Jeffrey Wu

Email me@wuthejeff.com
Website wuthejeff.com
Github user WuTheFWasThat
Google Scholar user x4JAvwMAAAAJ

Summary: Researcher and full-stack engineer interested in keeping the world safe for everyone as technologies such as machine learning become more powerful

---

Work Experience

- **OpenAI** Research engineer — Aug 2018 – Jul 2024
  AI safety research. Worked on language modeling (including training GPT-2) and RLHF (trained initial GPT instruct series models). Managed teams and projects working on scalable oversight, generalization, and interpretability.
- **Google Research** Software engineer — Oct 2016 – Aug 2018
  Built general infrastructure for supporting models for personalization from cross-product user history (basically learning giant embedding spaces for all users, Chrome pages, Youtube videos, etc). Experimented with RNN models to replace bag-of-words models, and contributed to launch of news feed trends personalization.
- **Terminal.com** Founding engineer — Jan 2013 – Oct 2016
  Building cloud-based container infrastructure, for scientific computing and online education. Helped design and implement many core systems across the stack and oversaw their security and scalability. Saw company grow from 2 to 12, and managed a small team of engineers. Interfaced with clients, including Crunchbase, Stanford University, Codecademy, and Udacity. Company was sold to Udacity.
- **Probabilistic Computing Project** Master's student — Nov 2011 – Jan 2013
  Implemented a probabilistic programming language. Explored a new Gibbs sampling algorithm to make inference more efficient in very general settings. Work presented [at NIPS 2012 probabilistic programing workshop]. [Source code] and [thesis].

Research

I'm broadly interested in training AI systems that are honest and kind to humans, and assistive systems that make humans smarter or wiser.

- **Training large language models** I scaled up OpenAI's largest language models 100x (from 110M parameters to 12B parameters). This included training GPT-2 [blog] and early stepping stones towards GPT-3 (NeurIPS 2020 best paper award),
- **Reinforcement learning from human feedback** I worked on helping language models learn from human feedback and specified human values. Early proofs of concept trained models that produced summaries better than human-written ones. This culminated in instruction following models (InstructGPT), which were the predecessor to ChatGPT.

- **Alignment in the superhuman regime** I managed projects working on studying ML in the regime where humans cannot evaluate model outputs. Our work on scalable oversight investigated whether models trained to self-critique could strengthen human supervision. Our work on generalization investigated whether models could behave as intended despite weak supervision.
- **Interpretability at scale** I managed projects working on understanding how large language models work. We worked on disentangling building blocks at scale and automating their interpretations.

Side Projects

I enjoy building software. Here is a selection of projects I worked on mainly for fun. For more, see my website.

- **Vimflowy** Note taking tool inspired by Vim and Workflowy. [Demo] [Source (Typescript)]
- **Hanabi simulation** Game engine for simulating hanabi strategies, and state of the art bots. Cited in DeepMind work and interviewed for in WSJ. [Source (Rust)]
- **Send A Damned Message** A simple puzzle game. [Play here] [Source (ReasonML)]

Education

Massachusetts Institute of Technology **Cumulative GPA**: 4.8/5
B.S. in Mathematics, **B.S.** in Computer Science May 2012
M.Eng. in Computer Science January 2013

Skills

- Deep learning and machine learning (especially language modeling, RL, reward learning)
- Front end development
- Devops, e.g. linux, cloud platforms, containers, databases
- Algorithms and distributed systems design
- CS theory (e.g. complexity theory, cryptography)
- Mathematics (2006-2008 USAMO, 2010 Putnam top 200)

Personal qualities

- I enjoy working with people who share my mission, and who bring some thinking that I don't
- I like to keep an eye on the big picture, and I tend to think abstractly/idealistically
- I strive to act with integrity
- I try to be open-minded