

钟浩田

haotian7.zhong@gmail.com

教育背景

清华大学

双学位项目

数理基础科学理学学士；软件工程工学学士

中国，北京

2024 年 8 月 – 至今

GPA: **3.73/4.00**

加州大学伯克利分校

校级本科交换生（即将入学）

入选清华大学资助的校级交换项目（每院系 1 个名额）

美国，加利福尼亚州

2027 年 1 月 – 2027 年 5 月

专业技能

编程语言

Python、C++、SQL

机器学习与科研

PyTorch、Hugging Face Transformers、CUDA；大语言模型后训练（监督微调/强化学习）、智能体评测、对比学习

系统与数据

Linux、Git、NumPy、Pandas、Weights & Biases、Docker、SQLite、Slurm、LaTeX

核心课程

面向对象程序设计、计算机语言与程序设计、数据结构、算法、计算机组成原理、数据库系统、深度学习、离散数学、概率论与数理统计

语言能力

中文（母语）；英语（托福 iBT 100，口语 26）

科研经历

科研实习生——AI 水印与签名机制鲁棒性

2026 年 6 月 – 至今

导师：步宇恒教授

加州大学圣塔芭芭拉分校计算机科学系

- 研究 AI 生成内容的水印与签名机制在常见后处理及对抗变换下的鲁棒性。
- 搭建涵盖水印/签名生成、攻击模拟、验证与鲁棒性度量的完整实验流程。
- 分析攻防权衡与失效案例，提高内容溯源机制在实际部署环境中的可靠性。

本科生研究员——大语言模型智能体提示词注入防御

2026 年 2 月 – 至今

第一作者；导师：董胤蓬教授

清华大学人工智能学院，北京

- 提出一种非侵入式推理时防御方法，通过检测工具调用前与工具响应后隐藏状态间的意图漂移，识别间接提示词注入攻击。
- 基于 Qwen3-8B 构建工具调用智能体流程，在 Transformer 最后 4 层设置钩子；构造干净/注入配对轨迹，并实现余弦探针、对比学习探针和攻击成功自动标注。
- 构建包含 1,220 项任务的基础基准和覆盖 10 个应用领域、4,000 个样本的跨领域评测集；评估任务、用户工具、攻击者及跨数据集泛化能力。
- 自动化生成注入攻击，并在含 3,072 种配置的策略空间中进行遗传搜索；第一作者项目。

科研实习生——大语言模型智能体安全

2026 年 2 月 – 2026 年 5 月

腾讯玄武实验室

远程

- 与团队探索早期防御方案，并使用腾讯提供的算力支持以上清华项目的大规模实验。

AI for Science 科研助理——晶体结构大语言模型后训练

2025 年 8 月 – 2026 年 2 月

导师：刘洋教授

清华大学计算机科学与技术系

- 参与晶体结构预测的三阶段大语言模型后训练流程：科学数据准备、监督微调和基于强化学习的优化。
- 收集、预处理并清洗晶体结构数据；对齐提示词、模型输入与训练目标输出，支持训练及下游评测。
- 设计面向推理的奖励函数，探索奖励塑形与提示词对齐策略，提升模型的推理和预测性能。

Understanding Is Done Early: A Depth Division of Labor in Large Language Models and Its Use for Unbounded-Context Memory

共同第二作者

arXiv:2607.28263 2026 年 7 月

- 通过逐层探针、因果截断和分层深度扫描研究大语言模型的层深分工，发现语义信息在较浅层即可被读取，而上层会进一步将表征特化以服务预测。
- 共同开发 CoMem：一种基于 Qwen3-8B 的记忆架构，缓存中间残差状态，检索固定的 Top-12 文本块，并使用自蒸馏的秩 32 LoRA 对查询相关的上层网络进行重计算。
- 在 5 个长上下文与对话记忆基准上进行评测：RULER 得分 97.05，LoCoMo 得分 38.27 (KV-Direct 为 34.59)；在无需适配器的 H20 128K 测试中，显存占用由 89.36 GB 降至 18.26 GB，预填充速度提升 7.83 $\times$ 。

FinAgent OS：多智能体财富管理系统

2026 年 5 月 – 2026 年 6 月

小组项目，软件工程课程大作业

- 基于 LangGraph 与 FastAPI 构建多智能体系统，协同完成用户画像、资产筛选、市场择时、批判性审查与合规审计，覆盖 9 类资产、54 种投资标的。
- 将大语言模型推理与确定性量化计算解耦；实现马科维茨优化、风险平价、3,000 路径蒙特卡洛模拟、VaR/CVaR、压力测试、回测及绩效归因。
- 接入 AkShare 与 yfinance 数据，实现 SQLite 会话持久化、Docker 部署及 GitHub Actions 持续集成，并保持 269 项测试全部通过。

TASIVE：基于 Transformer 孪生网络的签名验证

2025 年 11 月 – 2026 年 1 月

主要开发者，深度学习课程大作业

- 开发 Swin-T 孪生网络，配备轻量级 MLP 投影头、 L_2 归一化的 128 维嵌入及缩放余弦二元交叉熵目标函数。
- 设计严格的书写者互斥训练/验证划分，避免身份信息泄漏；在包含 5,748 个签名对的固定全配对验证集上评测。
- 准确率达 99.98%，ROC-AUC 达 1.0000，EER 为 0.0000；在相同评测协议下优于复现的 SigNet 基线。

2025 年清华大学高性能计算大赛

团队排名第 10 名（共 40 支队伍）