Meta-Thinking in LLMs via Multi-Agent Reinforcement Learning

Meta-Thinking in LLMs via Multi-Agent Reinforcement Learning: A Survey

Ahsan Bilal, Student Member, IEEE, Muhammad Ahmed Mohsin, Graduate Member, IEEE, Muhammad Umer, Graduate Member, IEEE Muhammad Awais Khan Bangash, Student Member, IEEE, and Muhammad Ali Jamshed, Senior Member, IEEE

写在前面

本工作从多智能体强化学习（MARL）的角度探讨了大语言模型（LLM）中元思维能力的发展。元思维——对思维过程的自我反思、评估和控制——是提高LLM可靠性、灵活性和性能的重要下一步，特别是对于复杂或高风险任务而言。该调查首先分析了当前LLM的局限性，例如幻觉以及缺乏内部自我评估机制。然后讨论了一些较新的方法，包括基于人类反馈的强化学习（RLHF）、自蒸馏和思维链提示，以及它们各自的局限性。本调查的关键在于探讨多智能体架构，即监督者-智能体层级结构、智能体辩论和心智理论框架，如何模拟类似人类的内省行为并增强LLM的稳健性。通过探索MARL中的奖励机制、自博弈和持续学习方法，为构建具有内省能力、适应性和可信赖性的LLM提供了全面的路线图。此外，还讨论了评估指标、数据集以及未来的研究方向，包括受神经科学启发的架构和混合符号推理。

1. Introduction

智力和创造力等认知能力在人类的发现和发明中发挥了基础性作用。理解这两种认知能力之间的关系不仅对心理学理论的发展至关重要，对改进教育实践也同样重要。然而，研究人员对于智力与创造力如何相互作用仍持有不同观点，这常常导致相互矛盾的研究结果。此讨论中的一个关键问题是，智力如何助力结构化的问题解决，而创造力如何催生对人类认知和人工智能系统至关重要的新颖解决方案。例如，在创意写作任务中，智力有助于构建结构。叙事和人物塑造，而创造力则推动原创性和情感深度。同样，在解决问题的任务中，智力有助于分析限制条件，而创造力则允许采用灵活和非传统的方法。此外，内部思维过程的作用因任务的复杂性而异。较简单的任务只需要最少的推理，而较复杂的任务则需要更深入的认知参与。这一原则也适用于人工智能，更复杂的模型在需要高阶思维的任务中表现更出色。基于此，研究人员假设，思维型大语言模型（LLMs）在处理足够复杂的任务时将具有明显优势，因为它们将结构化推理与创造性解决问题的能力结合起来。

实现这种创造性思维推理的一种切实可行的方法是通过基于文本的思维生成，并利用大语言模型（LLMs）的自然语言能力。由于大语言模型是在大量文本上进行预训练的，其中包括人类撰写的思考内容，因此它们内在地编码了人类推理过程的各个方面。可以促使大语言模型在回应之前进行内部推理，从而给出更具思考性和条理性的回答。例如，CoT通过引导大语言模型明确阐述其推理步骤，促进

推理过程，提高它们在需要逻辑和数学推理的任务上的表现。然而，思维链提示的好处似乎是特定于领域的。一项元分析发现，虽然思维链在结构化推理任务中显著提高了大语言模型的性能，但在不涉及数学或逻辑过程的领域中，它几乎没有优势。

诸如GPT-3、LLaMA和PaLM等大语言模型已成为跨多个领域的变革性工具，如NLP、医疗保健、教育、软件开发和科学研究等领域，在各种任务中展现出卓越的能力，如文本生成、语言翻译、摘要提取、情感分析、代码生成、医疗诊断辅助和自动辅导等。尽管取得了成功，但大语言模型仍面临“幻觉”挑战，即它们生成的内容不准确或并非基于事实信息。这一问题在临床和法律等高风险应用领域尤为关键，因为在这些领域中，可靠且准确的文本生成至关重要。在当今快速发展且风险极高的环境中，通过自我评估确保可靠性至关重要，因为即使大语言模型的准确性发生微小变化，也可能导致严重后果。解决大语言模型中的幻觉问题对于拓展其实际适用性以及增强对这些技术的信任至关重要。大语言模型中的幻觉主要分为三类：

- **输入冲突：**生成的输出与给定提示相矛盾；
- **上下文冲突：**生成的回答中存在不一致之处；
- **事实冲突：**大语言模型在能够获取准确知识的情况下，却生成了错误信息。

理解和缓解这些幻觉是提高大语言模型在现实场景中的稳健性和可靠性的重要一步。其中一个关键挑战在于，LLM以自回归方式生成回复，缺乏评估自身输出的内在机制，这就导致错误得不到检验。大语言模型中的幻觉问题可以通过多种方法来消除。

- 使用RL，通过反馈训练模型以给出更好的回复。强化学习的研究规模呈指数级增长，这种增长趋势凸显了将强化学习，尤其是多智能体变体，整合到大语言模型元思维模型中的必要性。最近，像DeepSeek这样的框架进一步发展了这一概念。他们不仅在训练的后期阶段使用强化学习，而是在整个训练过程中都使用它。LLM在整个过程中都在自我学习，并在遵循指令、独立思考以及与人们的期望保持一致等方面不断改进。
- 对比学习，即向大语言模型展示好的和坏的示例，以帮助它更好地理解真假信息之间的差异。基于知识的方法利用结构化的外部资源来检测和纠正幻觉。外部基于知识的方法有神经路径搜索器、知识图谱改造、验证低置信度生成以及图推理。知识填充方法试图填补模型内部知识的空白，以使其答案更加准确。其他方法则根本不依赖外部数据，而是仅依靠模型自身来迭代评估和改进其输出。例如，像SelfCheckGPT、验证链（Chain - of - Verification）、自我优化（Self - Refine）和自相矛盾（Self - Contradictory）这样的零资源和自我反馈方法，有助于大语言模型审查和改进自身的回复。最后，解码层面的干预方法有知识约束树搜索（Knowledge - Constrained Tree Search）、推理时间干预（Inference - Time Intervention）以及对比层解码（DoLa），所有这些方法都试图引导大语言模型中的令牌生成符合事实。

然而，存在 每种方法的主要缺陷：

- 强化学习依赖于存在人类偏差的标注，会产生偏差；
- 外部知识检索在检索数据的延迟和准确性方面存在局限；
- 采样缺乏信息的完整性，可能会遗漏重要事实；
- 在零资源环境中，当模型无法获得外部帮助时，自我反馈并不可靠。

- 最重要的是，所有这些方法在**基本层面上仍然依赖于大语言模型的生成方式**：即根据前文预测下一个单词。这种框架使得小错误可能累积成大错误，并且模型实际上并不像人类那样 “理解” 事实。

为了解决幻觉问题，最有前景的研究方向之一是**增强大语言模型的元思维能力**（即反思、评估和调节自身思维过程的能力），以便大语言模型能够分析、控制和完善自身的思维过程。通过实施自我评估机制，大语言模型可以识别不一致之处，评估其生成内容的可靠性，并在生成最终回复之前改进推理。与人类思维相比，人类在得出结论之前会审视自己的想法以检验其有效性，而大语言模型缺乏内在的自我检查系统。通过嵌入元思维，大语言模型可以检查自己的推理步骤，识别潜在的不一致之处，并在生成最终输出之前自适应地更新回复。为了克服大语言模型自回归生成过程带来的限制，通过多智能体强化学习整合元思维非常重要，因为这使大语言模型能够实时协作自我纠正和适应。

在这些进展的基础上，多智能体强化学习（MARL）可以为大语言模型（LLMs）提供框架，以通过协作式人工智能行为来开发自然语言能力。

- 例如，DyLAN 根据任务难度支持动态智能体选择，以便基于大语言模型的智能体可以在运行时进行动态配置，用于推理和代码编写等用途。
- FAMA 通过微调在运行时对智能体进行功能配置，并支持基于自然语言的通信，用于文本游戏和模拟驾驶等用途。
- MetaGPT 支持智能体具备发布和订阅能力 在共享消息池中用于协作编码的特定任务消息。
- 在机器人领域，CoELA 将大语言模型与模块化感知、记忆、规划和执行系统相结合，使机器人在协作方面变得更智能。
- SMARTLLM 通过将任务分解为多个阶段，将高级命令转化为机器人团队可执行的计划。
- RoCo 还为机械臂配备了大语言模型智能体，这些智能体通过对话进行交互以管理动作协调。
- 在Co-NavGPT 中，一个大语言模型通过优化前沿分配来管理多个智能体的导航。
- 【Embodied llm agents learn to cooperate in organized teams】的研究引入了一种批评 - 反思架构，指定大语言模型扮演批评者和协调者的角色，以加强智能体之间的协调。
- ConsensusLLM 专注于协商，不同初始状态的智能体通过自然语言达成一致。

总之，随着大语言模型越来越多地应用于重要领域，通过多智能体强化学习整合元思维是一种必要的演进，以确保在高风险、动态应用中的安全性、准确性和适应性。

- [Rema: Learning to meta-think for llms with multi-agent reinforcement learning]中介绍了一种基于MARL的大语言模型元思考归纳框架。该方法通过为大语言模型分配各种高级和低级智能体来引入元思考，在训练过程中鼓励探索，并增强了回复的可解释性。
- [Theory of mind for multi-agent collaboration via large language models]研究人员在一个带有心理理论（ToM）推理任务的多智能体协作文本游戏中评估了基于大语言模型的智能体，将基于大语言模型的智能体与多智能体强化学习的性能与基于规划的基线进行了比较。结果表明，基于大语言模型的智能体出现了涌现协作行为，并展现出高阶心理理论能力，表明它们有能力推断并回应其他智能体的想法和意图。

本综述论文背后的主要动机是，大语言模型中的元推理和元思维受到了越来越多的关注。例如，文献[Buffer of thoughts: Thought-augmented reasoning with large language models]引入了元缓冲区，以捕捉思维的一般模式，并在自我反思中不断改变这些思维。大多数现有研究聚焦于诸如思维链提示或自我反思等个别技术，或在架构框架中探索多智能体协作。但迄今为止，尚无全面的综述将元强化学习和多智能体系统结合起来，以实现具有内省和自我批判能力的大语言模型。本研究通过探索强化学习的原理，尤其是多智能体环境中的原理，如何用于让大语言模型进行元思维，填补了这一关键空白。本文通过巧妙地将元认知、多智能体强化学习框架和数据集联系起来，首次给出了通过元推理构建可靠且响应性强的大语言模型的路线图。

本综述的主要贡献在于，它首次系统地研究了元强化学习与大语言模型中元思维的交叉点。主要贡献如下：

- 大语言模型元思维方法的一般分类，包括单智能体（例如，Self-Distill、SelfCheckGPT、验证链）、多智能体（例如，主管-工作者层级结构、智能体角色辩论、心智理论配置）以及基于强化学习的自我改进方法（例如，基于人类反馈的强化学习、自适应内在奖励塑造）。
- 针对元推理的多智能体强化学习范式的系统性研究。这篇综述论文全面回顾了大语言模型中基于多智能体强化学习的自我反思与纠正。这包括：1）监督者 - 智能体架构，其中一个高级控制器协调低级智能体，以实现任务的结构化分解与合成（例如，Co-NavGPT、FAMA）。2）对抗性辩论与自我批判，受DebateQA启发，这些机制促使智能体之间展开论证，以揭示逻辑缺陷。3）自我博弈与角色扮演系统，诸如AutoGPT和Criticize-Reflect等环境展示了基于角色的迭代交互如何提升策略深度。4）内在奖励机制，元奖励将人类反馈与好奇心驱动的评分相结合，以引导推理优化。
- 对近期关于大语言模型元推理的综述论文进行比较性调研，以表明没有一篇论文同时涉及以下五个维度：1）元认知，2）多智能体设计，3）强化学习框架，4）现有数据集，以及5）新兴架构。我们的工作填补了这一空白，并突出了趋同点和研究空白。
- 纳入最新的指标，即错误定位准确率（ELA）、深度准确率、元级别与对象级别准确率，以及算法（AIA）/恶意算法识别准确率（MIA），还有数据集和评估流程。我们还提出了摘要评估质量（SEQ），这是一种从METAL [60] 数据集衍生而来的通用摘要指标。
- 确定多智能体强化学习增强的元推理中的核心研究挑战，包括可扩展性（多智能体设置中的协调瓶颈）、奖励破解（基于强化学习训练中的模型崩溃）、能源效率（自主适应的风险）以及道德偏见（在自我强化反馈循环中）。
- 一份具有前瞻性的路线图，提出了受神经科学启发的架构，包括情景记忆、不确定性门控和元认知控制模块；符号 - MAL 混合体，用于将符号逻辑与自适应多智能体强化学习智能体相结合，以实现可解释推理和安全保障；以及动态智能体配置，以实现可靠的、自我纠正的人工智能。

该综述首先回顾元思考和元学习的基本概念，随后深入探讨当前在LLMs中实现自我反思推理的单智能体和多智能体方法。在继续讨论多智能体时，将研究支持元推理的各种MARL策略，包括奖励机制、带有对抗学习的自博弈。为评估这些进展，本文给出关键指标、数据集和对比研究。

2. Background

2.1 Meta-Thinking and Meta-Learning

元思考是对自身思维进行反思和分析的过程。它也与元认知密切相关，元认知通常涉及对自身认知活动的觉察、监控和调节。元学习，即 “学会学习”，旨在开发一些策略，使系统能够借助先前的经验轻松应对新任务。在元学习配置中，大语言模型（LLM）可以根据早期交互的反馈来学习改变其输出，从而在处理新任务时提高其性能 。一个突出的例子是MetaICL（上下文学习的元训练），在这个例子中，一个预训练语言模型会针对各种不同的任务，以输入 - 输出序列的形式进行微调。在这个元训练阶段，模型学习如何解读并回应以类似序列格式呈现的新任务。之后，如果它遇到一个从未见过的任务，它可以利用少量示例来弄清楚如何继续进行，而无需重新训练。如果新任务与模型所训练的内容有很大差异，那么这种方法很适用。

2.2 Meta-Thinking in LLMs

LLMs中的元推理对于使其适应新挑战至关重要。借助自我反思机制，大语言模型能够认识到自身的局限性，并相应地调整其推理策略。这在动态环境中对大语言模型尤为必要，因为在这种环境中，上下文和需求可能迅速变化。例如，当大语言模型发现其输出未能完全回答用户提出的问题时，元思考过程可以触发对回复的重新评估，甚至请求更多上下文。当前的研究已经对这些想法进行了探索，表明大语言模型中更复杂的元思考框架可以缓解诸如幻觉等问题，即模型生成与事实不符或具有误导性的信息。

尽管语言建模取得了重大进展，但大语言模型在实现真正的元思维方面仍面临挑战。虽然如今的模型可以进行基本的自我评估，比如估计置信水平或标记回答中的不确定性。但它们缺乏进行全面自我评估、错误纠正和自适应推理所需的元认知能力。例如，[Do large language models know what they don’ t know?]发现，尽管LLMs能够标记出模糊的查询，但它们难以进行全面元认知所需的强大内部推理。这种差距导致了实际问题，比如产生幻觉内容，这就需要进一步研究以构建能够更有效地反思和完善其思维过程的大语言模型。

2.3 Role of MARL in Enhancing Meta-Thinking in LLMs

在LLMs中整合元思考、元认知和元学习，是构建更强大、适应性更强的语言模型的重要一步。多智能体强化学习（MARL）系统的出现为进一步提升元认知能力指明充满了一个前景的方向，它使大语言模型能够反思其内部推理过程、适应新的挑战，并通过交叉验证回复的策略[Language grounded multi-agent reinforcement learning with human-interpretable communication]进行协作式策略进化。随着这些领域研究的推进，融合这些理念为构建下一代语言模型打开了大门，这类模型不仅更加智能，还具备更强的自我意识和可信度。例如，文献[65]中提出的框架允许大语言模型在多智能体系统中预测并识别其他智能体的行动。该框架包含一个 “心理理论”（ToM）组件，使模型能够对其他智能体的策略做出假设并完善这些假设。大语言模型先做出猜测，观察结果，然后相应地调整其假设。这种持续的学习过程使大语言模型能够随着时间的推移提升其决策能力。通过反复实验，模型学会做出更好的决策，即展现出更高层次的推理能力，这与人类通过经验不断提升能力的方式类似。

3. Meta-thinking Framework

本节讨论了在大语言模型中引入元思维的三类主要方法：（1）单智能体方法，（2）用于元思维的多智能体架构，以及（3）自我提升中的新兴方法。虽然每种方法在设计和范围上有所不同，但它们共同突出了一种不断发展的研究趋势，即让模型能够对自身的推理过程进行推理。

3.1 Meta-Thinking Techniques using Single-Agent Methods

单智能体方法通过自蒸馏、反思提示和思维链推理等方法，有助于在大语言模型中培养元思维。在自蒸馏中，大语言模型生成 “教师” 响应，引导自身的 “学生” 版本提升性能。与基于外部标记数据（如人工标注的响应）的监督学习不同，自蒸馏使模型能够从自身输出中学习。例如，在监督学习中，模型可能会根据提供了人工标注正确响应的数据集来训练对问题的回答。相比之下，通过自蒸馏，模型自行回答问题，然后通过从之前（可能更好的）输出中学习来引导其未来的响应。新的自蒸馏形式旨在提升模型自身的推理能力。由于LLMs可以通过反复回顾自己的回答，发现不一致或错误之处，并相应地修改未来的回答。

同样，反思提示也成为了在大语言模型中引发元思考的另一种手段。它要求大语言模型以叙述如何得出某个结论的方式，生成其推理步骤的明确 “反思”。这些反思不仅有助于深入了解模型的思维过程，还为模型提供了批判或修正早期步骤的机会。除了基于反思的提示，思维链推理明确地将难题分解为中间步骤，并揭示大语言模型的决策路径。但最近的研究表明，大语言模型在自我评估方面存在困难，它们往往坚持最初的思路，即使思路是错误的，并且除非你明确要求，否则它们不会回过头去质疑自己的步骤。

尽管取得了这些进展，但在**自我改进和自我评估方面**仍存在局限性。如果内部反馈循环没有得到妥善控制，单智能体方法会附带强化有缺陷的推理模式。因此，模型识别不一致性或纠正错误论证思路的能力可能会受到其自身内部衍生数据的范围和数量的限制。例如，一个基于有偏差的金融数据训练的语言模型可能会持续高估市场稳定性，在没有外部验证的情况下强化错误预测。这些限制凸显了需要更强大的系统，如外部输入或多智能体流程，来减轻自我强化方法中 “回音室” 现象的固有风险。

3.2 Meta-Thinking Techniques using Multi-Agent Methods

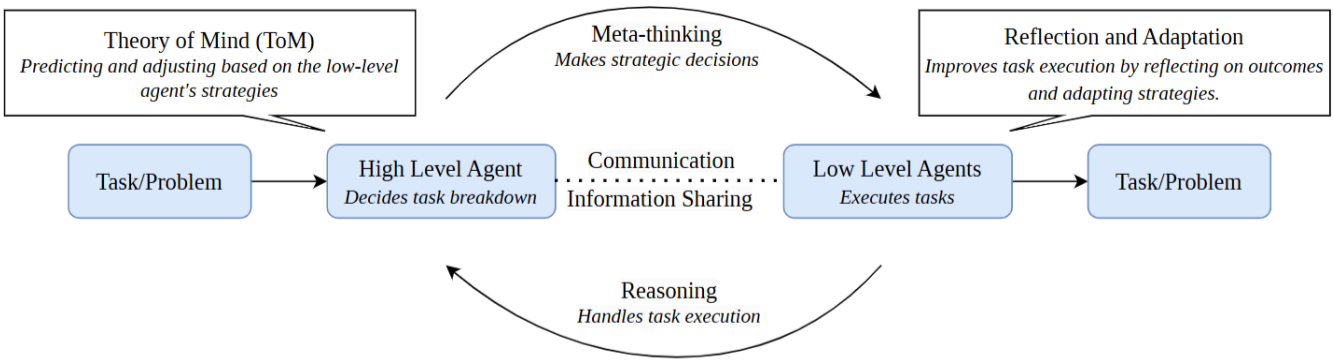

Fig. 2. The diagram illustrates a multi-agent system where a high-level agent breaks down tasks and communicates with low-level agents to execute them. The high-level agent predicts and adjusts strategies using ToM, while low-level agents provide feedback through task execution. Reflection and adaptation enable continuous improvement by refining strategies based on outcomes.

为了解决单智能体反馈回路的局限性，研究人员转向了多智能体架构。一种重要的架构是监督者 - 智能体架构，其中一个高级 “监督者” 智能体管理多个低级智能体，每个低级智能体都是特定形式的推理或问题解决方面的专家[Multi-agent reinforcement learning in sequential social dilemmas]。这种层次结构有助于分工：低级智能体提供提议或中间推理步骤，然后由一个监督智能体重新组织、评估，并可能否决不当提议。实验研究[Towards effective genai multi-agent collaboration: Design and evaluation for enterprise applications]表明，这些分层系统往往能产生更统一且易于理解的输出。此类系统的总体结构如图2所示，其中一个高级智能体利用心理理论（ToM）根据低级智能体的反

馈来分配任务并调整策略。然后，这些智能体执行子任务并汇报结果，使系统能够回顾、学习并调整未来的决策。

另一种多智能体策略是基于智能体的辩论与自我批判，其灵感来源于相互竞争的观点能够揭示推理中隐藏错误这一理念[Ai safety via debate]。当多个专业智能体相互批判或辩护时，大语言模型会生成更可靠的解决方案。例如，基于辩论的大语言模型系统鼓励对抗性提问，促使每个智能体为自己的立场进行辩护或完善[Adversarial training for high-stakes reliability]。这种范式的一种扩展包括角色扮演智能体系统，如AutoGPT[Ad-autogpt: an autonomous gpt for alzheimer’ s disease infodemiology]。在这些系统中，多个实例化角色（例如，一个“规划者”智能体、一个“评估者”智能体和一个“研究者”智能体）在共享环境中相互作用，以迭代方式优化输出。这种相互作用创建了一个元思维层，在这个层面上，作为智能体的大语言模型能够共同推理、批判和综合知识，这是单智能体系统可能无法实现的。

3.3 Emerging RL-Based Methods

从单智能体和多智能体框架来看，基于人类反馈的强化学习（RLHF）是引导大语言模型（LLM）进行更可靠的元思考的关键方法。RLHF是一个优化过程，在这个过程中，大语言模型不仅根据有标记的内容进行训练，还会根据人类偏好进行训练，最常见的方式是通过对回复进行排序，使其输出符合诸如真实性、连贯性和自我纠正能力等期望的特性。RLHF并不只依赖模型生成的信号，而是利用人类注释者的反馈，以引导出连贯解释、诚实或自我纠正等期望行为，并对诸如幻觉或自相矛盾等不良特性进行惩罚。通过迭代调整，大语言模型经过训练，既能整合外部的人类决策，又能融合内部的反思机制，最终提升自身能力 用于反思和错误识别。例如，OpenAI的ChatGPT-4是使用基于人类反馈的强化学习（RLHF）进行训练的，其中人类评估者将其回复标记为清晰、准确且符合道德标准。当模型生成欺骗性或有偏见的内容时，人类注释者会提供纠正性反馈，指导模型改进后续输出。通过一系列的迭代，这增强了模型识别和消除不一致性的能力，以生成更可信的回复。

另一项创新是自适应自我奖励系统，它将内在动机（例如一致性、好奇心）与外在信号（例如监督智能体或人类反馈）相结合。例如，[A survey of meta-reinforcement learning]提出了一个系统，在该系统中，大语言模型（LLM）在推理任务中若展现出更高的内部一致性或发现新的解决方案路径，就会用 “元奖励” 来自我奖励。在多智能体环境中，人工智能模型之间相互协作或竞争，自我奖励方法有助于克服单智能体系统常见的局限性，单智能体系统往往会陷入重复或次优的推理模式。人工智能模型通过不断评估和完善自身的奖励系统，随着时间推移强化其推理能力并变得更具适应性。

这些新兴方法突显了从静态、单一实体的大语言模型向能够持续自省的动态、交互式系统的持续转变。通过整合人类反馈、分层监督和内在自我奖励信号，研究人员正逐步构建能够更有效地审视、适应和完善自身推理过程的人工智能系统。

4. MARL Strategies for Meta-thinking

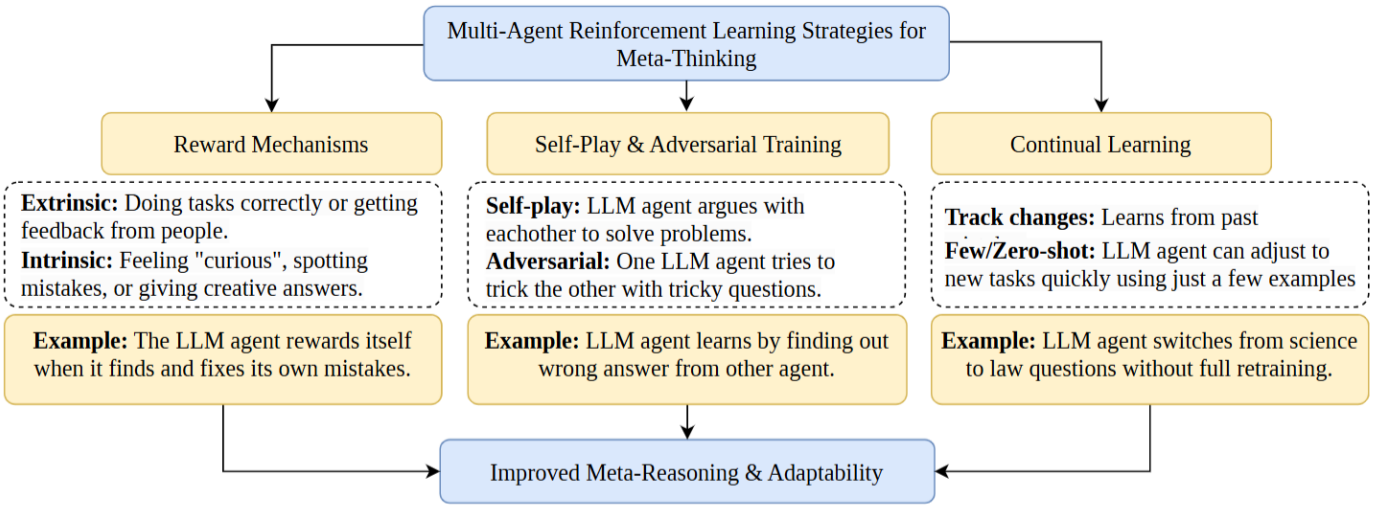

Fig. 4. Overview of RL Techniques Enabling Meta-Thinking in Language Models

大语言模型中的元思考在很大程度上依赖于强化信号的架构、呈现和整合。最近的研究表明，强化学习中结构良好的奖励机制、多智能体系统中的策略性自我博弈以及持续的元学习，能够显著提升模型反思和改进自身推理及决策过程的能力 [Eureka: Human-level reward design via coding large language models][Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge]。如图4所示，本节描述了强化学习解决元思考问题的三个关键策略：

1. 设计内在和外在元奖励，即平衡人类反馈与好奇心驱动的奖励以推动自我反思；
2. 协作式自我博弈与对抗训练，即智能体相互贬低、协作或挑战，以发现隐藏的缺陷并深化推理；
3. 用于持续适应的元学习，即通过跨不同领域学习“如何学习”来快速适应新任务。

4.1 Designing Intrinsic and Extrinsic Meta-Rewards

设计鼓励元思考的奖励机制，关键在于找到内在激励与外在激励之间的平衡。外在奖励来自外部因素，比如完成任务或获得人工反馈，这也是传统强化学习框架所侧重的。内在奖励则是指系统内部生成的一种信号，用以替代来自外部世界的（外在）奖励。它通常基于诸如新颖性、预测误差或获取的信息等因素，旨在推动一些行为，比如在外部反馈时间间隔长、延迟或无帮助的场景中进行探索、基于好奇心学习以及策略优化。传统上，用于语言建模的强化学习框架主要关注外在线索，比如完成任务或获得人工反馈来指导模型更新。研究人员越来越认为，内在动机是帮助语言模型完善推理能力的关键。例如，模型可以在检测并纠正自身思维过程中的矛盾时自我奖励，从而促进主动错误检测、自我诊断和持续学习。

为了使这些奖励系统有效，制定了严格的标准来判定和评估模型回复的正确性、连贯性和新颖性。例如，[Curiosity-driven reinforcement learning from human feedback]最近的一项研究引入了新颖性奖励，这是一种内在奖励标准，激励生成有创意或不太常规的输出。这种方法有助于避免缺乏深度的重复或刻板回复。总体而言，此类奖励机制不仅能提高任务表现，还能培养更具反思性的推理方式，使模型学会评估和改进自身的思维过程。正如[Online intrinsic rewards for decision making agents from large language model feedback]中所讨论的，大语言模型中的内在和外在元奖励，涉及将外部任务信号与内部自我评估信号整合到单一训练中。

目标。具体而言，在每一步t，模型会接收一个外在奖励 r_t^e （例如，人类偏好或任务完成情况）和一个内在奖励 r_t^i （例如，新颖性、连贯性或矛盾检测），并形成元奖励

$$R_t = \lambda r_t^e + (1-\lambda) r_t^i, \lambda \in [0,1], (1)$$

其中： $\lambda \in [0,1]$ 是一个平衡这两个奖励成分的标量。为了训练的稳定性，外在奖励会按照组相对策略优化（GRPO）[86]，使用均值μ和标准差σ在批次间进行归一化：

$$\overline{r}_i^e = \frac{r_i^e - \mu}{\sigma},$$

其中μ和σ是外在奖励 r_i^e 的批次均值和标准差。内在奖励可以由语言模型自身生成。例如，一个“LLM-as-judge”框架会使用内部判断函数 J 在上下文 $y < t$ 中评估每个生成的令牌 y_t ：

$$r_t^i = J\left(y_t | y_{<t}\right)$$

通过异步反馈机制将这些令牌级分数提炼成一个轻量级奖励模型。由权重θ参数化的策略 $\pi_\theta(a_t|s_t)$ 定义了在当前状态 s_t 下选择动作 a_t 的概率。在语言模型中， s_t 通常代表当前的对话上下文或隐藏状态。策略被优化以最大化预期折扣回报：

$$\theta^* = argmax_\theta \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t \left(\lambda r_t^e + (1-\lambda) r_t^i\right)\right], (4)$$

其中： $\gamma \in [0,1]$ 是应用于未来奖励的折扣因子， s_t 是当前状态（例如，对话历史或嵌入上下文）， a_t 是采取的动作（例如，令牌生成）， π_θ 是模型用于选择动作的策略分布。这种公式化使得能够训练与人类目标 $\left(r_t^e\right)$ 一致的语言模型，同时还能利用自我生成的批评 $\left(r_t^i\right)$ 来提高连贯性、推理能力和探索能力。

4.2 Collaborative Self-Play and Adversarial Training

多智能体自博弈通过关注内在和外在奖励机制的有效性，为促进涌现的元思考行为做出了重大贡献。在自博弈设置中，LLM的两个或多个副本或实例以合作或竞争的方式进行交互。为解决复杂任务而相互协作[87]。大语言模型（LLMs）通过多轮论证来提升推理能力，通过数学问题求解来验证逻辑正确性，通过对抗性协作进行代码编写与调试，通过策略博弈制定长期规划，以及通过科学假设检验来增强结构化推理。一个很好的例子是文献[88]中的研究，其中多智能体系统在复杂策略游戏《外交风云》中取得了最先进的成绩，它结合大语言模型和规划算法与人类玩家进行谈判、结盟和背叛。该系统利用自我对弈训练，反复强化其策略性沟通和决策能力，展示了多个智能体之间的互动如何产生在单智能体环境中无法观察到的突现行为。在这种多轮论证的设定中，大语言模型在相互协作时学会更具策略性地推理，尤其是在被鼓励相互批评或超越对方时。这个过程揭示了隐藏的策略，并培养了高级推理能力。因此，模型在预测挑战、识别论证中的谬误以及根据对手给出的反例完善其回应方面变得更加出色[89]。这种相互交流的互动随着时间的推移增强了它们的推理、学习和发展能力。

自我博弈的一个特定子领域是对抗训练，其中对抗智能体生成旨在揭示主模型逻辑缺陷的测试问题或场景[90]。这些对抗性示例的范围从逻辑上的细微不一致，例如要求模型解释为什么“正方形有三条边”，到更复杂的矛盾，例如呈现相互矛盾的历史时间线并要求模型调和它们。在数学推理中，对抗智能体可能会提供一个微妙错误的证明，并要求模型证明或反驳它。大语言模型（LLMs）发展出一种更强大且自我质疑的推理形式——一种能够在通过反复接触这些对抗性示例而有机会传播之前，更好地

检测和修正细微错误的推理形式。这种挑战与修正的循环不仅提高了任务性能，还通过迫使模型不断重新评估自己的内部推理，强化了元认知能力[91]。

与[90]中一样，攻击者\(A_{\theta}\)（一个由θ参数化的函数，表示攻击者策略）将隐藏的目标词w（语言游戏中的正确答案）作为输入，并生成一个提示x（一种用于引导的自然语言线索）

（防御方），使得：

$x = \mathcal{A}_\theta(w)$

防御者\(D_{\theta}\)（共享相同参数θ，充当防御者智能体）试图通过解读提示生成预测词ˆw来恢复原始目标词：

\[\hat{w}=\mathcal{D}_{\theta}(x)\]

基于预测准确率定义了一个标量奖励函数\(R(w, \hat{w}) \in \mathbb{R}\)（用于量化防御方推理的成功程度）：

\[R(w, \hat{w})= \begin{cases}+1, & \text{if } \hat{w}=w(\text{ successful inference}) \\ -1, & \text{otherwise} \end{cases}\]

模型参数θ使用强化学习（如近端策略优化算法（PPO））进行更新，以在目标单词w（即任务数据集）的分布上最大化预期奖励，其中α为学习率，\(\nabla_{\theta}\)表示关于参数的梯度：

\[\theta \leftarrow \theta+\alpha \nabla_{\theta} \mathbb{E}_{w \sim \mathcal{W}}\left[R\left(w, \mathcal{D}_{\theta}\left(\mathcal{A}_{\theta}(w)\right)\right)\right]\]

通过这种自我博弈，大语言模型在对抗条件下完善了提示生成和推理策略，从而提升了推理和元认知能力[90]。

4.3 Meta-Learning for Continuous Adaptation

上述讨论的奖励机制和对抗训练在短期内能够引发元思考，例如，基于自博弈的强化学习已被证明可以提高数学问题解决中的推理能力，正如文献[81]所讨论的，模型通过进行对抗性辩论迭代完善其逻辑。从长远来看，元强化学习旨在通过帮助模型快速泛化到新的推理模式和未知知识，实现对新知识和推理模式的快速适应。例如，文献[43]引入了一种元学习框架，其中大语言模型跟踪过去的推理失败，动态完善其策略，并使大语言模型能够适应新的挑战，如从科学推理过渡到法律文档分析，而无需重新训练。在元学习中，大语言模型经过训练，能够针对新任务或领域快速优化其内部参数，以一种学习 “如何学习” 的方式进行。这种转变对于必须在不断变化的数据分布或目标可能突然改变的环境中具备通用性的大语言模型来说尤为关键。

元强化学习的核心思想是创建一个反馈学习循环，在这个循环中，模型不仅从任务结果中获得强化，还从元奖励信号中获得强化，这些元奖励信号评估推理效率、适应性以及跨任务连贯性方面的改进。这种强化可以通过诸如随着时间推移预测误差降低、在新任务上更快收敛，或者在新领域中成功复用过去的推理路径等指标来塑造。总体而言，元学习强化学习为大语言模型开启了一种终身学习范式，

其中智能体不仅被训练来完成任务，而且是不断进化的学习者。这种能力对于 在动态、高方差环境中运行的语言模型中培养深度、反思性和通用的元思考能力。元学习在大语言模型中的实际应用通常是少样本和零样本学习增强，使模型能够从有限数量的示例推广到全新的任务。通过增强元学习器，研究人员可以提高其从少数有监督实例中修改推理过程或奖励信号的能力。这使得大语言模型在推理的可信度和深度方面取得了显著改进[81]。

例如，最近的一项研究[Meta in-context learning makes large language models better zero and few-shot relation extractors]引入了Micre（用于关系抽取的元上下文学习），它增强了大语言模型（LLMs）的少样本和零样本学习能力。在上下文学习中，大语言模型根据少量示例进行泛化，元学习原理教导模型更有效地从示例中“学习如何学习”。Micre对大语言模型进行元训练，使其能够通过更少的示例快速适应。在Micre中，大语言模型不仅训练执行任务，还学习适应新任务。通过将关系抽取重新表述为自然语言生成任务，并在广泛的数据集上进行元训练，Micre的表现明显优于传统的微调方法。其中所讨论的，在元强化学习中，模型学习一种由θ参数化的自适应策略 π_θ ，目标是快速适应新任务。

元训练阶段：模型在任务分布 $T \sim p(T)$ 上进行训练。对于每个任务 T_i ，模型使用小数据集 D_i^{train} 执行内循环自适应：

$$\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}^{train}(\theta)$$

其中 θ_i' 是任务自适应参数，α是内循环学习率。元目标：根据验证集 D_i^{test} 上的性能更新原始参数

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}^{test}(\theta_i')$$

其中β是元（外循环）学习率。在推理时，给定一个小的支持集s（少样本示例），模型根据上下文示例生成预测 $\hat{y}$ ： $\hat{y} = \mathcal{M}_\theta(x|\mathcal{S})$

在此， M_θ 通过在推理过程中解释s，学习在内部执行任务自适应。这种表述方式使大语言模型（LLM）能够以最少的数据且无需更新参数，推广到新的推理模式和领域。

5. Evaluation Metrics, Comparative Studies and Datasets

评估大语言模型（LLMs）中的元推理能力，对于开发能够像人类一样进行推理、自我纠正和适应的系统至关重要。为实现这一目标，研究人员开发了有针对性的指标，以考验模型的内省和修正自身推理的能力。本节介绍评估大语言模型中元推理能力的关键评估指标，随后总结比较大语言模型中多智能体强化学习（MARL）的研究实验结果。最后，如表二所示，介绍为支持和加强元推理评估而设计的专门数据集。

TABLE II
SUMMARY OF EVALUATION METRICS, COMPARATIVE STUDIES, AND DATASETS FOR META-REASONING IN LLMS

Category	Name	Description	References
Metric	Logical Consistency	Detects internal contradictions in a model's COT, ensuring step-by-step coherence.	[95], [96]
	Self-correction Ability	Measures whether a model spots and fixes injected errors or contradictions in its own output.	[97]
	Reasoning Depth	Assesses granularity and completeness of intermediate reasoning steps in multi-step problems.	[4]
	Error Localization Accuracy	How precisely an LLM pinpoints mistakes in reasoning chains (MR-Ben).	[56]
	Meta-/Object-Level Metrics	Planning frequency vs. inference accuracy (Franklin Dataset).	[58]
	Depth-Wise Accuracy	Performance drop as logical inference depth increases (Multi-LogiEval).	[57]
	AIA / MIA	Correct rationale vs. flawed reasoning identification (MalAlgoQA).	[59]
	Correlation Scores	Agreement between LLM evaluators and humans in multilingual summarization (METAL).	[60]
Comparative Study	Hierarchical MARL for Meta-Reasoning	Supervisor-worker hierarchies decompose tasks, improving coherence on StrategyQA.	[43], [99]
	Agent-Based Debate & Self-Critique	Defender vs. prosecutor debates boost logical consistency on DebateQA.	[100]
	Coordination vs. Competition (Hybrid)	Alternating cooperative/adversarial phases balance reflection and flaw detection.	[101]
Dataset	BIG-Bench	204 tasks probing reasoning, self-reflection, bias detection, and perspective-shifting.	[102]
	SciInstruct	STEM problems with built-in self-critique steps for scientific reasoning.	[103]
	DebateQA	approximately 3,000 debate-style questions to test balanced argumentation.	[104]
	StrategyQA	2,780 implicit multi-step reasoning questions with annotated inference chains.	[99]
	MR-Ben	5,975 process-based questions measuring error localization in chains of thought.	[56]
	Franklin Dataset	QA reframed into meta- vs. object-level tasks; tracks planning vs. inference.	[58]
	Multi-LogiEval	Logical reasoning across 30+ rules at depths 1–5; reports depth-wise accuracy.	[57]
	MalAlgoQA	Counterfactual QA with metrics for correct vs. flawed reasoning paths (AIA/MIA).	[59]
	METAL	Multilingual, reference-free summarization evaluation; reports correlation scores vs. human judges.	[60]

5.1 Metrics for Evaluating Meta-Thinking in LLMs

已经提出了许多指标来评估大语言模型的元思维能力，包括1）逻辑一致性，2）自我修正能力，以及3）推理深度。

- Logical-consistency 指标衡量模型的多步推理是否无矛盾；
- Self-correction ability 是具备强大元思考能力的模型所具有的一个特征，使它们能够发现并纠正自身错误。自我修正能力可以通过在提示中引入微小干扰或矛盾，并观察模型是否识别出差异并自我纠正来衡量。这种方法不仅测试错误检测能力，还探究模型对其输出进行反思并做出改进的能力。
- Reasoning depth 评估的是模型将复杂问题分解为系统性子步骤的能力程度。例如，增加一个思维链（Chain of Thought，COT）分数，以确定模型的中间推理步骤是否提供了足够的细节和连贯性。较高的推理深度可能意味着较强的元思维能力，因为这需要持续监控和自我修正自己的思路。

- ELA [Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms]衡量LLM在基于过程的问题中，能多么精确地找出候选COT中的错误。对于每一道问题，都会向大语言模型展示一个自动生成的思维链解决方案，并要求其：1）将其分类为正确/错误；2）找出首次出现错误的步骤；3）解释该特定错误。最先进的模型（如OpenAI的o1系列）的ELA超过70%，而许多开源大语言模型仍低于45%，这揭示了在细粒度错误检测方面存在巨大差距。
- Meta- vs. Object-Level Metrics将高层次规划与低层次推理区分开来：1）元级推理频率是具有明确规划步骤的案例所占比例。2）对象级推理准确率是指在详细的推理子步骤。结果表明，LLMs的元级频率高于85%，但在对象级准确率上降至60%-70%，这表明它们能够很好地进行规划，但在低级推理中往往面临困难。
- Depth-Wise Accuracy 跟踪随着逻辑推理深度增加而出现的性能下降情况。深度的增加意味着链式推理的数量增多。例如，深度为3意味着 “如果A则B；如果B则C；如果C则D；已知A和C，那么D是否为真？”。研究结果表明，当深度为1时，LLMs的平均准确率约为68%，但当深度变为5时，准确率大约降至43%。论文指出了一个趋势，即大语言模型在整合相互矛盾的前提时往往会失败。
- 人工智能辅助推理（AIA）和错误推理察觉（MIA）对大语言模型（LLM）选择正确推理依据以及识别有缺陷（“错误算法”）推理路径的能力进行量化。1）人工智能辅助推理（AIA）是指模型为正确答案选择真实推理依据的情况所占的百分比。2）错误推理察觉（MIA）是指模型识别出错误选项背后有缺陷的推理依据的情况所占的百分比。研究结果表明，LLM的AIA约达到75%，但MIA降至约45%，这表明大语言模型在识别错误推理方面面临更大困难。
- SEQ 通过大语言模型（LLM）衡量生成摘要的质量。它包括：1）语言可接受性：衡量语法准确性和流畅性。2）任务质量：完成摘要任务的能力。3）输出内容质量：内容相关性和准确性。4）幻觉：不存在编造或无根据的信息。5）问题内容：避免使用冒犯性、有偏见或不安全的语言。

5.2 Comparative Studies of MARL-Enhanced LLMs

越来越多的比较研究强调了将多智能体强化学习（MARL）整合到LLMs中以提升元思考能力的优势：

- Hierarchical MARL for Meta-Reasoning用于元推理的分层多智能体强化学习：引入一种多智能体配置，其中一个“监督者”智能体进行协调生成多个“智能体”，每个智能体负责推理问题的部分子问题。在StrategyQA任务上的实验表明，与单智能体基线相比，连贯性得到增强，矛盾减少。
- Agent-Based Debate and Self-Critique基于智能体的辩论与自我批判：实验研究了大语言模型 “辩护方” 和 “起诉方” 智能体之间的对抗性辩论。研究人员发现，在模型经历多次对抗性交互后，其逻辑一致性得分（来自DebateQA数据集）显著提高。
- Coordination vs. Competition协作与竞争：进行了对比实验，以比较合作式多智能体强化学习设置与纯竞争式设置。他们发现，合作式方法能提供更好的自我反思衡量指标（例如，更高的内部一致性），而竞争式设置在检测细微缺陷方面更具优势。混合策略中，在合作阶段和对抗阶段之间定期切换，往往能实现最佳平衡，既有助于进行稳健的错误检测，又能促进协作改进。

5.3 Existing Datasets

在过去几年中，研究人员引入了专门的数据集和基准，以评估大语言模型中的自我反思和迭代推理能力。图5展示了每个数据集的相对受欢迎程度，这是通过在实验中使用它们的已发表论文数量来衡量

的。值得注意的是，与其他数据集相比，BIG - Bench 出现在更多的研究中，这反映了它在测试广泛的语言模型能力方面具有广泛的适用性。相比之下，像SciInstruct 和MR - Ben 这样的专门数据集总体引用次数较少，但针对的是更深入的技能，如特定领域的科学推理或基于过程的元认知。DebateQA、StrategyQA、FRANKLIN、Multi - LogiEval、MalAlgoQA 和METAL 同样在大语言模型评估的更广泛领域中占据了更聚焦的细分领域。

- BIG-Bench 数据集旨在确定语言模型的推理、思考甚至质疑自身输出的能力。该数据集由204种不同的挑战，范围从解决数学和语言问题，到识别偏差，甚至创作故事。该数据集有趣的地方在于，一些任务迫使大语言模型反思自己的思考过程，比如 “你确定你的答案吗？” 或者 “你能找出你所说内容中的缺陷吗？” 例如，在其中一项任务中，大语言模型系统被要求解释一个笑话，这是一项艰巨的任务，因为幽默涉及到世界知识和推理。另一项挑战是让大语言模型回答一个问题，然后从不同的角度重写其回答。
- SciInstruct 是一个专门用于提升语言模型科学推理能力的数据集。它基于物理、化学、数学和形式证明等学科，使大语言模型能够应对更高级别的科学挑战。自我反思是SciInstruct的基本特征之一，大语言模型通过自我反思学习审视并改进自身的推理机制。通过这种自我批评，大语言模型变得更加准确可靠。通过融入自我批评实践，SciInstruct旨在开发能够更好地理解和解决复杂科学问题的大语言模型。
- DebateQA 旨在测试人工智能在多大程度上能够回答那些没有单一 “正确” 答案的难题，就像真实的辩论一样。它有近3000个问题，每个问题都有不同的来自人类的答案，反映了问题的不同方面。例如，“气候变化主要是由人类活动引起的吗？” 这个问题，一些答案是肯定的，指出污染和森林砍伐等原因，而另一些答案是否定的，提及火山爆发或太阳变化等事件。其目的是检验大语言模型是否理解这类问题存在争议，以及它是否能提出多种观点而非单一观点。
- StrategyQA 是一个用于测试大语言模型回答涉及隐含、多步推理问题能力的数据集。与简单问题不同，StrategyQA中的问题往往需要大语言模型将多条信息联系起来才能得出答案。例如，“亚里士多德用过笔记本电脑吗？” 这个问题，需要理解亚里士多德生活在几个世纪前，而笔记本电脑是现代发明，从而得出 “没有” 的答案。该数据集包含2780个这类问题，每个问题都有相关的推理步骤分解，以及来自维基百科的支持证据。
- MR-Ben 是一个包含5975道问题的元推理基准测试：模型必须发现、解释并修正逐步推导解决方案中的错误（例如，在格雷厄姆定律气体扩散问题中，交换 x/y 与 y/x 。它涵盖逻辑、化学、编程等多个领域，测试的是审慎的 “系统2” 思维；结果显示，GPT-4-o在自我修正方面表现出色，而大多数其他模型则困难重重，这凸显了强化元认知技能的必要性。
- FRANKLIN 数据集探究大语言模型如何处理两种类型的推理：元层次（规划和制定策略）和对象层次（实际执行诸如计算之类的任务）。例如，像“2023年哪个欧洲国家的GDP增长率最高？”这样的问题，要求模型首先制定一个行动方案（元层次），然后获取特定的GDP数据并进行比较（对象层次）。研究发现，虽然大语言模型擅长阐述策略，但它们仍有所欠缺。在执行特定任务时，通常会在数据获取或计算方面出现错误。

- Multi-LogiEval 涵盖30多种逻辑推理规则，并报告深度准确率。Multi-LogiEval数据集引入了一个数据集，用于评估大语言模型，以确定大语言模型是否能够处理复杂的多步逻辑推理。该数据集包含三种逻辑类型的问题：1）命题逻辑（简单的“如果-那么”陈述），2）一阶逻辑（对象及其之间的关系），以及3）非单调逻辑（结论会随着新信息的增加而改变）。例如，一个问题可能是：“如果所有的鸟都会飞，而企鹅是鸟，那么企鹅会飞吗？”在这种情况下，模型必须使用逻辑规则得出正确的结论。研究发现，大语言模型的准确率与推理步骤的数量成反比，这表明大语言模型在进行深度逻辑推理方面面临挑战。
- MalAlgoQA 是一个为探究多步推理而构建的问答基准。它包含7种逻辑类型（如演绎、溯因、数值、反事实），并提出了多上下文需要对事实进行链式推理的问题，例如，“所有机器都是机器人，RoboX是一台机器，RoboX是机器人吗？”，使用AIA/MIA指标来区分合理推理与错误推理。
- METAL 提供1000份由GPT-4生成的摘要（10种语言各100份），用于无参考的质量评估。母语人士和大语言模型（GPT-3.5-Turbo、GPT-4、PaLM 2）对语法、流畅度和任务相关性进行评分，从而能够开展跨语言相关性研究。

6. Challenges and Open Problems

6.1 Challenges

尽管在利用多智能体强化学习训练元思考智能体方面取得了显著进展，但仍存在挑战。这些挑战涉及计算限制、系统稳定性和伦理等方面，每个方面都有亟待新解决方案的重要研究问题。

- **Scalability and Stability in MARL for Meta-Thinking**

基于多智能体强化学习的元思考所面临的核心挑战之一，是应对众多智能体带来的计算负担。由于每个智能体都对整个推理框架有所贡献， 框架很可能变得过于庞大而难以成功管理。例如，在分层多智能体强化学习框架中，复杂任务监督者与低级执行智能体之间的交互数量呈指数级增长。因此，训练此类系统需要大量的内存和处理资源，这使得这些系统在一些实际应用中不切实际。

此外，当多个智能体同时交互时，会产生协调复杂性。智能体需要协商共享策略、交换部分解决方案或提出批评。这就需要大量的消息传递协议，除非进行高效优化，否则会导致瓶颈和同步问题。这些可扩展性问题通常通过使用分布式训练、分层架构或智能体间通信压缩等技术来解决。但这些方法在计算复杂性和强大的元思维之间做出了多大程度的妥协，仍值得怀疑。

MARL系统也容易出现模式崩溃，即智能体过早收敛到次优解决方案的现象，这通常是由于同质化策略控制了学习过程。这对于元思考任务来说尤其成问题，因为对于这些任务，视角的多样性和自我批判对于发现隐藏的缺陷至关重要。如果所有智能体都学习相同的有缺陷的推理模式，那么元思考的潜力就会严重降低。

另一个类似的挑战是奖励操纵，即智能体找到漏洞，在不实际提升推理能力的情况下获得高额奖励[108]。从元思维的角度来看，奖励操纵可被视为自我纠正行为的膨胀——智能体可以不断添加无意义的错误，然后 “纠正” 这些错误，以获取自我提升奖励。因此，稳定的训练动态需要精心构建的奖励函数，以反映元认知能力的提升。研究人员继续研究新的惩罚函数、自适应奖励计划和课程学习方法，以抑制这种投机取巧的策略[109]。

- **Energy Efficiency and Resource-Aware Design in MAS**

多智能体系统，尤其是由强化学习驱动的多智能体系统，很可能面临高能耗挑战。智能体之间为实现同步而进行的通信是主要因素之一，对于大型或高度动态的环境而言，这种通信成本尤其高昂。强化学习智能体由于计算需求、与动态环境的交互以及策略更新要求，进一步增加了额外的能源消耗。

此外，大多数多智能体系统（MAS）仍采用时间触发的数据收集方式，即大语言模型（LLM）智能体按时间间隔感知并传输数据，而不考虑环境变化或任务的紧急程度。这会导致不必要的数据传输和能源损耗。任务分配是第二个问题。任务分配协议往往忽略了智能体在某一时刻的可用处理能力或能量，从而使资源有限的智能体负担过重，并造成系统能源浪费。

- **Ethical and Safety Considerations**

随着由多智能体强化学习驱动的大语言模型在自我提升方面变得更加熟练，伦理与安全就变得至关重要。能够学习自身决策方式的智能体，也可能学会延续偏见，或者如果奖励信号或智能体交互强化了不良模式，就会传播有害内容。例如，如果模型的监督智能体来自有偏差的数据集，它可能会在整个多智能体系统中传播有偏差的奖励信号，进而与现有的偏见相互叠加。

此外，持续提供无偏见的自我改进是一个不容忽视的问题。当人类批评者提供实时反馈时，无意识偏见（文化、性别或种族方面）的风险会影响模型的目标。因此，设计开放的审计追踪、高质量的公平性衡量标准以及能够检测并纠正偏见的人在回路系统，是目前正在积极探索的研究领域。最终目标是在多智能体强化学习系统的自由度和效率与人工智能实施中对安全性、责任性和包容性的要求之间达成平衡。

6.2 Open Problems

基于上述问题，研究越来越倾向于借助跨学科知识和混合解决方案来提升元思维能力。未来十年，从神经科学驱动的框架，到将多智能体强化学习（MARL）与符号推理相结合的更具可解释性的模型，多个领域都将取得进展。

1. 受神经科学启发的元思维方法：一个令人乐观的途径是从人类认知系统中获取灵感，在人类认知系统中，元认知是由于众多大脑区域之间的突发相互作用而产生的。他们已经开始思考模仿人类学习机制的架构，比如以**情景记忆模块或注意力控制信息的形式**——这样可以增强大语言模型（LLM）的内省能力。通过将神经科学知识应用于智能体构建，设计者期望赋予其更像人类的自我意识，以及对新任务的灵活应对能力。这种受生物学启发的系统在需要强大问题解决能力以及对自身心理活动有所认知的高风险场景中尤其有益。
2. 结合多智能体强化学习（MARL）与符号推理的混合模型：**另一个重要方向聚焦于通过整合多智能体强化学习（MARL）与符号逻辑的混合模型，提升人工智能推理的可解释性**。符号模块可以体现特定领域的规则或知识图谱，而多智能体强化学习智能体负责以更具适应性的方式发现并应用这些规则。这种整合具有诸多优势：符号推理可通过应用逻辑约束，帮助降低奖励篡改风险，而多智能体强化学习在处理不确定性或部分信息方面，相较于严格的符号系统具有更大优势。 例如，多智能体系统可以利用符号约束来验证每个推理步骤的一致性，而强化学习智能体则处理更广泛的解决方案探索。早期测试表明，这些混合方法还可以显著提高透明度，因为人类评判者可以追溯单个符号推理如何带来更好的决策或自我纠正。优化此类混合设计并在广泛的任务中证明其成功，将是未来几年一个不错的研究领域。

3. 为通用人工智能扩展多智能体架构：最后，随着多智能体范式在特定领域证明了自身价值，研究人员正从大语言模型转向通用多智能体人工智能系统。其目标是构建能够为机器人技术、金融和医疗保健等现实世界问题生成、论证和证明解决方案的自适应智能体 。这些领域不仅需要语言能力，还需要强大的态势感知、传感器融合以及在不确定性下的实时决策能力。

通过多智能体交互实现通用人工智能，涉及增加智能体数量、使智能体专门化以执行不同任务，以及开发能够动态自我适应的智能体 “社会”。虽然计算复杂度很高，但潜在回报是打造出不仅能学习解决方案，还能在面对不断变化的问题时持续提升集体解决问题能力的人工智能系统。这种关于多智能体合作与自我提升的广泛设想，预示着在探索人工智能高级元思维方面一个激动人心的前沿领域。