

ZIHAN DONG

🏠 [Homepage:zihandong02.github.io](https://zihandong02.github.io)

🎓 [Google Scholar](#)

🌐 [LinkedIn](#)

✉️ zd182@stat.rutgers.edu

EDUCATION

Rutgers University
Ph.D. in Statistics

Piscataway, NJ
2023–Present

**School of the Gifted Young,
University of Science and Technology of China (USTC)**
B.S. in Mathematics, Major GPA: 4.17/4.30

Hefei, China
2019–2023

INDUSTRY EXPERIENCE

Applied Scientist Intern
Amazon AGI

2026.09 – Present

- Design and build scalable pipelines for efficient RL data generation and training, targeting SWE and long-horizon agentic tasks.

Research Intern

2026.05 – 2026.08

Qwen, Tongyi Lab, Alibaba Group

- Contributed to the post-training of Qwen3.7 and Qwen3.8 for long-horizon agentic coding (terminal-use) tasks, spanning CPT, SFT, RL and OPD.
- Built and curated CPT, SFT, and RL data from real-world logs, third-party sources, and synthetic tasks; established quality standards and developed automated trajectory-cleaning pipelines.
- Developed and scaled SFT recipes from small-model pilots to full-scale training, using downstream evaluations to guide recipe iteration.
- Designed RL reward-shaping strategies for solution efficiency and submission behavior; investigated fine-grained credit assignment for long-horizon tasks.

RECENT PROJECTS

Just Reuse the Rollout Log-Probs: Stable MoE RL

2026.05 – Present

- MoE RL is destabilized by a **train–inference mismatch** on the same tokens, which existing fixes remove only at the cost of bias.
- Analyzed the per-token distribution and built **VoKL clip**, a distributional top- k KL trust region that reuses the rollout’s own log-probs to fix the mismatch without bias, keeping the ratio clip only as a numerical safety net.
- On Qwen3-30B-A3B (DAPO-Math-17k), reached **64.5% AIME24** (ave@32), outperforming DAPO (59.8%) and GSPO (57.2%).

Evaluating LLMs When They Do Not Know the Answer

2025.10 – 2026.01

- Studied **evaluation-as-estimation**: how to estimate an LLM’s ability when the verification signal is noisy or the ground-truth answer is unknown, exploiting the generation–verification gap.
- Built a statistical framework that ranks models from **comparative signals**, giving significance-tested rankings with lower variance than single-answer accuracy.

Rubric-Based Alternating Reward Modeling (ARM)

2025.11 – 2026.04

- For open-ended problems and long-horizon coding, correctness cannot be checked by rules, so the **reward model itself must be learned**.
- Proposed an **alternating RL framework** that co-trains a rubric-based reward model with the policy for non-verifiable post-training.
- Designed pointwise rubric reward modeling (**RUBRIC-ARROW**) that scores open-ended responses against interpretable rubrics, improving reward accuracy over pairwise-preference reward models.

- Contributed to AutoResearchClaw (**13.5K+ GitHub stars**), a multi-agent autonomous research pipeline that turns the linear idea-to-paper process into an iterative loop: multi-agent debate, a self-healing executor that turns failed experiments into information, and cross-run evolution that carries lessons across runs.
- Outperforms AI Scientist v2 by **54.7%** on ARC-Bench; its human-in-the-loop design shows targeted collaboration at high-leverage decisions beats both full autonomy and exhaustive oversight.

PREPRINTS/PUBLICATIONS

- Conference Papers

1. Tianci Liu, **Zihan Dong**, Tianchun Li, Yi-Chung Chen, Qiming Cao, Xingchen Wang, Shiyang Wang, Zichen Miao, Linjun Zhang, Haoyu Wang, and Jing Gao (2026). Hybrid-Policy Self-Editing for Composable Unstructured Knowledge Editing. Submitted to *ICLR 2027*. <https://arxiv.org/abs/2608.11660>.
2. **Zihan Dong**, Zhixian Zhang, Yang Zhou, Can Jin, Ruijia Wu, Linjun Zhang (2026). Evaluating LLMs When They Do Not Know the Answer: Statistical Evaluation of Mathematical Reasoning via Comparative Signals. *ICML 2026*.
3. Pei Tian, **Zihan Dong**, Tianci Liu, Linjun Zhang, and Haoyu Wang (2026). RIMS: Preference Optimization via Smoothed Multi-pair Aggregation for Small-Scale LLM Retrieval-Augmented Generation. *COLM 2026*.
4. Tianci Liu, **Zihan Dong**, Linjun Zhang, Haoyu Wang, Jing Gao, Emre Kiciman, Ranveer Chandra, and Wei-Ting Chen (2026). Personalization as Inverse Planning: Learning Latent Design Intents for Agentic Slide Generation via Structural Denoising. *ECCV 2026*.
5. Ran Xu, Yuchen Zhuang, **Zihan Dong**, Ruiyu Wang, Yue Yu, Joyce C. Ho, Linjun Zhang, Haoyu Wang, Wenqi Shi, and Carl Yang (2025). AceSearcher: Bootstrapping Reasoning and Search for LLMs via Reinforced Self-Play. *NeurIPS 2025 Spotlight*.
6. Ran Xu, Tianci Liu, **Zihan Dong**, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tuo Zhao, Haoyu Wang (2026). Alternating Reinforcement Learning for Rubric-Based Reward Modeling in Non-Verifiable LLM Post-Training. *ICML 2026*.
7. Shuhang Lin, Chuhao Zhou, Xiao Lin, **Zihan Dong**, Kuan Lu, Zhencan Peng, Jie Yin, Dimitris N. Metaxas (2026). Conformal Path Reasoning: Trustworthy Knowledge Graph Question Answering via Path-Level Calibration. *ICML 2026*.
8. Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, **Zihan Dong**, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, Shujie Liu, Yan Lu, and Huaxiu Yao. (2026). MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning. *ICLR 2026*.
9. Tianci Liu, Ruirui Li, **Zihan Dong**, Hui Liu, Xianfeng Tang, Qingyu Yin, Linjun Zhang, Haoyu Wang, and Jing Gao (2025). Mitigating Heterogeneous Token Overfitting in LLM Knowledge Editing. *ICML 2025*.
10. **Zihan Dong**, Shuo Yang, Chiyu Ma, Haozhe Li, Yang Zhou, Zihang Meng, Hao Zhou, Linjun Zhang, and Guoyin Wang (2026). Just Reuse the Rollout Log-Pros: Stable Mixture-of-Experts Reinforcement Learning with a Vocabulary-KL Clip. Submitted to *ICLR 2027*.
11. Jiaqi Liu, Shi Qiu, Mairui Li, Bingzhou Li, Haonian Ji, Siwei Han, Meng Chen, Xinyu Ye, Peng Xia, **Zihan Dong**, Congyu Zhang, Letian Zhang, Hardy Chen, Haoqin Tu, Xinyu Yang, Lu Feng, Xujiang Zhao, Haifeng Chen, Jiawei Zhou, Xiao Wang, Weitong Zhang, Hongtu Zhu, Yun Li, Jieru Mei, Hongliang Fei, Jiaheng Zhang, Linjie Li, Linjun Zhang, Yuyin Zhou, Sheng Wang, Caiming Xiong, James Zou, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao (2026). **AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration**. Submitted to *ICLR 2027*. <https://arxiv.org/abs/2605.20025>.
12. Yang Zhou, Can Jin, **Zihan Dong**, Zhepeng Wang, Yanting Yang, Shiyu Zhao, Lei Li, Runxue Bao, Yaochen Xie, and Dimitris N. Metaxas (2026). DARE: Difficulty-Adaptive Reinforcement Learning with Co-Evolved Difficulty Estimation. Submitted to *ICLR 2027*. <https://arxiv.org/abs/2605.09188>.
13. Yang Zhou, **Zihan Dong**, Zhenting Wang, Can Jin, Shiyu Zhao, Bangwei Guo, Difei Gu, Zhepeng Wang, Linjun Zhang, Mu Zhou, and Dimitris N. Metaxas (2026). Evidence Over Plans: Online Trajectory Verification for Skill Distillation. Submitted to *ICLR 2027*. <https://arxiv.org/abs/2605.09192>.

14. Haoxiang Jiang, **Zihan Dong**, Tianci Liu, Wanying Wang, Ran Xu, Tony Yu, Linjun Zhang, and Haoyu Wang (2026). RUBRIC-ARROW: Alternating Pointwise Rubric Reward Modeling for LLM Post-training in Non-verifiable Domains. Submitted to *EMNLP 2026*. <https://arxiv.org/abs/2605.29156>.
15. Can Jin, Yang Zhou, Qixin Zhang, Hongwu Peng, Di Zhang, **Zihan Dong**, Marco Pavone, Ligong Han, Zhang-Wei Hong, Tong Che, Dimitris N. Metaxas (2026). AIRL-S: Unifying Reinforcement Learning and Search-Based Test-Time Scaling via Adversarial Inverse Reinforcement Learning. Submitted to *EMNLP 2026*. <https://arxiv.org/abs/2508.14313>.
16. Difei Gu, Yunhe Gao, Gerasimos Chatzoudis, **Zihan Dong**, Guoning Zhang, Bangwei Guo, Yang Zhou, Mu Zhou, Dimitris N. Metaxas (2026). LUCID-SAE: Learning Unified Vision-Language Sparse Codes for Interpretable Concept Discovery. <https://arxiv.org/abs/2602.07311>.

- Journal Papers

1. **Zihan Dong**, Ruijia Wu, and Linjun Zhang (2026). Labels or Preferences? Budget-Constrained Learning with Human Judgments over AI-Generated Outputs. To be submitted to *JASA Theory and Methods*. <https://arxiv.org/abs/2601.13458>.
2. **Zihan Dong**, Xin Zhou, Ryumei Nakada, Lexin Li, and Linjun Zhang (2025). Contrastive Network Representation Learning. Submitted to *JASA Theory and Methods*. <https://arxiv.org/abs/2509.11316>.

TECHNICAL SKILLS

- Proficient with verl, SGLang, vLLM, and Megatron-LM for large-scale LLM post-training, distributed training, and high-throughput inference.
- 4+ Years of research experience in data science, statistics, and machine learning, with expertise in Python (PyTorch), R and SQL. Experienced in generative and foundation models, applying supervised and unsupervised learning methods in AI.
- Programming Languages: Python, R, C/C++, SQL
- Libraries and Tools: PyTorch, Scikit-Learn, Pandas, Numpy, Scipy, Matplotlib, Excel

HONORS AND AWARDS

- | | |
|---|-----------|
| • Best Ph.D. Qualifying Exam Performance Award, Rutgers University | 2024 |
| • National Scholarship, Top 1% nationwide, Chinese Ministry of Education | 2022 |
| • Cyrus Tang Scholarship, Top 3%, Cyrus Tang Foundation, Nationwide | 2019-2022 |
| • The 12 th Chinese Mathematics Competitions, Second Prize in Professional Group | 2020 |