Steven Y. Feng

Stanford University
PhD in Computer Science

syfeng@stanford.edu
styfeng.github.io
linkedin.com/in/steven-feng
https://github.com/styfeng

EDUCATION

- **Stanford University** — Stanford, United States
  PhD in Computer Science — 2022-Present
- **Carnegie Mellon University (CMU)** — Pittsburgh, United States
  Master of Language Technologies (MLT) – Fully-funded research master's in NLP — 2020-2022
- **University of Waterloo** — Waterloo, Canada
  Bachelor of Mathematics, Statistics Major, Computer Science Minor — 2015-2020
- **Wilfrid Laurier University** — Waterloo, Canada
  Bachelor of Business Administration, Finance Concentration — 2015-2020

PUBLICATIONS AND CONFERENCE PROCEEDINGS

- Emmy Liu, Varun Gangal, (…), Sachin Kumar, and **Steven Y. Feng**. "HalluWorld: A Controlled Benchmark for Hallucination via Reference World Models". *Accepted to **NeurIPS 2026***. https://arxiv.org/abs/2605.19341
- Emmy Liu, Varun Gangal, (…), Sachin Kumar, and **Steven Y. Feng**. "A Unified Definition of Hallucination: It's the World Model, Stupid!". *Proceedings of **International Conference on Machine Learning (ICML) 2026***. https://arxiv.org/abs/2512.21577
- Long Phan, (…), **Steven Y. Feng**, (…), Alexandr Wang, and Dan Hendrycks. "Humanity's Last Exam: A benchmark of expert-level academic questions to assess AI capabilities". ***Nature, 2026***. https://www.nature.com/articles/s41586-025-09962-4
- Bria Long, Robert Z. Sparks, (…), Alvin W. M. Tan, **Steven Y. Feng**, Auddithio Nag, (…), Daniel L. K. Yamins, and Michael C. Frank. "The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences." Proceedings of ***Cognitive Computational Neuroscience (CCN) 2025.*** https://openreview.net/forum?id=bbv4CA2noZ
- **Steven Y. Feng**, Noah D. Goodman, and Michael C. Frank. "Is Child-Directed Speech Effective Training Data for Language Models?" Proceedings of ***Empirical Methods in Natural Language Processing (EMNLP) 2024***. https://aclanthology.org/2024.emnlp-main.1231/
- **Steven Y. Feng**, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard Hovy. "CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models." Proceedings of ***European Chapter of the Association for Computational Linguistics (EACL) 2023.*** https://aclanthology.org/2023.eacl-main.24/
- Sedrick Scott Keh, **Steven Y. Feng***, Varun Gangal*, Malihe Alikhani, and Eduard Hovy. "PANCETTA: Phoneme Aware Neural Completion to Elicit Tongue Twisters Automatically." Proceedings of ***European Chapter of the Association for Computational Linguistics (EACL) 2023.*** https://aclanthology.org/2023.eacl-main.36/ (* Equal Contribution)
- Sedrick Scott Keh, Kevin Lu, Varun Gangal*, **Steven Y. Feng***, Harsh Jhamtani, Malihe Alikhani, and Eduard Hovy. "PINEAPPLE: Personifying INanimate Entities by Acquiring Parallel Personification data for Learning Enhanced generation." Proceedings of ***International Conference on Computational Linguistics (COLING) 2022***. https://aclanthology.org/2022.coling-1.547/ (* Equal Contribution)
- **Steven Y. Feng**, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. "Retrieve, Caption, Generate: Visual Grounding for Enhancing Commonsense in Text Generation Models." Proceedings of ***AAAI Conference on Artificial Intelligence 2022***. https://ojs.aaai.org/index.php/AAAI/article/view/21306

- Varun Gangal*, **Steven Y. Feng***, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. "NAREOR: The Narrative Reordering Problem." Proceedings of ***AAAI 2022***. https://ojs.aaai.org/index.php/AAAI/article/view/21309 (* Equal Contribution)
- **Steven Y. Feng**, Jessica Huynh, Chaitanya Prasad Narisetty, Eduard Hovy, and Varun Gangal. "SAPPHIRE: Approaches for Enhanced Concept-to-Text Generation." Proceedings of ***International Conference on Natural Language Generation (INLG) 2021*** **[Best Long Paper]**. https://aclanthology.org/2021.inlg-1.21/
- **Steven Y. Feng***, Varun Gangal*, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. "A Survey of Data Augmentation Approaches for NLP." Proceedings of ***Association for Computational Linguistics (ACL) 2021 Findings***. https://aclanthology.org/2021.findings-acl.84/ (* Equal Contribution)
- **Steven Y. Feng***, Varun Gangal*, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. "GenAug: Data Augmentation for Finetuning Text Generators." Proceedings of ***EMNLP 2020 Deep Learning Inside Out (DeeLIO) Workshop***. https://aclanthology.org/2020.deelio-1.4/ (* Equal Contribution)
- Aaron W. Li, Veronica Jiang*, **Steven Y. Feng***, Julia Sprague, Wei Zhou, and Jesse Hoey. "ALOHA: Artificial Learning of Human Attributes for Dialogue Agents." Proceedings of ***AAAI Conference on Artificial Intelligence 2020*** [Oral]. https://ojs.aaai.org/index.php/AAAI/article/view/6328/ (* Equal Contribution)
- **Steven Y. Feng***, Aaron W. Li*, and Jesse Hoey. "Keep Calm and Switch On! Preserving Sentiment and Fluency in Semantic Text Exchange." Proceedings of ***Empirical Methods in Natural Language Processing (EMNLP) 2019***. https://www.aclweb.org/anthology/D19-1272/ (* Equal Contribution)

PREPRINTS, ABSTRACTS, SUBMITTED & IN PREPARATION

- **Steven Y. Feng**, Noah Goodman, Michael C. Frank, Evan Hubinger, Paul C. Bogdan, Andrew Kyle Lampinen. "When Do Models Admit They Are Wrong? Failure Disclosure Is Unstable Under Reinforcement Learning". *Submitted to ICLR 2027.* arxiv.org/abs/2609.33220
- **Steven Y. Feng***, Shrimai Prabhumoye*, (…), Mohammad Shoeybi, and Bryan Catanzaro. "Maximize Your Data's Potential: Enhancing LLM Accuracy with Two-Phase Pretraining." Arxiv Preprint. https://arxiv.org/abs/2412.15285 (* Equal Contribution)
- **Steven Y. Feng**, Alvin W.M. Tan, and Michael C. Frank. "Baby Scale: Investigating Models Trained on Individual Children's Language Input". Arxiv Preprint. *Submitted to NAACL 2027*. https://arxiv.org/abs/2603.29522
- Linda Zeng, **Steven Y. Feng**, and Michael C. Frank. "Bringing Up a Bilingual BabyLM: Investigating Multilingual Language Acquisition Using Small-Scale Models". Arxiv Preprint. *Submitted to NAACL 2027*. https://arxiv.org/abs/2603.29552
- Patrick Peixuan Ye, **Steven Y. Feng**, Chen Shani, and Noah Goodman. "Lexical Bootstrapping for Uncertainty Estimation." *Submitted to NAACL 2027.*
- Karan Singh, Michael Yu, (…), Emmy Liu, and **Steven Y. Feng**. "To Memorize or to Retrieve: Scaling Laws for RAG-Considerate Pretraining". Arxiv Preprint. *Submitted to ACL Rolling Review (ARR)*. https://arxiv.org/abs/2604.00715
- Ryan Marten, (…), **Steven Y. Feng**, (…), Ludwig Schmidt. "Terminal Bench 4.0." Project website: https://www.tbench.ai/, blogpost: https://www.tbench.ai/news/terminal-bench-4-0/. *Paper in preparation.*
- Jennifer Hu, Alvin Wei Ming Tan, **Steven Y. Feng**, and Michael C Frank. "Language production is harder than comprehension for children and language models". Proceedings of ***Annual Meeting of the Cognitive Science Society (CogSci) 2025***. Abstract link: https://escholarship.org/uc/item/5rz8b9jg [paper in preparation].
- Karan Singh and **Steven Y. Feng**. "Stable Rankings Are Not Enough: Influence Dynamics and Data Selection in Language Models". *In preparation for ICML 2027.*

WORK EXPERIENCE

- **Anthropic** Bay Area (CA), United States
  Research Fellow *June – Sept. 2026*
  - LLM alignment, RL post-training, and reasoning research. Investigated selective behavioral instability under RL post-training, focusing on failure disclosure: whether a model honestly communicates that it cannot solve a problem. Traced this divergence to small numerical and stochastic differences, and showed that KL reference anchoring helps to stabilize reporting (paper).

- **Contextual AI** Mountain View (CA), United States
  Research Intern *June – Sept. 2025*
  - Enhanced the long-term memory capabilities of LLM-based agents.

- **NVIDIA** Santa Clara (CA), United States
  Research Intern *June – Sept. 2024*
  - LLM research team in the Applied Deep Learning Research group. Investigated optimal data selection, mixing, ordering, and curriculum learning strategies for pretraining LLMs to enhance their performance, robustness, and generalizability (paper).

- **Amazon** Sunnyvale (CA), United States
  Applied Scientist Intern *June – Sept. 2023*
  - Alexa AI team (LLM group). Investigated ideas to improve and generalize chain-of-thought (CoT) reasoning for LLMs.

GRANTS, GIFTS, AND RESEARCH FUNDING

- **NSF ACCESS Compute Grant [~$300,000]** 2026
  Awarded 4.5M in credits (~$300k in value of GPU hours) for research projects.
- **NVIDIA Academic Compute Grant [~$120,000]** 2026
- **Modal Academic Compute Grant [$15,000+]** 2026
- **Lambda Labs Academic Compute Grant [$15,000]** 2026
- **Google DeepMind Compute Grant [$16,300]** 2024-2025
  Awarded for a collaborative project with GDM: *"Studying transfer from formal domains via generalization from code in language models".*
- **Amazon Science Gift Fund [$100,000]** 2024-2025
  Gift fund from Amazon to work on impactful projects. Initiated the idea and collaboration after my Amazon Summer 2023 internship.
- **Microsoft Accelerating Foundation Models Research (AFMR) Program [$20,000]** 2023-2024
  Awarded by Microsoft to work on a research project titled *"Bridging the Data Gap Between Large Language Models and Human Children"* [listed on the AFMR website]. Lead student contributor and main grant writer. Relevant publication.

AWARDS AND SCHOLARSHIPS

- **Top Contributor – Humanity's Last Exam** [Scale.AI and Center for AI Safety] *2025*
  Awarded prize money and a leading authorship position for contributing a substantial portion of the most challenging questions in mathematics, computer science, and other domains. Our team ranked top 10 out of 1000+ contributors to this leading benchmark, designed to stump the most capable AI models and LLMs. See the paper (Nature, Arxiv), website/blog, NYTimes and Reuters articles.
- **Research Fellow – Bain Capital Ventures (BCV) [**https://bcvairesearchfellows.framer.website/] *2024*
- **Honorable Mention – Jane Street Graduate Fellowship** 2024
- **Stability Fellow** 2023
- **NSERC Postgraduate Scholarships – Doctoral (PGS D)** *2022*
  Prestigious NSF equivalent fellowship for the top Canadian researchers going into a PhD program.

- **NSERC Canada Graduate Scholarships – Doctoral (CGS D)** [Declined] *2022*
- **Best Long Paper Award – International Conference on Natural Language Generation (INLG) 2021** *2021*
- **CMU Graduate Research Fellowship** *2020-2022*
- **Jessie W.H. Zou Memorial Award for Excellence in Undergraduate Research (2020) – Honorable Mention** *2020*
  Among the top three undergraduate researchers in the Faculty of Mathematics at the University of Waterloo. Article here.
- **CRA Outstanding Undergraduate Researcher Award (2020) – Honorable Mention** *2020*
  Award for the top North American undergraduate researchers in computing. Article here.
- **NSERC Alexander Graham Bell Canada Graduate Scholarship – Master's (CGS M)** [Declined] *2020*
- **Vector Scholarship in Artificial Intelligence** [Declined] *2020*
- **Ontario Graduate Scholarship and Queen Elizabeth II Graduate Scholarship in Science & Technology** [Declined] *2020*
- **NSERC Undergraduate Student Research Award (USRA)** *Jan. – Apr. 2019, May – Aug. 2019*
- **President's and Math Endowment Fund Research Awards** *2019 - 2020*
- **EMNLP and AAAI Student Scholarships** *2019 - 2020*
- **University of Waterloo Alumni Scholarship** *2015 - 2020*
- **University of Waterloo President's Scholarship of Distinction** *2015*

TALKS AND INTERVIEWS

- "Fueling Intelligent AI: How Data Drives LLM Training." Invited keynote talk for the ***Global AI Pitch Summit Silicon Valley***. Jan. 2025.
  - Keynote speaker for one of the largest AI events in the Bay Area, attended by thousands of researchers, startup founders, VCs, and general AI enthusiasts. Discussed several of my recent works on data-centric AI for both small and large-scale LLM pretraining.
- "Exploring Optimal Data Use for Language Models at Different Scales." Talk for the ***Stanford NLP Group***. Nov. 2024.
- "Language Models: Emergent Abilities, Trends, and Applications." Invited talk for ***University of Connecticut***. April 2023. Video *here*.
- "A Survey of Data Augmentation Approaches for NLP". Invited talk with **Varun Gangal** for ***Google Research***. Aug. 2021. *Video here*.
- Guest (with Dr. **Eduard Hovy**) on the ***Data Exchange Podcast with Ben Lorica.*** July 2021. *Video here*.

TEACHING EXPERIENCE

- **Course Assistant – CS 22A: The Social & Economic Impact of Artificial Intelligence (Stanford University)** *Jan. – Mar. 2026*
- **Course Assistant – CS 349H: Software Techniques for Emerging Hardware Platforms (Stanford University)** *Sept. – Dec. 2025*
- **Course Assistant – CS 224V: Conversational Virtual Assistants with Deep Learning (Stanford University)** *Sept. – Dec. 2024*
- **Co-Instructor – CS 25: Transformers United (Stanford University)** *2023 - 2026*

  Lead instructor of one of Stanford's hottest seminar courses. Responsibilities include lecturing, organizing speakers, and grading. In-depth talks each week about cutting-edge research in Transformers. Past speakers include Andrej Karpathy, Geoffrey Hinton, Jan Leike, Jim Fan, Ashish Vaswani, Hongyu Ren, Denny Zhou, and Jason Wei. Achieved worldwide recognition, with millions of YouTube views and hundreds of attendees. 2026 iteration sponsored by AGI House, Modal, and MongoDB ($12k). https://web.stanford.edu/class/cs25/

MENTORSHIP AND ADVISING

- **Patrick Peixuan Ye** – Stanford Undergrad, Computer Science, Class of 2026 [Now: Software Engineer @ Meta AI]
  PEFT/LoRA, chain-of-thought reasoning, and uncertainty estimation for LLMs. Paper in prep for NAACL 2027.
- **Ashley Malkin** – Stanford Undergrad, Symbolic Systems, Class of 2029
  Spatial and analogical reasoning for small language models. Paper in prep.

- **Linda Zeng** – The Harker School, Class of 2026 [Now: Undergrad @ Stanford]
  Code-switching and the multilingual capabilities of LLMs. Paper in prep for NAACL 2027.
- **Shijia Yang** – Stanford Master's of Computer Science, Class of 2025 [Now: Research Scientist @ AMD]
  Multimodal chain-of-thought reasoning with vision-language models (VLMs).
- **Sedrick Scott Keh** – CMU Master's of Machine Learning (MSML), Class of 2022 [Now: Member of Technical Staff @ Anthropic]
  Creative text generation. Two publications: one at COLING 2022 and another at EACL 2023.
- **Kevin Lu** – University of Waterloo Undergrad, Computer Science, Class of 2026 [Now: Co-founder & CEO @ Sweep]
  Controllable, creative, and visually-grounded text generation. First two publications at AAAI and COLING 2022, and an abstract at 2021.
- **Zhuofu (Derek) Tao** – UCLA Ph.D. in Electrical Engineering, Class of 2025 [Now: Compiler Engineer @ Amazon]
  Controllable and visually-grounded text generation. First NLP publication at AAAI 2022.

PROFESSIONAL SERVICE AND VOLUNTEERING

- **Reviewer – International Conference on Learning Representations (ICLR) 2027** *2026*
- **Reviewer – Conference on Language Modeling (COLM) 2026** *2026*
- **Reviewer – ACL Rolling Review (2026 Cycles, EMNLP 2025, ACL 2025, Dec. 2024 Cycle, EMNLP 2024)** *2024-2026*
- **Reviewer – CogInterp Workshop @ NeurIPS 2025, BabyLM Workshop @ EMNLP 2025** *2025*
- **Reviewer – ACL & EMNLP 2023** *2023*
- **Reviewer – Language Resources and Evaluation Conference (LREC) 2022** *2022*
- **Organizer – GEM Benchmark and Workshop for NLG Evaluation** [https://gem-benchmark.com/] *July 2021 – July 2022*
- **Lead Organizer – CtrlGen Workshop (NeurIPS 2021)** *Dec. 2020 – Dec. 2021*
  CtrlGen: Controllable Generative Modeling in Language and Vision workshop at NeurIPS 2021. Initiated idea and effort. In charge of planning & logistics. Main workshop host (recording). Primary writer/creator of the workshop proposal and website.
- **Reviewer – IEEE Transactions on Affective Computing (Journal)** *2021*
- **Reviewer – NeurIPS 2021 Workshop Proposals** *2021*
- **Reviewer – AAAI Conference on Artificial Intelligence 2021** *2020*

RESEARCH EXPERIENCE

DegenAI Labs
https://degenai-labs.github.io/

- Founder & Research Lead *Sept. 2025 - Present*
  - Founded an independent AI research group bringing together collaborators from Stanford, CMU, and other institutions.
  - Coordinate collaborations among PhD students, researchers, and engineers; oversee project scoping, experimental design, and writing.
  - Main writer and applicant for several research and compute grants including NSF, Modal, NVIDIA, and Lambda Labs [all awarded].
  - Leading multiple research projects (accepted to ICML and NeurIPS 2026, and targeting ICLR & ICML 2027), including hallucination in LLMs [ICML, NeurIPS], scaling laws for RAG [preprint], CoT reasoning, JEPA & world models, and model curiosity & welfare.

Stanford University

- PhD Graduate Research Assistant — Machine Learning, NLP, Computer Vision, Psychology, Cognitive Science — Stanford, United States
  Advisors: Dr. **Michael C. Frank**, **Noah Goodman** | *Other:* **Leonidas Guibas, Douwe Kiela** *Sept. 2022 - Present*
  - **Efficient Reasoning:** Several projects supported by NSF, Modal, NVIDIA, and Lambda Labs to enhance and optimize the reasoning capabilities of models, especially small and data-constrained LMs.
  - **BabyLM + VLM:** Projects supported by Amazon and Microsoft AFMR studying why and how human children learn and reason with

far less (language) data than current models. Investigating optimal training data sources and cognitively inspired benchmarks. First-author paper at EMNLP 2024 and submissions [1, 2] to NAACL 2027; additional papers at CogSci and CCN 2025.

- **Code Pretraining:** Collaborative project with Google DeepMind. Investigated the features of code data that support better learning for small models.
- **Diffusion Models & Text-to-Image Generation:** Worked on projects to enhance the efficiency and controllability of text-to-image generation.

Carnegie Mellon University's Language Technologies Institute (LTI)

- Master's Graduate Research Assistant — Machine Learning and Natural Language Processing (NLP) — Pittsburgh, United States
  Advisor: Dr. **Eduard Hovy** | *Other:* **Malihe Alikhani, Teruko Mitamura, Graham Neubig** — *Aug. 2020 – Aug. 2022*
  - **Commonsense Reasoning for NLG**: Devised effective approaches for generative commonsense reasoning and concept-to-text generation called SAPPHIRE. First-author paper published at INLG 2021 [**best long paper**]. Investigated visual-grounding for improving the commonsense reasoning of Transformer models for concept-to-text generation (first-author paper published at AAAI 2022) and commonsense QA.
  - **Creative Text Generation:** Investigated the automatic generation of personifications, tongue twisters, and narrative reordering (NAREOR). Publications at AAAI 2022, COLING 2022, and EACL 2023. Also explored approaches for visual storytelling (VIST).
  - **Clinical Generative Reasoning:** Explored the generative reasoning capabilities of language models for the clinical healthcare domain. Investigated data augmentation and automatic template construction and infilling. First-author paper published at EACL 2023.
  - **Data Augmentation for NLP**: Comprehensively surveyed existing data augmentation for NLP work and explored augmentation methods for text generation models. Co-first author papers published at ACL 2021 Findings and EMNLP 2020 DeeLIO Workshop.

University of Waterloo

- Undergraduate (USRA) Research Intern — Machine Learning and Natural Language Processing (NLP) — Waterloo, Canada
  Dr. **Jesse Hoey** and the Computational Health Informatics Lab — *Jan. 2019 – Apr. 2020*
  - **Semantic Text Exchange:** Proposed semantic text exchange (STE) to adjust the semantic content of text while preserving its sentiment and fluency. Developed a pipeline called SMERTI for STE. Co-first author of a paper published at EMNLP 2019. News article.
  - **Personalized Dialogue Agents:** Proposed Human Level Attributes (HLAs) to model human personality based on character tropes. Helped design and develop ALOHA, a personalized dialogue agent. Main writer and presenter (oral) of a paper published at AAAI 2020.

- Undergraduate (USRA) Research Intern — Machine Learning and Natural Language Processing (NLP) — Waterloo, Canada
  Dr. **Pascal Poupart** and the Artificial Intelligence Lab — *May 2019 – Apr. 2020*
  - **Robust Embeddings:** Devised a novel method for making word embeddings robust to noise using Bayesian statistics, where out-of-vocabulary words are modeled as Gaussian mixture distributions over vocabulary words weighted by their Levenshtein distance.

- Undergraduate Research Assistant (URA) — Robotics and Human-Computer Interaction (HCI) — Waterloo, Canada
  Dr. **Edith Law** and the HCI Lab — *Sept. - Dec. 2018*
  - **Autonomous Teaching Robots:** Worked on NAOqi robots that teach and interact with children, particularly those with learning disabilities.

EXTRACURRICULARS

- **Co-Founder and Co-President – Stanford Piano Society (SPS)** (https://piano.stanford.edu/) — *Sept. 2022 - Present*
- **Vice President of Education – University of Waterloo Data Science Club** — *Jan. – Apr. 2020*

CODING LANGUAGES

- Python, Java, C++, C, SQL, VBA, PyTorch, TensorFlow, Keras, etc.
- HuggingFace Transformers + Diffusers, DeepSpeed, FSDP, Megatron-LM, LitGPT, FAISS, etc. — Last Updated: ***10/1/2026***