arXiv:2602.01017v1 [cs.LG] 1 Feb 2026

不忠实推理如何从自回归训练中产生？一项合成实验研究

Fuxin Wang*、Amr Alazali、Yiqiao Zhong*

美国威斯康星大学麦迪逊分校，威斯康星州麦迪逊市

摘要

大语言模型（LLMs）生成的思维链（CoT）推理往往是不忠实的：中间步骤可能在逻辑上不一致，或无法反映导向最终答案的因果关系。尽管已有大量实证观察，但人们对 CoT 的基本理解仍然不足——什么构成忠实的 CoT 推理，以及不忠实性如何从自回归训练中产生。我们通过受控合成实验研究这些问题：在噪声数据上训练小型 Transformer，逐步求解模运算算术表达式，这项任务我们称为 *Arithmetic Expression Reasoning*。我们发现，模型能够学习忠实推理，以因果方式遵循潜在的算术规则，但前提是训练噪声低于临界阈值；这一现象可归因于简单性偏差。在较高噪声水平下，训练动态表现出从忠实的逐步推理向不忠实的跳步推理的转变，其间经过一种混合模式；该模式的特征是预测熵暂时增加。机制分析表明，模型通过解决不一致的推理步骤来学习编码内部不确定性，这表明隐式自我验证可能源自自回归训练。代码可在此处获取[2]。

1 引言

大语言模型（LLM）近期在推理任务上取得了卓越表现。这一成功很大程度上归功于思维链推理（Wei 等人，2022b）：模型在给出最终答案之前，会先生成中间步骤。通过使用包含显式推理轨迹的数据进行训练（Lightman 等人，2023年），可以鼓励这种行为，并促进模型学习更简单、具复合性的推理步骤。

然而，思维链推理的机制在很大程度上仍不透明，目前尚不清楚大语言模型是否以及何时真正掌握了支配推理轨迹的规则。尽管在基准测试中表现出色，大语言模型也存在稳定性问题：即使基准测试受到轻微扰动或发生数据污染，评估结果也可能变得不可靠（Alzahrani 等人，2024年；Gulati 等人，2024年；Hosseini 等人，2024年）。此外，思维链推理有时是多余的：大语言模型有时会为自己已经知道答案编造推理步骤，作为事后解释；甚至会在提示引导其得出错误解答时，捏造看似合理的推理（Turpin 等人，2023年；Lanham 等人，2023年；Barez 等人，2025年）。这种模糊性招致了批评，Shojaee 等人（2025年）认为，推理型大语言模型制造了“思维幻觉”。更广泛地说，缺乏对思维链推理的原则性理解，将对人工智能安全和对齐产生深远影响（Betley 等人，2025年）。

能否清晰地定义忠实的思维链推理？ 大语言模型中的不忠实推理已有大量文献记录。例如，Wei 等人（2022b）；Bao 等人（2024年，2025年）表明，模型可能在给出正确解答的同时产生错误推理，也可能在给出错误解答的同时进行正确推理。Turpin 等人（2023年）发现，生成的推理可能受到提示线索或多项选择选项顺序等伪特征的影响。然而，对于何为忠实推理，目前仍未达成共识。

[1]联系邮箱：fwang297@wisc.edu, yiqiao.zhong@wisc.edu

[2]https://github.com/jwtr297/Arithmetic_Expression_Reasoning

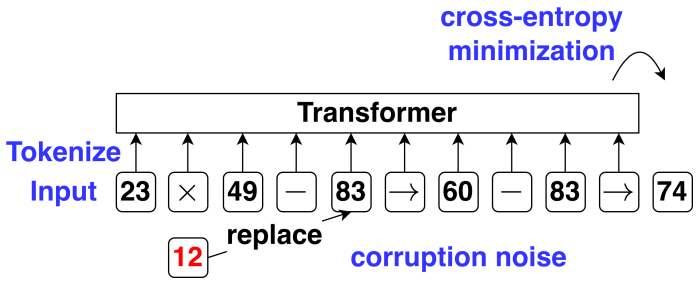


图 1：用于思维链推理的算术表达式推理（AER）任务示意图。对一个符合格式 (1) 的输入序列进行采样和词元化。随后，从头开始采用标准的自回归方式，在此类数据上训练一个小型 Transformer。

现有评估思维链忠实性的方法分为两类。（i）一致性检查，即考察推理步骤在逻辑上是否与最终答案一致（Lyu 等人，2023年）。（ii）对推理步骤的干预，即测试更改推理步骤（例如改述或注入错误）是否会改变最终解答（Turpin 等人，2023年；Lanham 等人，2023年）。这些方法针对思维链推理的两个关键问题：模仿推理轨迹并不能保证逻辑一致性，而推理轨迹的顺序结构也不意味着存在因果关系。受这些实证研究的启发，我们从这两个角度研究忠实推理。

模型对不忠实推理的内部不确定性。大语言模型的一个令人惊讶之处在于其元推理能力：能够反思自身的推理轨迹、检测不忠实推理，从而提高准确率。自我反思和自我验证（Shinn 等人，2023年；Madaan 等人，2023年）会促使模型生成诸如“我不确定”之类的陈述，从而将模型的内部不确定性外显。一个根本性问题是：模型如何通过自回归训练形成这种反思行为？经验证据表明，GPT-3 等基础模型已展现出初步的自我反思能力（Weng 等人，2023年），但其涌现机制仍缺乏充分理解。

通过合成任务对思维链推理进行建模。我们的合成任务抽象了数学推理基准测试 GSM8K（Cobbe 等人，2021年）：在该任务中，大语言模型通过逐步串联多个算术表达式，生成一系列推理词元（即 #### 之前的文本），最终得出最终解答（即 #### 之后的整数）。

提示。“蜡烛每燃烧一小时就会熔短 2 厘米。从下午 1 点燃烧到下午 5 点后，蜡烛会短多少厘米？”

答案。“蜡烛燃烧了 5 - 1 = <<5-1=4>>4小时。因此，蜡烛会短 2 * 4 = <<2*4=8>>8厘米。#### 8。”

我们的合成任务称为算术表达式推理（*AER*）任务，它在保留每个问题算术结构的同时去除了自然语言。形式上，训练数据由按算术表达式链格式排列的序列组成，其中每条链都遵循以下格式：

$$\underbrace{a \times b - c}_{e_1:\text{prompt}} \rightarrow \underbrace{d - c}_{e_2:\text{reasoning}} \rightarrow \underbrace{o}_{e_3:\text{solution}} . \tag{1}$$

图1展示了一个训练样本，该样本被分词后作为模型进行下一词元预测训练的输入序列。在此示例中，我们从 $\mathcal{V}_N := \{0, 1, 2, \ldots, N-1\}$ 中均匀抽取 a、b、c，其中 $N = 97$是一个固定的素数，然后基于模数 N 运算确定 d 和 o。随后，我们通过随机选择 a, b（概率为 ε_1）以及 d（概率为 ε_2），再将其替换为 $\mathcal{V}_N$中的一个均匀随机数，从而引入“扰动噪声”。从概念上看，这种噪声代表大型语料库中的随机损坏错误，以及自然语言在描述数学对象时的歧义。本文聚焦于自回归训练（即下一词元预测），这是预训练与监督微调的标准范式。

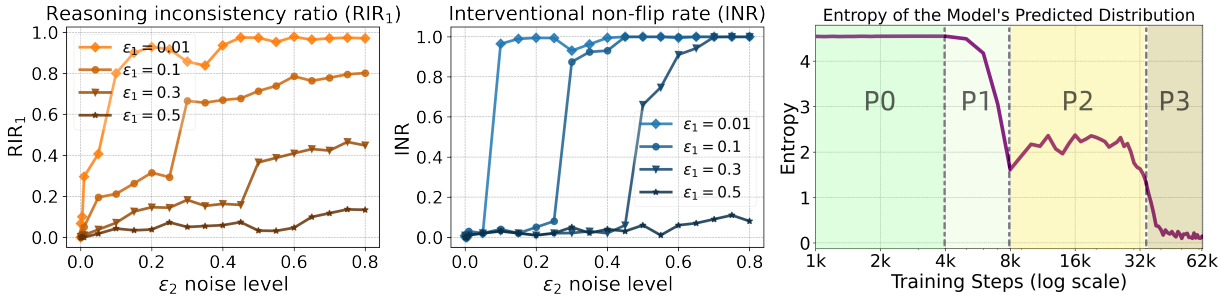


图2：为 AER 任务训练小型 Transformer。**左/中：根据两种忠实性定义评估思维链推理**。我们分别在不同噪声水平的数据集上训练多个 Transformer（ε_1：提示噪声， ε_2：推理噪声）。在噪声较低时，两种不忠实性指标都较小（即推理忠实），直到达到临界阈值 ε_2，超过该阈值后，不忠实性急剧上升——橙色曲线显示生成的推理链变得不那么一致，蓝色曲线则显示对推理的干预几乎不会改变生成的解答。**右：训练过程中的预测熵**。模型经历四个阶段：格式遵循（P0）、逐步推理（P1）、混合推理（P2）和跳步推理（P3）。在 P2 阶段，模型的预测不确定性暂时上升，因为它正在解决推理链中的信息冲突，这表明出现了隐式自我验证行为。

我们的AER任务为思维链推理提供了一种自然且最小（单一推理步骤）的组合结构。与已有文献中的合成任务相比，AER类似于顿悟中的模运算（Power等人，2022），但其关注点在于完整推理链 （包括 e_2 和 e_3），而非仅仅最终输出。这一设计为分析推理组合打开了大门。

主要贡献。

1. 最小化测试平台与形式化定义。我们提出合成任务 AER，并将两种忠实性概念形式化：数学一致性，以及对推理步骤进行反事实干预。2. 由简单性偏差导致的忠实性阈值。我们发现，只有当训练噪声低于临界阈值时，推理才可能保持忠实；算法简单性偏差可以解释这一现象。3. 三种推理模式与涌现式自我验证。在中等噪声下，训练动态呈现出三种不同的推理模式：忠实的逐步推理逐渐转变为不忠实的跳步推理，其间经过一种混合模式；该模式的特征是预测熵暂时上升，表明隐式自我验证行为涌现。

2 设置与定义

我们将训练小型 Transformer 来解决 AER 任务。遵循顿悟论文（Power 等人，2022）中的任务设置，在本文的大多数实验中，我们将模数固定为一个素数 $N = 97$ 。我们的输入序列是由取自 $\mathcal{V}_N = \{0, 1, \ldots, N-1\}$的整数、运算符 $\{+, -, \times\}$以及格式为 (1) 的推导符号 $\{\rightarrow\}$ 组成的字符串。词汇表为 $\{0, 1, \ldots, N-1, +, -, \times, \rightarrow\}$，共计 $N+4$ 个词元。

默认情况下，我们采用标准的 Transformer 架构，包含 3 层、2 个注意力头、128 个嵌入维度，并使用旋转位置嵌入（RoPE）（Su 等人，2024年）。模型使用 AdamW 优化器（Loshchilov 和 Hutter，2019年）进行训练，学习率为 0.001，权重衰减为 0.01，批次大小为 512，并采用标准的自回归损失。实验详情见 A 节。所有实验均使用 A30 图形处理器进行训练。我们还考虑了不同的实验变体，改变模数 N、模型规模、数据格式和训练样本量。结果大体相同，详见 B 节。

2.1 数据生成

下面介绍一个训练样本的生成过程。首先，按照以下规则拼接各个组成部分 e_1^0、 $\to$、 e_2^0、 $\to$、 e_3，生成一条无噪声链： e_1^0包含 a、op_1、 b、op_2、 c，其中 a, b, c是取自 $\mathcal{V}_N$ 的均匀随机整数，op_1 固定为 $\times$，而 op_2 从 $\{+, -\}$ 中均匀抽样； e_2^0和 $e_3 \in \mathcal{V}_N$分别是一个表达式和一个整数，二者通过在模数 N 下计算 e_1^0 得到。

接下来，对于给定的噪声水平 $\varepsilon_1, \varepsilon_2 \in [0, 1]$，我们按如下方式向表达式中注入扰动噪声：(i) 在e_1^0 中，我们以相同概率选择 a 或 b，并以概率 ε_1将选中的变量替换为 $\mathcal{V}_N$中的均匀随机整数，得到一个记作 e_1 的新表达式；(ii) 以概率 ε_2，我们将 e_2^0 中的 d 替换为 $\mathcal{V}_N$中的独立均匀随机整数，得到一个记作e_2的新表达式。作为输入序列提供给 Transformer 的结果链 $e_1 \to e_2 \to e_3$ 由 11 个词元组成。注意， $e_1 = e_1^0$ 的概率为 $1 - \varepsilon_1$，而 $e_2 = e_2^0$ 的概率为 $1 - \varepsilon_2$，因此一部分训练样本包含扰动噪声。

训练分布。令 $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ 分别为 e_1、 e_2、 e_3 的所有可能表达式构成的空间。因此，训练分布 $\mathcal{P}$ 是根据数据生成过程所述、定义在 $\mathcal{E}_1 \times \mathcal{E}_2 \times \mathcal{E}_3$ 上的 (e_1, e_2, e_3) 的联合分布。训练样本总数为 2,000,000。

作为复合的算术规则。令 $f_1 : \mathcal{E}_1 \to \mathcal{E}_2$ 为正确推导， $f_2 : \mathcal{E}_2 \to \mathcal{V}_N$ 为正确解。例如，若 $e_1 = 2 \times 9 - 19$, $e_2 = 18 - 19$ 和 $e_3 = 96$，则 $f_1(e_1) = e_2$ 且 $f_2(e_2) = e_3$。思维链推理会明确地将算术规则表示为复合 $f_2 \circ f_1$，记作 f。目标是学习推理链：给定 $e_1 \in \mathcal{E}_1$，模型理想情况下会生成

$$e_1 \to f_1(e_1) \to \underbrace{f_2(f_1(e_1))}_{f}. \tag{2}$$

若链 $e_1 \to e_2 \to e_3$ 满足 $e_2 = f_1(e_1)$ $e_3 = f_2(f_1(e_1))$，即满足上述算术规则，我们就称其一致。推理链也可以等价地写成元组 (e_1, e_2, e_3)，即 $e_1 \to e_2 \to e_3$。

2.2 忠实推理的两种定义

Transformer 是一种概率模型 p，它根据上下文给出概率分布。具体而言，条件概率 $p(e_2|e_1)$ 和 $p(e_3|e_1, e_2)$ 刻画了模型的行为。设 $\mathcal{F}_1$ 为从 $\mathcal{E}_1$ 映射到 $\mathcal{E}_2$的函数空间， $\mathcal{F}_2$ 为从 $\mathcal{E}_1 \times \mathcal{E}_2$ 映射到 $\mathcal{V}_N$的函数空间。我们分别定义 $\widehat{\ }$ $\widehat{\ }$两个随机函数 f_1， f_2，它们是取值于 $\mathcal{F}_1$和 $\mathcal{F}_2$的随机元素，并满足对于所有e_1， e_2， e_3，

$$\mathbb{P}\big(\widehat{f}_1(e_1) = e_2\big) = p(e_2|e_1),$$
$$\mathbb{P}\big(\widehat{f}_2(e_1, e_2) = e_3\big) = p(e_3|e_1, e_2).$$

$\widehat{\ }$ $\widehat{\ }$在现实中，由于多种因素（例如噪声、欠拟合、过拟合），我们预期 f_1， f_2 会以正概率偏离 f_1, f_2，也就是说，未能准确学习 f_1 和 f_2的（真实）算术规则。

基于一致性的忠实性。对于测试提示 e_1，我们根据给定的采样方案（例如贪心解码或在特定温度下采样）生成一条链 (e_1, e_2', e_3')。一种常见的刻画推理行为的方法，是将 (e_1, e_2', e_3') 与一致的真实链 (e_1, e_2, e_3) 进行比较（表 1 ）。如果对于每个提示 e_1， $e_2 = e_2'$ 且 $e_3 = e_3'$，我们就称思维链推理是忠实的。

这是一个不涉及因果依赖的忠实性观察性概念。尽管概念上很简单，但它有两个缺点：(i) 一致链的集合基数为 $|\mathcal{E}_1|$，仅占所有可能链（基数为 $|\mathcal{E}_1| \times |\mathcal{E}_2| \times |\mathcal{E}_3|$）的一小部分。因此，我们无法保证分布外鲁棒性。(ii) 即使模型生成了一条一致链，我们也无法区分真正的逐步推理与事后解释；后者缺少预期的因果联系。

表1：**基于一致性的推理分类**。 给定测试实例 e_1，将模型链（e_1， e'_2， e'_3）与真实标注链（e_1， e_2， e_3）进行比较。此处 e'_2 为推理步骤， e'_3 为解答；R = 推理，S = 解答。

Type	一致性检查	结果
忠实	$e'_2 = e_2$， $e'_3 = e_3$	推理正确；解答正确
不忠实	$e'_2 \neq e_2$， $e'_3 = e_3$	推理错误；解答正确
	$e'_2 = e_2$， $e'_3 \neq e_3$	推理正确；解答错误
	$e'_2 \neq e_2$， $e'_3 \neq e_3$	推理错误；解答错误

基于干预的忠实性。我们通过考虑 e_2 的反事实来提出一种更强的忠实性定义，使得 e_1 和 e_2 不一定一致，尤其是在低噪声设置下。该定义要求对推理 e_2 进行干预，并观察解答 e_3 随之产生的变化——特别是，当 e_1 和 e_2 不一致时，预测 e_3 在多大程度上取决于 e_1 和 e_2。以下定义刻画了两种极端模式。

定义 2.1。 (i) **（完美）逐步推理**：若模型的思维链推理以概率 1 满足以下条件，我们就称其遵循（完美）$\widehat{}$逐步推理： $f_2(e_1, e_2) = f_2(e_2)$ 对所有 $e_1 \in \mathcal{E}_1$， $e_2 \in \mathcal{E}_2$ 均成立。(ii) **（完美）跳步推理**：若模型的思维链推理以概率 1 满足以下条件，我们就称其遵循（完美）跳步推理 $\widehat{}$： $f_2(e_1, e_2) = f(e_1)$ 对所有 $e_1 \in \mathcal{E}_1$， $e_2 \in \mathcal{E}_2$ 均成立。

$\widehat{}$完美逐步推理意味着， f_2 仅取决于推理表达式 e_2 ，而与 $\backsim e_1$ 无关。同样，完美跳步推理意味着， f_2 仅取决于 e_1 ，而忽略推理表达式e_2。根据此定义，若模型的思维链推理遵循完美逐步推理 $\widehat{}$，并且对所有 $e_1 \in \mathcal{E}_1$ ，以概率 1 满足 $f_1(e_1) = f_1(e_1)$，则该模型的思维链推理完全忠实。显然，这种基于干预的忠实性意味着基于一致性的忠实性。

2.3 指标

训练期间，我们每经过 1,000 个训练步数，就在 1,000 个留出测试样例上使用一组指标评估模型。我们将给定 e_1 和指定解码方案下生成的推理链记作 (e'_2, e'_3)。

基于一致性的指标。 我们通过采样$n = 1,000$ 个遵循训练分布的测试提示（$e_{i,\,1}$）来计算表1中四类情况的比例，并使用模型在固定温度 1 下生成 $(e'_{i,\,2},\ e'_{i,\,3})$。我们主要考虑推理不一致率（RIR）的两个指标：

$$\mathrm{RIR}_1 := \frac{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2}, e_{i,3} = e'_{i,3}\}}{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2}\}} \tag{3}$$

$$\mathrm{RIR}_2 := \frac{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2}, e_{i,3} = e'_{i,3}\}}{\sum_{i\le n} \mathbf{1}\{e'_{i,3} = e_{i,3}\}}. \tag{4}$$

这两个指标的分母不同，在量化不忠实性方面互为补充（值越高表示越不忠实）。直观而言，RIR_1 衡量生成错误推理 e_2时恢复出正确解 e_3的比例；而 RIR_2 衡量解答正确的测试样本中推理错误的比例。

基于干预的指标。 我们将对推理 e_2 施加干预，并测量模型在多大程度上符合定义 2.1 中的两种分类。我们将重点关注均匀随机干预，即用 $\mathcal{V}_N$ 中的均匀随机变量替换 e_2 中的 d ，从而得到 $\tilde{e}_2$。对于 n 独立测试提示

$(e_{i,\ 1})_{i\le n}$、模型生成的推理 $(e'_{i,\ 2})_{i\le n}$，以及替换后的推理 $(\tilde{e}_{i,\ 2})_{i\le n}$，我们计算两个指标：干预分布敏感度（IDS）和干预非翻转率（INR）：

$$\mathrm{IDS} := \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}\big(\widehat{f}_2(e_{i,1}, e'_{i,2}), \widehat{f}_2(e_{i,1}, \tilde{e}_{i,2})\big);$$

$$\mathrm{INR} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\arg\max_{e\in\mathcal{V}_N} p(e|e_{i,1}, e'_{i,2}) = \arg\max_{e\in\mathcal{V}_N} p(e|e_{i,1}, \tilde{e}_{i,2})\}.$$

其中， D_{KL} 表示KL散度。较小的 IDS 和较大的 INR 表明 e_2 对 e_3的因果效应较弱，这是跳步推理的迹象。反之，较大的 IDS 和较小的 INR 则是逐步推理的证据。

预测不确定性指标。 我们使用预测熵来衡量模型的不确定性：对于测试提示（$e_{i,\ 1}$）$_{i\le n}$ 以及模型生成的推理链（$e'_{i,\ 2}$，$e'_{i,\ 3}$）$_{i\le n}$，我们计算预测熵（PE）：

$$\mathrm{PE} := \frac{1}{n}\sum_{i=1}^{n}\sum_{e\in\mathcal{V}_N} -p(e|e_{i,1}, e'_{i,2}) \log\, p(e|e_{i,1}, e'_{i,2}). \tag{5}$$

3 主要结果

3.1 噪声如何导致不忠实性?

我们构造了具有不同噪声水平 ε_1 和 ε_2的训练数据。具体而言，当$\varepsilon_1 = \varepsilon_2 = 0$时，训练数据不含噪声。对于（$\varepsilon_1$，$\varepsilon_2$）的每种组合，我们训练一个 Transformer 共 62,500 步，并在最终检查点对模型进行评估。

我们的目标是探究影响推理忠实性的因素，尤其是噪声水平。虽然我们预期较低的噪声水平会提升基于一致性的忠实性，但基于干预的忠实性是否可能实现，事先并不明朗。考虑使用不含噪声的训练数据$\widehat{\ }$（$\varepsilon_1 = \varepsilon_2 = 0$）训练模型。仅观察一致链时，模型理论上可以将解答 f_2（e_1，e_2）学为 f_1（e_1）、 f_2（e_2），或两者的混合形式——这些情况都能实现零损失和完美的基于一致性的忠实性。然而，只有学会 f_2（e_2）才符合基于干预的忠实性标准。因此，我们提出以下问题——基于干预的忠实性究竟能否实现？如果能，哪些因素会对此有所影响？

在低于临界噪声水平时，基于干预的忠实性是可能的。 我们发现，忠实推理是可以实现的，但前提是噪声水平足够低。在图2（左/中）中，我们将提示噪声水平设为 $\varepsilon_1 \in \{0.01, 0.1, 0.3, 0.5\}$，并改变推理噪声水平 $\varepsilon_2 \in \{0.001, 0.005, 0.01, 0.05, 0.1$,0。3、0。5、 $0.9\}$。我们使用两个指标来衡量不忠实性，即基于一致性的 RIR_1 和基于干预的 INR，并发现这两个指标都会随着噪声水平的升高而不降低。

此外，图2揭示了临界阈值 $\tau_c(\varepsilon_1)$ 存在于 ε_2 中，且每个 ε_1 都对应一个这样的阈值。事实上，在阈值 $\tau_c(\varepsilon_1)$ 以下，两个不忠实性指标都近似保持平坦并接近 0；超过阈值后，它们则会大幅上升。这表明，在最终检查点， e_1 与 e_2 作为降低损失的预测信号之间的竞争，在很大程度上决定了推理行为。

有趣的是，在无噪声设置下（$\varepsilon_1 = \varepsilon_2 = 0$），这两种不忠实性指标都接近 0，这表明仅使用一致链进行训练也能实现忠实推理。更一般地说， $\varepsilon_1 < \tau_c$（ε_1），这意味着在低噪声条件下，模型表现出惯性：即使跳步推理的损失略低，它仍会继续采用逐步推理，而不是立即切换到跳步推理。这促使我们研究噪声水平以外的算法效应。

忠实性受益于简单性偏差（低复杂度偏好）。我们假设，如果两种机制产生的准确率相近，训练算法会倾向于学习较简单的规则（涉及单个算子的逐步推理），而非较复杂的规则（涉及两个算子的跳步推理）。这一假设基于奥卡姆剃刀，以及近期关于过参数化模型中归纳偏置的理论研究

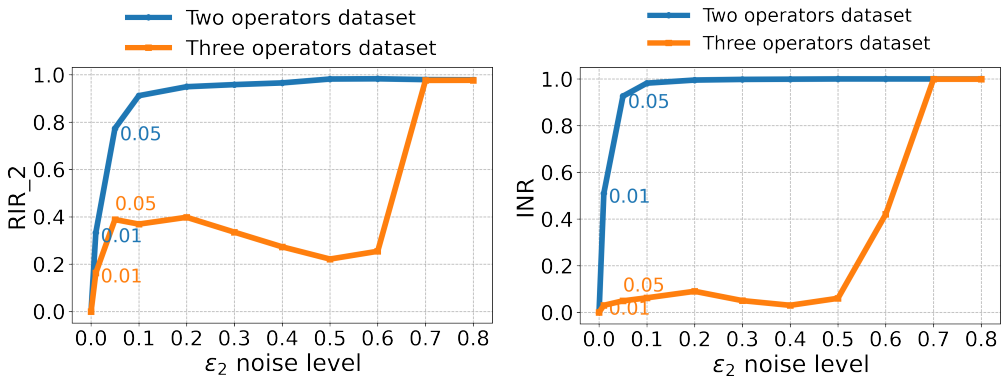


图 3：**与 e_2 之间的复杂度差距越大，$\varepsilon_1 = 0.01$不忠实性越低**。固定 $\varepsilon_1 = 0.01$，我们遍历 ε_2，并比较在双运算符提示 e_1 和三运算符提示上训练 Transformer 的结果。两项指标均表明，难度差距越大，越能促进逐步推理并降低不忠实性。

神经网络（So udry 等人，2018年；Montanari 和 Zhong，2022年；Abbe 等人，2024年），其中模型更倾向于 r 平滑、低次、 以及低复杂度函数。

为验证这一假设，我们修改实验以增大提示与推理之间的复杂度差距：在数据生成过程中，将 $e_1^0 = a \times b - c$ 替换为 $(a + b) \times c - d$，并在 e_1 中以噪声水平 ε_1 随机替换 a、b或 c 来注入噪声，同时保持针对 e_2 的噪声注入方案不变；我们还更改了模数为 $N = 29$。这一修改提高了 $e_1 \to e_2$的推理复杂度，因为它每一步会计算两个运算符。图 3 显示，e_1 与 e_2之间的复杂度差距越大，不忠实性越低；而且，模型在转变为跳步推理之前能够容忍更高的推理噪声。结果表明，推理链中连续步骤之间更大的复杂度差距会增强简单性偏差，从而为模型发展出不忠实的跳步推理设置一道算法障碍。

最后，我们表明，推理步骤之间的复杂度差距——而非相对位置——构成了算法归纳偏置，从而排除了一个混杂因素。具体而言，我们交换提示和推理的相对位置，使修改后的链变为 $e_2 \to e_1 \to e_3$。B.1节表明，推理忠实性取决于复杂度，而非相对位置。

实际意义。训练数据中的噪声较低，且推理步骤之间的复杂度差距较大，这对于在思维链推理中诱导出较高的忠实性至关重要。

3.2 训练过程中的推理模式阶段

推理模式的阶段特征。我们发现，在一项实验中，模型的推理模式会随着训练而演变。为便于说明，我们将噪声水平固定为 $\varepsilon_1 = 0.01$，$\varepsilon_2 = 0.1$ 以生成训练数据，并对模型进行 62,500 步训练。我们主要关注模型在训练过程中的概率分布 ^，即 $f_2(e_1, e_2)$。我们假设，模型会根据训练步数 t呈现以下四个阶段。令 (e_1, e_2) 服从与训练数据相同的分布。

- 阶段*0*（格式遵循）：在正确位置预测运算符标记“+”或“−”，以及演绎“→”。⌢
- 阶段*1*（逐步推理）：预测主要取决于 e_2，也就是说，$f_2(e_1, e_2) \approx f_2(e_2)$。⌢
- 阶段*2*（混合推理）：记指示函数 $c(e_1, e_2) = \mathbf{1}\{f(e_1) = f_2(e_2)\}$。那么，$f_2(e_1, e_2) \approx f_2(e_2)c(e_1, e_2) + U(1 - c(e_1, e_2))$，其中 U 是 $\mathcal{V}_N$ 中的均匀随机变量。也就是说，如果 e_1 和 e_2 不一致，则预测最不确定（服从均匀分布）；否则，预测与两者均一致：e_1, e_2。⌢
- 阶段 *3*（跳步推理）：预测主要取决于 e_1，也就是说，$f_2(e_1, e_2) \approx f(e_1)$。

我们使用若干评估指标验证这一假设性刻画（图 4 和图 11）。**(i)格式得分**：给定测试提示后，我们生成完整的推理轨迹和解答，并计算

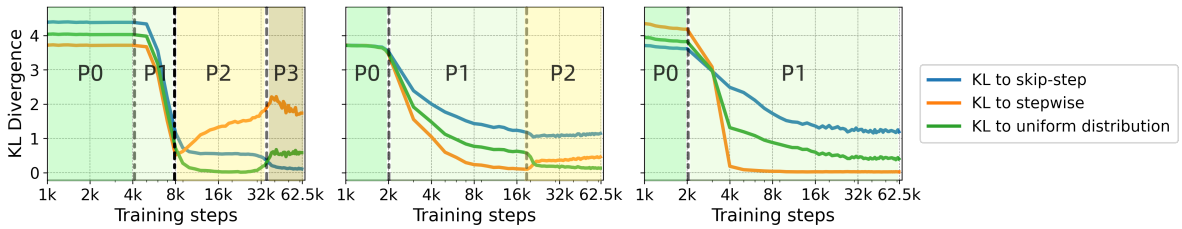


图 4：**训练动态呈现出推理模式的不同阶段**。我们在不同噪声配置下进行了三项实验，并使用三种指标评估模型（对数横轴）。**左图**：噪声水平（ε_1，ε_2）=（0.01，0.1）。模型呈现出四个不同的推理阶段，验证了我们提出的阶段刻画。**中图**：（ε_1，ε_2）=（0.1，0.1）。模型仅呈现前三个阶段，因为相等噪声水平会促使模型停留在混合推理模式 P2。**右图**：（ε_1，ε_2）=（0.1，0.01）。模型仅呈现前三个阶段，因为较低的提示噪声会促使模型采用逐步推理 P1。

在正确位置生成运算符和推理标记的比率。(ii) 与逐步推理的 KL 散度：我们计算模型的预测概率与完美逐步推理 $n^{-1}\sum_{i\le n} D_{\mathrm{KL}}(\,\widehat{f}_2(e_i,1,e'_i,2)\;\; f_2(e'_i,2))$ 之间的 KL 散度，其中 $(e_i,1,e'_i,2)_{i\le n}$ 基于 $n = 1,000$ 个测试提示 $e_i,1$ 生成。(iii) 与跳步推理的 KL 散度：我们计算模型的预测概率与完美跳步推理 $n^{-1}\sum_{i\le n} D_{\mathrm{KL}}(\,\widehat{f}_2(e_i,1,e'_i,2)\;\; f(e_i,1))$ 之间的 KL 散度，其中 $(e_i,1,e'_i,2)_{i\le n}$ 同样通过抽样生成。(iv) 与均匀分布的 KL 散度：我们计算模型的预测分布与均匀分布 $n^{-1}\sum_{i\le n} D_{\mathrm{KL}}(\,\widehat{f}_2(e_i,1,e'_i,2)\;\; U)$ 之间的 KL 散度，且仅在不一致推理链上评估，其中 $f(e_1) \neq f_2(e_2)$。我们指出，阶段1–3中的推理模式具有因果解释；更多说明请参见 B.2 节和图 10。

我们的结果支持所提出的阶段刻画。B.2节中的图11显示，模型在阶段0快速学会格式遵循。在图 4（左图）P1中，橙色曲线迅速降至最小值，且在三种KL散度指标中数值最低，这表明模型在训练早期阶段更倾向于逐步推理。在P2中，橙色曲线上升，而蓝色曲线下降至约0。这表明模型不再采用逐步推理，并学会为不一致推理链（e_1，e_2）预测均匀分布（信息量最低）。在P3中，蓝色曲线进一步下降至最小值，且在三种指标中数值最低，这表明模型最终学会了由损失最小化驱动的跳步推理。

我们还发现，在不同的噪声配置下，模型可能只达到阶段1或阶段2。图 4（中）显示，使用相等噪声水平训练的模型最终会保持在混合推理模式，因为同时结合提示表达式和推理表达式会产生最小的损失。图 4（右）展示了另一项实验，其中推理噪声较小，这会抑制模型依赖提示表达式 e_1。

决定推理模式的因素。基于上述实验，我们推测，决定模型推理模式的两个因素是数据质量（即噪声水平）和计算量（即训练步数）。我们固定 $\varepsilon_1 = 0.1$，并让推理噪声水平 ε_2 在 0.001 到 0.9 之间变化。对于每个噪声水平和训练期间的评估检查点，我们根据三项 KL 指标中哪一项 KL散度最小，将模型的推理模式划分为 P1、P2 或 P3。B.2节中的图 9 显示，更多计算量和更高的提示噪声通常会促使模型转向跳步推理。作为一项附带发现，我们还发现阶段边界并不规则，且会随不同的随机种子而变化，这表明优化景观十分复杂，也印证了大语言模型中涌现现象的不可预测性（Wei等人，2022a）。

实际意义。长时间进行思维链推理的自回归训练（尤其是监督微调）可能会促使模型采用跳步推理，从而降低其忠实性。

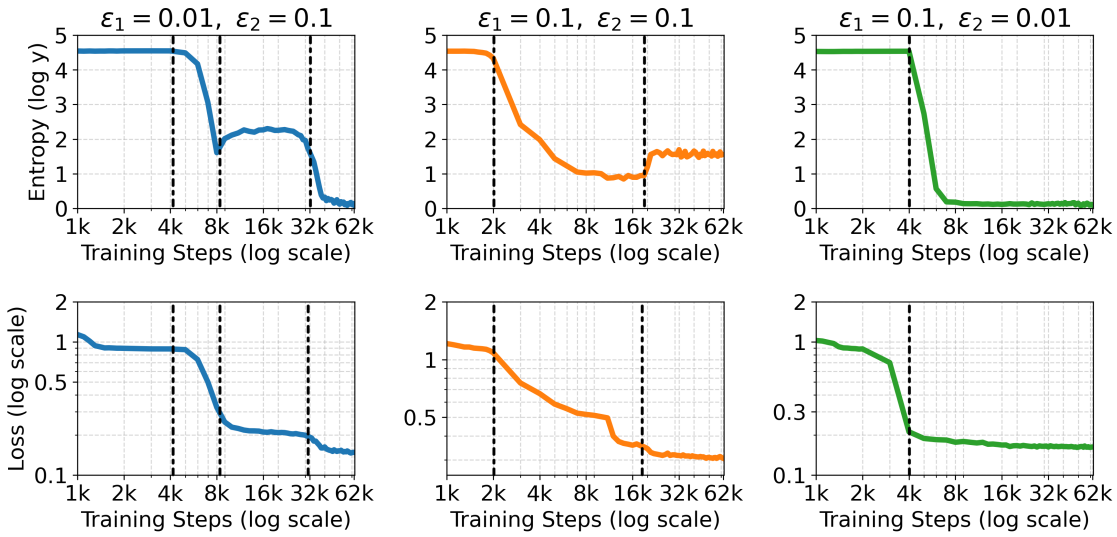


图 5：**尽管训练损失持续下降，预测熵仍会暂时上升**。模型在混合推理模式下编码推理不确定性，这与损失的大多数急剧下降相吻合。

3.3 隐式自我验证

自我反思和自我验证等元推理能力，是推理扩展中涌现出的有趣技能（Guo 等，2025）。我们旨在深入考察混合推理模式，并探讨以下问题——模型是否学会了编码推理不确定性？如果是，其内部机制是什么？

混合推理模式下的预测熵会暂时上升。我们使用预测熵 (5) 来量化模型的不确定性，这一方法已用于近期的实证论文（Fu 等，2025；Zhao 等，2025）。存在不一致推理链时（e_1，e_2），我们在阶段2对混合推理模式的刻画表明，模型输出趋近于近似均匀分布，对应于最大不确定性。图 5（上）证实了这一现象：在图 4 所示的三种噪声配置下，预测熵呈现出明显的非单调趋势，在模型进入混合推理模式时上升。

图 5（下）展示了训练损失的变化。与预测熵不同，损失如预期般单调下降。然而，损失的大多数急剧下降都与推理模式之间的转变密切吻合。这种对应关系表明，推理行为和不确定性的变化反映在损失景观的几何结构中。

机制分析揭示了不确定性编码特征。 为了探究模型如何在内部表征不确定性，我们考察了训练过程中顶层隐藏状态的演变。图 6 中的主成分分析可视化显示，隐藏状态编码了不同的推理阶段，其轨迹方向发生了明显变化，并与阶段转变相对应。

此外，我们还考察隐藏状态的某些坐标是否编码了不确定性特征，以及注意力的影响。为此，我们引入：

$$\text{Hidden State Contrast (HSC)} :=$$
$$\frac{1}{n}\sum_{i=1}^{n}\left|\left\|h_{L,e_3}(e^+_{i,1}, e^+_{i,2})\right\|_\infty - \left\|h_{L,e_3}(e^-_{i,1}, e^-_{i,2})\right\|_\infty\right|,$$
$$\text{Attention Contrast (AC)} :=$$
$$\frac{1}{n}\sum_{i=1}^{n}\left\|h^+_{L-1,i}W^{(L)}_{QK}(h^+_{L-1,i})^\top - h^-_{L-1,i}W^{(L)}_{QK}(h^-_{L-1,i})^\top\right\|_{\text{op}},$$

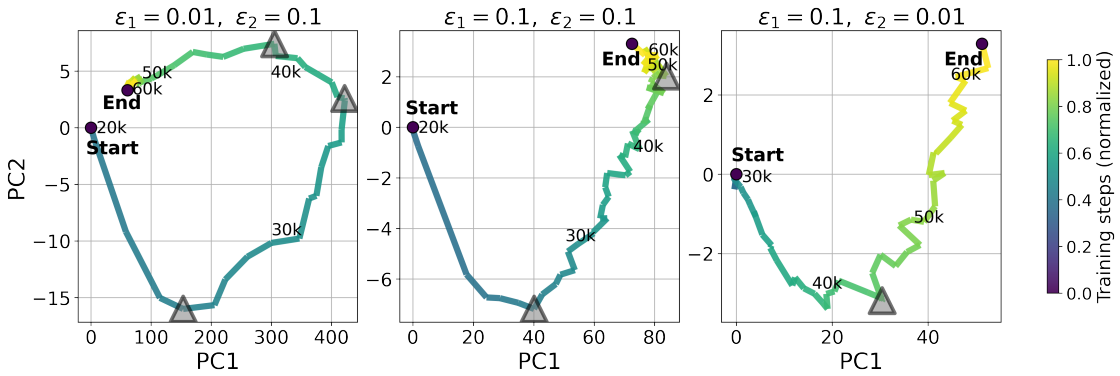


图 6：**隐藏状态的 PCA轨迹可视化了推理阶段**。在三种实验设置下，我们提取顶层隐藏状态，并将其投影到主子空间。**左图**：噪声水平（ε_1，ε_2）=（0.01，0.1）。三角形标示运动方向的变化，对应三个阶段。**中图**：噪声水平（ε_1，ε_2）=（0.1，0.1）。两个三角形标示阶段变化。**右图**：噪声水平（ε_1，ε_2）=（0.1，0.01）。一个三角形标示唯一一次方向变化。

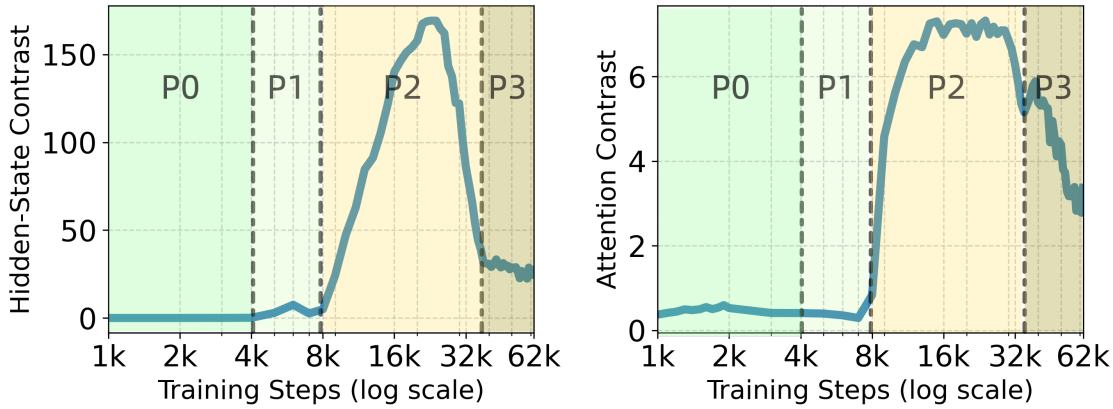


图 7：**通过机制分析揭示内部不确定性的涌现：**噪声水平为 $\varepsilon_1 = 0.01$ 和 $\varepsilon_2 = 0.1$时 HSC 和 AC 的变化。两个指标均经历了急剧上升，随后又下降，这与预测熵的阶段变化相吻合。

其中，$\|\cdot\|_{\text{op}}$ 表示矩阵算子范数（见 B.3 节）。这里，我们将一致（无噪声）的链（e_1^+，e_2^+）作为**正**样本，并将不一致（有噪声）的链（e_1^-，e_2^-）作为**负**样本，其噪声水平为$\varepsilon_1 = 0$，$\varepsilon_2 = 1$（即，e_1 和 e_2 总是得出不同结果）。我们用$h_{L,\ e_3}(e_1^+,\ e_2^+)$表示最后一层 L中解答 e_3 位置处的 FFN 后输出向量，用$h^+_{L-1,\ i}$表示第 i个正样本在倒数第二层 $L-1$ 的隐藏状态矩阵。$h_{L,\ e_3}(e_1^-,\ e_2^-)$和$h^-_{L-1,\ i}$的定义与此类似。这两个量衡量正样本与负样本之间的区分程度；值越大，表明内部区分越强，其作用类似于一致性检查。

图 7 显示，两个指标在推理模式的相同阶段阈值处都出现了急剧变化。这表明模型在内部编码了 e_1 和 e_2 之间的一致性检查，类似于一种自我验证机制，并在外部表现为预测不确定性。

实际意义。推理不确定性通过自回归训练（包括预训练和监督微调）编码在模型中，成为自我验证等元推理能力的基石。

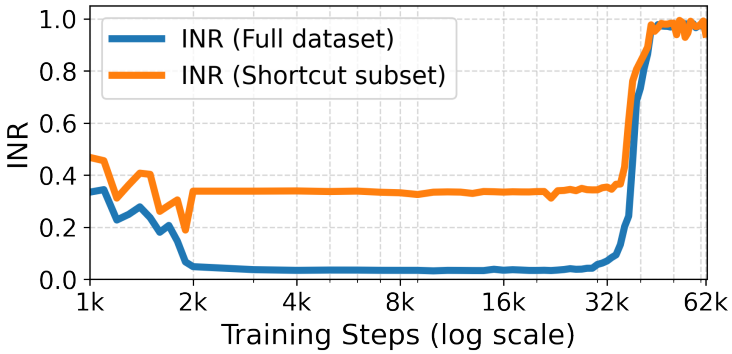


图 8：**捷径数据集和完整数据集上基于干预的指标的训练动态**。作为不忠实性度量的 INR 会在早期变大，这表明利用捷径特征会损害推理忠实性。

4 捷径特征如何加剧不忠实性

实证研究表明，大语言模型容易利用与解答存在虚假相关性的捷径特征（Turpin 等人，2023），这会促使模型绕过逐步推理。例如，在提示中加入“某位著名教授认为解答是选项 C”这一线索，会促使模型不论答案是否正确，都遵从所建议的答案，并对该答案进行不忠实的事后合理化。我们将指令提示视为可用于绕过推理的捷径特征。

我们通过修改实验设置来研究捷径特征的影响。我们将模数替换为一个合数 $N = 38$，并以不同的格式采样提示和推理轨迹，例如：

$$(a - b) \times c \;\to\; d \times c \;\to\; o\,,$$

其中，e_1 中的第一个算子从 $\{+, -\}$中均匀采样，第二个算子为 $\times$。与之前一样，所有表达式均在模数 N下确定。

捷径特征。我们将 $\mathbf{1}\{c = 0$ 或 $c = 2$ 或 $c = N/2\}$ 视为捷径特征，它与解答直接相关。事实上，当 $c = 0$ 或 $c = 2$ 或 $c = N/2$时，$o = 0$的概率很高，无需推导中间推理步骤 e_2。为评估捷径特征对推理忠实性的影响，我们使用两个测试数据集：(i) 完整测试数据集：$n = 1,000$ 个符合训练分布的测试提示；(ii) 捷径测试子集：从额外测试集中选出的 1000 个样本，其中$c \in \{0, 2, N/2\}$ 出现在提示表达式 e_1中。

在图 8 中，我们报告了噪声设置（ε_1，ε_2）$=$(0.01，0.1) 下训练过程中的基于干预的不忠实性指标。在整个训练过程中，INR 在捷径测试集上始终低于完整测试集，这表明只要存在捷径，模型就会更加依赖跳步推理。我们还计算了基于一致性的指标，例如 RIR_1 (B.4 节)，从互补的角度印证了这一行为。

5 相关工作

合成任务上的训练动态。为理解非推理型大语言模型的学习动态，研究者提出了受控良好的合成任务，作为开展大规模实验的替代方案。例如，解决模运算（如加法、乘法）需要学习代数规则，并会呈现出引人入胜的顿悟现象（Power 等人，2022）：模型早早便记住了训练数据，却要过很久才突然实现泛化。对于复制任务，Elhage 等人（2021）、Olsson 等人（2022）和 Song 等人（2025）指出，归纳头是从上下文中复制任意序列的关键组件。通过上下文线性回归（Garg 等人，2022；Von Oswald 等人，2023），研究表明，上下文学习的行为类似于一种

元学习算法，与梯度下降相似。然而，这些工作都没有关注思维链推理或测试时组合。相比之下，我们的 AER 任务旨在填补这一空白。

思维链推理的忠实性。 先前的工作对预训练大语言模型的推理忠实性给出了经验性描述。研究表明，伪特征和提示线索会塑造所生成的推理，这表明思维链可能是对答案的事后合理化，而非因果推理过程（Turpin 等人，2023年；Arcuschin 等人，2025年；Chen 等人，2025年；Chua 与 Evans，2025年；Guan 等人，2025年）。为应对这一问题，Turpin 等人（2023年）、Lanham 等人（2023年）以及 Bao 等人（2024年、2025年）提出了基于干预的指标，用以评估推理轨迹内部的因果关系。然而，由于缺乏针对思维链推理的干净的实验设置，现有工作缺少一个关于忠实性的有原则的框架。此外，它们关注的是静态模型，而非不忠实性的涌现，而后者对于理解非预期推理行为的成因至关重要，并对错位与人工智能安全具有启示意义。相比之下，我们的工作采用干净的实验设置来研究学习动态。

自我反思与不确定性度量指标。 自 GPT-3 以来，大语言模型展现出惊人的元推理能力（Brown 等人，2020；Weng 等人，2023）。通过迭代精炼（Madaan 等人，2023）和后训练（Guo 等，2025），自我反思和自我验证能力进一步增强，显著减少了不一致性和不忠实性。近期研究展示了利用此类指标改进大语言模型推理的潜力（Fu 等，2025；Zhao 等，2025）。然而，现有的实证进展对于底层机制几乎没有提供什么见解：这些元推理能力如何通过训练发展起来，以及模型如何在内部表征这些能力。我们的工作通过提供一种机制性解释来弥补这一空白：随着训练的进行，模型在解决推理轨迹中的不一致性时，学会编码一致性检查和内部不确定性。

6 局限性与未来工作

我们的工作聚焦于自回归训练，这是预训练和监督微调（SFT）的标准训练范式。未来，研究带验证器奖励的强化学习（RLVR）（Guo 等，2025）将很有意义；这种方法正成为增强大语言模型推理能力的标准方案。我们也有兴趣将我们的指标应用于合成设置之外的大语言模型。

影响声明

我们的工作通过受控合成实验分析思维链推理的基本机制。我们没有训练新模型，也没有研究可能导致有害应用的实用改进。我们的工作带来的主要积极影响是增进对思维链推理的科学理解，尤其是对自回归训练中不忠实性机制的理解。通过明确的定义、评估指标和有原则的分析，我们的结果或可为提升大语言模型推理能力的可靠性提供启示。具体而言，我们关于模型内部不确定性和隐式自我验证的发现，有助于理解元推理能力，并可能推动更安全的微调技术。潜在的负面影响是对我们的发现作出超出合成任务范围的过度泛化。在将结论用于政策制定或部署决策之前，还需要针对大语言模型和各类任务开展更广泛、深入的评估。

致谢

Y.Z. 部分获得美国国家科学基金会数学科学部（NSF-DMS）拨款 2412052 和 Coefficient Giving（前身为 Open Philanthropy）拨款的支持。我们也感谢余彬提供有益的讨论。

参考文献

Emmanuel Abbe、Samy Bengio、Aryo Lotfi 和 Kevin Rizk。未见数据上的泛化、逻辑推理与学位课程。机器学习研究期刊，25(331):1–58，2024年。

Norah Alzahrani、Hisham Alyahya、Yazeed Alnumay、Sultan Alrashed、Shaykhah Alsubaie、Yousef Al- mushayqih、Faisal Mirza、Nouf Alotaibi、Nora Al-Twairesh、Areeb Alowisheq 等。当基准测试成为目标：揭示大语言模型排行榜的敏感性。载于 第*62*届年会暨计算语言学协会会议论文集（第*1*卷：长篇论文），第13787–13805页，2024年。Iv´an Arcuschin、Jett Janiak、Robert Krzyzanowski、Senthooran Rajamanoharan、Neel Nanda 和 Arthur Conmy。在真实环境中的思维链推理并不总是忠实。*arXiv* 预印本 *arXiv:2503.08679*，2025年。Guangsheng Bao、Hongbo Zhang、Linyi Yang、Cunxiang Wang 和 Yue Zhang。带思维链的大语言模型是非因果推理者。*CoRR*，2024年。Guangsheng Bao、Hongbo Zhang、Cunxiang Wang、Linyi Yang 和 Yue Zhang。带思维链的大语言模型有多大可能会模仿人类推理？载于 第*31*届国际计算语言学会议论文集，第7831–7850页，2025年。Fazl Barez、Tung-Yu Wu、Iv´an Arcuschin、Michael Lan、Vincent Wang、Noah Siegel、Nicolas Collignon、Clement Neo、Isabelle Lee、Alasdair Paren 等。思维链并不等同于可解释性。预印本，*alphaXiv*，第v1页，2025年。Jan Betley、Daniel Tan、Niels Warncke、Anna Sztyber-Betley、Xuchan Bao、Mart´ın Soto、Nathan Labenz 和 Owain Evans。涌现失调：狭窄的微调可能导致大语言模型广泛失调。*arXiv* 预印本 *arXiv:2502.17424*，2025年。Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901，2020年。Yanda Chen、Joe Benton、Ansh Radhakrishnan、Jonathan Uesato、Carson Denison、John Schulman、Arushi Somani、Peter Hase、Misha Wagner、Fabien Roger 等。推理模型并不总是说出它们所想的内容。*arXiv* 预印本 *arXiv:2505.05410*，2025年。James Chua 和 Owain Evans。DeepSeek R1 及其他推理模型是否更加忠实？*arXiv* 预印本 *arXiv:2501.08156*，2025年。Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano 等。训练验证器来解决数学应用题。*arXiv* 预印本 *arXiv:2110.14168*，2021年。Nelson Elhage、Neel Nanda、Catherine Olsson、Tom Henighan、Nicholas Joseph、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly、Nova DasSarma、Dawn Drain、Deep Ganguli、Zac Hatfield-Dodds、Danny Hernandez、Andy Jones、Jackson Kernion、Liane Lovitt、Kamal Ndousse、Dario Amodei、Tom Brown、Jack Clark、Jared Kaplan、Sam McCandlish 和 Chris Olah。Transformer 电路的数学框架。*Transformer Circuits Thread*，2021年。https://transformer-circuits.pub/2021/framework/index.html。Yichao Fu、Xuewei Wang、Yuandong Tian 和 Jiawei Zhao。充满信心地深度思考。*arXiv* 预印本 *arXiv:2508.15260*，2025年。Shivam Garg、Dimitris Tsipras、Percy S Liang 和 Gregory Valiant。Transformer 能在上下文中学到什么？简单函数类案例研究。神经信息处理系统进展，35:30583–30598，2022年。Melody Y Guan、Miles Wang、Micah Carroll、Zehao Dou、Annie Y Wei、Marcus Williams、Benjamin Arnav、Joost Huizinga、Ian Kivlichan、Mia Glaese 等。监控可监控性。*arXiv* 预印本 *arXiv:2512.18311*，2025年。Aryan Gulati、Brando Miranda、Eric Chen、Emily Xia、Kai Fronsdal、Bruno de Moraes Dumont 和 Sanmi Koyejo。Putnam-axiom：用于衡量高阶数学推理能力的函数型静态基准测试。载于*NeurIPS'24* 数学推理与人工智能第*4*届研讨会，2024年。

Daya Guo，Dejian Yang，Haowei Zhang，Junxiao Song，Ruoyu Zhang，Runxin Xu，Qihao Zhu，Shirong Ma，Peiyi Wang，Xiao Bi等。Deepseek-r1：通过强化学习激励大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*，2025年。Arian Hosseini，Alessandro Sordoni，Daniel Kenji Toyama，Aaron Courville和Rishabh Agarwal。并非所有大语言模型推理者都生而平等。见第一届规模化系统*2*推理研讨会，*NeurIPS'24*，2024年。Andrej Karpathy。NanoGPT。https://github.com/karpathy/nanoGPT，2022年。Tamera Lanham，Anna Chen，Ansh Radhakrishnan，Benoit Steiner，Carson Denison，Danny Hernandez，Dustin Li，Esin Durmus，Evan Hubinger，Jackson Kernion等。衡量思维链推理中的忠实性。$CoRR$，2023年。Hunter Lightman，Vineet Kosaraju，Yuri Burda，Harrison Edwards，Bowen Baker，Teddy Lee，Jan Leike，John Schulman，Ilya Sutskever和Karl Cobbe。让我们逐步验证。见第十二届国际学习表征会议，2023年。Ilya Loshchilov和Frank Hutter。解耦权重衰减正则化。见国际学习表征会议，2019年。网址 https://openreview.net/forum?id=Bkg6RiCqY7。Qing Lyu，Shreya Havaldar，Adam Stein，Li Zhang，Delip Rao，Eric Wong，Marianna Apidianaki和Chris Callison-Burch。忠实思维链推理。见第十三届国际自然语言处理联合会议与第三届计算语言学协会亚太分会会议（*IJCNLP-AACL 2023*），2023年。

Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang 等。自我精炼：通过自我反馈进行迭代精炼。神经信息处理系统进展，36:46534–46594，2023年。

Andrea Montanari 和 Yiqiao Zhong。神经网络中的插值相变：惰性训练下的记忆与泛化。统计年鉴，50(5):2816–2847，2022年。

Catherine Olsson、Nelson Elhage、Neel Nanda、Nicholas Joseph、Nova DasSarma、Tom Henighan、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen 等。上下文学习与归纳头。*arXiv* 预印本 *arXiv:2209.11895*，2022年。

Alethea Power、Yuri Burda、Harri Edwards、Igor Babuschkin 和 Vedant Misra。顿悟：在小型算法数据集上实现超越过拟合的泛化。*arXiv* 预印本 *arXiv:2201.02177*，2022年。

Noah Shinn、Federico Cassano、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。Reflexion：通过言语强化学习实现的语言智能体。神经信息处理系统进展，36：8634–8652，2023年。

Parshin Shojaee、Iman Mirzadeh、Keivan Alizadeh、Maxwell Horton、Samy Bengio与Mehrdad Farajtabar。思维幻觉：从问题复杂度的视角理解推理模型的优势与局限。*arXiv*预印本 *arXiv:2506.06941*，2025年。Jiajun Song、Zhuoyan Xu与Yiqiao Zhong。通过复合实现分布外泛化：以Transformer中的归纳头为视角。美国国家科学院院刊，122(6): e2417182122，2025年。doi: 10.1073/pnas.2417182122。URL https://www.pnas.org/doi/abs/10.1073/pnas. 2417182122。Daniel Soudry、Elad Hoffer、Mor Shpigel Nacson、Suriya Gunasekar与Nathan Srebro。梯度下降在可分数据上的隐式偏差。机器学习研究期刊，19(1):2822–2878，2018年。Jianlin Su、Murtadha Ahmed、Yu Lu、Shengfeng Pan、Wen Bo与Yunfeng Liu。RoFormer：采用旋转位置编码的增强型Transformer。神经计算，568:127063，2024年。

Miles Turpin、Julian Michael、Ethan Perez与Samuel Bowman。语言模型并不总是言其所想：思维链提示中的不忠实解释。神经信息处理系统进展，36:74952–74965，2023年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 与 Illia Polosukhin。注意力就是你所需要的一切。神经信息处理系统进展，30，2017。Johannes Von Oswald、Eyvind Niklasson、Ettore Randazzo、Jo˜ao Sacramento、Alexander Mordvintsev、Andrey Zhmoginov 与 Max Vladymyrov。Transformer 通过梯度下降在上下文中学习。载于 国际机器学习会议，第 35151–35174 页。PMLR，2023年。Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、Denny Zhou、Donald Metzler 等。大语言模型的涌现能力。机器学习研究汇刊，2022年a。ISSN 2835-8856。Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示激发大语言模型中的推理。神经信息处理系统进展，35：24824–24837，2022年b。Yixuan Weng、Minjun Zhu、Fei Xia、Bin Li、Shizhu He、Shengping Liu、Bin Sun、Kang Liu 与 Jun Zhao。借助自我验证，大语言模型能成为更好的推理者。载于 计算语言学协会研究成果：*EMNLP 2023*年，第 2550–2575 页，2023年。

Xuandong Zhao、Zhewei Kang、Aosong Feng、Sergey Levine 和 Dawn Song。《无需外部奖励的推理学习》。*arXiv* 预印本 *arXiv:2505.19590*，2025年。

A 实验细节

模型。我们的模型遵循标准的 Transformer 架构（Vaswani 等人（2017）），并采用标准做法：将层归一化置于自注意力和前馈神经网络（FFN）之前（Karpathy（2022））。我们还使用了 RoPE（Su 等人（2024）），这是许多大语言模型默认的位置嵌入。在训练过程中，我们没有使用任何调度器来调整学习率。

- 层数（L）：3
- 注意力头数量（H）：2
- 模型维度：128
- 前馈网络维度：512
- 丢弃率：0.01
- Theta 参数为 RoPE：10000
- 权重衰减：0.01

训练。每个训练样本都遵循公式 1 中的格式，由 11 个词元和一个序列结束标记组成。我们从第一个 $\rightarrow$ 词元（第 6 个词元）开始，直到最终解答词元，计算自回归损失。训练期间会在线计算两项不忠实性指标以及预测熵等辅助量，以避免产生额外开销。在给定噪声配置下训练单个模型（包括所有指标评估），在一块 NVIDIA A30 GPU 上约需 50 分钟。

B 附加实验结果

B.1 验证简单性偏差：交换提示与推理

为清晰起见，我们在无噪声设置下研究位置顺序对推理忠实性的影响。具体而言，我们比较两种配置：

- 不交换： $e_1 \rightarrow e_2 \rightarrow e_3$
- 交换： $e_2 \rightarrow e_1 \rightarrow e_3$

其中（e_1、e_2、e_3）表示一个一致推理链。除原始（不交换）设置外，我们还使用相同的超参数在交换后的数据上训练 Transformer。训练完成后，我们比较四项忠实性指标，结果汇总于表 2。为完整起见，我们重述正文中定义的不忠实性指标：

$$
\begin{aligned}
\mathrm{RIR}_1 &:= \frac{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2},\, e_{i,3} = e'_{i,3}\}}{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2}\}} \\
\mathrm{RIR}_2 &:= \frac{\sum_{i\le n} \mathbf{1}\{e'_{i,2} \neq e_{i,2},\, e_{i,3} = e'_{i,3}\}}{\sum_{i\le n} \mathbf{1}\{e'_{i,3} = e_{i,3}\}} \\
\mathrm{IDS} &:= \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}\big(\widehat{f}_2(e_{i,1}, e'_{i,2}),\, \widehat{f}_2(e_{i,1}, \tilde{e}_{i,2})\big) \\
\mathrm{INR} &:= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big\{ \arg\max_{e\in\mathcal{V}_N} p(e \mid e_{i,1}, e'_{i,2}) = \arg\max_{e\in\mathcal{V}_N} p(e \mid e_{i,1}, \tilde{e}_{i,2})\big\}
\end{aligned}
$$

对于交换后的思维链格式，我们相应地重新定义 INR 和 IDS。从概念上讲，这两个指标仍然是通过对 e_2施加扰动来定义的；然而，交换后， e_2 现在出现在推理链的第一步

。类似地，在 RIR_1 和 RIR_2, 的定义中， e_2 和 e_3 的功能作用保持不变，只是 e_2 现在位于链的开头。

表 2：在无噪声数据上交换 e_1 和 e_2 前后的不忠实性指标。

Data layout	RIR_1	RIR_2	IDS	INR
$e_1^0 \to e_2^0 \to e_3$	0	0	24.7	0
$e_2^0 \to e_1^0 \to e_3$	1.00	0.99	26.0	0

基于干预的指标（IDS 和 INR）几乎保持不变，这表明观察到的效应并非由位置顺序导致：模型本身更倾向于依赖更简单的 e_2 进行预测。相比之下，交换后，两个 RIR 分数都从原格式下的 1 降至 0，因为在交换后的位置顺序下，依赖 e_2 进行预测就变成了一种“跳步”推理。

B.2 推理阶段

对混合推理模式的进一步解释。我们假设，在训练过程中，模型会进入一个混合推理阶段，在此阶段，其预测行为可以近似表示如下。定义指示函数

$$c(e_1, e_2) = \mathbf{1}\{f(e_1) = f_2(e_2)\},$$

并推测模型实际上按照以下方式进行预测

$$\widehat{f}_2(e_1, e_2) \;\approx\; f_2(e_2)\, c(e_1, e_2) \;+\; U\,(1 - c(e_1, e_2)),$$

其中， U 表示 $\mathcal{V}_N$上的均匀分布。换言之，模型形成了一种自我验证机制：它会比较根据 e_1 和 e_2计算出的结果，并在这些结果不一致时，输出一个近似均匀的预测分布。

为量化这一现象，我们考虑以下指标

$$\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} D_{\mathrm{KL}}\big(\widehat{f}_2(e_{i,1},\, e'_{i,2}),\, U\big),$$

其中， U 是 $\mathcal{V}_N$上的均匀分布，且

$$\mathcal{I} = \{\, i : f(e_{i,1}) \neq f_2(e_{i,2}) \,\}$$

是由不一致推理链组成的集合。该指标的值越小，表示在不一致样例上，模型的预测分布越接近均匀分布，这更有力地表明模型已转变到混合推理（自我验证）阶段。

相图的定性特征。 对于噪声水平 $\varepsilon_1 = 0.01$ 和 $\varepsilon_1 = 0.1$，我们构建相图，以训练计算量为横轴、推理噪声 ε_2 为纵轴。对于每一对（计算量 ε_2），我们根据模型的预测分布与三个参考分布——逐步解、跳步解或均匀分布（仅在不一致推理链上评估）——中哪个分布的 KL散度最小，来确定其所属阶段。

如图 9所示，我们观察到阶段演变呈现出明显的定性规律：随着 ε_2 增加，模型会从呈现两个阶段转变为三个阶段，最终发展为四个不同的阶段。对于较大的 ε_2，进一步增加计算量会推动模型逐渐从逐步推理转向跳步推理。直观而言，较大的 ε_2 （相对于 ε_1）会使不同推理模式所能达到的预测准确率之间产生更大的差距，而模型最大化准确率的驱动力因此会促使其采用跳步推理策略。

我们还注意到，模型的阶段转变和训练动态对训练数据和优化过程都很敏感。图 9 展示了通过在

由同一底层分布生成、但使用不同随机种子的三个不同训练集。尽管共享相同的数据分布，所得轨迹在不同种子之间仍有显著差异，包括每次相变的时机，乃至模型最终达到的阶段。

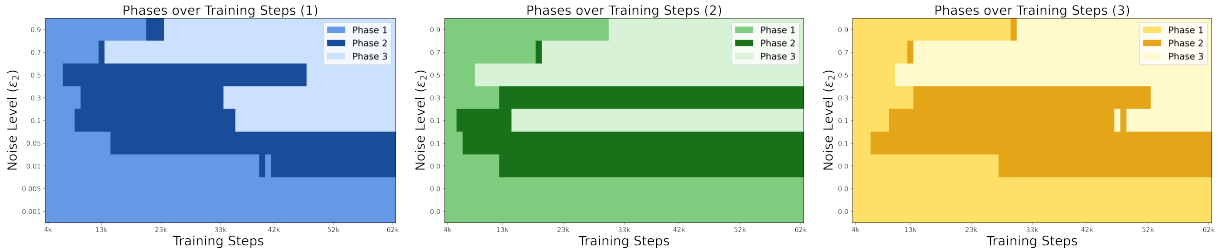


图 9：**数据质量与计算量作为推理模式的决定因素**。我们在固定 $\varepsilon_1 = 0.1$ 且变化 ε_2的条件下训练模型。更多的训练步数（计算量）和更高的噪声（数据质量）往往会诱发跳步推理。三个面板展示了同一数据分布下三个随机种子的相图，说明训练动态对数据和优化中的随机性较为敏感。

三种推理模式的因果解释。图 10 展示了三种定性上截然不同的推理模式。第一行描绘了完全忠实的逐步推理，这是我们在训练中希望从模型引出的理想行为。第二行展示了一种混合推理模式：模型在生成 e_2时依赖 e_1 ，随后在生成 e_3时同时使用 e_1 和 e_2 。这构成了一种不忠实推理过程，其中模型在推理时实际上将 e_2 与 e_3 耦合起来。第三行呈现了另一种形式的不忠实推理，即跳步演绎，其中 e_3 的生成仅依赖于 e_1，而完全独立于 e_2。

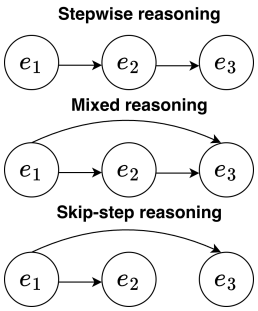


图 10：**表示三种推理模式的因果图**。 类似的因果视角可参见 Bao 等人（2024年）。

B.3 其他结果

注意力对比（AC）中的算子范数。 我们简要回顾注意力对比（AC）的定义中所用的矩阵算子范数概念。设 $\|\cdot\|$ 是 $\mathbb{R}^d$上的一个范数。对于矩阵 $A \in \mathbb{R}^{d\times d}$， A 相对于 $\|\cdot\|$ 的（诱导）算子范数定义为

$$\|A\|_{\mathrm{op}} := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

直观而言， $\|A\|_{\mathrm{op}}$ 衡量在所选范数下， A 对向量进行拉伸的最大倍数。在欧几里得情形 $\|\cdot\| = \|\cdot\|_2$下，这等于 A的最大奇异值，即 $\|A\|_{\mathrm{op}} = \sigma_{\max}(A)$。

在我们的 AC 中，算子范数内的矩阵为

$$h^+_{L-1,i} W^{(L)}_{QK} (h^+_{L-1,i})^\top - h^-_{L-1,i} W^{(L)}_{QK} (h^-_{L-1,i})^\top,$$

它表示样本索引 i 处正（一致）链（e_1^+，e_2^+）与负（不一致）链（e_1^-，e_2^-）对应的最后一层注意力分数矩阵之差。算子范数$\|h^+_{L-1,\,i}W_{(L)QK}(h^+_{L-1,\,i})^\top - h^-_{L-1,\,i}W_{(L)QK}(h^-_{L-1,\,i})^\top\|_{\mathrm{op}}$由此捕捉了从一致链转向不一致推理链时，表征空间中任意方向上的注意力分数最大变化。对 AC 定义中的 i 求平均，可得到一个全局指标，衡量模型的注意力机制在多大程度上区分正样本与负样本，从而为内部不确定性编码或自我验证行为提供证据。

阶段0与格式得分。在第3.2节中，我们假设训练初期存在一个阶段0，在此期间模型主要学习数据格式。图11展示了在噪声水平为 $\varepsilon_1 = 0.01$ 和 $\varepsilon_2 = 0.1$的数据集上进行训练时，格式得分如何变化。我们观察到，格式得分在最初的1k步内从0迅速上升到1，表明在这一早期阶段，模型确实首先专注于匹配数据格式。在图11中，我们还绘制了模型的预测准确率随训练的变化。在阶段1中，由于简单性偏差，模型首先学会从e_2逐步预测 e_3，准确率约为 $1 - p_2 = 90\%$。在阶段3中，模型转向跳步推理，准确率提高到约 $1 - p_1 = 99\%$。相比之下，阶段2以自我验证为主：根据我们的假设，对于满足 $f(e_1) = f_2(e_2)$ 的样本，模型输出近似均匀分布，因此无法预测正确结果。与这一图景一致，图 9中的准确率曲线在阶段2期间并未上升，而是保持平坦，为我们所提出的阶段2机制提供了间接证据。

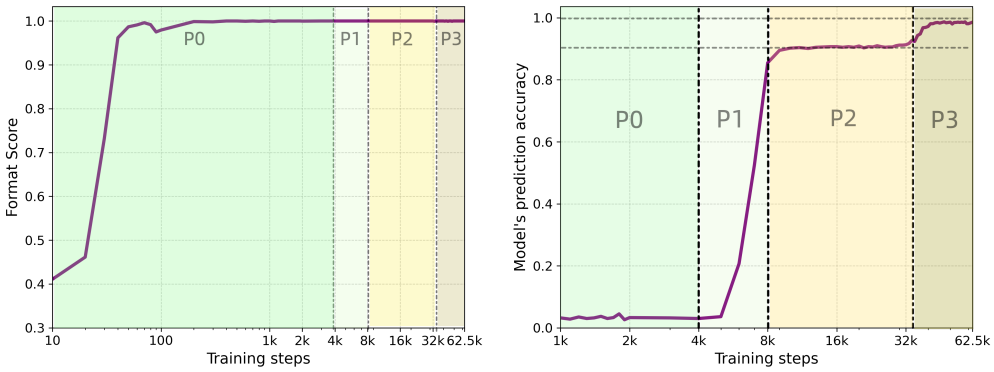


图11：**模型格式得分和预测准确率的训练轨迹。左：训练步数中的格式得分**。在阶段0，模型迅速提升了对结构位置（运算符和推理标记）的预测准确率。**右：模型在训练步数中的预测准确率（在e_3上）**。模型首先学会依赖 e_2，并在 P1 中达到约 90% 的准确率。在 P2 中，模型发展出自我验证机制，但这并未进一步提升预测准确率。最终，模型转变为跳步推理，准确率提升至约 99%。

其他不忠实性指标。我们提出了用于量化模型不忠实性的其他指标。图 12 展示了在噪声水平 $\varepsilon_1 \in \{0.01, 0.1, 0.3, 0.5\}$ 和 $\varepsilon_2 \in \{0, 0.01, 0.1, 0.3, 0.5, 0.7\}$构成的网格上，模型训练后的 一致性比率（CR）、干预分布敏感度（IDS）和推理不一致率（RIR）。由于我们要比较不同 ε_2取值下的 RIR_2，因此报告了一个经 ε_2中的有效噪声水平归一化的 RIR_2版本：归一化 $\mathrm{RIR}_2 = \mathrm{RIR}_2/(n^{-1}\sum_{i\le n}\mathbf{1}\{e'_{i,2} \neq e_{i,2}\})$。图 12 还表明，对于任意固定的 ε_1，随着 ε_2 增大，模型的不忠实性也会增加。此外，随着ε_1 增大，成功推理链的比例会下降，而模型也越来越倾向于逐步推理而非跳步推理，因为在这种情况下，后者已无法显著提升预测准确率。

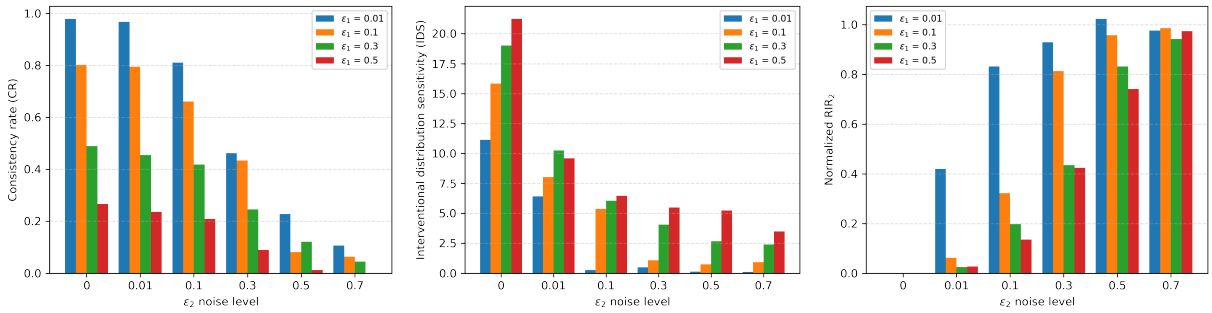


图 12：**在不同噪声机制下训练的模型的其他不忠实性指标结果。左：**随着 ε_2 增加，一致性比率逐渐降低，表明模型成功进行思维链推理的能力越来越弱。**中/右：**随着 ε_2增加，IDS 降低，而 RIR_2 增加。这些趋势共同印证了，噪声水平越高，模型的不忠实程度越高。

对优化轨迹的进一步分析。训练期间，我们还跟踪模型学习分布与相应理论最优分布之间的 KL散度： $p(e_3 \mid e_1, e_2)$，以便更好地理解优化轨迹。图 13 展示了正文中讨论的三种噪声配置下，该 KL散度的变化过程。对于噪声水平 $(\varepsilon_1, \varepsilon_2) = (0.01, 0.1)$ 和 $(0.1, 0.01)$，KL散度最终趋近于零，这表明在这些噪声设置下，模型收敛到了各自的最优目标分布。相比之下，对于 $(\varepsilon_1, \varepsilon_2) = (0.1, 0.1)$，图 12 显示，在 62,500 个训练步数内，模型仍处于阶段2（自我验证）。从优化角度来看，我们推测这一阶段对应于损失景观中的局部极小值或鞍点。当 ε_1 和 ε_2 接近时，这类临界结构会更频繁地出现，使模型更难脱离，因此需要显著增加训练步数。

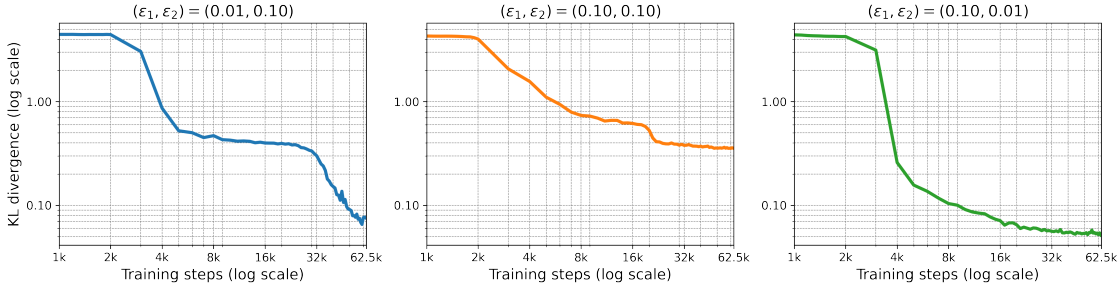


图 13：**模型学习分布与最优分布之间的 KL散度$p(e_3 \mid e_1, e_2)$ 在三种噪声配置下随训练进程的变化。**对于 $(\varepsilon_1, \varepsilon_2) = (0.01, 0.1)$，模型收敛到跳步推理，且与理论最优分布之间的 KL散度趋于 0。同样，对于 $(\varepsilon_1, \varepsilon_2) = (0.1, 0.01)$，模型收敛到逐步推理，KL散度也趋于 0。相反，对于 $(\varepsilon_1, \varepsilon_2) = (0.1, 0.1)$，模型陷入自我验证机制，无法达到最优预测分布。

B.4 捷径特征

图 14 左侧面板展示了基于一致性的 RIR_1 指标在捷径数据集和完整数据集上的训练轨迹。RIR_1 越大，表示对应数据上的不忠实性越明显。我们观察到，在整个训练过程中，捷径数据上的 RIR_1 始终高于完整数据集上计算出的值，这与图 8 一致。值得注意的是，RIR_1 在训练过程中呈现阶梯状变化，而 INR 仅在训练接近结束时上升一次。这种差异源于 RIR_1 的定义明确考虑了模型在 e_2上的学习准确率；随着模型在 e_2 上的准确率在训练过程中稳步提升，RIR_1 也相应地呈阶梯状上升。

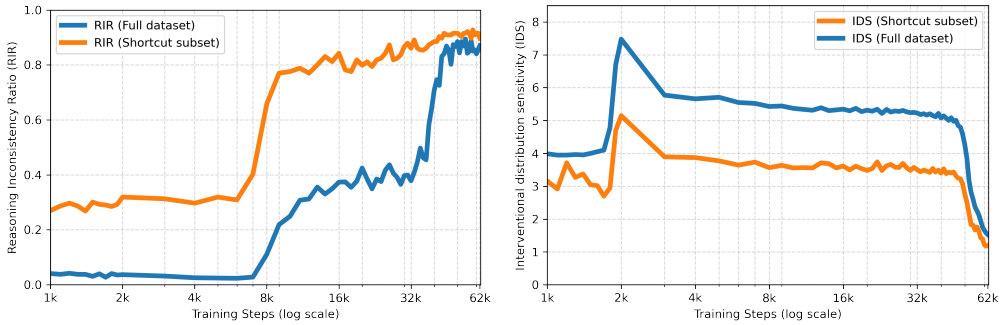


图 14：**捷径数据集和完整数据集上不忠实性指标的其他动态变化**。 左：在整个训练过程中，在捷径数据集上测得的 RIR_1 始终高于在完整数据集上的测量值，这表明模型在捷径数据上更不忠实，并且很可能利用了捷径。右： 在整个训练过程中，捷径数据集上的 IDS 始终低于完整数据集上的值，这表明模型对 e_2 的依赖有所减弱，并且在存在捷径时更倾向于采用跳步推理。

如下所示，我们绘制了不同噪声水平下指标在捷径子集和完整数据集上的训练轨迹。

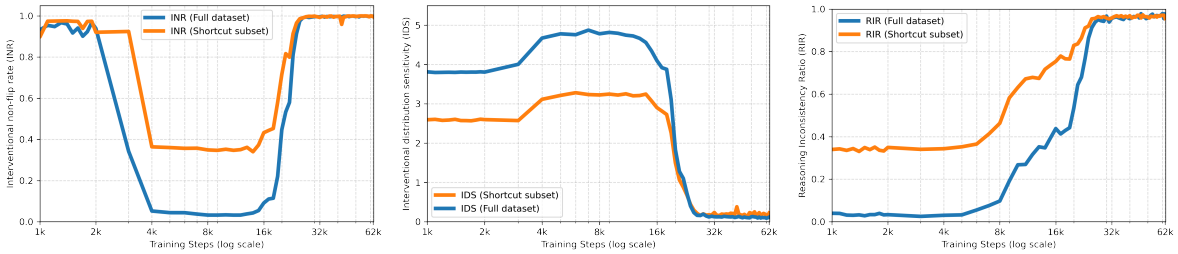


图 15：在设置 $N = 94$ 和噪声配置 $(\varepsilon_1, \varepsilon_2) = (0.01, 0.3)$ 下，INR、IDS 和 RIR_1 的训练动态。

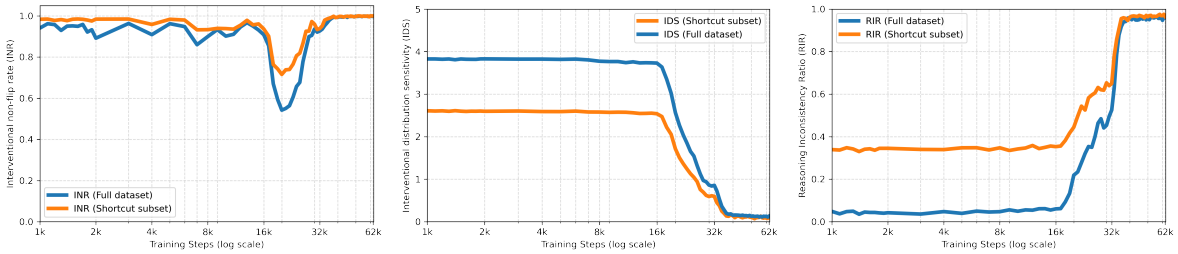


图 16：在设置 $N = 94$ 和噪声配置 $(\varepsilon_1, \varepsilon_2) = (0.01, 0.5)$ 下，INR、IDS 和 RIR_1 的训练动态。

我们将模数从 94 降至 38，并将噪声水平固定为 $(\varepsilon_1,\ \varepsilon_2) = (0.01,\ 0.1)$，重复进行了这些实验。相应的训练指标见图 17。

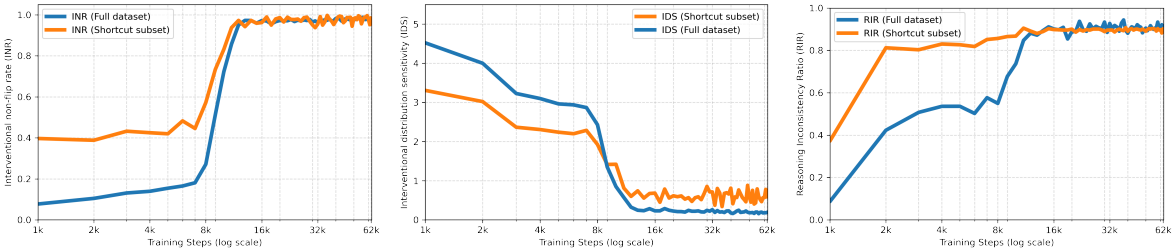


图17：在设定 $N = 38$ 和噪声配置（$\varepsilon_1, \varepsilon_2$） = （0.01,0.1）下，INR、IDS和$\text{RIR}_1$ 的训练动态。

C 实验设置的消融实验与变体

我们展示了在不同实验设置变体下的更多结果。

更大的模型规模。我们首先固定数据设置（回顾式 1），将模型深度从 3 层增加到 5 层，同时保持其他所有超参数不变。图 18 比较了噪声水平 (ε_1，ε_2) =(0。01，0。1) 下 5 层Transformer与 3 层模型的训练动态，展示了 INR 的变化以及模型预测分布的熵。5 层模型最终也学会了跳步推理策略；不过，其熵曲线基本呈单调变化，不再出现与阶段 2 自我验证机制相关的非单调“隆起”。与此一致，5 层模型的 INR 急剧下降，随后在 5k 至 10k 训练步数之间迅速回升至 1，这表明容量更大的模型在阶段 1 和阶段 2 中停留的时间很短，并迅速转入能够达到最优解的跳步阶段。相比之下，3 层模型在逐步推理和自我验证机制中停留的时间要长得多，实际上被困在这些次优局部极小值附近。

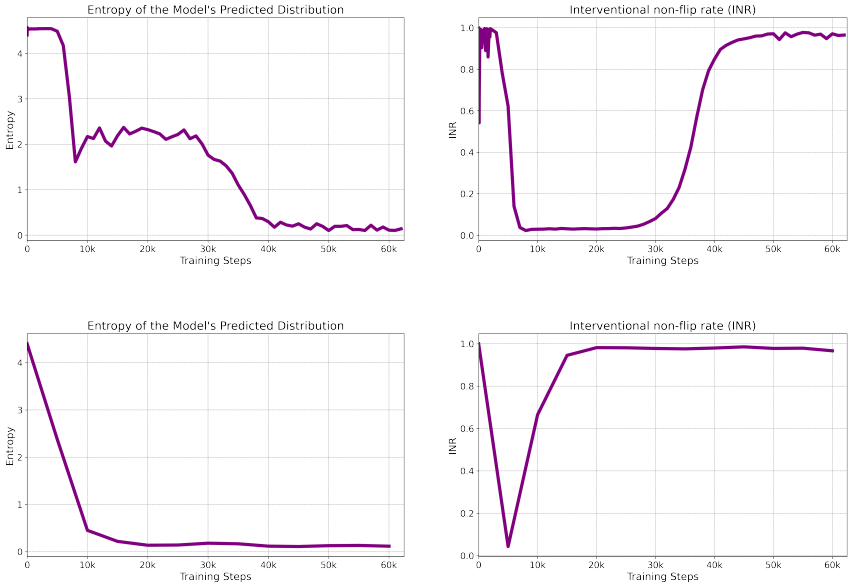


图18：**5层和3层Transformer的INR与熵的训练动态**。**第一行：**3层Transformer在训练过程中的INR以及模型预测分布的熵。**第二行：**5层Transformer因其更大的容量，迅速习得跳步推理，并在熵未出现非单调上升的情况下经历阶段1和阶段2。这体现为训练早期INR中剧烈的瞬态波动。

随着容量增加，转变可能变得非常迅速，以至于中间阶段不那么明显。为了更好地评估我们的定性现象在更大规模下是否依然存在——也就是说，这些现象是否可重标度，而非特定模型或问题规模造成的假象——我们进一步通过增大模数来提高任务复杂度。具体而言，我们将数据集模数增大至 $N = 113$ ，同时保持其他所有超参数不变，并按照相同的实验协议训练同一个 5 层Transformer。如果在这一更大模数机制下，典型的多阶段动态再次出现且在定性上保持一致，就能为我们的观察结果在模型容量和问题规模扩展时仍然稳健提供证据。

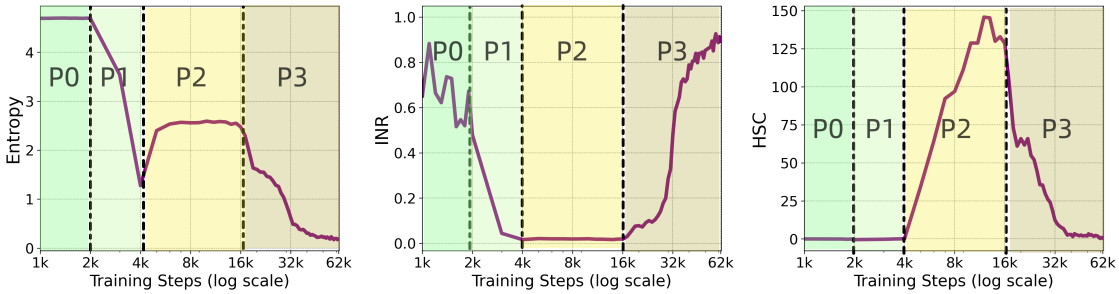


图19：**噪声水平为**$(\varepsilon_1,\ \varepsilon_2)\ =(0.01,\ 0.1)$**且模数为** $N = 113$**时，5层Transformer的训练动态**。左图：在5层Transformer设置中，模型预测分布的熵仍清晰地将训练过程划分为四个不同阶段。**中图**：为5层模型计算的INR指标捕捉到模型在三种推理模式之间的转变，从逐步推理到混合机制，最终到跳步推理。**右图**：在训练过程中，5层模型也形成了支持自我验证的内部指标，这一点可从HSC的相应变化中看出。

图19总结了5层Transformer在 $N = 113$ 设置下、噪声为（$\varepsilon_1,\ \varepsilon_2$）$=(0.01,\ 0.1)$时的训练轨迹。我们监测模型预测分布的熵、基于干预的不忠实性度量，以及用于量化自我验证行为的HSC指标。值得注意的是，即使在模数更大的任务中，且模型深度有所增加，学习过程仍遵循与3层模型、 $N = 97$ 实验基本相同的定性模式。模型从初始的逐步推理阶段，经过中间的混合机制，最终收敛到跳步策略。当模型容量和任务规模同时扩展时，这些机制变化依然存在，这支持了这样一种观点：我们观察到的现象反映的是一种稳健、可重标度的训练动态，而非某种特定配置下的特殊情况。

不同的模数。 我们将 N 从97改为83。图20和图21展示了模数83设置下，基于一致性和基于干预的不忠实性度量的训练动态，以及模型的输出熵、预测准确率和隐藏状态对比（HSC）。这些指标呈现的趋势与模数97设置下观察到的趋势十分相似，训练过程也再次显示出明显的阶段转变，且模型在各阶段的行为与模数97设置下基本相同。

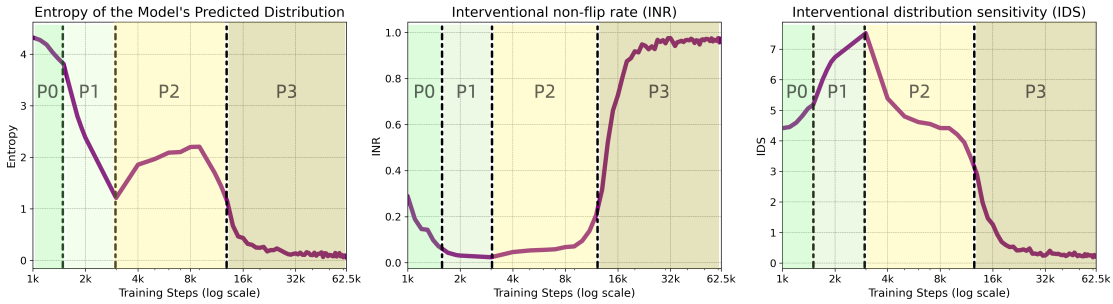


图20：**噪声水平下的训练动态**$(\varepsilon_1,\ \varepsilon_2) = (0.01,\ 0.1)$，**模数为** $N = 83$。**左：**模型输出预测的熵。**中：**在训练步数推进过程中，INR 先降至零，表明对 e_3 的预测完全依赖于 e_2；随后升至接近 1，表明最终仅根据 e_1预测 e_3 。**右：**IDS 先升高后降低，反映了模型对 e_2的依赖程度变化。

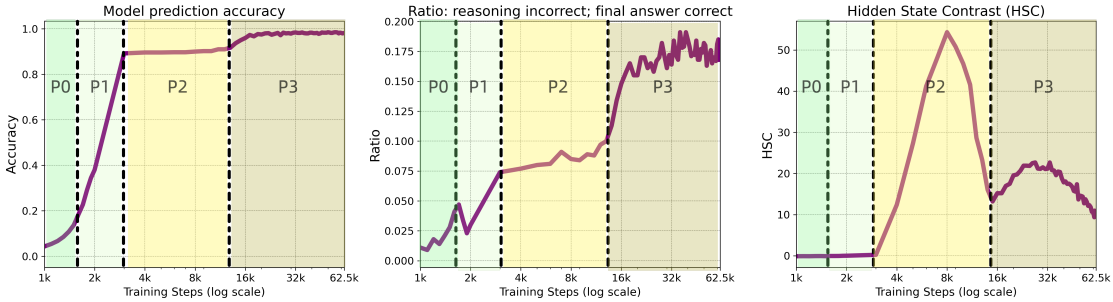


图21：**噪声水平下的训练动态**$(\varepsilon_1,\ \varepsilon_2) = (0.01,\ 0.1)$，**模数为** $N = 83$。**左**：模型的预测准确率在阶段1达到90%，在整个阶段2基本保持不变，随后在阶段3升至约99%，与模型内部推理机制的演变一致。**中**：不忠实比率：$e_2' \neq e_2,\ e_3' = e_3$。（推理错误；最终答案正确。）**右**：隐藏状态对比（HSC）在阶段2期间迅速升高，但在从阶段2转变到阶段 3期间大幅下降，这表明模型正在形成一种自我验证机制。

图22展示了模型在四种不同噪声水平下的不忠实性指标趋势：$\varepsilon_1 = (0。01,\ 0。1,\ 0。3,\ 0。5)$。与模数97设置下一样，对于任意固定的 ε_1，当其低于临界阈值 τ_c（ε_1）时，模型仍保留一定程度的逐步推理；一旦 ε_2 超过该阈值，不忠实性度量便会急剧上升。

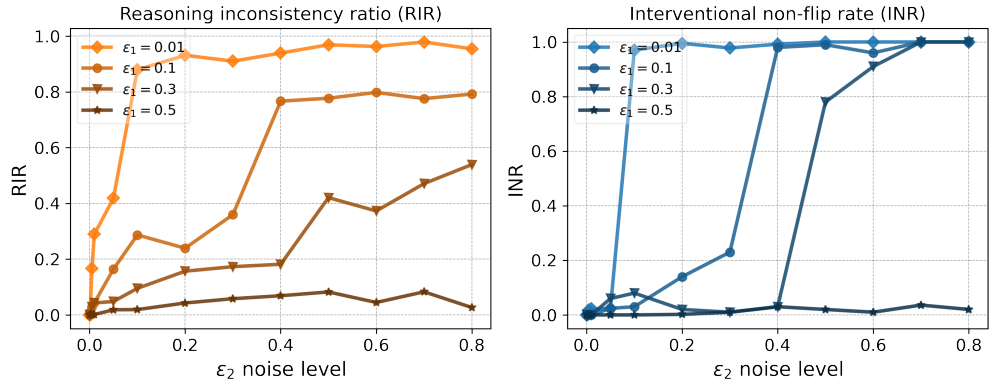


图22：**在模数为 $N = 83$ 的数据集上训练的 Transformer 模型，噪声水平各不相同**。对于模数 83，我们在噪声水平各不相同的数据集上训练 Transformer，以两种定义评估思维链忠实性（ε_1：提示噪声，ε_2：推理噪声）。在低噪声区间，两种不忠实性度量都保持较小；但超过临界 ε_2 值后，它们会急剧上升——橙色曲线表示不一致推理链，而蓝色曲线表明，对推理步骤进行干预几乎不会影响最终答案。

不同的数据格式。我们考虑一种替代的训练样本格式：在提示表达式中添加括号，同时保持模数与正文中的一致（即 $N = 97$）。修改后的数据格式如下：

$$\underbrace{(a \times b) - c}_{E_1\text{:提示}} \to \underbrace{d - c}_{E_2\text{:推理}} \to \underbrace{e}_{E_3\text{:解答}} .$$

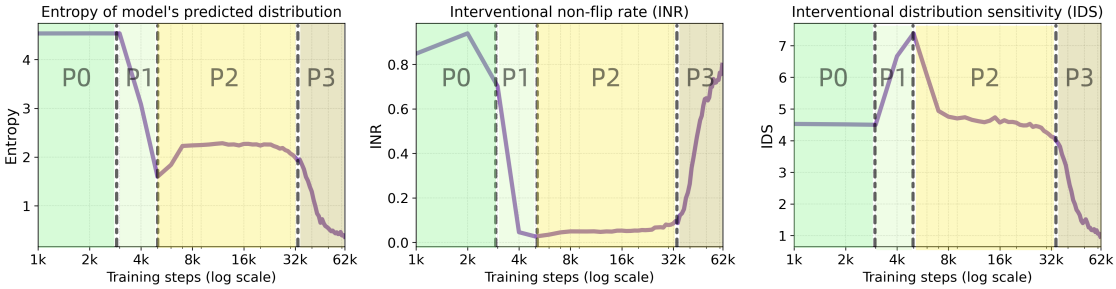


图23：**带括号数据变体上的训练动态**。**左：**模型输出预测的熵。**中：**在训练步数的推进过程中，INR 先降至零，这表明 e_3的预测完全依赖于 e_2，随后又升至接近一，说明最终仅根据 e_1预测 e_3 。**右：**IDS 先升后降，反映了模型对 e_2的依赖程度变化。

图23展示了带括号数据变体上的训练动态。总体而言，模型表现出的行为和相变与无括号设定中的观察结果几乎完全相同。这是意料之中的，因为在我们的构造中，括号只是使表达式的表层形式统一，并不会改变底层计算。因此，聚焦于最简单的格式——即不带括号的版本（见公式1）——进行训练和分析是合理的。