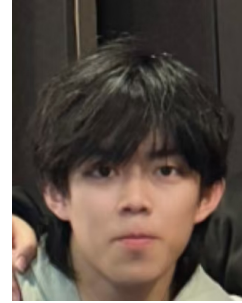


# 张桂嘉

手机：18820183845 · 邮箱：271660183@qq.com

男 · 广东深圳



## 教育背景

深圳大学，电子信息工程拔尖创新实验班（院士班），本科 2022.09 - 2026.06

- 核心课程：机器学习，人工智能，计算机网络，数据结构，数学分析，高等代数，随机过程
- 获得荣誉：院士奖学金（1w），文体之星（1k）

## 研究经历

深圳大学可视计算研究中心（VCC），研究实习生 2024.10 - 2025.3

- 在学长的指导下，深入学习了 NLP 和 LLM 领域，为个人未来的研究和职业发展奠定基础。
- 目前进展：通过文献阅读、理论学习和上手实验，初步了解 NLP 和 LLM 的最新进展，主动学习 Embedding 技术，并通过实战和网课资料加深理解，积极与研究生学长和老师探讨大模型技术的应用和未来发展创新点。
- 学习成果：掌握了 NLP 的基础知识，理解了 embedding 在 LLM 的关键作用，并学会了使用 Python、PyTorch、huggingface 和 GitHub 来帮助实现和训练模型。

清华大学人机交互实验室，remote intern 2025.3 - 至今

- 在清华大学博士学长指导下，参与到他们的实际项目之中，主要研究 CoT、Agent 和 RAG，这些项目在学术和工业界都有发展潜力，为我后续发展奠定了基础。
- 目前进展：通过实际体验 RAG 和与学长交流，我逐渐了解到 Base RAG 的缺陷和不足，并逐步学习结合知识图谱的 RAG 方法，包括 GraphRAG、KAG 和 LightRAG 等。同时，也跟学长交流了解 CoT 和智能体相关的内容。
- 学习成果：掌握了 KAG 和 CoT 的基础知识，同时意识到现有的 RAG 和 CoT 在解决实际问题上还有很长的路要走，许多方法未能很好地解决 RAG 的根本问题，CoT 也多停留在理论层面，其本质结构和思维模式仍未有改变。
- 远期规划：参与到课题组的实际项目，融合学术和工业界优点，改进现有 RAG 不足。同时与学长及其他实习生合作撰写学术论文，提升个人学术影响力。

目前 onsite 实习的话，是在南方科技大学-华为诺亚方舟可信人工智能实验室跟着研究生学长学姐学习

## 得奖经历

- 第三作者，《A Diffusion-Based Approach for Restoring Face-swapped Images》（APSIPA ASC 2024）
- 第三作者，《Document image tampering localization based on foundation models》（ICDAR 2025 投稿中）
- 华数杯全国大学生数学建模竞赛：全国二等奖。
- 全国大学生数学建模竞赛：广东省三等奖。
- 中国高校计算机大赛-网络技术挑战赛：广东省三等奖。

## 技能特长

- 掌握 Python，学习了数据结构与算法等内容，对 AI、NLP 和 LLM 有一定的了解。
- 了解并使用过 PyTorch、PaddlePaddle、huggingface 等库。
- 对大语言模型（LLAMA，QWen，chatGLM）和 LLM 框架（longchain，MetaChat）有一定的了解
- 通过 CET-4（500+ 分）且班级为全英教育，具备一定的英文论文阅读能力。
- 在班级担任组织委员，同时在多次比赛中担任队长，具有良好的团队协作精神和沟通能力。

## 自我评价

作为一名对自然语言处理（NLP）和大型语言模型（LLM）充满热情的学徒，我不断追求掌握新知识，以保证我能迅速了解该领域的前沿内容。在日常学习和工作中，我展现出高度的责任感并能迅速适应新环境和挑战。我熟练使用 GitHub 和 ChatGPT 等工具来快速掌握新技能和知识，以此来提高工作效率。面对复杂问题，我勇于挑战，积极寻找解决方案。同时，我保持乐观开朗的态度和出色的抗压能力，即使在高压环境下也能保持冷静，确保任务的高效完成。

## 项目经历

---

### 稳健联邦推荐：异常数据干扰下的智能推荐 | 2024.06 - 2025.06

作为队长，负责项目的构想、组队、与指导老师对接、答辩及后续项目推动。在项目中，我们以用户隐私和稳健性为切入点，提出了稳健联邦推荐这一想法，该项目成功获得了学校的认可和资助，项目最终成功申报为校级立项。

### langchain+qwen1.5-7b-chat 搭建本地 RAG 系统 | 2024/10

这是一个基于 langchain 和 qwen1.5-7b-chat 搭建而成的本地 RAG 系统，旨在通过结合语言模型和信息检索技术，生成更准确且与上下文相关的输出，通过该项目我学习到了如何解决知识局限和模型容易出现幻觉等问题。

### 基于 unsloth 微调 qwen2.5 的问答系统优化 | 2025/3

在 NLP 领域，预训练模型的微调能提升其在特定领域的问答性能。本项目旨在利用 unsloth 框架对 qwen2.5 模型进行微调，测试其回答问题的能力是否得到提升，数据集是使用通过 kimi-128k 蒸馏《吞噬星空》小说得到。项目分别使用了 0.5B、1.5B、3B、7B 来进行微调，其主要目的是测试多大参数下的模型可以有效的学习到专用数据集里的内容，经检测，0.5b 丢失了大量关键信息，其余 3 个模型都能较好地完成任务。通过该项目，我深入学习到 unsloth 框架的使用和模型微调技术，也积累了优化问答系统在特定领域的宝贵经验，为未来进一步发展奠定了坚实基础。

## 其他信息

---

- 2025/1/12-20 参与了北京大学国际机器学习中心文再文老师举办的“AI4MATH 数学形式化证明”培训班。培训班主要讲解了 Lean 语言的基本使用。此外，培训班上北大文老师团队介绍了他们在凸优化领域和数学形式化的工作，培训班结束有幸跟文老师一对一交流。
- 目前也在学习深度强化学习、分布式算法、多模态 AI 等内容, 主要通过赵世钰老师的《强化学习的数学原理》、王树森老师的《深度强化学习》、王金桥老师的《多模态人工智能》进行学习
- 有过几段初创公司的实习经验，主要做的是 comfyflow 工作流的搭建、vtuber 虚拟数字人的开发以及大模型课程的设计，此外也跟创业团队交流探讨过大模型课程设计创业的问题，得到创业团队的高度认可
- 平常的兴趣爱好主要有散步、爬山、篮球、听歌和骑车！