实验：社交游戏+大模型

请于 2024 年 4 月 28 日 **11:59** 之前提交至课程邮箱 ustcweb2022@163.com ，并于 4 月 29 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**大模型能够参与的狼人杀游戏**，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

该环节的目的在于设计一款大模型能够参与的社交游戏，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社交关系，并由此产生相应的社交互动。对不同角色的分配应随机进行，游戏过程中不应暴露玩家所扮演的角色。
- 游戏应包括 1 位管理者（DM）和若干名玩家（人类或大模型）。其中：
  - 由 DM 随机抽取**不少于** 2 名玩家，由该玩家操纵大模型代替本人参加。
  - 其他玩家**可自主选择**由本人参加或者由大模型参加，并在游戏结束后告知 DM。
  - 大模型玩家的身份在每一轮游戏结束后再公开，但**找出哪位玩家由大模型扮演并不是游戏目标之一，请不要因为玩家身份（人类或大模型）影响游戏进程**。
  - 由于大模型玩家的存在，建议游戏线上进行。
- 游戏应有明确的目标，可采用狼人杀原版胜利条件，也可以自行改进。
- 可根据团队意愿修改游戏规则，并请在实验报告中进行说明。
- 请在游戏过程中充分展现逻辑推理、社会博弈、团队合作与信任等行为。
- 玩家发言允许说谎，并欢迎在游戏后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

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。考虑到目前大模型存在“胡言乱语”“不说人话”的情况，可对大模型的输出进行一定的**改写**，但请**不要改变大模型的原意**。
  - （**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。
  - 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。
- 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

在完成游戏规则设计与大模型准备后，请基于上述设定进行至少 1 次游戏，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 DM 通过私聊进行必要的身份信息传递（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，特定身份的玩家（如狼人）也可就游戏内容通私聊进行讨论，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 DM 揭示玩家角色并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心情决定。

（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
   - ➢ 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结盟/合作内容。
   - ➢ 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，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，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。
- 提交报告的同时请提交**源代码**以供检查。
- **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 2024 年 4 月 29 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。

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。

如有未尽事宜，将对本说明进行进一步更新。

附录：

以下是一些可能用到的信息，供各组参考。

1. 7 人局狼人杀规则（仅供参考，可自行调整）

- 玩家角色。游戏分为狼人阵营和好人阵营，狼人阵营通常由 3 名玩家组成，好人阵营则由 2 名神职玩家（预言家+女巫，或者，预言家+猎人）和 2 名普通村民组成。
- 身份分配。游戏开始时，所有玩家随机分配角色，每个玩家知道自己的角色，但不知道其他玩家的身份。
- 夜晚行动：
  - 狼人在夜晚可以睁眼交流并选择一名村民作为杀害目标；
  - 女巫拥有一瓶解药和一瓶毒药，分别可以救一个人或杀一个人；两瓶药不可以在同一天晚上同时使用，并且解药是不可以自救的；例如，上帝会说（通过打手势/私聊等方式，仅告诉女巫）：“今天晚上被刀的人是一号，是否使用解药，是否使用毒药”，但是如果女巫的解药用完之后，上帝就不可以告诉女巫，今天晚上死的人是谁了（但还是要问“是否使用解药，是否使用毒药”，以保持其他人无法判断女巫是否使用解药或者毒药）；
  - 预言家可以查验一名玩家的阵营身份；只能查看是好人还是狼人，而不能知道具体的身份，比如预言家要查二号的身份，二号如果是个女巫的话，上帝只会跟预言家说二号是个好人，但是不会说二号是个女巫；注意：可以通过打手势/私聊等方式，仅让预言家知道查验结果；
  - 猎人被狼刀或者被投票出局，可以选择开枪带走一名玩家；这个带走一名玩家的技

能，不是一定要发动，也就是可以选择发动，也可以选择不发动

- ◆ 白天行动。所有玩家在白天一起讨论，分析谁可能是狼人，然后进行投票，票数最多的玩家将被放逐出局，如果投票结果为平局，则被投票的玩家不会被放逐，游戏继续进入下一轮夜晚。
- ◆ 游戏的目标是找出所有狼人，或者让所有村民被杀，狼人阵营的胜利条件是消灭所有好人，而好人阵营的胜利条件是驱逐所有狼人。游戏过程中，玩家需要利用自己的特殊能力或逻辑推理来找出对方阵营的成员，并采取相应的行动。

2. 常见的文本生成大模型及部分可参考的论文

在游戏中，玩家可采用常见的大模型客户端（如星火、文心一言、ChatGPT 等），也可根据自身技术积累自行选择其它开源大模型，部分资料可见：

- ✧ https://huggingface.co/models?pipeline_tag=text-generation&sort=trending
- ✧ https://github.com/chatanywhere/GPT_API_free

同时，部分大模型+狼人杀游戏的论文如下：

（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