

Puxuan (Martin) Yu *Ph.D. Candidate*

Center for Intelligent Information Retrieval
College of Information and Computer Sciences
University of Massachusetts Amherst

✉ pxyuwu@gmail.com
☎ +1 (339) 999-7566 // +44 07931-938660
🌐 [Personal Site](#) – [Google Scholar](#) – [in](#)

Professional Experience

- Dataminr**, New York, NY, Research Intern (Oct 2023 - Jan 2024). Hosts: Hemank Lamba and Daniel Cohen.
- Conduct research on the scale calibration of neural ranking models via large language model generated natural language explanations. Our method leads to up to 25% reduction in calibration error while maintaining or improving ranking performance ([Preprint](#)).
- Amazon**, Boston, MA, Applied Scientist Intern (May - Aug 2022). Hosts: Antonio Mallia and Matthias Petri.
- Developed and improved the machine learning pipeline of Alexa’s web-scale retrieval based on learned sparse retrieval (up to 12% effectiveness and 50% efficiency) by optimizing language model vocabulary ([ECIR’24](#)).
- Baidu Research USA**, Sunnyvale, CA, Research Intern (May - Dec 2020). Hosts: Hongliang Fei and Ping Li.
- Proposed and implemented retrieval-oriented pretraining of multilingual language models, leading to up to 29.7% improvement on cross-lingual IR and 9.8 point F1 improvement on cross-lingual QA ([WWW’21](#)).
- University of Virginia**, Charlottesville, VA, Research Intern (Jul - Sep 2017). Host: Prof. Hongning Wang.
- Developed a web browser plugin with JavaScript for obfuscation-based privacy protection in personalized web search. It submits obfuscation queries in the background and visualizes obfuscation history ([SIGIR’18](#)).
- University of Massachusetts Amherst**, Amherst, MA, Research Assistant (Sep 2019 - Present)
- Explainable Search: Performed task formulation, datasets creation, ML model implementation and evaluations for three tasks – generating explanation of search results in context ([SIGIR’22](#)), explainable search result diversification ([CIKM’23](#)), and data augmentation via LLM-generated natural language explanations for training large-scale retrieval/ranking models (work in progress).
 - Multilingual Search: Developed ML solutions for improving multilingual information retrieval via multilingual word embeddings ([SIGIR’20](#)) and knowledge distillation ([WSDM’23](#), [WSDM’23 Cup](#)).
 - Entity-based Search: Improved the effectiveness and efficiency of entity set expansion methods using unsupervised ([SIGIR’19](#), [SIGIR’21](#)) and supervised ML methods ([ICTIR’20](#)).
-

Education

University of Massachusetts Amherst, PhD in Computer Science (2018 - **May 2024 Expected**)

Advisor: James Allan; Thesis: “Leveraging Explanations for IR Systems under Data Scarcity.”

University of Massachusetts Amherst, MS in Computer Science (2018 - 2021); GPA 4.0/4.0

Information Retrieval · Machine Learning · NLP · Advanced Algorithms · Database Systems · Neural Networks

Wuhan University, BEng in Software Engineering (2014 - 2018), GPA 3.7/4.0

Skills

- **Technical:** Python · PyTorch (Distributed, Lightning) · Transformers · Slurm · AWS · Large Language Models
 - **Research & Engineering Experiences:** Information Retrieval (Search, Retrieval & Ranking) · Natural Language Processing · Machine Learning · Self-Supervised Learning · Language Model Fine-tuning · Interpretability · Explainability · Data Augmentation · Calibration · Transfer Learning
-

Honors and Awards

- Winner of the ACM WSDM Cup 2023 – [MIRACL](#), Runner-up.
 - Thesis/Proposal Writing Fellowship, UMass Amherst CICS, Spring 2023/2024.
 - ACM Conference Student Travel Grant: SIGIR 2018-22, ICTIR 2020, WWW 2021.
 - Outstanding Student Award, Wuhan University, 2015-2017.
-

Professional Service (PC Member)

WSDM’24, WWW’23, SIGIR (22-24), CIKM (21-23), EMNLP’21, CCL’21; ACM TOIS; IEEE BigData ’24.

Publications

- Puxuan Yu, Daniel Cohen, Hemank Lamba, Joel Tetreault and Alex Jaimes. “Explain then Rank: Scale Calibration of Neural Rankers via Natural Language Explanations from Large Language Models.” **Preprint**.
- Puxuan Yu, Antonio Mallia and Matthias Petri. “Improved Sparse Retrieval with Corpus-Specific Vocabularies.” **ECIR 2024**.
- Puxuan Yu, Razieh Rahimi, Zhiqi Huang and James Allan. “Search Result Diversification Using Aspects as Bottlenecks.” **CIKM 2023**.
- Zhiqi Huang, Puxuan Yu and James Allan. “Cross-lingual Knowledge Transfer via Distillation for Multilingual Information Retrieval.” **Technical Report (Runner-up of WSDM’23 Cup – MIRACL)**.
- Zhiqi Huang, Puxuan Yu and James Allan. “Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation.” **WSDM 2023**.
- Hongliang Fei, Puxuan Yu and Ping Li. “Cross-lingual Language Models and Pretraining of Cross-lingual Language Models.” **US Patent #11886446**.
- Puxuan Yu, Razieh Rahimi and James Allan. “Towards Explainable Search Results: A Listwise Explanation Generator.” **SIGIR 2022**.
- Puxuan Yu, Hongliang Fei and Ping Li. “Cross-lingual Language Model Pretraining for Retrieval.” **WWW 2021**.
- Zhiqi Huang, Razieh Rahimi, Puxuan Yu, Jingbo Shang and James Allan. “AutoName: A Corpus-Based Set Naming Framework.” **SIGIR 2021**.
- Puxuan Yu, Razieh Rahimi, Zhiqi Huang and James Allan. “Learning to Rank Entities for Set Expansion from Unstructured Data.” **ICTIR 2020**.
- Puxuan Yu and James Allan. “A Study of Neural Matching Models for Cross-lingual IR.” **SIGIR 2020**.
- Puxuan Yu, Zhiqi Huang, Razieh Rahimi and James Allan. “Corpus-based Set Expansion with Lexical Features and Distributed Representations.” **SIGIR 2019**.
- Puxuan Yu, Wasi Uddin Ahmed and Hongning Wang. “Hide-n-Seek: An Intent-aware Privacy Protection Plugin for Personalized Web Search.” **SIGIR 2018**.