Binxuan Huang

Email: binxuanhuang@gmail.com
binxuan@amazon.com
Homepage: https://binxuan.github.io
Phone: +1-412-652-8661

Summary

Principal Scientist focused on LLM pretraining. Led end-to-end development of Amazon Rufus foundation models—from architecture, scaling laws, and training recipes to data mixture/curriculum and productionization—shipping 4 generations including an ultra-sparse MoE model that matches or exceeds leading open models on key benchmarks. 10+ years of experience across academia and industry building LLMs, knowledge graphs, and large-scale data/ML systems, with a strong publication record and extensive service as reviewer and area chair.

Research Interests

Large Language Model, Natural Language Processing, Deep Learning, Data Mining

Education

Carnegie Mellon University *Pittsburgh, U.S.*
Ph.D., Computer Science *2020*
Zhejiang University *Hangzhou, China*
B.S., Physics *2015*
B.E., Computer Science *2015*

Experience

Principal Scientist *Amazon Rufus, Feb 2023 - Present*

- Lead end-to-end foundation model development for Amazon Rufus from day one, setting technical vision and roadmap across architecture, pretraining scaling laws, training recipes, data mixture and curriculum design.
- Delivered four generations of Rufus LLMs (dense and ultra-sparse MoE) with step-function quality improvements at each iteration.
- Designed and delivered the latest ultra-sparse MoE model (3.25% active parameters) that matches or exceeds leading open models (e.g., DeepSeek-V3-Base, Kimi-K2-Base) on key benchmarks including MMLU, MMLU-Pro, Math, plus internal shopping benchmarks.
- Collaborated closely with infrastructure, inference, and post-training teams to land research into Rufus production and align capabilities with core shopping use cases.

Senior Applied Scientist *Amazon Product Graph, June 2020 - Feb 2023*

- Led knowledge extraction for Amazon Product Graph, building automatic extraction for a global product catalog.
- Improved extraction recall by 70% at similar precision vs. the previous production system, materially improving coverage of structured product knowledge.
- Designed and shipped end-to-end text-based and web-based extraction pipelines: data preprocessing, model design, large-scale inference, post-processing, and evaluation.

- Worked cross-functionally with product, engineering, and evaluation teams to integrate extraction outputs into downstream catalog system.

Applied Scientist Intern *Amazon Alexa AI, May 2019 - Aug 2019*

- Proposed and implemented a multi-grained matching approach leveraging knowledge graphs to improve Alexa NLU performance.
- Built an entity-linking system achieving state-of-the-art results on a public benchmark and scaled it across the entire Amazon Music knowledge graph.

Research Assistant *CASOS, Carnegie Mellon University, 2015 - 2020*

Advisor: Prof. Kathleen M. Carley

- Proposed multiple methods for semantic modeling sentence pairs, with an application on aspect-level sentiment classification [EMNLP'19, EMNLP'18, SBP-BRiMS'18].
- Modeling social media users with various types of features, eg. text, categorical features, network, for user attributes prediction [EMNLP'19, TNSE'18, SBP-BRiMS'17, CIKM'17].
- Developed a high throughput tweet collection system. Collected and analyzed more than 40 billion tweets (16TB after compression) from 20 million users using Spark [ASONAM'19, ASONAM'17]

Research Assistant *AI Lab, Zhejiang University, 2014 - 2015*

Advisor: Prof. Xiaogang Jin

- Conducted research on analyzing Wikipedia editing pattern.

Publications

- Lee, Jing Yang JY, Hamed Bonab, Nasser Zalmout, Ming Zeng, Sanket Lokegaonkar, Colin Lockard, **Binxuan Huang**, Ritesh Sarkhel, and Haodong Wang. "DocTalk: Scalable graph-based dialogue synthesis for enhancing LLM conversational capabilities." In Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 658-677. 2025.
- Jin, Hongye, Pei Chen, Jingfeng Yang, Zhengyang Wang, Meng Jiang, Yifan Gao, **Binxuan Huang** et al. "END: Early Noise Dropping for Efficient and Effective Context Denoising." arXiv preprint arXiv:2502.18915 (2025).
- Zhuang, Yuchen, Jingfeng Yang, Haoming Jiang, Xin Liu, Kewei Cheng, Sanket Lokegaonkar, Yifan Gao et al. "Hephaestus: Improving fundamental agent capabilities of large language models through continual pre-training." ACL (2025).
- Chen, Yangyi, **Binxuan Huang**, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. "Scaling laws for predicting downstream performance in llms." TMLR (2025).
- Cheng, Kewei, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, **Binxuan Huang**, Ruirui Li, Shiyang Li et al. "Inductive or deductive? rethinking the fundamental reasoning abilities of llms." ACL 2024 Workshop on Natural Language Reasoning and Structured Explanations (2024).
- Huang, Zijie, Daheng Wang, **Binxuan Huang**, Chenwei Zhang, Jingbo Shang, Yan Liang, Zhengyang Wang et al. "Concept2Box: Joint geometric embeddings for learning two-view knowledge graphs." ACL (2023).
- Cheng, Kewei, Xian Li, Zhengyang Wang, Chenwei Zhang, **Binxuan Huang**, Yifan Ethan Xu, Xin Luna Dong, and Yizhou Sun. "Tab-cleaner: Weakly supervised tabular data cleaning via pre-training for E-commerce catalog." ACL (2023).

- Sarkhel, Ritesh, **Binxuan Huang**, Colin Lockard, and Prashant Shiralkar. "Self-training for label-efficient information extraction from semi-structured web-pages." Proceedings of the VLDB Endowment 16, no. 11 (2023): 3098-3110.
- Xiang Deng, Prashant Shiralkar, Colin Lockard, **Binxuan Huang**, Huan Sun. "DOM-LM: Learning Generalizable Representations for HTML Documents." arXiv preprint arXiv:2201.10608 (2022).
- Tai-Long He, Dylan BA Jones, Kazuyuki Miyazaki, **Binxuan Huang**, Yuyang Liu, Zhe Jiang, E Charlie White, Helen M Worden, John R Worden. "Deep learning to evaluate US NOx emissions using surface ozone predictions." In Journal of Geophysical Research: Atmospheres 127, no. 4 (2022).
- Daheng Wang, Prashant Shiralkar, Colin Lockard, **Binxuan Huang**, Xin Luna Dong, Meng Jiang. "TCN: Table Convolutional Network for Web Table Interpretation." In Proceedings of the Web Conference 2021, pp. 4020-4032. 2021.
- **Binxuan Huang**, and Kathleen M. Carley. "Disinformation and Misinformation on Twitter during the Novel Coronavirus Outbreak." arXiv preprint arXiv:2006.04278 (2020).
- **Binxuan Huang**, Han Wang, Tong Wang, Yue Liu, Yang Liu. "Entity Linking for Short Text Using Structured Knowledge Graph via Multi-grained Text Matching." INTERSPEECH, 2020
- Sumeet Kumar, **Binxuan Huang**, Ramon Alfonso Villa Cox, and Kathleen M. Carley. "An anatomical comparison of fake-news and trusted-news sharing pattern on Twitter." Computational and Mathematical Organization Theory, 2020
- **Binxuan Huang** and Kathleen M. Carley. "Discover Your Social Identity from What You Tweet: A Content Based Approach." In Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities, 2020.
- **Binxuan Huang**, and Kathleen M. Carley. "Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
- Tai-Long He, Dylan Jones, **Binxuan Huang**, Yuyang Liu, Kazuyuki Miyazaki, Zhe Jiang, E. Charlie White, Helen M. Worden, and John R. Worden. "Recurrent U-net: Deep learning to predict daily summertime ozone in the United States." arXiv preprint arXiv:1908.05841 (2019).
- **Binxuan Huang**, and Kathleen M. Carley. "A Hierarchical Location Prediction Neural Network for Twitter User Geolocation." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.
- **Binxuan Huang**, and Kathleen M. Carley. "A Large-Scale Empirical Study of Geotagging Behavior on Twitter." In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2019
- **Binxuan Huang**, and Kathleen M. Carley. "Residual or Gate? Towards Deeper Graph Neural Networks for Inductive Graph Representation Learning." NeurIPS Graph Representation Learning Workshop, 2019.
- **Binxuan Huang**, and Kathleen M. Carley. "Location Order Recovery in Trails with Low Temporal Resolution." IEEE Transactions on Network Science and Engineering (TNSE) ,2018.
- **Binxuan Huang**, and Kathleen Carley. "Parameterized convolutional neural networks for aspect level sentiment classification." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

- **Binxuan Huang**, Yanglan Ou, and Kathleen M. Carley. "Aspect level sentiment classification with attention-over-attention neural networks." In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), 2018.
- **Binxuan Huang**, and Kathleen M. Carley. "On predicting geolocation of tweets using convolutional neural networks." In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), 2017.
- Yu Zhang, Wei Wei, **Binxuan Huang**, Kathleen M. Carley, and Yan Zhang. "RATE: Overcoming Noise and Sparsity of Textual Features in Real-Time Location Estimation." In Proceedings of the Conference on Information and Knowledge Management (CIKM), 2017
- Felicia Natali, Kathleen M. Carley, Feida Zhu, and **Binxuan Huang**. "The role of different tie strength in disseminating different topics on a microblog." In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2017
- Yujie Qian, Jie Tang, Zhilin Yang, **Binxuan Huang**, Wei Wei, and Kathleen M. Carley. "A probabilistic framework for location inference from social media." arXiv preprint arXiv:1702.07281 (2017).
- William Frankenstein, **Binxuan Huang**, and Kathleen M. Carley. "NATO Trident Juncture on Twitter: Public Discussion." Available at SSRN 2720320 (2016).

Talks

- A Large-Scale Empirical Study of Geotagging Behavior on Twitter, ASONAM 2019
- Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks, Amazon Alexa AI, 2019/06
- Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks, SBP-BRiMS 2018
- A Large-Scale Empirical Study of Geotagging Behavior on Twitter, CASOS Summer Institute, 2018/06
- On Predicting Geolocation of Tweets Using Convolutional Neural Networks, SBP-BRiMS 2017

Teaching

Teaching Assistant, Dynamic Network Analysis	*Spring, 2017 & 2018*
Teaching Assistant, CASOS Summer Institute	*June, 2016 & 2017 & 2018*
Teaching Assistant, Introduction to Computing System	*Summer, 2014*

Professional Service

- Area Chair: Amazon Machine Learning Conference (2023, 2024)
- Reviewer/PC Member: COLM (2024), KDD (2021, 2023, 2024), CIKM (2021), ACL (2020), ICWSM (2019, 2020), Web Science Conference (2020), Journal of Computational Social Science (2020), Computational and Mathematical Organization Theory (2018, 2020), NAACL (2019), SBP-BRiMS (2018, 2019), IEEE Intelligent Systems (2019), IEEE Transactions on Network Science and Engineering (2018), Journal of Neural Network (2019)

Awards and Honors

SBP-BRiMS Travel Grant	*2017 & 2018*
GuSH Research Grant Awards	*2016*
National Scholarship of China (Top 2%, twice)	*2012 & 2013*
First-Class Scholarship for Outstanding Students (Top 3%, twice)	*2012 & 2013*
First-Class Scholarship for Outstanding Merits (Top 3%, twice)	*2012 & 2013*
Excellent Student Awards (Top 3%)	*2013*
First Prize of the National Talents Training Base (Top 3%)	*2012*
Scholarship for Excellence in Arts and Sports	*2012*

Technical Skills

- Languages: Python(Extensive), C/C++, Java, Matlab
- ML/DL: PyTorch, TensorFlow, LLM pretraining & fine-tuning, MoE architectures, graph neural networks, representation learning
- Data & Systems: Spark, large-scale distributed training, high-throughput data pipelines, web-scale data processing