Hanyu Wang

hbw5365@psu.edu · homepage · wwwhy · wwwhy725

Education

The Pennsylvania State University (PSU), State College, USA — Aug. 2024 – Present

Ph.D. student in Informatics
Advisor: Dr. Jinghui Chen

University of Science and Technology of China (USTC), Hefei, China — Sep. 2020 – Jun. 2024

B.S. in Information and Computing Science (Mathematics)
Advisor: Dr. Jingrun Chen

Research

Research Interest

My research lies at the intersection of Natural Language Processing and Trustworthy AI. I am broadly interested in a range of topics aimed at making AI systems more capable and reliable, including:

- Understanding and mitigating hallucinations in LLMs.
- Enhancing the multi-step reasoning abilities of language models.
- Building advanced agentic systems capable of reflection.

From Retrospective to Prospective Reflection in LLM Agents — Sep. 2025 – Jan. 2026

Lead Researcher | Manuscript in preparation for submission
Advisor: Dr. Jinghui Chen The Pennsylvania State University

Brief introduction:
This research developed a framework that shifts reflection from reactive, post-hoc correction to proactive, pre-execution foresight. By anchoring a planning-phase reflection loop with a distilled taxonomy of historical errors, the system identifies and mitigates potential failures before execution.

- Proposed a prospective reflection mechanism that enables agents to critique and refine plans before committing to actions, effectively preventing critical errors from happening.
- Distilled a domain-agnostic taxonomy of planning errors from agent trajectories to serve as experiential priors, allowing for grounded, precise prospective reflection.
- Integrated a dynamic re-planning loop that monitors execution-time state to trigger strategic pivots, ensuring robustness against environmental uncertainty.
- Outperformed reflection-related methods and complex agentic architectures across different benchmarks, including GAIA and SimpleQA.
- Further analysis shows an outstanding cost-performance trade-off and transferability across agentic frameworks (e.g., Smolagents and OWL).

Truthful LLM Generation via Representation Flow Correction — Sep. 2024 – Jan. 2025

Research Assistant
Advisor: Dr. Jinghui Chen The Pennsylvania State University

Brief introduction:
Developed **TruthFlow**, a novel representation intervention framework that mitigates LLM hallucination by generating query-specific correction vectors via flow matching. Unlike prior methods that rely on a universal intervention vector, this targeted approach is enhanced by a truth-related subspace projection using SVD that purifies correction signals. The method achieved a 7% average improvement in truthfulness on the TruthfulQA benchmark and demonstrated strong transferability to unseen domains.

- Led the design and development of TruthFlow, a novel framework tested across a wide range of LLMs, including the Llama, Mistral, and Gemma families.
- Engineered a query-specific correction mechanism using the Flow Matching technique to overcome the limitations of universal, one-size-fits-all intervention vectors.
- Designed and implemented a truth-related subspace projection method using Singular Value Decomposition to denoise correction vectors and further enhance model truthfulness.

- Improved open-ended generation truthfulness by over 7% on average on the TruthfulQA benchmark, outperforming various baselines.
- Demonstrated the framework's transferability by successfully applying it to unseen datasets, including HaluEval, Natural Questions, and TriviaQA.

Empirical Understanding of Generalizability in Diffusion Models
Jul. 2023 – Jun. 2024

Research Assistant
Advisor: Dr. Difan Zou *The University of Hong Kong*

Brief introduction:
This research addresses the inherent paradox in diffusion models, where the closed-form optimal solution for the denoising score suggests perfect memorization of the training data, yet empirically trained models demonstrate powerful generalization capabilities. We resolve this discrepancy by conducting a multi-faceted comparative analysis of the theoretical "optimal score" and the practical "trained score," ultimately attributing the model's ability to generalize to the superior geometric smoothness of the trained score function.

- Quantified the geometric properties of the score functions using Jacobian SVD and difference quotients, demonstrating the trained score is significantly smoother than its sharp, theoretical optimum.
- Established that generated images act as local minima within the trained score's objective, confirming that the model learns a structured landscape for creating novel data rather than just interpolating.
- Engineered a novel score-mixing experiment that injects the optimal score into the sampling process, proving it causally redirects generation to a memorization regime that replicates training data.

On Mitigating Memorization in Diffusion Models
Jan. 2023 – Dec. 2023

Undergraduate Student Research Program
Advisor: Dr. Jingrun Chen *University of Science and Technology of China*

Brief introduction:
This research connected generalization, memorization, and smoothness, based on which we proposed a regularization method to theoretically and empirically mitigate memorization in diffusion models.

- Proposed a regularization-based method to mitigate memorization in diffusion models by enforcing a smoother trained score function.
- Empirically, we achieved a good trade-off between generalization and generation quality, demonstrating the effectiveness of the proposed regularization method.
- Theoretically, we proved the lower bound control of generalization when the training dataset is large enough, showing that memorization risk increases as the training dataset size decreases.

Papers in Submission

- **PreFlect: From Retrospective to Prospective Reflection in Large Language Model Agents** [Paper]
  Hanyu Wang, Yuanpu Cao, Lu Lin, Jinghui Chen

Publications

- **TruthFlow: Truthful LLM Generation via Representation Flow Correction** [Paper]
  Hanyu Wang, Bochuan Cao, Yuanpu Cao, Jinghui Chen
  Proceedings of the 42nd International Conference on Machine Learning (**ICML 2025**), Vancouver, Canada.
- **On the Discrepancy and Connection between Memorization and Generation in Diffusion Models** [Paper]
  Hanyu Wang, Yujin Han, Difan Zou
  In Proceedings of the *ICML 2024 Workshop on Foundation Models in the Wild.*

Skills

- Programming Languages: Python(Pytorch), C/C++, MATLAB, Mathematica, R, LaTeX
- Agent Frameworks: Smolagents, OpenManus, Camel, OWL, AWorld.