# 柳志轩

lzx993124494@sjtu.edu.cn · 上海

chadlzx · 个人主页 · Google Scholar · OpenReview

## 教育经历

**上海交通大学**，计算机科学与技术博士 2022–2027

计算机学院；导师：乔宇；预计 2027 年 6 月毕业

**上海交通大学**，计算机科学与技术学士（ACM 班，致远学院） 2018–2022

## 科研与实习经历

**上海人工智能实验室**，博士联合培养 2022–至今

- 从事大语言模型后训练与推理研究，重点关注模型行为与安全、推理时对齐及高效生成；以第一作者身份主导多个独立研究课题，负责问题定义、方法设计、系统实现、实验分析与论文写作。
- 研究涵盖隐藏特质迁移机制与控制、推理时偏好引导及对比解码加速，分别面向模型后训练安全、推理时行为控制和推理效率多个层面开展研究。

**字节跳动 AI Lab**，研究实习生 2020.09–2021.06

- 针对神经模型难以稳定完成精确数值推理的问题，研究如何统一神经类比推理与可执行的符号逻辑推理；独立负责选题、方法设计、代码实现、数据与实验组织、结果分析和论文写作。
- 设计共享编码器、神经预测与“程序生成–符号执行”双分支及门控路由，组织 **32,011** 条 DROP 与 **722** 条 AWPNLI 程序监督标注，完成 QA/NLI 模型训练与评估。
- 在 DROP 与 AWPNLI 上分别达到 **84.01 F1** 与 **92.24%** 准确率；相关成果发表于 EMNLP 2022，并完成代码开源。

## 代表性研究项目

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

一作 · NeurIPS 2026 Poster

- 研究 SFT 蒸馏为何会通过与目标特质语义无关的数据，将教师模型的隐藏偏好迁移给学生模型；现有工作能够观察该现象，但缺少可预测的机制解释与定向控制。
- 提出特质方向漂移机制，以样本中的偏好差距解释特质对齐更新的累积，并设计探针空间走廊正则化，将恶意响应迁移率从 **29.55%** 降至 **6.45%**，同时保持主任务性能。

*Decoupled Contrastive Decoding via Expert-Aligned Drafting* 一作 · EMNLP 2026 主会

- 针对对比解码中弱模型串行生成候选造成的高延迟问题，研究如何在不改变原始对比解码目标分布的前提下引入推测式加速。
- 以 expert-aligned EAGLE3 生成候选，将弱模型移至并行验证阶段，并在 SGLang 中实现三模型推理与独立 KV cache 管理；8B 设置获得 **1.65–1.95**× 加速，70B 达到约 **2**×。

*Inference-Time Language Model Alignment via Integrated Value Guidance* 共同一作（第一署名） · EMNLP 2024 Findings

- 针对冻结大模型难以在推理阶段注入偏好、词元级引导缺少全局判断的问题，探索无需继续微调目标模型的细粒度生成对齐方法。
- 以 DPO 隐式价值逐词元修正生成分布，并用显式前缀价值进行分块级搜索，将 AlpacaEval 2.0 长度控制胜率从 **19.51%/25.58%** 提升至 **26.51%/33.75%**。

*Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models*
共同一作（第二署名）· NeurIPS 2024

- 研究如何用能力较弱的小模型对齐更强且冻结的大模型，同时避免重新训练目标模型、共享词表以及访问目标模型 logits 等限制。
- 将小模型的 tuned/reference 对数概率差转化为稠密奖励与价值估计，并据此执行分块级束搜索；将 Llama-3-70B 从 **34.42%** 提升至 **39.09%**，黑盒 GPT-3.5 从 **16.00%** 提升至 **20.07%**。

*A Neural-Symbolic Approach to Natural Language Understanding* 一作 · EMNLP 2022

- 针对神经模型擅长模式匹配却难以稳定完成精确数值推理的问题，研究如何统一神经类比推理与可执行的符号逻辑推理。
- 提出共享编码器、神经预测与“程序生成–符号执行”双分支的神经符号处理器，并以门控网络动态选择结果；构建 **32,011** 条 DROP 与 **722** 条 AWPNLI 程序标注，分别达到 **84.01 F1** 与 **92.24%** 准确率。

## 其他论文

*Scaling Model-Generated Distillation Data Can Make Latent Teacher Traits More Recoverable* 二作 · 2026 · 进行中的工作

- 研究扩大相互独立的模型生成蒸馏数据规模，是否会使教师模型的隐藏特质在学生模型中更易恢复；结果覆盖多种模型家族、特质类型与跨模型迁移设置。

*Native Reasoning Models: Training Language Models to Reason on Unverifiable Data* 二作 · ICLR 2026

- 提出原生推理训练（NRT），仅使用普通问答对让模型自行生成推理轨迹，将推理过程建模为潜变量，并内生奖励更可能导向正确答案的路径；无需专家推理示范或外部验证器，在 Llama、Mistral 模型上优于监督微调与既有无验证器强化学习方法。

*Adaptive Prompt Optimization for Open-Ended Tasks: Uncertainty Preference as a Secondary Signal* 二作 · ACL 2026

- 以语义熵衡量任务不确定性，并据此自适应选择候选提示：创造性任务偏好高熵提示，安全合规等保守任务偏好低熵提示；方法无需训练，支持黑盒模型，并可接入现有提示优化器。

*Emergent Response Planning in LLMs* 三作 · ICML 2025

- 使用轻量探针从早期隐藏表征预测响应长度、推理步数、内容选择、答案置信度与事实一致性等全局属性，并分析该规划能力随模型规模和生成过程的变化，揭示大模型在正式生成前已编码未来响应信息。

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

- SafeWork-R1 通过渐进式安全强化学习与多原则验证器协同提升能力和安全；本人负责第 4.1 节的推理时干预模块，将 IVG 扩展为安全、价值、知识三类原则价值模型，并以输入自适应门控动态路由，将安全得分从 **77.1** 提升至 **93.8**。

## 技能与荣誉

- Python, PyTorch, HuggingFace Transformers, SGLang, SFT/DPO, reward/value modeling, speculative decoding, contrastive decoding, search-based generation.
- 全国青少年信息学奥林匹克竞赛 (NOI) 银牌, 2017。