BINFENG XU

+1 (336) 744-6809 || billxbf@gmail.com || Page ||

EDUCATION

New York University – *M.S. in Data Science* , 3.9/4.0 SEP 2020 – MAY 2022
- Relevant Coursework: Deep Learning (*Yann Lecun*), Natural Language Processing & Understanding.

Wake Forest University – *B.S. doubling Computer Science and Statistics*, 3.74/4.0 SEP 2016 – MAY 2020
- Relevant Coursework: Data Structure & Algorithms, Machine Learning, Computer Vision, Parallel Programming, Numerical Computation, Mathematical Statistics, Probability Theory, Optimization.

EXPERIENCE

Gentopia.AI APR 2023 – PRESENT

I designed and developed an open-source framework Gentopia for researchers and developers to customize hierarchical ALM agents through config, and a companion platform GentPool to register and evaluate specialized agents. This project enables efficient tuning, evaluation, and hierarchical interaction of task-driven agents.

Full Stack Applied Researcher @ eBay JUL 2022 – PRESENT

I focus on optimizing Click-Through-Rate (CTR) in recommendation through both algorithmic and programmatic techniques. Here are some selected end-to-end projects I've implemented independently:
- **Product Comparison Table.** A tabular view of related/alternative products, each vertically displaying comparable aspects or features. This product improves page-level CTR by 20%+ in A/B tests. Key steps in my solution: (a) A two-tower user-product model for diverse candidate recalls. (b) A batch-updated Aspect Importance model, serving to order Displayed Aspects in vertical view. (c) Using LLM to cache offline Pivot Aspects for k-dup filtering, buyer funnel, and dynamic Catchphrase as placement subtitles.
- **Semantic Book Recommendation.** Recommending semantically similar books based on content. (a) I used external ISBNDB database to build ISBN embedding. (b) I built an auto-scheduled KNN Index Search service. (c) Finally, I created a downstream recall pipeline mapping ISBN KNNs into recommendation candidates.
- **Rec Sys Research Validator.** Reproducing benefits claimed in SoTA papers in Recommender System is usually hard under different industrial contexts. Thereby I built a repo to implement and validate new model/techniques on sampled traffic, quickly checking the potential of new research under eBay's recommendation context.

Research Intern @ eBay MAY 2021 – AUG 2021
- Built and deployed an online detection model for fraudulent seller activities in chat sessions. I trained RoBERTa with concatenation of tokenized messages and user features against fraud labels, and served it with ONNX.

Research @ New York University DEC 2020 – MAR 2021
- *Prediction and Policy-Learning under Uncertainty* Advisor: *Yann LeCun, Alfredo Canziani*
  A policy learning model for self-driving agents. I experimented with an extra penalization on environmental cars in dense traffic, updating their trajectories accordingly in order to avoid colliding into studied agent.

Research @ Wake Forest University SEP 2018 – DEC 2019
- *Tucker Decomposition with f-MRI Neural Activity Tensor* Advisor: *Grey Ballard*
  Used Tucker Decomp. to compress high dimensional fMRI tensors, and discussed robustness with ML models.

PAPERS

ReWOO: Decoupling Reasoning from Observations for Efficient Augmented Language Models 2023
Binfeng Xu*, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, Dongkuan Xu*

Gentopia: A Collaborative Platform for Tool-Augmented LLMs 2023
Binfeng Xu*, Xukun Liu, Hua Shen, Zeyu H, Yuhan L, Murong Y, Zhiyuan P, Yuchen L, Ziyu Y, Dongkuan X*

HONORS

Kaggle: Competitions Master - Global Ranking Top 1% MAY 2018 – JUN 2021
- Santander Customer Transaction Prediction (Banking, Classification): Rank #24 of 8,802 (Gold);
- Santa 2020 Contest (Competitive Reinforcement Learning): Rank #17 of 792 (Silver);
- BirdCLEF 2021 Birdcall Identification (Signal Processing, Time-Series): Rank #15 of 816 (Silver);

ACM ICPC: Regional 4th Place in North Carolina MAR 2019 – MAR 2019

Udacity Nanodegree in Deep Learning SEP 2018 – OCT 2018

TECH STACKS

Areas of Strength: MLOps, NLP Research, Search & Rec System, Augmented Language Model, Data Science.
Programming: Python3*, Scala*, Java, SQL, C#, C++; Spark*, Pytorch*, Spring framework, React.