BINFENG XU

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

EDUCATION

New York University – *M.S. in Data Science* , 3.8/4.0 SEP 2020 – MAY 2022

- Relevant Coursework: Deep Learning (with *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

Senior Research Engineer @ Samsung Research America OCT 2023 – PRESENT

Large Foundation Model. Contributed to multiple frontier AI research and products, including:

- Leading contribution to a **multi-turn retrieval-augmented chatbot** for (Bixby) on-device customer service. This involves massive dialogue curation and finetuning; Optimizing embedding and reranking models for fast retrieval; Integrating guardrail modules for safety; and Parameter-efficient quantization for NPU inference.
- Instruction finetuning of on-device LLMs for personal database queries. Enhanced **coding** (text2sql, text2sparql) **ability** and inference throughput on custom benchmarks. Integrated into an on-device RAG model.
- PoC of a **visual QA** model trained towards taking **autonomous actions/API calls** (for AI eyewear).

Applied Researcher @ eBay JUL 2022 – OCT 2023

Recommender System. Focused on optimizing Click-Through-Rate (CTR) in recommendation through both algorithmic and programmatic techniques. End-to-end projects I've led or independently contributed include:

- **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-updated Fast KNN Index service. (c) I created a downstream recall pipeline mapping ISBN KNNs into recommendation candidates.
- **RecSys Model Validator.** A fast evaluation framework to validate SoTA model/techniques on sampled real-time 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*

Research @ Wake Forest University SEP 2018 – DEC 2019

- *Tucker Decomposition with f-MRI Neural Activity Tensor* Advisor: *Grey Ballard*

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: Large Foundation Models, NLP, MLOps, Search & Rec System, Data Science.

Programming: Python3*, Pytorch*, Scala*, Java, SQL, C#, C++; Spark, Spring, React, Unity3D (simulation).