**Improving Factuality and Reasoning in Language Models through Multiagent Debate**

通过多代理辩论提高语言模型的真实性和推理能力

**太长不看版：**

本篇论文通过让多个模型实例参考“心智社会（*society of minds*）”的方法来提高大语言模型的真实性和推理能力。作者让多个语言模型就同一问题先各自得出答案，再让这些语言模型读取并批判其他语言模型的答案来更新自己的答案，再多次重复这一辩论过程，最终得到一个收敛的答案。实验结果表明，这一方法能显著提高语言模型在许多不同类型下的答案的准确率。这一发现对提高语言模型的推理能力与事实性提供了一种新的实现方法，也能使语言模型生成的文本更具可靠性。

**摘要**

近些年来大语言模型在语言生成、语言理解、少样本学习中展现出了非凡的能力。大量工作研究如何通过提示词进一步提高模型的表现。在本文中，作者提出一种补充方法，让多个语言模型在多轮中提出自己的答案与推理过程，并与其他模型进行辩论，以此来得出共同的最终的答案。研究结果证明，这种方法能显著增强许多任务中的数学及策略推理，且增强内容的真实与有效性，减少现在模型容易出现的错误答案和幻觉。此方法可以直接应用在现有黑箱模型，并且对所研究的所有任务使用相同的程序和提示。总的来说，“心智社会（*society of minds*）”这种方法有潜力显著提高大语言模型的能力，并能助力语言生成和语言理解的进一步突破。

1. **引言**

近些年来大语言模型通过在互联网上的大量文本语料库训练，但是互联网上提取的自然语言的质量和准确性无法得到保证，因此模型在经过训练后可能自信地产生错误的事实，或者在推理链上做出不可思议的跳跃。最近的大量工作集中于提高语言模型的事实准确和推理能力上。而作者注意到这些工作都是在单个语言模型上进行的，于是，他们受到“*The Society of Mind*”的启发，利用多个单模型分别先得到各自的结果，再让这些模型读取、批判其他模型的结果，最后参考其它模型的结果更新自己的结果，然后重复几轮这个过程，最终得到所有模型共同认可的答案。

在实验过程中，作者发现即使使用相同的基类模型，最初每个模型实例也会给出不同范围的回答。辩论过后，所有模型实例的回复基本上收敛于一个更准确的答案，且这个答案中不太可能包括错误的事实。并且辩论的过程不是从模型群体中放大正确的答案，而是随着辩论的进展从错误的回复中推理出正确的答案。

作者将该方法应用在六种不同的推理、事实和问题回答任务上进行测试，效果均好于单一模型。同时为了能够评估该方法的现实准确率，作者引入了一个著名计算机科学家的传记作为新的基准和数据集，结果发现该方法可以有效的纠正和删除单一模型实例提出的不一致的错误的事实。

该方法的步骤和prompt适用于所有的任务，并且可以使用所有的模型作为基类模型。

1. **多代理辩论中的语言生成**
   1. **多代理语言生成**

首先将问题分别输入给不同的语言模型，并让他们独立解决，得到第一轮每个代理的回复。然后，将所有代理的第一轮回复利用提示词附加到第二轮的提问中，并让每个代理根据第一轮其它代理的回复更新自己的回复。重复这个过程，每个代理通过对其他模型的答案进行批判和借鉴，不断完善自己答案。图1为三轮两个代理辩论的示例，图2为不同长度的辩论prompt。

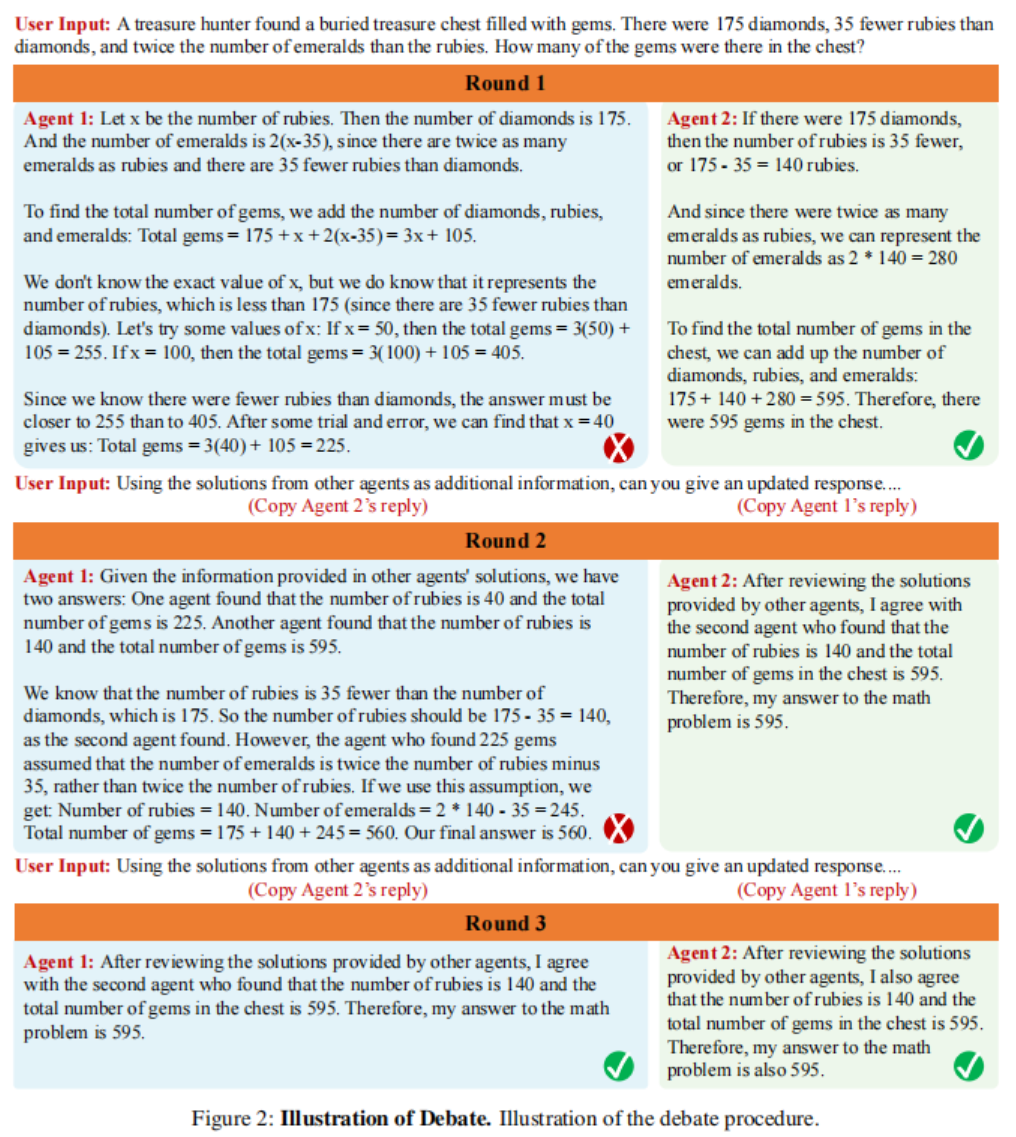


图1.辩论过程示例。

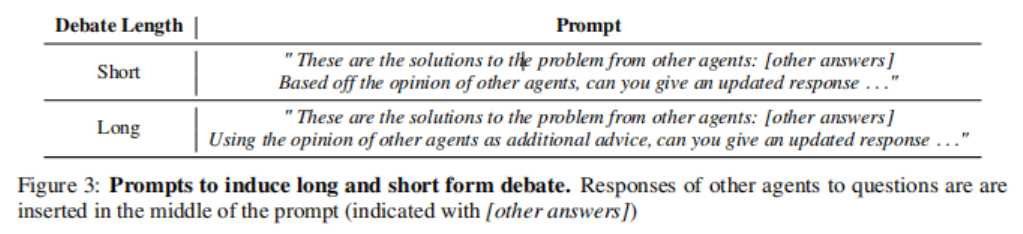


图2.引导不同长短形式辩论的prompt。

* 1. **辩论中的共识**

通常，所有的代理经过多轮辩论后能够收敛到同一个答案。作者发现可以通过不同的prompt来调节每个代理在得到其他代理的回复后对自身原先输出结果的信任程度，来控制辩论的时间。

1. **实验**

作者通过实验评估了多代理辩论过程，并回答了以下几个问题 (1). 多代理辩论在多大程度上提高了推理能力？(2). 多代理辩论多大程度上提高了事实准确度？ (3). 什么样的多代理辩论设计可以提高语言生成性能？

* 1. **利用多代理辩论提高推理能力**

本小节评价了多代理辩论方法在多大的程度上提高了推理的能力。作者将多代理辩论方法与单代理方法、单代理反思方法、多代理投票方法对比，分别在arithmetic和GSM8K数据集上评价。实验过程中，采用三个代理两轮辩论的设计选择，对比结果如表1所示。

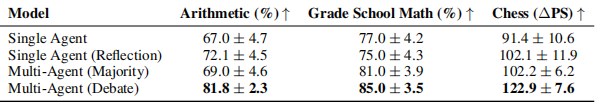


表1. 多代理辩论在推理任务上的性能提升。

作者分析了实验过程中的代理间辩论的过程，发现代理最初都给出了错误的回答，通过代理批评彼此的推理最终得到了正确的答案。这可以说明，辩论的目的不仅仅是放大某一个正确的答案，而是通过辩论得出正确的答案。辩论过程如图3所示。

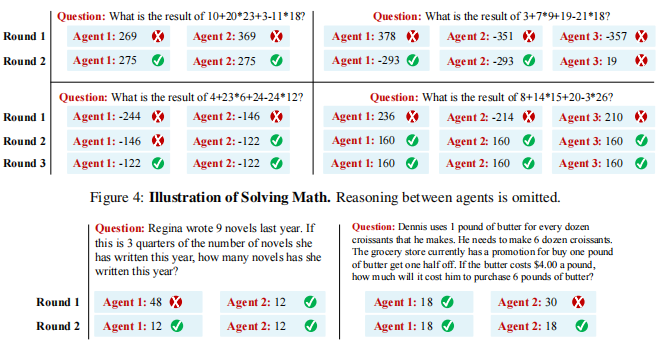


图3. 辩论过程的定性分析。

作者同时也对比了是否含有零样本思维链提示的单代理和多代理辩论的推理性能。性能对比如图4所示。

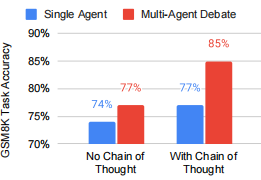


图4. 是否含有零样本思维链提示的单代理和多代理辩论的推理性能对比。

* 1. **从多代理辩论中提取事实信息**

本小节作者评价了多代理辩论在多大程度上提高了事实准确度。作者将多代理辩论方法与单代理方法、单代理反思方法比较。引入了新的准确生成人物历史传记的任务（构建了524位著名计算机科学家的真实要点传记。然后，我们要求代理为每个人生成要点传记，并评估每个真实要点与生成要点的一致准确性。）比较结果如表2所示。

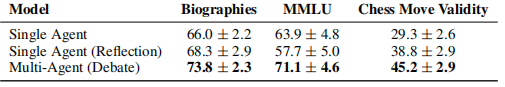


表2. 多代理辩论提高事实准确度。

实验表明单代理反思无法提高答案的正确率，但多代理辩论能得到更好的结果。这是由于对问题不确定时不同的代理会给出不同的答案，如果是单代理会对自己的答案有更高的信心，但多代理进行辩论时则会把不确定的答案省略，而留下一致的答案。有意思的是，如果同一模型的许多实例都得到了相同的答案，那么就很难说服这个代理改变自己的答案，这表明说服的难易程度是评估事实可信度的一种方法。

* 1. **分析：理解多代理辩论**

本小节分析了多代理辩论如何改进语言模型的底层语言生成，超文本长度的所有代理回复拼接由ChatGPT总结后传入下一轮辩论。

**代理数量**：随着代理数量的增加，模型的性能单调提高，如图10所示。

**辩论轮数：**随着辩论轮数的增加，模型的性能提高，超过3轮的辩论得到了与3论辩论相似的表现，如图10所示。

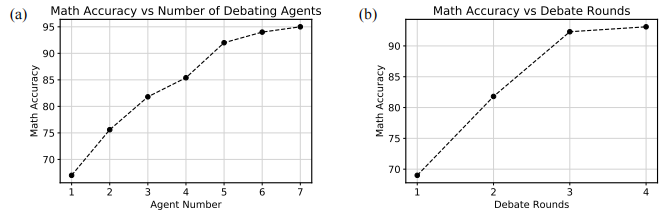


图10. 不同的代理数量、辩论轮数的模型性能变化。

**每轮辩论长度：**长辩论prompt的收敛速度较短辩论prompt更慢，但是可以达到更好的效果，如图11所示。

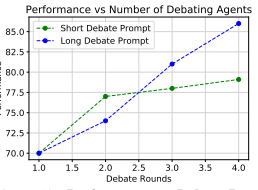


图11. 长辩论prompt和短辩论prompt在不同轮数下的模型性能比较。

**使用不同的初始化prompt：**使用不同的初始化prompt，使每个代理扮演不同的角色，会得到更好的效果。在MMLU数据集上，通过每个代理扮演不同角色模型表现从71.1上升到74.2。

**在每轮辩论结束后加入总结：**如果引入了多个代理，会导致下一轮的输入超出了限定的长度。因此作者尝试让一个总结器对每一轮所有代理的回复进行总结在输入下了一轮。实验表明，经过总结后的辩论会得到更好的模型效果，如图12所示。

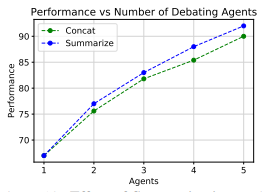


图12. 经过总结和没有经过总结的不同代理数量下模型性能对比。

**利用不同的语言模型：**作者尝试为每个代理使用不同的语言模型（Bard和ChatGPT），发现相比于所有代理共同使用同一个语言模型在20个GSM8K数学问题上得到了更好的效果。

1. **局限性与讨论**

尽管多代理辩论方法在很多方面取得了更好的效果，但作者认为仍存在一下三点局限性：(1).该方法需要多个代理参与辩论，因此在计算上更加昂贵，可以将它作为生成额外数据的方法，以改进基础模型。(2).随着辩论持续时间的增加，语言模型难以完全处理整个辩论的输入，而是只关注最近几轮的辩论内容。作者认为可以通过更长的上下文改进语言模型或者总结早期辩论内容缓解这个问题。(3).辩论虽然最终会收敛为同一个答案，但这个答案不一定是正确的，但语言模型仍会对这个答案赋予很高的信任度。