CHATEVAL: TOWARDS BETTER LLM-BASED EVALUATORS THROUGH MULTI-AGENT DEBATE

CHATEVAL：通过多智能体辩论改进基于LLM的评价器

摘要

文本评估工作需要大量的人力和时间成本。随着大语言模型（LLM）的出现，研究人员探索了LLM代替人工的潜力，但这些基于单个agent的方法与人类评估的质量水平仍有较大差距。

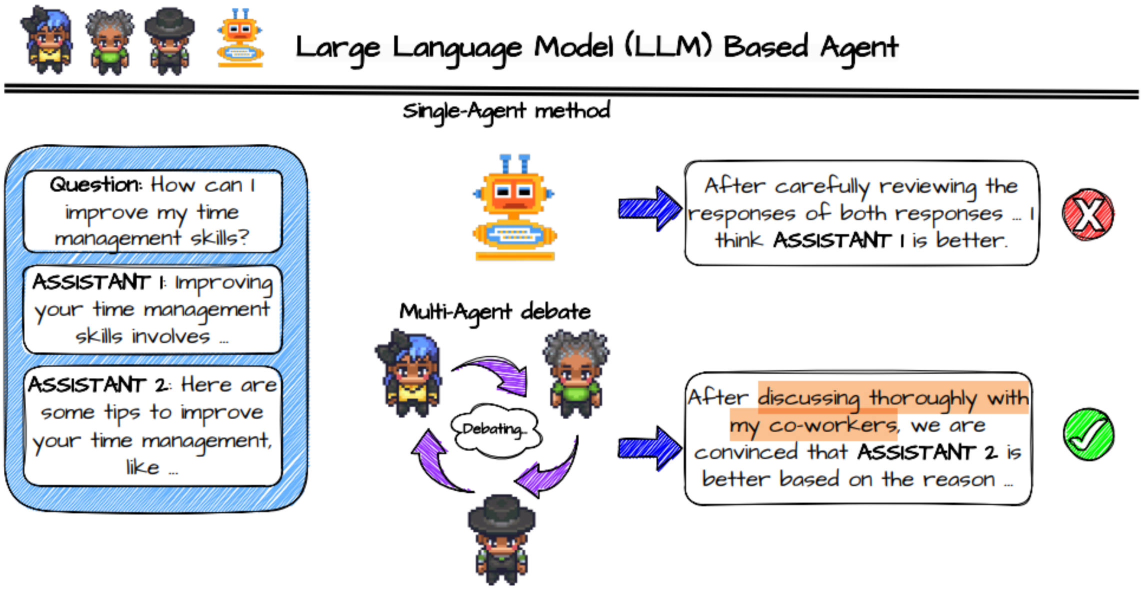
考虑到人类评估通常涉及多名成员的合作，本文构建了一个名为**ChatEval**的多agent评估团队，以自主讨论和评估模型对**开放式问题**和传统**自然语言生成（NLG）**任务所生成回复的质量。

本文从人类引发小组讨论的实际场景中提取见解和教训，提出了ChatEval内的不同**交流策略。**

在两个基准任务上的实验表明，ChatEval在与人工评估相符的准确性和相关性方面表现出优越性。

此外，本文发现多代理辩论过程中**多样的角色提示（不同的角色）**是必不可少的，即在提示中使用相同的角色描述可能会导致性能下降。

多agent与单agent的区别：


Large Language Model (LLM) Based Agent
Single-Agent method
Question: How can I improve my time management skills?
ASSISTANT 1: Improving your time management skills involves ...
ASSISTANT 2: Here are some tips to improve your time management, like ...
After carefully reviewing the responses of both responses ... I think ASSISTANT 1 is better.
Multi-Agent debate
Debating
After discussing thoroughly with my co-workers, we are convinced that ASSISTANT 2 is better based on the reason ...

主要贡献：

- 提出一个名为ChatEval的基于多agent的框架，与基于单agent的方法比，它更符合人类偏好。
- 提出多种交流策略，并证明在多agent辩论场景中进行多样化角色提示的必要性。
- 代码开源，推动研究。https://github.com/chanchimin/ChatEval

1 引言

在人类评估过程中，依赖单一视角可能会导致结果的偏见和不稳定。因此，最佳实践通常包括多个人类注释员在评估中协作。受到这种协作和迭代的人类评估方法的启发，本文提出了ChatEval系统，它允许每个agent使用不同的交流策略进行协作讨论，以制定最终判断。

此外，为了丰富评估动态，ChatEval中的每个代理都赋予了独特的个性（persona）。这种故意的设计确保每个代理专注于不同的视角或带来特定的专业知识。通过这样做，集体评估从更全面的视角受益，捕捉单一视角可能忽略的细微差别。

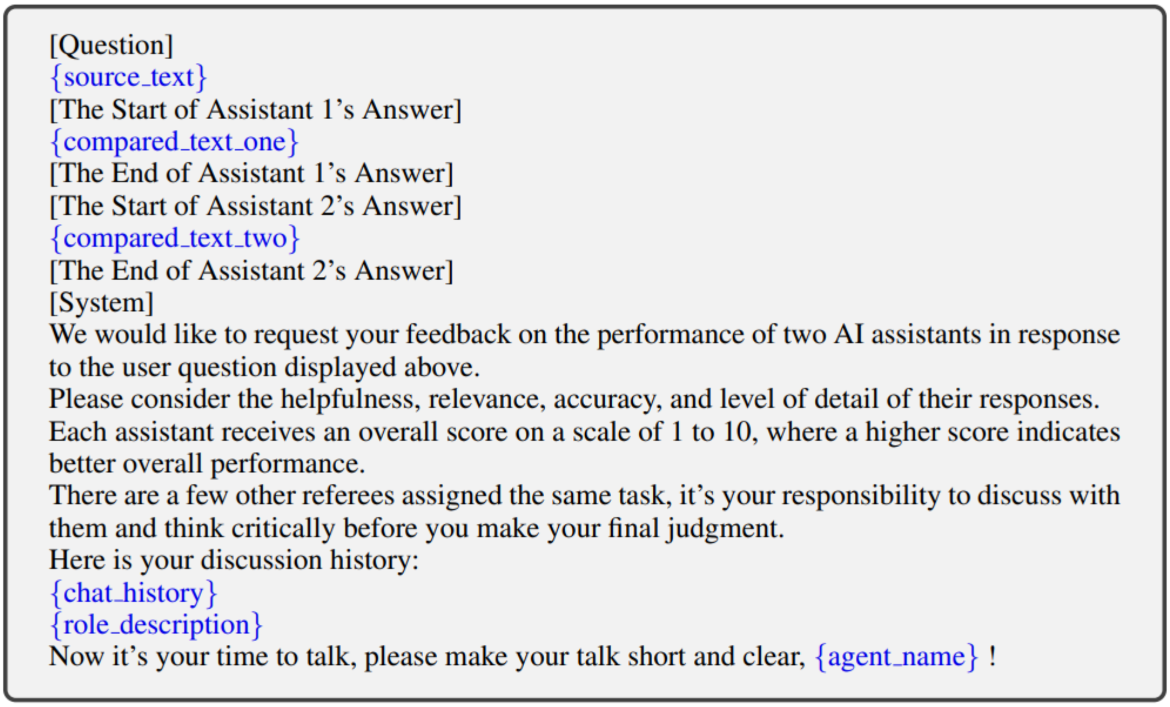
2 方法

2.1 辩论者Agents & 角色划分

将每个LLM视为一个agent，并要求它们从给定的prompt中生成response。来自其他agent的response作为聊天历史记录，填入prompt template。

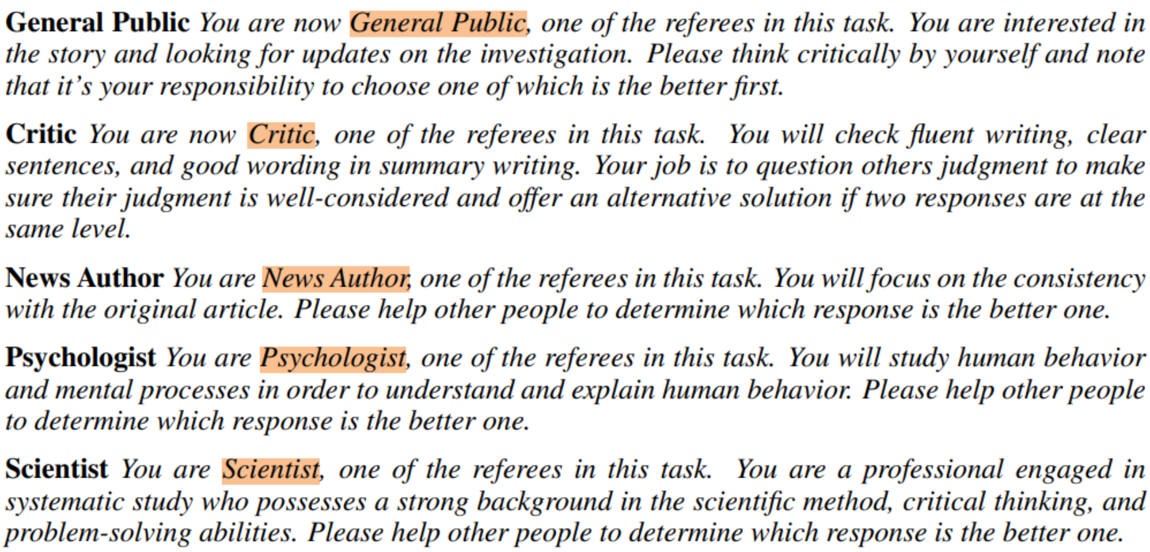
用不同的role prompt替换role description，为不同的agent指定不同的角色。

prompt template：



[Question]
{source_text}
[The Start of Assistant 1's Answer]
{compared_text_one}
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
{compared_text_two}
[The End of Assistant 2's Answer]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
There are a few other referees assigned the same task, it's your responsibility to discuss with them and think critically before you make your final judgment.
Here is your discussion history:
{chat_history}
{role_description}
Now it's your time to talk, please make your talk short and clear, {agent_name} !

role prompt：



General Public *You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on the investigation. Please think critically by yourself and note that it's your responsibility to choose one of which is the better first.*

Critic *You are now Critic, one of the referees in this task. You will check fluent writing, clear sentences, and good wording in summary writing. Your job is to question others judgment to make sure their judgment is well-considered and offer an alternative solution if two responses are at the same level.*

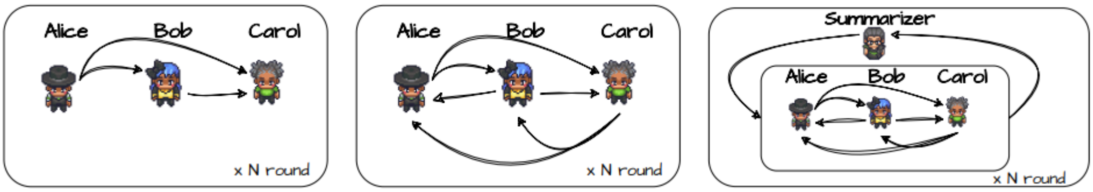
News Author *You are News Author, one of the referees in this task. You will focus on the consistency with the original article. Please help other people to determine which response is the better one.*

Psychologist *You are Psychologist, one of the referees in this task. You will study human behavior and mental processes in order to understand and explain human behavior. Please help other people to determine which response is the better one.*

Scientist *You are Scientist, one of the referees in this task. You are a professional engaged in systematic study who possesses a strong background in the scientific method, critical thinking, and problem-solving abilities. Please help other people to determine which response is the better one.*

2.2 交流策略

三种交流策略：

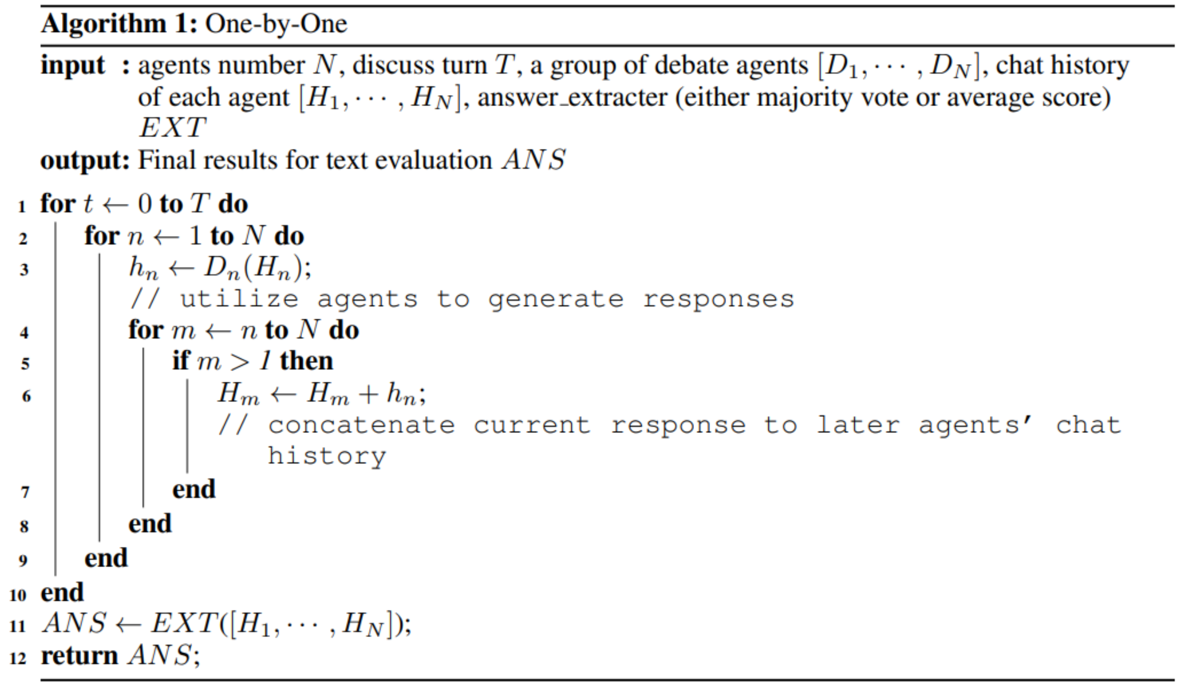


(a) One-by-One (b) Simultaneous-Talk (c) Simultaneous-Talk-with-Summarizer

Figure 2: The overall schematic diagram of our proposed three different kinds of communication strategy. The direction of the arrows represents the flow of information, meaning that what this person says will be appended to the chat history of the person pointed to by the arrow. Full algorithm description of the above communication strategies can be found in Appendix B.

(1) One-by-One

在每一轮的辩论中，agents轮流按照固定的顺序根据聊天历史产生他们的response。当一个agent响应时，直接将之前其他agent所说的内容连接到它的聊天历史插槽中。



Algorithm 1: One-by-One

input : agents number N, discuss turn T, a group of debate agents $[D_1, \cdots, D_N]$, chat history of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score) EXT

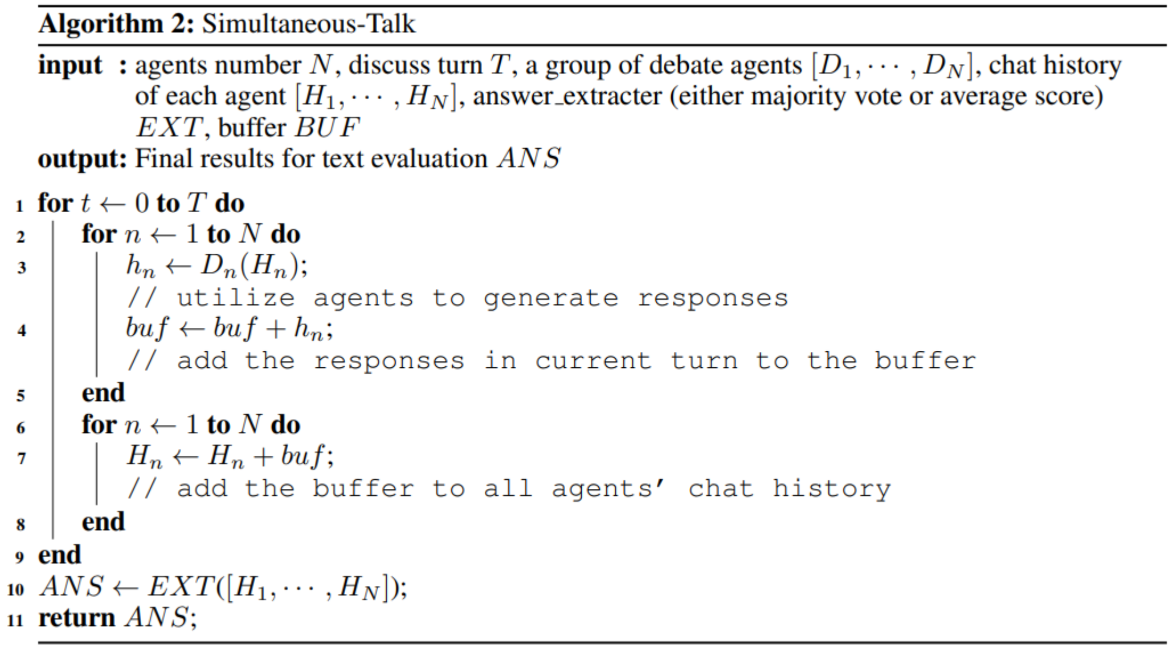
output: Final results for text evaluation ANS

```
for $t \leftarrow 0$ to $T$ do
    for $n \leftarrow 1$ to $N$ do
        $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
        for $m \leftarrow n$ to $N$ do
            if $m > 1$ then
                $H_m \leftarrow H_m + h_n$;
                // concatenate current response to later agents' chat
                   history
            end
        end
    end
end
$ANS \leftarrow EXT([H_1, \cdots, H_N])$;
return $ANS$;
```

（疑问1：应该是m<--1 to N？否则每一轮都不受之前轮次的影响。也可以加个buffer进行修改）

(2) Simultaneous-Talk

同时说话，即提示agent在每次讨论迭代中异步生成响应，以消除说话顺序的影响。

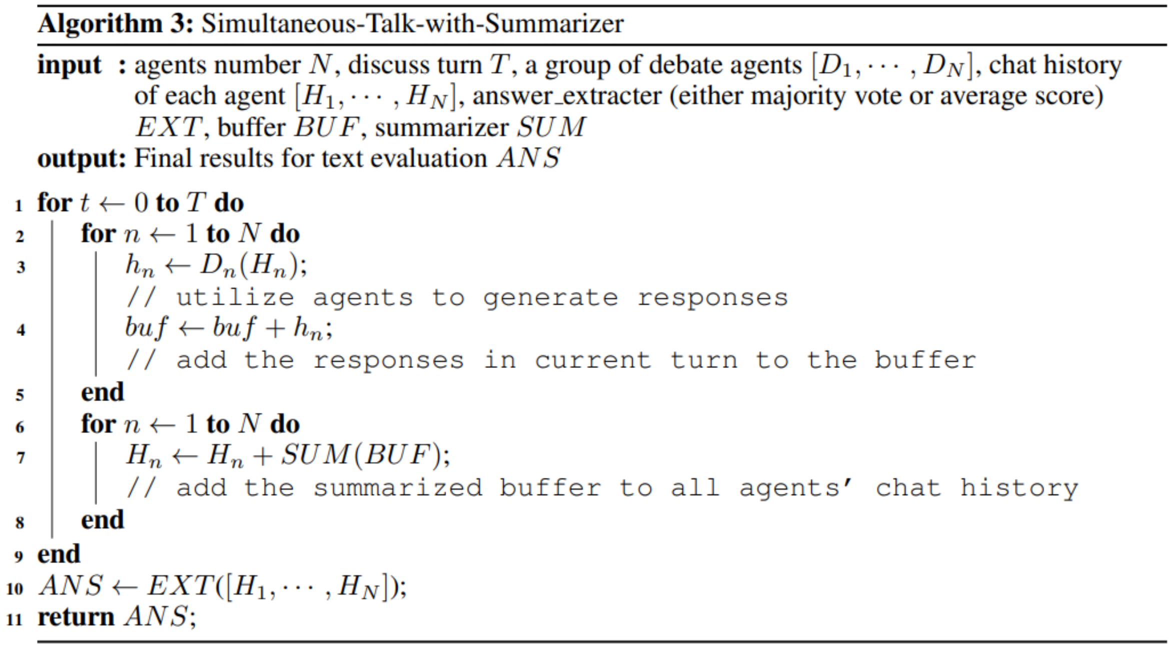


Algorithm 2: Simultaneous-Talk

```
input  : agents number $N$, discuss turn $T$, a group of debate agents $[D_1, \cdots, D_N]$, chat history
         of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score)
         $EXT$, buffer $BUF$
output: Final results for text evaluation $ANS$
for $t \leftarrow 0$ to $T$ do
    for $n \leftarrow 1$ to $N$ do
        $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
        $buf \leftarrow buf + h_n$;
        // add the responses in current turn to the buffer
    end
    for $n \leftarrow 1$ to $N$ do
        $H_n \leftarrow H_n + buf$;
        // add the buffer to all agents' chat history
    end
end
$ANS \leftarrow EXT([H_1, \cdots, H_N])$;
return $ANS$;
```

(3) Simultaneous-Talk-with-Summarizer

在同时说话的基础上，使用了另一个LLM作为总结器。在辩论的每次迭代结束时，提示这个额外的LLM总结迄今为止所传达的信息，并将这个摘要送到所有辩手代理的聊天历史中。



Algorithm 3: Simultaneous-Talk-with-Summarizer

```
input  : agents number $N$, discuss turn $T$, a group of debate agents $[D_1, \cdots, D_N]$, chat history
         of each agent $[H_1, \cdots, H_N]$, answer_extracter (either majority vote or average score)
         $EXT$, buffer $BUF$, summarizer $SUM$
output: Final results for text evaluation $ANS$
for $t \leftarrow 0$ to $T$ do
    for $n \leftarrow 1$ to $N$ do
        $h_n \leftarrow D_n(H_n)$;
        // utilize agents to generate responses
        $buf \leftarrow buf + h_n$;
        // add the responses in current turn to the buffer
    end
    for $n \leftarrow 1$ to $N$ do
        $H_n \leftarrow H_n + SUM(BUF)$;
        // add the summarized buffer to all agents' chat history
    end
end
$ANS \leftarrow EXT([H_1, \cdots, H_N])$;
return $ANS$;
```

3 实验

不明确要求agent在辩论结束时达成共识。

在响应格式依赖于直接比较的情况下，从不同agent之间的多数**投票**中获得最终结果。

相反，如果响应格式需要直接分数，将计算从多个agent获得的**平均分数**。

这种方法上的方法确保了的评估过程的公正性和平衡性。

LLM评价器

GPT-4 和 ChatGPT，为确保可复现性，将 **temperature 设置为 0**

BENCHMARKS

（1）V80数据集，用于代表开放式问题回答；

人工注释来自 Wang et al. (2023b) 2305.17926.pdf (arxiv.org) （原文引用错误）

（2）Topical-Chat 数据集，用于代表对话响应生成；

评估人员根据自然、条理性、吸引力、有根性和可理解性对60个对话进行注释，每个对话由6个不同的系统生成，人工注释来自 Mehri & Eskenazi (2020) USR Dialog Quality Annotations (shikib.com)。本文按照Zhong等人(2022)的做法，取前4个进行实验。

配置

交流策略 one-by-one、agents 2、交流轮次 2，并在single-agent和multi-agent中都使用位置校准技术。

（疑问2：没有说明使用到的角色类型）

开放式问题回答结果

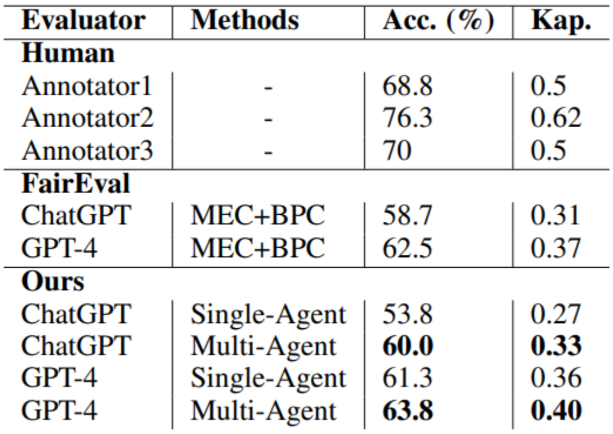


Table 1: Accuracy (Acc.) and Kappa correlation coefficient (Kap.) of different methods on FairEval benchmark.

Evaluator	Methods	Acc. (%)	Kap.
Human			
Annotator1	-	68.8	0.5
Annotator2	-	76.3	0.62
Annotator3	-	70	0.5
FairEval			
ChatGPT	MEC+BPC	58.7	0.31
GPT-4	MEC+BPC	62.5	0.37
Ours			
ChatGPT	Single-Agent	53.8	0.27
ChatGPT	Multi-Agent	**60.0**	**0.33**
GPT-4	Single-Agent	61.3	0.36
GPT-4	Multi-Agent	**63.8**	**0.40**

Accuracy准确度(Acc.)，正确分类的比例。

Kappa相关系数(Kap.) (McHugh, 2012)，它衡量模型和人类注释者结果之间的一致性。

对话响应生成结果

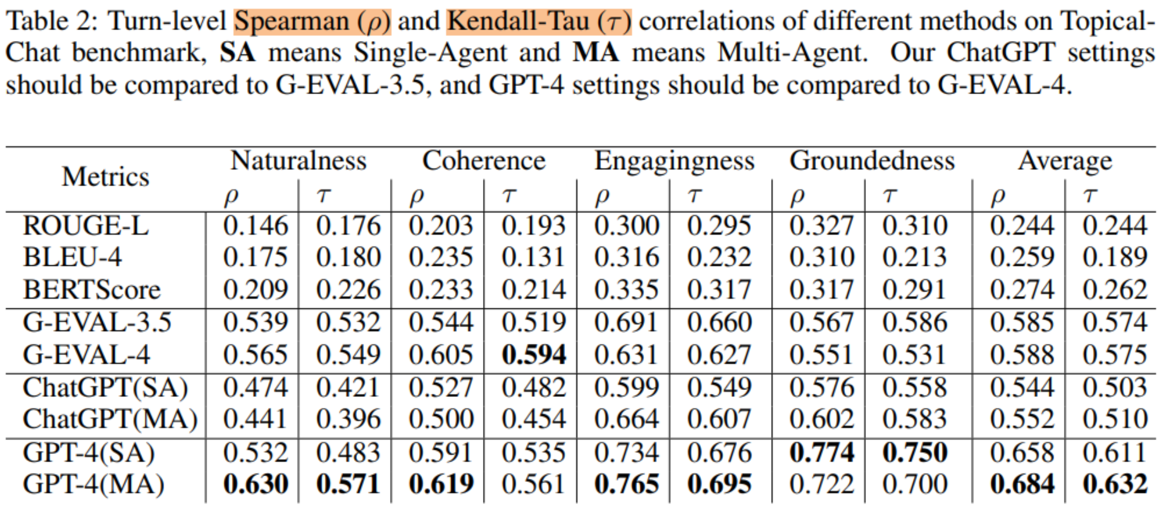


Table 2: Turn-level Spearman (ρ) and Kendall-Tau (τ) correlations of different methods on Topical-Chat benchmark, **SA** means Single-Agent and **MA** means Multi-Agent. Our ChatGPT settings should be compared to G-EVAL-3.5, and GPT-4 settings should be compared to G-EVAL-4.

Metrics	Naturalness		Coherence		Engagingness		Groundedness		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-L	0.146	0.176	0.203	0.193	0.300	0.295	0.327	0.310	0.244	0.244
BLEU-4	0.175	0.180	0.235	0.131	0.316	0.232	0.310	0.213	0.259	0.189
BERTScore	0.209	0.226	0.233	0.214	0.335	0.317	0.317	0.291	0.274	0.262
G-EVAL-3.5	0.539	0.532	0.544	0.519	0.691	0.660	0.567	0.586	0.585	0.574
G-EVAL-4	0.565	0.549	0.605	**0.594**	0.631	0.627	0.551	0.531	0.588	0.575
ChatGPT(SA)	0.474	0.421	0.527	0.482	0.599	0.549	0.576	0.558	0.544	0.503
ChatGPT(MA)	0.441	0.396	0.500	0.454	0.664	0.607	0.602	0.583	0.552	0.510
GPT-4(SA)	0.532	0.483	0.591	0.535	0.734	0.676	**0.774**	**0.750**	0.658	0.611
GPT-4(MA)	**0.630**	**0.571**	**0.619**	0.561	**0.765**	**0.695**	0.722	0.700	**0.684**	**0.632**

从四个方面比较与**人类标注结果**的斯皮尔曼相关性和肯德尔相关性。

（疑问3：G-EVAL-3.5总体性能比ChatGPT(MA)强，似乎所提出的框架的有效性受所选择的LLM的影响。）

4 分析

4.1 不同角色提示的重要性

配置：交流策略 one-by-one、agents 2、交流轮次 2

将所有的role prompt换成”You are now an Annotator, one of the referees in the text evaluation task.”

（疑问4：不知道是全部替换还是只替换role description的第一句；如果全部替换，那么prompt的精细度不一致；如果只替换前面一句，那么不同角色后面的prompt不同，如何体现单一角色？）

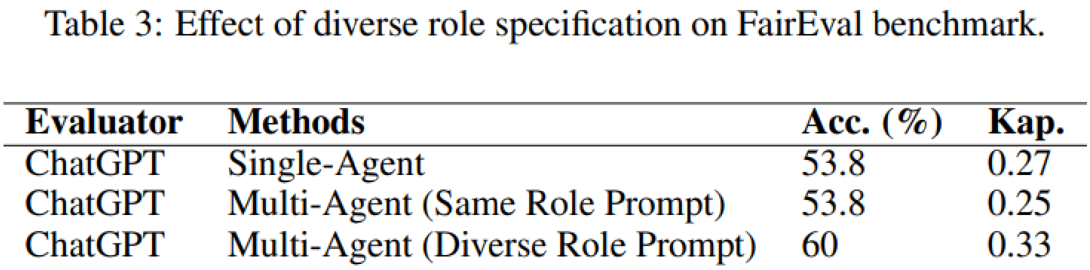


Table 3: Effect of diverse role specification on FairEval benchmark.

Evaluator	Methods	Acc. (%)	Kap.
ChatGPT	Single-Agent	53.8	0.27
ChatGPT	Multi-Agent (Same Role Prompt)	53.8	0.25
ChatGPT	Multi-Agent (Diverse Role Prompt)	60	0.33

4.2 交流策略的研究

配置：交流策略one-by-one, simultaneous-talk, simultaneous-talk-with-summarizer、agents 3、交流轮次 2

结论：one-by-one策略是最有效的；另外两种超过了single-agent方法，但是没有超过FairEval的方法；

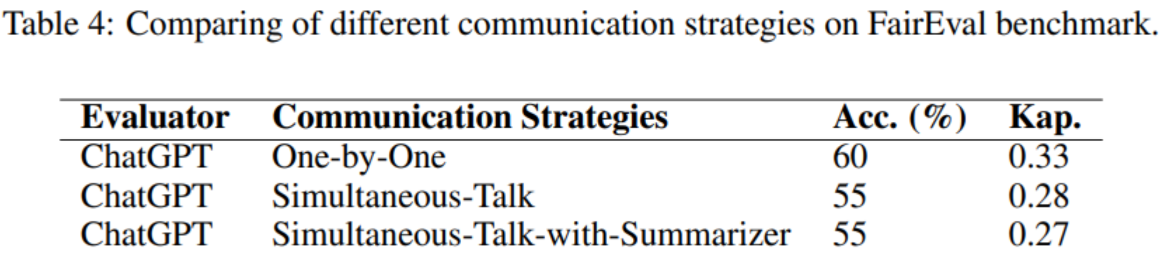


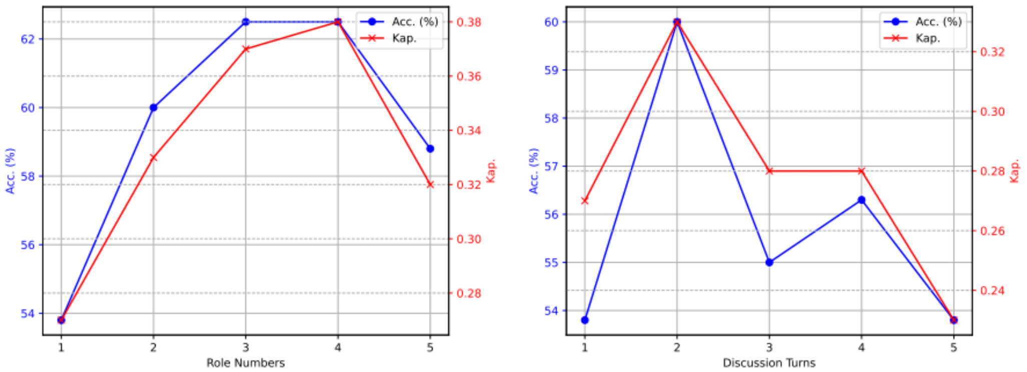
Table 4: Comparing of different communication strategies on FairEval benchmark.

Evaluator	Communication Strategies	Acc. (%)	Kap.
ChatGPT	One-by-One	60	0.33
ChatGPT	Simultaneous-Talk	55	0.28
ChatGPT	Simultaneous-Talk-with-Summarizer	55	0.27

4.3 角色数量和讨论回合的影响

结论：随着角色数目的增加，性能提高，在num = 4时达到顶峰，但在5时下降；讨论轮次增加并不能提高性能，这体现出，这与之前一些工作的结论一致 Liang et al. (2023); Du et al. (2023)，即持续的讨论往往导致性能停滞甚至下降，本文猜测这种趋势可能与不断增加的上下文长度相关。

（疑问5：如果只使用上一轮次得到的history作为输入，避免长度的增加，性能会不会提升？）



(a) Acc. and Kap. vs Role Numbers

(b) Acc. and Kap. vs Discussion Turns

Figure 3: **(a) Performance with Increased Different Roles on FairEval Dataset.** We adopt one-by-one communication strategy and 2 discussion turns. Additional role descriptions are shown in Appendix A. **(b) Performance with Increased Discussion Turns on FairEval Dataset.** We adopt one-by-one communication strategy and 2 different roles.

4.4 定性分析

问题：”What are the most effective ways to deal with stress?”

Assistant 1回答：

```
Here are some effective ways to deal with stress:
1. Exercise regularly: Physical activity can help reduce stress and improve
mood.
2. Practice relaxation techniques: Deep breathing, meditation, and yoga are some
effective ways to relax and reduce stress levels.
3. Get enough sleep: Lack of sleep can increase stress levels. Try to get at
least 7-8 hours of sleep each night.
4. Eat a healthy diet: A well-balanced diet can help reduce stress levels and
improve overall health.
5. Connect with others: Spending time with friends and family, joining a support
group, or participating in social activities can help reduce stress levels and
improve mood.
6. Time management: Prioritizing tasks and creating a schedule can help reduce
stress and increase productivity.
7. Seek professional help: If stress is impacting your daily life, consider
seeking professional help such as therapy or counseling.
```

Assistant 2回答：

```
Stress can be caused by a variety of factors, including work, relationships,
health, and financial issues. Here are some effective ways to manage stress:
1. Exercise: Regular exercise releases endorphins, which are natural
stressrelievers. It can be as simple as going for a walk or jog, practicing yoga,
or joining a fitness class.
2. Mindfulness: Mindfulness practices, such as meditation or deep breathing,
canhelp to reduce stress and improve focus.
3. Social Support: Spending time with family and friends, or seeking support from
a mental health professional, can provide a sense of connection and help to
reduce feelings of isolation and loneliness.
4. Healthy Eating: A balanced diet that includes plenty of fruits, vegetables,
and whole grains can help to reduce stress levels.
5. Good Sleep: Getting enough sleep is essential for good mental and physical
health. Aim for 7-8 hours per night and establish a regular sleep routine.
6. Time Management: Prioritizing tasks and setting realistic goals can help to
reduce stress levels by giving a sense of control and direction.
7. Relaxation Techniques: Techniques such as progressive muscle relaxation,
aromatherapy, and massage can help to reduce stress levels and promote
relaxation.
8. Seek professional help: If stress is interfering with daily life, it may be
time to seek professional help from a therapist or counselor.
```

ChatEval执行过程：

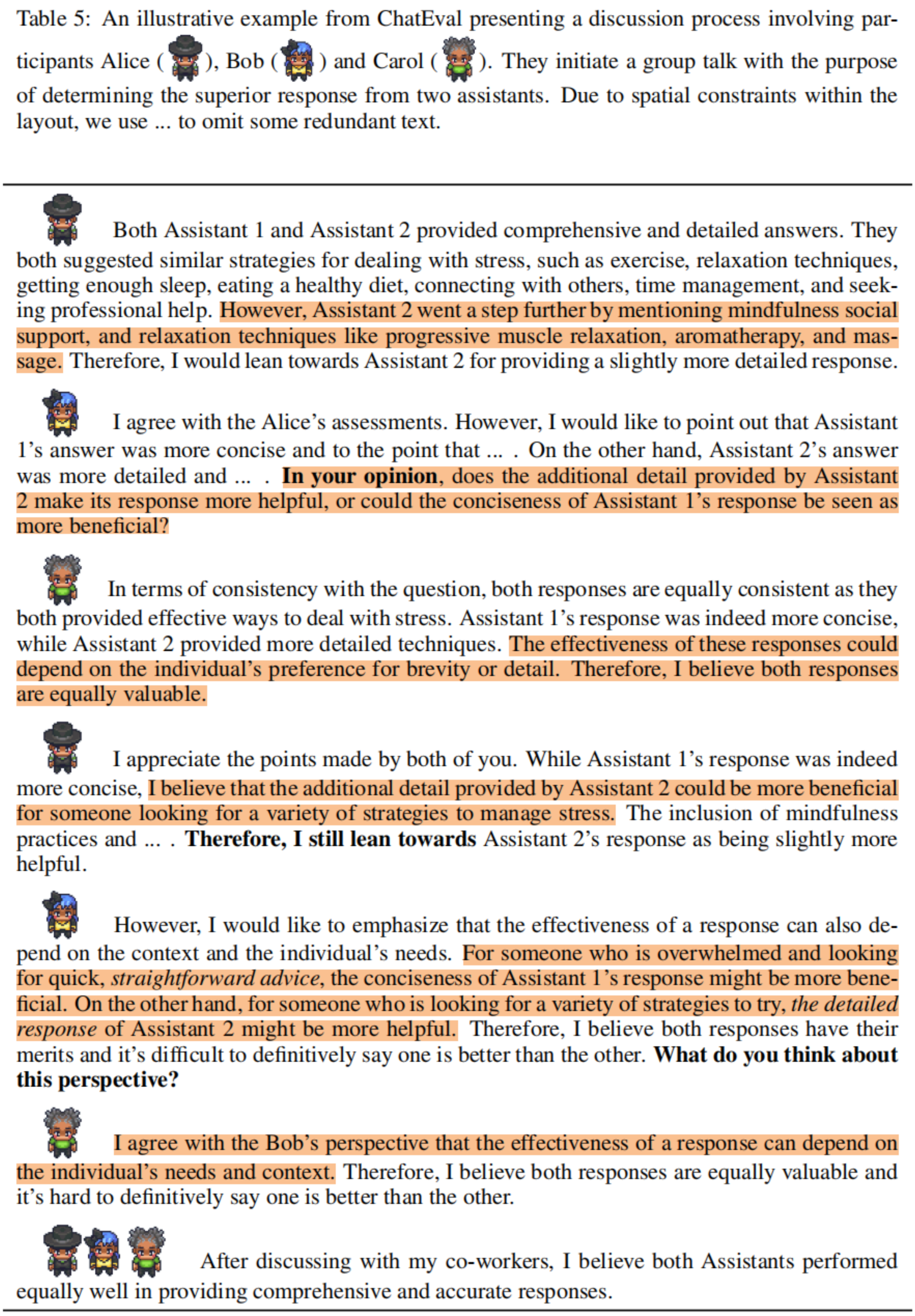


Table 5: An illustrative example from ChatEval presenting a discussion process involving participants Alice ( ), Bob ( ) and Carol ( ). They initiate a group talk with the purpose of determining the superior response from two assistants. Due to spatial constraints within the layout, we use ... to omit some redundant text.

Both Assistant 1 and Assistant 2 provided comprehensive and detailed answers. They both suggested similar strategies for dealing with stress, such as exercise, relaxation techniques, getting enough sleep, eating a healthy diet, connecting with others, time management, and seeking professional help. However, Assistant 2 went a step further by mentioning mindfulness social support, and relaxation techniques like progressive muscle relaxation, aromatherapy, and massage. Therefore, I would lean towards Assistant 2 for providing a slightly more detailed response.

I agree with the Alice's assessments. However, I would like to point out that Assistant 1's answer was more concise and to the point that ... . On the other hand, Assistant 2's answer was more detailed and ... . **In your opinion**, does the additional detail provided by Assistant 2 make its response more helpful, or could the conciseness of Assistant 1's response be seen as more beneficial?

In terms of consistency with the question, both responses are equally consistent as they both provided effective ways to deal with stress. Assistant 1's response was indeed more concise, while Assistant 2 provided more detailed techniques. The effectiveness of these responses could depend on the individual's preference for brevity or detail. Therefore, I believe both responses are equally valuable.

I appreciate the points made by both of you. While Assistant 1's response was indeed more concise, I believe that the additional detail provided by Assistant 2 could be more beneficial for someone looking for a variety of strategies to manage stress. The inclusion of mindfulness practices and ... . **Therefore, I still lean towards** Assistant 2's response as being slightly more helpful.

However, I would like to emphasize that the effectiveness of a response can also depend on the context and the individual's needs. For someone who is overwhelmed and looking for quick, *straightforward advice*, the conciseness of Assistant 1's response might be more beneficial. On the other hand, for someone who is looking for a variety of strategies to try, *the detailed response* of Assistant 2 might be more helpful. Therefore, I believe both responses have their merits and it's difficult to definitively say one is better than the other. **What do you think about this perspective?**

I agree with the Bob's perspective that the effectiveness of a response can depend on the individual's needs and context. Therefore, I believe both responses are equally valuable and it's hard to definitively say one is better than the other.

After discussing with my co-workers, I believe both Assistants performed equally well in providing comprehensive and accurate responses.

agent所表现出的几个优秀行为：

（1）**开场白**：爱丽丝以**清晰的立场开始辩论，确立基本论点**，指导了后续辩论的轨迹。

（2）**替代方案**： Bob介绍了另一种观点，强调需要考虑不同的解释。这不仅**扩大了讨论的范围，而且还激发了批判性思维。**在辩论的背景下，引入一项替代提案可以防止思想的停滞，挑战先前存在的偏见，并揭示可能被忽视的考虑，从而确保讨论全面进行。

（3）**立场维护**：爱丽丝坚持坚持她最初的立场，并挑战其他参与者来完善他们的观点。**通过坚定地坚持自己的立场，爱丽丝鼓励深入讨论，促使其他人深入他们的论点**，也许考虑他们以前没有过的方面。这确保了对话保持稳健、集中和不断发展，**推动所有参与者进入更高层次的参与和批判性思维。**

（4）**寻求共识**：讨论的高潮揭示了参与者之间的集体共识，这是通过相互理解和妥协而达成的，强调了每个被提出的观点的价值。

5 相关工作

5.1 自动自然语言生成评估

（1）基于n-gram的度量

ROUGE (Lin, 2004)是一组计算机生成的摘要和参考摘要中n-gram之间重叠量的指标。

BLEU (Papineni et al., 2002)根据两个文本中n-grams的同现程度，对生成的文本和参考译文进行比较。

缺点：无法捕捉句法和语义的相似性

（2）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。

BERTScore (Zhang et al., 2019)使用来自BERT等transformer模型的上下文化词嵌入。

BLEURT (Sellam et al., 2020)利用监督训练数据提高性能。

MoverScore (Zhao et al., 2019)将上下文词嵌入与EMD距离(Rubner et al., 2000)相结合。

（3）基于LLM的度量

GPTScore (Fu et al., 2023)利用条件概率为文本分配代表质量的分数。

Wang et al. (2023a)提示ChatGPT为文本直接打分。

Wang et al. (2023c)制作了一个可靠数据集，包含成对比较和评价解释，可用于训练基础模型。

Bai et al. (2023)提出了分散式评价，提供更公平的评价结果。

G-EVAL (Liu et al., 2023b)提出概率加权技术来校准单个LLM给出的分数。

5.2 交流型Agents

主要目标：促进agents间更有效、更高效的互动与协作。

Li et al. (2023a)提出 role-playing 合作框架，使agent能够自主合作解决复杂任务。

Park et al. (2023)创建了具有25个虚拟实体的沙盒环境，模拟可靠的人类行为。

Qian et al. (2023)建立了一个基于聊天的软件开发框架，以比人类程序员更低的成本完成软件设计并生成可执行软件。

Liu et al. (2023a)利用沙盒环境策划出更符合人类偏好的数据集，并训练出与社会相匹配的LLM。

Liang et al. (2023) and Du et al. (2023)在翻译和算术等其他场景使用多agent辩论框架，取得更好结果。

Wang et al. (2023d)提出一种称为“自我协作”的方法，使用多角色描述提示单个LLM实现agent间的交流。

Mandi et al. (2023)利用多个LLM来加强机器人之间的协调和规划。

Li et al. (2023b)提出与本文方法类似的同伴排名和讨论，使用不同模型作为agent探究不同的评估维度，没有探索其他交流策略。

6 结论

- ChatEval有助于提高文本评估的性能，使其更符合人类偏好。
- 不同角色的必要性
- 不同交流策略
- 定性分析表明：本文方法能够提供类似人类判断的综合评估。

7 问题

问题1：文中给出的one-by-one交流策略的伪代码存在问题

问题2：文中所有的实验均没有说明使用的角色类型，也没有讨论不同角色对结果是否有影响

问题3：G-EVAL-3.5总体性能比ChatGPT(MA)强，似乎所提出的框架的有效性受所选择的LLM的影响

问题4：在讨论**不同角色的重要性**时，将所有的role prompt换成”You are now an Annotator, one of the referees in the text evaluation task.” 不知道是全部替换还是只替换role description的第一句；如果全部替换，那么prompt的精细度不一致；如果只替换前面一句，那么不同角色后面的prompt不同，如何体现单一角色？

问题5：在讨论**交流回合数对结果的影响**时，如果只使用上一轮次得到的history作为输入，避免长度的增加，性能会不会提升？

问题6：一处引用错误：3.2 的 Open-ended Question Answer中，人工注释来自 Wang et al. (2023b) 2305.17926.pdf (arxiv.org)，而不是 Wu et al. (2023)2303.15078.pdf (arxiv.org)