Zhong, Wei

Ph.D. in information retrieval & full-stack system engineer

w32zhong@uwaterloo.ca
(CAN) +01 226-600-8357
github.com/w32zhong
w32zhong.github.io
linkedin/w32zhong

Summary

Crafted a search engine (`https://approach0.xyz`) from frontend to backend, I've navigated millions of search queries. Proud recipient of two Best Paper Awards during my Ph.D. program. I am eager to learn new things and am happy to optimize myself through experience.

Education

University of Waterloo 2021 May. – 2023 Sep.
Ph.D. in Computer Science (Cum GPA: 89.25 / 100) *Waterloo, ON, CA*
(Advisor: Prof. and ACM Fellow Jimmy Lin)

Rochester Institute of Technology 2017 Aug. – 2019 Jul. (Transferred Out)
Ph.D. Candidate in Computer Science (Cum GPA: 3.780 / 4.0) *Rochester, NY, USA*
(Advisor: Prof. Richard Zanibbi)

University of Delaware 2013 Aug. – 2015 Aug.
M.S. in Electrical and Computer Engineering (Cum GPA: 3.867 / 4.0) *Newark, DE, USA*
(Advisor: Prof. Hui Fang)

China Ji Liang University (CJLU) 2009 Aug. – 2013 Jun.
B.S. in Information and Computation Science *Hangzhou, P.R. China*

Selected Publications

2023 Ph.D. Thesis **Wei Zhong.** *Effective Math-Aware Ad-Hoc Retrieval based on Structure Search and Semantic Similarities*

2023 SIGIR **Wei Zhong**, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin. *One Blade for One Purpose: Advancing Math Information Retrieval using Hybrid Search.*

2022 CLEF (Best Paper and SOTA results) **Wei Zhong**, Yuqing Xie, Jimmy Lin. *Applying Structural and Dense Semantic Matching for the ARQMath Lab 2022, CLEF.*

2022 EMNLP Findings (1st Math NLP workshop) **Wei Zhong**, Jheng-Hong Yang, Yuqing Xie and Jimmy Lin. *Evaluating token-level and passage-level dense retrieval models for math information retrieval.*

2021 CLEF (The Best Formula Search System) **Wei Zhong**, Xinyu Zhang, Ji Xin, Richard Zanibbi, Jimmy Lin. *Approach zero and anserini at the CLEF-2021 arqmath track: Applying substructure search and BM25 on operator tree path tokens.*

2021 SIGIR **Wei Zhong**, Jimmy Lin. *PyAo: A Python Toolkit for Accessible Math-Aware Search.*

2020 ECIR (My favorate paper) **Wei Zhong**, Shaurya Rohatgi, Jian Wu, C. Lee Giles and Richard Zanibbi. *Accelerating Substructure Similarity Search for Formula Retrieval.*

2019 ECIR (Best Application Paper) **Wei Zhong** and Richard Zanibbi. *Structural Similarity Search for Formulas using Leaf-Root Paths in Operator Subtrees.*

Experience

Microsoft Research Jul. 2023 – Oct. 2023
Internship (Augmented Learning and Reasoning) *Redmond, Washington, USA*

- Improved math answering using retrieval augmentation and Large Language Models (LLMs), outperforming the concurrent SoTA models like Mammoth and WizardMath of the same size.
- Identified the key issue in math retrieval augmentation when search results are mostly false positives. Proposed a method to boost the baseline accuracy by at least 10%.
- Hands-on experience with model parallelism using DeepSpeed.

University of Waterloo Apr. 2021 – Present
Research Assistant and Teaching Assistant *Waterloo, Ontaria, Canada*

- Obtained #1 performance in two recent math information retrieval tasks: CLEF ARQMath-2 and ARQMath-3. As a result, I was awarded the Best Paper at CLEF 2023, a major evaluation forum in Information Retrieval.
- Successul in building advanced neural retrievers like CoCondenser and MAE, boosting 10% in NDCG, 14% in MAP, and 12% in top-result precision compared to the previous best model.

DMAI Apr. 2020 – Sep. 2020
NLP Researcher (internship) *Guangzhou, P.R. China*

- Developed a math expression simplifier that reduces the number of simplifying steps by a 50% improvement compared to the company's online version.
- Applied technologies including Mutual information, RNN with attention, GBDT+LR, LDA, and SVM.
- Implemented a lock-free MCTS agent with different decision strategies using Reinforcement Learning.

Rochester Institute of Technology Aug. 2017 – Aug. 2019
Research Assistant *Rochester, NY, USA*

- Created the state-of-the-art structure search engine for math, achieving top results on the NTCIR dataset, and received a best application paper award at ECIR (the top IR conference in Europe).
- Improved the efficiency of the structure search engine using binary programming by a factor of 3, and making real-world effective math structure search feasible for under half a second.

  (Discontinued and transfered to Canada due to U.S. visa issues caused by the COVID-19 pandemic)

Huawei Technologies Sep. 2016 – July. 2017
Full-time software developer *Shenzhen, P.R. China*

- STB (TV Box) Hardware Abstraction Layer C/C++ code maintenance. Fixed more than 20 non-trivial bugs.
- Participated Peach Fuzzing testing for Android-based system interface.

  (Good communication with my colleagues. Left the team to pursue a PhD. degree.)

SevOne (2015 Glassdoor best places to work) Jun. 2015 – Aug. 2015
WEB backend (S.M.A.R.T.S program internship) *Wilmington DE, US*

- PHP, C/C++, MySQL code maintenance.
- Search engine back-end rewriting using CLucene.

Selected Projects

Approach Zero **Math-aware search engine** 2015 – Present

- Ranked as the #1 Community Promotion Ad of Math StackExchange in both 2020 and 2021 (higher than the rank of OverLeaf), one of the largest math Q&A communities.
- Obtained the state-of-the-art scores in the NTCIR-12, CLEF-2021 and CLEF-2022 Tasks.
- An online version of the search engine is made available, capable of searching tens of millions of structured math formulas in real-time. The search engine is deployed by myself and is hosted by five low-end Linode instances (at a cost about only $50/mo.). I maintain the full software stacks from front-end to back-end.

PyAo **Evaluation toolkit for math IR systems and neural retrievers** 2021 – 2023

- Created 1.7 million effective training data pairs for math IR from scratch.
- Implemented DPR, ColBERT, CoCondenser, MAE deep neural retrievers, covering inference and evaluation in one maintainable code framework.

TinyNN and MNN **Educational deep learning frameworks** Sep. – Oct. 2019

- My open-source new-code contributions include denoising autoencoder, Restricted Boltzmann Machine (w/ CD-k training) and activation maximization (AM) visualization.
- Refactored CNN convolutional layer (the *im2col* function) with maintainable code.
- Hand-written gradient/Jacobian matrix derivation using GPU acceleration based on CuPy.

Mathsteps-v2 **A step-by-step math solver** June. 2020

- Designed a declarative macro language using compiler languages for math transformations.
- Efficient lock-free MCTS math solver in C language, search space is reduced by a policy network.

Search engine UI **A modern Web UI for search engine** June. 2020

- Modern, responsive, and single-page UI application written in Vue 3.
- Under 500 ms website response time, served half a million real user queries with a bundle size of 91 KB.

Gateway **An API gateway service** Nov. 2020

- Solid and minimal API gateway router based on Nginx, Lua and OpenResty.
- Technologies: Docker Swarm service discovery, JWT login, rate limit for unique IP, Prometheus, Grafana metrics, and TLS automatic renewal.

CalaBASH **An orchestration layer for Docker Swarm operated by BASH or Web UI** Jan. 2021

- Experience in bootstrapping a modern web app with a simple and maintainable DevOps approach.
- Deployed highly available search services that utilize sharding, load balancing, and service discovery.
- Technologies: Shell script, Node-js, Docker Swarm, VPS/Cloud APIs. .

Tech. Skill

Software Linux/Shell, Git, WEB stack, docker, C/C++, Python, PyTorch, HuggingFace transformers and TRL.

Hardware Embedded system design, VHDL (See my 8bits TTL CPU in Multisim and Flappy Bird game in VHDL!)

Communication Skills

Teaching Skill Instructed an algorithm colloquium on Dantzig's simplex algorithm and attended a credited Teaching Skills Workshop in Rochester Institute of Technology. Hosted UWaterloo TA office hours for CS136.

Presentation Skill Presented 10+ research papers as the first author, including two award-winning best papers. Maintained effective communication with my Master and Ph.D. program advisors.