Chongyang Gao

chongyanggao2026@u.northwestern.edu

Education Background

Northwestern University Sept.2021-now
Ph.D. in Computer Science, advised by Prof. V.S. Subrahmanian GPA: 3.9/4.0

Dartmouth College Sept.2019-Jun.2021
MS in Computer Science GPA: 12/14 HP

Tianjin University Sept.2015-Jun.2019
B.E. in Computer Science and Technology GPA: 3.8/4.0

Nanyang Technological University Jan.2019-Jun.2019
Visiting Student in Computer Science and Engineering

Websites

- **https://gcyzsl.github.io/**
- **Google Scholar: https://scholar.google.com/citations?user=HEAgatAAAAAJ&hl=en**

Internships

- **Student Researcher at Google DeepMind** July.2025-Dec.2025
- **Research Scientist Intern at Google X** Jun.2023-Feb.2024
- **Research Scientist Intern at Adobe Research** Mar.2023-Jun.2023

Research Experiences

Reference-Gated Surgical Corrective Learning 2025-2025

- Introduced a novel pipeline that utilizes a stronger teacher model to decompose negative responses into fine-grained "error tokens," enabling corrective learning to focus exclusively on the root cause of errors and improving data efficiency compared to coarse-grained sequence-level penalties.
- Proposed a computationally efficient approximation of the KL-divergence constraint via a Sparse KL formulation that discards the vocabulary tail and localizes corrective regularization to the contested token, reducing the KL overhead to O(1) per token.
- Introduced a dynamic, non-negative reference-gated mechanism that acts as a validity filter, applying token-level corrective penalties only when the reference model confirms the token is contextually anomalous; otherwise, the loss is gated to zero to preserve fluent syntax and common words.
- Presented a unified corrective learning objective where the positive supervision signal is explicitly scaled by the KL-penalty coefficient, balancing reinforcement of correct behaviors with reference-constrained correction on targeted error tokens.

On Large Language Model Continual Unlearning 2024-2025

- Studied the practical setting of continually arriving unlearning requests under no retained-data access and formalized the goal of balancing unlearning effectiveness with utility preservation in long-horizon deployments.
- Proposed the O3 framework which includes an unsupervised OOD-style detector to measure the similarity between input and unlearning data, an Orthogonal low-rank adapter for continuously unlearning and a dedicated unlearning-knowledge detection module to control when unlearning should be activated.

Higher Layers Need More LoRA Experts 2023-2024

- Introduced a novel parameter-efficient MoE method, MoE-LoRA with Layer-wise Expert Allocation for Transformer-based models, where each layer has the flexibility to employ a varying number of experts.
- Found that allocating more LoRA experts to higher layers further enhances the effectiveness of models with a certain number of experts in total. With much fewer parameters, this allocation strategy outperforms the setting with the same number of experts in every layer.

Selected Publications & Patterns (* Co-first author)

Chongyang Gao, Diji Yang, Shuyan Zhou, Xichen Yan, Luchuan Song, Shuo Li, Kezhen Chen, 'Classroom Final Exam: An Instructor-Tested Reasoning Benchmark', https://arxiv.org/abs/2602.19517

Chongyang Gao*, Lixu Wang*, Kaize, Ding Chenkai Weng, Xiao Wang, Qi Zhu, 'On Large Language Model Continual Unlearning', published in **ICLR 2025**, https://arxiv.org/abs/2407.10223

Chongyang Gao, et al. 'Higher Layers Need More LoRA Experts', published in **Findings at NAACL 2025**, https://arxiv.org/abs/2402.08562

Chongyang Gao, Marco Postiglione, Julian Baldwin, Natalia Denisenko, Isabel Gortner, Luke Fosdick, Chiara Pulice, Sarit Kraus, V. S. Subrahmanian, 'Context and Transcripts Improve Detection of Deepfake Audios of Public Figures', Submitted to Nature Communication**,** https://arxiv.org/abs/2601.13464

Chongyang Gao, Marco Postiglione, Isabel Gortner, Sarit Kraus, V. S. Subrahmanian, 'Perturbed Public Voices (P2V): A Dataset for Robust Audio Deepfake Detection', Submitted to KDD 2026**,** https://arxiv.org/abs/2508.10949

Qing, Peijun, **Chongyang Gao**, Yefan Zhou, Xingjian Diao, Yaoqing Yang, and Soroush Vosoughi. 'AlphaLoRA: Assigning LoRA Experts Based on Layer Training Quality.' **EMNLP 2024.** https://aclanthology.org/2024.emnlp-main.1141/

Chongyang Gao, Yiren Jian, Natalia Denisenko, Soroush Vosoughi, V.S. Subrahmanian, 'GEM: Generating Engaging Multimodal Posts', published in **IJCAI 2024.** https://www.ijcai.org/proceedings/2024/847

Chongyang Gao, Kang Gu, Soroush Vosoughi and Shagufta Mehnaz, 'Semantic-Preserving Adversarial Example Attack against BERT', published in **TrustNLP Workshop at NAACL 2024**, https://aclanthology.org/2024.trustnlp-1.17/

Chongyang Gao, Sushil Jajodia, Andrea Pugliese, V.S. Subrahmanian, 'FakeDB: Generating Fake Synthetic Databases', published in **IEEE Transactions on Dependable and Secure Computing.** https://www.computer.org/csdl/journal/tq/5555/01/10475547/1VrCU9KqFTq

Li Li, Jiawei Peng, Huiyi Chen, **Chongyang Gao,** Xu Yang, 'How to Configure Good In-Context Sequence for Visual Question Answering', published in **CVPR 2024.** https://arxiv.org/abs/2312.01571

Yiren Jian, **Chongyang Gao**, Soroush Vosoughi, 'Bootstrapping Vision-Language Learning with Decoupled Language Pre-training', (***Spotlight*** paper) published in **NeurIPS 2023**. https://arxiv.org/abs/2307.07063

Weiyi Wu, **Chongyang Gao**, Joseph DiPalma, Soroush Vosoughi1, Saeed Hassanpour, 'Improving Representation Learning for Histopathologic Images with Clustering', published in **ICCV 2023**. https://www.computer.org/csdl/proceedings-article/iccv/2023/071800v1347/1TJfxAf9vLW

Yiren Jian*, **Chongyang Gao***, Soroush Vosoughi, 'Non-Linguistic Supervision for Contrastive Learning of Sentence Embeddings', published in **NeurIPS 2022**. https://arxiv.org/abs/2209.09433

Yiren Jian*, **Chongyang Gao***, Soroush Vosoughi, 'Embedding Hallucination for Few-Shot Language Fine-tuning', published in **NAACL 2022**. https://arxiv.org/abs/2205.01307

Xu Yang*, **Chongyang Gao***, Hanwang Zhang, Jianfei Cai, 'Automatically Parsing Network for Image Captioning and Visual Question Answering', published in **ICCV 2021**. https://www.computer.org/csdl/proceedings-article/iccv/2021/281200c177/1BmIITZDI6A

Xu Yang*, **Chongyang Gao***, Hanwang Zhang, Jianfei Cai, 'Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning', published in **ACM MM 2020**. https://dl.acm.org/doi/abs/10.1145/3394171.3413859

Services

- **PC Member & Reviewer: ACL, AAAI, IJCAI, ICML, NeurIPS, ICLR, EMNLP, CVPR, ECCV, AISTATS, ACM MM, WACV, COLM, CoNLL, IJCV, IEEE TCSVT, IEEE Multimedia**