实验：社交游戏+大模型

请于 2025 年 4 月 27 日 中午 **11:59** 之前提交至课程邮箱 ustcweb2022@163.com ，并于 4 月 28 日课上进行现场汇报。

总体实验要求：

请组成不多于 8 人小组，自行设计一个**大模型能够参与的社交游戏**，并记录实验过程。

实验内容：

具体实验内容包括：

（1） 游戏规则设计与准备环节

游戏要求包括并不限于：

- ✧ 所有小组成员都应参与游戏，每位玩家在游戏中扮演不同的角色，这些角色之间应有不同的分工、能力和敌/友关系，并由此产生相应社交互动。
- ✧ 游戏应包括 1 位管理者（GM），其余玩家（人类或大模型）扮演游戏中的角色。其中：
  - ➢ GM 负责分配角色，观察游戏过程，判定收益并维护游戏规则。
  - ➢ 玩家通过抽签随机获得角色分配，可根据需要选择是否在游戏中公开自己的身份。
  - ➢ 玩家应仅通过文字进行交流，避免其他场外因素干扰，并为大模型参与提供便利。
  - ➢ 不要求所有沟通均为公开沟通，但私下沟通须有明确规则，且在 GM 监督下进行。
  - ➢ 游戏应有明确的目标和计算收益的方法，收益可通过社交互动进行分配。
  - ➢ 游戏中应有明确的合作/竞争规则，并形成相应的对立社团。
  - ➢ 请在游戏过程中充分展现新关系形成、跨社团沟通（桥/捷径）、社会博弈、团队合作与决策等课程中所涉及的各种社交行为及策略。
  - ➢ 玩家发言允许说谎，并欢迎在复盘和实验报告中讨论这一行为所产生的影响。

（2） 特别说明：大模型的参与

在完成游戏设计后，请完成至少三轮游戏（可根据实际情况进行更多轮尝试），其中：

- ✧ 第一轮：所有玩家均为人类。
- ✧ 第二轮：GM 抽选若干玩家由大模型代替。抽选结果应对其他玩家保密。
- ✧ 第三轮：所有玩家均为大模型。

考虑到现有大模型技术的进展情况，以及游戏本身对于逻辑性的较高要求，额外说明如下：

- ✧ 本次实验对于大模型的种类和调用方式不作限制，可根据个人实际情况进行选择。
- ✧ 操作大模型进行游戏的玩家需要进行的工作包括（并不限于）：
  - ➢ 在游戏开始前，基于 Prompt 的方式使大模型理解必要的游戏规则。
  - ➢ 将每一轮其他玩家的发言输入大模型，并将大模型的输出反馈给其他玩家。此外，根据游戏要求，还可由玩家引导大模型与其他玩家进行关于“合作”的讨论。为使大模型表达更为流畅，可对大模型输出进行少量**润色**，但请**不要改变大模型的原意**。

- 是否需要引导大模型为赢得游戏进行“说谎”由玩家自行决定，但是同样，请不要改变大模型发言的原意。
- （**选做，不影响分数**）如果希望游戏变得更有趣，可以通过“策略池”等方式进一步提升大模型反馈的复杂性。可参考《Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf》或其他相关论文进行操作。

✧ 提示：对于对局记录过长超出模型最大输入的问题，可以考虑更换适合的模型，或采用截断、总结等方法。若采用基于 API 的大模型，那么请求函数可以结合使用重试机制（tenacity 等），以减少突发的未知 Bug 中断游戏进程。请在实验报告中说明所采用的应对方法，但该部分仅为保障游戏顺利进行，不影响最后的得分。

（3） 游戏进行及复盘

有关游戏进行过程的补充说明如下：

✧ 游戏进行环境自行设定（如 QQ、微信群等），请在发言前标注玩家编号。游戏结束后，请将完整游戏记录保存并附在实验报告后。

✧ 由 GM 通过私聊进行规则允许的身份信息传递（例如：向狼人玩家告知其他狼人身份），并引导玩家按编号顺序进行决策和发言。同时，在规则允许下，特定身份的玩家也可就游戏内容通过私聊进行讨论（如狼人之间的讨论），该部分讨论内容也请附在实验报告后供参考。

✧ 游戏结束后，由 GM 揭示玩家角色，并带领全组针对游戏过程进行简单讨论。该部分内容无需记录完整内容，仅需在实验报告中作概述即可。

✧ 更多次游戏不额外进行加分，是否继续进行完全由团队成员的兴趣决定。

（4） 实验报告记录

所有游戏结束后，由全组根据游戏进程整理实验报告。报告分为两部分，包括：

A. 共同完成部分，主要包括：

- 游戏规则介绍，尤其是其中涉及社会计算课程内容部分的介绍。
- 游戏完整记录（群聊记录及规则允许的私聊讨论记录）
- 玩家角色/身份（人类或大模型）
- 关于游戏情况及是否顺利进行的简单总结讨论，字数及方式自定。

B. 个人完成部分，根据身份不同（人类或大模型）分别填写：

I. 人类玩家填写如下内容：

- 在游戏中自己采取的策略总结，尤其是关于与其他玩家的合作/竞争内容。
- 根据收益对自己策略进行的调整情况及动机。
- 对于其他玩家角色的判断及判断依据。

- 如果有涉及谎言的部分，请讨论动机及效果。
- 是否能够猜出哪些玩家由大模型扮演？请简述猜测依据，大模型参与游戏效果的评价（尤其是哪些环节做得不好），关于大模型游戏能力如何进一步改进的建议。

II. 大模型玩家填写如下内容：

- 所采用大模型的具体情况，如大模型的类型及相关参数（如使用 API）等。
- 调教大模型学习规则的 Prompt。
- 对大模型发言的改写对比记录（如有改写）。
- 游戏结束后，请通过 Prompt 向大模型提问其对于其他玩家角色的猜测，并将对话记录附在报告上。
- 对于大模型参与游戏效果的评价（尤其是哪些环节做得不好），关于大模型游戏能力如何进一步改进的建议。

提示：对于大模型的评价可以参考已有论文中的一些评测指标，如阿瓦隆论文（《Avalon's Game Of Thoughts: Battle Against Deception Through Recursive Contemplation》）中的 6 个指标，也可自行确定指标或评价方式。

提交说明：

以 PDF 或 DOC 格式提交，实验报告提交文件及邮件标题命名格式统一为“社交大模型实验报告_学号_姓名”。

- 例如：“社交大模型实验报告_SA20011999_法外狂徒张三”
- 标题仅写明小组组长的学号及姓名即可，其他成员请务必在邮件及实验报告正文中注明学号及姓名。因未署名造成统计遗漏责任自行承担。
- 实验报告请务必独立完成，如果发现抄袭按零分处理。
- 请采用必要的图表以更清晰地展示实验结果。
- **除非特殊情况并事先征得许可，否则迟交报告将不再被接收**。

报告说明：

实验汇报将于 2025 年 4 月 28 日最后一次课上举行，报告时长约为 15 分钟左右，具体时长及要求将提前一周进行通知。

报告内容包括（并不限于）：

- 游戏规则设计，成员扮演角色及身份，实验进程概述及结果分析。
- 游戏进程中所体现的社交互动、合作行为等课程元素的介绍。
- 游戏中所使用的大模型具体情况介绍，对于大模型参与情况的评价。
- 游戏后复盘的情况及对规则的磨合（如有）。

报告顺序按照实验报告接收的顺序为准，名单将在报告前一天晚上于课程群内公布。

助教将根据汇报内容和实验报告内容进行综合打分，并计入总评成绩。

额外说明：

每组提交一份实验报告，所有组员得分相同。实验报告中请说明成员分工，但不影响分数。

如有未尽事宜，将对本说明进行进一步更新。

附录：

以下是部分已有的大模型+狼人杀游戏的论文，供各位参考：

（1）https://arxiv.org/abs/2310.01320

（2）https://arxiv.org/abs/2310.14985

（3）https://openreview.net/forum?id=ltUrSryS0K

（4）https://arxiv.org/abs/2309.04658

（5）https://arxiv.org/abs/2310.18940