Hugh Zhang

Last updated: October 9, 2024

hughbzhang@gmail.com
https://hughbzhang.com

Work Experience

Scale AI
Senior Research Scientist — *Oct 2024 - present*
Research Engineer — *Jan 2024 - Oct 2024*

Google DeepMind
Research Intern — *Nov 2023 - Jan 2024*

Facebook AI Research
Visiting Research Engineer (with Noam Brown) — *Jul 2020 - Jul 2021*

Google Brain
Research Intern (with Daniel Duckworth and Arvind Neelakantan) — *Jul 2019 - Mar 2020*

Stanford NLP Group
Research Intern (with Percy Liang and Tatsunori Hashimoto) — *Jul 2018 - Jul 2019*

Asana
Software Engineer — *Sep 2015 - Sep 2016*

Research (braces denote equal contribution)

Planning In Natural Language Improves LLM Search For Code Generation.
Evan Wang, 8 others, and **Hugh Zhang**.
Arxiv 2024.

A Careful Examination of Large Language Model Performance on Grade School Arithmetic.
Hugh Zhang, 12 others, and {Michele Lunati, Summer Yue}.
NeurIPS D&B 2024 (Spotlight).

LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet.
Nathaniel Li, 7 others including **Hugh Zhang**, and Summer Yue.
Arxiv 2024.

Learning Goal-Conditioned Representations for Language Reward Models.
{Vaskar Nath, Dylan Slack}, Jeff Da, Yuntao Ma, **Hugh Zhang**, and {Spencer Whitehead, Sean Hendryx}.
NeurIPS 2024.

NATURAL PLAN: Benchmarking LLMs on Natural Language Planning.
Huaixiu Steven Zheng, and 9 others including **Hugh Zhang**, and Denny Zhou.
Arxiv 2024.

Q-Probe: A Lightweight Approach to Reward Maximization for Language Models.
Kenneth Li, Samy Jelassi, **Hugh Zhang**, Sham Kakade, Martin Wattenberg, David Brandfonbrener.
Arxiv 2024.

Easy as ABCs: Unifying Boltzmann Q-Learning and Counterfactual Regret Minimization.
{Luca D'Amico Wong, **Hugh Zhang**}, and David C. Parkes.

Arxiv 2023.

Chain-of-Thought Reasoning is a Policy Improvement Operator.
Hugh Zhang, David C. Parkes.
ITIF Workshop at NeurIPS 2023.

No-regret Learning Dynamics for Sequential Correlated Equilibria.
Hugh Zhang.
AAMAS 2023 (Extended Abstract).

Human-level play in the game of Diplomacy by combining language models with strategic reasoning.
Meta FAIR Diplomacy Team and {25 others including **Hugh Zhang**}.
Science 2022.

Equilibrium Finding in Matrix Games Via Greedy Regret Minimization.
Hugh Zhang, Adam Lerer, and Noam Brown.
AAAI 2022.

Trading Off Diversity and Quality in Natural Language Generation.
{**Hugh Zhang**, Daniel Duckworth}, Daphne Ippolito, and Arvind Neelakantan.
HumEval Workshop at EACL 2021.

A Simple Adaptive Procedure Converging to Forgiving Correlated Equilibria.
Hugh Zhang.
Senior Thesis (John G. Sobieski Award for Creative Thinking).

Unifying Human and Statistical Evaluation for Natural Language Generation.
{Tatsunori Hashimoto, **Hugh Zhang**}, and Percy Liang.
NAACL 2019 (Oral).

Generating Transferable Adversarial Examples via Smooth Max Ensembling.
Yuchen Zhang and 6 others including **Hugh Zhang**, and Percy Liang.
Adversarial Attacks And Defenses Competition at NeurIPS 2017.

Invited Talks

A Careful Examination of LLM Performance on Grade School Arithmetic
University of Washington - Online - June 2024
AI Base Camp - Healdsburg, CA - June 2024
Scale AI Webinar - Online - September 2024

Chain-Of-Thought Reasoning is a Policy Improvement Operator
Kempner Institute - Cambridge, MA - November 2023
Scale AI - San Francisco, CA - December 2023

Generative AI and the Future of AI Research
Harvard Cabot House - Cambridge, MA - October 2023

Easy as ABCs: Unifying Boltzmann Q-Learning and Counterfactual Regret Minimization
Qualification Exam - Cambridge, MA - May 2023

A Simple, Adaptive Procedure Converging to Forgiving Correlated Equilibria
Thesis Presentation - Stanford, CA - May 2020

Unifying Human and Statistical Evaluation for Natural Language Generation

NAACL Oral - Minneapolis, MN - June 2019
Oxford Applied and Theoretical Machine Learning Group - Oxford, UK - June 2019
Apple Siri Lab - Cambridge, UK - June 2019
Naver Machine Learning - Seoul, KR - June 2019
Seoul National University Vision and Learning Lab - Seoul, KR - July 2019

Education

Harvard University (advised by David Parkes) *PhD Candidate in Computer Science (on leave)*
2021 -

Stanford University (advised by Gabriel Carroll) *BA with Honors in Economics*
2016 - 2020

Honors

Go (also known as weiqi or baduk)

US Open, 10th place	2022
North American Master's Tournament, 6th place (tied)	2013
Youngest Team Member on the US Team at the World Mind Sports Games	2008

Other

Kempner Institute Graduate Research Fellowship	2023
National Science Foundation Graduate Research Fellowship	2022
John G. Sobieski Award for Creative Thinking (for my senior honors thesis in Economics)	2020
USA Computing Olympiad Finalist (top 24 high school students nationally)	2014

Mentorship

Evan Wang, Scale AI internship (2024)
Luca D'Amico-Wong, Harvard Undergraduate (2023)
Mason Meyer (currently OpenAI), Harvard Undergraduate Senior Thesis (2021)