Peiran Yao — PhD Candidate

+1 (825) 440 3379 peroyao@gmail.com xavieryao

Experience

Researching large language models, knowledge graphs, and general NLP *since 2019*

- Conducted original research in symbolic LLM reasoning, agentic LLMs, obtaining reliable and factual answers from LLMs and knowledge graphs for question answering, by employing Bayesian networks, crowdsourcing models, sampling, prompting, and active learning
- In-depth knowledge of LLMs, Transformers, and PyTorch;
- Formulated findings as research papers and patents, and presented at top-tier conferences including NeurIPS, ACL, EMNLP, EACL, WWW, and KDD.

Building ML- and LLM-powered systems and developing full-stack applications *since 2021*

- Developed open source end-to-end text analysis system for data ingestion, NLP pipelines (ETL), and visualization;
- Applied to news and social media analysis for financial institutions and humanities researchers;
- Managing full life-cycles from requirements to development, testing, MLOps, and CI/CD;
- Core technologies: Python, Docker, Celery/AirFlow, Elasticsearch, Neo4j, Vue.js;
- Various other mobile and web apps as independent and contract work, gaining 130k+ downloads and 900+ stars.

Applying data science for financial texts, academic graphs, and IT operations *since 2017*

- Cleaned, modeled, and analyzed large-scale real-world structured data Open Academic Graph, time series data for AIOps, and unstructured data such as news and call centre notes at financial institutions
- Applied machine learning models (BERT, GBDT, Isolation Forest, etc.) and statistical methods;
- Managed large-scale data with SQL, Spark, and MongoDB.

Education

University of Alberta, Canada

PhD candidate, Computing Science *2019 – January 2025*

Thesis: Assessing and Improving Factual Answers from Knowledge Graphs and Language Models

Tsinghua University, China

BEng, Computer Science and Technology *2015 – 2019*

BA, English Language and Literature (2nd degree)

Employment

Research Intern at Infinigence AI

Research on human-inspired learning, LLM agents, prompting, efficiency, and weak-to-strong generalization. *2024-now*

Research Assistant at Scotiabank

Collaboration between Scotiabank and UofA on modeling call intents and building systems for integrating state-of-the-art NLP tools for news and text analysis. *2020-2023*

Software Developer Intern at BizSeer

Implementing AIOps log analysis algorithms for site reliability. *2018-2019*

Software Developer Intern at Tencent

Building server metric monitoring and alerting algorithms and systems for ML-powered site reliability. *2018*

Services and Honours

Awards

J Gordin Kaplan Graduate Student Award: 2024

Alberta Graduate Excellence Scholarship: 2021

Fung Scholarship: 2017

THU Freshman Award: 2015

Teaching Assistant

Introduction to Natural Language Processing (2020, 2021, 2023)

Teaching and Research Methods (2022)

Introduction to Information Retrieval (2020)

Operating System Concepts (2019)

Professional Roles

Program Committee, ACL / Annual Meeting of the Association for Computational Linguistics (2023, 2024)

Program Committee, EMNLP / Empirical Methods in Natural Language Processing (2022, 2023, 2024)

President, Tsinghua University TUNA Association (2018-2019)

Publications

- **Peiran Yao**, Jerin George Mathew, Shehraj Singh, Donatella Firmani, and Denilson Barbosa. 2024. A bayesian approach towards crowdsourcing the truths from LLMs. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.
- Xuefei Ning*, Zifu Wang*, Shiyao Li*, Zinan Lin*, **Peiran Yao***, Tianyu Fu, Matthew B. Blaschko, Guohao Dai, Huazhong Yang, and Yu Wang. 2024. Can LLMs learn by teaching? A preliminary study. In *Advances in Neural Information Processing Systems*, volume 37.
- Natalie Hervieux*, **Peiran Yao***, Susan Brown, and Denilson Barbosa. 2024. Language resources from prominent born-digital humanities texts are still needed in the age of LLMs. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 85–104, Miami, USA.
- **Peiran Yao**, Kostyantyn Guzhva, and Denilson Barbosa. 2024. Semantic graphs for syntactic simplification: A revisit from the age of LLM. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 105–115, Bangkok, Thailand.
- **Peiran Yao** and Denilson Barbosa. 2024. Accurate and nuanced open-QA evaluation through textual entailment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2575–2587, Bangkok, Thailand.
- **Peiran Yao**, Matej Kosmajac, Abeer Waheed, Kostyantyn Guzhva, Natalie Hervieux, and Denilson Barbosa. 2023. NLP workbench: Efficient and extensible integration of state-of-the-art text mining tools. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 18–26, Dubrovnik, Croatia.
- Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, **Peiran Yao**, Jie Zhang, Xiaotao Gu, Yan Wang, Evgeny Kharlamov, Bin Shao, Rui Li, and Kuansan Wang. 2023. OAG: Linking entities across large-scale heterogeneous knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9225–9239.
- **Peiran Yao**, Tobias Renwick, and Denilson Barbosa. 2022. WordTies: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5959–5970, Abu Dhabi, United Arab Emirates.
- **Peiran Yao** and Denilson Barbosa. 2021. Typing errors in factual knowledge graphs: Severity and possible ways out. In *Proceedings of the Web Conference 2021*, WWW '21, page 3305–3313.
- Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, **Peiran Yao**, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, and Kuansan Wang. 2019. OAG: Toward linking large-scale heterogeneous entity graphs. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2585–2595, New York, NY, USA.
- Yutao Zhang, Fanjin Zhang, **Peiran Yao**, and Jie Tang. 2018. Name disambiguation in AMiner: Clustering, maintenance, and human in the loop. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1002–1011.
- Fanjin Zhang, Xiaotao Gu, **Peiran Yao**, and Jie Tang. 2018. Integration of heterogeneous data from multiple sources. *Journal of Chinese Information Processing*, 32(9):84–92.