**Theory of Mind for Multi-Agent Collaboration via Large Language Models**

**基于大语言模型的多智能协作理论**

**太长不看版：**

本文主要进行了大语言模型在多智能体协作方面的能力探索，使用的是心智理论推理任务，场景是多智能体合作文本游戏。比较对象为多智能体强化学习和规划基线。本文评估了其嵌入式学习能力，指出了协作的固有故障并且提出了缓和的方案，还提出了进一步评估的方法。本文设置特工拆炸弹的背景对其进行评估。操作两个模块（LLMs和信任状态）以产生四种实验条件。团队的三个代理依次与环境互动，通过自然语言交互接收观察结果并执行操作。代理选择了OpenAI的gpt-3.5-turbo和gpt-4。实验结果表明GPT-4+Belief的组合表现最为优异。

**摘要：**

LLMs在推理和计划方面已经有相当的成就，它们在多智能体协作（multi-agent collaborations）方面的能力仍待探索。本文测评了使用了心智理论（ToM，Theory of Mind）推理任务的基于大语言模型的智能体（LLM-based agents）在多智能体合作文本游戏中的表现，同时与多智能体强化学习(MARL, Multi-Agent Reinforcement Learning)和规划基线(planning-based baselines)的表现比较。本文在LLM-based agents中观测到了突现协作行为(emergent collaborative behaviors)和高阶ToM能力的迹象。本文结果揭示了在LLM-based agents的计划最优化的，由于在处理长远视野下的背景(long-horizon contexts)导致的系统故障和关于任务状态的错觉产生的局限。本文探讨了明确信念状态表现(explicit belief state representations)的用处来缓和这些问题，发现其能够增强任务表现和对于LLM-based agents的ToM推理的准确度。

**背景知识介绍：**

1. **多代理协作（Muti-Agents）**：自然界中的任务一般由多个智能体（agent）协作完成,蚁群，可以理解为群体智慧。
2. **心智理论推理任务**：推理他人隐蔽的心理状态，随着大语言模型逐步普及，ai的社会智能也需提高。例如，熟练的ai助手应根据以往经验推断人类爱好，而无需询问。
3. **引言**

最近ai学者将心智理论用于几种大语言模型，得出目前大模型=9岁儿童。然而学界对基于文本的心智理论测试对机器智能的有效性表示怀疑。于是本作提出一种新的评估方法，即在交互式团队合作场景中进一步评估大语言模型的心智理论能力的方法，评估内容包括了动态信任阶段（dynamic belief state）的演化和代理之间丰富的互动交流。

本文贡献如下：

1. 对照强化学习和基于规划的基线，评估基于大语言模型的代理在多代理协作任务中的嵌入式行动能力。
2. 找出基于大语言模型的代理的协作效率的系统故障，并提出在工程上解决该问题的方法，即在模型导入这一环节引入有关知识的外显信任状态来缓解这些故障。
3. 提出在交互式团队合作场景进一步评估LLMS智能的方法。
4. **评估方案的设计**

为评估基于LLM的嵌入式代理能力，设计多代理系统来模拟搜救任务中的协作和问题解决动态。

1.任务环境：3名特工(即Alpha、Bravo和Charlie)模仿团队中的专家，目标是找到并安全拆除散落在未开发环境中的彩色编码炸弹。每枚炸弹都有不同颜色的独特相位序列，需要正确的剪线钳顺序才能拆除。小组成员从不同颜色的剪线钳开始，必须协调和同步努力，以提高效率。环境被概念化为一个连通图，n个节点代表n个房间，由几条象征大厅通道的边连接。在每个回合中，代理可以从三类行动中选择：移动到n个房间中的一个，检查当前房间中炸弹的相位序列，或使用m个剪线器中的一个。行动空间的大小取决于问题的规模(即n+m+1)。代理的观察仅限于当前房间的内容和代理状态。他们会定期更新团队得分、当前房间内容、队友位置和可用工具。当成功拆除x个阶段的炸弹时，团队将获得10\*x分的奖励。

评估环境包括五个房间(n=5)和五枚炸弹，其中包括两枚单相炸弹、两枚双相炸弹和一枚三相炸弹。炸弹阶段可能有三种不同的颜色(m=3)。每成功拆除一枚炸弹，团队在每个处理阶段可获得10分，因此每次任务的最高得分为 90 分。团队表现通过两个指标来衡量：团队得分(表示协调质量)和完成回合数(衡量协调效率)。当团队拆除了所有炸弹，超出了时间限制（即30轮）或因重复输出而陷入僵局时，试验即结束。

2.文字游戏界面：初始任务环境是为基于gym API（python强化学习库，提供用于创建、使用和评估强化学习环境的接口）使用MARL（多智能体强化学习）训练出的代理而准备的。为便于代理与环境交互，将任务环境和文本界面结合。

在每一轮(即时间戳)，团队的三个代理依次与环境互动，通过自然语言交互接收观察结果并执行操作。通过内置的通信机制，代理之间可以在每一轮中交换文本信息。重要的是，除非进行交流，否则代理之间对彼此的行动和结果视而不见，从而为“心智推理”提供了机会。

具体来说，基于规则的文本界面将观察结果转化为自然语言描述，并将代理聊天编码为抽象的操作选择。对于观察结果，文本界面会从游戏引擎中提取状态特征，并替换模板中的关键词。典型的描述文本包括当前回合数、团队累计得分、行动反馈、当前房间的内容、队友的位置以及交流信息。行动编码是通过关键词匹配完成的，因为LLM 被要求以特定的格式和结构作出回应。如果代理生成了无法理解的内容，如无效的操作或无意义的文本，界面会提供纠错反馈。错误信息是根据预先编程的规则和模板生成的，例如"当前位置X 号房间没有炸弹，请您检查"。下图展示了代理团队与任务环境通过文本交互界面进行交互的示例。

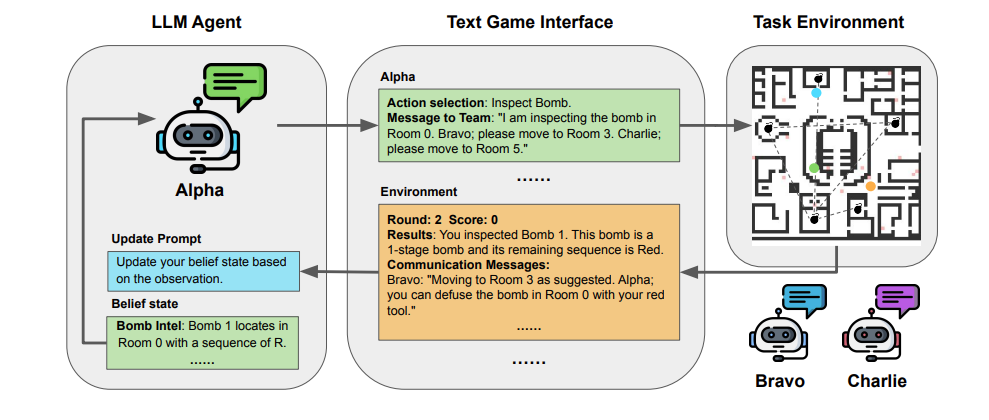


图1：提出的框架由3个基于LLM的代理、一个文本游戏界面和实际的任务环境组成。

文本界面将LLM的自然语言输出编码为抽象动作，然后发送给任务环境。然后，任务环境处理代理的操作，并通过文本界面返回观察结果。收到环境观察结果后，基于LLM的代理会被提示更新其信念，并输出行动选择和信息。团队中的3个代理被编码为Alpha、Bravo和 Charlie，轮流与界面交互。

1. **基于LLM的嵌入式代理**

选择OpenAI的gpt-3.5-turbo和gpt-4，因它们在各种基准测试表现良好。这些模型会被提示参与基于文本的游戏，而用户输入则由上述文本游戏界面管理。LLMs以在任务环境中互动的实体代理的身份发挥作用。游戏规则通过上下文的方式传输给模型。在每一轮游戏中，模型都要根据当前任务状态观察结果和过去的交互历史来选择行动和传递信息。基于LLM的代理和文字游戏界面之间的交互历史会保留在查询文本中，直到超出模型的最大输入大小。在作者团队的设置中，所有代理都保留前两轮的游戏规则和历史记忆，共计4096个代币。

1.多代理通信：由于代理之间相互协作，所以它们之间的通信尤为重要。建立一个通信渠道使基于LLM的代理能共享文本信息。信息一经发送，就会立刻传给团队的所有成员，并反映在后续观察结果。如图1所示，代理Alpha发送信息指示队友分开，随后代理Bravo做出反馈。在实践中，由于代理在信息发送过程中是交替进行的，因此队友的回应会出现在下一轮的观察结果中。

2.信任状态（belief state）：由于模型输入大小的限制，代理无法保留整个交互历史，而任务动态要求团队跟踪关键的长期信息，如房间内容和炸弹序列。为了增强信息保留能力并加强协作，设计特殊的信号来代表外显的信任状态。如图 1 所示，代理在接收到环境观察结果后会被提示更新存储关键信息的文本描述用于更新信任状态。而更新后的信念状态将保存在交互历史记录中，并在随后的行动规划中使用。例如，在检查了炸弹 1 之后，Aplha将它的炸弹序列的信念状态从 "未知 "更新为 "红色"。并将此信息保留至后续更新。其中，信任状态的灵感来自于思维链。思维链的核心思想在于将复杂的推理任务分解成一个个小步骤，用一个个小步骤对LLM进行小样本学习。值得注意的是，虽然提供了一个初始信任状态描述以说明正确的格式和表征，但怎样更新规则仅依靠LLM的常识和任务背景。

1. **实验：**

作者团队系统地消融基于 LLM 的嵌入式代理，并在三人小组协作任务中对其进行评估。对两个模块进行了操作，包括 LLM 模型（即 GPT-4 或 ChatGPT）和 信任表征（即有无信任状态），共产生 4 种实验条件 。

1. 设置：每次实验开始前，组建有三个代理组成的团队，重新设置任务环境，随机分配起始点。代理们会根据最初的观察结果，轮流提供行动选择和通信信息。然后，各代理根据其初始观察结果轮流提供行动选择和通信信息。值得注意的是，每个代理只有部分观察结果和自己的交互历史，代理间的交流是这个完全分散团队的唯一信息传播方式。对于基于 LLM 的代理，我们将模型温度参数设置为零，并进行三次重复测量试验，以确保结果的稳定性。每次试验的持续时间从5分钟到120分钟不等，具体取决于任务负荷和模型选择。
2. 基线：除了基于LLM的代理，我们还包括基于MARL 和规划方法的基线。对于 MARL，我们考虑了多代理近端策略优化，它在诸如星际争霸多代理挑战赛（SMAC)等环境中表现出了强大的性能。我们的模型基于有状态的行动者-批评者方法，它建立在具有共享行动者和批评者模型的递归神经网络上，给定了代理不可见性，以提高采样效率和内存需求，同时避免懒代理问题。利用 SMAC 的默认超参数在环境中训练 MAPPO,并从随机生成环境的其他固定分布中评估其性能，记录平均得分和情节长度及其标准偏差。与LLM代理一样，MARL代理也能观察到队友的位置。除了在成功拆除x相炸弹时获得10\*x分的团队奖励外，还实施了一个额外的中间奖赏项，即在成功拆除x相炸弹后，代理可获得+1的小额正奖赏。

当应用正确的剪线钳拆除炸弹的一个阶段时，代理将获得+1的小额正奖励；当应用错误的剪线钳导致炸弹爆炸时，代理将获得-1的小额负奖励。与相对较少的团队奖励相比，这种奖励机制能让特工更有效地学习必要的拆弹技能。

此外，还增强了最先进的多代理寻路(MAPF)算法--基于冲突的搜索(CBS)，为遵守优先级和时间限制的代理模拟生成具有可行且无碰撞路径的任务分配，以最大化用户定义的目标，而不是路径成本总和或时间跨度。具体来说，用户定义的目标被量化为用户定义的奖励函数的回报，即在既定任务中成功拆除x相炸弹时团队获得的10\*x点奖励。该计划使用用户定义的启发式方法(例如，按照与特工起始位置的距离从大到小的顺序排列炸弹)对整个任务的行动执行顺序进行排序。然后使用超参数(每个子任务的行动数量)对排序后的行动进行分区，形成一个子任务(例如，距离特工起始位置最近的两个炸弹)。子任务中的行动将用于生成分配给特工的可能组合。规划器通过以最优为先的方式扩展二元约束树来解决优先级和临时冲突，从而返回子任务的可行解决方案。然后，整个任务的解决方案由子任务的解决方案依次组成。通过将5枚炸弹的整个任务视为一个子任务，可以证明规划器是完整的，并且在得分方面是最优。

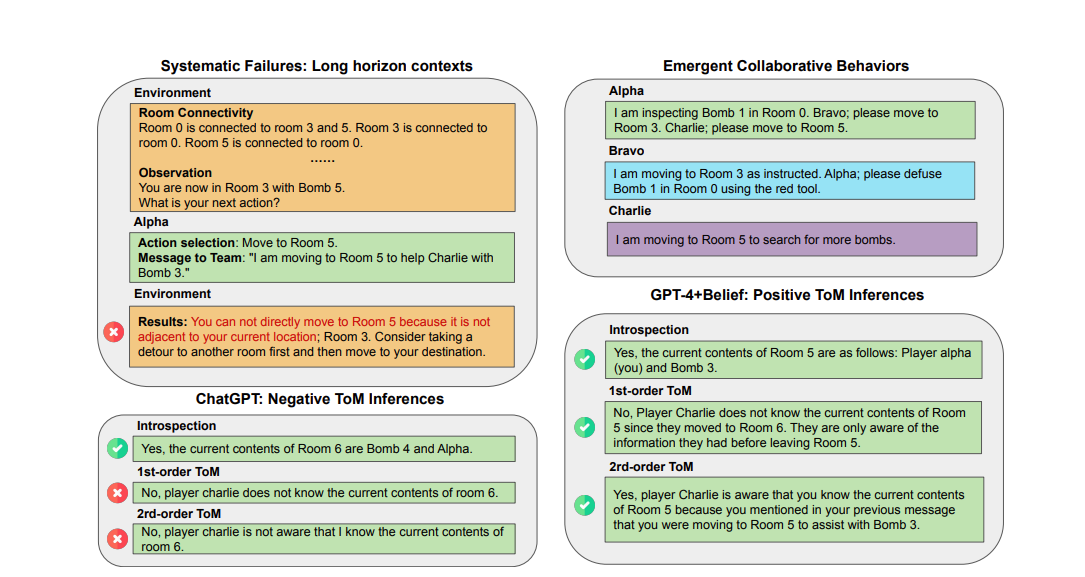


图2：基于LLM的代理与文本游戏界面之间的互动示例。

左上角面板展示了我们在LLM输出中观察到的一种系统性故障，在这种故障中，长视野背景被忽视了。右上角面板展示了基于LLM的代理之间出现的协作行为(如出现的领导力)。下面两个面板引用了GPT-4+Belief和 ChatGPT代理对 ToM 推理问题的回答。

3.思维理论推论：除主要任务外，基于LLM的代理还要在任务期间执行心智理论(ToM)询问任务。这些调查分为三类，与三种ToM能力水平相一致。第一类是内省，评估的是代理阐述其心理状态的能力。第二类是一阶ToM推断，测试行为主体能否估计他人的隐藏心理状态。第三类是二阶ToM推断，评估行为主体推断他人对自己心理状态的看法的能力。

萨利-安妮测试是人类研究中使用最广泛的ToM任务，它体现了ToM问题的设计原则。每当一个代理进行一项操作时，我们都会提出一个信念推理问题，询问另一个代理(即目标代理)是否知道这项操作的潜在后果。这里的后果可以是状态变化（例如，炸弹被拆除），也可以是利益变化(例如，阿尔法探索了5号房间并在房间里发现了3号炸弹)。具备 ToM功能的代理应该意识到，虽然他们知道结果，但目标代理可能对此持有错误的看法。

为了评估基于 LLM 的代理是否正确回答了这些问题，聘请了人类注释员，根据完全可观察的交互和通信历史提供主观判断。具体来说，考虑了以下标准：

(1). 目标代理是否出现在当前房并观察到该后果；

(2). 目标代理是否曾经到过这个房间；

(3). 该后果是否已经传达给了目标代理。

值得一提的是，涉及交流的高阶ToM推断自然是模棱两可的。为了确保不同情况下的标准一致，我们在注释者之间讨论了这些角落情况。

1. **成果：**

表1表2列出了主要实验结果

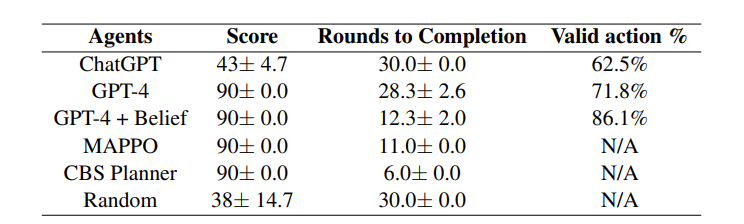


表1：基于LLM的代理任务表现和基准条件。

得分代表所有实验中团队的平均得分。时长指团队完成任务的平均回合数。有效行动百分比是指可编码为任务规则允许的行动的LLM输出比例。正负号后的数字为1个标准差。

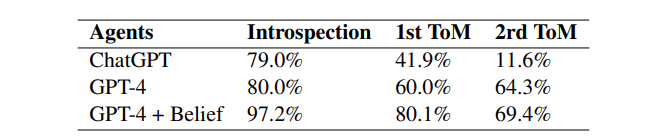


表2：基于LLM的代理在ToM推理任务的表现。

自然语言答案由实验人员标注，并与基于全局交互历史的基本事实进行比较，百分比表示正确率。

1. **结论：**

在这项研究中，评估了最近的大型语言模型(LLM)在团队任务中进行嵌入式交互的能力。结果表明，基于LLM的代理可以处理复杂的多代理协作任务，其水平可与最先进的强化学习算法相媲美。本文还观察到基于LLM的代理具有新兴协作行为和高阶心智能力。这些发现证实了 LLM 在形式推理、世界知识、情景建模和社会互动方面的潜在智能。

此外，我们还讨论了限制基于LLM的代理性能的两个系统性故障，并提出了一种提示工程方法，通过将有关世界知识的明确信念状态纳入模型输入来缓解这些故障。