Conference of the North American Chapter of the Association for Computational Linguistics Jun. 2024

[6] CIRAL: A Test Collection for CLIR Evaluations in African Languages. [Best Paper Nominee]
Mofetoluwa Adeyemi, Akintunde Oladipo, **Xinyu Zhang**, Jimmy Lin, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, ... (17 authors)
In the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval Jul. 2024

[7] 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, and Jimmy Lin
Transactions of the Association for Computational Linguistics, Sept. 2023

[8] Toward Best Practices for Training Multilingual Dense Retrieval Models
Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, Jimmy Lin
ACM Transactions on Information Systems 42 (2), 1–33, Sept. 2023

[9] Spacerini: Plug-and-play search engines with Pyserini and Hugging Face
Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, **Xinyu Zhang**, Akintunde Oladipo, Jimmy Lin, Martin Potthast
In the 2023 Conference on Empirical Methods in Natural Language Processing

[10] GAIA Search: Hugging Face and Pyserini Interoperability for NLP Training Data Exploration
Aleksandra Piktus, Odunayo Ogundepo, Christopher Akiki, Akintunde Oladipo, **Xinyu Zhang**, Hailey Schoelkopf, Stella Biderman, Martin Potthast, Jimmy Lin
In the 2023 Conference of the Association for Computational Linguistics

[11] Evaluating Embedding APIs for Information Retrieval
Ehsan Kamalloo, **Xinyu Zhang**, Odunayo Ogundepo, Nandan Thakur, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Jimmy Lin
In the 2023 Conference of the Association for Computational Linguistics

[12] Certified Error Control of Candidate Set Pruning for Two-Stage Relevance Ranking [Oral]
Minghan Li*, **Xinyu Zhang***, Ji Xin, Hongyang Zhang, and Jimmy Lin
In the 2022 Conference on Empirical Methods in Natural Language Processing Dec. 2022

[13] AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages
Odunayo Ogundepo, **Xinyu Zhang**, Shuo Sun, Kevin Duh, and Jimmy Lin
In the 2022 Conference on Empirical Methods in Natural Language Processing, Dec. 2022

[14] Squeezing Water from a Stone: A Bag of Tricks for Further Improving Cross-encoder Effectiveness for Reranking
Ronak Pradeep*, Yuqi Liu*, **Xinyu Zhang**, Yilin Li, Andrew Yates, and Jimmy Lin
In the 44rd European Conference on Information Retrieval Apr. 2022

[15] Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval [Oral]
Xinyu Zhang, Xueguang Ma, Peng Shi, Jimmy Lin
In the 2021 Conference on Empirical Methods in Natural Language Processing Workshop on Multilingual Representation Learning, Nov. 2021

[16] Comparing Score Aggregation Approaches for Pretrained Neural Language Models
Xinyu Zhang, Andrew Yates, and Jimmy Lin
European Conference on Information Retrieval (ECIR), 2021

[17] Bag-of-Words Baselines for Semantic Code Search
Xinyu Zhang, Ji Xin, Andrew Yates, Jimmy Lin
Proceedings of the 1st Workshop on Natural Language Processing for Programming, 2021

[18] A Little Bit Is Worse Than None: Ranking with Limited Training Data
Xinyu Zhang, Andrew Yates, and Jimmy Lin
In the 2020 Conference on Empirical Methods in Natural Language Processing Workshop on Simple and Efficient Natural Language Processing, Nov. 2020

[19] Flexible IR Pipelines with Capreolus
Andrew Yates, Kevin Martin Jose, **Xinyu Zhang**, Jimmy Lin
In the 29th ACM International Conference on Information and Knowledge Management, Oct. 2020

PREPRINT	[20] Rank-without-GPT: Building GPT-Independent Listwise Rerankers on Open-Source Large Language Models **Xinyu Zhang**, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, Jimmy Lin
	[21] What do Llamas really think? Revealing preference biases in language model representations Raphael Tang, **Xinyu Zhang**, Jimmy Lin, Ferhan Ture
	[22] Zero-shot listwise document reranking with a large language model Xueguang Ma, **Xinyu Zhang**, Ronak Pradeep, Jimmy Lin
	[23] 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** (in alphabetical order)
	[24] HAGRID: A human-LLM collaborative dataset for generative information-seeking with attribution Ehsan Kamalloo, Aref Jafari, **Xinyu Zhang**, Nandan Thakur, Jimmy Lin
	[25] Better Than Whitespace: Information Retrieval for Languages without Custom Tokenizers Odunayo Ogundepo, **Xinyu Zhang**, Jimmy Lin

TALKS	**SIGIR 2024 Large Language Model Day:** Multilinguality in the Era of LLMs. 2024
	Google DeepMind: 80/20 Rule in Word Embedding: A Study on Multilingual Vocabulary and Embedding Redundancy in Encoder Models. 2024
	WSDM Cup 2023 Overview: Multilingual Information Retrieval Across a Continuum of Languages 2023
	Math and Computing Research Discovery Day, University of Waterloo: No Language Left Behind from the Lens of IR 2023

SERVICES	**Competition:** FIRE 2023: Cross-Lingual Information Retrieval for African Languages WSDM Cup 2023: Multilingual Information Retrieval Across a Continuum of Languages
	Area Chair: ARR 2023
	Conference Reviewer: ARR 2023-Present SIGIR 2024 ECIR 2023
	Conference Volunteer: EMNLP 2022 ACL 2023 NAACL 2024
	Departmental Service: Data System Group (DSG) Lab Committee Member, University of Waterloo (2022–)

MENTORING	Yuqi Liu. Undergrad and Master Student of UWaterloo. Published [14]	2021
	Odunayo Ogundepo. Master Student of Uwaterloo. Published [13,25]	2022
	Mofetoluwa Adeyemi. Master Student of UWaterloo. Published [3]	2023
	Andrew Liu. Undergrad Student of UWaterloo	2024
	Edward Xu. Undergrad Student of UWaterloo	2024

TEACHING EXPERIENCE	*TA, CS 135. Designing Functional Programs*	Fall 2022
	TA, CS 136. Elementary Algorithm Design and Data Abstraction	Winter 2021
		Spring 2022, Winter 2023
	TA, CS 246. Object-Oriented Software Development	Spring 2021
	TA, CS 343. Concurrent and Parallel Programming	Fall 2021, Winter 2022

Awards	Best Paper Nominee at SIGIR 2024	2024
	David R. Cheriton Graduate Scholarships in Computer Science Award	2023
	International Award of Excellence Scholarship	2020