arXiv:2601.00513v1 [cs.LG] 1 Jan 2026

小型模型因错误理由得出正确答案：可信智能体的过程验证

Laksh Advani

独立研究人员，美国华盛顿州西雅图，
laad8452@colorado.edu

摘要

将小型语言模型（7–9B 参数）部署为自主智能体，需要信任它们的推理，而不仅仅是其输出。我们揭示了一场严重的可靠性危机：这些模型给出的正确答案中，有 50–69% 包含根本性的有缺陷的推理——这种“因错误理由得出正确答案”现象无法通过标准准确率指标察觉。通过分析三个模型在各类任务中的 10,734 条推理轨迹，我们提出了推理完整性评分（RIS），这是一种基于过程的指标，并经验证具有显著的评定者间一致性（$\kappa = 0.657$）。我们的发现对传统做法提出了挑战：虽然检索增强生成（RAG）显著提升了推理完整性（Cohen 的 $d = 0.23$–0.93），但在所评估的任务中，自我批评等元认知干预往往会损害小型模型的表现（$d = -0.14$ 至 -0.33）。机制分析表明，RAG 通过以外部证据为计算提供依据而取得成效，使错误减少了 7.6%；而元认知在模型容量不足时会加剧困惑。为支持部署，我们将验证能力提炼到一个神经分类器中，其 F1分数达到 0.86，速度提升 100×。这些结果凸显了基于过程的验证对于可信智能体的必要性：当模型可能完全因错误理由得出正确答案时，仅凭准确率远远不够，甚至会带来危险。

引言

由小型语言模型（7–9B 参数）驱动的自主智能体有望让人工智能部署惠及大众：它们可在消费级硬件上运行，在毫秒内响应，并以极低的边际成本运作。然而，一个根本性问题威胁着这一愿景：我们能相信它们的推理吗？我们提出的证据表明，一场严重的可靠性危机正在显现：即使给出正确的输出，这些模型在 50–69% 的情况下仍会表现出根本性的推理缺陷，这种现象我们称为“因错误理由得出正确答案”（RWR）。

设想一个负责金融计算的智能体。对于“80 的 15% 是多少？”，它回答：*“第 1 步：要计算 80 的 15%，我将 80 乘以 0.2。第 2 步：* 80×0*。2 = 12。答案：12。”* 虽然答案正确，但其推理在数学上是错误的——使用了 0.2 而不是 0.15。在自主运行中，这类隐性失败会造成灾难性的累积后果：智能体可能会批准交易、提出医疗建议，或基于碰巧正确但存在根本性缺陷的逻辑来控制系统。

版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

这一问题对于将驱动大多数已部署智能体的小型模型尤为突出。与只能在集中式服务器上运行的前沿模型不同，拥有 7–9B 参数的模型能够实现边缘部署、隐私保护和高性价比扩展。然而，它们的容量有限，容易出现准确率指标无法检测到的推理失误。当前的评估范式只评判最终输出，造成了一个危险的盲点：这些智能体在基准数据集中看似可靠，部署后却会不可预测地失效。

本文对小型语言模型的推理完整性进行了大规模研究，将以往的诊断扩展到智能体场景，分析了三个模型（Llama-3-8B、Mistral-7B、Qwen- 2.5-7B）在三类不同任务（数学推理、多跳问答和常识推理）中的 10,734 条推理轨迹。我们探讨三个关键问题：**（RQ1）**小型模型作为智能体部署时，隐性推理失败问题有多严重？；**（RQ2）**哪些干预措施能够提升推理完整性——自我批评和检索增强等常见方法是否有帮助？；以及**（RQ3）**哪些机制可以解释干预措施为何成功或失败，又如何高效地检测失败?

我们的分析表明，理解干预措施为何 成功或失败，比了解它们做了什么 更为重要。我们发现，尽管**检索增强（RAG）**能显著提升推理完整性（$d = 0.23$–0.93），主要是通过借助外部证据减少计算错误，但**元认知提示**却持续损害性能（$d = -0.14$ 至 -0.33），因为它会加剧内部混乱。我们的贡献包括提出**推理完整性评分（RIS）**（$\kappa = 0.657$）、基于 10,734 条推理轨迹提供隐性推理缺陷占比达 50–69% 的大规模证据，以及一种**可部署的验证器**（F1分数为 0.86，提速 100×），可用于实时可信度评估。

这些结果对部署可信智能体具有直接意义。我们提供了可操作的指导：对于事实依据任务，在检索可行时优先为小型模型采用检索增强生成；对于知识密集型任务，避免在参数量低于 10B 的模型中使用元认知提示；并将基于过程的验证作为不可或缺的安全层。

相关工作

过程与结果评估。 近期研究认识到，基于结果的评估会掩盖推理失败（Lightman 等，2023；Yuan 等，2025）。然而，这些研究关注的是训练 改进，而非检测已部署模型中的隐性失败，也没有量化我们所揭示的因错误理由得出正确答案现象的普遍程度。

小模型可靠性。 参数量不足 100 亿的模型所面临的挑战已有充分记载（Magesh 等，2024年；Kalai 等，2025年；Cheng 等，2024年），包括幻觉和事实错误。然而，先前的研究缺乏对隐性失败的系统性测量——即模型通过有缺陷的推理得出正确答案的情况。我们发现，这类情况占“成功”案例的 50%–69%。

干预策略。虽然检索增强生成（Wang 等，2025年；Chen、Myrzakhan 和 Lee，2025年）和元认知技术（如自我批评）（Gou 等，2023年；Yang 等，2024b）较为常见，但此前鲜有研究系统比较它们对推理完整性而非准确率的影响。我们提供的证据表明，元认知可能会主动损害推理质量（$d = -0.14$ 到 -0.33），而检索增强生成则始终能带来帮助（$d = 0.23$–0。93）。

研究空白。尽管小型模型作为智能体的部署日益增多，但很少有大规模研究量化隐藏的推理失败、系统比较干预对过程质量的影响，或解释其作用机制。我们对 10,734 条推理轨迹的分析填补了这一关键研究空白，助力可信智能体部署。

方法论

为系统研究小型语言模型中的因错误原因得出正确结果（RWR）现象，我们开展了一项大规模实证研究。我们的方法论涵盖数据集选择、模型选择、推理轨迹生成、干预实施、基于 RIS 的推理评估、错误分类和统计验证。所有推理轨迹生成实验均通过 OpenRouter API 开展，而推理轨迹分析和蒸馏模型训练则在本地进行。

数据集与模型

我们选取了三个不同类型的基准数据集：**GSM8K** （Cobbe 等，2021）（1,319 道数学文字题）、**Hot-potQA**（Yang 等，2018）（1,000 个多跳问答样本），以及**ARC**（Clark 等，2018）（1,119 个常识科学问题）。我们对这些数据集进行了子采样，以兼顾计算可行性与多样性。

我们评估了三种广受欢迎的开源小模型：**Llama-3-8B**（Dubey 等，2024）、**Mistral-7B** （Jiang 等，2023）和 **Qwen-2.5-7B** （Yang 等，2024a）。模型均采用其基础指令微调变体，并使用贪心解码（温度=0）。

推理轨迹生成

对于每个模型—数据集对，我们在四种条件下生成了推理轨迹（基线 + 三种干预），共得到 10,734 条轨迹（3 个模型 $\times$ 3 个数据集，每个数据集每种条件约有 298 个样本）。这些轨迹被提示以结构化格式生成逐步推理：“步骤 1： [推理] 步骤 2： [推理] .....最终答案： [输出]”，该格式改编自标准思维链模板（Wei 等人，2022）。

干预措施

我们实施了三种轻量级干预措施：**(1)检索增强生成（RAG）**，通过提示词“使用提供的上下文逐步推理。”提供预言式真实上下文（例如，HotpotQA 的维基百科片段）；**(2) 自我批评**，提示模型审查其推理：“批评你之前的推理，找出错误；如有必要，请提供修正后的版本。”；**(3) 验证提示**，在初始提示中加入：“继续下一步之前，请验证每一步是否准确。”

推理完整性评分（RIS）

RIS通过对每个推理步骤按0.0–1.0的量表评分来衡量过程质量：1.0（完全正确）、0.5（存在部分缺陷）、0.0（错误）。步骤通过正则表达式解析提取。对于每条推理轨迹，RIS =为各步骤评分的平均值。评分由三位相互独立的大语言模型评审（GPT-4o-mini、Claude-3.5-Sonnet、Gemini-1.5-Flash）依据详细的评分标准完成。我们在500个步骤上验证了评审者间的一致性（Fleiss’ $\kappa = 0.657$，一致性良好）。最终RIS采用多数投票确定。若 $RIS < 0.8$，则将一条推理轨迹归类为“有缺陷”；该阈值通过敏感性分析确定。我们测试了0.7至0.9之间的阈值，并选取0.8，因为它在灵敏度（检测有缺陷的推理）和精确率（避免误报）之间实现了最佳平衡。

错误分析

为揭示其机制，我们手动将 1,000 个有缺陷的步骤归为四类：**计算错误**（算术、数字或事实应用错误）、**幻觉**（捏造信息）、**逻辑跳跃**（无效推理）或**其他**。我们按条件计算各类错误的分布，并与基线比较。我们还测量了错误位置（归一化至 0–1）、上下文误用（检索到的事实被错误应用的比例），以及使用 Pearson 的 r 计算相关性。

蒸馏验证系统

为支持高效部署，我们训练了一个轻量级多层感知机分类器，用于预测有缺陷的轨迹（$RIS < 0.8$），采用的混合特征包括 Sentence-BERT 嵌入（来自 all-MiniLM-L6-v2 的 384 维向量） + 7 结构指标（例如步骤数、轨迹长度）。模型（391 个输入特征、5 层、$\sim$300k 个参数）使用 80% 的轨迹进行训练（分层划分），采用焦点损失（$\gamma = 2.0$、 $\alpha = 0.25$）、AdamW（$lr = 5 \times 10^{-4}$）和早停。在 20% 留出的测试数据上评估，模型的宏平均 F1 为 0.86，“有缺陷”类别的精确率为 0.88；与 LLM 判定相比，速度提升 $\sim$100$\times$ 倍（在中央处理器上推理需 5–10 毫秒）。

统计分析

统计功效是在事后计算的，不同干预类型之间差异显著：检索增强生成效应的统计功效极佳（0.95–1.00），自我批判效应的统计功效良好

（0.76–0.99），验证效应的统计功效则参差不齐（0.56–0.96）。我们只报告达到足够统计功效的常规阈值 0.75 的发现，不过部分验证效应已接近这一阈值。

结果

我们对 10,734 条推理轨迹的分析揭示了小型语言模型中普遍存在的隐性失败，也发现了干预有效性的清晰规律。以下是我们的主要实证发现，均经统计分析验证。

隐性推理失败的发生率

如表 1 所示，理由错误但答案正确问题十分严重：50–69% 的正确最终答案存在有缺陷的推理（RIS < 0.8）。这一问题因模型和任务而异。尽管 Qwen-2.5-7B 相对较强，其平均发生率仍为最高（69.3%），这可能是因为其推理链更冗长，从而增加了出错的机会。HotpotQA 中的失败最为突出（平均 67.9%），这表明知识密集型任务会加剧模型对虚假模式而非稳健逻辑的依赖。

表 1：各模型和任务中存在有缺陷推理的正确输出所占百分比（RIS < 0.8）。

模型	ARC	GSM8K	HotpotQA	Avg
Mistral-7B	45.8%	44.3%	60.5%	50.2%
Llama-3-8B	47.0%	59.2%	59.4%	55.2%
Qwen-2.5-7B	61.4%	62.7%	83.8%	**69.3%**
平均值	51.4%	55.4%	67.9%	58.2%

干预效应

图 1 总结了干预措施对 RIS 的影响。检索增强生成（RAG）始终能提升推理完整性，在事实依据任务上的效应量为中到大（平均值为 $d = 0.41$，在 HotpotQA 上对 Qwen 的提升最高达 0.93）。相比之下，自我批评和验证提示在 78% 的条件下会损害性能（平均值分别为 $d = -0.14$ 和 -0.15），且对 Mistral 和 Llama 等较弱模型的负面影响最大。

任务依赖性显而易见：RAG 对纯推理（ARC，$d \approx 0$）的影响微乎其微，但对数学和问答（GSM8K $d = 0.23$–0.43；HotpotQA$d = 0.51$–0.93）有显著益处。元认知损害始终存在，这表明有一个能力阈值，超过该阈值后自我反思便会失效。

错误机制

表 2 展示了错误类型转移。基线中计算错误占主导（60.3%），而 RAG 最有效地减少了这类错误（−7.6%），同时增加了幻觉（+4.5%）和逻辑跳跃（+3.3%）。这种权衡带来了正向 RIS，因为“推理尝试”（得分为 0.5）代表部分得分，显著高于彻底失败——
干预措施（得分为 0.0）。基础计算

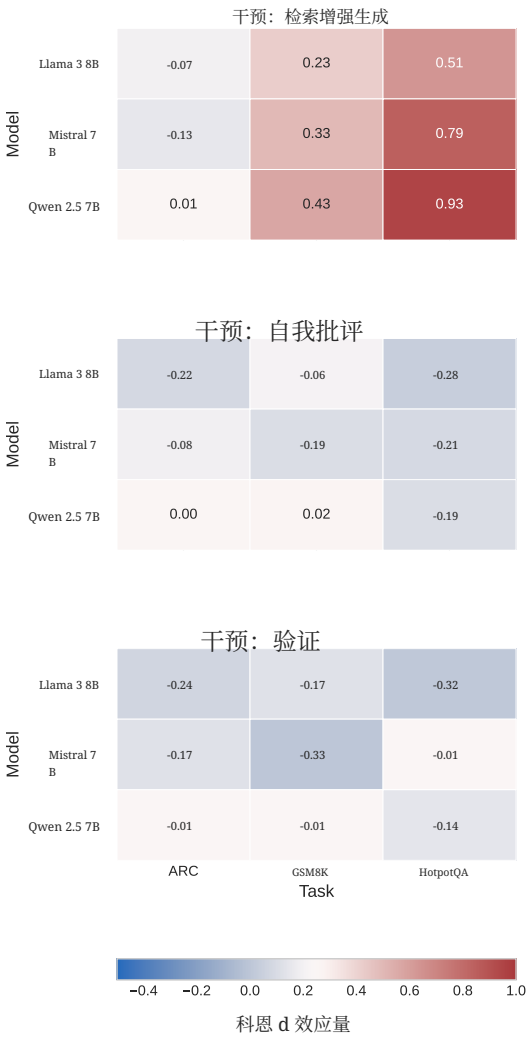


图 1：干预措施的 Cohen d 效应量。红色表示推理完整性有所改善（正向 d），蓝色表示有害影响（负向 d）。

错误（完全不正确）的减少，超过了带有部分缺陷的推理尝试的增加。元认知干预带来的减少幅度较小（−4.2%），而增加幅度相近，因此总体上弊大于利。

补充分析显示，上下文误用可强烈预测检索增强生成失败（$r = -0.951$），较弱的基线模型从检索增强生成中获益更多（$r = 0.671$），错误会在推理过程后期累积（平均位置=0.56–0.71）。所有相关性均具有统计显著性（$p < 0.001$）。

蒸馏验证器性能

该多层感知机分类器在留出数据上达到 0.86 的宏平均 F1，在“有缺陷”类别上的精确率为 0.88，召回率为 0.87。这一表现加上较低的延迟（~5–10 毫秒），使其适用于自主智能体的实时生产告警。

表 2：相对于基线的错误分布变化（百分点）。

错误类型	基线	+RAG	+自我批判	+验证
计算错误	60.3%	52.7%	56.1%	56.1%
变化	–	↓7.6	↓4.2	↓4.2
幻觉	25.2%	29.7%	27.2%	27.9%
变化	–	↑4.5	↑2.0	↑2.7
逻辑跳跃	14.3%	17.6%	16.7%	16.0%
变化	–	↑3.3	↑2.4	↑1.7

讨论

我们的实证结果揭示了小型语言模型中的一个关键问题，同时也指明了清晰的改进方向。这些发现挑战了关于智能体可靠性的核心假设，表明哪些 干预措施有效，不如它们为何 有效来得重要。

元认知为何会失败：伪反思

最引人注目的发现是，元认知提示所产生的影响始终有害。我们的分析表明，这不仅仅是中性的失败，而是主动引入了新的错误。我们认为，这是由于“伪反思”：小型模型缺乏进行内省所需的真正高层次元认知能力。当受到“批评”或“验证”的提示时，它们并不会进行 反思；它们会生成看起来像反思的文本。

这种伪反思会通过编造错误的论证来放大错误。模型缺乏可供核对的内部“真实答案”，因此会编造听起来合理（但实际上错误）的论证，正如权衡分析（表 2）所示。尽管元认知干预有时会减少计算错误（-4.2%），但同时也会增加幻觉和逻辑跳跃，最终对推理完整性产生净负面影响（均值 $d \approx -0.15$）。这表明，小型模型能够识别并纠正某些错误，但缺乏在批评过程中避免引入新错误的能力，从而支持了这样一种观点：有效自我反思存在一个“能力阈值”，而 7–9B 模型低于这一阈值。

RAG为何成功：外部支架

RAG的成功（均值 $d = 0.41$）或许可以理解为提供了“外部支架”。弱基线表现与RAG带来的高收益之间存在很强的相关性（r =0.671），这表明RAG可能发挥着认知矫具的作用，或许能够弥补模型内部知识和推理能力的不足。

错误位置数据支持了这一分析。错误会在推理轨迹的后期不断累积（平均位置=0.56–0.71），此时模型的内部状态会逐渐偏离原始事实。检索增强生成提供了一个恒定的外部锚点，在每一步重新将模型锚定，从而防止这种偏移。上下文误用与检索增强生成有效性之间近乎完美的负相关（$r = -0.951$）证实了这一点：检索增强生成的益处几乎完全取决于模型能否正确整合这一外部支架。需要注意的是，我们的研究采用了神谕式检索，这为检索增强生成的有效性提供了上限。使用噪声检索器的现实世界检索增强生成系统可能表现出较小的收益。

对智能体信任的启示

50–69%的RWR率（表1）表明，仅依据输出准确率来评估可靠性是极其危险且不足的。这要求转向“持续审计”；为此，我们蒸馏出的验证器（F1分数为0.86，推理耗时5–10毫秒）可作为实用的“信任警报”，实时标记高风险、有缺陷的推理链，供人工审查；而缓慢的基于大语言模型评审的评估则无法做到这一点。

局限性

我们承认本研究存在若干局限：研究采用了**预言机检索增强生成**，代表RAG收益的最佳情况上限；关于元认知在7–9B**模型** 上失效的结论，未必适用于更大的模型（例如70B+）；我们的**RIS指标**对各步骤的分数取平均，可能无法捕捉整体性失效；所有实验均以**英语**进行，并使用了**大语言模型评审** ，这些评审可能存在偏差；此外，我们的发现基于三个特定模型和三个任务领域，未必能推广到所有小型模型或任务。

未来工作

未来工作将使用含噪的“真实世界”RAG检索器验证这些发现，确定元认知干预可能变得有效的“能力阈值”（例如，在40B-70B+模型中），并改进蒸馏后的验证器，还可能使用基于图的网络将推理轨迹建模为依赖图。

结论

本研究量化了小型语言模型智能体中一种关键且隐蔽的失效模式：其正确答案中有 50-69% 属于“因错误理由得出正确答案”，源于根本上有缺陷的推理。我们表明，这一信任鸿沟无法通过准确率指标显现。我们的 10,734 条推理轨迹分析为缓解这一风险提供了清晰且可操作的路径：检索增强生成（RAG）充当必要的认知支架，能够稳健地提升推理完整性（d=0.23-0.93）。相反，我们提供了有力证据表明，诸如自我批评之类的常见“最佳实践”实际上有害（d=-0.14 至 -0.33）**当应用于所评估领域中的小型模型时**，会导致“伪反思”，从而放大错误。

我们贡献了经验证的基于过程的指标——推理完整性评分（RIS），以及一个快速、高精确率的蒸馏分类器（F1分数为 0.86），以便大规模部署这项验证。为了推动该领域迈向可信赖的自主智能体，我们应考虑转变评估范式。我们的研究结果表明，仅凭准确率并不足够；对于部署在需要事实知识或多步推理的任务上的小型语言模型，基于过程的验证或许需要成为必不可少的安全层。

参考文献

Chen, S.S.; Myrzakhan,A.; andLee, K. 2025年。DRAG：从大语言模型中蒸馏检索增强生成，以便小型语言模型跨领域迁移知识。*arXiv* 预印本 *arXiv:2506.01954*。

Cheng, X.; Li, S.; Wen, K.; Wang, H.; Zhou, Y.; Tang, L.; Li, C.; and Wang, J. 2024年。《小型智能体也能大放异彩！赋能小型语言模型成为幻觉检测器》。*arXiv*预印本 *arXiv:2406.11277*。

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018年。《你以为自己解决了问答问题?试试 ARC：AI2 推理挑战》。*arXiv* 预印本 *arXiv:1803.05457*。

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021年。《训练验证器来解决数学文字题》。*arXiv* 预印本 *arXiv:2110.14168*。

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.;Letman,A.; Mathur,H.;Bedoui, A.;Zhang,A.; 等. 2024年. Llama 3 模型家族 . *arXiv* 预印本 *arXiv:2407.21783*。

Gou, Z.; Shao, Z.; Gong, Y.; Yang, Y.; Huang, M.; Duan, N.; and Chen, W. 2023年. CRITIC：大语言模型可通过工具交互式批评进行自我纠正. *arXiv* 预印本 *arXiv:2305.11738*。

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot,D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; 等. 2023年. Mistral 7B. *arXiv* 预印本 *arXiv:2310.06825*。

Kalai, A. T.; Nachum, O.; Vempala, S. S.; 和 Zhang, E. 2025年。为什么语言模型会产生幻觉。*arXiv* 预印本 *arXiv:2509.04664*。

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards,H.; Branwen, G.; Biles, J.; Sutskever, I.; Schulman, J.; 和 Liang, P. 2023年。让我们逐步验证。*arXiv* 预印本 *arXiv:2305.20050*。

Magesh, V.; Surani, F.; Dahl, M.; Suzgun, M.; Manning, C. D.; 和 Ho, D. E. 2024年。无幻觉？评估领先的人工智能法律研究工具的可靠性。*arXiv* 预印本 *arXiv:2405.20362*。

Wang, L.; Chu, Z.; Chen, X.; Li, Y.-F.; Tran, Q. H.; Knop- man, D.; Zhang, W.; He, Q.; Qiu, X.; Guo, Q.; 和 Wang, Y. 2025年。MiniRAG: 迈向极简的检索增强生成。*arXiv* 预印本*arXiv:2501.06713*。

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; 等. 2022年. 思维链提示能够激发大语言模型的推理能力。神经信息处理系统进展, 35: 24824–24837. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; 等. 2024a. Qwen2 技术报告。*arXiv* 预印本 *arXiv:2407.10671*。
Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhut- dinov, R.; and Manning, C. D. 2018年. HotpotQA：一个用于多样化、可解释的多跳问答的数据集。载于*2018*年自然语言处理实证方法会议论文集, 2369–2380。

Yang, Z.; Zhang, Y.; Wang, Y.; Xu, Z.; Lin, J.; Sui, Z.; and Liu, T. 2024b. 置信度与批评：大语言模型自我纠错能力的分解。*arXiv* 预印本 *arXiv:2412.19513*。

Yuan, X.; Zhang, X.; Xu, K.; Xu, Y.; Yu, L.; Wang, J.; Dong,Y.; and Wang,H. 2025年。通过策略博弈追踪 TracingLLM 的推理过程：一种用于规划、修订和资源受限决策的框架。arXiv:2506.12012。

补充材料

A. 生成提示与干预措施

为确保可复现性，我们提供了所有生成任务所使用的确切系统提示和用户提示。基础系统提示会根据干预类型动态调整。

- **基础系统提示：**

  ”请逐步解决用户的请求。对于数学问题，请将最终答案放在方括号中，[例如这样]。对于选择题，请将最终答案（例如，[A] 或 [1]）放在方括号中。”

- **干预：基于提示的验证**
  预先 d 至系统提示：

  ”继续之前先验证每一步。 [基础系统提示]”

- **干预：检索增强生成（RAG）**对用户提示的修改：

  ”上下文：{检索到的上下文} {用户问题}”

  注：上下文取自 *GSM8K/ARC* 的真实答案分解，以及 *HotpotQA* 的支持句。

- **干预：自我批判**附加到系统提示：

  ”[基础系统提示] 解答后，请检查推理过程是否存在任何缺陷。”

B. 评判器与验证器提示

我们使用了一个强大的评判模型（DeepSeek-V3/Gemini-Flash）来评估推理完整性并对错误类型进行分类。

1. 推理完整性评分（RIS）二元评判器 用于判断在给定上下文 C的情况下，特定步骤 S_t 是否有效。

”你是一名严格验证器。你的任务是确定“生成步骤”是否在逻辑和事实层面得到“上下文”的支持。上下文：{context} 生成步骤：{step} “生成步骤”是否得到“上下文”完整且正确的支持？仅回答“Yes”或“No”。”

2. 失败模式分类 用于对特定步骤失败的原因进行分类。

”你是一名错误分析员。’生成步骤’被判定为有缺陷（不正确）。请根据’上下文’，对’生成步骤’中的主要错误进行分类。类别： [l。事实错误、2. 逻辑跳跃、3. 数值计算错误、4. 其他] 上下文： {context} 生成步骤： {step} 只输出类别名称。”

3. 检索增强生成误用分类 用于检测模型是否忽略了检索到的上下文，或基于检索到的上下文产生幻觉。

”你是一名错误分析员。请判断’生成步骤’是否误用了’上下文’。误用：引用了上下文，但使用不当（例如，逻辑错误、错误引用、误解）。正确：正确使用了上下文。无关：完全没有使用上下文。仅回复’误用’、’正确’或’无关’。”

C. 实施细节

- **生成器模型：**
  – mistralai/Mistral-7B-Instruct-v0.2–
  meta-llama/Llama-3-8B-Instruct–
  qwen/qwen-2.5-7b-instruct
- **数据集：** GSM8K（数学）、HotpotQA（多跳问答）、ARC-Challenge（推理）。
- **采样：** 温度设为默认值（因提供商而异，通常为 0.7-1.0），并采用标准核采样。
- **评判模型：**使用 deepseek/deepseek-v3.1-terminus 和 google/gemini-2.5-flash-lite 进行自动评估。

•蒸馏验证器： 一个经过训练的 4 层多层感知机（512-256-128-1），其输入为句子嵌入（all-MiniLM-L6-v2）与冗长度特征（步骤数、轨迹长度）的拼接。