# 自生成兵棋 AI：基于大语言模型的双层 Agent 任务规划

孙宇祥[1,2†]，赵俊杰[1]，解宇轩[1]，喻车澄[1]，周献中[1,2†]

(1. 南京大学 工程管理学院，南京市 210093；2. 南京大学 智能装备新技术研究中心，南京市 210093)

**摘　要:** ChatGPT 所代表的大语言模型对 AI 领域产生了颠覆性影响。但它主要关注自然语言处理、语音识别、机器学习和自然语言理解。这篇论文创新性地将大语言模型应用于智能决策领域，将大语言模型置于决策中心，并构建以大语言模型为核心的 Agent 体系结构。基于此，进一步提出了双层 Agent 任务规划，通过自然语言的交互，发出和执行决策指令，并通过兵棋推演模拟环境进行仿真验证。通过兵棋对抗模拟实验，发现大语言模型的智能决策能力明显优于常用的强化学习 AI，并且其智能性、可理解性都更强。通过实验证明，大语言模型的智能性与 Prompt 密切相关。这项工作还将大语言模型从以往的人机交互领域拓展到智能决策领域，对智能决策的发展具有重要的参考价值和意义。

**关键词:** 自生成兵棋 AI; 大型语言模型;ChatGPT; 智能决策; 兵棋推演; 强化学习

**中图分类号:**　　**文献标志码:** A

**DOI:** 10.13195/j.kzyjc.2023.1497

**引用格式：**孙宇祥，赵俊杰，解宇轩，喻车澄等. 自生成兵棋 AI：基于大型语言模型的双层 Agent 任务规划 [J]. 控制与决策.

# Self generated Wargame AI: Double layer Agent Task Planning Based on Large Language Model

*Yuxiang Sun*[1,2†]，*Junjie Zhao*[1]，*Yuxuan Xie*[1]，*Checheng Yu*[1]，*Xianzhong Zhou*[1,2†]

(1. School of Management and Engineering,Nanjing University, Nanjing 210093,China； 2. School of Intelligence Science and Technology, Nanjing University, Nanjing 210093, China)

**Abstract:** The Large Language Model, exemplified by ChatGPT, has brought a disruptive impact to the field of artificial intelligence, with a primary focus on natural language processing, speech recognition, machine learning, and natural language understanding. This paper innovatively applies the Large Language Model to the field of intelligent decision-making, places the Large Language Model in the decision-making center, and constructs an agent architecture with the Large Language Model as the core. Building on this, it further introduces a two-tier agent task planning strategy, issuing and executing decision commands through natural language interactions, and conducting simulation validations within a wargame simulation environment. Through game confrontation simulation experiments, we found that the intelligent decision-making capability of Large Language Models is significantly superior to that of commonly used reinforcement learning AI. This superiority is apparent in terms of intelligence, comprehensibility, and generalizability. And through experiments, it was found that the intelligence of the Large Language Model is closely related to Prompt Engineering. This work also expands the application of Large Language Models from previous human-computer interactions to the realm of intelligent decision-making, providing valuable insights and significance for the advancement of intelligent decision-making.

**Keywords:** Self generated wargame AI; Large language model; ChatGPT; Intelligent decision-making; wargame; Reinforcement learning

## 0 引　言

自 ChatGPT 于 2022 年 11 月 30 日正式推出以来，它迅速成为最受欢迎的智能 Chatbot 之一[1-2]。自问世以来，ChatGPT 已经在多个领域得到应用，如代码纠正[3]、公共卫生和全球变暖[4-5]。2023 年 7 月，OpenAI 发布了 Code Interpreter 插件，进一步增强了

**收稿日期:** 2023-10-26；**录用日期:** 2024-03-11.

**基金项目:** 国家自然科学基金青年项目“基于人机融合的深度强化学习智能博弈决策机理研究”（62306135）；教育部人文社会科学研究规划基金, 青年基金“面向智能博弈的行为决策一致性机理研究”（23YJC630156）；江苏省自然科学青年基金“面向兵棋智能博弈的智能决策与人机融合范式研究”（BK20230783）

†通讯作者. E-mail: sunyuxiang@nju.edu.cn, zhouxz@nju.edu.cn

ChatGPT 的数据解析能力，并解决了大语言模型在数学和语言领域的固有缺点。这些发展为改进智能决策推演领域的 AI 智能性和泛化性提供了新的启发，即利用 ChatGPT 自生成的 AI 在兵棋推演中做出智能决策。

尽管基于知识驱动的 AI 的发展和应用是智能兵棋领域的起点，但近年来数据驱动 AI[6]、知识和数据混合驱动 AI[7]，逐渐成为研究的热点，其中强化学习 AI 取得了一系列突破。在知识和数据混合驱动 AI 方面，刘满等人[7] 设计了一个平衡规则和数据的兵棋决策框架。在强化学习和深度学习 AI 领域中，针对强化学习，张健等人[8] 提出一种基于 QMC 的蒙特卡洛聚类扩展算法，以解决多智能体决策问题中的高计算复杂度和内存占用大的挑战。南京理工大学的李琛团队[9] 设计了一种基于 Actor-Critic 框架的多 Agent 决策方法，并取得了出色的智能表现。Chen 等人[10] 将人类经验引入指导智能体，提出了基于主动强化学习的目标分类人机交互智能体，以提高分类准确性。后续该团队[11]] 提出了一种深度学习架构，用于处理不完整信息的战争游戏中的意图识别，提高识别稳定性，和准确度。徐佳乐、张海东等人[12] 设计了一个基于卷积神经网络的策略学习模型，以提高战局预测的准确性。Sun 等人[13] 创新性地将强化学习、深度学习和自然语言处理技术应用于游戏 AI，实现了高准确度的语义文本输出。Xu 团队[14] 构建了一个用于多智能体不完整信息游戏的人机博弈平台，庙算战争游戏平台，用于智能体的训练和评估。腾讯 AI 实验室[15-16] 采用深度强化学习在《王者荣耀》游戏中实现了游戏对抗，并战胜了职业选手。Dong 等人[17] 提出了一种改进方法，通过利用专家演示以提高深度强化学习的学习效率。总之，通过深度学习、强化学习和智能兵棋的深度结合，Agent 的智能水平不断提高[18-21]。

尽管基于规则的 AI 无需长时间的训练，但由于受规则限制，其智能水平的上限难以突破；而数据驱动型 AI 和强化学习型 AI 通过处理大量数据并采用强化学习算法来提高智能性和灵活性，但它们的可解释性较差，难以在场景和关键点变化下实现模型迁移[22-25]。因此，在智能兵棋领域提升 AI 的智能性和可理解能力成为下一步研究的重点。

此外，对抗性游戏的决策是复杂且连续的。为了使决策更具备更强的智能性和泛化性，本文着重介绍了一种基于大语言模型的自生成 AI 兵棋架构。构建了一个涉及战略 Agent 和战术 Agent 互相交互的决策机制，该机制也可以拓展到涉及多个生成 Agent 交互的决策机制，通过该机制生成具备可信、可解释的兵棋对抗智能决策。

本文的核心工作和创新点如下：

1) 自生成的 AI 兵棋架构是以大语言模型为中心的智能 Agent 体系结构。该架构由多个生成 Agent 组成，每个 Agent 都拥有自己的大语言模型（本文使用 ChatGPT 作为驱动工具）。这些智能 Agent 可以通过反射流和记忆流进行通信和合作，共同制定决策。通过相互对话，它们可以共享信息、分析情况，并基于对话内容进行推断和决策。

2) 建立一个双层 Agent 任务规划模型，分别面向战略 Agent 和战术 Agent，以规划兵棋对抗过程中的任务。战略 Agent 描述了所有当前 Agent 观察到的具体情况，规划涉及基于所有观察到的情境信息，进行任务分配和执行。战术 Agent 仅关注单个 Agent 算子所观察到的情况，并根据战略规划 Agent 执行相关任务。同时，战术 Agent 也可以根据战略 Agent 发布的提示自行判断并提供反馈。

3) 以兵棋作为实验平台，通过三种推演想定进行实验验证，实验结果表明，大语言模型的智能决策能力、稳定性和泛化性明显强于强化学习 AI，而且其智能性、可理解性都更出色，还发现专门为大语言模型提供领域专家的先验知识可以显著提高它们的智能化水平。

## 1 兵棋推演环境与想定介绍

兵棋推演系统是一个典型的智能博弈应用仿真平台，通常由四个基本组件组成：推演、地图、规则和想定。本文的模拟实验环境是一个用于战术级别的智能兵棋推演平台，名为“先胜一号”[28]。该平台提供了用于红蓝双方人机/机机对抗兵棋的功能设计，为基于规则和强化学习的智能 Agent 的开发提供了支持环境。本文的兵棋环境支持三种想定，并且遵循相同的规则和标准，每一种想定任务由打击、夺控、存活三种子任务构成，三个子任务的权重不同，三种想定分别以一种任务为关键任务。分别是以打击为关键任务的想定、以夺控为关键任务的想定和以存活为关键任务的想定。该平台包括地图编辑、推演人员管理、规则编辑、场景编辑、推演设置、数据分析和系统功能模块，可以实现智能指挥模型算法的作战行动序列的自生成以及有效性评估，以下是总体的实验设计介绍。

总体的实验设计：

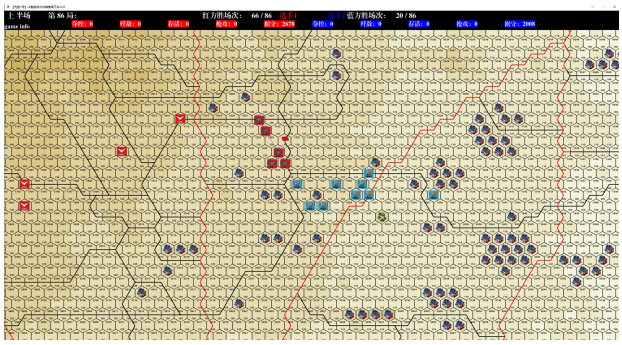

**图 1　兵棋实验场景展示**

1) 实验场景：在“先胜 1 号”推演平台上进行模拟和推演，如图 1 所示。主战区呈规则六边形网格，代表城市居住区。地图上标有坐标，例如 1224（”12” 代表横轴，”24” 代表纵轴），主要夺控点标有红旗。每格下方标有高度信息，颜色深浅表示不同高度，相邻颜色块高度差为 10 米。城市住宅区带有房子图标，可提供推演人员隐蔽和防御优势。黑色和红色线分别表示一级和二级公路，两者的行驶速度不同。

2) 任务想定：博弈双方为红方和蓝方，各有 10 个算子。胜利条件为首先占领夺控点或摧毁对方所有算子。推演人员通过控制算子进行移动，每次移动耗费一定的燃料。算子可以在城镇居民地内移动、互相射击或隐蔽，瞄准和射击时显示十字标志。直射火力受距离、能见度和其他因素的影响，这些因素通过随机数反映。

3) 动作空间：行动空间包括机动、隐蔽和射击。并且算子可执行向南、北、东南、西南、东北、西北六个方向的战术级移动。

4) 状态空间：地图上的坐标、高度信息以及算子的状态（完好、受损、战损）构成了环境状态空间。

从本文作者所调研的已发表文献来看，本文将大语言模型引入智能兵棋，是首次在兵棋环境进行了大模型智能决策的实验验证，开拓了大语言模型在智能决策中的应用领域。

## 2　生成式兵棋 AI 架构

整个核心流程如图 2 所示，其核心是将兵棋推演中的情景图像信息转化为语义信息，包括描述信息和情景信息，然后以 Prompt 的形式将这些信息发送给兵棋 Agent，然后 Agent 反馈相应的目标导向的规划语义。规划语义进而被转化为动作序列（例如 1、2、3、4、...10，其中数字代表具体行动。具体来说，这些数字被转化为相应的行动，比如进攻、防御、回避、加速、射击、向左移动等），这些行动会影响兵棋环境并生成新的态势。最终，这些行动产

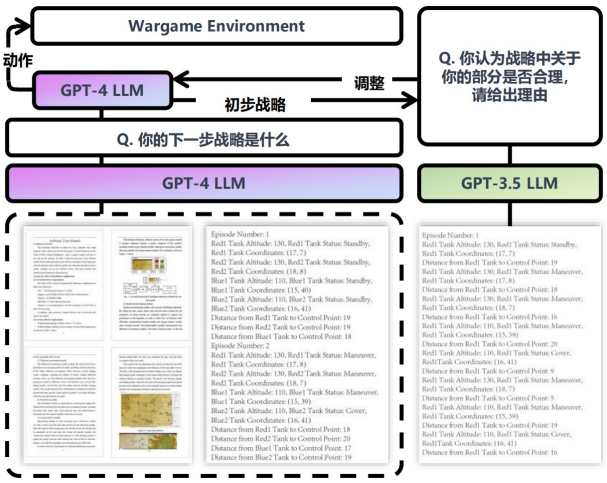


**图 2　基于大语言模型的双层 Agent 任务规划决策框架**

生的新态势被回收成为起始情景图像并转化为语义。

在此基础上，为了减少计算能力和内存需求，并提高操作效率，本文中战略 Agent 使用 GPT-4 LLM，而战术 Agent 使用 GPT-3.5 LLM。与完全使用 GPT-4 或 GPT-3.5 LLM 相比，这样可以全面提高智能决策的智能性，而不需要过多的计算能力和内存空间。首先，将专家的先前知识文档输入到战略 Agent 中，通过 GPT-4 LLM 进行学习，然后提供适当的提示输入，使战略 Agent 通过 GPT-4 LLM 做出决策并将其转化为影响兵棋环境的行动输出。然后，战略 Agent 将相应的指令发送给每个战术 Agent 以供执行。战术 Agent 通过结合适当的提示，使用 GPT-3.5 LLM 提供反馈，判断当前 Agent 是否适合执行任务，并将推荐的执行结果提供给战略 Agent 进行调整。战略 Agent 和战术 Agent 之间的关系如图 3 所示。

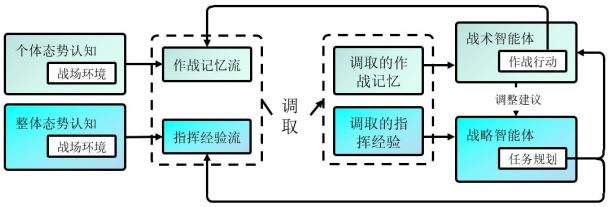


**图 3　战略 Agent 和战术 Agent 之间的关系**

战略 Agent：基于任务规划流程，战略 Agent 综合各个个体 Agent 的 state 并提供了任务规划序列，即每个兵棋在步骤分配中应采取的行动；战术 Agent：战术 Agent 接收任务规划并根据自身 state 提供修改建议和分配任务的原因；战略 Agent 根据修改建议再次进行规划，直到所有战术 Agent 不再提供修改建议。

在兵棋环境中，本文已经实现了十个红方棋子和十个蓝方棋子之间的对抗。红方和蓝方棋子在不同的集群中具有不同的由 ChatGPT 生成的语义交互

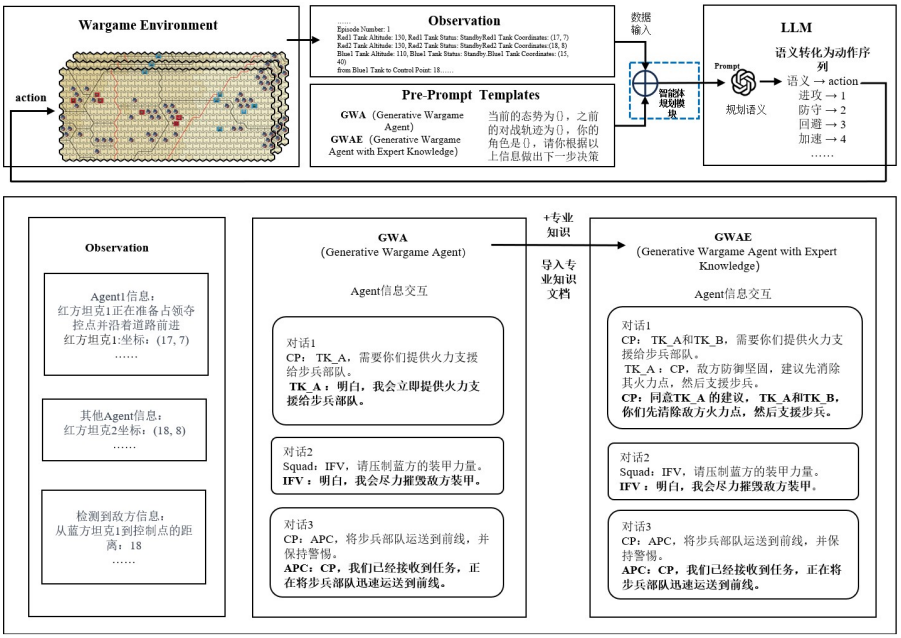


**图 4 自生成的 AI 兵棋架构**

信息。为了实现上述提到的决策机制，我们已经开发了一个包括三个主要组件的 Agent 架构：一个用于存储和分配缓冲区以及生成 batches 的记忆流；一个将 batches 作为提示用于大语言模型理解其在决策过程中的角色反思流；以及一个任务规划流，用于从 batches 中综合更高级的推理，使 Agent 能够整合想定信息并制定更好的战前计划。这个 Agent 架构的设计旨在存储、综合和应用过去的战场经验，以使大语言模型生成可信赖的决策。

## 3 生成式兵棋 AI 模型

### 3.1 兵棋 Agent 交互与通信

在上述所描述的架构中，兵棋 Agent 获取情境信息并通过自然语言互相交互，以维持协作。每个 Agent 用自然语言描述他们的动作，比如“红方 Agent1 正在准备占领夺控点并沿着道路前进”，“蓝方 Agent2 准备瞄准敌方目标 1”。然后，这些语句被翻译成具体的行动，直接影响兵棋环境。同时，所有的动作和移动都将显示为一系列数字符号，出现在每个头像的上方，以提供对行动的抽象表示。为了实现这一点，该架构利用语言模型将语言翻译成行动，同时在每个兵棋上方表示一个简洁的符号，代表 ChatGPT 为该 Agent 提供的行动建议。例如，“红方 Agent1 正在准备占领夺控点”会显示为出现在兵棋上方的“!”，而“红方 Agent1 正在准备瞄准敌方”则显示为“→”。

在这个环境中，Agent 相互使用人类完全能理解的自然语言进行交流。他们通过句子的语义获取其他推演人员和环境的情境信息。如图 8 所示，每个对话框是战略 Agent 与战术 Agent 进行交流的示例，图中 CP、TK、Squad、IFV 和 APC 分别代表指挥所、坦克、战术小队、步兵战车和装甲人员运输车。

### 3.2 构建模型

生成式兵棋 AI 旨在为兵棋环境中的智能决策提供一种新颖的决策框架。与传统的基于规则的 AI、数据驱动的 AI 或强化学习 AI 相比，本文所提的架构如图 4 所示，利用 ChatGPT 进行智能决策和与兵棋环境的交互，以当前环境和过去的经验作为输入，并以生成的行动的形式产生输出。具体通过 GPT 插件将兵棋环境的情景图像信息转化为语义信息 observation，GWA（Generative Wargame Agent）算法通过输入当前的态势、之前的对战轨迹和当前角色等信息，通过图 5 所示的包含记忆流，反思流和任务规划流的智能体规划模块，以 Prompt 的形式将这些信息发送给以 GPT 为决策中心的兵棋 Agent，然后 Agent 反馈相应的目标导向的规划语义。规划语义然之后被转化为动作序列。最后这个动作输出反馈给兵棋环境中的算子，进而帮助决策。GWAE（Generative Wargame Agent with Expert Knowledge）算法在 GWA 的基础上以文件的形式输入了兵棋推演的专家经验。在该架构的基础上，我们构建了一个包括战略 Agent 和战术 Agent 的双层 Agent 系统。战略 Agent 将自己位置信息和观察到的对手状态的所有信息作为输入，然后将其与整体环境和输入结合起来作为提示，生成宏观层面的战术智能任务规划流程。战略 Agent 以提示的形式分配任务给战术

Agent，战术 Agent 基于自身状态提供修改建议和修改原因。然后，战略 Agent 根据这些建议进行重新规划，直到所有战术 Agent 不再提供进一步的建议。

当使用像 GPT-4 LLM 这样的最先进大语言模型时，战略和战术 Agent 仍然面临许多挑战。由于这两个 Agent 生成了大量事件和记忆，在该架构中最关键的挑战在于在从记忆流中检索和综合相关数据时生成最相关的记忆片段。因此，本文尝试降低计算能力和内存需求。由于 GPT-3.5 LLM 输入的长度限制是 4k-tokens，大约两千个汉字，GPT-4.0 LLM 输入的长度限制是 32k-tokens，便于整体战略输入和专家知识文档输入，而 GPT-3.5 LLM 无法处理宏观战略，所以战略 Agent 只能用 GPT-4.0 LLM。由于战术 Agent 需要相互交互并提供快速提供反馈，且只需要关注一个算子的行为。本文通过将测试样本中的兵棋算子信息分别输入到战术 Agent 进行 GPT-3.5 LLM 和 GPT-4 LLM 的性能对比，如表 1 所示，GPT-3.5 LLM 降低了计算能力和内存需求，而不显著影响智能水平，损失很微小的胜率和智能性换取更快的反应时间和更低的成本对战术 Agent 来说是可以接受的，因而战术 Agent 采用 GPT-3.5 LLM。

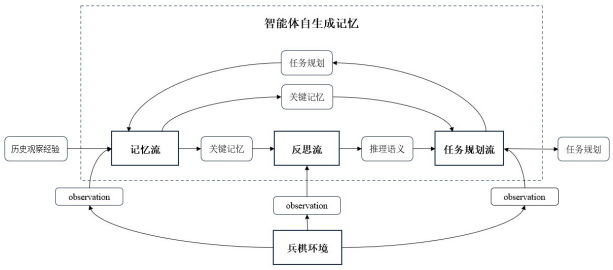


**图 5　智能体规划模块**

**表 1　战术 Agent 使用 GPT-3.5 LLM 和 GPT-4 LLM 性能对比**

| | 计算时间 | 成本 | 平均胜率 |
|---|---|---|---|
| GPT-3.5 LLM | 34ms per token | $0.002 per 1K tokens | 76% |
| GPT-4 LLM | 196ms per token | $0.06 per 1K tokens | 79.5% |

### 3.2.1　记忆流

作为该架构的核心组件，记忆流直接影响了决策的效率和准确性。整个记忆流是一个记忆对象的列表，每个对象由自然语言描述、创建时间戳和最近访问时间戳组成。记忆列表中的基本元素是观察，包括 Agent 观察到的所有情境信息。由于存在战争迷雾，战场环境不允许完全了解和意识。在特定状态下，Agent 观察到的共同信息受到某些限制，包括个体行动、红方 Agent 的行动以及在我方可见范围内对手蓝方 Agent 的行动。

**例 1**　Observation1（态势 1）：红方 $Agent_1$ 观察到红方 $Agent_2$ 靠近夺控点并试图夺控。

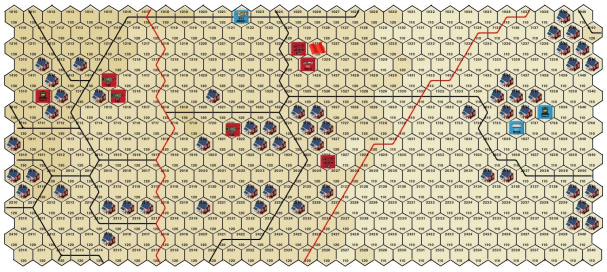

**图 6　态势 1：Agent 观察到自己方的 Agent 靠近夺控点并试图夺控**

**例 2**　Observation2（态势 2）：红方 $Agent_1$ 观察到蓝方 $Agent_2$ 靠近城市居住区并试图射击。

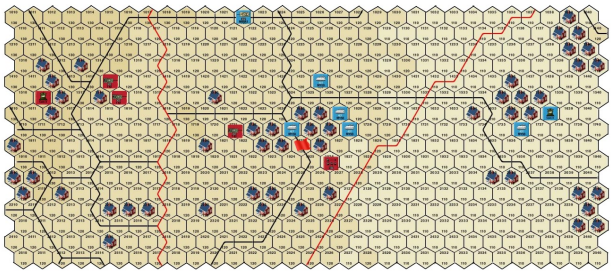

**图 7　态势 2：Agent 观察到对方 Agent 靠近城市居住区并试图射击**

本文在整个记忆流架构中构建了一个检索功能，并利用它从 Agent 的历史经验中提取 Observation，为生成合理提示和语言模型产生理性决策提供基础。检索功能可以是有选择性的，优先提取最近的 Observation、之前设置的重要节点以及相关记忆，以产生有效的结果。

”Recency”（近因性）：评估记忆的新近程度，越新近的记忆分数越高。为最近添加的 Observation 分配更高的分数。这种情况下，Agent 优先考虑最近几个步骤生成的记忆信息。为了考虑时间因素的影响，我们采用一个时间衰减系数来计算分数。

”Importance”（关键性）：评估记忆内容与当前情境的相关度，相关性越高的记忆分数越高。评估记忆内容的重要性，核心记忆的分数高于常规记忆。将记忆流中的数据分为常规记忆和核心记忆。并为 Agent 生成的核心记忆分配更高的分数。例如，一个红方 Agent 向左移动并接近道路可以被归类为常规记忆，而一个红方 Agent 接近夺控点并消灭蓝方 Agent 可以被归类为核心记忆。在这个架构中，我们要求语言模型直接输出在 1 到 10 的范围内的关键性整数分数，其中 1 表示纯粹的常规记忆，比如在道路上移动，而 10 表示最重要的核心记忆，比如占领夺控点或成功射击。具体的实现过程可以描述如下：从记

忆流中检索相应的记忆以形成提示，允许 Agent 相应地生成关键性分数并将其存储回记忆流中。

”Relevance”（相关性）：为与当前情况相关的对象分配更高的分数。由于不同记忆对象之间存在相关性。例如，红方 Agent 以较高速度到达道路并接近夺控点，这个记忆与红方 Agent 夺取夺控点有很强的相关性。在本文中，要求 ChatGPT 生成相关性分数，用一个从 1 到 10 的刻度来描述记忆对象之间的相关性程度。

Agent 的最终分数是基于三个指标——近因性、关键性和相关性来计算的。每个指标都有一个相应的系数，这个系数是由推演者人为指定的，反映了不同推演者的决策风格偏好。每个指标都有一个分数，这些分数在 1 到 10 的范围内。为了计算 Agent 的最终分数，我们首先需要对每个指标的分数进行归一化处理，以便将它们转换到统一的 0,1 区间内。归一化的计算方法可以用以下公式表示：

$$X'_{ij} = \frac{X_{ij} - m_j}{M_j - m_j} \tag{1}$$

其中，$x_{ij}$ 代表原始指标分数，$M_j$ 为 $x_{ij}$ 最大值，$m_j$ 为 $x_{ij}$ 最小值，其中 i 保持不变。归一化处理后，我们将近因性分数、关键性分数和相关性分数与它们各自的系数相乘，然后加总这些乘积来得到最终分数。本文用 $\alpha_{recency}$ 代表近因性系数，$\alpha_{importance}$ 代表关键性系数，$\alpha_{relevance}$ 代表相关性系数，$score_{recency}$ 代表归一化后的近因性分数，$score_{i}mportance$ 代表归一化后的关键性分数，$score_{relevance}$ 代表归一化后的相关性分数，那么最终分数 $score_{final}$ 的计算公式如下：

$$score_{final} = \alpha_{recency} \times score_{recency} + \alpha_{importance} \times score_{importance} + \alpha_{relevance} \times score_{relevance} \tag{2}$$

我们通过近因性、关键性和相关性三个指定变量来计算出总分数。同时，近因性系数、关键性系数及相关性系数是人为指定的。可以让推演者根据自己的风格偏好，这样好处是可以让 LLM 进行的决策体现出推演者的决策风格。假设，如果推演者倾向于看重近期的历史信息对当前决策的影响，即决策更加侧重长远的历史信息还是更加侧重近期的历史信息，就可以提高近因性的权重系数。最终使用这个总分数来全面确定应提取的提示，并指导 Agent 基于这些提示生成相应和合理的行动计划。

### 3.2.2 反思流

然而，在实际兵棋环境中，记忆流的观察性能在决策过程中存在局限性。基于原始观察的推理不足以使大语言模型生成高级别的决策结果。因此，需要通过信息观察和行动规划进行推断和生成高级别的推理语义。本文将这个推理过程定义为更高级别的记忆流，称为反思。它本质上是一个更高级别，更抽象的思考过程。反思与记忆流一同生成，但在前面的记忆流中的检索功能中区分了反思的生成。当检索函数中的分数超过一定的阈值时，就会触发反思。这个反思过程涉及到对先前观察到的信息的更高级别的抽象和理解。它本质上是通过提示生成的观察语义和计划语义的结合，并定期生成，为兵棋 Agent 提供推理语义。

反思的第一步是基于兵棋 Agent 的先前经验流提出问题并澄清反思过程。例如，蓝方 Agent 正在接近道路并加速朝夺控点前进。规划建议红方 Agent 应该到达网格 1403 并在该点射击蓝方 Agent。基于这个情况，产生了反思语义：蓝方 Agent 构成了重大威胁，并可在合理位置进行打击。反思过程使 Agent 不仅能够反思他们当前的观察，还可以反思其他反思。因此，Agent 生成的记忆在反思机制下可以分为不同的层次，从而允许在抽象水平上进行更准确的决策。

### 3.2.3 任务规划流

战略 Agent 基于我方所有 Agent 观察到的当前情况将其描述为一个按照特定格式的提示：< 总结：... 观察：... 规划：...>，如例 3 所示。总结的目标是将当前情况从视觉信息转化为语义信息[20-25]。

观察描述了所有 Agent 观察到的具体情况，进一步丰富了基于总结的语义信息。规划涉及基于观察到的情况进行任务分配和执行。

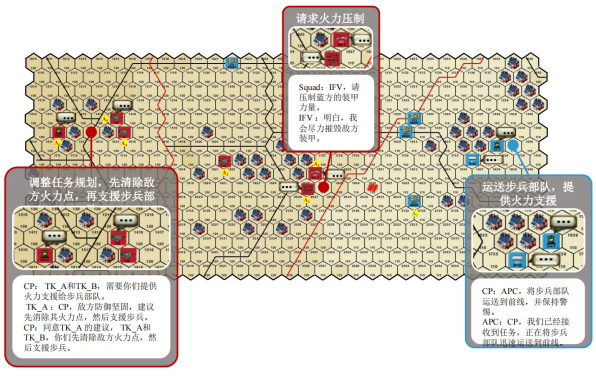


**图 8 展示自生成的兵棋 AI 在环境中的特定交互**

**例 3** 总结：我方的 6 名 Agent 正在前往夺控点，已经发现了 3 名蓝方 Agent。观察：蓝方代理

1 正在接近夺控点。规划：红方 $Agent_{1-3}$ 将优先与蓝方 $Agent_1$ 交战，而 $Agent_{4-6}$ 将迅速前往夺控点，如图 8 所示。

## 4 模拟实验环境验证

### 4.1 实验环境展示

本文使用“先胜一号”推演平台，如图 1，来演练和验证本文提出基于大语言模型的自生成 AI 在复杂战术决策中的有效性和适应性。在兵棋实验中，红方和蓝方将采用智能算法进行战术决策和执行动作[26]，实验环境复现了战术级的城镇居民对抗，基于博弈双方的基本行动规则和对抗规则，其中具体的博弈目标和行动、对抗规则请参阅第 1 小节。

### 4.2 大模型在智能决策领域的优越性

在先前的实验中，我们主要通过规则 AI 和强化学习 AI 做出决策。这项工作创新性的使用大语言模型为 Agent 做出决策，并在这个平台上进行了验证。有趣的是，这项工作发现大语言模型与强化学习之间存在较大的差异。首先，经过训练的大语言模型可以无需等待训练收敛做出决策，并直接达到较高水平的智能。本文考虑可能是大模型本身已经过大量预训练，在不熟悉想定的情况下，依然具有较强的智能性。而强化学习算法针对某个想定通常需要大量的训练来逐渐适应新任务，在刚开始的试错阶段智能性较低。同时，与强化学习算法相比，使用大语言模型做出的决策可以在多个不同任务中直接达到较高的智能水平（见表 2），而不需要为不同任务重新训练，这对实际应用具有很高的价值。本文提出了两种算法，GWA 算法和 GWAE 算法。GWA 算法采用了本文提出的双层模型，并利用 ChatGPT 进行大语言模型的决策。GWAE 在 GWA 的基础上输入了专家经验。本文以文档形式输入了兵棋推演的专家经验。

实验比较了本文提出的 GWAE 和 GWA 算法，并与 RNM-PPO[28]、3WMADM-PPO[29]、PPO、PK-DQN[27] 和 DQN 算法的胜率进行了比较。所有这些算法作为红方，由 DQN 强化学习算法驱动一个蓝方在对抗兵棋中行动。这可以确保所有红方对手保持一致，然后通过比较其胜率来确定每个红方算法的智能性。打击中的数字代表的含义是消灭掉敌蓝方算子的积分，例如一局推演中消灭一个蓝方算子，红方获得 5 分，累积推演 200 episode，累积获得的分数。夺控中的数字代表的含义是占据夺控点获得的分数，存活中的数字代表的含义是我方剩余算子获得的分数。本文做了五次实验，每次 200episode，±后面数字的含义是这五次实验的标准差，来看是否稳定。通过图 9,10,11,12 可以发现在固定蓝方智能和固定推理情景的前提下，本文提出的算法的累积胜率、平均胜率明显优于以前的经典强化学习算法，使用大语言模型进行智能决策的总体效果也相对稳定，总体胜率波动较小。

**表 2 不同算法在三个想定（关键任务分别为打击、夺控、存活）中的得分**

| 方法 | 想定 | | |
|---|---|---|---|
| | 打击关键 | 夺控关键 | 存活关键 |
| GWAE(无训练) | 3320±9 | 10504±64 | 6238±28 |
| GWA(无训练) | 2980±11 | 9106±99 | 5102±33 |
| 3WMADM-PPO(训练 10h) | 761±16 | 9001±50 | 4996±31 |
| RNM-PPO(训练 10h) | 1285±7 | 9102±141 | 4985±44 |
| PPO(训练 10h) | 850±19 | 7804±44 | 5068±38 |
| PK-DQN(训练 10h) | 792±14 | 7732±60 | 5026±53 |
| DQN(训练 10h) | 745±9 | 6948±161 | 4154±57 |

本文对提出的 GWAE 和 GWA 算法与 3WMADM-PPO、RNM-PPO、PPO、PK-DQN 和 DQN 算法实验对比胜率如图 9，10。实验训练回合数为 200 episodes，合计胜率 100%。实验表明使用大语言模型做出决策的 GWA 算法优于强化学习算法。只有 RNM-PPO 算法接近 GWA。如果为 GWA 提供专家经验文档，GWAE 算法的胜率将显著提高。总体而言 GWAE 和 GWA 算法具有更高的胜率和稳定性。

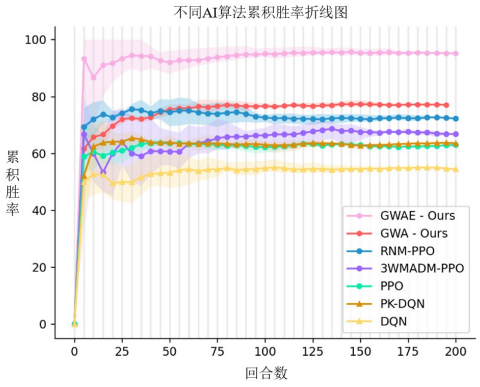


**图 9 所有算法在不同轮次中的胜率折线图**

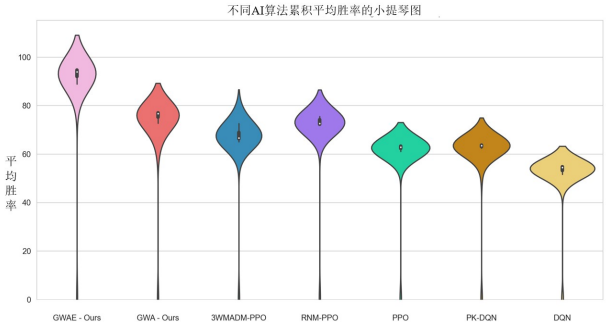


**图 10 所有算法在不同轮次中平均胜率的小提琴图**

本文比较了 GWAE、GWA、PPO、DQN 算法的总体胜率，以及算法间的相关性，对角线上的柱状图表明对应算法的胜率分布，图 11 可以看到结合专家先验知识的 GWAE 算法胜率达到 80%，而 GWA 算法胜率也有 75% 左右，均显著高于其他算法。其余散点图对应横纵坐标算法的相关性，点分布越集中分布在 y=x 线上，对应算法相关性越强，点集中分布在 y=x 线下方，表示 x 坐标对应算法表现好；反之表示 y 坐标对应算法表现好。总体而言，GWAE 和 GWE 算法的相关性较强，GWAE 算法平均胜率高。

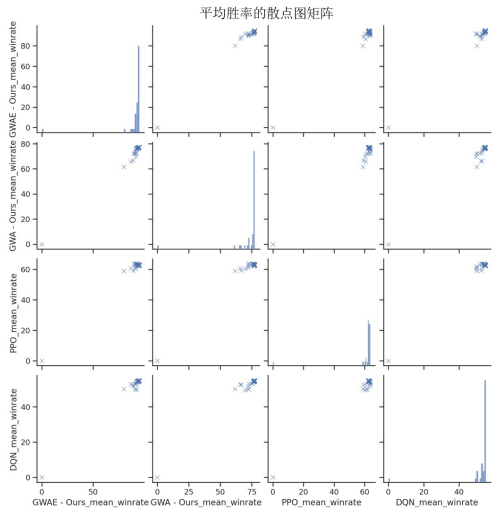


**图 11　主要算法的平均胜率散点图**

本实验一共有三种想定，分别为打击想定，夺控想定，存活想定，每个想定都有三个子任务（打击、夺控、存活），但每个想定的侧重点都有不同（如打击想定的侧重点为打击任务）。而图 12 采用的是夺控

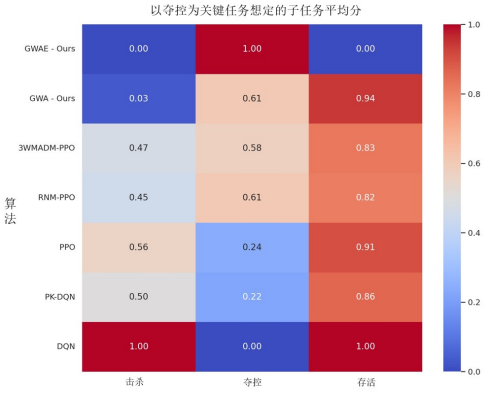


**图 12　任务平均分数热度图**

想定，展示了 GWAE、GWA、3WMADM-PPO、RNM-PPO、PPO、PK-DQN 和 DQN 算法在三个典型任务（打击、夺控、存活）中的表现。颜色越深，表示算法在该任务中对侧重点能力的表现越好，即智能性越强。可以看见的是 GWAE 算法和 GWA 算法在夺控想定中对于夺控的得分高于其他算法。

## 5　结论

本研究探索了大语言模型在智能决策领域的应用，并在兵棋模拟中验证了其有效性。结果表明，相对于强化学习方法，预训练的大语言模型能够迅速适应新任务和环境，并且在决策能力上具有显著优势。这种模型无需针对特定任务进行大量迭代训练即可展现出高智能性和广泛的泛化能力。实验亦证实，大语言模型的表现受到适当提示的显著影响，提示的优化可进一步提升其智能决策性能。未来的工作将探索大语言模型在更多场景下的适用性，例如其在复杂兵棋对抗环境和多智能体协作中的表现，并计划引入国产大模型以测试其在本地化兵棋应用中的性能。这些成果不仅展示了大语言模型在智能决策中的潜力，也为未来研究提供了新的方向和视角。

## 参考文献 (References)

[1] Van Dis E A M, Johan B, Willem Z, et al. ChatGPT: Five Priorities for Research[J]. Nature, 2023, 614(7947): 224-226.

[2] Stokel-Walker C, Van Noorden R. What ChatGPT and Generative AI Mean for Science[J]. Nature, 2023, 614(7947): 214-216.

[3] Surameery N M S, Shakor M Y. Use Chat GPT to Solve Programming Bugs[J]. International Journal of Information technology and Computer Engineering, 2023:17-22.

[4] Biswas S S. Role of Chat GPT in Public Health[J]. Annals of Biomedical Engineering, 2023, 51(05): 868-869.

[5] Biswas S S. Potential Use of Chat GPT in Global Warming[J]. Annals of Biomedical Engineering ,2023, 51(6):1126-1127.

[6] 程恺, 陈刚, 余晓晗等. 知识牵引与数据驱动的兵棋 AI 设计及关键技术 [J]. 系统工程与电子技术,2021,43(10):2911-2917.
(Cheng K, Chen G, Yu X H, et al. Design and Key Technologies of Knowledge Driven and Data Driven Military AI[J]. Systems Engineering and Electronic Technology, 2021-43(10): 2911-2917.)

[7] 刘满, 张宏军, 郝文宁, 等. 战术级兵棋实体作战行动智能决策方法 [J]. 控制与决策, 2020, 35(12): 2977-2985.
(Liu M, Zhang H J, Hao W N, et al. Intelligent Decision Method for Tactical Level wargame Entity Operations[J]. Control and Decision, 2020,35(12): 2977-2985.)

[8] 张健, 潘耀宗, 杨海涛, 等. 基于蒙特卡洛 Q 值函数的多智能体决策方法 [J]. 控制与决策, 2020, 35(03): 637-644.
(Zhang J, Pan Y Z, Yang H T, et al. Multi-agent decision making using Monte Carlo Q-value function[J]. Control and Decision, 2020,35(3): 637-644.)

[9] 李琛, 黄炎焱, 张永亮, 等. Actor-Critic 框架下的多智

能体决策方法及其在兵棋上的应用 [J]. 系统工程与电子技术, 2021, 43(03): 755-762.
(Li C, Huang Y Y, Zhang Y L, et al. A multi-Agent decision-making method under the Actor Critic framework and its application in wargame[J]. Systems Engineering and Electronic Technology, 2021,43(3): 755-762)

[10] Chen L, Zhang Y, Feng Y, et al. A Human-Machine Agent Based on Active Reinforcement Learning for Target Classification in Wargame[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 7515-7528.

[11] Chen L, Liang X X, Feng Y H, et al. A Human-Machine Agent Based on Active Reinforcement Learning for Target Classification in Wargame[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 卷号:1-13.

[12] 徐佳乐, 张海东, 赵东海, 等. 基于卷积神经网络的陆战兵棋战术机动策略学习 [J]. 系统仿真学报, 2022, 34(10): 2181-2193.
(Xu J L, Zhang H D, Zhao D H, et al. Tactical maneuver strategy learning of land war based on Convolutional neural network[J]. Journal of System Simulation, 2022,34 (10): 2181-2193.)

[13] Sun Y X, Yuan B, Xiang Q, et al. Intelligent Decision-Making and Human Language Communication Based on Deep Reinforcement Learning in a Wargame Environment[J]. IEEE Transactions on Human-Machine Systems, 2022, 53(1): 201-214.

[14] Xu J L, Hu J, Wang S X, et al. MiaoSuan Wargame: A Multi-Mode Integrated Platform for Imperfect Information Game[C]. 2022 IEEE Conference on Games (CoG). Beijing: IEEE, 2022: 457-464.

[15] Ye D H, Liu Z, Sun M F, et al. Mastering Complex Control in MOBA Games with Deep Reinforcement Learning[J]. AAAI Conference on Artificial Intelligence, 2020, 34(04): 6672-6679.

[16] Ye D H, Chen G B, Zhang W, et al. Towards Playing Full MOBA Games with Deep Reinforcement Learning[J]. Neural Information Processing Systems, 2020, 34(04): 6672-6679.

[17] Dong L W, Li N, Yuan H T, et al. Accelerating wargaming reinforcement learning by dynamic multi-demonstrator ensemble[J]. Information Sciences, 2023, 648: 119534.

[18] Mnih V, Kavukcuoglu K, Silver D, et al. Human-Level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540): 529-533.

[19] Silver D, Huang A, Maddison CJ, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search[J]. Nature, 2016, 529(7587): 484-489.

[20] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster Level in StarCraftII Using Multi-Agent Reinforcement Learning[J]. Nature, 2019, 575(7782): 350-354.

[21] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述 [J]. 智能科学与技术学报,2020,2(04):314-326.
(Liu C Y, Mu C X, Sun C Y. Overview of Deep reinforcement learning algorithm and application research[J]. Journal of Intelligent Science and Technology, 2020, 2(04): 314-326.)

[22] 孙宇祥, 彭益辉, 李斌, 等. 智能博弈综述：游戏 AI 对作战推演的启示 [J]. 智能科学与技术学报, 2022, 4(02): 157-173.
(Sun Y X, Peng Y H, Li B, et al. Review of Intelligent Games: Implications of Game AI for Combat Inference[J]. Journal of Intelligent Science and Technology, 2022, 4(2): 157-173.)

[23] Wurman P R, Barrett S, Kawamoto K, et al. Outracing Champion Gran Turismo Drivers with Deep Reinforcement Learning[J]. Nature, 2022, 602(7896): 223-228.

[24] Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering Atari, Go, Chess and Shogi by Planning With a Learned Model[J]. Nature, 2020, 588(7839): 604-609.

[25] Silver D, Hubert T, Schrittwieser J, et al. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-play[J]. Science, 2018, 362(6419): 1140-1144.

[26] Sun Y X, Yuan B, Xiang Q, et al. Intelligent Decision-Making and Human Language Communication Based on Deep Reinforcement Learning in a Wargame Environment[J]. IEEE Transactions on Human-Machine Systems, 2023, 53(1): 201-214.

[27] Sun Y X, Yuan B, Zhang T, et al. Research and Implementation of Intelligent Decision Based on a Priori Knowledge and DQN Algorithms in Wargame Environment[J]. Electronics, 2020, 9(10): 1668.

[28] Xue Y F, Sun Y X, Zhou J W, et al. Multi-attribute decision-making in wargames leveraging the Entropy-Weight method in conjunction with deep reinforcement learning[J]. IEEE Transactions on Games, 2023:1-12

[29] Zhang Q, Ju Y, Luis M, et al. The SMAA-TWD model: A novel stochastic multi-attribute three-way decision with interrelated attributes in triangular fuzzy information systems[J]. Information Sciences, 2022, 618: 14-38.

## 作者简介

孙宇祥(1990−), 男，助理研究员，博士，从事智能博弈与决策、智能决策可解释性等方向的研究，E-mail: sunyuxiang@nju.edu.cn;

赵俊杰(1999−), 男，学生，硕士研究生，从事量化金融、强化学习、智能博弈等研究，E-mail: junjiezhao@smail.nju.edu.cn;

解宇轩(2001−), 男，学生，硕士研究生，从事强化学习、智能博弈与决策等，E-mail: 522023150059@smail.nju.edu.cn;

喻车澄(2003−), 男，学生，本科生，从事强化学习、智能博弈与决策等，E-mail：211870228@smail.nju.edu.cn;

周献中(1962−), 男，教授，博士，从事强化学习、智能博弈与决策等，E-mail：z houxz@nju.edu.cn.