

# BINFENG XU

[🏠 Profile](#) || [🐙 GitHub](#) || [🗨️ HuggingFace](#) || [🎓 Scholar](#) || [🏆 Kaggle](#) || [🌐 LinkedIn](#) || [🐦 X](#) || [✉️ email](#)

## EXPERIENCE

### **Senior Research Engineer @ Samsung Research America**

OCT 2023 – PRESENT

*Large Foundation Models; On-device AI.*

#### **§ Responsibilities and Contributions:**

**I. Post-Training.** Distributed training of 3 ~ 72B LLMs on 1~64×A100 GPUs for various tasks.

- Supervised Fine-tuning (SFT) LLMs for task-specific instruction following.
- Preference Alignment from feedback (RLHF); Reinforcement Learning from verified reward (RLVR).
- Knowledge Distillation from server-side LLMs to on-device light-weight LLMs.
- Parameter Efficient Fine-tuning with LoRA and QLoRA for on-device deployment.

**II. RAG and Agents.** End-to-end development of two agentic systems for conversational chat assistant.

- Designed and architected a modularized agent framework integrating LLM training, inference, prompts, vector database, embedding models, retrieval, ranking, and function calling.
- Built and benchmarked various light-weight NLP modules for optimal efficiency, such as Named Entity Recognition (NER), Entity Linking and Intent Classification as agent function-calls.
- Fine-tuned LLMs for enhanced function calling, context-aware conversation, and safety / rejection.

#### **§ Selected Projects:**

**I. On-device Chat Assistant over Personal Knowledge Graph.** (Galaxy AI)

- Led a group of 5 researchers and engineers to develop an efficient agent system over Personal Knowledge Graph (PKG) for Samsung Gallery, where an on-device 3B LLM is distilled to effectively locate and retrieve information from on-device knowledge graph database (RDFox) and converse with users.

**II. Customer Service Troubleshooting Chat Assistant.** (Samsung Web Shop)

- Independently developed a chatbot for electronics troubleshooting. Fine-tuned a series (7B, 27B, 70B) of LLMs to retrieve ranked context from PDF user manuals and past troubleshooting chat log. Hosted server instances for inference, interactive web demo and provided RESTful APIs for product.

**III. Research and Side Projects.**

- LLM Reasoning: Mentored a PhD intern on augmenting LLM Mathematical reasoning (on AIME) with step-wise data synthesis from Monte Carlo Tree Search (MCTS) and dynamic step prediction.
- LLM Pretraining: Dumped Yahoo Finance database, tokenized Nasdaq stocks candle history, and trained (from scratch) a 7B transformer on 5 million tokens to predict the market.
- Efficient Inference: Benchmarked Speculative Decoding to accelerate SPARQL code generation.

### **Applied Researcher @ eBay**

JUL 2022 – OCT 2023

*Recommender Systems; Natural Language Processing; MLOps.*

**I. Table-view Product Comparison Module.**

- End-to-end developed a recommendation algorithm for comparable and alternative items, popping up predicted key aspects and features users intend to compare.
- Trained and served (1) A two-tower User-Product BERT-based embedding model for product candidate recalls. (2) An Aspect Importance model to rank popping-up features.
- Collaborated with a front-end engineer to serve through production, which significantly improved page-level Click-Through-Rate (CTR) by 20%+ through multiple A/B tests.

**II. Semantic Book Recommendation.**

- Recreated Book Recommendation service on eBay with semantic embeddings from ISBN databases.
- Built and served a batch-updating ANN Index service for ISBN embeddings. Engineered downstream recall & ranking pipelines to merge ISBN top neighbors into recommendation candidates.

### **Research Intern @ eBay**

MAY 2021 – AUG 2021

Fine-tuned an RoBERTa-based fraud detection model using seller-buyer chat log, and served with ONNX.

### **Research Intern @ Baidu AI Lab**


MAY 2018 – AUG 2018


Benchmarked YOLO v3 for an internal face recognition model based on Siamese Network.

## EDUCATION

**NEW YORK UNIVERSITY** *M.S. in Data Science* SEP 2020 – MAY 2022  
Relevant Coursework: Deep Learning, Natural Language Processing & Understanding.



**WAKE FOREST UNIVERSITY** *B.S. doubling Computer Science and Statistics* SEP 2016 – MAY 2020  
Relevant Coursework: Machine Learning, Computer Vision, Data Structures & Algorithms, Parallel Programming, Numerical Computation, Probability & Statistics, Markov Processes, Convex Optimization.


**Research @ New York University** DEC 2020 – MAR 2021  
• *Prediction and Policy-Learning under Uncertainty*  Advisor: *Yann LeCun, Alfredo Canziani*  
Involved in a policy learning project for self-driving cars, where I experimented with introducing penalization term to the loss of environmental cars in dense traffic to avoid colliding into the studied agent.

**Research @ Wake Forest University** SEP 2018 – DEC 2019  
• *Tucker Decomposition with f-MRI Neural Activity Tensor*  Advisor: *Grey Ballard*  
Developed a novel algorithm to compress high dimensional f-MRI tensors with Tucker Decomposition for data efficiency, benchmarked against various statistical models.  
• *Object Recognition in Peru Forest* Advisor: *Paul Pauca*  
Trained YOLOv3 to detect illegal mining activities of drone-taken images above Peru forests.





## PUBLICATIONS

**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

**Dynamic Noise Preference Optimization for LLM Self-Improvement via Synthetic Data**  2024  
*Haoyan Yang\**, Ting Hua, Shangqian Gao, **Binfeng Xu**, Zheng Tang, Jie Xu, Hongxia Jin, Vijay Srinivasan

## HONORS

**Kaggle: Competitions Master** - Global Ranking Top 1% MAY 2018 – PRESENT  
• **Classification under noise:** Santander Customer Transaction Prediction  **Rank #24 of 8,802**  
• **LLM Reward Modeling:** LMSYS Chatbot Arena Human Preference Predictions  **Rank #30 of 950**  
• **Competitive Agent Reinforcement Learning:** Santa 2020 Contest  **Rank #17 of 792**  
• **Time Series Modeling:** BirdCLEF 2021 Birdcall Identification  **Rank #15 of 816**

**ACM ICPC: Regional Gold in North Carolina** MAR 2019 – MAR 2019

**Udacity Nanodegree in Deep Learning**  SEP 2018 – NOV 2018

## TECH STACK

**Areas of Strength:** NLP\*, LLM Post-training\* (SFT, RLHF, RLVR), Retrieval-Augmented Generation (RAG), Agentic Systems, Search & Ranking, Recommender Systems, MLOps.

**Programming Languages:** Python3\*, Scala\*, Java, SQL, SPARQL, Kotlin, C#, C++.

**Infrastructure & Frameworks:** PyTorch\*, HuggingFace\* (Transformers, TRL, PEFT, Datasets), DeepSpeed, PyTorch Lightning, Verl, vLLM, LMDeploy, llama.cpp, open-instruct, Ray, Weights & Biases, Unity3D (simulation).

**Data:** Apache Spark, Lucene, VectorDBs (Faiss, Weaviate), Databases (RDFox, Neo4j, SQLite, MongoDB).

**Serving & Deployment:** Docker, Kubernetes, Flask, FastAPI, Streamlit, Spring Boot, AWS (EC2, S3).