ChatEval

1.引言

- Q表示存在的问题，A表示答案或者相应的措施，B表示背景或者研究基础
- Q:文本评估耗时耗力
  - A:诞生了基于n-grams的自动评估指标
    - Q:与人类判断的相关性较弱,尤其在设计开放式生成或者需要特定领域专业知识的任务中
- B:大语言模型以及相关训练范式的出现
  - A:LLM-as-a-judge方法
    - LLM可提供与人类判断一致的评价
- B:新的研究表明多个LLM可以通过辩论和合作进一步提升彼此的能力
- A:将多个LLM纳入一个综合小组评设计特定互动机制
  - 提高了生成的反应的真实性
  - 改善了艰巨任务的完成情况
  - 解决和减轻了思维退化（DOT）问题
- Q:认识到人类单一视角可能会导致结果出现偏差和不稳定
  - 最佳做法
    - 多个人类合作评估
      - A:加入协作和迭代系统
        - ChatEval系统
- what we do in conclusion
  - ChatEval
    - 灵感来源
      - 一千个人眼中有一千个哈姆雷特
        - 不同视角形成了对其的全面和多方面评价
      - 集体智慧与认知协同
        - 多认知过程或者系统相配合的综合效应大于单独效应的总和1+1>2
  - communicate strategies
    - 并证明了多样化角色提示的必要性
  - release our library
    - 可拓展性和可组合型
      - 探索者可实施自己独特的交流策略

2.方法

- debater agents
  - 每个LLM都是一个代理，根据给定的代理生成自己的回复，其他代理的回复作为聊天记录，并在模板中替换
- diverse role specification
  - 所有代理有共同的提示模板，但是不同代理有不同个性(角色)
- communication strategy
  - 维护聊天历史记录
    - one-by-one
      - in a set order
    - simultaneous-talk
      - 非线性消除发言顺序的影响
    - simultaneous-talk-with-summarizer
      - 每次辩论迭代结束后，提示额外的LLM总结已有的所有信息，并将总结串联到代理的聊天记录槽中
- 并没有要求所有代理最后达成共识
  - 结果依赖直接比较
    - 取不同代理中的多数
  - 结果需要直接评分
    - 取多代理中的平均得分

3.实验

- 两个基准
  - 开放式问题解答
    - FairEval
      - 人工智能系统对没有预定义或固定答案集的问题提供全面详细类似人类的回答
        - 详见3.2部分
  - 对话回复生成
    - Topical-Chat
- 3.1实验准备
  - GPT系列模型作为LLMs
    - 这是世界上目前最强大，最先进的型号
    - temperature参数为0
      - 保证可重复性
  - 模型通过应用程序接口（API）实现可访问性与易用性
  - 本文的研究采取同质LLMs组
    - 所有LLMs均为GPT-4或者均为ChatGPT
    - 同时也点出了异质模型潜力：探究强弱模型多代理下的合作
- 3.3基准
  - 单代理与多代理都采用位置校准技术
  - ChatEval默认参数　strategy：One-by-One　agant*2　round*2
  - 对开放式问题解答任务
    - 方法也与FairEval进行比较
      - 他们提出的可以提升LLM评估性能的策略如
        - 平衡位置校准（BPC）
        - 多重证据校准（MEC）
  - 对于对话应答生成任务
    - 方法与G-EVAL进行比较
      - 他们采用了CoT与概率加权求和
    - 此外
      - 纳入了基于n-gram的指标结果
        - 如ROUGE和BLEU
      - 纳入了基于嵌入的指标
        - 如BERTScore
- 3.4开放式问题回答的结果
  - 如何评价
    - 准确度Acc.
      - 衡量正确分类的实例占比
    - 卡帕相关系数Kap.
      - 衡量模型与人类注释结果间的一致性
      - 同时考虑了偶然一致的可能性
  - 发现
    - ChatEval可以提升评估过程性能相比于单代理更符合人类偏好
    - ChatEval在ChatGPT与GPT-4中都超过了FairEval的最佳结果
- 3.5生成对话相应的结果
  - 如何评价
    - 计算了回合级的Spearman与Kendall-Tau相关性将其与人类对四个方面的判断相对应
  - 结果发现
    - 纳入了基于n-gram的度量方法与基于嵌入的度量方法在所有评估方面表现都很差
      - 这两个方法很难揭示人类偏好
    - G-EVAL的方法：要求LLM生成中间思维，最后根据概率计算输出分数加权和。这样的结果优于传统度量方法

4.分析

- 默认：以FairEval基准,ChatGPT为分析主干的LLM
- 4.1多样化角色的重要性
  - ChatEval参数　strategy：One-by-One　agant*2　round*2
    - 结果见表3
      - 结果证明了相同角色设计的ChatEval不如多样化角色设计的ChatEval，与单个代理相比不能有效提升性能
      - 凸显了多代理辩论框架中多样化角色提示设计的重要性
- 4.2传播方法的研究
  - 对于三种策略
    - ChatEval参数　agant*3　round*2
      - ChatGPT环境下一对一交流策略优于另外两种策略
      - 另外两种策略表现仍然超过了单代理方法
      - 三种策略的性能差异凸显了不同策略对评估过程有效性的影响
        - 未来可以更全面了解不同交流策略以及如何使用这些交流策略以提升绩效
- 4.3角色编号与讨论回合数的影响
  - 图3a，FairEval数据集中
    - Acc.随角色数量增加而增加
      - 角色编号3和4时为顶峰，峰值62.5%
      - 角色编号5时下降
    - Kap.在角色编号4时最大
    - 凸显了不同角色纳入ChatEval的有效性
  - 图3b中
    - 讨论次数增加并没有带来结果明显的上升趋势
      - 结果指向的现象：持续的讨论会导致成绩的停止甚至下降
      - 本文猜测语境长度的增加会因此降低绩效
- 4.4定性分析
  - 图5
    - 在对于同一个开放式问题的两个回复，三个代理展开了辩论，呈现了很多引人入胜的行为
      - 开场白
      - 替代提案
      - 立场维护
      - 寻求共识
    - 鉴于以上行为
      - ChatEval流程的意义
        - 模拟人类的争论互动
        - 动态的互动展示了语言的丰富与复杂性
        - 捕捉到了单一观点的细微差别
        - 提供了可靠的评估过程，突出了合作对话的变革力量

5.相关工作

- 自动NLG评估
  - 基于n-gram的度量：
    - ROUCE
      - 一组计算机器生成的摘要和参考摘要中n-gram之间重叠量的指标
    - BLEU
      - 根据两个文本中n-grams的共现程度对生成的文本与参考译文比较
    - 局限：
      - 无法捕捉句法和语义的相似性
  - 基于嵌入的度量
    - 词嵌入是词的矢量表示，可以捕捉词的语义属性，意义相似的词具有相似的嵌入
      - 许多研究利用词嵌入评估两端文本间的语义相似性
        - BERTScore
          - 使用来自BERT等转换器模型的上下文化词嵌入
        - BLEURT
          - 利用监督训练数据来提高性能
        - MoverScore
          - 将上下次嵌入与地球移动距离相结合
  - 基于LLM的度量
    - LLM蕴含着从大量训练数据中获得的丰富信息
      - GPTScore
        - 利用条件概率为文本分配代表其质量的分数
      - wang等人
        - 利用ChatGPT作为NLG评估器的潜力
          - 方法是提示ChatGPT直接为文本打分
        - 策划了一个可靠的数据集其中包含成对比较和评价解释
          - 可用于训练基础模型使其成为更可靠的数据集
      - Bai等人
        - 提出了分散式评价
          - 提供更公平的评价结果
      - G-EVAL
        - 提出了概率加权技术来校准单个LLM给出的分数
- 交际代理
  - 通常由ChatGPT或者GPT-4等LLMs驱动
  - 旨在用自然语言与其他代理或者人类用户进行有效互动与交流，共同完成更复杂任务
  - 相关研究
    - Li等人
      - 提出了一个角色扮演的合作代理框架
        - 代理能自主合作解决复杂任务
    - Park等人
      - 创建了一个25个虚拟实体的沙盒环境，配有角色描述和记忆系统
        - 每个智能代理都能自主与其他代理和环境互动模拟人类行为
    - Qian等人
      - 建立了一个基于聊天的软件开发框架
        - 相比于招聘人类程序员能以更低成本完成软件设计并生成可执行软件
    - Liang等人　Du等人
      - 利用沙盒环境策划出了更符合人类偏好的可靠数据集，并训练出了与社会相匹配的LLM
    - Wang等人
      - 自我协作的方法
        - 利用多角色描述引发的单个LLM实现代理间交流
    - Mandi等人
      - 设计用于多个机器人的协作
        - 利用多个LLM加强机器人之间的协调和战略规划
    - 与本文的工作同时Li等人
      - 提出了与本文相似的同伴排名与讨论
        - 但他们通过使用不同模型作为代理来探究不同的评估维度
        - 没有探索其他交流策略

6.结论

- 本文提出的证据表明ChatEval有助于提高有关文本质量的评估性能，使之更符合人类的偏好
- 本文强调了不同角色规范的必要性
- 本文提出了不同交流策略作为ChatEval的组成部分
- 本文对讨论过程的定性分析传达了有关ChatEval如何评估文本的深刻直觉
  - 证实了本文的方法能够提供类似人类判断的综合评估
    - 证明了本文框架的可靠性与有有效性