Wei Fang

AI Research Scientist

weifang@alum.mit.edu | peskotiveswf.github.io | wfangtw | wei-fang-tw | Wei Fang

Education

Ph.D. in EECS, Massachusetts Institute of Technology *2026*

S.M. in EECS, Massachusetts Institute of Technology *2021*

B.S.E. in EE, National Taiwan University *2018*

Experience

Subconscious *Cambridge, MA, USA*

Founding AI Scientist *Apr. 2026 - Present*

- Efficient, long-horizon agentic inference with context and kv cache compression.
- Building and serving large-scale inference and post-training systems.
- Efficient vision-language model inference and post-training.
- Speculative decoding.

MIT CSAIL Spoken Language Systems Group *Cambridge, MA, USA*

Research Assistant, supervised by Dr. James R. Glass *Sep. 2018 - Mar. 2026*

- **Agentic Tool-using** Designed a zero-shot framework for tool instruction tuning to enhance LLM tool-using [*ACL25*] , and a lightweight query-planning module for SOTA multi-step tool retrieval [*ACL26*] .
- **Information Retrieval** Proposed a joint re-ranking method based on pointwise mutual information for improved passage re-ranking [*EACL24*] , and query expansion ranking to boost open-domain QA [*ACL23*] .
- **Factuality & Verification** Factual and fair language generation with self-verification [*ASRU23W*] ; Automatic fact-checking system [*NAACL19*] ; Stance detection with multi-task transfer learning [*FEVER19*] .
- **Other** conversational QA [*ACL23W*] ; temporal information extraction [*Preprint*] ; speech synthesis [*Preprint*] .

Teaching Assistant *Feb. 2021 - May 2021*

- 6.864 Advanced Natural Language Processing

MIT-IBM Watson AI Lab *Cambridge, MA, USA*

Research Intern *Jun. 2024 - Aug. 2024*

- Collaboration with MIT-IBM researchers on the tool instruction optimization project [*ACL25*] .

Apple Inc. Frameworks NLP Team *Cupertino, CA, USA*

NLP Intern *Jul. 2018 - Aug. 2018*

- Developed a synthetic data generation pipeline using generative models for the QuickPath project.

NLP Intern *Jun. 2017 - Sep. 2017*

- Investigated multilingual input representations for neural models to enhance user keyboard experience.

NTU & Academia Sinica Vision & Learning Lab *Taipei, Taiwan*

Research Assistant, supervised by Prof. Yu-Chiang Frank Wang *Jul. 2016 - Jan. 2018*

- Incorporated knowledge graphs for multi-label zero-shot image recognition [*CVPR18*] .

National Taiwan University (NTU) Speech Lab *Taipei, Taiwan*

Research Assistant, supervised by Prof. Lin-Shan Lee & Prof. Hung-Yi Lee *Sep. 2015 - Jun. 2017*

- Proposed a hierarchical attention-based model for the TOEFL Listening Comprehension Test [*SLT16*] [*TASLP19*] .
- Researched on unsupervised audio word embeddings.

Teaching Assistant *Feb. 2016 - Jun. 2017*

- EE5177/EE5184 Machine Learning; CSIE5610 Data Analytics & Modeling; EE2011 Signals and Systems

Publications

[1] **Wei Fang**, James Glass. Beyond Single-Shot: Multi-step Tool Retrieval via Query Planning. *Findings of the Association for Computational Linguistics: ACL*, 2026.

[2] **Wei Fang**. Towards Scalable and Generalizable Agentic Large Language Models. *Ph.D. Thesis*, 2026.

[3] **Wei Fang**, Yang Zhang, Kaizhi Qian, James Glass, Yada Zhu. PLAY2PROMPT: Zero-shot Tool Instruction Optimization for LLM Agents via Tool Play. *Findings of the Association for Computational Linguistics: ACL*, 2025.

[4] **Wei Fang**, Yung-Sung Chuang, James Glass. Joint Inference of Retrieval and Generation for Passage Re-ranking. *Findings of the Association for Computational Linguistics: EACL*, 2024.

[5] Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, **Wei Fang**, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, James Glass. Interpretable Unified Language Checking. *Symposium for Celebrating 40 Years of Bayesian Learning in Speech and Language Processing and Beyond at ASRU*, 2023.

[6] Yung-Sung Chuang, **Wei Fang**, Shang-Wen Li, Wen-tau Yih, James Glass. Expand, Rerank, and Retrieve: Query Reranking for Open-domain Question Answering. *Findings of the Association for Computational Linguistics: ACL*, 2023.

[7] Tianhua Zhang, Liping Tang, **Wei Fang**, Hongyin Luo, Xixin Wu, Helen Meng, James Glass. ConvRGX: Recognition, Generation, and Extraction for Self-trained Conversational Question Answering. *DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering at ACL*, 2023.

[8] **Wei Fang** and James Glass. Cross-Document Temporal Relation Extraction with Temporal Anchoring Events. *Preprint*, 2021.

[9] **Wei Fang**. On End-to-end Automatic Fact-checking Systems. *S.M. Thesis*, 2021.

[10] **Wei Fang**, Moin Nadeem, Mitra Mohtarami, James Glass. Neural Multi-task Learning for Stance Prediction. *EMNLP-IJCNLP Workshop on Fact Extraction and VERification*, 2019.

[11] **Wei Fang**, Yu-An Chung, James Glass. Towards Transfer Learning for End-to-end Speech Synthesis from Deep Pre-trained Language Models. *arXiv Preprint*, 2019.

[12] Chia-Hsuan Lee, Hung-Yi Lee, Szu-Lin Wu, Chi-Liang Liu, **Wei Fang**, Jui-Yang Hsu, Bo-Hsiang Tseng. Machine Comprehension of Spoken Content: TOEFL Listening Test and Spoken SQuAD. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLPRO)*, 2019.

[13] Moin Nadeem, **Wei Fang**, Brian Xu, Mitra Mohtarami, James Glass. FAKTA: An Automatic End-to-End Fact Checking System. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[14] Chung-Wei Lee†, **Wei Fang**†, Chih-Kuan Yeh, Yu-Chiang Frank Wang. Multi-Label Zero-Shot Learning with Structured Knowledge Graphs. *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2018.

[15] **Wei Fang**†, Jui-Yang Hsu†, Hung-Yi Lee, Lin-Shan Lee. Hierarchical Attention Model for Improved Comprehension of Spoken Content. *IEEE Workshop on Spoken Language Technology (SLT)*, 2016.

† indicates equal contribution

Honors

Government Scholarship to Study Abroad, Ministry of Education, Taiwan — *May 2021*
Student Travel Bursary, Amazon — *Oct. 2019*
EECS Great Educators Fellowship, MIT EECS — *Feb. 2018*
Conference Grant, Foundation for the Advancement of Outstanding Scholarship — *Dec. 2016*
Dean's List, National Taiwan University — *2014-2016*

Professional Activity

Reviewing ACL Rolling Review (2021-), ACL (2021-), EMNLP (2021-), NeurIPS (2024-), ICML (2025-)
Affiliations Sigma Xi, the Scientific Research Honor Society (2025-)

Skills

Toolkits PyTorch, HuggingFace (Transformers, TRL), SGLang/vLLM/llama.cpp, PrimeRL/SkyRL, Nvidia Dynamo
ML Skills Inference optimization, Post training, Data generation, Agentic Search, KV Cache Compression