通过资格推理和分区感知强化学习可靠地使用引理

Zhikun Xu[1]*余晓东[2],周本[1],刘江[2],吴佳莲[2],王泽[2], 孙熙萌[2],陈浩[2],刘子成[2]

[1]亚利桑那州立大学, [2]超威半导体公司 zhikunxu@asu.edu

arXiv:2602.00998v1 [cs.CL] 1 Feb 2026

摘要

近期的大语言模型（LLM）在数学基准测试中表现出色，但经常误用引理，在未经验证假设的情况下直接套用结论。我们将引理判断形式化为一项结构化预测任务：给定一个陈述和一个候选引理，模型必须输出前提检查和结论效用检查，并据此得出有用性判定。我们提出 RULES，通过双章节输出对这一规范进行编码，并结合强化学习与分区感知损失掩码进行训练，从而对导致错误的章节施加惩罚。训练和评估采用多样化的自然语言与形式化证明语料库；通过留出扰动套件评估鲁棒性；端到端评估则涵盖不同大语言模型在竞赛风格、与扰动对齐及基于定理的问题上的表现。结果显示，相较于基础模型和单标签强化学习基线，模型在域内表现均有稳定提升；在破坏适用性的扰动下，提升幅度更大；在端到端任务上则持平或略有提升。消融实验表明，双章节输出和分区感知强化对于实现鲁棒性都不可或缺。

1 引言

近期的大语言模型（LLM）在数学基准测试中取得了优异成绩，并且往往能够生成类似证明的文本（Welleck 等人，2022年；OpenAI 等人，2024年；Yang 等人，2024年；Shao 等人，2024年；Li 等人，2025b；Guo 等人，2025年）。然而，它们的行为往往体现的是模式模仿，而非有原则的推理：模型倾向于复现熟悉的解题路径，也可能在未验证引理假设的情况下直接援引其结论。这种行为利用了训练分布中的表面相关性，削弱了数学推理的可信度。

受人类数学家的需求启发（Frieder 等人，2024年），我们将引理判断建模为结构化预测问题：给定一个陈述—引理对，模型必须输出： 前提检查（验证引理的假设是否在当前语境中成立），以及 结论效用检查（评估该引理是否确实有助于证明目标），并据此得出有用的判断。我们的假设是，要求模型明确给出资格推理，并按章节对齐信用分配，将减少模型对虚假表面线索的依赖，提高其应对破坏适用性的扰动时的稳健性与安全性，并能迁移至端到端任务。这一思路与标准证明实践相呼应，即先检查假设，再使用引理；它也解释了为何单标签监督会掩盖失败的具体环节。

我们提出一种训练框架，RULES （通过资格推理和章节感知强化学习实现引理的可靠使用）。该训练基于组相对策略优化（GRPO）（Shao 等人，2024年），并采用分节感知损失掩码机制，将惩罚分配给导致失败的具体章节，而非整个推演。该设计使监督与任务所要求的认知分解相一致，鼓励模型学习先检查后使用的规范，而不是过拟合于标签层面的捷径。

在多种模型和不同评测数据集上的实证结果表明，与基础模型和单标签强化学习基线相比，RULES 提升了域内引理判断的表现。在留出的、会破坏适用性的扰动测试集上，优势更大，最多达 23 个百分点，表明其对引理使用的门控效果更好。在端到端基准测试中，其平均表现总体优于或持平于单标签强化学习基线。消融实验表明，双章节输出和章节感知强化学习二者缺一不可：将细粒度的

*工作完成于在 AMD 的研究实习期间。

检查监督与训练时的扰动监督混为一谈，或排除训练时的扰动监督，都会大幅降低资格推理的鲁棒性。

2 相关工作

许多研究探索了在数学语料库上进行预训练，并结合思维链提示，从而使大语言模型具备特定领域的理解能力和逐步推理能力（Wang 等，2025a）。包括 GRPO 在内的强化学习进展，在 GSM8K 和 MATH 等基准测试上取得了显著提升（Shao 等人，2024）。然而，对前提顺序和问题表述的敏感性仍是一项挑战（Chen 等，2024；Huang 等，2025）。自校正方法也通过支持自我改进，提升了大语言模型的数学推理能力（Xiong 等，2025；Yan 等，2025）。开放证明语料库（Dekoninck 等，2025）和 MathArena（Balunovi´c 等，2025）等基准测试揭示了最终答案的准确性与证明有效性之间的差距。研究者已探索利用知识图谱智能体提高成功率（Li 等，2025a），而 LeanDojo（Yang 等，2023）及其他研究（Mikuła 等，2023；Tao 等，2025）则改进了前提选择。近期的创新（Mukherjee 等，2025；Pan 等，2025；Wang 等，2025b）进一步使监督与中间推理相协调，从而改进错误检测和反思推理。我们的工作建立在已有研究成果之上，聚焦于结构化的中间推理，旨在提升大语言模型在数学领域的泛化推理能力。

3 方法：RULES

RULES 包含两个组成部分：一个**双分节引理判定任务**，要求在使用前进行明确的资格检查；以及 **分节感知强化学习**，使信用分配与这些检查相一致。

3.1 引理判定任务

给定输入 $x = (L,\ S)$、引理 L 和陈述 S，模型会生成包含两个部分 $y = (r_{pre},\ l_{pre}) \oplus (r_{con},\ l_{con})$ 的结构化输出，其中 r_{pre} 和 r_{con} 是针对前提条件和结论效用检查的自由形式的推理说明，而各部分的（预测）标签，即 $l_{pre},\ l_{con} \in \{0, 1\}$，则是提取的判定。最终的有用性判定由前提条件检查和结论检查确定性聚合得出 $l_{use} = l_{pre} \wedge l_{con} \in \{0, 1\}$。该任务规范将前提条件检查和结论检查明确化，使其可供审计；采用带有分节标记的最简模式，可确保解析结果明确无歧义。附录 A 通过提示词对此作了说明。

3.2 分节感知强化学习

训练方法改编自 GRPO。具体来说，对于每个输入 x，给定一次推演 $y^i = y^i_{pre} \oplus y^i_{con} = (r^i_{pre},\ l^i_{pre}) \oplus (r^i_{con},\ l^i_{con})$，如果前提条件检查正确，但汇总最终标签错误，这就隐含地意味着结论检查必定有误，那么我们只保留结论检查部分，即将 $(r^i_{con},\ l^i_{con})$ 的损失掩码设为 1，其他部分设为 0。对于强化学习更新，每个群 G 中的梯度函数如下所示：

$$\nabla_\theta L_{section-aware} = -\frac{1}{|G|}\sum_{i=1}^{|G|}\frac{1}{|y^i|}\sum_{j=1}^{|y^i|} m^i_j \tilde{A}^i_j \nabla_\theta \log\ \pi_\theta(y^i_j | x,\ y^i_{<j}),$$

其中 m^i_j 是第 i 次推演中第 j 个令牌的掩码，该公式中的其他符号遵循 GRPO 的默认设置，例如估计优势 $\tilde{A}^i_j$。通过这种方式，我们只惩罚错误分节，并针对某些错误令牌，以更细粒度的信号监督模型，从而将“先检查后使用”落实为定理应用的安全过滤器。至关重要的是，这种目标层面的修改独立于底层模型架构或规模。计算开销仅限于损失计算期间可忽略不计的令牌掩码操作，因此 RULES 兼容任何大语言模型骨干网络。

4 实验设置

4.1 模型

我们主要对四种大语言模型进行了实验：**DeepMath-1.5B**（He 等人，2025年）、**OLMo-2-1124-7B-Instruct**（OLMo 等人，2025年）、**Qwen2.5-Math-7B-Instruct**（Yang 等人，2024年）和**Llama-3.1-8B-Instruct**（Grattafiori 等人，2024年）。对于每个模型，我们报告三种变体：**基础版**：使用我们的双分节模式进行直接推理。**GRPO**：使用原始提示，仅采用最终有用性标签（单标签目标）进行默认 GRPO 优化。**RULES**：基于双分节模式进行带分节感知掩码的 GRPO。

4.2 训练设置

训练数据来自自然语言和形式化证明语料库，即 **NaturalProofs**（Welleck 等人，2021年）、**NLPS**（自然语言前提选择）（Ferreira 和 Freitas，2020年）、**前提选**

模型\测试		域内测试				扰动测试	端到端基准					
		NP	NLPS	IS	IL	DT	PA	CM	IM	MP-hard	MP-simple	TQ-Math
DeepMath	基础版	60.8	51.5	51.6	31.9	24	39.1	45.7	**76**	53	72	55.4
	GRPO	77.5	77.2	67.6	65.8	35.6	37.9	**63.1**	73	54.5	72.8	56.1
	RULES	**88.5**	**86.2**	**75.8**	**89.5**	**47**	**39.5**	46	71	**56.3**	**72.8**	**57.7**
OLMO2	基础版	77.6	69.1	53.9	64.9	47.6	**8.4**	**52.3**	**29**	8.6	18.6	27.1
	GRPO	**85.2**	**84.2**	72.2	**83.6**	46.8	7.8	48.1	23	9.3	21.9	**30.3**
	RULES	84.9	83.4	**76.6**	82.5	**60.5**	6.7	48.2	24	**10.4**	**23.3**	26.9
通义千问 2.5	基础版	63.9	71.9	55.5	55.8	71.9	32.4	59.5	**55**	33.7	59.5	49.8
	GRPO	76.5	83.7	74.2	80.4	81.2	**35.6**	60.5	54	33.7	59.5	50
	RULES	**88.2**	**85.8**	**81.3**	**92.7**	**87.8**	32	**61.3**	53	**35.1**	**62.4**	**50.5**
Llama-3.1	基础版	61.3	75.1	54.7	52.3	89.2	11.5	**56.5**	**22**	**13.3**	**38.7**	**33.5**
	GRPO	88.6	**86.5**	85.9	89.8	80.9	**13.2**	50.7	18	13.3	36.2	32.6
	RULES	**93.2**	86.4	**86.3**	**93.6**	**90.9**	12.8	49.3	16	12.9	36.9	33

表 1： **主要结果**。对于域内测试，NP、NLPS、IS 和 IL 分别代表 NaturalProofs、NLPS、Isabelle 中的前提选择以及基于 IMO-lemma 构建的测试集。DT 代表从 DeepTheorem 采样的扰动测试集。PA、CM、IM、MP 和 TQ 分别代表 Putnam-Axiom、CounterMATH、IneqMath、MATH-Perturb 和 TheoremQA。MP-hard 和 MP-simple 分别代表基于论文的困难和简单扰动。TQ-Math 是 TQ 基准测试中的数学子集。由于 Li 等人(2025b) 中提到的数据分布不均衡，CounterMATH 使用 F1，其余数据则使用基于精确匹配的准确率。

消融实验	域内测试				扰动测试	端到端基准					
	NP	NLPS	IS	IL	DT	PA	CM	IM	MP-hard	MP-simple	TQ-Math
基础版	63.9	71.9	55.5	55.8	71.9	32.4	59.5	**55**	33.7	59.5	49.8
GRPO	76.5	83.7	74.2	80.4	81.2	**35.6**	60.5	54	33.7	59.5	50
双部分一次性	87	85.4	**84.4**	92.7	73.1	32.4	60.5	49	35.5	60.9	49.5
无扰动数据	**93.2**	**87.4**	82.4	94.4	19.1	32.2	56.4	49	34.4	59.1	50.5
去噪扰动数据	86	83.9	83.6	**95**	74.3	34.1	59.6	54	**35.8**	58.8	**50.9**
RULES	88.2	85.8	81.3	92.7	**87.8**	32	**61.3**	53	35.1	**62.4**	50.5

表 2： **消融结果**。 双部分一次性 指在最后输出两个检查预测。无扰动数据 和去噪扰动数据 分别指训练时移除扰动数据，以及使用 o3 模型和人工验证筛除不合格的扰动训练数据。实验使用 Qwen2.5-Math-7B-Instruct 完成。

择（Isabelle 中）（Mikuła 等人，2023年），以及 **IMO 引理**（Liang 等人，2025年），将陈述与候选引理配对，并以二元标签标示引理的有效性。更多训练设置和设计见附录B。

扰动设计。 我们的目标是将表面相似性与真正的适格性区分开来，因此我们还在训练中加入了一些扰动数据：在尽可能少地改动陈述并保持数学连贯性的前提下，使某个引理不再适用。构造扰动时仅为前提条件检查提供标签，原因有二： 前提条件满足可以直接根据陈述—引理对判定是否成立，且独立于证明策略，因此能以较低的歧义获得可靠的监督信号。 结论效用本质上取决于策略（例如反例证明），因此大规模标注成本高且噪声大。

奖励设计。 为了与基线进行公平比较，奖励采用基于规则的二元形式：正确得 1 分，错误得 -1 分。此外，对于那些无法按照双分节模式解析的输出，将受到 -2 分的格式惩罚。

4.3 评估设置

我们从三个维度进行评估，以对应预期用途：域内引理判断、对破坏适用性的扰动的鲁棒性，以及向端到端数学问题求解的迁移。所有结果均通过采样 7 个输出，并对提取出的答案进行自一致性多数投票计算，记为**SC@7**。更多评估细节见附录 复数域。

域内评估 我们使用从训练数据来源中抽取的留出划分，分别命名为**naturalproofs-test**、**nlps-test**、**isabelle-test**和**IMO-lemma-test**。

扰动测试与目标以及扰动数据整理的训练流程类似，我们也在一个域外扰动套件上进行评估。该套件通过对陈述进行最小幅度的编辑来违反引理前提条件而构建。但原始数据采样自 DeepTheorem（Zhang 等人，2025年），其

称为**deeptheorem-perturbation**。关于扰动数据质量的更多详情，请见附录D。

端到端基准测试 除了针对任务的评估外，我们还评估模型在多种数学任务上的迁移能力，包括竞赛风格的问题（**Putnam-Axiom** (Gulati 等人，2024年)）、不等式推理（**IneqMath** (Sheng等人，2025b)）、由反例驱动的问题（**CounterMATH**(Li 等人，2025b)）、与扰动相匹配的测试集（**MATH-Perturb**(Huang 等人，2025年)），以及以定理为核心的数学问答（**TheoremQA** (Chen 等人，2023年)。

5 分析

5.1 主要结果

域内引理判断。 在全部四种基础模型上，与基础版 和单标签*GRPO* 基线相比，RULES 均提升了域内引理判断表现。在大多数数据划分上，提升都很稳定，在源自 IMO 的测试集上尤为显著；OLMO2 在三个数据划分上的表现与 GRPO 相当，而在 Isabelle 数据划分上则有明显提升。这些结果表明，明确检查前提条件和结论有助于模型加深对引理适用性的理解。

对破坏适用性的扰动的鲁棒性。 在域外扰动套件上，RULES 在全部四个模型上均持续优于基础版 和 *GRPO*，最大优势可达23+ 分。这一模式支持训练时扰动数据与分节感知掩码的预期效果：在使前提条件失效的最小陈述编辑下，模型能更可靠地排除熟悉但不适用的引理。

混合端到端迁移。 在下游任务中，RULES 的整体表现大致与单标签*GRPO*持平：在与扰动相适应的子集上有所提升，但在竞赛风格和基于不等式的问题上偶尔出现退步。我们将这些退步解读为一种**性能与严谨性的权衡**：标准模型有时可能会“碰巧走运”，通过臆造适用性来强行求解，而 RULES 则严格执行前提条件检查。这种严谨做法优先保障精确率和安全性，避免无效推理；因此，在约束较宽松或许能意外得出正确答案的任务上，召回率自然会降低。因此，提升最显著的情况是在与扰动相适应的测试集上，因为抵御无效输入的鲁棒性至关重要。相反，CounterMATH 这类构造性任务需要综合示例，这种能力不同于资格判定。因此，RULES 充当必要的鲁棒性层，确保模型不会基于根本不适用的前提尝试证明。

5.2 消融实验

两次预测与一次预测的对比。 将顺序执行的双分节协议替换为“双部分一次性”变体（仅在末尾同时输出两项判断）可获得较高的域内准确率，但在扰动套件上的鲁棒性明显较弱。这表明分阶段结构与分节感知对齐对于学习在前提条件未满足时暂不使用引理至关重要。

训练时扰动数据的作用。 在训练中移除扰动数据可保持域内得分，但会使域外测试套件上的鲁棒性大幅下降，这证实精心构建的负例对于教会前置条件门控至关重要。然而，我们强调这种鲁棒性并非仅仅来自数据扩展。架构贡献显著：RULES 在扰动测试上比数据等价的“双部分一次性”基线高出 14.7 分（表 2）。这一显著差距证实，尽管数据提供了必要信号，**分节感知强化学习机制**对于有效利用这些信号、防止浅层模式匹配仍不可或缺。

“去噪”扰动数据的影响。过滤低质量扰动带来了喜忧参半的结果：某些端到端指标和域内测试略有提升，但在扰动套件上的表现仍不及 RULES。迄今为止，这些证据并不支持“干净的细粒度标签”总能提升整体数学推理能力；相反，严格清洗与鲁棒性覆盖之间似乎存在权衡。

6 结论

我们提出 RULES，通过明确的先检查后使用协议和分节感知强化学习，使引理使用与之保持一致。RULES 在域内取得稳定提升，在前提条件遭破坏的扰动下显著增强鲁棒性，并在端到端任务上达到持平或略有提升。消融实验表明，分阶段双节输出和与章节对齐的信用分配缺一不可。

局限性

任务范围与监督。 我们的研究聚焦于引理判断，而非完整的证明生成。监督信号针对特定任务量身设计，这可能会使模型产生意料之外的行为偏差。

标签覆盖率与噪声。 在扰动数据中，只有前提条件标签可用，而结论效用标签被省略，以避免依赖策略。这种非对称监督可能引入偏差，并限制对适格性之外的有用性的学习。此外，奖励是二元的，并带有格式惩罚，这简化了优化过程，但可能无法充分利用分级信号。

模型与数据覆盖范围。 实验涵盖四个中等规模模型，以及一组特定的数学语料库（NaturalProofs、NLPS、Isabelle 前提选择和 IMO-lemma）。这使数学各分支、自动定理证明技术以及英语以外的语言仍存在覆盖空白。尽管适格性推理机制在理论上与语言无关，但将其扩展到多语言环境或其他形式化证明助手（例如 Lean、Coq）仍是未来的研究方向。

参考文献

MislavBalunovi´c、Jasper Dekoninck、Ivo Petrov、Nikola Jovanovi´c 和 Martin Vechev。2025年。Matharena：评估大语言模型在未受污染的数学竞赛中的表现。*arXiv* 预印本 *arXiv:2505.23281*。

Wenhu Chen、Ming Yin、Max Ku、Pan Lu、Yixin Wan、Xueguang Ma、Jianyu Xu、Xinyi Wang 和 Tony Xia。2023年。TheoremQA：定理驱动问答数据集。载于*2023*年自然语言处理实证方法会议论文集，第7889–7901页，新加坡。计算语言学协会。

Xinyun Chen、Ryan A Chi、Xuezhi Wang 和 Denny Zhou。2024年。前提顺序对于使用大语言模型进行推理至关重要。载于第*41*届国际机器学习会议论文集，第6596–6620页。

Jasper Dekoninck、Ivo Petrov、Kristian Minchev、Mis- lav Balunovic、Martin Vechev、Miroslav Marinov、Maria Drencheva、Lyuba Konova、Milen Shumanov、Kaloyan Tsvetkov 等人。2025年。开放证明语料库：对大语言模型生成的数学证明进行大规模研究。*arXiv* 预印本 *arXiv:2506.21621*。

Deborah Ferreira 和 André Freitas。2020年。自然语言前提选择：为数学文本寻找支持性陈述。载于第十二届语言资源与评估会议论文集，第2175–2182页，法国马赛。欧洲语言资源协会。

Simon Frieder、Julius Berner、Philipp Petersen 和 Thomas Lukasiewicz。2024年。面向数学家的大语言模型。预印本，arXiv:2312.04556。

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan、Amy Yang、Angela Fan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark，以及其他542位作者。2024年。Llama 3模型家族。预印本，arXiv:2407.21783。

Aryan Gulati、Brando Miranda、Eric Chen、Emily Xia、Kai Fronsdal、Bruno de Moraes Dumont 与 Sanmi Koyejo。2024年。Putnam-AXIOM：用于衡量更高层次数学推理的功能性与静态基准测试。载于 *NeurIPS' 24* 第四届数学推理与人工智能研讨会。

Dadi Guo、Jiayu Liu、Zhiyuan Fan、Zhitao He、Haoran Li、Yumeng Wang 与 Yi R. Fung。2025年。以数学证明作为试金石：揭示先进大型推理模型的失效模式。预印本，arXiv:2506.17114。

Zhiwei He、Tian Liang、Jiahao Xu、QiuzhiLiu、Xingyu Chen、Yue Wang、Linfeng Song、Dian Yu、Zhenwen Liang、Wenxuan Wang、Zhuosheng Zhang、Rui Wang、Zhaopeng Tu、Haitao Mi 与 Dong Yu。2025年。DeepMath-103K：一个大规模、具有挑战性、已去污染且可验证的数学数据集，用于推进推理。预印本，arXiv:2504.11456。

Kaixuan Huang、Jiacheng Guo、Zihao Li、Xiang Ji、Jiawei Ge、Wenzhe Li、Yingqing Guo、Tianle Cai、Hui Yuan、Runzhe Wang、Yue Wu、Ming Yin、Shange Tang、Yangsibo Huang、Chi Jin、Xinyun Chen、Chiyuan Zhang 与 Mengdi Wang。2025年。MATH-perturb：针对强扰动对大语言模型数学推理能力的基准测试。载于第四十二届国际机器学习会议。

Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 与 Ion Stoica。2023年。基于 PagedAttention 的大语言模型服务高效内存管理。载于*ACM SIGOPS* 第*29*届操作系统原理研讨会论文集。

Vincent Li、Tim Knappe、Yule Fu、Kevin Han 与 Kevin Zhu。2025年a。利用大语言模型智能体与知识图谱自动生成数学证明。*arXiv* 预印本 *arXiv:2503.11657*。

YinghuiLi、JiayiKuang、Haojing Huang、Zhikun Xu、Xinnian Liang、Yi Yu、Wenlian Lu、Yangning Li、Xiaoyu Tan、Chao Qu、Ying Shen、Hai-Tao Zheng,

和 Philip S. Yu。2025b。展示一个示例，掌握多种概念！数学大语言模型中的反例驱动的概念推理。载于 第四十二届国际机器学习会议。

Zhenwen Liang、Linfeng Song、Yang Li、Tao Yang、Feng Zhang、Haitao Mi 和 Dong Yu。2025年。通过解耦推理与证明，解决更具挑战性的国际数学奥林匹克问题。*arXiv* 预印本。

Maciej Mikuła、Szymon Antoniak、Szymon Tworkowski、Bartosz Piotrowski、Albert Jiang、Jin Peng Zhou、Christian Szegedy、Łukasz Kuci´nski、Piotr Miło´s 和 Yuhuai Wu。2023年。Magnushammer：一种用于前提选择的基于Transformer的方法。发表于 *NeurIPS'23*数学推理与人工智能第三届研讨会。

Sagnik Mukherjee、Abhinav Chinta、Takyoung Kim、Tarun Anoop Sharma 和 Dilek Hakkani Tur。2025年。前提增强推理链可提升大语言模型数学推理中的错误识别能力。发表于第四十二届国际机器学习会议。

Team OLMo、Pete Walsh、Luca Soldaini、Dirk Groen- eveld、Kyle Lo、ShaneArora、Akshita Bhagia、Yuling Gu、Shengyi Huang、Matt Jordan、Nathan Lambert、Dustin Schwenk、Oyvind Tafjord、Taira Anderson、David Atkinson、Faeze Brahman、Christopher Clark、Pradeep Dasigi、Nouha Dziri及其他21人。2025年。2 OLMo 2 狂怒。预印本，arXiv:2501.00656。

OpenAI、Josh Achiam、Steven Adler、SandhiniAgarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Ale- man、Diogo Almeida、Janko Altenschmidt、Sam Alt- man、Shyamal Anadkat、Red Avila、Igor Babuschkin、Suchir Balaji、Valerie Balcom、Paul Baltescu、Haim- ing Bao、Mohammad Bavarian、Jeff Belgum及其他262人。2024年。GPT-4技术报告。预印本，arXiv:2303.08774。

Zhuoshi Pan、Yu Li、Honglin Lin、Qizhi Pei、Zinan Tang、Wei Wu、Chenlin Ming、H. Vicky Zhao、Conghui He 和 Lijun Wu。2025年。LEMMA：通过从错误中学习，推动大语言模型的数学进步。载于计算语言学协会 *ACL 2025* 年会论文集：*ACL 2025*，第 11615–11639 页，奥地利维也纳。计算语言学协会。

Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。2024年。DeepSeekMath：突破开放语言模型数学推理的极限。预印本，arXiv:2402.03300。

Guangming Sheng、Chi Zhang、Zilingfeng Ye、Xibin Wu、Wang Zhang、Ru Zhang、Yanghua Peng、Haibin Lin 和 Chuan Wu。2025a。Hybridflow：一种灵活高效的基于人类反馈的强化学习框架。载于第二十届欧洲计算机系统会议论文集，EuroSys '25，第 1279–1297 页，美国纽约州纽约市。美国计算机协会。

Jiayi Sheng、Luna Lyu、Jikai Jin、Tony Xia、Alex Gu、James Zou 和 Pan Lu。2025b。利用大语言模型求解不等式证明。预印本，arXiv:2506.07927。Yicheng Tao、Haotian Liu、Shanwen Wang 和 Hongteng Xu。2025年。学习一种有效的前提检索模型，以实现高效的数学形式化。在第 2 届 AI for Math研讨会 @ 国际机器学习会议 2025。Peng-Yuan Wang、Tian-Shuo Liu、Chenyang Wang、Yi-Di Wang、Shu Yan、Cheng-Xing Jia、Xu-Hui Liu、Xin-Wei Chen、Jia-Cheng Xu、Ziniu Li 等。2025a。大语言模型数学推理综述。arXiv 预印本 arXiv:2506.08446。Pinzheng Wang、Juntao Li、Zecheng Tang、Haijia Gui 等。2025b。通过自我对弈提升语言模型推理过程的合理性。发表于第四十二届国际机器学习会议。Sean Welleck、Jiacheng Liu、Ronan Le Bras、Hannaneh Hajishirzi、Yejin Choi 和 Kyunghyun Cho。2021。NaturalProofs：自然语言数学定理证明。发表于神经信息处理系统大会数据集与基准赛道第三十五届会议（第一轮）。Sean Welleck、Jiacheng Liu、Ximing Lu、Hannaneh Hajishirzi 和 Yejin Choi。2022。NaturalProver：利用语言模型生成有依据的数学证明。发表于神经信息处理系统大会进展，第 35 卷，第 4913–4927 页。Curran Associates, Inc. Wei Xiong、Hanning Zhang、Chenlu Ye、Lichang Chen、Nan Jiang 和 Tong Zhang。2025年。数学推理的自奖励校正。arXiv 预印本 arXiv:2502.19613。Yuchen Yan、Jin Jiang、Yang Liu、Yixin Cao、Xin Xu、Mengdi Zhang、Xunliang Cai 和 Jian Shao。2025年。Sˆ 3cmath：自发的步骤级自校正让大语言模型成为更优秀的数学推理者。发表于AAAI人工智能会议论文集，第 39 卷，第 25588–25596 页。An Yang、Beichen Zhang、Binyuan Hui、Bofei Gao、Bowen Yu、Chengpeng Li、Dayiheng Liu、Jian-hong Tu、Jingren Zhou、Junyang Lin、Keming Lu、Mingfeng Xue、Runji Lin、Tianyu Liu、Xingzhang Ren 和 Zhenru Zhang。2024年。通义千问2.5-Math 技术报告：通过自我改进迈向数学专家模型。预印本，arXiv:2409.12122。Kaiyu Ya ng、Aidan Swope、Alex Gu、Rahul Chala mala、Peiyang Song、Shixing Yu、Saad Godil、Ryan J Prenger 和 Animashree Anandkumar。2023年。LeanDojo：利用检索增强语言模型进行定理证明。神经信息处理系统大会进展，36:21573–21612。Ziyin Zhang、Jiahao Xu、Zhiwei He、Tian Liang、Qiuzhi Liu、Yansi Li、Linfeng Song、Zhenwen Liang，

Zhuosheng Zhang、Rui Wang、Zhaopeng Tu、Haitao Mi 和 Dong Yu。2025年。 DeepTheorem：通过自然语言和强化学习推进大语言模型推理与定理证明。 预印本，arXiv:2505.23754。

A 提示

双部分提示

引理：{lemma} 陈述：
{statement}

请逐步分析该引理是否有助于证明或证伪给定陈述。你应依次进行两项检查：

步骤 1：前提条件检查 首先，分析该引理的前提条件是否在给定陈述中得到满足。推理后，请在 boxed{} 中给出答案，填写 boxed{True}（前提条件已满足）或 boxed{False}（前提条件未满足）。

步骤 2：结论检查
接下来，分析该引理的结论是否有助于证明或反驳给定的陈述。在推理之后，将答案以 boxed{True}（结论有帮助）或 boxed{False}（结论没有帮助）的形式放入 boxed{} 中。

请使用与上述相同的章节标题（即 **步骤 1：前提条件检查** 和 **步骤 2：结论检查**）分别开始两项检查，并严格遵循以下格式：- 首先，说明前提条件检查的推理过程 - 然后输出 boxed{True} 或 boxed{False} 作为前提条件检查的结果 - 接着，说明结论检查的推理过程 - 最后输出 boxed{True} 或 boxed{False} 作为结论检查的结果

原始提示

引理：{lemma} 陈述：
{statement}

请逐步推理上述引理是否有助于证明或反驳其后续陈述，然后将最终答案（即 True 或 False）放入 boxed 中。

双章节一次性提示

引理：{lemma} 陈述：
{statement}

请逐步思考该引理是否有助于证明或反驳给定陈述。具体而言，请 1）检查给定陈述是否满足该引理的前提条件；2）检查该引理的结论是否有助于证明给定陈述。然后将前提条件检查和结论检查的最终答案（即每项检查的 True 或 False）填入 boxed{} 中。（例如，boxed{True, False} 表示该引理的前提条件得到满足，但其结论无助于证明给定陈述。）

面向 o3（DeepTheo- rem）的前提条件扰动提示

根据给定陈述及其证明，能否帮我以最少的修改扰动该陈述，使得证明中使用的某个引理的一项前提条件不再满足？如果无法进行扰动，请输出 suitability: False。否则，请返回 suitability: True，然后提取证明中受影响的引理，生成扰动后的陈述，并说明扰动的推理依据。

关键要求：
1. 数学一致性：扰动后的陈述必须定义良好且具有数学意义。不要创建概念上不可能或毫无意义的陈述（例如，“半群的子群”）。

2. 真正不可适用性：所选引理必须在任何合理的证明方法中都无法用于扰动后的陈述。避免进行表面上的更改，例如：- 引理适用于新定义域的某个子集或限制范围 - 可以通过标准对应关系间接使用引理 - 可以将该陈述还原到引理适用的定义域

3. 验证测试：在最终确定之前，请问：“所选引理是否仍能通过任何数学路径，对证明或反驳扰动后的陈述做出有意义的贡献？”如果是，则该扰动无效。

规则：-

除该片段外，逐字复制原始陈述。- 不要强化假设，不要引入无关的概念/符号，也不要进行会使陈述变得平凡的更改。- 优先采用能对引理适用性构成根本障碍的最简单更改。- 扰动应针对引理的核心假设或定义域，且其方式无法被规避。

应避免的无效扰动模式：- 存在标准对应关系的定义域更改（例如，在对实变量求导时的$\mathbb{R}$ 指数 $\to \mathbb{C}$ 指数）- 仍允许目标概念成立的结构降级（例如，群→半群，但仍保留“子群”这一术语）- 仅影响表面术语而不影响数学实质的更改

仅返回此模式：适用性：{bool} 受影响引理：{lemma-content} 原始陈述：{original statement} 扰动后陈述：{perturbed statement} 推理依据：{rationale explaining why the lemma becomes fundamentally inapplicable, not just surface-level different}

现在处理：

陈述：{statement} 证明：{proof}

o3 的前提条件扰动验证提示词

你正在评估数学陈述扰动的质量。你的任务是确定该扰动是否有效地破坏了所选引理的适用性。

给定扰动数据：引理：{lemma} 原始陈述：{statement} 扰动后的陈述：{perturbed statement} 推理依据：{rationale}

评估标准（必须全部通过）：

1. 数学一致性检查：- 扰动后的陈述在数学上是否定义明确？- 所使用的数学概念是否始终保持一致且具有明确意义？- 数学对象及其关系是否能够相互协调？

常见失败情况：- 混用不兼容的概念（例如，“半群的子群”）- 未定义明确的运算或结构- 概念矛盾

2. 引理不可适用性验证：系统检查该引理是否仍可通过以下方式使用：

a) 直接应用：- 在新的语境中，是否仍能满足引理的假设？

b) 限制/子集应用：- 引理是否适用于新域中的相关子集？- 我们能否将问题限制在引理条件成立的情形？

c) 对应/同构：- 旧域与新域之间是否存在标准的数学对应？- 我们能否将问题映射回引理适用的情形？

d) 间接/基础性使用：- 能否利用该引理证明前置结果？- 该引理是否是所需理论的基础，而该理论又是证明扰动后陈述所必需的？- 它是否可以应用于有限商、局部分析等？

e) 类似结果：- 新域中是否存在该引理的直接类似结果？- 原引理是否会成为证明这些类似结果的关键步骤？

3. 失败模式检测：检查以下已知的无效模式：

- 表面领域变更：看起来有所不同，但保留了数学实质的变化 - 术语替换：更改标签，但不影响底层数学内容 - 平凡泛化：转向更宽泛的类别，但限制仍然适用 - 不完整领域迁移：更改领域中部分但非全部相关方面

4. 证明路径分析：请思考：“在任何合理的尝试中，无论是证明还是反驳这个经过扰动的陈述，给定的引理能否对论证作出实质性贡献？”

如果任一路径的答案为是 →无效扰动 如果所有路径的答案均为否 → 有效扰动

请按以下格式给出评估结果：

有效性： [布尔值]推理依据： [说明该扰动为何有效或无效]

现在请评估给定的扰动：

B 更多训练详情

我们的训练实验主要在 AMD MI250 和 Nvidia H200 GPU 服务器上进行。训练批大小为 128，推演次数设为 8。推理模型的响应长度为 8000，而非推理模型为 2048。当单个批次中启用分节感知损失掩码的实例比例接近零时，使用训练检查点。我们的训练框架基于*verl* (Sheng等人，2025a）。训练数据分布如下：Natural-Proofs：4744，NLPS：2858，Isabelle 中的前提选择：4484，以及 IMO-lemma：986。

扰动设计详情。具体而言，我们针对证明中使用的引理的某个特定假设，对陈述进行最小幅度的修改，使该假设不再成立（$l_{pre}^{gold} = 0$），同时不引入无关概念，也不使主张变得 trivial。这些数据源自 NaturalProofs，并由 o3 模型[1]处理。扰动数据是唯一带有用于分节感知掩码的黄金前提标签的部分。

C 更多评估详情

我们的评估框架基于 *vllm* (Kwon等人，2023年）。最大词元长度设为 10k，以避免过早截断；所有模型的 SC@7 温度均设为 0.7。域内数据和扰动评估数据的分布分别为 **naturalproofs-test**(1002), **nlps-test**(676), **isabelle-test** (256), **IMO-lemma-test**(342)，以及 deeptheorem-perturbation（839）

D 扰动数据质量

为验证我们的扰动套件（其中仅翻转前提条件标签）的质量，我们进行了多阶段验证。对于训练数据，我们使用 o3 模型和 GPT-OSS-120B 对其进行交叉验证，两个模型对 84.8% 的训练数据达成一致。我们仅将这 84.8% 的数据用作最终的扰动训练。此外，我们作者还从中抽取了 100 个数据点进行人工验证，通过率为 90%。对于扰动评估数据，我们作者在经过基于 o3 模型的筛选和人工筛选后，抽取了 100 个数据点，以验证最终数据的质量。我们的评估指南严格遵循提示中的原则：(i) 数学一致性；(ii) 真正不可适用性；(iii) 通过等价引理验证相似路径。抽样评估数据的通过率为 95%。

[1] https://platform.openai.com/docs/models/o3

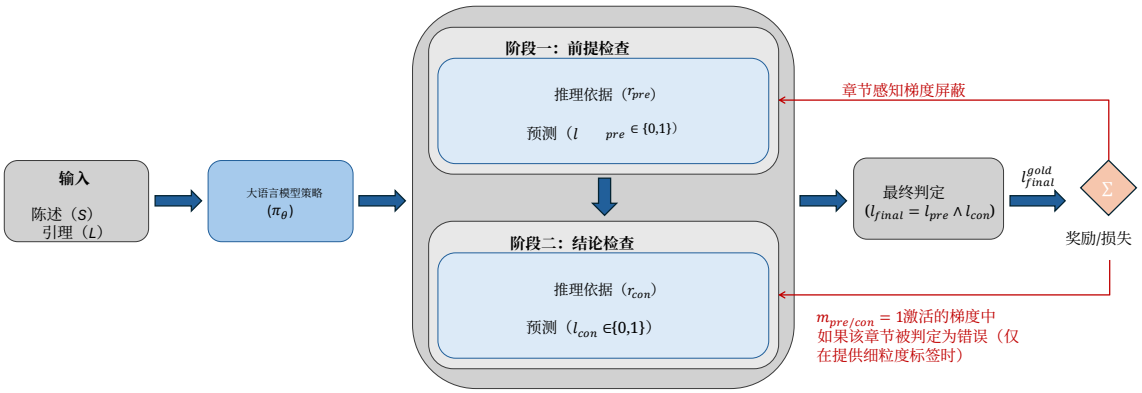

图 1：RULES 训练框架概览。如第 3.2 节所述，如果提供细粒度标签（前提条件/结论检查），我们就能知道哪些 token 片段应因负向奖励而受到归责。

E 案例研究：防止引理误用

为了说明 RULES 如何防止对引理适用性的幻觉，我们展示一个来自黎曼几何领域的案例，其中标准基础模型表现不佳。

情境。陈述： 设 $G = SU(n)$，且令 g 为一个左不变度量。证明任何满足线性化爱因斯坦方程的对称 2-张量 h .....必为平凡张量。

候选引理： “由于 g 是双不变的，因此它是一个爱因斯坦度量.....对于某个 $\lambda > 0$，$Ric_g = \lambda g$。”

分析。 该引理要求度量为双不变。然而，题目陈述仅保证 g 是左不变的。在微分几何中，左不变性弱于双不变性；因此，该引理的前提条件并未满足。

模型行为。

- **基线（基础版）：** 该模型忽略了“左不变”和“双不变”之间的细微区别，臆测条件已满足，输出 True，并尝试使用引理。

- **RULES：** 模型正确识别出前提条件检查部分中的不匹配：“推理依据：该引理要求双不变性，但该陈述仅假设左不变性.....”它针对前提条件输出 $False$，从而有效阻断了无效路径。

F 大语言模型的使用

我们仅在论文写作中使用大语言模型进行语言润色。我们并未使用大语言模型进行想法形式化，也未达到可将其视为贡献者的程度。