Zhefu Shi

Email: shizhefu3@gmail.com **Personal Cell:** 425-761-7871
LinkedIn Website: https://www.linkedin.com/in/zhefushi/

SUMMARY

- **Solid Academic Accomplishment:**

1) Ph.D. in Computer Science; Master in Math/Statistics.
2) Research: Mathematical Modeling, Optimization. 11 publications, 300+ citations.

- **Solid Professional Experience:**

1) Worked in 3 domains: a) Applying Machine Learning in Search/Ads/Recommendation; b) ML Infrastructure; c) Cloud Computing platform.
2) End to end deliver Machine Learning models to production.

PROFESSIONAL EXPERIENCE

Research Institute, HongKong 2024/01 – 2025/06
Title: Software Architect in AI and ML

- LLM inference platform performance improvement 30%
- Ship an automated optimization system for LLM inference, widely being used by tech companies
- Profiling LLM inference traffic, build ML models for prefilling/decoding performance optimization, within TTFT/TPOT latency constraints. This is applicable for different LLM inference platforms.
- LLM training instability problems root cause analysis and solution proposal
- LLM kernel development, solving memory corruption caused kernel calculation inaccuracy
- LLM kernel Quantization, implement and optimize Duquant and Hadamard transpose algorithms, improved kernel performance, without losing accuracy

Microsoft, Mountain View 2021/09 – 2023/12
Title: Principal Applied Scientist

- Tech lead for Shopping Recommendation and Ads projects, with teams from US, Beijing, India, etc.
- Significantly improve CTR and DAU, e.g., 10% on en-us, and 30% increase on international markets.
- Recommendation L1/L2 state of the art (SOTA) DL models training, feature engineering, etc.
- Persists in optimizing SOTA DL models, e.g., DeepFM/DIN models, applying Multi Modality models to combine text/image features in L1/L2 process.
- Recommendation L0 index improvement. International markets 10% CTR gain.
- Significantly reduce incorrect contents in Shopping recommendation, 90% reduction, by using Multi Modality models
- Design personal shopping assistant agent, applying Chain of Thought (CoT) and Tree of Thought (ToT) in agent, combining with RAG, for personalized shopping assistant
- Consistent research applying LLMs in conversational AI system design, evaluation, and ML ops.

Environment: Python, TensorFlow, PyTorch, C#

Coupang, Mountain View **2018/09 – 2021/09**

Title: Principal Machine Learning Engineer

- Search and Ads DL model training, feature engineering, online A/B testing, online model serving.
- Search and Ads online serving framework build with high scalability
- L0 index and evaluation data pipeline build up. Data quality monitor dashboard build up.
- Constantly improve GMV, Revenue, Profit significantly, e.g., experiments constantly have tens of millions $ per year gain.
- Solved model performance significantly different between Online A/B test and Offline evaluation

-Amazon-A9, Palo Alto **2015/09 – 2018/03**

Title: Software Engineer, Search Relevance Core Ranking Team.

- Amazon search relevance Machine Learning model (GBDT, XGBoost, etc.).
  Introducing significant profit gain in millions of dollars.
- Information Retrieval and Feature extraction.
  Owner of a data pipeline of which output is used by 95% of Amazon relevance ranking models.
- Search engine optimization.
  Optimizing query traffic to search engine and reducing engine workloads by 10%.
- ML model training pipeline development and optimization.
  Reducing model training manual workloads by 20%, and model training time by 20%.

Environment: C++, Python, Java, TensorFlow, Scikit Learn, Keras, Linux, Hadoop, Spark.

Microsoft Corporation, Redmond **2014/04 – 2015/09**

Title: Software Engineer

- Power Business Intelligence (PBI).

1) Auto insights engine design and development. This engine automatically extracts/analyzes useful signals in data.
2) Implementation of ML algorithms in PBI engine. Reducing manual workloads by 20%.

- Azure Cloud Computing Platform, Compute Core Team (fundamental team in Azure)
  Design and develop Must-Ship key components: Computer Resource Provider, Billing pipelines.

Bloomberg, NYC **2012/10–2014/04**

Title: Senior Software Engineer

- Machine Learning (ML) and Natural Language Processing (NLP).

1) Design and development of finance news/data search solutions in timing critical finance domain.
2) Design and development of solutions such as query auto suggestions and auto complete.

Microsoft Corporation, Redmond **2010/01–2012/10**

Title: Software Engineer

- ML and NLP. Implementation of ML Algorithm, e.g., SVM, to office product.
- Cloud computing technology, virtual network.

EDUCATION

University of Missouri-Kansas City, USA

Ph.D. in Computer Science GPA: 3.96
Coordinating: Telecom and Computer Networking.
Research: Stochastic Process and Markov chain
Co-discipline: Math and Statistics
Master of Science in Math and Statistics GPA: 3.95 Graduated in 2009/12
Master of Science in Computer Science GPA: 3.98 Graduated in 2005/7

PUBLICATIONS (Totally 11 Publications, 300+ Citations. Selected Publications.)

Research focuses on Stochastic Process, Markov Chain and Optimization. All could be applied in ML and AI.

Book Published:

- **Zhefu Shi.** Stochastic modeling, correlation, competition, and cooperation in a CSMA wireless network. ProQuest, UMI Dissertation Publishing, ISBN: 9781243709219.

Book Chapter:

- **Zhefu Shi,** Cory Beard. Chapter: **QoS In The Mobile Cloud Computing**. Published in Book: Mobile Computing over Cloud: Technologies, Services, and Applications.

Papers:

- **Zhefu Shi,** Cory Beard, Ken Mitchell. **Analytical Models for Understanding Misbehavior and MAC Friendliness in CSMA Networks**. Performance Evaluation Archive Volume 66 Issue 9-10, September 2009.
- **Zhefu Shi,** Cory Beard, Ken Mitchell. **Analytical Models for Understanding Space, Backoff and Flow Correlation in CSMA Wireless Networks**. Wireless Networks, Springer, 2012.
- **Zhefu Shi,** Cory Beard, Ken Mitchell. **Competition, Cooperation, and Optimization in Multi-Hop CSMA Networks with Correlated Traffic**. International Journal of Next-Generation Computing 3
- **Zhefu Shi,** Cory Beard, Ken Mitchell. **Misbehavior and MAC friendliness in CSMA networks**. IEEE Wireless Communications and Networking Conference (WCNC) 2007.