Zheyu Zhang

Email: zheyu915@gmail.com
Phone: (+49) 15754742120
Website: oooranz.github.io
LinkedIn: THIS URL
Google Scholar: THIS URL

Education

Technical University of Munich — Munich, Germany
Ph.D. student in Computer Science, supervised by Prof. Dr. Gjergji Kasneci; — Dec. 2024–Mar. 2028 (expected)
Junior Member of Munich Center for Machine Learning (MCML)

University of Munich (LMU Munich) — Munich, Germany
M.Sc. in Computational Linguistics, supervised by Prof. Dr. Hinrich Schütze — Oct. 2022–Aug. 2024

University of Munich (LMU Munich) — Munich, Germany
B.Sc. in Computer Science and Mathematics — Apr. 2020–Sept. 2022

Research Interests

I work on **LLM post-training, model adaptation, and data-centric machine learning**, with a strong interest in translating methodological advances into real-world systems. My research explores how reward-based post-training and inference-time steering can shape LLM capabilities [1, 5, 7], and how generative modeling and data augmentation can improve learning from limited structured data [2, 3, 4, 6]. More recently, I have been also exploring world modeling for relational systems [9].

Selected Publications

* *indicates equal contribution.*

Published and Accepted

[1] **Zheyu Zhang**, Shuo Yang, and Gjergji Kasneci. *Consolidating Rewarded Perturbations for LLM Post-Training*. **NeurIPS 2026**. [arXiv] [code]

[2] **Zheyu Zhang**, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. *Active Tabular Augmentation via Policy-Guided Diffusion Inpainting*. **ICML 2026**. [arXiv] [code]

[3] Shuo Yang*, **Zheyu Zhang***, Bardh Prenkaj, and Gjergji Kasneci. *SAGE: Sparse Adaptive Guidance for Dependency-Aware Tabular Data Generation*. **ACL 2026**. [arXiv] [code]

[4] **Zheyu Zhang**, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. *Not All Features Deserve Attention: Graph-Guided Dependency Learning for Tabular Data Generation with Language Models*. **EMNLP 2025**. [arXiv] [code]

[5] Chenchen Yuan, **Zheyu Zhang**, and Gjergji Kasneci. *Where Paths Split: Localized, Calibrated Control of Moral Reasoning in Large Language Models*. **ACL 2026**. [arXiv]

[6] Shuo Yang, **Zheyu Zhang**, Bardh Prenkaj, and Gjergji Kasneci. *Doubling Your Data in Minutes: Ultra-Fast Tabular Data Generation via LLM-Induced Dependency Graphs*. **EMNLP 2025**. [arXiv] [code]

[7] Chenchen Yuan, **Zheyu Zhang**, Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. *Probabilistic Aggregation and Targeted Embedding Optimization for Collective Moral Reasoning in Large Language Models*. **ACL 2025**. [arXiv]

[8] **Zheyu Zhang**, Han Yang, Bolei Ma, David Rügamer, and Ercong Nie. *Baby's CoThought: Leveraging Large Language Models for Enhanced Reasoning in Compact Models*. **EMNLP 2023 Shared Task BabyLM Challenge**. [arXiv] [code]

Preprints

[9] **Zheyu Zhang** and Gjergji Kasneci. *Relational Learning as World Modeling*. **Under Review**. [arXiv] [code] [model].

Experience

National Institute of Informatics (NII) — Tokyo, Japan
Research Intern; working on LLM post-training for self-improvement — Oct. 2026–Jan. 2027

German Research Center for Artificial Intelligence (DFKI) — Berlin, Germany
Research Assistant; working on explainability for LLM-based RAG systems — Apr. 2024–Aug. 2024

Bosch Research — Stuttgart, Germany
Research Intern; working on agentic RAG systems for industrial applications — Nov. 2023–Apr. 2024

LMU Munich — Munich, Germany
Teaching Assistant; for course *Logic and Discrete Structures* — Apr. 2021–Sept. 2023

Projects

Enabling Agentic Tool-Use Capabilities in a 1.6B On-Device LLM through Post-Training Tabularis.AI
Applied Research Collaboration Dec. 2025–June 2026

- Led the post-training and capability optimization of `Faust-1`, a **1.6B German-first small LLM** designed for local deployment, enabling reliable tool use and function calling through **SFT, DPO, GRPO, and self-distillation**.
- Developed a data-centric training pipeline with multi-source curation, **12K+ targeted synthetic examples**, LLM-based filtering, and data-mixture optimization to improve tool-use accuracy and reduce hallucinated calls.
- Improved single-turn BFCL performance **from 26.7% to 74.2% (+47.5 pts)** via SFT, post-training, and schema-aware validation/retry, outperforming comparable 1B–4B models under the same evaluation.

Generative Modeling for Low-Resource Structured Business Data Creditreform
Industry-Sponsored Research Collaboration May 2025–July 2026

- Led a long-term research project on **scarce, sensitive, and heterogeneous structured business data**, working with multidimensional company and ESG attributes under limited-data settings.
- Developed **LLM- and diffusion-based synthetic data generation and augmentation methods** to capture complex dependencies across heterogeneous business attributes and expand limited training data.
- This project resulted in 3 first-author publications at **EMNLP, ACL, and ICML**.

Academic Service

- **Volunteering:** ACL 2025, EACL 2024
- **Reviewing:** ICLR 2026, ICML 2026, ACL-ARR 2025–2026, ACM MM 2026, ECML-PKDD 2025

Skills

- **Programming:** Python, C, Java, JavaScript, SQL
- **ML/NLP:** PyTorch, Hugging Face, DeepSpeed, vLLM, VeRL
- **Languages:** Chinese (Native), English (Fluent), German (Fluent)