专题：智能体通信与网络技术

基于冻结主干双回路进化的多智能体潜空间策略研究

赵袁赫[1]，王昱[1]，李昊忱[2]，杨士铎[1]，郭柯楠[1]，高晨[1]，白小高[1]，管孜然[1]，王继业[1]
（1. 华北电力大学控制与计算机工程学院，北京 102206；
2. 交通运输部天津水运工程科学研究院，天津 300072）

摘　要：针对大语言模型多智能体面临的“稳定性—可塑性困境”及微调引发的灾难性遗忘风险，提出一种基于冻结主干的双回路进化框架。该框架将通用推理与策略偏好解耦：行为回路利用离散价值迭代优化即时动作；语义回路引入预测误差调制的语义反思机制，将自然语言反馈转化为高维语义梯度，驱动外部潜向量在连续语义空间中演化。在GridWorld博弈环境实验中，该方法在完全冻结参数前提下实现了85.6%的任务成功率，路径加权成功率达到最高，且具备极低的推理延迟（2.1 s/step）和Token消耗（3.2 k/session）。实验还观察到基于隐式因果推理的社会协作行为，验证了该范式在低资源消耗与高可解释性方面的显著潜力。

关键词：大语言模型；多智能体系统；持续学习；双回路进化
中图分类号：TP181；TN915；TP39
文献标志码：A
doi: 10.11959/j.issn.1000−0801.DXKX260121

Research on multi-agent latent space strategies based on frozen main branch dual-loop evolution

Zhao Yuanhe[1], Wang Yu[1], Li Haochen[2], Yang Shiduo[1], Guo Kenan[1], Gao Chen[1], Bai Xiaogao[1], Guan Ziran[1], Wang Jiye[1]
1. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China
2. Tianjin Research Institute for Water Transport Engineering, M.O.T., Tianjin 300072, China

Abstract: To address the “stability-plasticity dilemma” faced by multi-agent systems based on large language models and the risk of catastrophic forgetting caused by fine-tuning, a frozen main branch dual-loop evolution framework was proposed. General reasoning was decoupled from policy preferences by this framework. Discrete value iteration was utilized to optimize immediate actions by the behavioral loop. A prediction error-modulated semantic reflection mechanism was introduced by the semantic loop that transformed natural language feedback into high-dimensional semantic gradients, driving the evolution of external latent vectors in a continuous semantic space. In experiments conducted in the GridWorld game environment, this method achieves a 85.6% task success rate under fully frozen parameters, with the highest path-weighted success rate, while exhibiting extremely low inference latency (2.1 s/step) and token consumption (3.2 k/session). The experiments also observe socially cooperative behavior based on implicit

收稿日期：2026−02−25；修回日期：2026−05−20
通信作者：王继业，jiyewang@sgcc.com.cn

causal reasoning, validating the significant potential of this paradigm in terms of low resource consumption and high interpretability.

Key words: large language model, multi-agent system, continuous learning, dual-loop evolution

0　引言

基于大语言模型（LLM）的自主智能体系统已从单体任务规划迈向更复杂的群体社会模拟与动态博弈[1-3]。现有研究表明，基于 LLM 的多智能体系统在模拟人类社会行为及开放世界探索中展现了超越传统强化学习的泛化潜力[4]。然而，构建能够在动态、非平稳环境中长期生存的系统，仍受制于“稳定性—可塑性困境”，即智能体既需要保持预训练获得的通用认知（稳定性），又必须具备根据环境反馈快速重构策略以适应任务需求的能力（可塑性）[5]。为了明确本文研究的定位，有必要回顾现有解决这一困境的主流范式及其面临的瓶颈。

在智能体推理与反思机制方面，随着模型规模增长，基于提示工程的范式（如 CoT[6] 和 ReAct[7]）取得了显著进展。此类方法主要通过最大化条件概率来生成反思文本并优化后续动作，依赖存储在滑动窗口中的自然语言反馈[8-9]。现有基于上下文的方法（如 Reflexion[10] 和 CoALA[11]）试图通过堆叠历史记忆引导行为。然而，这类方法面临两大瓶颈：一是随着时间推移，累积历史易超出窗口限制，导致早期关键信息被忽略[12-13]；二是自然语言作为策略载体过于稀疏低效，受限于离散空间的组合爆炸，难以像数值向量那样精确编码连续变量的梯度变化，并在长程任务中产生昂贵的 Token 开销。

为了克服离散符号优化的局限性，另一类研究转向连续潜空间的参数微调。参数高效微调技术将策略搜索空间扩展到连续潜空间，相关研究证明软提示向量的有效性[14-15]。虽然基于参数微调的方法能注入新知，但在持续微调中存在显著的灾难性遗忘风险，往往会破坏模型原有的通用语义表征[16-17]，且多智能体同时更新会导致环境状态不稳定，致使联合策略难以收敛[18]。主流商业大模型通常仅提供应用程序接口（API）访问，导致内部梯度不可见，这使得依赖梯度的优化方法在黑盒多智能体博弈中应用受限，而现有黑盒优化方法往往缺乏语义归因，样本效率极低[19]。

在多目标资源采集等复杂场景中，传统多目标强化学习通常采用线性标量化方法将多目标集合简化为单一价值函数的优化[20]。然而，现有分层架构存在显著的耦合失效问题，权重向量通常是静态预设的，难以应对动态非平稳环境。当前研究缺乏一种有效机制，能够利用深层语义反思来动态调制行为策略的偏好权重。

针对上述问题，本文受认知科学双重处理理论启发[21]，提出一种基于冻结主干的双回路进化框架。核心理念是将通用推理能力与特异性策略偏好解耦：LLM 被视为冻结的通用认知引擎，而进化策略被外置为可学习的连续潜向量。不同于 LSPO[22] 将文本聚类为离散策略簇，该方法允许策略在连续语义空间中进行平滑梯度更新。通过行为回路和语义回路的协同，该框架在规避昂贵微调成本的同时，实现了稳定的策略收敛。

本文主要贡献包括：（1）提出一种基于潜空间的解耦进化架构，利用双回路机制将策略优化从模型参数更新中剥离，有效规避了灾难性遗忘；（2）定义了反思驱动的连续策略表征，构建了可经由强化学习更新的外部潜空间，填补了符号化反思与数值化强化学习之间的鸿沟；（3）揭示了异构智能体间的隐式因果协作，证实了系统在缺乏显式共享奖励约束下，能够涌现出基于因果推断的协作模式。

1 基本定义

定义1 异构多智能体系统

本文将多地形资源采集任务建模为一个任务特化、部分可观测的马尔可夫决策过程，形式化为元组$\mathcal{G}=\langle \mathcal{W},\mathcal{U},\mathcal{S},\Omega,\mathcal{R}\rangle$。其中，交互环境$\mathcal{W}$被定义为一个非均匀势能场，由映射函数$\mathcal{T}:\mathcal{W}\to\mathcal{R}$决定空间中任意坐标的移动阻尼与能量耗散率；智能体集合$\mathcal{U}=\{u_1,\cdots,u_N\}$包含$N$个异构的功能算子，每个节点$u_k$被预设为优化特定的子目标函数$\mathcal{J}_k$；系统状态$\mathcal{S}$包含全局物理坐标与机体能量状态；而观测空间$\Omega$限制了每个智能体仅能获取与其功能偏好相关的局部多模态信息；异构奖励函数集合$\mathcal{R}$定义了各算子的内在驱动力，系统通过元控制器对各私有奖励的动态加权，隐式地逼近全局帕累托前沿[23]。

定义2 连续潜空间

针对异构大语言模型在多智能体协作中面临的通信与交互瓶颈，本文将潜空间定义为一个高维、连续的共享语义流形$\mathcal{H}\subset R^d$。不同于离散且极易引发组合爆炸的自然语言通信，该空间作为多智能体底层交互的统一介质，允许不同模型间的文本反馈与环境观测转化为连续可微的数值表示。

定义3 潜向量表征

本文将每个异构智能体u_k的功能优先级与行为偏好具体实例化为可学习的潜向量$\boldsymbol{h_k}\in\mathcal{H}$。通过将策略表征为连续潜向量，系统能够在完全冻结主干参数的前提下，利用高维语义梯度驱动策略在数值空间中平滑演化。智能体在时间步t的策略函数形式化表示为：

$$\pi_k\left(a|\boldsymbol{o}_{k,t};\boldsymbol{h}_{k,t}\right)=\mathrm{LLM}_{\mathrm{Dec}}\left(\boldsymbol{h}_{k,t}\oplus\mathrm{Enc}\left(\boldsymbol{o}_{k,t}\right);\boldsymbol{\Phi}\right)\quad(1)$$

其中，$\boldsymbol{\Phi}$为冻结的模型主干参数，Dec(·)与Enc(·)分别为解码器与编码器，⊕表示向量拼接操作。

问题定义 参数冻结下的双回路多目标优化

给定异构算子集合$\mathcal{U}$与冻结模型$\mathcal{M}_\Phi$，本文目标是在时间视界T内，通过双回路机制协同优化行为策略与潜在表征。该问题被分解为两个耦合的时间尺度：（1）在行为层中，每个智能体u_k通过时序差分学习寻找最优动作价值$Q_k^*(s,a)$以最大化其私有子目标期望回报；（2）在认知层中，系统寻找最优潜在轨迹$\mathcal{H}^*$，以最小化行为层的预测误差δ。最终优化目标是找到一组策略组合，使得在满足物理硬约束的前提下，最大化由元控制器动态加权的多目标累积收益：

$$\max_{\mathcal{H}}\ \sum_{t=0}^{T}\sum_{u_k\in\mathcal{U}}\boldsymbol{S_{k,t}}\cdot\mathcal{R}_k(s_t,a_t)\quad\text{s.t.}\quad\boldsymbol{\Phi}=\text{const.}\quad(2)$$

其中，$S_{k,t}$为基于预测误差调制的动态显著性权重。

2 双回路潜空间演化方法

双回路进化框架如图1所示，旨在通过快慢两个更新通道协调稳定性与可塑性。与传统微调方法通过反向传播直接修改模型权重$\boldsymbol{\Phi}$、导致新知识覆盖旧有表征从而引发灾难性遗忘不同，本文框架将所有策略变化外置为潜向量$\boldsymbol{h_k}$的更新，模型主干参数$\boldsymbol{\Phi}$始终保持冻结，新环境信息只改变外部潜向量而不触及内部权重，从机制上切断灾难性遗忘的发生路径。在此基础上，行为回路负责基于离散价值函数优化的即时物理决策，而语义回路引入预测误差调制的自适应更新机制，将行为层的预测误差转化为语义更新信号，驱动潜向量在连续空间中进行长程演化。两者协同工作，使系统能够在多地形资源采集任务中，动态平衡多目标冲突并实现策略收敛。

2.1 基于离散价值更新的行为回路

行为回路旨在冻结大语言模型主干参数$\boldsymbol{\Phi}$的约束下，通过异构功能算子将全局任务分解为局部的即时价值评估，实现高频、低延迟的物理交互。

2.1.1 异构算子模块设计

为了在多地形资源采集任务中逼近帕累托最

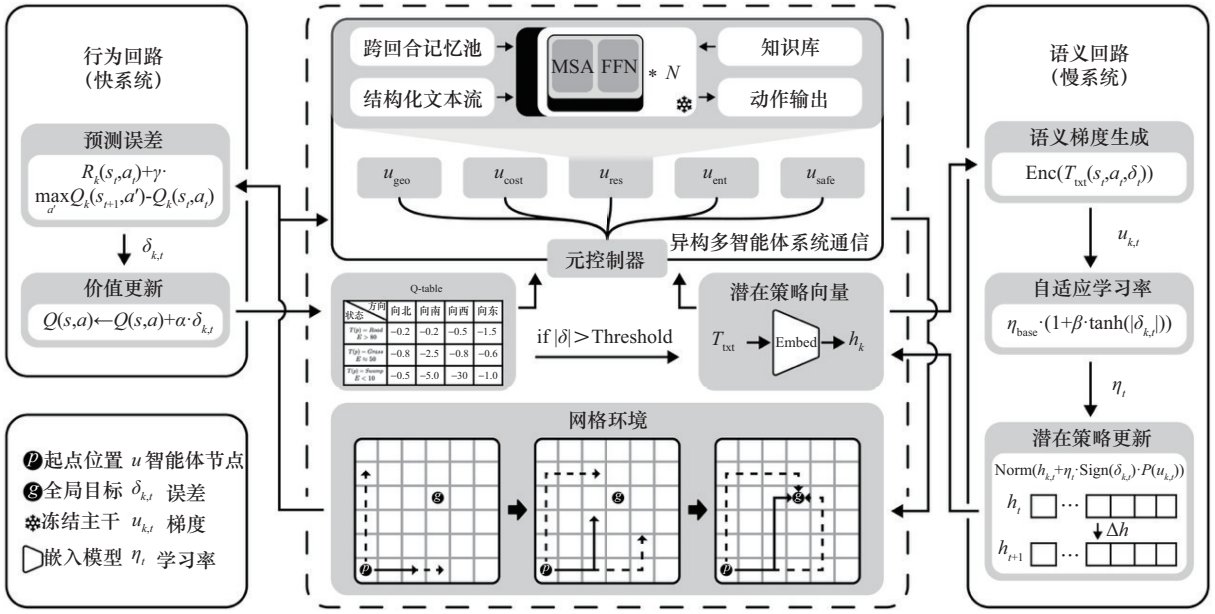


图1　双回路进化框架

优前沿，本文将抽象智能体集合$\mathcal{U}$具体实例化为5个功能特异的算子模块：$\mathcal{U}=\{u_{geo},u_{cost},u_{res},u_{ent},u_{safe}\}$。异构智能体结构如图2所示，这5个模块共同构成一个多目标动态博弈系统。

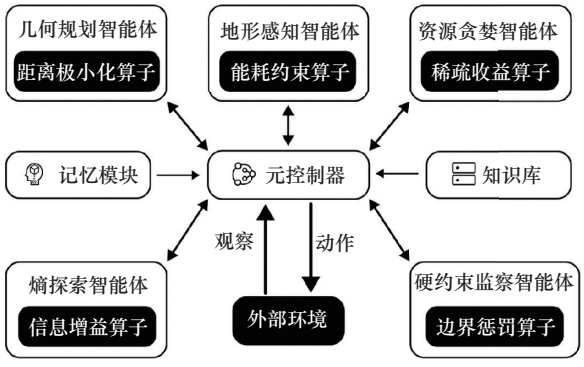


图2　异构智能体结构

（1）几何规划智能体（u_{geo}）——距离极小化算子：该模块旨在寻找连接当前位置p_t与全局目标g的最短路径。其奖励函数定义为距离势能的差分：

$$R_{geo}(s_t,a_t)=\zeta\cdot\left(|p_t-g|_2-|p_{t+1}-g|_2\right)\tag{3}$$

其中，ζ为几何缩放因子，$|\cdot|_2$表示L2范数。

（2）地形感知智能体（u_{cost}）——能耗约束算子：该模块旨在最小化移动过程中的能量损耗，其目标是优化系统的续航能力，规避高阻力区域。奖励函数与地块能耗系数$\omega(\cdot)$呈负相关：

$$R_{cost}(s_t,a_t)=-\psi\cdot\omega\left(\tau_{p_{t+1}}\right)\tag{4}$$

其中，$\tau(p)$表示坐标处的地形类别，ψ为成本敏感系数。

（3）资源贪婪智能体（u_{res}）——稀疏收益算子：该模块模拟局部贪婪算法，专注于最大化任务过程中的累积资源获取量。其奖励函数由稀疏的离散信号触发：

$$R_{res}(s_t,a_t)=\begin{cases}\Lambda, & p_{t+1}\in S_{resource}\\ 0, & \text{其他}\end{cases}\tag{5}$$

其中，$S_{resource}$为资源坐标集合，Λ为单次采集的奖励增益。

（4）熵探索智能体（u_{ent}）——信息增益算子：该模块基于状态访问频率驱动系统探索未知领域。它利用基于计数的内在动机来奖励新颖状态：

$$R_{ent}(s_t,a_t)=\frac{\chi}{\sqrt{N(p_{t+1})+1}}\tag{6}$$

其中，$N(p)$记录了历史轨迹中该坐标的被访问次数，χ为探索激励常数。

（5）硬约束监察智能体（u_{safe}）——边界惩罚算子：该模块作为系统的安全约束模块，负责监测不可行状态。不同于其他模块的软引导，它施加较大的惩罚信号：

$$R_{\text{safe}}(s_t,a_t)=-\Omega\cdot I\left(p_{t+1}\in O_{\text{obstacle}}\vee E_{t+1}<0\right) \quad (7)$$

其中，O_{obstacle}为障碍物集合，E为系统剩余能量，$I(\cdot)$为示性函数，$\Omega\gg 1$为惩罚阈值。

2.1.2　基于显著性的元控制器策略

元控制器充当系统的中央仲裁枢纽，负责在异构算子的策略建议π_k之间寻找动态平衡。系统维护一个动态显著性向量$\boldsymbol{S_t}$，最终动作a_{meta}通过显著性加权的软投票机制生成：

$$a_{\text{meta}}=\underset{a\in\mathcal{A}}{\text{argmax}}\sum_{u_k\in\mathcal{U}}\frac{\exp\left(\boldsymbol{S}_{k,t}/\tau\right)}{\sum_j\exp\left(\boldsymbol{S}_{j,t}/\tau\right)}\cdot P(a|\pi_k) \quad (8)$$

其中，τ为温度系数，该机制类似于加权投票机制：当前环境越紧迫（如能量告急），系统越倾向于赋予安全类智能体更高的投票权重，从而动态调整最终执行动作。

2.1.3　离散价值迭代与快速反馈

为了赋予行为回路基本的试错学习能力，每个功能算子u_k维护一个私有的离散价值表$Q_k(s,a)$。该价值表通过时序差分误差$\delta_{k,t}$进行高频更新：

$$\delta_{k,t}=\mathcal{R}_k(s_t,a_t)+\gamma\cdot\max_{a'}Q_k(s_{t+1},a')-Q_k(s_t,a_t) \quad (9)$$

其中，$\mathcal{R}_k$为异构奖励函数。随后，价值表依据$\delta_{k,t}$进行即时更新：

$$Q_k(s_t,a_{t+1})\leftarrow Q_k(s_t,a_t)+\alpha\cdot\delta_{k,t} \quad (10)$$

这一价值反馈回路确保智能体能够快速锁定高价值行为，并为语义回路提供关键的认知惊奇信号。

2.2　基于潜空间策略更新的语义回路

语义回路构成了系统的长期记忆与个性化进化通路。当系统遭遇较大预测误差时，语义回路被激活，利用预测误差调制的语义梯度对潜向量$\boldsymbol{h_k}$进长期更新，从而实现策略在连续语义空间中的自适应演化。

2.2.1　基于预测误差的语义反思机制

为了连接物理反馈与策略表征，本文引入基于预测误差的语义反思机制。只有当预期结果与实际体验产生显著偏差时，系统才会激活慢通道进行结构化归因分析。在每轮交互结束后，系统首先计算行为回路的时序差分误差：

$$\delta_t=r_{t+1}+\gamma\max Q(s_{t+1},a')-Q(s_t,a_t) \quad (11)$$

该误差值充当语义反思的门控触发器：当$|\delta_t|$超过自适应阈值时，触发智能体u_k基于当前轨迹τ_t生成自然语言反思文本T_{txt}。反思文本的生成不再是简单的状态描述，而是被设计为一种结构化的归因分析。

2.2.2　基于预测误差调制的策略进化

为了将上述非结构化的因果逻辑转化为可计算的优化信号，本文首创语义梯度投影机制。首先，利用预训练的编码器将反思文本映射到高维语义空间，生成语义梯度向量$\boldsymbol{u_{k,t}}$：

$$\boldsymbol{u_{k,t}}=\text{Enc}\left(T_{\text{txt}}(s_t,a_t,\delta_t)\right)\in R^D \quad (12)$$

向量$\boldsymbol{u_{k,t}}$表示反思文本中与策略偏差相关的语义信息，指明策略在语义空间中优化的具体方向，它将作为语义梯度，在双回路架构中直接驱动潜向量的更新。

为了平衡策略的稳定性与可塑性，定义动态学习率η_t为预测误差绝对值的非线性函数：

$$\eta_t=\eta_{\text{base}}\cdot\left(1+\beta\cdot\tanh\left(|\delta_{k,t}|\right)\right) \quad (13)$$

其中，η_{base}为基础代谢学习率，β为敏感度系数。当$\delta_{k,t}\to 0$时，η_t维持在低位，系统处于策略固化期；当$\delta_{k,t}$剧增时，η_t迅速放大，系统进入策略重塑期。基于此，潜向量$\boldsymbol{h_{k,t}}$利用生成的语义梯

度进行投影更新：

$$\boldsymbol{h}_{k,t+1} \leftarrow \mathrm{Norm}\left(\boldsymbol{h}_{k,t}+\eta_t \cdot \mathrm{Sign}\left(\delta_{k,t}\right) \cdot P\left(\boldsymbol{u}_{k,t}\right)\right) \quad (14)$$

其中，$P(\cdot)$是将通用语义空间映射到特定策略子空间的投影函数，$\mathrm{Sign}\left(\delta_{k,t}\right)$决定新方向是增强还是抑制该语义特征。该更新机制通过三重设计保障长期稳定性：$\mathrm{Norm}(\cdot)$约束潜向量模长恒定，防止数值爆炸或消亡；动态学习率 η_t在预测误差趋近于 0 时自动退火，避免策略收敛后的无效震荡；投影函数$P(\cdot)$ 将更新限制在特定策略子空间内，防止潜向量长期迭代中发生无约束语义漂移。三重机制协同作用，确保潜空间演化的有界稳定性。

2.2.3　能量动力学与记忆迁移

为了模拟受限资源环境下的非平稳交互特性，本文将环境W建模为一个非均匀势能场，其势能梯度$T\left(p\right)$的幅值反映了智能体在该位置移动时所受的阻尼强度与能量耗散率。为将连续势能场转化为可计算的离散决策信号，本文将$T\left(p\right)$按地形类别进行分段线性量化：梯度越大对应的移动阻力越强，能量耗散率越高，由此得到非线性移动代价函数$C_{\mathrm{move}}\left(p\right)$如下：

$$C_{\mathrm{move}}\left(p\right)=\begin{cases}1.0, & T(p)=\mathrm{Road}\\ 3.0, & T(p)=\mathrm{Grass}\\ 5.0, & T(p)=\mathrm{Swamp}\\ \infty, & T(p)=\mathrm{Obstacle}\end{cases} \quad (15)$$

系统演化严格受制于能量动力学约束。在每个时间步t，能量状态E_t遵循以下方程演化，其中C_{base}为基础代谢消耗，ξ为地形敏感度系数：$E_{t+1}=E_t-\xi\cdot C_{\mathrm{move}}\left(p_{t+1}\right)-C_{\mathrm{base}}$。若 $E_{t+1}\leqslant 0$，系统触发硬终止约束并判定任务失败，迫使元控制器权衡时间与能量消耗。

为了实现跨回合的知识迁移，系统在语义回路中维护情景记忆池。不同于存储原始动作序列，系统提取每回合反思文本的语义质心$\boldsymbol{m}_k$作为记忆向量：

$$\boldsymbol{m}_k=\frac{1}{|\tau_k|}\sum_{t\in\tau_k}\mathrm{Enc}\left(\mathcal{T}_{\mathrm{txt},t}\right) \quad (16)$$

在新回合开始时，系统通过计算当前地形嵌入$\boldsymbol{e}_{\mathbf{map}}$与记忆池向量的余弦相似度，检索历史最优策略模式，从而加速潜空间H的收敛。

3　实验与结果分析

3.1　智能体参数与实验环境配置

为了验证双回路进化框架的有效性，构建一个 15×15 的中型动态网格世界（初始能量 Emax=100），地图包含公路、草地、沼泽及障碍物等多种复合地形，具体生成细节见第 3.3 节。实验采用异构配置：由 Qwen3-30B-Instruct 作为元控制器处理全局逻辑，轻量级的 Qwen3-14B-Instruct 负责各功能智能体的高频执行。所有实验均在 NVIDIA A100（80 GB）节点上完成，每组实验独立重复 5 次以确保结果的稳健性，取均值与标准差作为最终报告数值。训练共包含 50 个回合，以确保策略收敛的稳定性。超参数方面设定潜向量维度为 1 024，行为回路 Q-learning 学习率为 0.1，语义回路基础更新步长为 0.05。

为了构建系统内部“收益—成本—风险”的动态张力，5 个异构智能体被赋予特定的参数阈值。几何规划智能体设定为标准化的单位权重（ζ=1.0）以提供基础方向引导；地形感知智能体（ψ=2.0）将不同地质的能耗权重严格映射为 1∶3∶5，迫使系统在长程规划中权衡能量性价比；资源贪婪智能体赋予较高的单次采集增益（Λ=10.0）以测试系统对局部诱惑的理性控制；熵探索智能体设定较小的激励常数（χ=0.5）以防止死锁；而硬约束监察智能体则施加极大的惩罚阈值（Ω=100.0），确保策略演化始终被限制在物理可行性边界之内。

3.2　对比基线

为了全方位评估双回路进化框架的综合优势，本文选取 3 类代表性方法作为基线。

首先，在基础推理与文本记忆方面，选取ReAct和Reflexion。ReAct作为依赖静态预训练知识的标准范式，旨在反衬持续学习应对动态环境的必要性；而Reflexion代表通过堆叠文本反思进行修正的主流框架，引入该基线是为了证明本文研究的潜向量表征能有效规避长程任务中的“迷失中间”效应及昂贵的Token开销。

其次，在参数优化与记忆检索层面，对比了白盒微调（Soft Prompt Tuning）、黑盒无梯度优化（BBT）以及显式结构化检索（RoboMemory）[24]。这些对比旨在验证本文提出的语义反思算子在黑盒条件下能否超越盲目的随机扰动和传统的显式检索，实现媲美白盒梯度优化的样本效率与策略迁移能力。

最后，针对高性能但高成本的方法，选取全参数微调（MAPoRL2）[25]和蒙特卡洛树搜索规划（LATS）[26]。与这两者的核心对比在于验证本文框架的效能边界：即在完全冻结主干参数且保持极低推理延迟的前提下，双回路机制能否通过轻量级的潜空间策略调整，达到相近的协作性能与规划成功率。

3.3 数据集

本文采用程序化生成方法构建动态交互数据集，而非使用静态语料库。实验环境被建模为15×15的二维离散流形，基于分层噪声逻辑随机生成复合地形，其中公路占比约20%、草地占比约45%、沼泽占比约15%、障碍物占比约10%，剩余10%为资源点分布区域。每张地图随机放置8~12个资源采集点，起点与终点均由程序随机生成，且保证二者之间至少存在一条可行路径。为了全面评估智能体的泛化能力，测试集由30张独立随机生成的地图构成，训练集与测试集之间不存在地图复用，确保评估结果反映模型的真实泛化性能而非记忆能力。每次实验会话均重置拓扑结构，迫使智能体必须通过实时探测与策略调整适应环境，避免对特定地图布局的过拟合。所有随机地图的生成均基于固定的随机数种子序列，并在实验结束后完整保存以便后续复现验证。

在观测数据流方面，系统构建了结构化语义观测流以降低推理延迟并适配大模型特性。智能体的5×5局部视野被实时转译为包含当前坐标、地形类别、能量读数及资源方位的结构化文本，模拟具身智能体的感知过滤过程。数据集完整保留了双重演化轨迹，既包含用于评估任务效能的行为轨迹链，也包含记录自然语言反思与潜空间策略向量序列的认知演化链。

3.4 评估指标

本文依据实验目标将评估体系划分为3个层面：任务效能与基线对比、潜空间策略演化分析以及社会动力学与涌现行为观测。

3.4.1 任务效能与基线对比指标

平均累积奖励（average cumulative reward，ACR）[27]用于衡量智能体在单次会话中的整体表现，定义为会话内所有K个回合的物理奖励总和的平均值：

$$\mathrm{ACR}=\frac{1}{K}\sum_{k=1}^{K}\sum_{t=0}^{T_k}R\left(s_{t,k},a_{t,k}\right) \tag{17}$$

任务成功率（success rate，SR）定义为智能体在规定的最大时间步$T_{\max}=50$内成功到达目标位置g的回合比例：

$$\mathrm{SR}=\frac{1}{N}\sum_{n=1}^{N}I\left(s_{T,n}=g\right) \tag{18}$$

为了区分低效试错与高效规划，引入路径加权成功率（success weighted by path length，SPL）[28]，它同时考量成功率与路径最优性：

$$\mathrm{SPL}=\frac{1}{N}\sum_{n=1}^{N}S_n\frac{L_n^*}{\max\left(L_n,L_n^*\right)} \tag{19}$$

上下文消耗量（context token consumption，CTC）是统计完成单次回合所需的平均输入Token数量，用于对比显式文本记忆方法与本文方法的潜向量记忆：

$$\mathrm{CTC}=\frac{1}{N}\sum_{n=1}^{N}\mathrm{Count}\left(\mathcal{T}_{\mathrm{prompt},n}\right) \tag{20}$$

推理延迟（inference latency，IL）记录智能体在单步决策中所需的平均挂钟时间，用于衡量系统的实时响应能力[29]：

$$\mathrm{IL}=\frac{\sum_{t=1}^{T_{\mathrm{total}}}\mathrm{Time}\left(a_t\right)}{T_{\mathrm{total}}} \tag{21}$$

3.4.2　潜空间策略演化指标

初始状态余弦相似度通过计算当前时刻向量$\boldsymbol{h}_t$与初始随机向量$\boldsymbol{h}_0$的夹角余弦值，评估策略演化的趋势：

$$\mathrm{Sim}_{\cos}\left(\boldsymbol{h}_t,\boldsymbol{h}_0\right)=\frac{\boldsymbol{h}_t\cdot\boldsymbol{h}_0}{|\boldsymbol{h}_t|_2|\boldsymbol{h}_0|_2} \tag{22}$$

单步 L2 语义漂移用于检测策略更新过程中的剧烈调整，定义为连续两个时间步之间潜向量的欧氏距离变化：

$$\Delta\mathcal{L}_t=|\boldsymbol{h}_t-\boldsymbol{h}_{t-1}|_2 \tag{23}$$

$\Delta\mathcal{L}_t$的局部峰值通常对应关键的反思事件，反映了语义更新的敏锐度。为了定性分析异构智能体的策略分化，采用主成分分析将高维向量集合$\mathcal{H}=\{\boldsymbol{h}_{i,t}\}$投影至二维平面[30]，观测不同角色智能体是否形成线性可分的聚类簇C_i。

3.4.3　社会动力学指标

建议采纳率用于衡量特定智能体v_i的建议被元控制器选为最终执行动作a_{meta}的频率，反映其在当前环境下的功能主导地位[31]：

$$\mathrm{AR}_i=\frac{\sum_{t=1}^{T}I\left(a_{\mathrm{meta},t}=\pi_i\left(s_t\right)\right)}{T} \tag{24}$$

信任分数轨迹用于记录元控制器对各智能体信任权重$\boldsymbol{w}_{i,t}$随时间变化的情况。通过追踪$\boldsymbol{w}_{i,t}$的波动，可以分析系统是否涌现出隐式因果推理能力[32]。

上述 3 类指标在多目标优化场景下形成层次化的互补评估体系。在任务效能层面，SR 与 SPL 共同刻画规划质量但并不完全正相关，ACR 进一步引入多目标奖励的综合维度，CTC 与 IL 则从计算代价角度对效能指标形成约束，5 项指标共同反映系统的任务表现与实际部署潜力。在潜空间演化层面，余弦相似度与 L2 语义漂移分别从全局趋势与局部突变两个维度刻画策略更新过程，PCA 投影则从定性角度验证不同角色智能体的策略分化结构。在社会动力学层面，建议采纳率与信任分数轨迹共同揭示元控制器的协调机制，二者结合可验证系统是否涌现出隐式因果推理能力。3 类指标从任务效能、策略演化与社会协作 3 个维度协同构成对双回路框架的全面评估闭环。

3.5　实验结果与分析

本节展示了双回路进化框架在 GridWorld 复杂博弈环境中的实验结果。首先通过与 SOTA 基线的横向对比，验证系统在任务效能与计算成本上的综合优势；随后深入潜空间内部，解析策略向量的结构化演化轨迹；再探讨元控制器在缺乏显式规则指导下涌现出的社会动力学特征；继而通过潜向量维度消融实验验证关键超参数的选取依据；最后在同等计算资源约束下与全参数微调及参数高效微调方法进行收敛效率对比，量化本文框架的综合效能边界。

3.5.1　总体性能与基线对比

为了验证本文方法的有效性，在包含 30 个随机生成地图的测试集上与第 3.2 节定义的 7 种基线方法进行对比实验。不同方法在动态 GridWorld 任务上的综合性能对比见表 1。

与基础推理方法相比，本文方法在动态环境中的适应能力表现更优。ReAct 主要依赖静态预训练知识，在动态地形变化的 GridWorld 环境中成功率为 45.2%，表明缺乏反馈更新机制会限制其对非平稳环境的适应能力。Reflexion 通过累积自然语言反思文本进行策略修正，成功率达到 78.5%，但 Token 消耗为 18.3 k/session，约为本文方法（3.2 k/session）的 5.8 倍，且推理延迟更高。该结果说明，以自然语言文本作为长期策略记忆

表 1　不同方法在动态 GridWorld 任务上的综合性能对比

方法	任务成功率(SR)↑	路径加权成功率(SPL)↑	平均累积奖励(ACR)↑	推理延迟(s/step)↓	Token消耗(k/session)↓
ReAct	45.2%	0.31	12.3	1.7	4.2
Reflexion	78.5%	0.65	28.4	3.5	18.3
LATS	88.1%	0.72	32.1	12.4	12.6
RoboMemory	81.2%	0.68	29.6	4.2	11.8
Soft Prompt	82.3%	0.69	30.2	1.3	3.5
MAPoRL2	84.1%	0.70	30.5	1.8	6.8
BBT	65.4%	0.52	21.6	1.6	3.4
本文方法	85.6%	0.75	31.0	2.1	3.2

载体时，容易引入上下文长度增长和计算开销增加的问题；相比之下，将反思信息压缩为连续潜向量，有助于在维持较高任务成功率的同时降低资源消耗。

在参数优化与记忆检索类方法中，本文方法在黑盒设定下取得较好的综合性能。Soft Prompt Tuning 依赖模型内部梯度信息，在仅提供 API 访问的商业大模型场景中部署受限；本文方法在主干参数冻结的条件下取得更高的成功率（85.6% vs 82.3%），说明基于语义反思的潜向量更新能够在不访问模型梯度的情况下实现有效的策略迁移。BBT 采用无语义归因的随机扰动搜索，成功率为 65.4%，表明其搜索效率相对有限。RoboMemory 通过显式检索机制实现策略记忆，成功率为 81.2%，但 Token 消耗（11.8 k/session）和推理延迟（4.2 s/step）均高于本文方法，说明隐式潜向量记忆机制在实时交互任务中具有一定的计算效率优势。

与高性能但计算成本较高的方法相比，本文方法在性能与资源开销之间取得较好的平衡。LATS 基于蒙特卡洛树搜索取得 88.1% 的最高成功率，但其推理延迟为 12.4 s/step，约为本文方法的 5.9 倍，难以适用于对实时性要求较高的交互场景。MAPoRL2 采用全参数微调与多智能体强化学习联合训练，成功率为 84.1%，与本文方法接近，但该方法依赖可访问的模型参数，并需要额外的联合训练过程；在多智能体并行更新条件下，也可能面临策略收敛稳定性问题。相比之下，本文方法在冻结主干参数的前提下取得 85.6% 的成功率，表明外置潜向量更新机制适用于低资源、黑盒模型访问条件下的多智能体策略优化任务。

3.5.2　潜空间的结构化演化

为了验证语义反思是否真正驱动了策略更新，对潜向量 $\boldsymbol{h}$ 的演化过程进行可视化分析。

潜空间向量余弦相似度变化如图 3 所示，验证了预测误差调制机制的有效性，交互初期（0~15 step），高时序差分误差通过自适应学习率驱动策略快速脱离随机空间。其中硬约束与地形感知智能体的余弦相似度下降幅度最为显著，回溯交互日志可发现，硬约束智能体在 8 step 附近首次触发能量临界约束，产生幅值最大的时序差分误差，导致该时刻语义梯度显著增大；地形感知智能体则在 11 step 附近首次进入连续沼泽区域，累积的高能耗惩罚信号触发了深度语义反思，推动潜向量发生较大幅度的定向更新。上述现象表明语义梯度在初期优先响应与生存约束直接相关的负反馈事件。后期（35~50 step）随着误差收敛，系统触发学习率退火进入稳态；唯有熵探索智能体受基于计数的内在动机驱动维持高频震荡，这种持续的可塑性有效防止了局部最优，证实了架构在稳定性与适应性间的动态平衡。所有智能体的余弦相似度曲线在演化后期（35 step 之

后）均趋于稳定收敛，未出现持续下降或发散的趋势，这从实证角度验证了潜向量在长期迭代中并未偏离冻结主干的原始语义分布，式（14）中的三重机制将策略演化有效约束在与预训练空间相容的有界范围内。

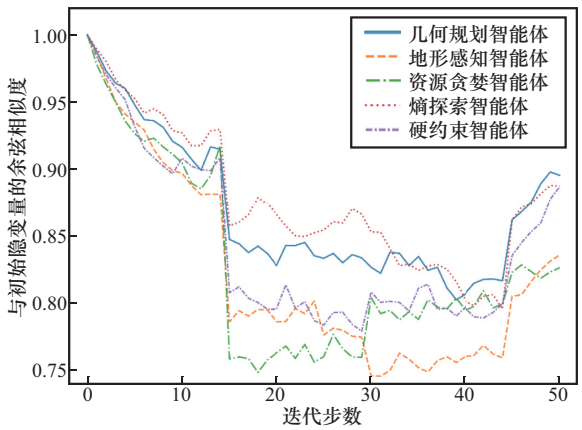


图3　潜空间向量余弦相似度变化

潜向量的单步L2语义漂移如图4所示。观察到3个显著的脉冲式峰值（约15、30和45 step），回溯交互日志可发现，这些时刻分别对应不同类型的关键负反馈事件：15 step峰值对应智能体首次进入连续沼泽区域，单步能耗显著超出预期；30 step峰值对应系统连续触发安全边界约束，硬约束监察智能体产生高幅值惩罚信号；45 step峰值则对应能量进入低阈值区间后的路径规划策略强制切换。上述事件均导致语义回路在单步内生成显著的语义梯度，驱动潜向量发生较大幅度的定向更新，体现了双回路框架对负反馈事件的即时自适应调节能力。

潜向量的PCA投影分布如图5所示，不同角色的智能体在潜空间中自动分化出清晰可辨的聚类簇。几何规划智能体的向量依然紧密聚集在代表直线趋向的语义区域，而地形感知智能体则占据了代表避险与绕行的区域。这种线性可分的流形结构证明双回路框架成功地将预定义的认知角色转化为大模型内部可解释的参数化表征，而非简单的提示词拼接。

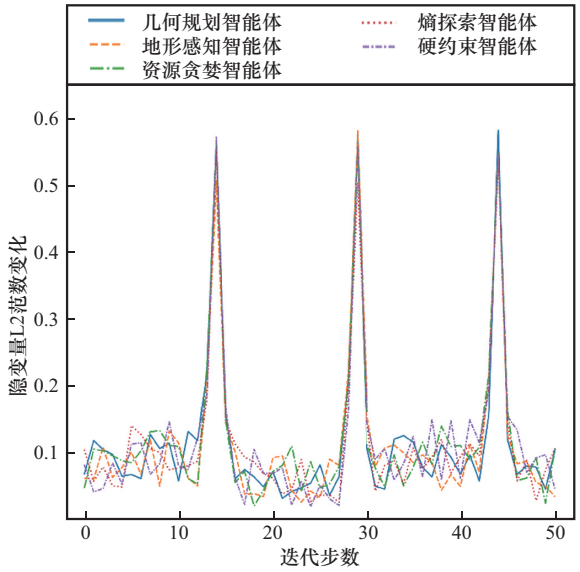


图4　潜向量的单步 L2 语义漂移

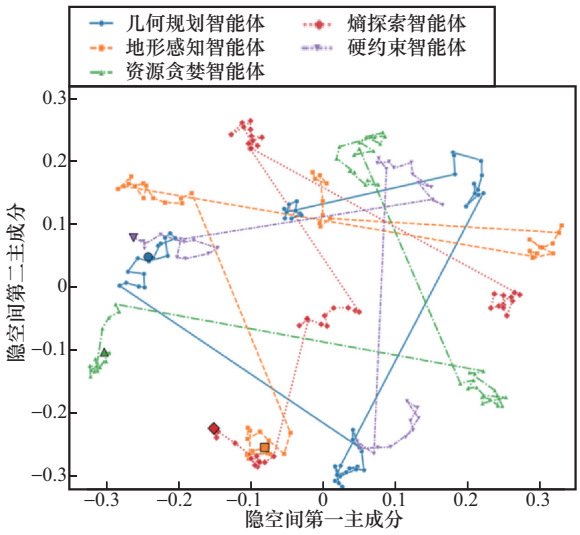


图5　潜向量的 PCA 投影分布

3.5.3　涌现的社会动力学与隐式因果推理

为了揭示元控制器如何协调多目标冲突，分析了各功能智能体的建议采纳率与信任权重变化。

每个智能体的总采纳次数如图6所示，揭示了系统涌现出的适应性生存策略。地形感知智能体的采纳次数（约43次）超越了几何规划智能体（约42次），这一反直觉的优先级反转证实元控制器习得了隐式因果法则：即主动绕行低阻尼区域以换取长程生存，而非盲目追求欧氏距离最短。

同时，硬约束监察智能体的适度介入（约13次）验证了其作为物理边界的有效性，将策略演化始终限制在可行域之内。

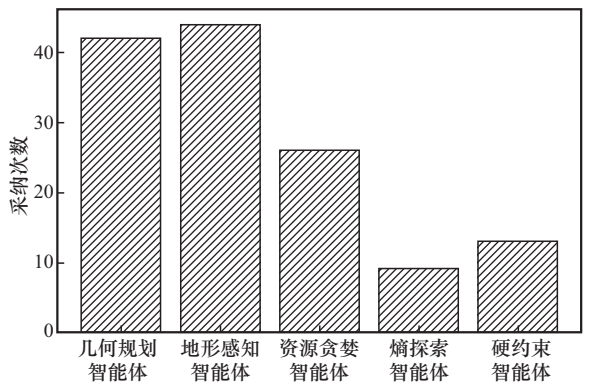


图6　每个智能体的总采纳次数

元控制器信任权重随能量的变化如图7所示，展示了元控制器涌现出的隐式因果推理能力。在能量充足阶段，系统高度信任几何规划算子以追求路径效率；而随着能量衰减，系统表现出前瞻性的风险厌恶，地形感知算子的信任权重在低能耗区间迅速攀升并实现反超。这种动态的优先级切换证实了双回路框架能够在无显式规则约束下，自主习得在效率与生存之间进行权衡的高阶策略，实现多目标冲突的动态平衡。

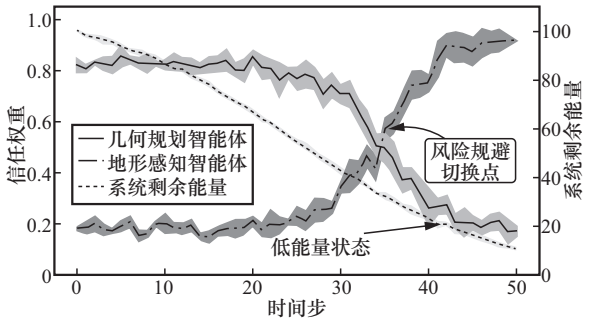


图7　元控制器信任权重随能量的变化

3.5.4　潜向量维度消融实验

为验证潜向量维度d的选取对系统性能的影响，并阐明d=1 024的理论依据，本文在4种典型维度配置下进行了消融实验。实验分别采用BGE-large-zh-v1.5（d=768）、BGE-M3（d=1 024）、text-embedding-ada-002（d=1 536）以及Qwen3-Embedding-4B（d=2 560）作为语义回路的编码器，其余实验配置保持不变。其中text-embedding-ada-002通过OpenAI API调用，其延迟受网络传输影响，不纳入延迟对比分析，仅作为高维度区间的性能参考。不同潜向量维度配置下的性能对比见表2。

实验结果表现出边际收益递减趋势。d=768时系统成功率仅为79.3%，SPL为0.65，表明该维度不足以完整编码多智能体博弈场景中的复杂策略信息。d从768增至1 024时，SR提升了6.3个百分点，SPL从0.65跃升至0.75，是性能提升的关键拐点。而d进一步增至1 536和2 560时，SR仅分别提升0.5%和0.8%，SPL不再增长甚至在2 560维时略有下降，本地推理延迟从2.1 s增至6.1 s，计算开销显著攀升。综合考量性能与计算效率，d=1 024在语义表达能力与推理开销之间取得了最优平衡，因此本文最终选定d=1 024作为潜向量的标准维度配置。

3.5.5　等计算资源下的收敛效率对比

为量化本文框架在相同资源约束下的收敛效率，本节在单张NVIDIA A100节点、物理训练时长统一为90 min的条件下，对本文方案与3类基线方法进行了系统对比，同等计算资源下各方法任务成功率收敛曲线如图8所示。

表2　不同潜向量维度配置下的性能对比

维度	编码模型	任务成功率（SR）	路径加权成功率（SPL）	推理延迟（s·step^{-1}）
768	BGE-large-zh-v1.5	79.3	0.65	1.6
1 024	BGE-M3	85.6	0.75	2.1
1 536	text-embedding-ada-002	86.1	0.75	-
2 560	Qwen3-Embedding-4B	87.4	0.76	6.1

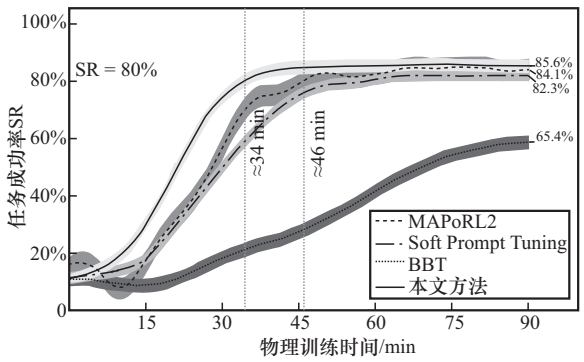


图8　同等计算资源下各方法任务成功率收敛曲线

本文方法每步仅需更新外置潜向量，计算路径轻量，约第 34 min 率先越过 SR=80% 阈值，此后曲线平稳上升并保持领先，终值达到 85.6%。MAPoRL2 因多智能体联合参数更新引发的持续策略震荡，有效进展被反复拉回，约第 46 min 才首次突破 SR=80%，终值为 84.1%。Soft Prompt Tuning 依赖白盒梯度，终值为 82.3%，在黑盒部署场景中应用受限。BBT 全程依赖随机扰动搜索，在 90 min 内未能突破 SR=80% 阈值，终值仅为 65.4%。上述结果表明，在同等计算资源条件下，外置潜向量更新机制有助于降低优化开销，并在收敛速度和最终任务成功率方面表现出一定优势。

4　结束语

本文提出基于冻结主干的双回路进化框架，通过解耦通用推理与策略偏好，利用语义反思驱动的潜向量更新机制，实现了离散价值迭代与连续语义梯度的深度融合。实验结果显示，该框架能够支持策略收敛，并观察到与隐式因果线索相关的协作行为。这一发现表明，将因果推理信号引入强化回路，能有效引导智能体超越局部贪婪，向着更具鲁棒性的自组织通用智能演进。

作为一种低成本、抗遗忘且可解释的新范式，本文证实了外置进化路径的潜力。当前实验主要在离散网格博弈环境下进行，该框架的跨场景泛化能力尚待进一步验证。从理论层面分析，双回路框架对具体任务环境的依赖主要体现在奖励函数的设计与异构算子的功能划分上，而潜向量演化机制本身与任务类型无关，具备迁移至其他场景的理论基础。展望未来，将致力于验证该框架在非结构化开放世界、多模态感知及真实物理环境等场景下的泛化边界，并深入探究策略基因在社会网络中的传播与变异规律。期待这种对硬件资源友好的技术路线，能为构建高效可持续的通用人工智能系统提供坚实的理论支撑。

参考文献：

[1] Xi Z H, Chen W X, Guo X, et al. The rise and potential of large language model based agents: a survey[J]. Science China Information Sciences, 2025, 68(2): 121101.

[2] Wang L, Ma C, Feng X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[3] Chen X. A survey on large language model based autonomous agents[J]. Computational Linguistics, 2024(2): 141-150.

[4] Guo T, Chen X, Wang Y, et al. Large language model based multi-agents: a survey of progress and challenges[C]//Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24). San Francisco: Morgan Kaufmann, 2024: 8048-8057.

[5] 李文斌, 熊亚锟, 范祉辰, 等. 持续学习的研究进展与趋势[J]. 计算机研究与发展, 2024, 61(6): 1476-1496.
Li W B, Xiong Y K, Fan Z C, et al. Advances and trends of continual learning[J]. Journal of Computer Research and Development, 2024, 61(6): 1476-1496.

[6] Bosma M, Chi E, Ichter B, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the Advances in Neural Information Processing Systems 35. Red Hook: Curran Associates, 2022: 24824-24837.

[7] Yao S, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models[C]//Proceedings of the Eleventh International Conference on Learning Representations. Amherst, Massachusetts: OpenReview, 2023.

[8] Wang G, Xie Y, Jiang Y, et al. Voyager: an open-ended embodied agent with large language models[J]. Transactions on Machine Learning Research, 2024.

[9] Park J S, O' Brien J, Cai C J, et al. Generative agents: interac-

tive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 2023: 1-22.

[10] Cassano F, Gopinath A, Narasimhan K, et al. Reflexion: language agents with verbal reinforcement learning[C]//Proceedings of the Advances in Neural Information Processing Systems 36. Red Hook: Curran Associates, 2023: 8634-8652.

[11] Sumers T, Yao S, Narasimhan K R, et al. Cognitive architectures for language agents[J]. Transactions on Machine Learning Research, 2023.

[12] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173.

[13] Chen B D, Chen R J, Liu S W, et al. Found in the middle: how language models use long contexts better via plug-and-play positional encoding[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Red Hook: Curran Associates, 2024: 60755-60775.

[14] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL Press, 2021: 3045-3059.

[15] Liu X, Zheng Y N, Du Z X, et al. GPT understands, too[J]. AI Open, 2024, 5: 208-215.

[16] Luo Y, Yang Z, Meng F D, et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 3776-3786.

[17] Huang J H, Cui L Y, Wang A T, et al. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL Press, 2024: 1416-1428.

[18] Li H Y, Ding L, Fang M, et al. Revisiting catastrophic forgetting in large language model tuning[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg, PA: ACL Press, 2024: 4297-4308.

[19] Sun T, Shao Y, Qian H, et al. Black-box tuning for language-model-as-a-service[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR Press, 2022: 20841-20855.

[20] Hayes C F, Rădulescu R, Bargiacchi E, et al. A practical guide to multi-objective reinforcement learning and planning[J]. Autonomous Agents and Multi-Agent Systems, 2022, 36: 26.

[21] Song Y F, Yin D, Yue X, et al. Trial and error: exploration-based trajectory optimization of LLM agents[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL Press, 2024: 7584-7600.

[22] Xu Z, Gu W, Yu C, et al. Learning strategic language agents in the werewolf game with iterative latent space policy optimization[C]//Proceedings of the 42nd International Conference on Machine Learning. New York: PMLR Press, 2025.

[23] 王若男, 董琦. 基于学习机制的多智能体强化学习综述[J]. 工程科学学报, 2024, 46(7): 1251-1268.
Wang R N, Dong Q. Multiagent game decision-making method based on the learning mechanism[J]. Chinese Journal of Engineering, 2024, 46(7): 1251-1268.

[24] Lei M, Cai H, Cui Z, et al. RoboMemory: a brain-inspired multi-memory agentic framework for lifelong learning in physical embodied systems[C]//Proceedings of the NeurIPS 2025 Workshop on Space in Vision, Language, and Embodied AI. Red Hook: Curran Associates, 2025.

[25] Park C, Han S, Guo X Z, et al. MAPoRL: multi-agent post-co-training for collaborative large language models with reinforcement learning[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL Press, 2025: 30215-30248.

[26] Zhou A, Yan K, Shlapentokh-Rothman M, et al. Language agent tree search unifies reasoning, acting, and planning in language models[C]//Proceedings of the 41st International Conference on Machine Learning. New York: PMLR Press, 2024: 1-23.

[27] Liu X, Yu H, Zhang H, et al. AgentBench: evaluating LLMs as agents[C]//Proceedings of the Twelfth International Conference on Learning Representations. Amherst, Massachusetts: OpenReview, 2024.

[28] Anderson P, Chang A, Chaplot D S, et al. On evaluation of embodied navigation agents[PP]. V1. arXiv (2018-07-18) [2026-06-10]. arXiv: 1807.06757.

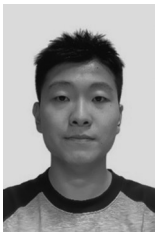
[29] Wu Y, Tang X, Mitchell T, et al. SmartPlay: abenchmark for LLMs as intelligent agents[C]//Proceedings of the 12th International Conference on Learning Representations. Amherst, Massachusetts: OpenReview, 2024.

[30] Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: ACL Press, 2019: 55-65.

[31] Hong S, Zhuge M, Chen J, et al. MetaGPT: meta programming for a multi-agent collaborative framework[C]//Proceedings of the 12th International Conference on Learning Representations. Amherst, Massachusetts: OpenReview, 2024.

[32] Ghanem B, Hammoud H, Itani H, et al. CAMEL: communicative agents for "mind" exploration of large language model society[C]//Proceedings of the Advances in Neural Information Processing Systems 36. Red Hook: Curran Associates, 2023: 51991-52008.

[作者简介]



赵袁赫（2002– ），男，华北电力大学控制与计算机工程学院博士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

王昱（1998– ），男，华北电力大学控制与计算机工程学院博士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

李昊忱（1996– ），男，交通运输部天津水运工程科学研究院工程师，主要研究方向为计量测试。

杨士铎（1998– ），男，华北电力大学控制与计算机工程学院博士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

郭柯楠（2000– ），男，华北电力大学控制与计算机工程学院硕士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

高晨（2003– ），女，华北电力大学控制与计算机工程学院硕士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

白小高（2003– ），男，华北电力大学控制与计算机工程学院硕士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

管孜然（2002– ），女，华北电力大学控制与计算机工程学院硕士生，主要研究方向为电力信息安全、人工智能安全、多智能体协同。

王继业（1965– ），男，博士，华北电力大学控制与计算机工程学院教授、博士生导师、正高级工程师，主要研究方向为电力信息安全、人工智能安全、多智能体协同。