Jerry Wei

jerrywei@alumni.stanford.edu
https://www.jerrywei.net

Employment

2024 *Anthropic*
Member of Technical Staff (Alignment Science / Safeguards Research).
Apr. 2024 - present.

2023 *Google DeepMind*
Research Engineer (Gemini Post-training.).
Jul. 2023 - Apr. 2024.

2022 *Google Brain*
Student Researcher (LLM reasoning).
Nov. 2022 - Jul. 2023.

2022 *Meta*
Software Engineer Intern (Android Messenger).
Jun. 2022 - Sep. 2022.

Education

2021 *Stanford University*
Bachelor of Science, Computer Science, AI Track.
Sep. 2021 - Dec. 2022 (unfinished).
GPA: 3.89/4.00.

Selected Publications

2025 *Constitutional classifiers: defending against universal jailbreaks across thousands of hours of red teaming.*
{M. Sharma, M. Tong, J. Mu, J. Wei, J. Kruthoff, S. Goodfriend, E. Ong}, and 36 others.
arXiv preprint

2024 *Long-form factuality in large language models.*
{J. Wei, C. Yang, X. Song, Y. Lu}, and 8 others.
Conference on Neural Information Processing Systems (NeurIPS).

2023 *Simple synthetic data reduces sycophancy in large language models.*
J. Wei, D. Huang, Y. Lu, D. Zhou, and Q. V. Le.
arXiv preprint.

2023 *Symbol tuning improves in-context learning in language models.*
J. Wei, L. Hou, A. Lampinen, 3 others, X. Chen, Y. Lu, D. Zhou, T. Ma, and Q. V. Le.
Empirical Methods on Natural Language Processing (EMNLP).

2023 *Larger language models do in-context learning differently.*
J. Wei, J. Wei, Y. Tay, 3 others, X. Chen, H. Liu, D. Huang, D. Zhou, and T. Ma.
arXiv preprint.

2019 *Generative image translation for data augmentation in colorectal histopathology images.*
J. Wei, A. Suriawinata, L. Vaickus, B. Ren, X. Liu, J. Wei, and S. Hassanpour.
ML4H Workshop at NeurIPS.

Code

2024 *Benchmarking long-form factuality in large language models.*
GitHub: 575+ stars, `https://github.com/google-deepmind/long-form-factuality`.

2023 *Simple synthetic data for reducing sycophancy in large language models.*
GitHub: 100+ stars, `https://github.com/google/sycophancy-intervention`.

Blog Posts

Jul. 2023 *Symbol tuning improves in-context learning in language models.*
Jerry Wei & Denny Zhou, Google AI Blog.

May 2023 *Larger language models do in-context learning differently.*
Jerry Wei & Denny Zhou, Google AI Blog.

Honors

2021 Regeneron Science Talent Search (STS) Semifinalist.
2021 National Merit Scholarship Semifinalist & Leidos Corporation Scholarship Winner.
2019 Spotlight presentation at the NeurIPS ML4H Workshop.