

# Jeffrey Wu

Email [me@wuthejeff.com](mailto:me@wuthejeff.com)  
Github [user WuTheFWasThat](#)

Website [wuthejeff.com](http://wuthejeff.com)  
Google Scholar [user x4JAvwMAAAAJ](#)

**Summary:** Machine learning researcher and full-stack engineer interested in keeping the world safe as technologies such as machine learning become more powerful.

---

## Work Experience

- **OpenAI** Research engineer Aug 2018 – Present  
Focused on aligning models to human values. Researched language modeling (including training GPT-2, a state-of-the-art model in 2019) and reward modeling (trained initial beta version of the instruct series).
  - **Google Research** Software engineer Oct 2016 – Aug 2018  
Built general infrastructure (data pipelines, libraries, custom Tensorflow ops, a fun Lisp-like DSL) 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 helped launch news feed personalization experiments. In 20% time, studied properties of generalization error.
  - **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 programming workshop]. [Source code] and [thesis].
- 

## Research

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

- **Large-scale language model training** I trained GPT-2 (blog) and the early iterations of GPT-3 (NEURIPS 2020 best paper award), scaling up OpenAI's largest models from 110M parameters to over 6B parameters.
- **Reinforcement learning from human-specified values** I have been working on the problem of learning from humans. Early work trained models that produced summaries better than human-written ones. I then worked on book summarization and instruction following.

---

## Selected Side Projects

- **Vimflowy** Vim inspired outlining tool with many features. [Source] (Typescript) and [Demo].
- **Hanabi simulation** Game engine for simulating hanabi strategies, and state of the art bots. Cited in [DeepMind/Brain paper] and subsequently [interviewed for WSJ]. [Source] (Rust).
- **Send A Damned Message** A simple puzzle game (Play here!)

---

## Education

Massachusetts Institute of Technology  
B.S. in Mathematics, B.S. in Computer Science  
M.Eng. in Computer Science

Cumulative GPA: 4.8/5  
May 2012  
January 2013

---

## Skills

- Deep learning and machine learning (especially language modeling, RL, reward learning)
- Front end
- 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)
- Keeping an eye on the big picture
- Acting with integrity