arXiv:2603.23517v1 [cs.LG] 6 Mar 2026

超越准确率：引入一种用于可解释评估的符号-机制方法

Reza Habibi[1*], **Darian Lee**[1*],
Magy Seif El-Nasr
[1] 加利福尼亚大学圣克鲁兹分校，606
Engineering Loop，Santa Cruz，CA 95064，美国
rehabibi@ucsc.edu，daeilee@ucsc.edu

摘要

基于准确率的评估无法可靠地区分真正的泛化与记忆、数据泄漏或脆弱的启发式规则等捷径，尤其是在小数据场景下。在这篇**立场论文**中，我们主张采用机制感知评估，将与任务相关的符号规则与机制可解释性相结合，从而得出算法式通过/失败评分，准确揭示模型在哪些方面实现了泛化，又在哪些方面利用了模式。我们通过自然语言到 SQL 任务对此进行了演示：在不同条件下训练两个架构相同的模型，其中一个不提供数据库模式信息（迫使模型依赖记忆），另一个提供模式信息（使模型能够落地）。标准评估显示，记忆型模型在未见数据上的字段名称准确率达到 94%，这会错误地暗示其具备相应能力。我们的符号-机制评估则揭示，该模型违反了核心的模式泛化规则，而这一失败无法通过准确率指标察觉。

引言与背景

标准的自然语言处理评估依赖留出测试集，并使用精确匹配、F1 或 BLEU 等指标（Papineni 等人，2002；Nakache、Metais 和 Timsit，2005；Gehrmann、Clark 和 Sellam，2023）。然而，近期研究表明，这些方法无法可靠地区分记忆与泛化。污染现象十分普遍，例如 QNLI 等基准测试在热门预训练语料库中的污染率超过 50%（Lee 等人，2022；Dodge 等人，2021）；即使面对干净的数据，模型也会利用伪相关启发式规则（Gururangan 等人，2018；McCoy、Pavlick 和 Linzen，2019；Kim 等人，2025）。LLM-as-a-Judge 等替代方法则受到自我偏好偏差和对抗脆弱性的影响（Chen 等人，2024；Raina、Liusie 和 Gales，2024）。我们认为，这些缺陷源于一个根本问题：当前的评估衡量的是表层表现，而非潜在的计算机制。我们提出采用机制可解释性（MI）评估，直接检验模型是否使用了正确的算法。

污染与可利用的启发式规则

污染研究揭示了这一问题的普遍程度。（Zhang 等，2024）发现，记忆概率与 GSM8K 变体上的表现之间的相关系数为 0.32，而记忆概率与 GSM8K 变体上的表现之间的相关系数为 0.32，而（Li 和 Flanigan，2024）发现，无上下文的精确匹配生成与测试表现之间的相关系数为 0.88。（Aiyappa 等，2023）观察到，在评估 ChatGPT 的论文发表后，其基准测试表现有所提升，这表明存在由基于人类反馈的强化学习驱动的污染。值得注意的是，（Samuel、Zhou 和 Zou，2025）发现不同污染检测方法之间的一致性低得惊人，表明目前不存在可靠的标准；这意味着在测试前缓解和检测污染并不可行。此外，现代训练流程会利用用户数据进行基于人类反馈的强化学习，并持续抓取网页，这些做法造成了无法消除的持续污染途径（Zhang 等，2024；Aiyappa 等，2023）。

即使数据完美无瑕，模型也会利用伪相关模式。(Gururangan 等人，2018)通过利用标注伪影，仅凭假设句就达到了 67% 的 SNLI 准确率。(Kim 等人，2025年)从机制层面表明，基准测试往往未能检验其声称要评估的能力，反而奖励替代机制。对于数据集规模较小的任务（低资源语言和专业领域中很常见），测试集缺乏足够的统计效力，无法区分真正的能力与模式匹配。

从顿悟式泛化到验证

关于顿悟式泛化的研究表明，泛化对应着不同的内部结构（Power 等人，2022年；Nanda 等人，2023年；Liu 等人，2022年）。(Nanda 等人，2023年)指出，泛化电路在记忆过程中、尚未得到充分利用之前便已形成，这表明机制探测可以检测模型运行于记忆主导还是泛化主导状态。然而，顿悟式泛化的见解仍未应用于真实的自然语言处理评估。

这促使人们开展符号验证。符号人工智能提供了用于将正确性属性表示为具有语义意义的规则的框架（Newell、Shaw 和 Simon，1957年、1959年；McCarthy，1995年；Xie、Kersting 和 Neider，2022年）。通过检查内部计算是否符合预期的泛化模式，符号规则能够实现比单纯依靠准确率更严格的评估。

提案

我们的方法采用符号逻辑中的“不可协商规则”，用于描述任何能够可靠解决

[*]这些作者贡献相同。版权所有 © 2026，人工智能促进协会 (www.aaai.org)。保留所有权利。

该任务的电路必须满足的性质。对于任务 T，用户指定规则 R ={r 1, ...,r r k} ，以捕捉任务语义层面的基本要求：必须使用哪些信息，必须满足哪些不变性。这些规则必须能够通过常见的机制干预进行验证，例如激活修补、Logit Lens 分析或注意力可视化。

对于小型评估集中的每个示例，我们检验规则是否得到满足，并赋予通过/失败分数。汇总后的百分比反映模型使用恰当泛化电路的一致程度。

本文是一篇立场论文，主张对具有明确定义的预期算法的任务（例如落地、检索或解析）采用机制感知评估。我们并不声称这种方法适用于开放式或创造性生成任务；在这类任务中，多种在性质上截然不同的算法可能都同样有效，因此不适合采用单一的机制标准。

案例研究

我们在一个简单的自然语言到 SQL 任务上，将我们的评估框架与标准准确率进行比较，并聚焦于作为核心算法的模式落地。我们使用相同的架构、但不同的训练条件训练两个模型：一个不使用模式信息（无模式），因此只能依赖浅层启发式规则；另一个使用模式信息（有模式），从而能够实现真正的泛化。我们表明，标准评估认为两者的能力大致相当，而符号机制评估则揭示了无模式模型缺乏真正的泛化能力。

数据集 我们的案例研究采用 TinySQL CS1 同义词数据集（Harrasse 等人，2025）。每个示例由一条英文提示（english prompt）、以 CREATE TABLE 语句（create statement）形式提供的数据库模式，以及目标 SQL 查询（sql_statement）组成。关键在于，数据库模式使用了英文提示中所提列名的同义词，[1] 因此，仅凭自然语言而没有模式落地，就无法生成正确的 SQL。

该数据集仅包含遵循 SELECT x FROM y 模式的简单查询。鉴于这种结构上的简单性，我们将评估重点放在模式落地上，即正确地将英语列名映射到其在数据库模式中的同义词这一核心算法能力。所有示例中，其余查询结构都遵循可预测的模板。

训练条件： 我们基于 TinyStories-33M（Eldan 和 Li，2023）微调了两个模型，它们具有相同的架构和优化设置，并遵循 Harrasse 等人（2025）的方法。两个模型唯一的区别在于训练期间是否提供数据库模式。基线模型仅使用 TinyStories-33M 短篇故事进行预训练，不包含任何自然语言到 SQL 示例，从而确保没有任务污染（Eldan 和 Li，2023）。

数据库**模式** 条件会在每个训练示例的上下文中包含创建语句。**无模式** 条件则完全省略该语句，同时保持目标 SQL 不变，迫使模型依赖浅层启发式规则，而非查询模式。英文提示中的列名和表名是同义词，而非精确匹配，因此在模式信息可用时，模型能够从中受益（Harrasse 等人，2025）。

我们采用统一的提示格式：### 指令： {英文提示} ### 上下文： {CREATE 语句或留空}###响应： {SQL 语句}。解码从“### Response:”之后开始，并将预测的 SQL 与 SQL 语句进行比较。

基于标准准确率的测试： 我们在 TinySQL CS1 同义词测试集上，将我们的方法与标准评估进行比较，并报告精确匹配准确率（预测结果与标准 SQL 完全一致）和字段名准确率（正确的表名/列名占标准字段名总数的比例，不包括 SQL 关键字）。每个模型都在两种配置下进行评估：包含数据库模式（测试提示中包含数据库模式）和无模式（不包含数据库模式）。

符号机制评估

我们评估了 100 组匹配的提示对，每组都包含一个带有正确数据库模式字段的干净提示，以及一个将某个词替换为其他词的受损版本，同时保持格式不变。

我们测试了五种扰动类型，以探究模型对数据库模式的 grounding 能力。**DB-Synonyms**将列名替换为语义相关的替代词，这些词也曾在训练中作为列名出现（“price”/“cost”），以检验模型是否能在训练过的数据库模式中映射语义等价词。**非数据库同义词**使用语义相关但从未作为列名出现过的词（“mouse”/“rat”），以检验这种语义灵活性是否能够泛化。其余三种扰动用于检验模型是否会对数据库模式违规作出反应，即评估模型能否在出现意外词语时进行检测并作出相应反应。**DB 扰乱**将两个互不相关、但都曾作为列名出现的词配对（“price”/“brand”），**非数据库扰乱**将两个互不相关、且都未曾作为列名出现的词配对（“mouse”/“car”），而**Super-Scramble**则将列名与非列名词语混合（“price”/“rat”）。如果模型能够检测到这些扰乱式扰动，就说明它在主动追踪数据库模式。

符号规则为评估是否存在模式检查电路，我们采用符号规则方法，将这一问题拆解为三个可检验的条件。首先，我们检验因果敏感性：破坏模式标记是否会改变模型偏好的答案？其次，我们检验定位：能否通过在模式标记位置进行激活修补来恢复这一效应？第三，我们检验一致性：多个示例是否都能通过同一层恢复，从而表明存在可复用的机制？我们将这些规则称为 R1 至 R3。由于只有在 R1 和 R2 均通过时，R3 才能通过，因此我们将 R3 通过视为满足所有条件的标志。我们使用 logit 差值来衡量模型对正确标记的偏好：logit diff$(x; i) = \ell(x)_{\text{correct}_i} - \ell(x)_{\text{incorrect}_i}$，其中 $\ell(x)$ 表示最终位置处的模型 logits。令 x_i^{clean} 和 x_i^{corr} 分别表示仅在模式标记上有所不同的干净提示和受损提示

[1] 例如，数据库模式为 CREATE TABLE backup (website TEXT)，提示为“从备份中调出网址”，目标为 SELECT website FROM backup。

。我们将模式敏感性差距定义为

$$\Delta_i = \text{logit_diff}(x_i^{\text{clean}}; i) - \text{logit_diff}(x_i^{\text{corr}}; i).$$

正值 Δ_i 表示模型在选择答案时因果性地依赖模式标记。

激活修补协议 为了测试是否可以从特定层恢复 Δ_i，我们在模式标记位置 (p) 执行残差流激活修补。我们缓存来自 x_i^{clean} 的激活值，然后对于每一层 $L \in \{0, \ldots, L_{\max}\}$，重新运行 x_i^{corr}，同时将 (L, p) 处的残差激活替换为缓存的干净激活，从而得到偏移量 $_i(L)$ = 修补后的差值 $_i(L)$ –校正后的差值 $_i$。我们将最佳有符号恢复值及实现该值的层定义为

$$\text{best_recovery}_i = \max_L \ \max\big(0,\ \text{sign}(\Delta_i) \cdot \text{shift}_i(L)\big)$$

$$L_i^* = \text{argmax}_L \ \max\big(0,\ \text{sign}(\Delta_i) \cdot \text{shift}_i(L)\big)$$

我们将恢复值上限设为 $|\Delta_i|$，并将分数恢复率定义为 RecFrac_i = min(最佳恢复值$_i$, $|\Delta_i|$)/$|\Delta_i|$。我们将 TopLayers_i 表示为达到 ≥90% 恢复率的层集合，同时考虑到可能存在叠加和冗余电路（Elhage 等，2022）。

分层规则系统 直观而言，我们的评估提出了三个问题：(1) 改变模式标记是否会因果性地影响模型的答案偏好，(2) 能否通过修补将此效应定位到某项特定的内部计算，以及 (3) 同一项计算是否会在不同输入之间重复使用？以下定义使用 logit 差值测量和残差流干预，将这些问题形式化。从关于电路存在性的顶层主张出发，我们将其分解为逐例因果测试（R1、R2、R3）。

令 $\tau_{\text{gap}} \geq 0$ 为最小差距阈值，$\alpha \in (0, 1]$为所需的恢复比例。

R1（模式敏感性）。如果数据库模式标记遭到破坏，模型的偏好变化至少达到τ_{gap}：

$$R1_i \ = \ \mathbb{I}\big[|\Delta_i| \geq \tau_{\text{gap}}\big].$$

这项测试用于检验模型对数据库模式标记的因果依赖。满足 $|\Delta_i| < \tau_{\text{gap}}$ 的示例表明，模型忽略了数据库模式信息。通过 R1 表明，模型的答案偏好取决于当前存在的是哪个数据库模式标记。

R2（恢复效能）。如果 R1 成立，并且在位置（p）修补干净的数据库模式标记激活，那么至少可恢复差距的 α比例：

$$R2_i \ = \ R1_i \ \wedge \ \mathbb{I}\big[\text{RecFrac}_i \geq \alpha\big].$$

这项测试用于判断因果效应是否定位于特定层。将恢复率表示为比例（RecFrac_i），可在不受差距大小影响的情况下评估机制是否存在。通过 R2 表明，在数据库模式位置的单一层足以恢复偏好，这说明存在局部化电路。由于叠加或冗余电路的存在，多个层都可能达到 ≥90% 的恢复率，这一点由 TopLayers_i 捕获。

R3（电路可复用性）。如果 R1 和 R2 均成立，且所有样例中最常见的高恢复率层为 L^*，则 $L^* \in \text{TopLayers}_i$：

$$R3_i \ = \ R1_i \ \wedge \ R2_i \ \wedge \ \mathbb{I}\big[L^* \in \text{TopLayers}_i\big].$$

这项测试用于检验不同示例中是否使用一致的层进行恢复。通过 R3 表明，模型对不同查询使用相同的计算路径，说明它会重复利用一致的计算路径，而非依赖针对特定输入的启发式规则。

我们在经过数据库模式训练的模型和无模式基线模型上运行完全相同的评估流程，并在相同的提示集和阈值下比较每个示例的通过率（$\frac{1}{N}\sum_i \overline{R3_i}$）。请记住，只有 R1 和 R2 均通过，R3 才可能通过。

该框架给出示例级通过率，用以估计模型接近完全泛化的程度。将解耦差距显著性（R1）与恢复质量（R2）分开，可独立衡量机制是否存在，而不受任务难度或损坏强度影响。一致性指标（R3）则能区分特定位置的效应与真正的电路复用，从而为局部化计算提供更有力的证据。

结果

标准准确率

TinySQL CS1 同义词测试集的结果揭示了表层评估的一项关键局限：即使测试时不提供数据库模式，未基于模式训练的模型仍达到 93.5% 的表/列准确率。由于该模型无法获取模式信息，这一高性能只能是利用了数据中的伪相关模式。尽管基于模式训练的模型表现更好（提供模式时准确率为 99.1%），但仅凭这些指标无法区分模型究竟真正使用了模式信息，还是只记住了训练模式（表 1）。

训练 数据库模式	Eval 数据库模式	精确 匹配	字段 Acc
✗	✗	**72.1%**	**93.5%**
✗	✓	10.4%	83.6%
✓	✗	0.0%	40.4%
✓	✓	**94.0%**	**99.1%**

表1：不同模型配置的标准准确率。无模式模型通过利用模式实现了93.5%的字段名称准确率，表明传统评估并不可靠。

符号机制结果

在我们的 100 组受控提示对上，基于模式训练的模型对模式破坏表现出显著更高的敏感度（均值 $\Delta_{\text{schema}} = 1.88$ [95% 置信区间：1.44, 2.32]），相比之下，无模式基线为（均值 $\Delta_{\text{NO Schema}} = 0.65$ [95% 置信区间：0.39, 0.90]），由此得到的训练效应为+1.23 [95% 置信区间：0.79, 1.67] 逻辑值（增加 190%）。按类别分析（表2）显示，该效应在

打乱式破坏（Super-Scramble：$+2.35$，非数据库打乱：$+1.61$）中最大，在同义词（$+0.53$ 至$+0.57$）中较小，这表明模型会部分接受语义相关的替换。

应用我们的符号规则框架（R1：$\Delta > 0.4$；R2：恢复 $\geq 90\%$ 的 $|\Delta|$；以及 R3：$L^* \in \text{TopLayers}_i$）**模式模型的总体通过率达到 76%（100 个示例中有 76 个通过），而无模式模型为 59%**[2]（表2）。对于模式模型，各类别的通过率介于 55%（DB 同义词）和 95%（非 DB 扰乱）之间，并在所有类别中始终超过无模式基线。五个类别中均展现出的稳定优势，以及不重叠置信区间所对应的大效应量，支持了该框架的区分能力。

通过的示例表现出很强的电路一致性：在满足 R3（表明三条规则均通过）的 76 个模式模型示例中，第 0–2 层合计占 89%。相比之下，无模式模型呈现出分散式处理（模态层使用率：34%）。这种层级层面的收敛，加之较高的修补恢复率，为基于模式训练的模型中存在局部化、可复用的模式检查电路提供了证据。

类别	效应（$\Delta_S - \Delta_N$）	通过$_S$	通过$_N$
DB 同义词	+0.53	55%	25%
DB 扰乱	+0.51	75%	70%
非 DB 同义词	+0.57	80%	45%
非 DB 扰乱	+2.28	95%	85%
Super-Scr	+0.77	75%	70%
总体	**+0.93**	**76%**	**59%**

表2：100个示例（每类20个）的分类结果。

讨论

我们的结果表明，与标准指标相比，机制评估能更有效地区分算法学习与模式匹配。模式模型的通过率达到 76%，而无模式基线为 59%，两者相差 17 个百分点（相比之下，标准准确率测试中的差距仅为 5 个百分点）。这一反转表明，表面指标严重低估了真正的算法学习与模式利用之间的差异。

在标准指标失效时，符号机制评估能够区分这些情况。符号规则系统提供了可解释的诊断结果：R1 失败表明模型忽略了关键输入特征，R2 失败揭示了模型采用分布式而非局部化计算的方式，R3 失败则表明模型在不同示例中的处理方式不一致。

当前基准测试普遍存在数据泄漏问题，且缺乏可靠的检测共识。模型会利用伪相关模式，即使是在干净测试集中也是如此。我们的无模式基线便证明了这一点：尽管缺少预期算法，字段名称准确率仍达到 93.5%。然而，标准评估无法检测出这种失效模式，而我们的符号—机制方法则能正确识别这一局限。

结论与未来工作

本研究提出了一种机制感知评估框架，将符号规则与机制可解释性相结合，以区分真正的泛化与模式利用。将其应用于自然语言到 SQL 任务后，我们发现，一个字段名称准确率达到 93.5% 的模型，在机制一致性检查中的得分却只有 59%，这表明标准指标无法可靠地评估模型是否采用了预期算法。分层规则系统能够提供可解释的诊断，并可作为处理具有明确定义算法原语的任务（如检索、落地或结构化解析）的模板。

三项局限性指明了富有成效的未来研究方向。第一，我们的规则依赖于需要审慎校准的超参数；要在不同领域系统地进行阈值选择，还需要采用已知真实机制的控制任务。第二，激活修补可能导致分布偏移，从而使模型得以通过混杂路径恢复表现；尽管我们通过多种干预类型进行汇聚验证，未来研究仍应结合路径修补、因果追踪或消融研究进行交叉验证。第三，对于缺乏清晰算法原语的复杂任务，符号规则难以定义，因此我们的框架更适用于针对特定能力的评估，而非整体评估。

更广泛地说，机制感知评估能够揭示准确率指标无法发现的失效模式。随着模型污染日益普遍，模式匹配日趋复杂，验证模型是否实现了预期算法，而非仅仅近似地产生正确输出，变得至关重要。未来的基准测试应要求在报告准确率的同时进行机制认证，尤其是在推理透明度至关重要的高风险领域。

参考文献

Aiyappa, R.; An, J.; Kwak, H.; and Ahn, Y.-y. 2023. 我们能信任对 ChatGPT 的评估吗？载于 Ovalle, A.; Chang, K.-W.; Mehrabi, N.; Pruksachatkun, Y.; Galystan, A.; Dhamala, J.; Verma, A.; Cao, T.; Kumar, A.; and Gupta, R. 编，第 *3* 届可信自然语言处理研讨会论文集（*TrustNLP 2023*），47–54。加拿大多伦多：计算语言学协会。

Chen, G. H.; Chen, S.; Liu, Z.; Jiang, F.; and Wang, B. 2024. 人类还是大语言模型作为评判者？关于判断偏差的研究。载于 Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N. 编，*2024*年自然语言处理实证方法会议论文集，8301–8327。美国佛罗里达州迈阿密：计算语言学协会。

Dodge, J.; Sap, M.; Marasovi´c, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. 文档化大型网页文本语料库：巨型清洁抓取语料库案例研究。收录于 Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t. 编，论文集

[2]替代阈值（R1：$\Delta \in \{0.25, 0.3\}$；R2/R3：$\geq$80%，95%）显示，在所有组合下，模式模型的表现均优于基线 10–20 个百分点，表明这些规则对阈值噪声具有稳健性。

*2021*年自然语言处理实证方法会议，1286–1305。线上及多米尼加共和国蓬塔卡纳：计算语言学协会。

Eldan，R.；和 Li，Y. 2023。TinyStories：语言模型可以有多小，同时仍能说出连贯的英语？arXiv:2305.07759。

Elhage，N.；Hume，T.；Olsson，C.；Schiefer，N.；Henighan，T.；Kravec，S.；Hatfield-Dodds，Z.；Lasenby，R.；Drain，D.；Chen，C.；Grosse，R.；McCandlish，S.；Kaplan，J.；Amodei，D.；Wattenberg，M.；和 Olah，C. 2022。叠加的玩具模型。ArXiv:2209.10652 [cs.LG]。

Gehrmann，S.；Clark，E.；和 Sellam，T. 2023。修复破裂的基础：生成文本评估实践中障碍的综述。载于 Rogers，A.；Boyd-Graber，J.；和 Okazaki，N. 编，计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），14007–14027。加拿大多伦多：计算语言学协会。

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. 自然语言推理数据中的标注伪影。载于 Walker, M.; Ji, H.; and Stent, A. 编，*2018*年北美计算语言学协会分会会议论文集：人类语言技术，第*2*卷（短篇论文），107–112。路易斯安那州新奥尔良：计算语言学协会。

Harrasse, A.; Quirke, P.; Neo, C.; Nathawani, D.; Marks, L.; and Abdullah, A. 2025. TinySQL：用于机制可解释性研究的渐进式文本到 SQL 数据集。载于 Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V. 编，*2025*年自然语言处理实证方法会议论文集，29256–29284。中国苏州：计算语言学协会。国际标准书号 979-8-89176-332-6。

Kim, D.; Shim, G.; Chun, Y.; Kim, M.; 和 Seo, M. 2025. 基准测试剖析：大语言模型基准测试的机制诊断。载于*2025*年自然语言处理实证方法会议论文集，14226–14246。美国佛罗里达州迈阿密：计算语言学协会。

Lee, K.; Ippolito, D.; Nystrom, A.; Zhang, C.; Eck, D.; Callison-Burch, C.; 和 Carlini, N. 2022. 去重训练数据能让语言模型变得更好。载于 Mure- san, S.; Nakov, P.; 和 Villavicencio, A. 编，计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），8424–8445。爱尔兰都柏林：计算语言学协会。

Li, C.; and Flanigan, J. 2024. 任务污染：语言模型可能不再适用于少样本学习。载于*AAAI*人工智能大会论文集，第 38 卷，18471–18480。

Liu, Z.; Kitouni, O.;Nolte, N.; Michaud, E. J.; Tegmark, M.; and Williams, M. 2022. 朝着理解顿悟式泛化迈进：表征学习的有效理论。载于第*36*届神经信息处理系统国际会议论文集

神经信息处理系统，NIPS ’22。美国纽约州雷德胡克：Curran Associates Inc. 国际标准书号 9781713871088。

McCarthy, J. 1995. 具有常识的程序，479–492。美国：美国人工智能协会。国际标准书号 0262621010。McCoy, R. T.; Pavlick, E.; and Linzen, T. 2019. 因错误理由而正确：诊断自然语言推理中的句法启发式。载于 Korhonen, A.; Traum, D.; and M`arquez, L., 编，计算语言学协会第*57*届年会论文集，3428–3448。意大利佛罗伦萨：计算语言学协会。Nakache, D.; Metais, E.; and Timsit, J. F. 2005. 评估与自然语言处理。载于国际数据库与专家系统应用会议，626–632。施普林格。

Nanda, N.; Chan, L.; Lieberum, T.; Smith, J.; and Steinhardt, J. 2023. 通过机制可解释性研究顿悟式泛化的进展度量。载于第十一届学习表征国际会议。

Newell, A.; Shaw, J. C.; and Simon, H. A. 1957. 对逻辑理论机的经验性探索：启发式方法的案例研究。载于*1957*年*2*月*26*日至*28*日西部联合计算机会议论文集：可靠性技术，IRE-AIEE-ACM ’57（西部），218–230。美国纽约州纽约：计算机协会。国际标准书号 9781450378611。

Newell, A.; Shaw, J. C.; and Simon, H. A. 1959. 关于通用问题求解程序的报告。载于国际信息处理会议论文集，256–264。联合国教科文组织。

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU：一种用于机器翻译自动评估的方法。载于第*40*届计算语言学协会年会论文集，311–318。

Power, A.; Burda, Y.; Edwards, H.; Babuschkin, I.; and Misra,V.2022. 顿悟式泛化：在小型算法数据集上超越过拟合的泛化。*arXiv* 预印本 *arXiv:2201.02177*。

Raina, V.; Liusie, A.; and Gales, M. 2024. LLM-as-a-Judge 是否稳健？研究零样本大语言模型评估中的通用对抗攻击。载于 Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., 编，*2024*年自然语言处理实证方法会议论文集，7499–7517。美国佛罗里达州迈阿密：计算语言学协会。

Samuel, V.; Zhou, Y.; and Zou, H. P. 2025. 面向现代大语言模型的数据污染检测：局限性、不一致性与预言机挑战。载于第三十一届计算语言学国际会议论文集，5058–5070。

Xie, X.;Kersting, K.; andNeider, D. 2022. 深度神经网络的神经符号验证。载于第三十一届国际人工智能联合会议（*IJCAI-22*）论文集，3622–3628。

Zhang, H.; Da, J.; Lee, D.; Robinson,V.; Wu,C.; Song, W.; Zhao, T.; Raja, P.; Zhuang, C.; Slack, D.; 等。2024。《

对大语言模型在小学算术方面表现的仔细考察。神经信息处理系统进展，37：46819–46836。