# Zifeng Wang

Google, 12181 Bluff Creek Drive, Playa Vista, CA 90094

zifengwangking@gmail.com kingspencer.github.io Google Scholar

## About Me

I am a Senior Research Scientist at Google Cloud AI Research, where I specialize in large language models (LLMs) and their applications in real world settings. My current work centers on multi-LLM / agent collaboration, synthetic data generation for post training and evaluation, and cost effective inference. During my PhD at Northeastern University, I investigated continual learning and a range of machine learning applications. I hold a bachelor's degree from Tsinghua University.

## Education

**Northeastern University** Boston, MA

Ph.D. in Computer Engineering, GPA: 4.0/4.0 Sep 2018 – Aug 2023

- **Advisor**: Prof. Jennifer Dy
- **Thesis**: "Effective and Efficient Continual Learning"

**Tsinghua University** Beijing, China

B.Eng. in Electronic Engineering, GPA: 92/100 (*top 5% out of 233*) Aug 2014 - July 2018

## Work Experience

**Cloud AI Research, Google** Sunnyvale, CA

*Research Scientist* July 2023 - Present

- Initiated and led multiple research projects on large language models:
  - Multi-LLM / Agent collaboration and optimization [ICML25, ArXiv]
  - Synthetic data generation for post-training (*e.g.*, instruction-following, tool use) [NAACL24, ACL25]
  - Cost-effective LLM inference (*e.g.*, distillation, routing) [ICLR25, ArXiv]
  - Reasoning (*e.g.*, structured data reasoning, backward reasoning) [NeurIPS24, NAACL25]
- 15+ papers published at top-tier ML and NLP venues.

**Cloud AI Research, Google** Remote / Sunnyvale, CA

*Research Intern; Advised by: Zizhao Zhang, Vincent Perot, Jacob Devlin* May 2022 – Jan 2023

- QueryForm: A zero-shot document entity extraction (DEE) Framework [ACL23].
- Large-scale webpage-based pre-training for document understanding.

**Cloud AI Research, Google** Remote / Sunnyvale, CA

*Research Intern; Advised by: Zizhao Zhang, Chen-Yu Lee* June 2021 – May 2022

- Proposed the first prompting-based continual learning framework.
  - Learning to prompt for continual learning [CVPR22].
  - DualPrompt [ECCV22].
- Open-sourced prompting-based continual learning framework in JAX [GitHub, 400+ stars].

## Academic Experience

**Machine Learning Group, Northeastern University** Boston, MA

*Graduate Research Assistant; Advised by: Prof. Jennifer Dy* Sep 2018 – July 2023

- Effective and efficient continual learning [NeurIPS22, ICML23].
- Robust learning of neural networks [NeurIPS21, ICDM20].

**Channing Laboratory, Harvard Medical School** Boston, MA

*Student collaborator; Advised by: Prof. Peter J. Castaldi, Prof. Jennifer Dy* Sep 2018 – Sep 2021

- Smoking prediction via isoform-aware RNA-seq deep learning models [PLOS Computational Biology].

**i-Vision Group, Tsinghua University** Beijing, China

*Undergrad Research Assistant; Advised by: Prof. Jiwen Lu* Sep 2017 – Mar 2018

- Multi-object tracking via deep reinforcement learning [ECCV18].

## Selected Publications (Google Scholar)

(* *indicates equal contribution*)

**Preprints:**

[1] Shangbin Feng, **Zifeng Wang**, Palash Goyal, Yike Wang, Weijia Shi, Huang Xia, Hamid Palangi, Luke Zettlemoyer, Yulia Tsvetkov, Chen-Yu Lee, Tomas Pfister. "Heterogeneous Swarms: Jointly Optimizing Model Roles and Weights for Multi-LLM Systems". *ArXiv 2025*.

[2] Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, **Zifeng Wang**, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, Sanjiv Kumar. "Universal Model Routing for Efficient LLM Inference". *ArXiv 2025*.

[3] Shangbin Feng, Wenxuan Ding, Alisa Liu, **Zifeng Wang**, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, Yulia Tsvetkov. "When One LLM Drools, Multi-LLM Collaboration Rules". *ArXiv 2025*.

**Conference Papers:**

[4] Shangbin Feng, **Zifeng Wang**, Yike Wang, Sayna Ebrahimi, Hamid Palangi, Lesly Miculicich, Achin Kulshrestha, Nathalie Rauschmayr, Yejin Choi, Yulia Tsvetkov, Chen-Yu Lee, Tomas Pfister. "Model Swarms: Collaborative Search to Adapt LLM Experts via Swarm Intelligence". *ICML 2025*.

[5] Fan Yin, **Zifeng Wang**, I-Hung Hsu, Jun Yan, Ke Jiang, Yanfei Chen, Jindong Gu, Long T. Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, Tomas Pfister. "Magnet: Multi-turn Tool-use Data Synthesis and Distillation via Graph Translation". *ACL 2025*.

[6] Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, **Zifeng Wang**, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, Tomas Pfister. "In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents". *ACL 2025*.

[7] Justin Chih-Yao Chen, **Zifeng Wang**, Hamid Palangi, Rujun Han, Sayna Ebrahimi, Long Le, Vincent Perot, Swaroop Mishra, Mohit Bansal, Chen-Yu Lee, Tomas Pfister. "Reverse Thinking Makes LLMs Stronger Reasoners". *NAACL 2025*.

[8] Wenda Xu, Rujun Han, **Zifeng Wang**, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, Tomas Pfister. "Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling". *ICLR 2025*.

[9] Zilong Wang, **Zifeng Wang**, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, Tomas Pfister. "Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting". *ICLR 2025*.

[10] Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, **Zifeng Wang**, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, Boqing Gong. "OmnixR: Evaluating Omni-modality Language Models on Reasoning across Modalities". *ICLR 2025*.

[11] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, **Zifeng Wang**, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, Tomas Pfister. "Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding". *ICLR 2024*.

[12] **Zifeng Wang**, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, Tomas Pfister. "CodecLM: Aligning Language Models with Tailored Synthetic Data". *Findings of NAACL 2024*.

[13] **Zifeng Wang**, Zizhao Zhang, Jacob Devlin, Chen-Yu Lee, Guolong Su, Hao Zhang, Jennifer Dy, Vincent Perot, Tomas Pfister. "QueryForm: A Simple Zero-shot Form Entity Query Framework". *Findings of ACL 2023*.

[14] **Zifeng Wang**, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy. "DualHSIC: HSIC-Bottleneck and Alignment for Continual Learning". *ICML 2023*.

[15] **Zifeng Wang***, Zheng Zhan*, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, Jennifer Dy. "SparCL: Sparse Continual Learning on the Edge". *NeurIPS 2022*.

[16] **Zifeng Wang**, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, Tomas Pfister. "DualPrompt: Complementary Prompting for Rehearsal-free Continual Learning". *ECCV 2022*.

[17] **Zifeng Wang**, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, Tomas Pfister. "Learning to Prompt for Continual Learning". *CVPR 2022*.

[18] Tong Jian*, **Zifeng Wang***, Yanzhi Wang, Jennifer Dy, Stratis Ioannidis. "Pruning Adversarially Robust Neural Networks without Adversarial Examples". *ICDM 2022*.

[19] **Zifeng Wang***, Tong Jian*, Aria Masoomi, Stratis Ioannidis, Jennifer Dy. "Revisiting Hilbert-Schmidt Information Bottleneck for Adversarial Robustness". *NeurIPS 2021*.

[20] **Zifeng Wang***, Tong Jian*, Kaushik Chowdhury, Yanzhi Wang, Jennifer Dy, Stratis Ioannidis. "Learn-Prune-Share for Lifelong Learning". *ICDM 2020*.

[21] **Zifeng Wang**, Batool Salehi, Andrey Gritsenko, Kaushik Chowdhury, Stratis Ioannidis, Jennifer Dy. "Open-World Class Discovery with Kernel Networks". *ICDM 2020*. **Best Paper Candidate**.

[22] Aria Masoomi, Chieh Wu, Tingting Zhao, **Zifeng Wang**, Peter Castaldi, Jennifer Dy. "Instance-wise Feature Grouping". *NeurIPS 2020*.

[23] Andrey Gritsenko*, **Zifeng Wang***, Jennifer Dy, Kaushik Chowdhury, Stratis Ioannidis. "Finding a 'New' Needle in the Haystack: Unseen Radio Detection in Large Populations Using Deep Learning". *DySPAN 2019*, **Best Paper Award**.

[24] Liangliang Ren, Jiwen Lu, **Zifeng Wang**, Qi Tian, Jie Zhou. "Collaborative Deep Reinforcement Learning for Multi-Object Tracking". *ECCV 2018*.

**Journal Papers:**

[25] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, **Zifeng Wang**, Sayna Ebrahimi, Hao Wang "Continual Learning of Large Language Models: A Comprehensive Survey". *ACM Computing Surveys (2025)*.

[26] Ruoxi Sun, Sercan Ö Arik, Alex Muzio, Lesly Miculicich, Satya Gundabathula, Pengcheng Yin, Hanjun Dai, Hootan Nakhost, Rajarishi Sinha, **Zifeng Wang**, Tomas Pfister. "SQL-PaLM: Improved Large Language Model Adaptation for Text-to-SQL". *Transactions on Machine Learning Research (2024)*.

[27] **Zifeng Wang**, Aria Masoomi, Zhonghui Xu, Adel Boueiz, Sool Lee, Tingting Zhao, Russell Bowler, Michael Cho, Edwin K. Silverman, Craig Hersh, Jennifer Dy, Peter J. Castaldi. "Improved Prediction of Smoking Status via Isoform-Aware RNAseq Deep Learning Models". *PLoS Computational Biology 17(10): e1009433*.

[28] Tingting Zhao*, **Zifeng Wang***, Aria Masoomi, Jennifer Dy. "Deep Bayesian Unsupervised Lifelong Learning". *Neural Networks 149, 95–106*.

[29] Tong Jian, Yifan Gong, Zheng Zhan, Runbin Shi, Nasim Soltani, **Zifeng Wang**, Jennifer Dy, Kaushik Chowdhury, Yanzhi Wang, Stratis Ioannidis. "Radio Frequency Fingerprinting on the Edge". *IEEE Transactions on Mobile Computing 21(11), 4078–4093*.

## Honers and Awards

**Outstanding Graduate Student in Research**, Northeastern University, 2023 — Boston, MA
**Scholar Award**, NeurIPS 2022 — New Orleans, LA
**Best Paper Candidate**, ICDM 2020 — Sorrento, Italy
**Best Paper Award**, IEEE DySPAN 2019 — Newark, NJ
**Travel Award**, NeurIPS 2019 — Vancouver, Canada
**Dean's Fellowship**, Northeastern University, 2018 — Boston, MA
**Outstanding Undergraduate Scholarship**, Tsinghua University, 2016 — Beijing, China

## Academic Services

- **Area Chair**: ACL ARR
- **Conference Reviewer**: NeurIPS, ICML, ICLR, CVPR, ICCV
- **PC Member**: SDM
- **Journal Reviewer**: TPAMI, TMLR, Neural Networks

## Invited Talks

- *Effective and Efficient Continual Learning*
  - Centre for Frontier AI Research (CFAR), A*STAR, Remote, July 2023
- *SparCL: Sparse Continual Learning on the Edge*
  - ContinualAI, Remote, Feb 2023
- *QueryForm: A Simple Zero-shot Form Entity Query Framework*
  - Google Cloud AI Research, Sunnyvale, CA, Nov 2022
- *Prompting-based Continual Learning*
  - The AI Talks, Nanyang Technological University, Remote, Mar 2022
  - ContinualAI, Remote, Sep 2022
  - Google Cloud AI Research, Sunnyvale, CA, May 2022
- *Revisiting Hilbert-Schmidt Information Bottleneck for Adversarial Robustness*
  - INFORMS Annual Meeting, Indianapolis, IN, Oct 2022
  - AI Times, Tsinghua University, Remote, June 2022