

Yang Bai, Ph.D.

tony.bai.yang@gmail.com +1 (352) 284-2346 tonyby.github.io LinkedIn

Summary

Machine Learning Research Scientist specializing in large-scale modeling and decision systems. Experienced in training and deploying models on billions of samples for real-world optimization problems. Background in recommendation systems, information retrieval, and multimodal LLM-based modeling. Ph.D. in Computer Science with publications in NAACL, COLING, and SIGIR.

Experience

Meta <i>Research Scientist</i>	Feb 2025 – Present
- Built large-scale engagement prediction models for Instagram Feed Ads, trained on billions of user interactions to estimate scroll probability under ad exposure. - Designed causal inference models to estimate counterfactual engagement sensitivity, improving robustness under changing ad policies. - Led migration of training pipelines into Meta's NRML framework, enabling scalable training and deployment across production ranking systems. - Deployed models via large-scale A/B experiments, increasing IG Feed revenue by 0.6% (~\$0.6M/day) while maintaining neutral impact on top-line engagement metrics. - Developed modeling frameworks to capture session-level user behavior dynamics, enabling data-driven decision systems for ads supply optimization. - Worked with large-scale distributed training pipelines handling billions of samples with strict latency and reliability constraints.	
Meta <i>Machine Learning Engineer Intern</i>	Summer 2024
- Built neural ranking models for large-scale search relevance, improving ranking performance over production baselines. - Designed data pipelines to generate synthetic labels for training at scale. - Ran large-scale offline experiments to evaluate model performance across multiple architectures.	
Nokia Bell Labs <i>Machine Learning Intern</i>	Summer 2022
Built dense retrieval models for large-scale log analysis (10M+ entries), improving retrieval performance by 16.1% over BM25.	
University of Florida <i>Graduate Researcher</i>	2019 – 2024
- Developed multi-hop dense retrieval systems for open-domain fact verification on the FEVER benchmark. - Built datasets and annotation platforms for multi-answer open-domain QA and stance mining. - Developed an event-centric knowledge graph search system for DARPA AIDA, ranked #1 in NIST TAC SM-KBP 2020.	

Education

University of Florida Ph.D. in Computer Science GPA: 3.96/4.0	2019 – 2024
University of Florida M.S. in Electrical and Computer Engineering GPA: 3.5/4.0	2016 – 2018
Sichuan University B.S. in Microelectronics	2012 – 2016

Skills

Programming
Python, Java, C++, SQL

Machine Learning
Deep Learning, Causal Inference, Learning-to-Rank, Retrieval Systems

Frameworks
PyTorch, TensorFlow, Transformers, Scikit-learn

Data & Infrastructure
Spark, Linux, Docker, Distributed Training

Domains
Ranking Systems, Information Retrieval, NLP, Causal Inference

Selected Impact

- Models deployed to production systems serving billions of daily impressions.
- Experience designing end-to-end ML systems combining modeling, data, and large-scale infrastructure.

Selected Publications

NAACL Findings 2025
RAMQA: Retrieval-Augmented Multi-Modal Question Answering (LLM-based multimodal QA)

COLING 2024
M3: Multi-Hop Dense Sentence Retrieval

SIGIR 2023
MythQA: Query-Based Claim Detection