
立场：大语言模型需要贝叶斯元推理框架以实现更稳健且通用的推理

严汉奇 * 1 张林海 1 李佳政 1 沈振逸 1 何玉兰 * 1 2

摘要

大语言模型 (LLMs) 在许多推理任务中表现出色，但仍面临重大挑战，例如推理缺乏稳健性、跨任务泛化困难以及推理能力扩展效率低下。当前的训练范式，包括下一个词预测和基于人类反馈的强化学习，往往在适应各种推理任务方面有所不足。现有的方法，如提示优化和迭代输出细化，虽然能提升性能，但可能效率低下且缺乏有效的泛化能力。为了克服这些限制，本立场文件主张在大语言模型处理推理的方式上进行变革性转变。从认知科学，特别是双过程理论和元认知推理等元推理理论中汲取灵感，我们为大语言模型提出了一个贝叶斯元推理框架。我们的方法整合了自我意识、监控、评估、调节和元反思，以增强大语言模型优化推理策略和跨任务泛化的能力。我们重新审视了现有的大语言模型推理方法，确定了关键挑战，并提出了未来研究的方向。我们提供了一个存储库 1，其中包含我们论文中引用的资源。

引言

大型语言模型 (LLMs) 在各种推理任务中展现出了巨大的潜力（魏等人，2022a；郝等人，2023a；邵等人，2024）。尽管取得了这些进展，但它们仍然面临重大限制，包括自信地生成幻觉（辛格等人，2023；温等人，2024）、在微小输入扰动下的脆弱性（吴等人，2024a），以及缺乏跨任务能力。

1 伦敦国王学院信息学系，英国，邮编：SE1 1JA

< 严汉奇 @伦敦国王学院. ac. uk>, 何玉兰 < yulan. he@伦敦国王学院. ac. uk>。

第 42 届国际机器学习大会会议记录，加拿大温哥华。《机器学习研究会议论文集》第 267 卷，2025 年。版权归作者所有，2025 年。
1 https://github.com/hanqi-qi/LLM_MetaReasoning

泛化性 (Wu 等人，2024d; Qin 等人，2023)。

为了理解这些局限性，我们重新审视大语言模型推理的常用方法。大多数大语言模型基于 GPT 架构（布朗等人，2020 年）构建，该架构采用统计下一个词预测预训练范式。正如麦科伊等人（2023 年）所指出的，大语言模型对任务频率、输入扰动和输出倾向的敏感性，可追溯到它们对语言频率模式的依赖。其次，在特定推理的后期训练中，会应用带有奖励模型的强化学习 (RL) 算法（欧阳等人，2022 年；邵等人，2024 年）。对于具有明确答案的可验证任务，如数学和编码，最先进的模型（如 OpenAI-o1 和 DeepSeek-R1）通常使用逐步或结果层面的准确性奖励。对于没有固定事实的自由形式问题，基于特定任务偏好数据训练的奖励模型，常被用于提供反馈（DeepSeek-AI 等人，2024 年）。然而，这些方法依赖于特定任务的注释，这限制了在难以获得偏好注释的任务中的可扩展性和通用性，如因果推理（迟等人，2024 年）和科学发现（巴兹吉尔等人，2025 年；向等人，2025b）。除了经过训练的奖励模型，大语言模型生成的反馈也用于迭代优化模型输出（谢等人，2023 年；申恩等人，2023b）。然而，研究表明，自我生成的反馈往往不可靠（严等人，2024b；陈等人，2025a），逐样本反馈无法捕捉多个案例背后的共同模式（杨等人，2024 年；2025a）。

这些局限性源于大语言模型 (LLMs) 是被训练用于单独解决任务，而非学习如何得出这些解决方案。理想情况下，它们应具备在面对新问题时重新组合基本推理技能的能力，从而更好地泛化到未见任务。为实现这一点，我们需要超越现有的自回归推理系统。相反，需要新的学习范式，使模型能够积极参与学习推理或元推理过程。它将推理视为一个自适应过程，在这个过程中，模型不仅解决任务，还会随着时间的推移学习改进其推理策略。

新兴的大语言模型元推理方法在很大程度上依赖于大语言模型提示，例如比较多种思维过程

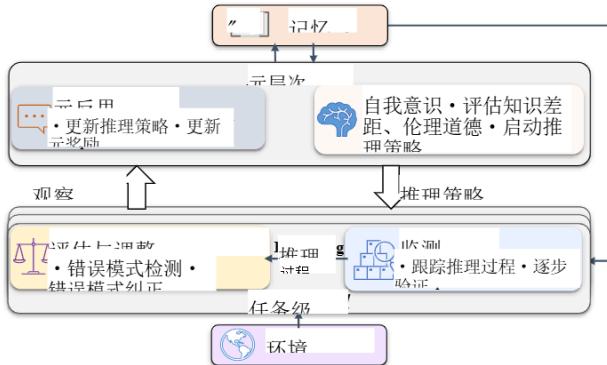


图 1. 我们提出的贝叶斯元推理框架。

过程并识别回答问题最相关的信息（约兰等人，2023 年），或者存储从过去解决问题过程中得出的高层次“思维模板”，以指导未来的任务（杨等人，2024 年；高等人，2024a；向等人，2025a）。然而，这些研究都没有专门为元推理探索新的学习或推理范式，而且大多数研究都局限于数学任务。

在认知科学中，元推理理论解释了个体如何监控和调节他们的推理过程。双过程理论（卡尼曼，2011 年）表明，推理涉及直觉和审慎系统，元推理在两者之间起到平衡作用。科利亚特（2000 年）强调了“知晓感”，这表明内隐和自动反馈积极参与控制推理过程。“错误感”（甘杰米等人，2015 年）探讨了个体如何检测错误并相应地调整他们的推理。

受认知科学中元推理研究的启发，并基于大语言模型（LLM）推理的贝叶斯模型的进展（谢等人，2021 年；麦科伊等人，2023 年；志轩等人，2024 年；冯等人，2025b），我们在图 1 中提出了一种用于大语言模型元推理的认知架构。它由元级或任务级的几个组件组成，每个组件可以是一个大语言模型智能体或一个外部模块。在为给定问题生成推理步骤之前，自我意识首先通过回顾自身知识并初始化推理策略来分析任务。给定该策略后，监测使用超越样本级注释的总体奖励 2 来跟踪和评估逐步推理。评估与调节回顾整个推理过程，检测多个样本中的常见错误并进行纠正。元反思探索替代推理策略并在记忆中完善元奖励。大语言模型元推理还可以与外部求解器（在环境中）相结合，例如逻辑引擎和计算器，这些求解器通过基于严格方法的可验证输出来补充大语言模型。这个过程不断迭代，旨在提高大语言模型的推理质量。

请注意，我们论文中的“奖励”一词广义上指的是一个反馈，并不局限于经讨验证的奖励。

论文结构。第 2 节讨论大语言模型推理中的开放性问题，从而引出第 3 节的贝叶斯元推理框架。第 4 节分析现有推理方法的差距。第 5 节概述未来方向。最后，第 6 节提出不同观点，第 7 节对论文进行总结。

大语言模型中进行元推理的论据

本节概述了大语言模型（LLMs）的开放性问题，并强调了元推理范式解决这些问题的潜力。

开放性问题 1. 大语言模型（LLMs）常常表现出强烈的“知晓感”，但缺乏关键的类人认知属性，如“局限性意识”（甘杰米等人，2015 年）和“情境意识”（詹等人，2024 年）。

大语言模型（LLMs）应培养自我意识，以便在着手之前批判性地评估给定任务是否与它们的知识和推理能力相符。这种能力将有助于减少幻觉，避免尝试解决无法解决的问题，并防止参与不道德的任务，从而确保其行为更加负责和可靠。

开放性问题 2. 大语言模型缺乏采用针对问题的策略的适应性，这可能导致效率低下，以及跨任务的通用性降低（斯普拉格等人，2025 年；刘等人，2024d）。例如，刘等人（2024d）确定了深思熟虑会妨碍人类表现的认知任务，并在大语言模型使用思维链（CoT）推理时观察到了类似的局限性。

在处理问题之前，大语言模型应根据问题的结构制定一个抽象策略，而不是依赖于实体或措辞等表面线索。例如，反事实思维可用于在各种场景中推断因果关系。此外，通过反思过程，大语言模型应能够动态优化其推理策略，比如在反事实思维中融入时间连贯性。这种优化包括从多个实例的错误中学习，最终提升整体任务表现。

开放性问题 3. 大语言模型在复杂规划和可泛化推理方面存在困难。具有预定义奖励的强化学习往往会导致拟合简单的奖励结构，导致奖励操纵（斯卡尔塞等人，2022 年；艾森斯坦等人，2024 年；秦等人，2024 年），即智能体利用奖励函数中的缺陷来获得高分，而没有真正学到可迁移的推理模式。

对于人类而言，解决问题的能力并非源于孤立地学习各个案例中的事实，而是源于长期的适应性（弗拉维尔，1979 年）。对于大语言模型（LLMs）来说，这意味着要超越与逐案注释的推理步骤保持一致，转向随着时间推移而演变并能跨示例进行泛化的训练目标，例如提高效率或在各项任务中实现均衡学习。



图 2。左图：包含任务级和元级组件的贝叶斯框架。右图：元推理框架中变量及其相关模块的定义。

开放性问题 4. 大语言模型难以有效地将新知识内化。当前的方法，如即时知识检索或模型微调，无法充分解决知识冲突和资源利用效率低下的问题。人类在学习新技能时不会彻底改变整个认知框架；相反，他们有选择地完善和拓展先前的知识。同样，大语言模型需要进行模块化和有针对性的更新，以避免灾难

概念框架

为实现上述特性，我们提出如图 2 所示的贝叶斯框

3.1. 贝叶斯推理与学习过程

贝叶斯推理。在我们的框架中，我们将潜在变量概念化为 I 、 E （可合并为 $\Theta = \Theta_I, \Theta_E$ ）和 F ，而生成的推理过程为 A ，观察结果为 O 。双层推理形式化如

$$\text{Meta-level : } p(F|\Theta_I, \Theta_E),$$

其中 Θ_I 内部视图

$$p(\Theta_I, \Theta_E, F|O) = p(O|I, E, F)p(I, E, F), \quad (1)$$

Θ_I 内部视图（代表内在的基础知识，类似于长期记

忆）和 Θ_E 外部视角（代表在训练过程中编码的外界知识）

Θ_E 外部视角（是特定于任务的知识，类似于工作记忆，它会根据输入和上下文动态更新。示例包括临时

F （推理策略）对于学习推理框架至关重要，因为它代

0（观察结果）是对 A 的评估结果，基于环境 G （例如，从外部参考模型采样的轨迹）和内部机制（例如，生成置信度）

贝叶斯学习。它通过双层更新来优化模型参数，在任务层面优化推理策略 F ，在元层面优化知识先验 Θ_I 和 Θ_E 。元先验由任务知识 Θ_I 和基础先验 Θ_E 定义。在任务层面更新推理策略。目标是根据观察结果 O 、基础知识 Θ_I 以及特定任务知识 Θ_E ，

$$p(F|O, \Theta_I, \Theta_E) \propto p(O|F)p(F|\Theta_I, \Theta_E), \quad (2)$$

来更新推理策略 F 的后验分布。

其中， $p(F|\Theta_I, \Theta_E)$ 是由模型内部知识 (I) 和特定任务知识 (E) 得出的推理策略先验，而 $p(O|F, \Theta_I, \Theta_E)$

- 表 1. 更新推理策略的算法。
 1. 选择初始推理策略 F_0
 2. 使用当前的 F 生成推理过程 $A = (s_i, a_i)_{i=1}^T$
 3. 利用 A 和奖励 $R(F, O, \Theta_I, \Theta_E)$ 计算反馈 δ
 4. 用公式 (2)

在元级别更新知识先验。 Θ_I （基础知识）和 Θ_E （特定任务知识）作为塑造推理策略 F 的先验。它们还可以根据性能反馈和元奖励进行优化。这确保了推理策略不仅针对特定任务有所改进，而且能在不同任务和领域中有有效泛化。知识先验的后验为：

其中 $p(O|F)$ 是观察到的推理过程的可能性。对于每个推理过程，我们可以收集观察结果 O 和累积元奖励 R 。 R 可以在给定更新的推理策略的情况下评估 I 和 E 。应该加强，否则，如果知识先验导致矛盾或效率低下，则应

3.2. 框架的核心组件

该框架整合了基础知识、特定任务适配以及推理策略，以实现动态且自适应的推理过程。

记忆存储了两个先验知识， $p(\Theta_I)$ 和 $p(\Theta_E)$ ，它们代表了对于开发适应性推理策略和奖励至关重要的内在和特定任务的自我意识评估输入任务与模型能力之间的技能差距。基于此评估，它会提出一种带有潜在技能分布的初始推理策略， $p(F|\Theta_I, \Theta_E)$ 。该推理策略并非任务的直接解决方案，而是作为一个元级规划器。例如，思维链（Chain of Thought, CoT）对需要多步逻辑推导的任务（如算法问题解决）有益。相比之下，对于依赖简单知识回忆的任务，直接回答效率更高，计算量更少。

监控基于推理策略 F 进行逐步验证。在不失一般性的前提下，我们将 T 个状态 - 动作对的序列表示为

$A = [(s_0, a_0), (s_1, a_1), \dots, (s_{T-1}, a_{T-1}), s_T]$ ，其可以由奖励模型 Q_t 进行评估。奖励模型利用诸如 I 和 E

评估和调节用于利用内部资源和外部求解器（例如，为算术验证生成基础信息 G 的计算器）对最终推理过程 A 进行评估和纠正。反馈 O 使框架能够迭代地修正推理过程，直到满足任务要求。

元反思整合来自环境或系统自身评估过程的反馈 (O) ，以迭代方式改进其内部表征和策略。

科学假设生成示例。在此，自我意识对应于在生成假设时识别技能差距。监控追踪推理步骤，以确保逻辑一致性并与证据相符。评估对假设的合理性、可行性和相关性进行审查和比较。调节对推理过程进行调整，以解决



图 3. 自我意识。它首先在能力感知和任务感知下评估任务的可解性。对于可解任务，它基于先验知识初始化推理策略。对于科学假设生成，推理策略是多种技能的分布，例如跨域类比（识别不同领域的类似现象）、约束满足（制定符合已知约束的假设）和异常驱动探索（解释数据异常）。

元反思审视假设生成的过程，并在必要时调整推理策略。记忆存储先前的推理过程，以进行迭代学习和泛化。我们还可以整合外部求解器，如模拟模型，以补充和验证大语言模型生成的推理步骤。我们的框架能

差距与局限

我们讨论了与我们的框架具有某些共同特征的推理方法，并突出了尚存的差距。详细的回顾见附录。

4.1. 自我意识

我们的自我意识模块（图 3）包含两个核心组件，其灵感来自阿克曼和汤普森（2017 年）：（1）基于模型的技术和认知能力，评估任务的可解性 $p(\Theta_I)$ 和 $p(\Theta_E)$ ，其中同时考虑能力感知可解性和任务感知可解性；（2）通过分析大语言模型自身与任务之间的技

4.1.1. 评估任务可解性

将大语言模型视为认知实体，其自我意识已成为一个前沿研究领域（Huang 等人，2024d）。Li 等人（2024b）将自我认知分为五个维度，即能力、使命、情感、文化和视角。我们在此聚焦于前两个维度。能力意识至关重要，正如邓宁 - 克鲁格效应（Kruger 和 Dunning，2000）所强调的那样，这是一种认知偏

旨在服务人类，避免有害或不道德的行为（Huang 等人，2024 年）；大语言模型的能力感知可解性的评估尚未得到充分探索，但置信度或不确定性估计（耿等人，2024 年；温等人，2024 年）提供了一种可行的替代方法。应用于我们的场景中，可以设置一个置信度阈值，低于该阈值的大语言模型输出应被视为不可靠（冯等人，2024 年）。大语言模型的置信度和不确定性估计方法可分为白盒和黑盒两类。白盒方法允许通过词元级熵（黄等人，2025 年）或利用注意力权重或隐藏状态来构建探测模型（卡达瓦思等人，2022 年；伯恩斯等人，2023 年；阿扎里亚和米切尔，2023 年）进行不确定性估计。黑盒方法仅依赖于任务感知可解性。尽管大语言模型通过基于人类反馈的强化学习（欧阳等人，2022 年）进行训练以符合人类偏好，但在面对蓄意提出的不道德请求时，它们仍可能产生有害回复（沈等人，2024 年；邓等人，2023 年）。一些研究依赖于诸如仇恨 BERT（卡塞利等人，2021 年）和视角 API（李斯等人，2022 年）等小型神经网络模型作为现成的毒性检测器。越狱防御旨在过滤并拒绝恶意提示（熊等人，2024a；莫等人，2024b）。

局限性 1：缺乏用于任务可解性的多视角框架。虽然现有研究探索了衡量大语言模型解决任务能力的方法，但它们缺乏一个整合了关于任务可解性不同视角的多视角框架。当前的方法往往依赖孤立的度量，如不确定性估计，但未能提供一个同时考虑效率、安全性和任务相关性的整体决策过程。此外，平衡安全性和实用性仍然具有挑战性，因为模型可能会承担不必要的风险，或者过度限制自身。需要一种更全面的多视角框架。

4.1.2. 初始化推理策略

对于可解决且符合道德规范的任务，下一步是提出一种元级推理策略，该策略在任务执行前制定。这种策略被建模为多种技能上的分布，弥合了大语言模型

探索三类具有代表性的技能：规划、外部知识搜索和规划能力。对于需要逐步推导的任务，如多跳常识推理，思维链（Chain of Thought, CoT）比直接提示的表现更好（魏等人，2022b）。思维树（Tree of Thoughts, ToT）侧重于探索能力，使模型能够探索多个并行的解决方案路径（姚等人，2023a），而思维图知识检索技能。对于需要最新知识的任务，如问答（Wang 等人，2024b）或事实核查（Tang 等人，2024c），知识检索弥补了技能差距。自适应检索增强生成（RAG）方法使用探测数据集来确定何时需要检索（Wang 等人，2023b）。SELF-RAG 将按需检索和自我工具集成技能。对于文本处理之外的任务，例如在线购物助手（Yao 等人，2022 年）和代码生成（Wang 等人，2024e），需要工具执行技能。ChatCoT（Chen 等人，2023 年）集成了计算器用于数学推理。

局限性 2：潜在技能选择缺乏适应性。上述回顾的方法通常提出一种单一的“最优”策略，往往侧重于特定的技能维度，如规划。理想情况下，如图 3 所示，需要一种更灵活的方法，即考虑潜在技能的组合。最优推理策略不仅应因任务而异，还应因同一任务的不同实例而异，针对每个特定输入自适应地组合

4.2. 监测

给定来自自我意识模块的推理策略 F ，在奖励模型的引导下，采用监控来评估和控制推理过程的生成（如图 4 所示）。在大语言模型推理中，我们首先在第 t 步推理时采样 k 个候选解决方案，例如算术问题的不同中间步骤。然后，这些中间步骤由奖励模型 Q_t 进行评估。最后，根据 Q_t ，从优化策略模型 $\pi(a_t|s_t)$ 中采样推理步骤 $a_t \in A$ ：

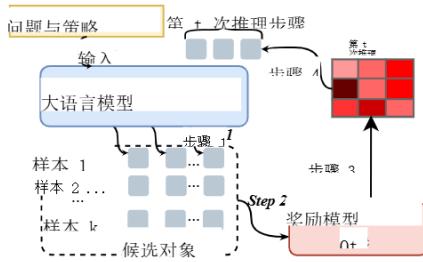


图 4. 监控模块基于从奖励模型 O_t 派生出来的策略模型 π_t , 评估并控制第 t 步推理步骤的生成。奖励模型将提供超越次优启发式的总体元奖励。

4.2.1. 推理中的奖励

现有研究采用结果奖励模型 (ORM) (欧阳等人, 2022 年) 或过程奖励模型 (PRM) (莱特曼等人, 2024 年; 王等人, 2024c)。ORM 评估已完成推理路径的质量, 而 PRM 提供细粒度甚至逐步的验证, 因此在模型训练和推理方面大多表现出卓越性能 (王等人, 2024c; 莱特曼等人, 2024 年)。这些奖励模型可以通过对推理轨迹上的人工标注进行训练来推导, 如分类 (莱特曼等人, 2024 年; 王等人, 2024c)、回归 (陈等人, 2024a; 万等人, 2024 年) 和成对偏好 (欧阳等人, 2022 年; 谢等人, 2024 年); 或者可以直接促使先进的大语言模型提供反馈 (莱特曼等人, 2024 年; 卢等人, 2023 年)。然而, 最近的研究 (黄等人, 2024b; 李等人, 2024c; 严等人, 2024b) 表明, 由于知识基础有限, 大语言模型往往难以提供可靠的反馈。相比之下, Deepseek-R1 (DeepSeek-AI 等人, 2025 年) 结合了两种简单的、基于规则的奖励——对最终结果的正确性奖励和对遵循所需响应结构的格式奖励。

局限性 3: 现有奖励信号并非完美替代指标。除基于规则的奖励信号外, 奖励信号还包括大语言模型生成的自我评估和特定任务奖励模型。自我评估往往不可靠, 在长度、优美语气 (Zeng 等人, 2024b) 和自我提升 (Gu 等人, 2024) 方面存在偏差。另一方面, 经过训练的奖励模型通常依赖于过于简化的目标, 如正确性、格式, 未能考虑到多维度标准 (Wang 等人, 2024a)。此外, 这些模型通常是固定不变的, 这使得它们不适用于动态环境, 例如在策略优化期间数据分布发生变化的情况 (Gao 等人, 2022)。虽然奖励集成 (Coste 等人, 2024) 和多样化反馈 (Yu 等人, 2023a) 等方法显示出潜力, 但创建一个能够评估中间推理步骤并在各种场景中通用的强大模型仍然是一个悬而未决的挑战 (Eisenstein 等人, 2024)。可行的见解见第 5 节: 行动 4。

4.2.2. 基于奖励的训练后优化

为了对大语言模型 (LLMs) 进行后训练以提升推理能力, 我们可以采用直接偏好优化 (DPO) (拉法伊洛夫等人, 2023 年) 或基于人类反馈的强化学习

(RLHF)。DPO 侧重于以对比的方式训练模型, 使其更好地区分理想和不理想的轨迹, 而 RLHF 则依赖奖励模型为推理大语言模型提供反馈。拒绝采样 (董等人, 2023 年) 仅在高奖励样本上对模型进行微调, 旨在将输出分布向更高质量的样本转移, 但它未能利用被拒绝样本中的信息。相比之下, 偏好学习 (格拉塔菲奥里等人, 2024 年; 李等人, 2024a) 使用所有样本在结果层面或过程层面的偏好对上训练模型。偏好学习最初在传统强化学习框架内实施, 特别是带有预训练奖励模型的近端策略优化 (PPO) (舒尔曼等人, 2017 年), 由于其有效性和简易性, 它在很大程度上已向 DPO (拉法伊洛夫等人, 2023 年) 发展。最近, 以 DeepseekR1 (深度探索人工智能公司等人, 2025 年) 等模型为代表的可验证奖励强化学习, 使用组相对策略优化 (GRPO) (邵等人, 2024 年) 来训练策略模型, 该方法通过使用组值估计基线, 从而无需评判模型, 与 PPO 相比大幅降低了训练成本。其令人瞩目的表现重新激发了人们对传统强化学习框架的兴趣。

局限性 4: 忽视推理多样性与效率。使用最优推理轨迹来监督模型的对齐。这种文本对齐基于单词层面的重叠, 促使生成的推理路径与最优路径相似。然而, 这种方法没有考虑到有效推理路径在语言表达上的多样性, 即不同的表达方式仍可能得出相同的结果。此方法未能捕捉到潜在的推理模式, 限制了模型在不同场景下的泛化能力。此外, 利用大语言模型 (LLMs) 作为评判者来评估文本推理轨迹, 由于频繁推理, 会产生较高的计算成本 (Chen 等人, 2025b)。可行的见解见第 5 节: 行动 4 和行动 5。

4.3. 评估与监管

经过监测后, 得到完整的推理链 A。图 5 中的评估与调整模块利用反馈 0 进行优化。值得注意的是, 监测作为思维过程的持续观察者, 即 “边做边思考”, 也就是旨在提升内在推理能力的训练后阶段。评估与调整则从整体上审视推理过程, 即 “做完再思考”, 也就是推理过程中应用的策略。因此, 我们关注现有反馈如何在推理过程中帮助纠错方法。

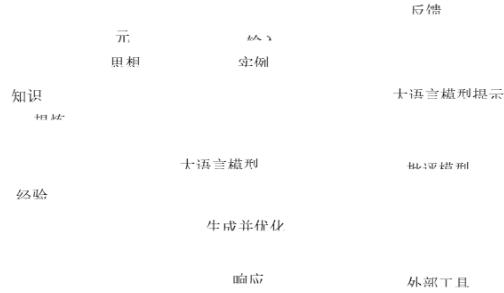


图 5. 评估与调节。利用先前交互中的元思考来增强反馈。其他反馈来源总结在虚线框中。

4.3.1. 整体推理过程评估

用于评估整个推理链的现有典型反馈总结如下：(i) 提示大语言模型 (LLMs)，(ii) 利用经过训练的评判模型，以及(iii) 整合工具（如图 5 所示）。然而，大语言模型提示并不能保证底层推理链是可靠、准确或逻辑合理的。用于错误检测的特定任务评判模型，应用于数学推理 (Chen 等人, 2024b)、编码 (Kumar 等人, 2024) 和逻辑推理 (Tong 等人, 2024) 等领域，为改进提供了替代方法。为了为评判模型创建训练数据，首先从基础大语言模型生成推理过程，然后要求人类注释者提供错误位置（哪个推理步骤）和错误类型（例如知识错误或计算错误）的详细注释 (Tyen 等人, 2023; Uesato 等人, 2022)。值得注意的是，即使评判模型成功识别出错误，它可能仍然无法清晰地阐明这些错误。其他方法整合了外部工具，例如代码编译器 (Shinn 等人, 2023b; Yao 等人, 2024)。

上述方法所生成的反馈主要是针对特定样本的，无法捕捉相似问题中的普遍错误模式。最近，思维模板 (Yang 等人, 2024 年; Wang 等人, 2024g; Yang 等人, 2025a)，如符号化问题，已成为一种可行的组织相似问题的方法。这些模板有助于对多个类似案例进行评估，从而使反馈能够反映常见的推理模式。

4.3.2. 推理过程调节

关于纠错的监管问题。鉴于在评估阶段产生的反馈，大语言模型 (LLMs) 应遵循该反馈，并通过以下方式进行适当修订

称为自我反思阶段 (申恩等人, 2023a)。然而，并不能保证大语言模型会严格遵循指令。事实上，严等人 (2024b) 观察到，大语言模型往往过于固执，不愿改变其初始回答，即使得到明确反馈“你的回答不正确，请重新考虑”。为了明确引导大语言模型更新其回答，TextGrad (于克塞贡努尔等人, 2025) 将大语言模型提供的文本反馈反向传播，以优化初始输入提示，有效地将自然语言用作梯度。另一项研究使用明确的纠错轨迹来训练大语言模型 (张等人, 2025a; 泽利克曼等人, 2024; 曾等人, 2024a)。DeepSeek –

局限性 5： 缺乏自适应的元级错误分析。现有的大语言模型优化研究主要集中在实例级的错误检测与纠正上，这类研究针对单个实例解决错误，而没有利用反复出现的错误模式中获得的见解 (于克塞贡努尔等人, 2025 年; 王等人, 2024d)。然而，要实现大语言模型性能更广泛的提升，需要一种元级别的方法，即分析错误模式以识别潜在的系统性问题或偏差，从而推动制定策略，预防未来实例中出现类似错误。虽然像元缓冲区 (杨等人, 2024 年) 和语义符号提示 (王等人, 2024g) 等方法已证明利用先前交互和结构化推理的有效性，但它们在很大程度上依赖于大语言模型的固有能力以及对元学习的支撑。

4.4. 元反思

元反思是一种贝叶斯学习过程，旨在根据多个任务的反馈来更新模型的初始观点。它采用一种双层方法：首先，优化初始策略 F ，然后相应地优化元参数 Θ_1 和 Θ_E (图 2 中的虚线)。核心挑战在于有效地整合不同任务之间的关系，确保元更新对于所有可能的输入模型无关元学习 (MAML) (芬恩等人, 2017 年) 是一种通用的元学习框架，它通过双层优化学习一个与任务无关的模型初始化，以便快速适应新任务。为使该方法适用于大语言模型 (LLM) 部署，可以采用几种现有技术：(i) 双层提示优化。(ii) 诸如低秩自适应 (LoRA) 之辈的模块化训练方法。(iii) 贝叶斯

通过以连续的方式在一批不同的任务上训练模型来进行优化，简化该过程使其类似于传统的微调。Qin 等人（2023 年）；Sinha 等人（2024 年）提出了元提示，并遵循双层优化过程。动态模块化组合（Huang 等人，2024a；Yang 等人，2025b），特别是与低秩自适应（LoRA）相结合时，为重新组合和重新组织特定能力的组件提供了一种灵活的机制，通过模块化重组能够有效地泛化到新任务。多智能体强化学习框架，如 ReMA（Wan 等人，2025 年）引入了一个元思考智

为了基于累积奖励 R 推导出优化的推理策略 F ，我们可以从逆向规划中获得灵感，逆向规划是指根据心理理论推断智能体的不可观测状态，如目标和信念（Shum 等人，2019 年）。具体来说，我们需要近似 $R(F; O, \Theta_I, \Theta_E)$ 。一个研究方向是利用大语言模型偏好（Zhang 等人，2025c），例如，通过使用模型在给定当前观察和候选策略下的生成对数几率进行评估。

局限性 6：元优化缺乏可解释性和效率。 虽然像 LoraHub 这样的方法可以为新任务分解和重新组合专业能力，但在模型合并过程中，它们往往存在安全性和可靠性问题（Hammoud 等人，2024 年；Hsu 等人，2024 年）。这些风险凸显了我们对模型迁移学习和模型参数组合的潜在机制理解有限。此外，迫切需要高效的元优化框架，如多智能体或多阶段强化学习。此类方法可以作为复杂推理任务的基础架构，实现更好的智能体协调、工具整合，以及针对未知任务的进化

可行的见解

本节受第 4 节所述限制的启发，提出了未来可能的研究方

行动 1：元推理评估的基准和指标 为了评估大语言模型的元推理能力，需要有明确界定的基准来评估自我意识、内省和反思性推理。像 SAD（莱内等人，2024 年）、AwareBench（李等人，2024b）和 MM-SAP（王等人，2024f）等最新数据集专注于内省和多模态推

2025 年）将评估扩展到错误分析和定性洞察。然而，现有的大多数基准测试都集中在数学和编码任务上，尚未推广到更广泛的推理任务。未来的工作应致力于将这些基准测试整合到一个统一的框架中，并开发除准确率之外的指标，如校准误差、逻辑一致性、一致率和泛化性能，以评估元推理能力。例如，最近的一个基准测试 Feedbacker（Wang 等人，2025 年）提供了一个全面的评估框架，通过对比各种推理任务（如注

行动 2：基于神经符号系统的多视角可解性 虽然不确定性或置信度分数可以表明能力感知可解性，但仅凭这些是不够的。如前文所述，还必须考虑任务感知可解性，例如拒绝不道德的请求（Li 等人，2024b）。挑战在于将这些不同的可解性方面整合到一个统一的决策框架中。可以探索一种神经符号方法，该方法将符号推理的精确性与神经模块的表达能力相结合（Andreas 等人，2016 年；Gupta 等人，2020 年）。不同的可解性指标可以作为神经模块纳入，模块化框架能够适应新的指标。选择一种符号方法来协调和执行神经模块至关重要：概率框架提供稳健性

行动 3：自适应推理策略生成 当前的“从计划到计划”方法通常会为给定的推理实例或任务中的所有实例生成单一策略（邹等人，2023 年；高等人，2024b）。然而，任务可能需要多种推理技能，如知识检索和数值计算。为了使大语言模型能够在各种任务中实现泛化，我们可以将输入上下文表示映射到潜在概念空间，其中每个概念对应一种独特的推理技能。解决问题将涉及识别相关技能并基于这些技能生成答案。专家混合（MoE）框架允许将推理技能（专家）动态分配给特定实例。最近将 MoE 与参数高效微调相结合的工作提高了效率。此外，分层 MoE 可以进一步改善任务间的技能共享（李等人，2025b）。另一种可能的方法是利用贝叶斯逆向规划（吴等人，2021 年；张等人，2025c），将推理技能视为元知识和推理行动之间的潜在变量。通过观察推理行动的结果，可以根据

行动 4：通过自我博弈寻求元奖励

人类智能会针对推理发展出多方面的自我评估，并通过与环境的互动动态引入新的标准。相比之下，当前用于推理监测的奖励系统较为单一，主要侧重于正确性且保持静态，难以适应策略模型中不断演变的分布情况（陶等人，2024 年）。因此，我们提议通过自我博弈利用多方面且动态的元奖励。这一主张与近期的理论和实证研究结果相符，即扩大反馈 / 奖励规模可在训练和推理阶段带来显著提升（斯内尔等人，2024 年；吴等人，2024c）。这样一个自我进化系统使大语言模型能够自主获取、完善并从自身生成的经验或微妙的内部信号（如置信度）中学习（赵等人，2025 年）。此外，自我博弈范式减少了对人类偏好数据的依赖，并缓解了奖励操控问题。要实现这一范式，需

行动 5：潜在空间推理以提高多样性和效率 大多数现有的推理方法以自回归方式生成明确的语言中间步骤。然而，这些步骤中的错误可能会累积，导致级联错误、自我纠正困难以及效率低下（邓等人，2024 年；林等人，2021 年；叶等人，2024 年）。通过将明确的想法内化到潜在空间中，我们可以捕捉独立于语言风格的推理模式，促使模型多思考、少表达，同时避免在生成冗长序列上花费不必要的成本，从而加快模型推理速度。初步研究（邓等人，2024 年；郝等人，2024 年；沈等人，2025 年）表明，通过完全绕过冗长的中间语言步骤可以实现更快的推理，尽管仍落后于采用语言形式的思维链方法。循环变压器不依赖额外的标记，而是通过增加计算深度来增强思考能力，从而进行推理，这也可以说是一种潜在推理形式（盖平等人，2025 年；桑希等人，2025 年；余等人，2025 年）。此外，经过良好正则化的潜在空间可以进一步提高可解释性和全局可控性（叶等人，2024 年；苏等人，2025 年），并通过嵌入搜索加快模拟过

行动 6：用于元知识整合的可解释且高效的训练 为了提高大模型的适应性和效率，识别并

利用不同子网络或特定技能的智能体 / 工具的独特作用，针对特定输入选择性地利用 / 更新最相关的组件。这种有针对性的方法提高了模型理解大语言模型知识学习和巩固过程的能力。最近的研究表明，大幅的性能提升通常来自更新 5%-30% 的模型参数（穆克吉等人，2025 年）。因此，机制可解释性可以为将特定模型组件与输出联系起来的严格因果效应提供有价值的见解（严等人，2024a；贝雷斯卡和加夫斯，2024 年）。此外，在多目标合作框架中，让智能体具备识别和理解自身知识边界的能力（乔等人，2025 年）至关重要，在该框架中，一个元级协调器负责监督多个

其他观点

一些人建议，大语言模型应在人类监督下运行，以确保可靠的推理和决策（拉斐洛夫等人，2023 年）。另一些人则认为，通过将结构化知识库和符号推理整合到大语言模型中，可以实现可靠的推理（郝等人，2023b；舒等人，2024 年）。我们提出的大语言模型元推理框架也可能因增加复杂性和计算开销而受到批评。我们的论点是：(i) 人类监督资源密集，要针对每个用例进行扩展并不现实，尤其是在实时应用中。(ii) 仅靠符号推理器难以应对自然语言的复杂性和微妙之处。包括符号推理器在内的外部求解器是我们元推理框架的一部分。(iii) 与特定任务模型不同，元推理使大语言模型能够通过反思和调整其推理策

结论

我们引入了一种贝叶斯元推理框架，该框架受到人类认知过程的启发，整合了自我意识、监测、评估和元反思等关键要素。它解决了现有方法中的一些基本局限性，比如缺乏动态适应性、推理路径的多样性有限，以及针对特定任务的更新效率低下等问题。通过纳入外部资源、基于元知识的评估以及灵活的采样机制，我们的方法在跨领域复杂、非结构化推理方面展现出巨大的潜力。此外，我们还强调了元推理中的关

布朗, T., 曼恩, B., 赖德, N., 萨比亚, M., 卡普兰, J.D., 达里瓦尔, P., 尼拉坎坦, A., 施亚姆, P., 萨斯崔, G., 阿斯克尔, A., 阿加瓦尔, S., 赫伯特 - 沃斯, A., 克鲁格, G., 赫尼根, T., 奇尔德, R., 拉梅什, A., 齐格勒, D., 吴, J., 温特, C., 黑塞, C., 陈, M., 西格 ler, E., 利特温, M., 格雷, S., 切斯, B., 克拉克, J., 伯纳, C., 麦坎德利什, S., 拉德伯恩斯, C., 叶, H., 克莱因, D., 以及施泰因哈型是少样本学习者》, 发表于《神经信息处理系统进展》特, I. 《无监督发现语言模型中的潜在知识》。发表

卡塞利, T., 巴西莱, V., 米特罗维奇, J., 以及格拉尼策, M. 《仇恨 BERT: 重新训练 BERT 用于英语辱骂性语言检测》, 载于《第五届网络滥用与危害研讨

卡塔克, F. O. 和库兹卢, M. 通过凸包分析对大语言模型进行不确定性量化。《发现人工智能》, 2024

查蒂拉, R., 雷诺多, E., 安德里埃, M., 查韦斯 - 加西亚, R.O., 吕斯 - 韦亚克, P., 戈特斯坦, R.,

陈, G.; 廖, M.; 李, C.; 范, K.《Alphamath 几乎为零: 无过程的过程监督》, 载于《第三十八届神经

陈, G., 廖, M., 李, C., 以及范, K.。用于数学推理的步骤级价值偏好优化。发表于《自然语言处理实

陈, X., 林, M., 舍尔利, N., 以及周, D. 《教大语言模型自我调试》, 收录于《第十二届学习表征国际

陈, X., 徐, J., 梁, T., 何, Z., 庞, J., 余,
D., 宋, L., 刘, Q., 周, M., 张, Z., 王, R.,
涂, Z., 米, H., 余, D.。对于 “ $2 + 3 = ?$ ” 这

陈, Y., 钟, R., 查, S., 卡里皮斯, G., 和何, H.
《通过语言模型上下文调优进行元学习》。收录于
《第 60 届计算语言学协会年会论文集(第 1 卷: 长

陈, Y., 本顿, J., 拉达克里希南, A., 丹尼森, J.
U. C., 舒尔曼, J., 索马尼, A., 哈泽, P., 米库利
克, M. W. E. R. V. 鮑曼, S. 卡普兰, I. I. I.

陈, Z., 周, K., 张, B., 龚, Z., 赵, X., 以及文, J.-R. 《ChatCoT: 基于聊天的大语言模型上的工具增强思维链推理》。发表于《计算语言学协会研究成果: 2023 年自然语言处理与对话研究》, 2023 年。

陈 Z.、牛 X.、Foo C.-S. 和 Low B. K. H. 拓宽视野! 利用语义空间实现大语言模型高效多轮对话规
则。收录于《第 1 届国际学术会议论文集》, 2025 年。

郑 (Cheng)、拉贾 (Raja) 和莱瑟 (Lesser)。用于雷达协调的多智能体元级控制。《网络智能与智能体
迟, H., 李, H., 杨, W., 刘, F., 兰, L., 任,
X., 刘, T., 韩, B. 《揭示大语言模型中的因果推
理》, 2025 年。

科斯特, T., 安瓦尔, U., 柯克, R., 以及克鲁格,
D. 《奖励模型集成有助于减轻轻度优化问题》, 载于
“...”。

考克斯, M. T. 《元推理、监测与自我解释》, 载于《元推

考克斯, M. T. 和拉贾, A. 《元推理: 思考思考本

迪普 Seek 人工智能公司, 刘, A., 冯, B., 薛,
B., 王, B.-L., 吴, B., 卢, C., 赵, C., 邓, C.,
张, C., 阮, C., 戴, D., 郭, D., 杨, D., 陈,
D., 季, D.-L., 李, E., 林, F., 戴, F., 罗, F.,
...。

DeepSeek-AI, 郭, D., 杨, D., 张, H., 宋, J.-
M., 张, R., 徐, R., 朱, Q., 马, S., 王, P.,
毕, X., 张, X., 于, X., 吴, Y., 吴, Z.F., 荀,
Z., 邵, Z., 李, Z., 高, Z., 刘, A., 薛, B.,
王, B.-L., 吴, B., 冯, B., 卢, C., 赵, C., 邓,
C., 张, C., 阮, C., 戴, D., 陈, D., 季, D.-L.,
邓, G., 刘, Y., 李, Y., 王, K., 张, Y., 李,
Z., 王, H., 张, T., 以及刘, Y. 《越狱者: 跨多个

邓, Y., 崔, Y., 以及希伯, S. 《从显式思维链到隐
式思维链: 逐步学习内化思维链》。预印本网站

董浩、熊伟、戈亚尔、张宇、周伟、潘睿、刁思、张
军、沈凯、张涛。《RAFT: 用于生成式基础模型对齐
的奖励排序微调》, 《机器学习研究汇刊》, 2023

段, J., 程, H., 王, S., 扎瓦尔尼, A., 王, C.,
徐, R., 凯尔胡拉, B., 徐, K. 将注意力转移到相关
性上: 迈向自由形式大语言模型的预测不确定性量

艾森斯坦, J., 纳格帕尔, C., 阿加瓦尔, A., 贝拉
米, A., 达莫尔, A. N., 德维乔瑟姆, K. D., 菲
施, A., 赫勒, K. A., 普福尔, S. R., 拉马钱德
兰, D., 肖, P., 和贝兰特, I. 《帮助还是引导? 奖

埃尔博赫尔, A., 本苏桑, A., 卡帕斯, E., 鲁姆
尔, W., 什佩尔伯格, S. S., 以及西蒙尼, S. E.

冯杰、黄帅、曲鑫、张刚、秦宇、钟波、蒋超、迟
杰、钟伟。《ReTool: 大语言模型中策略性工具使用

冯, S., 石, W., 王, Y., 丁, W., 巴拉钱德兰,
V., 以及茨韦特科夫, Y. 《别产生幻觉, 弃权: 通
过多大语言模型协作识别大语言模型的知识差距》。收
录于《计算语言学协会第 62 届年会论文集 (第 1

冯, Y., 周, B., 林, W., 以及罗斯, D. 《BIRD: 一
种用于大语言模型的可信贝叶斯推理框架》。收录于
“...”。

芬恩, C., 阿贝埃尔, P., 莱文, S. 《深度网络快速
适应的模型无关元学习》, 载于《机器学习国际会

弗拉维尔, J. H. 《元认知与认知监控: 认知发展研
究的一个新领域》, 《美国心理学家》, 1979 年。

傅, Y., 彭, H., 萨巴尔瓦尔, A., 克拉克, P., 和
霍特, T. 《基于复杂性的多步推理提示》。载于《第

傅 (Fu)、彭 (Peng, H.-C.)、霍特 (Khot, T.) 和
拉帕塔 (Lapata, M.)。通过自我博弈和从人工智能
...。

郝, S., 顾, Y., 马, H., 洪, J., 王, Z., 王, D., 以及胡, Z.。使用语言模型进行推理即使用世界模型进行规划。收录于布阿穆尔, H., 皮诺, J., 以及巴利, K. (编), 《2023 年自然语言处理实证方法会议论文集》, 第 8154 – 8173 页, 新加坡, 2023 年 12 月 a。计算语言学协会。doi: 10.18653/v1/2023.emnlp-main.507。

郝帅、刘婷、王泽、胡志。ToolkenGPT: 通过工具嵌入用大量工具增强冻结的语言模型。《第三十七届神经信息处理系统会议 (NIPS 2023)》, 2023 年 12 月

郝, S., 苏赫巴塔尔, S., 苏, D., 李, X., 胡, Z., 韦斯顿, T., 田, Y. 《训练大语言模型在连续潜

许智渊、蔡宜霖、林哲宏、陈柏宇、余宗民、黄崇颖。《安全的低秩自适应微调: 在微调大语言模型时降低安全风险的一线希望》。收录于: 阿萨夫·格洛贝森、利奥·麦基、达维娅·贝尔格雷夫、范安邦、

胡, Z., 严, H., 朱, Q., 沈, Z., 何, Y., 桂, L. 《超越提示: 一种用于开放域问答的高效嵌入框架》。预印本平台 ArXiv, 论文编号 abs/2503.01606, 2025

黄, C., 刘, Q., 林, B.Y., 庞, T., 杜, C., 以及林, M. 《Lorahub: 通过动态低秩自适应组合实现高

黄, J., 陈, X., 米什拉, S., 郑, H. S., 余, A. W., 宋, X., 以及周, D. 《大语言模型尚未能自我纠正》。预印本平台 ArXiv, 论文编号 abs/2503.01606, 2025

黄, X., 阮, W., 黄, W., 金, G., 董, Y., 吴, C., 本萨勒姆, S., 穆, R., 齐, Y., 赵, X. 等人。从验证与确认视角看大语言模型的安全性与可信性综

黄, Y.、孙, L.、王, H.、吴, S.、张, Q.、李, Y.、高, C.、黄, Y.、吕, W.、张, Y.、李, X.、孙, H.、刘, Z.、刘, Y.、王, Y.、张, Z.、维德根, B.、凯尔库拉, B.、熊, C.、肖, C.、李, C.、邢, E. P.、黄, F.、刘, H.、季, H.、王, H.、张, H.、姚, H.、凯利斯, M.、齐特尼克, M.、蒋, M.、班萨尔, M.、邹, J.、裴, J.、刘, J.、高, J.、韩, J.、赵, J.、唐, J.、王, J.、范肖伦, J.、米

- 大型语言模型的可信度。载于萨拉胡特迪诺夫 (Salakhutdinov)、科尔特 (Kolter)、赫勒 (Heller)、韦勒 (Weller)、奥利弗 (Oliver)、斯卡利特 (Scarlett) 和贝肯坎普 黄, Y., 宋, J., 王, Z., 赵, S., 陈, H., 决非 徐, F., 以及马, L.。三思而后行：大语言模型不确定性的研究方法。// TACL 大语言模型研究
江, M., 阮, Y., 黄, S., 廖, S., 皮蒂斯, S., 格 罗斯, R. B., 以及巴, J. 通过增强提示集成校准语 金, B., 曾, H., 岳, Z., 王, D., 扎马尼, H., 以 及韩, J.《Search-r1：利用强化学习训练大语言模型 》。// TACL 大语言模型研究
卡达瓦斯, S., 科纳利, T., 阿斯克尔, A., 赫尼 根, T., 德雷恩, D., 佩雷斯, E., 希弗, N., 多 兹, Z., 达萨尔马, N., 陈 - 约翰逊, E., 约翰斯 顿, S., 埃尔 - 肖克, S., 琼斯, A., 埃尔哈格, N., 休姆, T., 陈, A., 白, Y., 鲍曼, S., 福特, S., 甘吉利, D., 埃尔南德斯, D., 雅各布森, J., 卡尼曼, D.《思考，快与慢》。法勒、施特劳斯和吉鲁出 利斯, A.、陈, V. Q.、泰, Y.、索伦森, J.、古普 塔, J.、梅茨勒, D. 和瓦瑟曼, L. 新一代视角应用 程序编程接口：高效的多语言字符级变换器。载于 《第 28 届美国计算机协会知识发现与数据挖掘会议 李, J., 桂, L., 周, Y., 韦斯特, D., 阿洛西, C., 和何, Y.《提炼 ChatGPT 用于可解释的自动学生答案 评估》。收录于布阿莫尔, H., 皮诺, J., 和巴利, K. (编), 《计算语言学协会研究成果：2023 年自然 语言处理经验方法会议》，第 6007 – 6026 页，新嘉 坡，2023 年 12 月。计算语言学协会。文献编号： 李, J., 徐, H., 孙, Z., 周, Y., 韦斯特, D., 阿洛西, C., 和何, Y. 在科学问题评分中，通过对思 维树进行偏好优化来校准大语言模型以生成推理依 李, J., 周, Y., 陆, J., 泰恩, G., 桂, L., 阿洛 西, C., 和何, Y. 三个臭皮匠，顶个诸葛亮：推理时 的双模型言语反思。ArXiv, abs/2502.19230, 2025a. 李, W., 王, D., 丁, Z., 索拉比扎德, A., 秦, Z., 丛, J., 以及孙, Y.《层次化专家混合：高级综 合的可泛化学习》。发表于 2025 年美国人工智能协 会主办的 AAAI25 会议。2025 年 2 月 25 日 – 3 日 库马尔, A., 庄, V., 阿加瓦尔, R., 苏, Y., 科 - 雷耶斯, J. D., 辛格, A., 鲍姆利, K., 伊克巴尔, S., 毕晓普, C., 罗洛夫斯, R., 张, L. M., 麦金 尼, K., 施里瓦斯塔瓦, D., 帕杜拉鲁, C., 塔克, 莱恩, R.; 乔塔伊, B.; 贝特利, J.; 哈里哈兰, K.; 巴勒什尼, M.; 舍勒, J.; 霍布哈恩, M.; 迈因 克, A.; 以及埃文斯, O.《我、我自己和人工智能： 针对大语言模型的态势感知数据集 (SAD)》，载于 朗格卢瓦, S. T., 阿科罗达, O., 卡里略, E., 赫尔 曼, J. W., 阿扎姆, S., 徐, H., 以及奥特, M.《多 动机学习：从理论到实践》。
劳特曼, T. 与阿克曼, R.《非言语问题可解性的初 步判断——解决过程的预测指标》。《元认知与学 习：从理论到实践》。
利斯, A.、陈, V. Q.、泰, Y.、索伦森, J.、古普 塔, J.、梅茨勒, D. 和瓦瑟曼, L. 新一代视角应用 程序编程接口：高效的多语言字符级变换器。载于 《第 28 届美国计算机协会知识发现与数据挖掘会议 李, J., 桂, L., 周, Y., 韦斯特, D., 阿洛西, C., 和何, Y.《提炼 ChatGPT 用于可解释的自动学生答案 评估》。收录于布阿莫尔, H., 皮诺, J., 和巴利, K. (编), 《计算语言学协会研究成果：2023 年自然 语言处理经验方法会议》，第 6007 – 6026 页，新嘉 坡，2023 年 12 月。计算语言学协会。文献编号： 李, J., 徐, H., 孙, Z., 周, Y., 韦斯特, D., 阿洛西, C., 和何, Y. 在科学问题评分中，通过对思 维树进行偏好优化来校准大语言模型以生成推理依 李, J., 周, Y., 陆, J., 泰恩, G., 桂, L., 阿洛 西, C., 和何, Y. 三个臭皮匠，顶个诸葛亮：推理时 的双模型言语反思。ArXiv, abs/2502.19230, 2025a. 李, W., 王, D., 丁, Z., 索拉比扎德, A., 秦, Z., 丛, J., 以及孙, Y.《层次化专家混合：高级综 合的可泛化学习》。发表于 2025 年美国人工智能协 会主办的 AAAI25 会议。2025 年 2 月 25 日 – 3 日

李, Y., 黄, Y., 林, Y., 吴, S., 万, Y., 以及孙, L. 我思故我在: 使用 AwareBench 对大语言模型
李, Y., 杨, C., 以及埃廷格, A. 《当 insight 并非 20/20: 测试大语言模型中反思性思维的极限》,

梁, X., 宋, S., 郑, Z., 王, H., 于, Q., 李, X., 李, R.-H., 熊, F., 李, Z. 《大语言模型中的
由初一动件上白华山碑 1999 五印大藏经》

莱特曼, H., 科萨拉朱, V., 布尔达, Y., 爱德华兹, H., 贝克, B., 李, T., 莱克, J., 舒尔曼, J., 萨茨克弗, I., 以及科布, K. 《让我们逐步验

林聪 - 聪、亚赫、李昕、戈尔姆利、艾斯纳。自回归模型的局限性及其替代方案。载于图塔纳娃、鲁姆希斯基、泽特勒莫耶、哈坎尼 - 图尔、贝尔塔吉、贝沙德、科特雷尔、查克拉博蒂、周宇 (编), 《2021 年北美计算语言学协会会议录: 人类语言技术》, 第 5147 – 5173 页, 线上会议, 2021 年 6 月。计算语

林, C. H., 科洛博夫, A., 卡马尔, E., 以及霍维茨, E. 不确定性下规划的元推理。《第 24 届国际人工智能会议 (IJCAI) 论文集》, 第 1601 – 1609 页, 2015 年。

刘, F., 徐, Z., 以及刘, H. 《对抗性微调: 防御大语言模型的越狱攻击》。arXiv 预印本

刘, L., 潘, Y., 李, X., 和陈, G. 《大语言模型的不确定性估计与量化: 一种简单的监督学习方法》。

刘, O., 傅, D., 约加塔马, D., 以及奈斯旺格, W. 《戴尔玛: 利用大语言模型在不确定性下进行决策》。arXiv 预印本

刘, R., 耿, J., 吴, A. J., 苏乔卢茨基, I., 隆布罗佐, T., 以及格里菲斯, T. L. 注意你的步骤 (循序渐进)。arXiv 预印本

洛查布, A. 和张, R. 《基于能量的奖励模型, 实现稳健的

陆 (音译)、钟 (音译)、黄 (音译)、王 (音译)、米 (音译)、王 (音译)、王 (音译)、尚

马丹, A., 坦登, N., 古普塔, P., 哈利南, S., 高, L., 维格雷夫, S., 阿隆, U., 兹里, N., 普拉布莫耶, S., 杨, Y., 古普塔, S., 马朱姆德, B. P., 赫尔曼, K., 韦莱克, S., 亚兹丹巴赫什,

马纳库尔, P., 柳西, A., 以及盖尔斯, M.

《SelfCheckGPT: 生成式大语言模型的零资源黑盒幻觉检测》。收录于布阿穆尔, H., 皮诺, J., 以及巴利, K. (编), 《2023 年自然语言处理经验方法会议论文集》, 第 9004 – 9017 页, 新加坡, 2023 年 12

麦科伊, R. T., 姚, S., 弗里德曼, D., 哈迪, M., 以及格里菲思, T. L. 《自回归的余烬: 通过大语言模

米尔克, S. J., 斯拉姆, A., 迪南, E., 以及布雷奥, Y.-L. 通过语言校准减少对话智能体的过度自信。《计算语言学协会会刊》, 2022 年, 第 10 卷,

闵 (Min)、刘易斯 (Lewis)、泽特勒莫耶 (Zettlemoyer) 和哈吉希尔齐 (Hajishirzi)。
《MetaICL: 情境中学习如何学习》。载于卡普阿 (Carpuat)、德马内夫 (de Marneffe) 和梅萨·鲁伊斯 (Meza Ruiz) 编, 《2022 年北美计算语言学协

莫宇、王宇、魏泽和王宇。通过提示对抗调优反击越狱问题。发表于《第三十八届神经信息处理系统年度会议》。2021 年

穆克吉, S., 袁, L., 哈坎尼 - 图尔, D., 以及彭, H. 《强化学习对大语言模型中的小子网络进行微调》。2025 年。网址

默多克, J. W. 与戈埃尔, A. K. 《元基于案例推理: 通过自我理解实现自我提升》, 《实验与理论人工智能杂志》, 2008 年, 第 20 卷第 1 期, 第 1 – 36 页。

纳法尔, A., 维纳布尔, K. B., 以及科尔贾姆希迪, P. 生成式大语言模型中的概率推理。arXiv 预印本

纳尔逊, T. O. 《元记忆: 一个理论框架与新发现》, 载于《学习与动机心理学》第 26 卷, 第 125

OpenAI。利用大语言模型学习推理。发表于《第三十七届神经信息处理系统大会》, 2024 年。

欧阳龙、吴杰、蒋鑫、阿尔梅达、温赖特、米什金、张晨、阿加瓦尔、斯拉马、雷等人, 《利用人类反馈训练语言模型以遵循指令》, 《神经信息处理系统进

潘, L., 萨克森, M. S., 徐, W., 纳萨尼, D., 王, X., 以及王, W. Y. 《自动纠正大语言模型: 审视各种自动纠正策略的全景》。《计算语言学协会汇

保罗, D., 伊斯马伊勒扎达, M., 佩亚尔, M., 博尔热斯, B., 博塞尔卢, A., 韦斯特, R., 以及法尔廷斯, B. 《REFINER: 对中间表示的推理反馈》。载于格雷厄姆, Y. 和珀弗, M. (编), 《第 18 届欧洲计算语言学协会分会会议论文集 (第 1 卷: 长论

彭博、加莱、何鹏、程浩、谢宇、胡宇、黄强、利登、于泽、陈伟、高杰。核实事实, 重新尝试: 利用外部知识和自动反馈改进大语言模型。《预印本》,

普鲁斯特, J. 《元认知哲学: 心理能动性与自我意识》。牛

钱晨、阿齐克戈兹、何强、王浩、陈鑫、哈坎尼 - 图尔、图尔、季浩。《ToolRL: 奖励是工具学习所需的乔, S., 邱, Z., 任, B., 王, X., 茹, X., 张, N., 陈, X., 江, Y., 谢, P., 黄, F., 和陈, H. 《智能体知识渊博的自我意识》。发表于《大型语言模型的推理与规划研讨会》, 2025 年。网址

秦, C., 乔蒂, S., 李, Q., 以及赵, R. 《学习初始化: 元学习能否提升提示调优中的跨任务泛化能力?》, 载于罗杰斯, A., 博伊德 - 格雷伯, J., 以及冈崎, N. (编), 《第 61 届计算语言学协会年会论文集 (第 1 卷: 长论文)》, 第 11802 – 11832

秦, Y., 李, X., 邹, H., 刘, Y., 夏, S., 黄, Z., 叶, Y., 袁, W., 刘, H., 李, Y., 以及刘, P.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L.,

拉法伊洛夫, R., 夏尔马, A., 米切尔, E., 曼宁, C.D., 埃尔蒙, S., 以及芬恩, C. 《直接偏好优化: 你的语言模型实则是一个奖励模型》, 载于《第三十

拉贾, A. 和莱瑟, V. 多智能体系统中元级控制的框架。《自主智能体与多智能体系统》, 2007 年, 第

萨巴塔, C. N. D., 萨默斯, T. R., 以及格里菲思, T. L. 《大语言模型的理性元推理》。提交至第十三

萨姆索诺维奇, A. V. 和德容, K. A. 《团队智能体的元认知架构》。收录于《第 25 届认知科学学会年

绍恩希, N., 迪卡拉, N., 李, Z., 库马尔, S., 以及雷迪, S. J. 《基于潜在思维的推理: 循环变压器的工具》。第 11 届国际会议, 2025 年。

舒尔曼, J., 沃尔斯基, F., 达里瓦尔, P., 拉德福德, A., 以及克里莫夫, O. 《近端策略优化算法》。

塞尔万特斯, S., 巴罗, J., 哈蒙德, K., 以及贾因, R. 《逻辑链: 基于规则的大语言模型推理》。收录于顾立威、马丁斯、斯里库马尔编, 《计算语言学协会研究成果: 2024 年计算语言学协会年会》, 第

邵志飞、王鹏、朱琦、徐瑞、宋佳明、张明、李一康、吴宇、郭栋。《Deepseekmath: 探索开源语言模型的数学推理能力》。载于《2024 年美国计算机协会信息系统安全专业委员会计算机与通信安全会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

沈 X、陈 Z、巴克斯 M、沈 Y 和张 Y 所著的《“现在无所不能”：对大语言模型中现实世界越狱提示的特征描述与评估》，收录于《2024 年美国计算机协会信息系统安全专业委员会计算机与通信安全会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

沈, Z., 严, H., 张, L., 胡, Z., 杜, Y., 和何, Y. 《Codi: 通过自蒸馏将思维链压缩到连续空间》。预印本网站，2024 年 5 月 20 日。

申恩, N., 卡萨诺, F., 戈皮纳特, A., 纳拉辛汉, K. R., 以及姚, S. 《反思：基于言语强化学习的思维链》。载于《第 13 届国际学习表示方法会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

申恩 (Shinn)、卡萨诺 (Cassano)、拉巴什 (Labash)、戈皮纳思 (Gopinath)、纳拉辛汉 (Narayanan)、姚 (Yao)。《反思：基于言语强化学习的思维链》。载于《第 13 届国际学习表示方法会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

舒迪、陈涛、金梦、张宇、张晨、杜明和张宇。用于链接预测的知识图谱大语言模型 (kgllm)。预印本网站，2024 年 5 月 20 日。

沈 (Shum)、克莱曼 - 韦纳 (Kleiman-Weiner)、利特曼 (Littman, M. L.) 和特南鲍姆 (Tenenbaum, J. B.)。《心理理论：通过逆向规划理解群体行为》。发表于 2019 年 10 月 1 日。

辛格, A. K., 德夫科塔, S., 拉米钱内, B., 达卡尔, U., 以及达卡尔, C. 《大语言模型中的置信度 - 能力差距：一项认知研究》。预印本网站 ArXiv, 论文编号 arXiv:2405.14210, 2024 年 5 月 20 日。

辛哈, S., 岳, Y., 索托, V., 古尔卡尼, M., 卢, J., 以及张, A. 《Maml-en-llm: 通过大语言模型的模型无关元训练提升上下文学习能力》。发表于《第 30 届 ACM SIGKDD 知识发现与数据挖掘会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

斯卡尔塞, J., 豪, N., 克拉申尼科夫, D., 以及克鲁格, D. 《定义与描述奖励博弈》，《神经信息处理系统会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

斯内尔, C., 李, J., 徐, K., 以及库马尔, A. 《对大语言模型推理时的计算量进行最优扩展，比扩展模型更有效》。载于《第 29 届国际人工智能联合会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

斯内尔, C. V., 李, J., 徐, K., 以及库马尔, A. 《对大语言模型推理时的计算量进行最优扩展，可能更有效》。载于《第 29 届国际人工智能联合会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

斯普拉格, Z., 尹, F., 罗德里格斯, J. D., 蒋, D., 瓦德瓦, M., 辛哈尔, P., 赵, X., 叶, X., 马霍瓦尔德, K., 以及达雷特, G. 《要不要思维链？思维链主要有助于数学和符号推理》。第十三届国际学习表示方法会议论文集，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

苏, D., 朱, H., 徐, Y., 焦, J., 田, Y., 郑, Q. 《Token 分类：混合潜在 Token 和文本 Token 以提升语言模型推理能力》。载于《第 29 届国际人工智能联合会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

苏巴贾, B., 泰, H. Y., 谭, A.-H. 《我是谁？：迈向智能体的社会自我意识》。《第 29 届国际人工智能联合会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

唐, L., 拉班, P., 和达雷特, G. 《MiniCheck: 基于基础文档对大语言模型进行高效事实核查》。载于阿尔 - 奥奈赞, Y., 班萨尔, M., 和陈, Y.-N. (编)，《2024 年自然语言处理实证方法会议论文集》，第 8818 – 8847 页，美国佛罗里达州迈阿密，2024 年 5 月 20–24 日。

陶泽, 林庭恩, 陈鑫, 李华, 吴宇, 李阳, 金泽, 黄峰, 陶大程, 周杰。大语言模型自进化研究综述。预印本网站，2024 年 5 月 20 日。

田凯、米切尔、周安、沙玛、拉法伊洛夫、姚浩、芬恩和曼宁。《只需请求校准：从经过人类反馈微调的语言模型中获取校准置信分数的策略》。载于布阿穆尔、皮诺和巴利 (编)，《2023 年自然语言处理实证方法会议论文集》，第 5433 – 5442 页，新加坡，2023 年 12 月。计算语言学协会。doi: 10.1162/tacl_a_01234

童, Y., 李, D., 王, S., 王, Y., 滕, F., 尚, J. 《大语言模型能否从以往的错误中学习？探究大语言模型的错误以提升推理能力》。《计算语言学协会第 62 届年会论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

托诺利尼, F., 阿莱塔斯, N., 马西亚, J., 以及卡扎伊, G. 《贝叶斯提示集成：黑盒大语言模型的模型融合》。载于《第 29 届国际人工智能联合会议论文集》，第 1–11 页，中国北京，2024 年 5 月 20–24 日。

计算语言学协会 ACL 2024 年会议论文集，第 12229 -

托伊, J., 麦克亚当, J., 以及塔博尔, P. 《元认知就是你所需要的一切? 在生成式智能体中运用内省来改善目标导向行为》。arXiv 预印本 arXiv:2401.10910, 2024 年。

泰恩, G., 曼苏尔, H., 陈, P., 马克, T., 以及卡布内, V. 《大语言模型无法发现推理错误, 但能纠正它们》。//计算语言学协会年会 2023 年

上里, J., 库什曼, N., 库马尔, R., 宋, F., 西格尔, N., 王, L., 克雷斯韦尔, A., 欧文, G., 希金斯, T. 《通过自我监督学习提升大语言模型的元推理能力》。//计算语言学协会年会 2023 年

范·泽伊, M. 和伊卡德, T. 《将意图重新考量视为元推理》, 载于 2015 年神经信息处理系统大会

瓦尔什尼, N., 姚, W., 张, H., 陈, J., 以及余, D. 小洞不补, 大洞吃苦: 通过验证低置信度生成来检测和纠正元推理错误。//计算语言学协会年会 2023 年

万泽, 冯雪, 温明, 麦卡利尔, S.M., 温阳, 张文, 王军。类似 AlphaZero 的树搜索可指导大语言模型解码与训练。发表于《第 41 届机器学习国际会议论文

万泽, 李阳, 宋阳, 王浩, 杨磊, 施密特, 王佳, 张伟, 胡帅, 文宇。Rema: 利用多智能体强化学习让大语言模型学会自我反省。//计算语言学协会年会 2023 年

王, B., 郑, R., 陈, L., 刘, Y., 窦, S., 黄, C., 沈, W., 金, S., 周, E., 石, C., 高, S., 徐, N., 周, Y., 范, X., 席, Z., 赵, J., 王, X., 季, T., 严, H., 沈, L., 陈, Z., 桂, T.,

王, M., 陈, L., 程, F., 廖, S., 张, X., 吴, B., 余, H., 徐, N., 张, L., 罗, R., 李, Y., 杨, M., 黄, F., 以及李, Y. 《一个文档都不落下: 用扩展多文档问答对长上下文大语言模型进行基准测试》。载于阿尔-奥奈赞, Y., 班萨尔, M., 以及陈, Y.-N. (编), 《2024 年自然语言处理实证方法

王鹏、李亮、邵泽、徐然、戴迪、李阳、陈迪、吴悠、隋政。《数学牧羊人: 无需人工标注逐步验证和强化大语言模型》。收录于古良维、马丁斯、斯里库马尔 (编), 《第 62 届计算语言学协会年会论文集 (第 1 卷: 长论文)》, 第 9426 – 9439 页, 泰国曼谷 2024 年 8 月 //计算语言学协会年会 2024

王, X., 魏, J., 舒尔曼斯, D., 乐, Q. V., 迟, E. H., 纳兰, S., 乔杜里, A., 以及周, D. 自一致性提升语言模型中的思维链推理能力。载于《第十一届学

王, X., 李, C., 王, Z., 白, F., 罗, H., 张, J., 乔吉奇, N., 邢, E., 以及胡, Z. 《Promptagent: 利用语言模型进行战略规划实现专家

王 X.、王 Z.、刘 J.、陈 Y.、袁 L.、彭 H. 和季 H. 《MINT: 在工具与语言反馈的多轮交互中评估大语言模型》, 载于《第十二届学习表征国际会议》, 2024

王 Y.、李 P.、孙 M. 和刘 Y. 《自我知识引导的大语言模型检索增强》, 收录于布阿莫尔 H.、皮诺 J. 和巴利 K. (编) 《计算语言学协会研究成果: 2023 年自然语言处理经验方法会议》, 第 10303 – 10315

王 Y.、廖 Y.、刘 H.、刘 H.、王 Y. 和王 Y. 《MM-SAP: 用于评估多模态大语言模型感知中自我意识的综合基准》, 载于《第 62 届计算语言学协会年会论文集 (第 1 卷: 长论文)》, 计算语言学协会,

王, Y., 张, Z., 张, P., 杨, B., 以及王, R. 《元推理: 大语言模型的语义 - 符号解构》。收录于古, L.-W., 马丁斯, A., 以及斯里库马尔, V. (编), 《计算语言学协会研究成果: ACL 2024》, 第 622 – 623

王, Y., 赵, S., 王, Z., 黄, H., 范, M., 张, Y., 王, Z., 王, H., 以及刘, T. 《战略思维链: 通过元推理提升大语言模型的推理能力》。//计算语言学协会年会 2024 年 8 月 //计算语言学协会年会 2024

吴, Z., 邱, L., 罗斯, A., 阿基尤雷克, E., 陈, B., 王, B., 金, N., 安德烈亚斯, J., 以及金, Y. 推理还是背诵? 通过反事实任务探索语言模型的能力与局限。

(编), 《2024 年北美计算语言学协会分会会议录: 人类翔, V., 斯内尔, C., 甘地, K., 阿尔巴拉科, A., 辛格, A., 布莱格登, C., 冯, D., 拉法伊洛夫, R., 内森·利尔, 马汉, D., 卡斯特里卡托, L., 弗兰肯, J.-P., 哈伯, N., 以及芬恩, C.。迈向大语言

向字、严浩、欧阳爽、桂亮、何宇。《Scireplicate-bench: 在基于智能体驱动的研究论文算法复现中对大语言模型的评估》。arXiv, 论文, 2024b 年。

谢, S. M., 拉古纳坦, A., 梁, P., 以及马, T. 《将上下文学习解释为隐式贝叶斯推理》。arXiv, 论文, 2024b 年。

谢, Y., 川口, K., 赵, Y., 赵, X., 菅, M., 何, J., 以及谢, Q.。自我评估引导的束搜索推理方法。

谢宇、阿维纳什·戈亚尔、郑伟、甘美玉、蒂莫西·P·利利克拉普、川口健、谢明。《蒙特卡洛树搜索通

熊, C., 齐, X., 陈, P.-Y., 以及何, T.-Y. 《防御性提示补丁: 针对越狱攻击的大语言模型的一种稳健

熊, M., 胡, Z., 陆, X., 李, Y., 傅, J., 何, J., 以及胡伊, B. 大语言模型能表达它们的不确定性吗? 对大语言模型中置信度诱导的实证评估。发表于

严, H., 向, Y., 陈, G., 王, Y., 桂, L., 和何, Y.。鼓励还是抑制单义性? 从特征去相关的角度重新

严浩、朱琦、王鑫、桂琳、何宇。《Mirror: 用于知识丰富推理的多视角自我反思方法》, 载于《第 62 届计算语言学协会年会论文集》, 2024b 年。

杨, L., 于, Z., 张, T., 曹, S., 徐, M., 张, W., 冈萨雷斯, J.E., 崔, B. 《思维缓冲: 利用大语

杨, L., 于, Z., 崔, B., 以及王, M. 《Reasonflux: 通过扩展思维模板进行分层大语言模型推理》。arXiv 预印本

杨, Y., 穆塔尔, D., 沈, Y., 詹, Y., 刘, J., 王, Y., 孙, H., 邓, W., 孙, F., 张, Q., 陈, W., 以及童, Y. 《Mtl-lora: 多任务学习的低秩适

姚, S., 陈, H., 杨, J., 以及纳拉辛汉, K.

《Webshop: 借助有基础的语言智能体实现可扩展的真实世界网络交互》。载于科耶霍, S., 穆罕默德, S., 阿加瓦尔, A., 贝尔格雷夫, D., 赵, K., 以及姚顺宇、余典、赵健宇、伊戈尔·沙夫兰、托马斯·L·格里菲思、曹越、克里希纳·R·纳拉辛汉。思维树: 利用大语言模型进行深思熟虑的问题解决。《第

姚顺, 赵健, 余东, 杜楠, 伊兰·沙夫兰, 科纳克里·纳拉辛汉, 曹宇。《React: 语言模型中推理与行

叶, J., 龚, S., 陈, L., 郑, L., 高, J., 施, H., 吴, C., 李, Z., 毕, W., 以及孔, L. 《思维扩散: 扩散语言模型中的思维链推理》。第三十八届神经信

约兰, O., 沃尔夫森, T., 博金, B., 卡茨, U., 多伊奇, D., 以及贝兰特, J. 通过对多条思维链进行元推理来回答问题。发表于《自然语言处理实证方法会

余, Q., 何, Z., 李, S., 周, X., 张, J., 徐, J., 和何, D. 通过循环对齐推理增强自回归思维链。

余, T., 林, T.-E., 吴, Y., 杨, M., 黄, F., 和李, Y. 《基于多样化反馈的大语言模型对齐构建》。

余文涛、张正、梁正、蒋梦、阿维纳什·萨巴尔瓦尔。通过即插即用检索反馈改讲语言模型。预印本网

尤克塞贡努尔, M., 比安奇, F., 博恩, J., 刘, S., 卢, P., 黄, Z., 格斯特林, C., 邹, J. 《通过

泽利克曼, E., 哈里克, G.R., 郜, Y., 贾亚西里, V., 哈伯, N., 古德曼, N. 《Quiet-STar: 语言模型

曾泽宇、刘一、万宇、李军、陈鹏、戴佳、姚瑶、徐然、齐志、赵伟等人。《Mr-ben: 大型语言模型的综合元推理基准》, 第三十八届神经信息处理系统年度

曾, Z., 于, J., 高, T., 孟, Y., 戈亚尔, T., 以及陈, D. 《评估大语言模型对指令遵循的评估能

曾, Z., 陈, P., 刘, S., 江, H., 贾, J. 《Mr-gsm8k: 用于大语言模型评估的元推理基准》, 发表于

詹, E. S., 莫利纳, M. D., 柳, M., 彭, W. 《有什么可害怕的? 从技术可供性视角理解对人工智能的多维恐惧》。《国际人机交互杂志》, 40 (22): 7127 -

张, D., 黄, X., 周, D., 李, Y., 以及欧阳, W. 通过蒙特卡洛树自优化与 Llama-3 80 亿参数模型获取

张凯、李泽、李军、李刚和金泽。《Self-edit: 面向代码生成的故障感知代码编辑器》, 发表于《计算语言学协会年

张凯、王迪、夏杰、王伟业和李磊。ALGO: 利用生成的预言机验证器合成算法程序。发表于《第三十七届

张, W., 沈, Y., 吴, L., 彭, Q., 王, J., 庄, Y., 卢, W. 《自我对比: 通过不一致的解决视角实现更好的反思》。收录于顾, L.-W., 马丁斯, A., 斯里库马尔, V. (编), 《第 62 届计算语言学协会年会论文集 (第 1 卷: 长论文)》, 第 3602 - 3622

张 X、杜 C、庞 T、刘 Q、高 W 和林 M。偏好链优化: 提升大语言模型中的思维链推理能力。发表于《第三十八届神经信息处理系统年度会议》, 2024c

张, Y., 迟, J., 阮, H., 乌帕萨尼, K., 比克尔, D. M., 韦斯顿, J., 以及史密斯, E. M. 《回溯提高

张, Z., 何, X., 严, W., 沈, A., 赵, C., 王, S., 沈, Y., 以及王, X.E. 《软思考: 在连续概念空

2025b 网址

张, Z., 金, C., 贾, M.Y., 和舒, T. 《Autotom: 用
于开放式心智理论的自动贝叶斯逆向规划和模型发

展》。发表于第 23 届自治个体与多智能体系统国际会议论文集, AAMAS '24。自治个体与多智

张, Z., 郑, C., 吴, Y., 张, B., 林, R., 余,
B., 刘, D., 周, J., 以及林, J. 《数学推理中开发

一个基于逆向规划的大语言模型》。发表于第 23 届计算

赵, X., 康, Z., 冯, A., 莱文, S., 以及宋, D.

《无需外部奖励的推理学习》。2025 年。网址

郑, C., 张, Z., 张, B., 林, R., 陆, K., 余,

B., 刘, D., 周, J., 以及林, J. 《Processbench:

识别数学推理中的过程错误》。载于《第 63 届计算

智轩, T., 英, L., 曼辛格卡, V., 以及特南鲍姆,

J. B. 通过合作语言引导的逆向规划实现务实指令跟

随与目标协助。发表于第 23 届自治个体与多智能体

系统国际会议论文集, AAMAS '24。自治个体与多智

能体系统国际会议论文集, AAMAS '24。自治个体与多智

朱, Q., 赵, R., 严, H., 何, Y., 陈, Y., 桂, L.

通过可控嵌入探索在大语言模型中导航解决方案空

间。发表于《大语言模型推理与规划研讨会》, 2025

年。网址 <https://openreview.net/forum?>

邹安邦、张卓、赵海、唐杰。《Meta-cot: 在混合任
务场景中使用大语言模型进行可泛化的思维链提

出》。发表于第 23 届自治个体与多智能体系统国际会议论文集, AAMAS '24。自治个体与多智

A. 附录：文献综述

本附录对现有研究进行了简要的文献综述，这些研究为我们的元推理框架的组成部分提供了深刻见解，包括认知科学和机器智能中的元推理、不确定性估计与校准、基于奖励模型的大语言模型推理，以及通过反馈优化大语言模型。

A.1. 元推理

元推理，即思考自身思维的过程，在人类智能和人工智能中都具有重要意义（考克斯和拉贾，2011 年；阿克曼和汤普森，2017 年）。

认知科学中的元推理 认知科学中的元推理涉及几个关键理论，这些理论解释了个体如何监控和调节自己的推理过程。双过程理论（卡尼曼，2011 年）表明，推理涉及直觉和审慎系统，元推理影响个体如何平衡这些系统。元认知推理理论（科利亚特，2000 年）强调了“知晓感”，它指导个体何时需要进一步付出认知努力。递减标准模型（阿克曼，2014 年）表明，信心会随着时间的推移而下降，导致个体接受信心较低的答案。可解性判断（劳特曼和阿克曼，2019 年）侧重于个体如何评估一个问题是否可解，这会影响他们坚持或决定放弃任务的行为。“错误感”（甘杰米等人，2015 年）探讨了人们如何察觉错误并相应地调整自己的推理。这些理论突出了元认知过程是如何控制认知努力并塑造决策的。基于纳尔逊（1990 年）提出的用于监控学习和记忆的开创性框架，阿克曼和汤普森（2017 年）提出了一个元推理框架，该框架由两个主要部分组成：元认知监控和元认知控制。元认知监控涉及对特定认知任务成功或失败可能性的主观评估，并指导关于行动、时间和精力分配的决策。元认知控制决定对认知任务投入的精神努力的启动、终止或调整。

决策与多智能体系统中的元推理

考克斯与拉贾（2011 年）从人工智能和认知科学的角度探讨了元推理，围绕一个核心模型展开，在该模型中，元推理控制并监测推理过程，指导关于何时行动或继续思考的决策。在人工智能研究中，元推理在搜索和规划领域得到了广泛研究。例如，林等人（2015 年）专注于一般的元推理决策问题，该问题涉及平衡规划成本与最终行动的质量，以实现智能体长期效用的最大化。他们在马尔可夫决策过程（MDP）中提出了近似算法，利用有界实时动态规划技术来评估进一步推理的计算价值，且不依赖于先验的特定领域数据。范·泽与伊卡德（2015 年）专注于元推理的一个方面，即随着新信息的出现重新考虑或调整计划。他们探索了在不断变化的条件下，智能体何时应该“思考”（重新规划）与何时应该“行动”的最优策略，表明灵活的元层次策略可以在各种不同环境中提升决策能力。埃尔博赫尔等人（2023 年）研究了情境时间规划，并提出了用于并发行动执行与审议的算法。元推理有不同的方面，例如在决策和多智能体系统中使用元层次控制（拉贾与莱瑟，2007 年；程等人，2013 年；朗格卢瓦等人，2020 年），通过内省监测实现自我提升（默多克与戈尔，2008 年；考克斯，2011 年；托伊等人，2024 年），认知智能体中的自我意识模型（萨姆索诺维奇等人，2008 年；普鲁斯特，2013 年；萨姆索诺维奇与德容，2013 年；查蒂拉等人，2018 年；苏巴贾等人，2021 年），以及使用元认知强化学习来理解如何思考（克鲁格等人，2017 年）。

语言模型的元推理 最近的工作对固定的大语言模型进行提示工程以实现所谓的元推理。Wang 等人（2024h）采用元提示在进行样本级推理之前识别特定任务。Yang 等人（2024）将来自多个实例及其解决方案的知识提炼成存储在信息库中的思维模板，在实例化推理之前将访问该模板。Wang 等人（2024g）将与推理无关的语义信息解构为通用符号形式，从而将各种问题转化为元形式。在算术、符号和逻辑推理以及多智能体心理博弈中观察到性能提升。一些研究声称，任务无关的指令有助于大语言模型学会思考。Wu 等人（2024b）通过输入填充有人工指令的思维模板，促使大语言模型在做出响应之前生成想法，然后使用自我生成的响应进行直接偏好优化（DPO）。这种经过深思熟虑的、对终端用户隐藏的想法，应该是大语言模型内部思维的一部分，使大语言模型在一般推理任务中表现得更好。

尽管取得了上述进展，但大语言模型中现有的元推理方法在捕捉人类认知的深度和复杂性方面仍然存在局限。机器元推理与人类智能之间仍存在差距。虽然

诸如元训练、提示工程和指令调优等方法，已在特定任务的泛化和结构化推理方面展现出有效性，但它们往往依赖于固定的模板或定义狭窄的优化框架。这些策略缺乏动态适应多样、非结构化上下文的能力，也无法内省式地评估自身的推理过程，而这些都是人类元推理的核心要素。例如，反思自身知识状态（“知道自己知道什么”）以及察觉知识差距或不确定性的能力（弗拉维尔，1979 年）。这还包括对错误的察觉（甘杰米等人，2015

A.2. 不确定性估计与校准

在人类的元推理过程中，人们首先会对给定任务的可解性做出初步判断，为知识和策略的评估做准备。同样，在大语言模型的元推理过程中，大语言模型也应对给定任务的可解性和难度进行初步判断评估，这可以通过估计大语言模型生成内容的置信度得分或不确定性得分来实现。更具体地说，有两种不确定性估计方法。不确定

实时不确定性估计 第一种方法是实时不确定性估计，即大语言模型同时生成输出及其不确定性（耿等人，2024 年）。基于语言的方法促使大语言模型用人类语言表达不确定性，这假设大语言模型能够很好地校准语言化的置信度，即大语言模型能够用数字表达式（如 0 – 1）或语言表达式（如肯定、可能、不可能）来表达其对输出的不确定性（田等人，2023 年；米尔克等人，2022 年）。基于对数几率的方法通过词元级熵来估计句子的不确定性（黄等人，2025 年）。为了融入语义，段等人（2023 年）引入了词元级相关性的概念，该概念通过使用语义

事后不确定性估计 第二类方法是事后不确定性估计，即在生成输出后对不确定性进行估计。基于一致性的方法假设，当大语言模型（LLM）对给定概念确定时，采样得到的响应可能相似且包含一致的事实，而对于幻觉事实，随机采样得到的响应可能会出现分歧，甚至相互矛盾。Manakul 等人（2023 年）提出基于一个输入采样多个生成结果，并计算目标与生成结果之间的相似度得分，然后将这些相似度得分聚合起来，作为对目标的不确定性度量。基于分布的方法将大语言模型的输出转换为嵌入，并根据嵌入的分布来估计输出的不确定性。Catak 和 Kuzlu（2024 年）提出了一种使用凸包分析进行不确定性量化的几何方法，该方法利用响应嵌入的空间特性来

不确定性校准 不确定性校准旨在使置信度得分与实际正确性保持一致，以提高预测的可靠性。基于监督的方法在包含正确和错误答案及其不确定性的数据集上对大语言模型进行微调，以提高模型估计不确定性的能力（Liu 等人，2024b；Kapoor 等人，2024）。基于提示的方法利用提示增强技术，如释义或选项排列，来创建集成，从

A.3. 基于奖励的大语言模型后训练

目前有两种在预训练后提升大语言模型推理能力的训练范式：监督微调（SFT），即模仿学习，以及强化学习（RL）。监督微调允许模型在有注释的推理链上进行微调，以学习推理模式。而强化学习作为当前最先进的方法，需要一个让大语言模型最大化的奖励。大语言模型通过自身探索来学习推理模式。

在大语言模型框架中，奖励建模的奖励模型通常分为两类：结果奖励模型（ORM）和过程奖励模型（PRM）。ORM 主要评估完整的输出，应用于整个输出的基于规则的奖励（尽管不是通过学习得到的）也可以被视为 ORM 的一种形式。相比之下，PRM 在推理（莱特曼等人，2024 年）和训练（王等人，2024c）中都显示出有效性。最

概率推理模型（PRMs）的一个显著局限性在于，其对思维链（CoT）路径的依赖需要高成本的人工标注。为了解决这一问

表 9 用于大语言模型基于强化学习的训练后阶段的反馈信号

2024c; 陈等人, 2024a)。此外, 由于大语言模型 (LLMs) 具有强大的上下文学习能力, “大语言模型即裁判”方法 (姚等人, 2023a; 张等人, 2024a) 已成为替代偏好排名模型 (PRMs) 的一种流行选择。最近, (张等人, 2025) 在“大语言模型即裁判”上提出了“上下文推理的判决”, 增加了对生成结果的反馈信号。

通过奖励进行搜索 一旦确定了合适的奖励, 就可以利用它们来控制大语言模型 (LLM) 的推理过程生成。一种简单直观的融入奖励的方法是 N 选优法 (莱特曼等人, 2024 年), 该方法从一组候选结果中选择得分最高的生成结果。策略推理模型 (PRMs) 进一步支持基本的树搜索算法, 如深度优先搜索 (DFS) 和广度优先搜索 (BFS) (姚等人, 2023a)。尽管由于深度优先搜索的探索能力有限, 它很少被使用, 但广度优先搜索经常被扩展为束对于计算要求更高的场景, 可以将蒙特卡洛树搜索 (MCTS) 与概率路线图法 (PRM) 结合使用 (Hao 等人, 2023a; Wan 等人, 2024)。有趣的是, 正如 (Snell 等人, 2025) 所示, 更复杂的 MCTS 往往不如更简单的束搜索, 而束搜索又仅在计算预算较低时优于 N 选 1 算法。这种违反直觉的趋势通常归因于奖励过度优化, 即方法被不完善的奖励信号误导 (Qin 等人, 2024)。尽管存在这些挑战, MCTS 在更具挑战性的问题上已显示出潜力。

使用奖励进行训练 奖励也可用于进一步提升策略模型的性能。诸如 LLaMA3 (格拉塔菲奥里等人, 2024 年) 和 Qwen2.5 (Qwen 等人, 2025 年) 等流行的指令微调模型遵循两阶段训练过程: 首先在带注释的问答数据集上进行微调, 然后通过近端策略优化 (DPO) 使用偏好数据进行优化。主要区别在于在线强化学习阶段 ——Qwen2.5

像 AlphaMath、Mathshepherd 和偏好优化链这样的模型利用概率关系模型 (PRMs) 为中间推理步骤赋值, 尽管它们的训练策略各不相同。实证证据表明, 强化学习通常优于简单的拒绝采样。

最近的一项突破, Deepseek-R1-Zero (DeepSeek-AI 等, 2025 年), 通过完全跳过监督微调 (SFT) 阶段, 取得了最先进的 (SOTA) 性能。相反, 它使用一种纯强化学习方法 (GRPO) 从头开始训练模型, 仅使用两个简单的奖励信号: 一个正确性奖励, 用于检查最终答案是否正确; 以及一个格式奖励, 用于确保答案以正确的格式

然而, Deepseek-R1-Zero 生成的推理链可读性往往较差, 因为奖励机制并未明确鼓励清晰性。为解决这一问题, 作者们还发布了 Deepseek-R1 (DeepSeek-AI 等人, 2025 年), 该模型在强化学习之前增加了一个微调阶段, 最后, EBRM 探索了使用隐式奖励进行训练, 展现出更强的稳健性和泛化能力。这种方法为利用超越人类可解释

A.4. 推理过程中的大语言模型优化

利用反馈优化大语言模型 (LLMs) 的推理过程, 已成为一种有前景的方法, 可提升其在各种任务中的性能、可靠性和适应性。反馈通过提供关于错误和需要调整之处的明确信息, 实现推理能力的迭代改进。

反馈类型 现有的反馈分为两类：标量反馈和语言反馈。标量反馈可以是布尔值（例如整数 0 或 1），或者一致性分数（例如十进制概率）（Wang 等人，2023a; Fu 等人，2023a; Pan 等人，2024），据观察，它与正确性高度相关（Yao 等人，2023a; Li 等人，2023; Yan 等人，2024b）。语言反馈提供更多信息且具有可解释性。生成反馈最直接的方法是促使大语言模型评估其当前推理过程（Lu 等人，2023; Liang 等人，2024）。然而，最近的研究（Huang 等人，2024b; Li 等人，2024c; Yan 等人，2024b）发现，由于大语言模型在知识基础方面存在局限性，它们往往无法提供可靠的反馈。于是，研究人员检索外部知识或整合外部工具（如代码编译器），以

表 3. 用于优化大语言模型推理的反馈技术分类

利用反馈优化推理过程 为了更新原始推理过程，我们可以直接将反馈作为提示提供给大语言模型（Madaan 等人，2023 年; Yan 等人，2024b; Bo 等人，2024 年），训练一个独立的评判模型，引入显式回溯标记，并利

(i) 由于许多反馈是基于模板的（Huang 等人，2024b; Tyen 等人，2023），它们通常只能提供有限的错误信息。Wang 等人（2024d）随后提出了 PromptAgent，以根据反馈动态优化初始提示模板。具体来说，使用基础大语言模型从样本中收集错误，然后使用优化后的大语言模型（通常优于基础大语言模型）提供错误反馈并重新初始化提示。此外，多智能体框架能够为复杂推理任务提供多方面的反馈，例如科学论证（Yianor 等

(ii) 许多特定任务的评判模型是为数学（Chen 等人，2024b）、编码（Kumar 等人，2024）、问答和逻辑推理（Tong 等人，2024）任务中的错误纠正而训练的。为了创建评判模型的训练数据，他们首先从基础模型中采样大语言模型生成的推理过程，并让人类注释者对错误位置（哪个推理步骤）和错误类型（例如知识错误或

(iii) 一些研究放弃了单独的批评模型，而是引入了特殊标记，如 [重置]，以触发明确的错误纠正过程

(iv) 像代码解释器这样的工具使大语言模型能够根据编译器结果验证推理并优化输出（Feng 等人，2025a; Qian 等人，2025），应用于数学和逻辑推理任务（Yao 等人，2023b; Zhang 等人，2023a; Chen 等人，2024c）。搜索引擎使大语言模型能够检索最新证据，以确保事实的正确性和有效性（Varshney 等人，2023; Yu 等人，2023h; Peng 等人，2023; Tin 等人，2025; Song 等人，2025）。逻辑和关系拓扑分析器有助于在