

Zhenyu (Zach) Wang

Ph.D. Candidate in Statistics

wangzy0701@gmail.com | Website | GitHub | Google Scholar

My research spans statistical learning, optimization, and large language models. I develop mathematically principled approaches to budget-aware LLM inference and LLM post-training: deriving closed-form asymptotics for the accuracy–compute Pareto frontier of LLM inference and formulating optimal on-policy distillation as a bilevel optimization problem.

Education

Rutgers University — Ph.D. in Statistics, GPA: 4.0/4.0 Sept. 2022–May 2027 (expected)
Advisors: Zijian Guo and Yifan Hu.

Columbia University — M.S. in Applied Analytics, GPA: 4.0/4.0 Sept. 2019–May 2021

Shanghai Jiao Tong University — Bachelor’s in Finance, GPA: 3.76/4.0 Sept. 2014–June 2019

LLM Post-training & Inference

Dr. OPD: On-Policy Distillation Done Right 2026 · Preprint

Zhenyu Wang*; Tianze Wang*; Linjun Zhang; Yifan Hu. Paper · GitHub

- Defined **optimal on-policy distillation** as choosing token-level teacher weights to maximize student task performance. Formulated it as bilevel optimization: weighted distillation is the inner problem, and student outcome reward is the outer objective.
- Strong-to-weak distillation: Improved Qwen3-1.7B’s five-benchmark math average by **9.7 percentage points** over vanilla OPD (25.3% → 35.0%), surpassing its larger Qwen3-4B teacher (34.1%).
- Same-size distillation: Improved the math average by **2.2 / 3.0 percentage points** over vanilla OPD for Qwen3-4B / Qwen3-4B-Base (55.6% → 57.8% / 33.0% → 36.0%), surpassing both corresponding RL-trained teachers.

PRICE: The Accuracy–Compute Frontier of LLM Inference 2026 · Preprint

Zhenyu Wang; Xiaozhi Zhu; Yifan Hu. Paper · GitHub

- How much accuracy can one more token buy? Used large-deviation theory to derive a closed-form asymptotic accuracy–compute Pareto frontier, characterizing the accuracy ceiling and the rate of diminishing returns.
- Designed a budget-aware algorithm that jointly adapts rollout counts and voting rules per query, rather than fixing answer aggregation while allocating more samples.
- On MATH-500, outperformed six baselines at every evaluated budget on Qwen2.5-1.5B and Llama-3.2-3B; matched self-consistency accuracy with up to **2.8× / 3.5× smaller token budgets**, respectively.

Recognition: NeurIPS 2026 workshop **oral presentation** in FLLMPT; also accepted at MLxOR, RAAAI, and TTCL.

AdSteer: Adaptive Decoding-Time Steering for Native Advertising in LLMs 2026 · Preprint

Shuhang Lin; Chuhao Zhou; Zhenyu Wang; Xiao Lin; Changsheng Wang; Wenyue Hua; Saibo Geng; Dimitris N. Metaxas.

- Developed single-pass, controllable LLM decoding for native advertising: a learned SteerHead adapts token-level logit biases, while a keyword trie and finite state machine control when steering engages and stops, keeping the LLM backbone frozen.
- On GEM-Bench, improved the LLM-judged overall score by **35.4%** relative to GIR-P on CA-Prod, with approximately **25% fewer total tokens** on average than rewrite-based methods.

Statistical Learning & Optimization

DRoL: Robust Learning Across Multiple Sources Annals of Statistics, 2026

Zhenyu Wang; Peter Bühlmann; Zijian Guo. Paper

- Find the most robust model across data sources. Developed a minimax framework for multi-source domain adaptation that maximizes worst-case predictive performance using labeled sources and no target labels.
- Established statistical guarantees and demonstrated gains on real air-pollution data; the framework accommodates neural networks, gradient boosting, and random forests.

* Equal contribution.

Statistical Learning & Optimization (continued)

StablePCA: Robust Shared Representation Learning

2025 · Preprint

Zhenyu Wang; Molei Liu; Jing Lei; Francis Bach; Zijian Guo. [Paper](#)

- Learned low-dimensional representations that maximize worst-case explained variance across heterogeneous sources, avoiding embeddings dominated by a single population.
- Developed closed-form Mirror Prox updates, achieving approximately $39\times$ **faster** optimization than SDP-based FairPCA at dimension 300; evaluated shared representations on multi-batch single-cell RNA-seq data.

NegDRO: Causal Invariance via Nonconvex Optimization

2024 · Preprint

Zhenyu Wang*; Yifan Hu*; Peter Bühlmann; Zijian Guo. [Paper](#)

- Recast causal invariance learning as continuous nonconvex minimax optimization, replacing combinatorial feature search with gradient-based methods for out-of-distribution generalization.
- Under additive interventions, established a benign optimization landscape and gradient convergence to the causal outcome model, with rates in both sample size and iteration count.

CG-DRO: Robust Classification & Reliable Inference

2025 · Preprint

Zijian Guo; Zhenyu Wang; Yifan Hu; Francis Bach. [Paper](#)

- Developed conditional group DRO for robust classification with cross-entropy loss, using Mirror Prox optimization and double machine learning for nuisance estimation.
- Established sharp statistical rates and uniformly valid confidence intervals, addressing settings where standard Wald-type inference fails.

SIHR: High-Dimensional Statistical Inference

The R Journal, 2024

Prabrisha Rakshit*; Zhenyu Wang*; T. Tony Cai; Zijian Guo. [Paper](#) · [Software](#)

- Developed methods and software for confidence intervals and hypothesis testing in high-dimensional linear and logistic regression.

Teaching & Research Communication

Teaching Assistant, Rutgers University

- Statistical Learning for Data Science (Spring & Fall 2026); Data Mining (Spring 2025); Statistics Ph.D. core courses (Spring 2024).
- Delivered two invited lectures on diffusion models (Spring 2025). [Tutorial notes](#)

Invited Talks: NeurIPS / AoS collaboration (2025); INFORMS (2025); JSM (2025); NESS at Yale (2025); IMS ICSDS (2024).

Awards

ETH Zurich Young Researcher Travel Award (2025) · IMS ICSDS Student Travel Award (2024) · Rutgers University Research Fellowship (2024).