HAO ZOU

hz2999@columbia.edu ◇ haozou-official.github.io ◇ Google Scholar ◇ @haozou_official
Columbia University, Department of Computer Science ◇ New York, NY

EDUCATION

Columbia University — New York, NY
M.S. in Computer Science, Machine Learning Track — *09/2024 – 05/2026*

University of Minnesota, Twin Cities — Minneapolis, MN
B.S. in Computer Science — *09/2019 – 05/2023*

Research Interests: Reliable language generation, masked/diffusion language models, inference-time algorithms, faithful summarization, LLM agents, long-horizon computer-use, agent evaluation and training.

SELECTED PUBLICATIONS AND PREPRINTS

1. **Detect, Remask, Repair: Diffusion Editing for Faithful Summarization of Evolving Contexts**
Hao Zou, Zachary Horvitz, Chandhru Karthick, Zhou Yu, Kathleen McKeown
Under review, EMNLP 2026. [arXiv]
2. **OSWorld 2.0: Benchmarking Computer Use Agents on Long-Horizon Real-World Tasks**
OSWorld 2.0 Team, including **Hao Zou**
Under review, NeurIPS 2026.
3. **No Compute Left Behind: Rethinking Reasoning and Sampling with Masked Diffusion Models**
Zachary Horvitz, Raghav Singhal, **Hao Zou**, Carles Domingo-Enrich, Zhou Yu, Rajesh Ranganath, Kathleen McKeown
Under review, ICLR 2026. [arXiv]
4. **OpenForge: Training Harness-Based Agents End-to-End**
OpenForge Team, including **Hao Zou**
Manuscript in preparation, 2026.
5. **A Survey of Diffusion Models in Natural Language Processing**
Hao Zou, Zae Myung Kim, Dongyeop Kang
Preprint, 2023. [arXiv]
6. **You Make me Feel like a Natural Question: Training QA Systems on Transformed Trivia Questions**
Tasnim Kabir, Yoo Yeon Sung, Saptarashmi Bandyopadhyay, **Hao Zou**, Abhranil Chandra, Jordan Boyd-Graber
EMNLP 2024. [paper]
7. **Debiasing Language Models for In-Context Learning Using a Causal Inference-Inspired Method**
Hao Zou, Karin de Langis, Dongyeop Kang, Yohan Jo
Manuscript, 2023. [paper]
8. **Improving Question Answering with Generation of NQ-like Questions**
Saptarashmi Bandyopadhyay, Shraman Pal, **Hao Zou**, Abhranil Chandra, Jordan Boyd-Graber
Workshop manuscript, 2021. [arXiv]

RESEARCH EXPERIENCE

Columbia University, Department of Computer Science — 02/2025 – Present
Research Staff / Staff Associate I (06/2026 – Present); Graduate Research Assistant (02/2025 – 05/2026)
Advisors and collaborators: Kathleen McKeown, Zhou Yu, Zachary Horvitz, Xiao Yu

- **Faithful summarization with diffusion editing.** First-authored *Detect, Remask, Repair*, a localized repair framework for evolving-context summarization that detects unsupported spans, remasks them, and repairs them with masked diffusion language models instead of fully regenerating summaries.
- Built and evaluated faithfulness-oriented repair pipelines across DialogSum and StreamSum, combining uncertainty, discriminator signals, and faithfulness-steered decoding to study tradeoffs among factuality, preservation, and generation cost.

- **Masked diffusion language model reasoning.** Contributed to *No Compute Left Behind*, studying inference-time compute allocation in MDLMs across mathematical reasoning, coding, and structured tasks; explored reasoning-as-infilling, answer-conditioned posterior sampling, early exit, and adaptive multi-token decoding.
- **Long-horizon computer-use agents.** Contributed to OSWorld 2.0 by designing high-fidelity, long-horizon desktop/web tasks with realistic multi-application workflows, external dependencies, constraint-following requirements, and robust evaluation rubrics.
- **Harness-based agent training and environments.** Working with Xiao Yu on Azure/OpenForge-style agent environment scaling: curating executable tasks and environments for SFT/RL, improving reproducible evaluation, and connecting realistic agent harnesses with scalable training workflows.

Duke University 05/2024 – 01/2025
Graduate Research Assistant *Advisor: Enmao Diao*

- Developed a benchmarking pipeline comparing diffusion and flow-matching training paradigms under different prediction objectives across architectures and datasets.
- Introduced a triangle-distribution-based training objective to improve training stability and generation quality over Gaussian baselines.

University of Minnesota, Twin Cities, NLP Lab 08/2021 – 05/2023
Undergraduate Research Assistant *Advisor: Dongyeop Kang*

- First-authored a survey of diffusion models for NLP, synthesizing discrete/continuous diffusion formulations, denoising objectives, parallel generation, controllability, and robustness in language generation.
- Developed a causal-inference-inspired method for debiasing in-context learning, estimating the causal effect of input text on potential labels and improving pretrained language model accuracy across text classification tasks.

University of Maryland, CLIP Lab 06/2021 – 06/2022
Undergraduate Research Assistant *Advisor: Jordan Boyd-Graber*

- Developed pipelines for transforming complex trivia questions into natural information-seeking questions using syntactic simplification and question rewriting models.
- Fine-tuned dense passage retrievers on rewritten questions to better align retrieval embeddings with transformed query distributions and improve cross-domain QA evaluation.
- Analyzed 2k+ generated questions to characterize linguistic variation, abstraction patterns, and domain-shift behavior in generated QA data.

INDUSTRY EXPERIENCE

IBM 10/2024 – 01/2025
AI Algorithms Intern, Quantization / LLM Efficiency *Mentor: Naigang Wang*

- Proposed a Johnson–Lindenstrauss-based KV-cache quantization approach integrated with sparsity pruning, targeting lower LLM inference memory and latency on IBM platforms.
- Benchmarked quantized attention mechanisms across multiple LLM families and evaluated memory-speed trade-offs for efficient inference.

Sony Research 01/2022 – 05/2023
LLM RL Intern

- Developed a multi-style controllable generation pipeline using PPO and fusion-based reward/modeling strategies, improving human-judged style alignment in generated text.

ACADEMIC SERVICE

Conference Reviewer: ACL, EMNLP, EACL

SKILLS

Research Areas	Reliable generation, MDLMs, diffusion models, faithful summarization, LLM agents
Methods	Inference-time algorithms, decoding, SMC/FK steering, RL/SFT for agents, evaluation
NLP/Agents	Summarization, reasoning, RAG, QA, long-horizon planning, computer-use benchmarks
Programming	Python, PyTorch, CUDA, C/C++, Linux, Git, Docker, LaTeX