Zhixuan Liu

lzx993124494@sjtu.edu.cn · Shanghai, China

chadlzx · Homepage · Google Scholar · OpenReview

Education

Shanghai Jiao Tong University, Ph.D. in Computer Science 2022–2027

School of Computer Science; Advisor: Yu Qiao; expected Jun. 2027

Shanghai Jiao Tong University, B.Eng. in CS (ACM Class, Zhiyuan College) 2018–2022

Research & Professional Experience

Shanghai Artificial Intelligence Laboratory, Joint Ph.D. Training 2022–Present

- Conduct research on LLM post-training and inference, focusing on model behavior and safety, inference-time alignment, and efficient generation; lead multiple independent research projects as first author and own problem formulation, method design, system implementation, experimental analysis, and paper writing.
- Research spans latent-trait transfer mechanisms and control, inference-time preference guidance, and contrastive decoding acceleration, respectively addressing post-training safety, inference-time behavior control, and inference efficiency.

ByteDance AI Lab, Research Intern Sep. 2020–Jun. 2021

- Addressed the difficulty of reliable exact numerical reasoning with neural models by unifying neural analogical inference with executable symbolic reasoning; independently owned problem selection, method design, implementation, data and experiment organization, analysis, and paper writing.
- Designed a shared encoder, neural prediction and program-generation/execution branches, and gated routing; organized **32,011** DROP and **722** AWPNLI program annotations and completed QA/NLI model training and evaluation.
- Achieved **84.01 F1** on DROP and **92.24%** accuracy on AWPNLI; the resulting work appeared at EMNLP 2022 with open-source code.

Selected Research Projects

Subliminal Learning as Trait-Direction Drift: A Mechanism and Targeted Control under SFT Distillation

First Author · NeurIPS 2026 Poster

- Investigates why SFT distillation can transfer latent teacher preferences through data that is semantically unrelated to the target trait, a phenomenon previously observed without a predictive mechanism or targeted control.
- Introduces trait-direction drift to explain how sample-level preference gaps accumulate into trait-aligned updates, and probe-space corridor regularization to reduce malicious-response transfer from **29.55%** to **6.45%** while preserving task performance.

Decoupled Contrastive Decoding via Expert-Aligned Drafting First Author · EMNLP 2026 Main

- Addresses the serial latency of amateur-model proposals in contrastive decoding, seeking speculative acceleration without changing the target distribution of exact contrastive decoding.
- Uses an expert-aligned EAGLE3 drafter and moves the amateur model to parallel verification; implements three-model inference with independent KV caches in SGLang, achieving **1.65–1.95**× speedups at 8B and about **2**× at 70B.

Inference-Time Language Model Alignment via Integrated Value Guidance Co-first Author (First-listed) · EMNLP 2024 Findings

- Addresses preference control for frozen LLMs at inference time, where token-level guidance lacks global judgment and further fine-tuning of the target model is undesirable.
- Combines token-level DPO implicit values with explicit prefix-value-guided chunk search, raising AlpacaEval 2.0 length-controlled win rates from **19.51% / 25.58%** to **26.51% / 33.75%**.

Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models Co-first Author (Second-listed) · NeurIPS 2024

- Studies how weak small models can align stronger frozen models without retraining the target, sharing its vocabulary, or accessing target-model logits.
- Converts the tuned/reference log-probability difference into dense reward and value estimates for chunk-level beam search; improves Llama-3-70B from **34.42%** to **39.09%** and black-box GPT-3.5 from **16.00%** to **20.07%**.

A Neural-Symbolic Approach to Natural Language Understanding First Author · EMNLP 2022

- Addresses the difficulty of neural models in reliable exact numerical reasoning by unifying neural analogical inference with executable symbolic reasoning.
- Proposes a gated neural-symbolic processor that selects between neural predictions and executable programs; creates **32,011** DROP and **722** AWPNLI program annotations, reaching **84.01 F1** and **92.24%** accuracy.

Other Publications

Scaling Model-Generated Distillation Data Can Make Latent Teacher Traits More Recoverable Second Author · 2026 · Ongoing Work

- Studies whether scaling independently generated distillation data makes latent teacher traits more recoverable in student models, with evidence across model families, trait types, and cross-model transfer settings.

Native Reasoning Models: Training Language Models to Reason on Unverifiable Data Second Author · ICLR 2026

- Proposes Native Reasoning Training (NRT), which uses standard question-answer pairs to generate latent reasoning traces and intrinsically rewards paths that increase the likelihood of the correct answer, without expert demonstrations or external verifiers; it outperforms SFT and prior verifier-free RL methods on Llama and Mistral models.

Adaptive Prompt Optimization for Open-Ended Tasks: Uncertainty Preference as a Secondary Signal Second Author · ACL 2026

- Uses semantic entropy to estimate task uncertainty and adapt prompt selection: high-entropy candidates for creative tasks and low-entropy candidates for conservative tasks; the method is training-free, supports black-box models, and integrates with existing prompt optimizers.

Emergent Response Planning in LLMs Third Author · ICML 2025

- Uses lightweight probes to predict global structural, content, and behavioral properties from early hidden representations, and analyzes how this planning signal scales with model size and evolves during generation, showing that LLMs encode future response information before full generation.

SafeWork-R1: Coevolving Safety and Intelligence under the AI-45° Law arXiv 2025

- The paper combines progressive safety RL with multi-principled verifiers; I contributed the Section 4.1 inference-time intervention module, extending IVG into Safety, Value, and Knowledge PVM Guidance with input-adaptive routing and raising the safety score from **77.1** to **93.8**.

Skills & Honors

- Python, PyTorch, HuggingFace Transformers, SGLang, SFT/DPO, reward/value modeling, speculative decoding, contrastive decoding, search-based generation.
- Silver Medal, National Olympiad in Informatics (NOI), 2017.