通过采样进行推理：你的基础模型比你想象的更聪明

Aayush Karan[1], Yilun Du[1] [1]哈佛大学

网站 代码

论文标题与作者信息解读

这篇论文的标题《通过采样进行推理：你的基础模型比你想象的更聪明》本身就是一个非常大胆的宣言。它不仅仅是在描述一项技术，更是在挑战当前人工智能领域的一个主流趋势。作者们似乎在说，我们一直以来用来提升模型能力的方法可能走偏了，其实模型本身已经足够“聪明”，只是我们没有用对方法去“发掘”它的智慧。

论文发表在 `arXiv.org` 上，这是一个预印本服务器。在科研领域，这就像是电影上映前的“点映”，研究者可以在这里快速分享他们的最新成果，而无需等待漫长的期刊同行评审过程。这使得科学交流的速度大大加快 。

作者 Aayush Karan 和 Yilun Du 均来自哈佛大学，这是世界顶尖的学术机构之一，尤其在计算机科学领域有着深厚的研究实力。论文提供了项目网站和代码的链接，这体现了现代科学研究中“开放”和“可复现”的重要原则。这意味着任何人都可以访问他们的研究，并亲自验证或在他们的工作基础上进行新的探索，这极大地推动了整个领域的进步。

摘要

前沿的推理模型通过强化学习（RL）对大型语言模型（LLMs）进行后训练，已在众多学科中展现出惊人的能力。然而，尽管这一范式取得了广泛成功，但大部分文献都致力于厘清在强化学习期间涌现的、而基础模型中所不具备的真正新颖行为。在我们的工作中，我们从一个不同的角度探讨这个问题，转而探究是否可以在推理时通过纯粹的采样，从基础模型中引发出相当的推理能力，而无需任何额外的训练。受马尔可夫链蒙特卡洛（MCMC）技术从锐化分布中采样的启发，我们提出了一种利用基础模型自身似然度的简单迭代采样算法。在不同的基础模型上，我们表明我们的算法在推理能力上带来了显著提升，在多种单次任务（包括MATH500、HumanEval和GPQA）中几乎匹敌甚至超越了强化学习带来的提升。此外，我们的采样器避免了在多个样本上出现多样性崩塌的问题，而这正是强化学习后训练的典型特征。至关重要的是，我们的方法不需要训练、精选数据集或验证器，这表明它在那些难以验证的领域之外也具有广泛的适用性 。

摘要解读

这篇摘要信息量巨大，我们可以把它拆解成几个核心要点来理解：

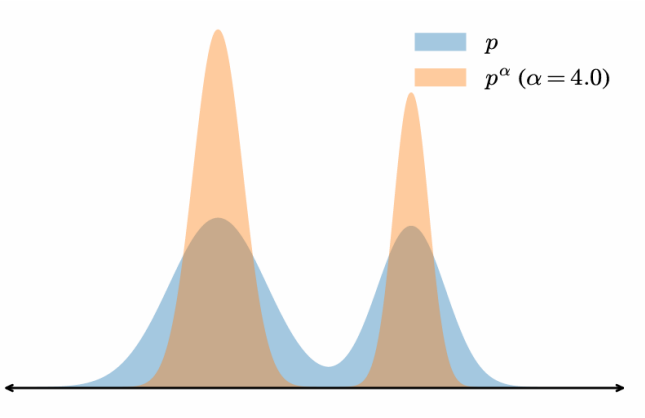

Figure 2: **A toy example of distribution sharpening.** Here p is a mixture of Gaussians, which we plot against p^α ($\alpha = 4.0$).

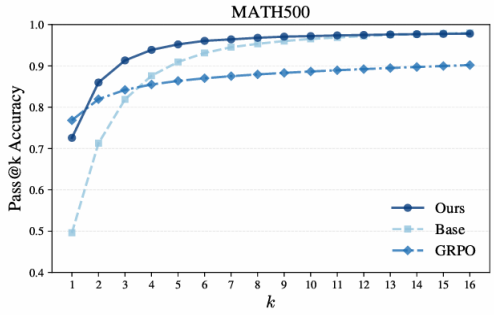

Figure 5: **Pass@k performance on MATH500**. We plot the pass@k accuracy (correct if at least one of k samples is accurate) of power sampling (ours) and RL (GRPO) relative to the base model (Qwen2.5-Math-7B). Our performance curve is strictly better than both GRPO and the base model, and our pass rate at high k matches the base model, demonstrating sustained generation diversity.

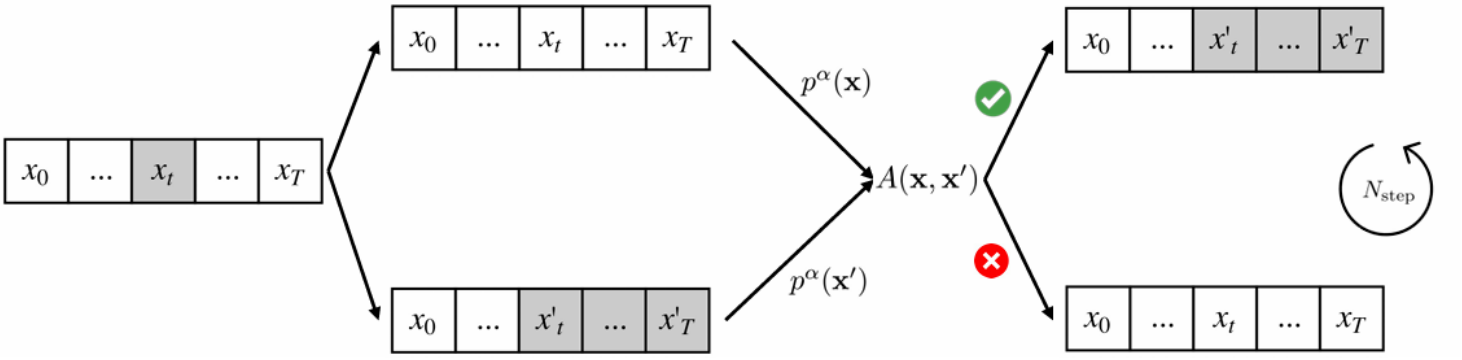

Figure 3: **Illustrating Metropolis-Hastings with random resampling.** A random index t is selected and a new candidate is generated by resampling. Based on the relative likelihoods, the candidate is accepted or rejected, and the process repeats.

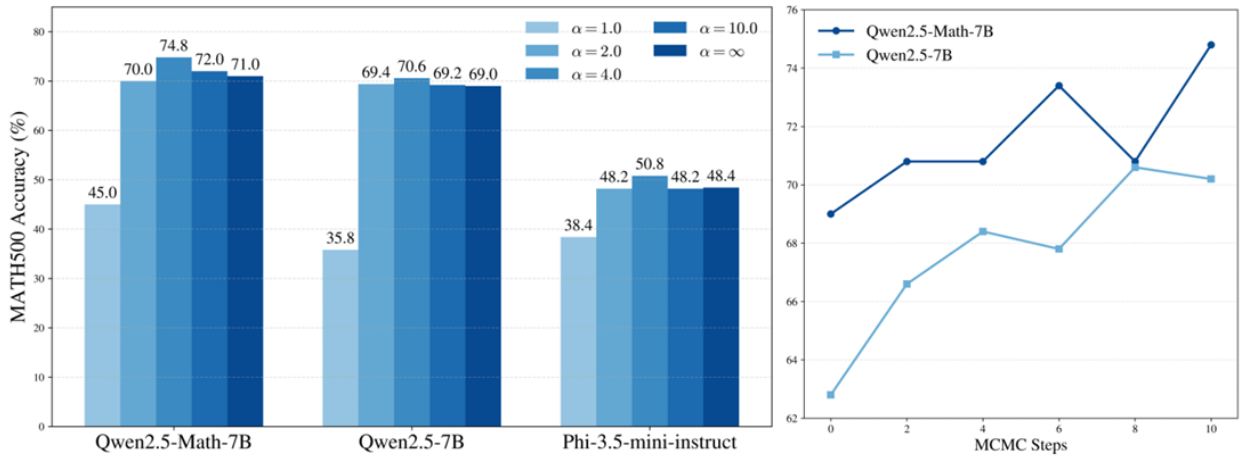

Figure 6: **Effect of hyperparameters on power sampling.** Left: We plot MATH500 accuracy across model families for various values of α. Right: We plot the increase in accuracy of power sampling on Qwen models as the number of MCMC steps increases.

1. **当前的主流方法及其问题**：目前，要让一个大型语言模型（LLM），比如你熟悉的一些对话AI，变得更擅长推理（如做数学题、写代码），最流行的方法是“强化学习（RL）后训练”。你可以把它想象成：先有一个“基础模型”（相当于一个刚学完基础知识的学生），然后通过强化学习的方式，像请一位严格的教练，通过不断的奖励和惩罚（比如题目做对了就奖励，做错了就惩罚）来“特训”它，让它变成一个推理高手。但这种方法有个问题，很多研究发现，经过特训的模型虽然在第一次尝试时表现很好，但似乎丧失了“创造力”或“变通能力”，总是给出相似的答案，这被称为“多样性崩塌” 。

2. **本文提出的颠覆性问题**：作者们提出了一个根本性的问题——我们真的需要那位“强化学习教练”吗？有没有可能，那个基础模型（学生）本身就已经很聪明了，只是我们提问的方式不对？作者认为，或许我们不需要花费巨大的资源去“训练”它，而只需要在“使用”它（即“推理时”）的时候，采用一种更聪明的“采样”方法，就能激发它已有的潜力 。

3. **核心技术方案**：作者的灵感来源于一种经典的统计学方法——“马尔可夫链蒙特卡洛（MCMC）”。这听起来很复杂，但核心思想很简单：让模型在生成答案的过程中，不断地自我反思和调整。它会生成一个初步的答案，然后利用模型自己对这个答案的“信心”（即“似然度”）来判断好坏，并进行迭代修改，最终得到一个高质量的答案。这就像一个学生做题时，写下一个步骤后会停下来检查一下，看看这个步骤是否合理，是否导向正确答案，然后再继续下一步 。

4. **惊人的成果与优势**：实验结果显示，这种“无需训练”的采样方法，在多个高难度的推理任务上（如数学竞赛题MATH500，编程题HumanEval），表现竟然能媲美甚至超过经过复杂强化学习训练的模型。更重要的是，它还解决了强化学习带来的“多样性崩塌”问题，能够生成多种不同的高质量答案 。

5. **深远的影响**：该方法最大的亮点在于它的“三无”特性：**无需训练、无需特制数据集、无需验证器**。强化学习通常需要一个“验证器”（比如一个能自动判断答案对错的程序）来提供奖励信号，这极大地限制了其应用范围，因为很多问题（如写一篇好文章、进行法律分析）没有简单的对错标准。而本文的方法完全摆脱了这一限制，为提升AI在更广泛、更主观领域的推理能力开辟了全新的可能性 。这不仅仅是技术的进步，更是对AI能力提升范式的一次深刻反思：与其投入无尽的资源去“训练”，不如用更巧妙的算法去“解锁”模型内在的智慧。

1 引言

强化学习（RL）已成为增强大型语言模型（LLM）推理能力的主流范式 。借助通常可自动验证的奖励信号，流行的强化学习技术已成功应用于前沿模型的后训练，在数学、编程和科学等领域带来了显著的性能提升 。

尽管强化学习在LLM上取得了广泛的经验性成功，但大量文献都围绕着以下问题展开：在强化学习后训练期间涌现的能力，是否是基础模型中不存在的根本性新行为？这就是分布锐化（distribution sharpening）的问题：即，后训练的分布是否仅仅是基础模型分布的一个“更锐利”的版本，而不是将概率质量置于基础模型不可能生成的推理路径上 。

一些研究指出了通过强化学习后训练学习新能力的困难。He等人 、Song等人 比较了基础模型与后训练模型的pass@k（多样本）得分，发现当k较大时，基础模型实际上表现更优，而后者则遭受生成多样性下降的困扰。在这种情况下，强化学习似乎是以牺牲多样本推理为代价，将pass@k的性能重新分配到了单样本性能上。Yue等人 也指出，强化学习后的推理路径紧密集中在基础模型下的高似然度/置信度区域，似乎是从现有的高似然度能力中汲取而来。我们在图4的实验中也说明了这一点。无论如何，迄今为止，强化学习后训练在单样本推理上的优势仍然是不可否认的 。

在本文中，我们提出了一个令人惊讶的结果：直接从基础模型采样可以实现与强化学习相媲美的单样本推理能力。我们为基础模型提出了一种采样算法，该算法在推理时利用额外的计算资源，实现了在领域内推理任务上几乎与强化学习后训练相匹配的单样本性能，甚至在领域外推理任务上能够超越它。此外，我们观察到，使用我们的采样器，生成多样性并未下降；事实上，我们的pass@k（多样本）性能远超强化学习。我们特别与组相对策略优化（Group Relative Policy Optimization, GRPO）进行了基准比较，这是增强LLM推理的标准强化学习算法 。

至关重要的是，我们的算法是**免训练、免数据集、免验证器**的，避免了强化学习方法的一些固有弱点，包括为避免训练不稳定而进行的大量超参数调整、需要策划一个多样化且庞大的后训练数据集，以及无法保证能获得一个真实的验证器/奖励信号 。

我们的贡献可以总结如下： i) 我们引入了幂分布（power distribution）作为一个有用的推理任务采样目标。由于它可以用基础LLM明确指定，因此不需要额外的训练。 ii) 我们进一步引入了一种针对幂分布的近似采样算法，该算法使用马尔可夫链蒙特卡洛（MCMC）算法，根据基础模型的似然度迭代地重采样词元子序列。 iii) 我们通过在一系列模型（Qwen2.5-Math-7B, Qwen2.5-7B, Phi-3.5-mini-instruct）和推理任务（MATH500, HumanEval, GPQA, AlpacaEval 2.0）上的实验，凭经验证明了我们算法的有效性。我们的结果表明，直接从基础模型采样可以取得与GRPO相媲美的结果。事实上，对于一些领域外的任务，我们的算法始终优于强化学习基准。此外，在多个样本上，我们避免了困扰强化学习后训练的多样性崩塌问题，在单样本到少样本的推理能力以及样本多样性方面实现了两全其美 。

我们的结果共同表明，现有的基础模型在单样本推理方面的能力远比当前采样方法所揭示的要强大得多 。

引言解读

引言部分为我们搭建了舞台，清晰地阐述了研究的背景、动机和核心贡献。

首先，作者再次确认了强化学习（RL）在提升LLM推理能力方面的“霸主”地位，尤其是在那些有明确对错标准（即“可自动验证”）的领域，如数学和编程。

接着，引言抛出了一个核心的学术辩论：**“分布锐化”**。这是一个非常关键的概念。我们可以用一个比喻来理解：假设一个基础模型（学生）对某个数学问题有A、B、C三种解法思路，其中A是正确答案，但模型对A的信心（概率）只有40%，对B和C的信心各有30%。经过强化学习后，模型对A的信心飙升到95%，而对B和C的信心则降至几乎为零。那么，强化学习是教会了学生一种全新的解题方法D，还是仅仅让他对原本就会的A方法变得极度自信？“分布锐化”理论倾向于后者，认为强化学习更像是一个“信心放大器”，它把模型原有的概率分布变得更“尖锐”，让高概率的答案变得更高，低概率的变得更低，但并未创造全新的知识 。

为了支撑这一观点，作者提到了 `pass@k` 指标。`pass@k` 的意思是“尝试k次内至少有一次成功”。研究发现，强化学习训练后的模型虽然在 `k=1`（第一次尝试）时表现很好，但随着 `k` 的增大（允许多次尝试），其表现的提升却很慢，甚至不如基础模型。这正是因为它丧失了生成多样性答案的能力——翻来覆去只会用那一种它最有信心的方法，一旦这种方法错了，就再也想不出别的办法了。这是一种以“多样性”换取“单次准确率”的权衡 。

在这样的背景下，作者亮出了本文的“王牌”——一个令人惊讶的发现：**我们根本不需要强化学习，仅仅通过一种更聪明的采样方法，就能在单次尝试中达到甚至超越强化学习的效果，同时还保留了多样性。** 这直接挑战了之前段落提到的强化学习“不可否认的优势”。

最后，引言清晰地列出了三点核心贡献，就像一份宣言，告诉读者接下来会读到什么：一种新的采样目标（幂分布）、一种实现该目标的算法（MCMC），以及证明该算法有效的充分实验。整篇引言的逻辑层层递进，从描述现状，到提出质疑，再到给出颠覆性的解决方案和成果，极具说服力。

图1：我们的采样算法能够匹敌并超越强化学习后训练

图表描述： 这张柱状图展示了三种方法在四个不同任务上的性能对比。这三种方法分别是：

- **Base (Qwen2.5-Math-7B)**：即原始的基础模型，未经过任何额外训练。
- **GRPO (RL)**：使用GRPO强化学习算法对基础模型进行后训练得到的模型。
- **Ours (Training-free)**：使用本文提出的免训练采样算法的基础模型。

左侧图表比较了三种可验证的推理任务：MATH500（数学）、HumanEval（编程）和GPQA（科学问答）。右侧图表比较了一种不可验证的通用任务：AlpacaEval2.0（通用对话能力）。

图表解读： 这张图可以说是整篇论文核心论点的“一图流”总结，它提供了强有力的视觉证据：

1. **在“主场”打成平手**：MATH500是GRPO模型训练时使用的数据集领域，可以看作是它的“主场”。在这个任务上，“Ours”方法的表现（74.8%）非常接近GRPO（78.5%），这说明即使在强化学习最擅长的领域，这种免训练方法也具有极强的竞争力 。
2. **在“客场”实现超越**：HumanEval（编程）和AlpacaEval（通用对话）对于在数学数据上训练的GRPO模型来说是“领域外”任务。在这些任务上，“Ours”方法的表现明显优于GRPO。特别是在HumanEval上，“Ours” (57.3%) 显著高于GRPO (53.7%)。这表明本文提出的采样方法具有更好的**泛化能力**，不会像强化学习那样因为过度拟合训练数据而导致在其他任务上表现下降 。
3. **巨大的提升**：相较于基础模型（Base），“Ours”方法在所有推理任务上都取得了巨大的性能提升。这证明了论文的标题——基础模型确实比我们想象的更聪明，其潜力是可以通过先进的采样技术来解锁的。

这张图直观地告诉我们，本文的方法不仅有效，而且可能比当前主流的强化学习范式更具鲁棒性和通用性，为提升AI推理能力提供了一条成本更低、适用范围更广的新路径。

2 相关工作

用于LLM的强化学习。强化学习在LLM的后训练中起到了关键作用。早期，带有人类反馈的强化学习（RLHF）被开发出来，作为一种使用训练好的奖励模型来使LLM与人类偏好对齐的技术。最近，带有可验证奖励的强化学习（RLVR）已成为一种强大的新后训练技术，许多工作发现，由自动验证器给出的简单的、在生成结束时的奖励，可以显著增强在数学和编程等困难推理任务上的性能。组相对策略优化（GRPO）算法是这些进展的核心。在这一成功的基础上，许多后续工作研究了使用源自内部信号的奖励信号，如自熵、置信度，甚至是随机奖励。与这些工作类似，本文也研究了将基础模型似然度作为提高推理性能的机制，但至关重要的一点是，我们的技术是免训练的 。

LLM的自回归MCMC采样。先前的工作已经探索了将经典的MCMC技术与自回归采样相结合。许多场景，包括红队测试、提示工程和个性化生成，都可以被构建为从基础LLM分布中进行采样，但向某个外部奖励函数倾斜。Zhao等人 提出了学习中间价值函数，并将其用于顺序蒙特卡洛（SMC）框架中，其中维护多个候选序列，并根据其预期的未来奖励进行更新。类似地，Faria等人 提出了Metropolis-Hastings（MH）算法，它不是维护多个候选序列，而是执行迭代重采样，同样根据预期奖励进行更新。在方法论上，我们的采样算法与后一项工作最为相似，但关键区别在于，我们的目标采样分布完全由基础LLM指定，避免了对外部奖励的需求 。

扩散模型的退火采样。在统计物理学和蒙特卡洛文献中，从 pα 中采样被称为从退火（annealed）或回火（tempered）分布中采样，并激发了扩散模型社区新一轮的兴趣。确实，在传统的MCMC采样中，退火被用作一种避免采样过程中模式崩塌、更准确地从复杂多模态分布中采样的方法。这已作为扩散模型的推理时采样方法重新出现，旨在将预训练模型引导向“倾斜分布”。传统强化学习技术表现出模式崩塌，而物理科学中的应用则需要多模态采样。为此，Du等人 、Wang等人 、Kim等人 等工作构建了一系列退火分布，以简化从基础扩散分布到倾斜分布的过渡。其他工作则有意地将从 pα（其中 α>1）采样作为目标，以此作为从基础扩散模型生成更高质量样本的手段，这在生成更具可设计性的蛋白质方面尤其流行 。

相关工作解读

这部分内容相当于一篇学术论文的“文献综述”，作者在此将自己的工作与前人的研究联系起来，并明确指出自己的创新之处。这展示了科学研究是如何站在巨人肩膀上前行的。

1. **与强化学习（RL）的联系与区别**：作者首先回顾了强化学习在LLM领域的发展历程，从依赖人类主观判断的RLHF，演变到依赖程序自动判断的RLVR。这反映了一个趋势：研究者们在寻求更客观、更可扩展的训练方法。作者承认，他们的工作与近期一些利用模型“内部信号”（如置信度）进行RL的研究有相似之处，都是在挖掘模型自身的信息。但他们强调了一个**本质区别**：所有这些RL方法都需要“训练”——即调整模型的参数，而本文的方法是**“免训练”**的，仅在生成答案时发挥作用 。

2. **与MCMC采样方法的继承与发展**：作者指出，将MCMC（一种经典的统计采样方法）用于语言模型的想法并非首创。但之前的研究大多是将MCMC作为一种工具，来引导模型生成更符合某个“外部奖励”的答案。例如，目标是生成更积极的评论，那么“积极性”就是一个外部奖励。而本文的**核心创新**在于，他们采样的目标不是由外部定义的，而是完全由模型**内部的概率分布** p 决定的（具体来说是 pα）。这意味着整个过程是自洽的，不需要任何外部的“指挥棒” 。

3. **从其他AI领域的交叉启发**：这一段展示了科学思想的“跨界”魅力。作者提到了“退火采样”和“扩散模型”。“扩散模型”是近年来在图像生成领域（例如DALL-E、Midjourney）大放异彩的技术。而“退火”是一个源自物理学（金属冶炼）的概念，比喻通过一个缓慢、渐进的过程来达到一个最优状态。在采样中，“退火”可以帮助算法避免陷入局部最优解（就像一个登山者不会只满足于一个小山丘，而是会继续寻找更高的主峰）。作者从这些看似不相关的领域中汲取灵感，设计了他们分阶段、逐步求精的采样算法（即后文的算法1）。这体现了作者广阔的学术视野，他们将不同领域的思想融合在一起，为LLM推理问题提供了一个全新的解决方案 。

总的来说，这一章节清晰地勾勒出本文工作的学术坐标：它根植于强化学习的研究背景，继承了MCMC的采样技术，并受到了扩散模型中退火思想的启发，最终通过“免训练”和“内部信号驱动”这两个关键创新，开辟了一条独特的技术路径。

3 预备知识

令 X 为一个有限的词元（token）词汇表，令 XT 表示由词元组成的有限序列 $x_{0:T}=(x_0,x_1,...,x_T)$ 的集合，其中对所有 i 都有 $x_i \in X$，且 $T \in Z_{\geq 0}$ 是某个非负整数。为方便起见，对于给定的 t，令 $x_{<t}=(x_0,...,x_{t-1})$，$x_{>t}=(x_{t+1},...,x_T)$，对于 $x_{\leq t}$ 和 $x_{>t}$ 也有类似的定义。通常， x 指的是一个词元序列 $x_{0:T}$，其中 T 是隐式给定的。

那么，一个LLM通过自回归地学习所有 t 的条件词元分布 $p(x_t|x_{<t})$，定义了词元序列 XT 上的一个分布 p，通过以下恒等式给出联合分布： $p(x_{0;T})=\prod_{t=0}^{T} p(x_t|x_{<t})$ (1)

要从 p 中采样一个序列，我们只需使用条件分布逐个词元地从LLM中采样，根据式(1)，这直接从联合分布中进行了采样 。

预备知识解读

这一章非常简短，但它为我们理解整篇论文的技术核心打下了最重要的数学基础。它就像是学习微积分前，必须先掌握函数的概念一样。

首先，我们来理解几个基本术语：

- **词元（token）**：你可以把它理解成一个单词或者一个汉字，是构成句子的基本单位。
- **词汇表（X）**：就是模型认识的所有词元的集合。
- **序列（$x_{0:T}$）**：就是由一连串词元组成的句子或段落。

本章最核心的内容是**公式(1)**： p(x0;T)=∏t=0Tp(xt|x<t)。 这个公式揭示了所有自回归语言模型（比如GPT系列）生成文本的根本原理，它基于概率论中的“链式法则”。让我们用一个简单的例子来拆解它：

假设我们要计算模型生成句子“猫坐在垫子上”的概率。

- p(x0:T) 就是 p(“猫坐在垫子上”)，代表整个句子的概率。
- ∏ 是连乘符号。
- p(xt|x<t) 是模型最核心的计算部分，它表示“在已经生成了前面所有词元（x<t）的条件下，下一个词元是 xt 的概率”。

所以，整个句子的概率可以这样计算： p(“猫坐在垫子上”)=p(“猫”)×p(“坐”|“猫”)×p(“在”|“猫坐”)×p(“垫”|“猫坐在”)×p(“子”|“猫坐在垫”)×p(“上”|“猫坐在垫子”)

语言模型的工作方式就是这样“一个接一个”地预测下一个最可能的词。这个看似简单的公式是后续所有复杂算法的基础。正是因为我们可以计算出任何一个文本序列的概率 p(x)，我们才能够比较不同答案的好坏，并在此基础上进行优化。本文提出的“幂分布采样”正是利用了这个可以被计算的 p(x) 值，来找到那些概率极高的、通常也代表着高质量推理的答案序列。

4 面向幂分布的MCMC采样

在本节中，我们介绍我们为基础模型设计的采样算法。我们的核心直觉源于第1节中提出的分布锐化概念。锐化一个参考分布指的是对该分布进行重新加权，使得高似然区域被进一步增权，而低似然区域被降权，从而使样本严重偏向于参考分布下的更高似然区域。那么，如果强化学习后训练的模型真的只是基础模型的锐化版本，我们应该能够明确指定一个能达到同样效果的目标采样分布。

图2：分布锐化的一个简单示例。 这里 p 是一个高斯混合分布，我们将其与 pα（α=4.0）进行对比。

我们将本节组织如下。第4.1节介绍了这个目标锐化分布，并为其样本适用于推理任务提供了一些数学动机。第4.2节介绍了一类旨在从该目标分布中实际采样的通用马尔可夫链蒙特卡洛（MCMC）算法。最后，第4.3节详细介绍了我们针对LLM的具体实现。

4.1 使用幂分布进行推理

锐化一个分布 p 的一种自然方法是从幂分布 pα 中采样。由于 p(x)>p(x′)⇒p(x′)αp(x)α>p(x′)p(x)（$\alpha\in。

4.1节 解读：远见卓识的“幂分布” vs. 目光短浅的“低温采样”

这是本文最核心、最关键的技术部分。作者在这里提出了他们的核心理念：**使用“幂分布” (pα) 作为采样目标，可以更好地进行推理。**

首先，让我们通过**图2**来直观理解“分布锐化”。想象原始分布 p 是一片连绵的丘陵，有高有低。经过 pα （α=4）处理后，这片丘陵变成了陡峭的山峰。原来最高的山丘变得尖锐无比，而其他地方则变得更加平坦。采样的过程就像是在这片地形上随机撒豆子，锐化后，绝大多数豆子都会落在最高的山峰上。这就是“幂分布”的作用：**极大地放大高概率事件的优势，让你更容易找到最优解** 。

接下来，作者做了一件非常重要的事情：将他们提出的“幂分布采样”与一个现有常用技术“低温采样”进行了对比。这两种方法看起来很像，都是想让模型更“确定”、更“自信”，但作者通过严谨的数学证明（**命题1**）和生动的例子（**示例1**）告诉我们，它们之间存在着天壤之别。

这个区别的关键在于**“远见”**。**我们可以用一个**迷宫寻宝**的比喻来理解： 假设你在迷宫的起点，面前有两条路：

- **A路**：这条路只有一个出口，但通向一个价值100分的巨大宝藏。
- **B路**：这条路有两条支路，分别通向两个价值30分的小宝箱。

低温采样就像一个**目光短浅的寻宝者**。在岔路口，他只看下一步。他看到B路后面有两个宝箱，总价值是 30+30=60 分，而A路只有一个宝藏。他会简单地认为B路的总期望更高，于是选择了B路（这对应公式8的“和的指数”：(∑p)α，先求和再放大）。结果他只得到了30分。

幂分布采样则像一个**有远见的规划师**。他会评估每一条**完整的路径**。他看到A路的最终价值是100分，B路的两条最终价值都只有30分。他会先对每条完整路径的最终价值进行放大（取幂），比如平方。那么A路的价值就变成了 1002=10000，B路的总价值变成了 302+302=1800（这对应公式7的“指数的和”：∑pα，先放大再求和）。在这种评估体系下，A路的优势被极大地凸显出来。他会毫不犹豫地选择A路，最终获得巨大宝藏。

为什么这对于推理至关重要？ 因为复杂的推理任务（如数学证明、编程）就像走迷宫。其中往往存在一些**“关键步骤”**（pivotal tokens）。一步走错，全盘皆输。一个看似不错的、有很多后续可能性的步骤（像B路），可能最终都导向平庸或错误的答案。而那条通往正确答案的路径可能很唯一、很狭窄（像A路），但它是唯一能成功的路。

低温采样这种“贪心”的策略很容易在关键步骤上犯错，而幂分布采样因为其“远见”，能够更好地识别并选择那条通往最终高价值（高似然度）答案的正确路径。这正是本文方法能够提升推理能力的根本原因。

4.2 Metropolis-Hastings算法

现在我们已经看到，从 pα 采样理论上如何能帮助底层LLM的推理能力，我们的目标现在转向提出一个算法来准确地从中采样。给定一个LLM p，我们可以获得任何序列长度上的 pα 值；然而，这些值是未归一化的。直接从真实概率进行采样需要对所有序列 (x0,...,xT)∈XT 进行归一化，这在计算上是不可行的。

为了解决这个问题，我们引入了一种称为Metropolis-Hastings（MH）的马尔可夫链蒙特卡洛（MCMC）算法，它恰好能实现我们想要的目标：从一个未归一化的概率分布中进行近似采样。MH算法使用一个任意的提议分布 q(x|xi) 来选择下一个候选 xi+1，从而构建一个样本序列的马尔可夫链 (x0,x1,...,xn)。以概率 A(x,xi)=min{1,pα(xi)·q(x|xi)pα(x)·q(xi|x)} (9) 接受候选 x 作为

xi+1；否则，MH设置 xi+1=xi。这个算法特别方便，因为它只需要由 pα 给出的相对权重（因为A中的归一化权重会抵消），并且可以与任何通用但易于处理的采样器 q 一起工作，限制极少。值得注意的是，对于足够大的 n，在提议分布满足以下（相当宽松的）条件下，这个过程会收敛到从目标分布 pα 中采样 。

定义1。如果对于任何在目标分布 pα 下具有非零质量的集合X，提议分布 q 都有非零概率最终从X中采样，则称 q 是**不可约的**。如果所引导的样本链不会在固定的间隔步数后返回到同一个样本，则称提议是**非周期的**。

因此，我们只需确保我们的提议分布满足不可约性和非周期性，Metropolis-Hastings就会处理剩下的事情。在实践层面上，我们还希望 q(x∣xi) 和其反向 q(xi∣x) 都易于计算。

图3：用随机重采样说明Metropolis-Hastings。 选择一个随机索引t，并通过重采样生成一个新的候选序列。根据相对似然度，接受或拒绝该候选，然后重复该过程。

考虑以下一类随机重采样提议分布（见图3）。令 pprop 为一个提议LLM。以均匀概率 T1，选择一个随机的 t∈，并使用 pprop 从索引 t 开始重采样序列。那么转移似然度 q(x′∣x) 就是重采样的似然度。注意，在每个候选选择步骤中，我们都有非零的概率在任意两个序列 x,x′∈XT 之间转换，因为总有一定概率我们可以从 x 的开头就进行重采样。这确保了我们的提议分布既是不可约的也是非周期的。此外，通过对称性，q(x∣x′) 也很容易计算，因为我们可以将 x 视为 x′ 的一个重采样版本。借助Metropolis-Hastings赋予的灵活性，我们可以选择提议LLM pprop 为任何 LLM，并采用任何采样策略（例如，低温采样） 。

4.2节 解读：借助百年算法，实现不可能的采样

上一节我们确定了一个绝佳的目标：从 pα 分布中采样。但这立刻带来了一个巨大的技术难题：理论上，要从一个概率分布中采样，我们需要知道所有可能事件的概率之和，以便进行归一化。对于语言模型来说，所有可能的句子组合是无穷无尽的，计算这个总和是“计算上不可行”的，也就是不可能完成的任务。

这时，作者请出了一位“老将”——诞生于20世纪50年代的**Metropolis-Hastings（MH）算法**。这个算法的绝妙之处在于，**它允许我们从一个我们只知道其“相对”概率、而不知道其“绝对”概率的分布中进行采样**。

为了理解这个算法，我们可以使用一个生动的比喻：**一个在山脉中寻找最高峰的盲人登山者**。

- **目标**：登山者希望最终能花大部分时间停留在山脉的最高区域（对应高概率区域）。
- **地形**：山脉的高度就代表了我们的目标分布 pα。
- **登山者的工具**：他看不见整片山脉，但他有一个高度计，可以测量当前位置和任何想去的新位置的高度。

MH算法就是这位登山者的行动策略，如**图3**所示：

1. **提议一个新位置（Proposal）**：登山者随机选择一个方向，迈出一步，到达一个候选位置 x′。在LLM中，这就相当于随机选择一个句子的中间位置 t，然后让模型重新生成后半部分，得到一个新的候选句子。

2. **做出决定（Accept/Reject）**：登山者使用他的高度计比较新旧两个位置的高度，并根据**MH法则（公式9）**做决定：
    - **如果新位置更高（$p_\alpha(x')>p_\alpha(x)$）**：这无疑是好事，登山者**总是会**移动到这个新位置。
    - **如果新位置更低（$p_\alpha(x')<p_\alpha(x)$）**：这就有趣了。登山者**不会**立刻拒绝，而是会以一个**特定的概率**（等于新旧高度的比值 $p_\alpha(x)p_\alpha(x')$）决定是否移动。比如，新位置比旧位置低一半，他就有50%的概率移过去。

为什么“下山”的步骤至关重要？ 这正是MH算法的精髓所在。如果登山者只往高处走，他很可能会被困在一个小山丘的顶上（局部最优解），而错过了不远处更高大的主峰。允许他有一定概率“下山”，给了他逃离局部陷阱、去探索整个山脉并最终找到全局最高峰的机会。

通过成千上万次的“提议-决定”迭代，这位盲人登山者的足迹最终会呈现出一个神奇的分布：他在高海拔区域停留的时间远远多于在低海拔区域的时间，完美地模拟了从目标地形（概率分布）中采样的效果。

这个算法的强大之处在于它的普适性和简洁性。作者巧妙地将LLM自身既作为“提议者”（生成新句子），又作为“裁判”（计算句子似然度），通过MH这个古老而智慧的规则，构建了一个优雅的自我优化循环。

4.3 使用自回归MCMC进行幂采样

直接为LLM实现Metropolis-Hastings算法将涉及用一个采样得到的长度为 T 的词元序列进行初始化，随后通过式(9)在许多次迭代中生成新的长度为 T 的候选序列。然而，由于需要对LLM进行重复的、完整的序列推理调用，这个过程在计算上是昂贵的。

事实上，MCMC算法在实践中的主要缺点是可能存在指数级的混合时间，其中糟糕的初始化或提议分布选择可能导致在收敛到目标分布之前需要指数级数量的样本。如果样本空间维度很高，这个问题会更加严重，而词元序列空间 X^T 正好表现出这一特性，特别是对于长序列/大的 T 值。

为了解决这个问题，我们提出了一种利用自回归采样序列结构的算法。我们定义了一系列中间分布，并逐步从中采样，直到收敛到目标分布 p_α。特别地，来自一个中间分布的样本会为下一个分布启动一个Metropolis-Hastings过程，这有助于避免病态的初始化。

算法1：自回归模型的幂采样 输入：基础模型 p；提议模型 p_{prop}；幂 α；长度 T **超参数**：块大小 B；MCMC步数 N_{MCMC}**输出**：$(x_0,...,x_T)\sim p_\alpha$

1. **符号说明**：定义未归一化的中间目标 $\pi_{k}(x_{0:kB})\propto p(x_{0:kB})^\alpha$。
2. **for** $k\leftarrow 0$ to $\lceil BT\rceil-1$ **do**
3. 给定前缀 $x_{0:kB}$，我们希望从 π_{k+1} 中采样。通过使用 p_{prop} 自回归地扩展来构建初始化 x_0：$x_t(0)\sim p_{prop}(x_t|x_{<t})$，对于 $kB+1\leq t\leq (k+1)B$。
4. 设置当前状态 $x\leftarrow x_0$。

5. **for** n←1 to NMCMC **do**
6. 均匀采样一个索引 m∈{1,...,(k+1)B}。
7. 构建提议序列 x′，其前缀为 x0:m−1，并重采样补全部分：xt′∼pprop(xt∣x<t′)，对于 m≤t≤(k+1)B。
8. 计算接受率(9) $A(x^{\prime},x)\leftarrow min\{1,\frac{\pi_{k+1}(x^{\prime})}{\pi_{k+1}(x)}\cdot\frac{p_{prop}(x_{0:(k+1)B}|x')}{p_{prop}(x'_{0:(k+1)B}|x)}}$。
9. 抽取 u∼Uniform(0,1)；如果 u≤A(x′,x) 则接受并设置 x←x′。
10. **end for**
11. 设置 x0:(k+1)B←x 以固定新前缀序列，用于下一阶段。
12. **end for**
13. **return** x0:T

固定块大小 B 和提议LLM pprop，并考虑一系列（未归一化的）分布 δ⟶p(x0,...,xB)α⟶p(x0,...,x2B)α⟶⋯⟶p(x0,...,xT)α (10) 其中 p(x0,...,xkB) 表示长度为 kB 的词元序列上的联合分布，对于任意 k。为方便起见，令 πk 表示由下式给出的分布 πk(x0:kB)∝p(x0:kB)α (11) 假设我们有一个来自 πk 的样本。为了获得一个来自 πk+1 的样本，我们通过用 pprop 采样接下来的 B 个词元 xkB+1:(k+1)B 来初始化一个Metropolis-Hastings过程。我们随后运行MCMC采样过程 NMCMC 步，使用前一节中的随机重采样提议分布 q。完整的细节在算法1中呈现。

请注意，算法1是**单样本**的：尽管进行了多次推理调用，但接受与拒绝新词元的决定纯粹由基础模型似然度做出，以模拟从 pα 采样单个序列。我们可以将其解释为推理时扩展的一个新轴线，因为我们在采样期间花费了额外的计算来获得更高质量/似然度的样本。

为了量化这种扩展，我们可以估计算法1生成的平均词元数。注意，当从 πk(x0:kB) 采样时，每个候选生成步骤平均重采样 2kB 个词元，NMCMC 次。对所有 k 求和，生成的预期词元数为 Etokens=NMCMC∑k=1⌈T/B⌉2kB≈4BNMCMCT2 (12) 这里的关键权衡在于块大小 B 和MCMC步数 NMCMC 之间。较大的 B 需要在中间分布之间进行更大的“跳跃”，需要更大的 NMCMC 来充分过渡。在第5节中，我们凭经验找到了一个 B 的值，使得算法1在相对较小的 NMCMC 值下也能表现良好 。

4.3节 解读：化整为零，积木式构建最优解

理论是完美的，但实践是骨感的。上一节介绍的MH算法虽然强大，但直接应用到长文本生成上会遇到一个致命问题：**效率太低**。想象一下，每修改一个词，就要重新评估和生成整篇几千字的文章，这个计算成本是无法承受的。

为了解决这个难题，作者设计了**算法1**，这是一种非常聪明的**“分而治之”**的策略。其核心思想是：**不要试图一步到位生成整篇完美的文章，而是像搭积木一样，一块一块地搭建，并确保每一块都尽可能完美。**

让我们来逐步解析这个算法的流程：

1. **分块（Block Size B）**：首先，将要生成的总长度 T 分成若干个小块，每块的大小为 B。比如，要生成一篇1024个词的文章，可以分成16块，每块64个词。

2. **生成第一块**：先正常地生成第一块（0-63词）。

3. **精炼第一块**：现在，对这生成的第一块内容，启动MH算法（就是上一节的“登山者”算法）。进行 NMCMC 次迭代（比如10次），在这一小块内部不断地进行“提议-修改-接受/拒绝”，目标是让这一块的质量（似然度）达到最高。

4. **“冻结”并扩展**：当第一块被“精炼”完毕后，就把它固定下来，当作不可更改的前缀。然后，在这个基础上，生成第二块（64-127词）。

5. **精炼“整体”**：现在，我们有了两块内容（0-127词）。再次启动MH算法，但这次的优化范围是**整个已生成的部分**。算法可能会修改第一块的某个词，也可能修改第二块的某个词，目标是让这两块组合在一起的整体质量最高。

6. **重复**：不断重复“冻结-扩展-精炼”这个过程，直到生成完整的文章。

这种渐进式的方法有几个巨大的好处：

- **避免病态初始化**：每一步都是在前一步已经优化的基础上进行的，这使得整个搜索过程非常稳定，不会一开始就走偏。

- **提高效率**：每次MH算法的计算量都限制在当前已生成的长度内，而不是一开始就要面对整个长序列，大大降低了计算复杂度。

推理时扩展（Inference-time Scaling） 这个算法引入了一个全新的概念。传统上，我们想提升模型性能，要么用更大的模型，要么用更多数据去训练。而这个算法提供了一个**第三种选择**：在**使用模型（推理）时**，通过增加计算量来换取更好的结果。从**公式(12)**可以看出，增加 MCMC的步数 NMCMC，就会增加计算成本，但（如下文实验所示）通常也会带来更好的性能。

这带来了一种极具弹性的应用模式：对于简单的任务，我们可以用较少的 NMCMC 快速生成答案；而对于高难度的推理任务，我们可以动态地增加 NMCMC，投入更多计算资源，“压榨”出基础模型的极限性能。这就像给一辆普通汽车装上了“氮气加速”系统，平时正常行驶，关键时刻可以瞬间爆发出强大的动力。这是一种更高效、更经济的AI能力扩展方式。

5 实验

5.1 实验设置

评估。我们使用一套标准的推理基准测试，涵盖数学、编程和STEM（科学、技术、工程和数学）领域（MATH500, HumanEval, GPQA），以及一个评估通用帮助性的非可验证基准测试（AlpacaEval 2.0）。我们对所有我们的方法和基准进行单样本评估；即，基于一个最终的响应字符串。

- **MATH500**：MATH数据集包含竞赛级别的数学问题，涵盖几何、数论和初等代数等七个类别。总共有12500个问题，其中7500个训练问题和5000个测试问题。MATH500是OpenAI标准化的一个从测试集中随机选择的特定子集 。
- **HumanEval**：HumanEval是一套包含164个手写编程问题的集合，涵盖算法、推理、数学和语言理解。每个问题平均关联7.7个单元测试，解决问题对应于通过所有单元测试 。
- **GPQA**：GPQA是一个包含多项选择科学问题（物理、化学和生物学）的数据集，需要高级推理技能才能解决。我们使用GPQA Diamond子集进行评估，该子集包含198个问题，代表了GPQA数据集中质量最高的部分 。
- **AlpacaEval 2.0**：AlpacaEval数据集是包含805个提示的集合，用于衡量通用的帮助性，问题例如要求写影评、提供建议和阅读电子邮件。模型响应由一个自动化的LLM评判员（GPT-4-turbo）进行评分，该评判员确定模型响应相对于基准（也是GPT-4-turbo）的偏好。最终得分是模型响应的胜率，并根据模型响应的长度进行了归一化 。

模型。为了展示我们采样算法的有效性，我们使用了基础模型Qwen2.5-Math-7B、Qwen2.5-7B和Phi-3.5-mini-instruct。对于我们的强化学习基准，我们使用了Shao等人 中GRPO的实现，该实现将这些模型在MATH训练集上进行后训练。对于Qwen2.5模型，我们使用了Shao等人 中用于基准测试其性能的默认超参数。对于Phi-3.5模型，我们使用了一组从Abdin等人 中选取的超参数，以避免训练不稳定，并在大量轮次中收敛到对基础模型的改进 。

采样算法。对于我们的幂采样（算法1）的实现，我们将最大 T 设置为 Tmax=3072（可以通过EOS词元提前终止），块大小 B=3072/16=192。根据经验，我们发现 α=4.0 与一个选择为基础模型且采样温度为 1/α 的提议LLM pprop 相结合，在推理任务上表现最佳。对于AlpacaEval 2.0，我们发现使用更高温度（τ=0.5）的提议分布可以提高性能 。

5.1节 解读：在严格的考场上验证实力

任何科学理论都需要通过实验来验证。这一节详细说明了作者们是如何设计他们的“考场”和“考生”的，以确保对比的公平性和结果的说服力。

考场（评估基准）： 作者选择了四个非常有代表性的公开“考场”，来全面评估模型的推理能力：

- **MATH500**：数学竞赛题，考验严谨的逻辑推理和计算能力。
- **HumanEval**：编程挑战，考验算法设计和代码实现能力。
- **GPQA**：研究生水平的科学问题，这是“地狱难度”的考试，专门用来为难最顶尖的模型。
- **AlpacaEval 2.0**：通用对话能力测试，由更强大的AI（GPT-4）来当考官，评估回答是否“有帮助”，考验的是更主观、更贴近日常应用的综合能力。

这些基准覆盖了从客观对错分明的理科问题，到主观判断的通用对话，能够非常全面地衡量一个模型的综合实力。

考生（模型）： 作者选择了三个不同“家族”、不同规模的模型作为“考生”：

- **Qwen2.5-Math-7B / Qwen2.5-7B**：这是一个在数学方面经过优化的模型和一个通用模型。
- **Phi-3.5-mini-instruct**：这是一个相对较小的模型。

选择不同类型的模型是为了证明本文提出的方法不是“挑食”的，它对不同架构和大小的模型都有效，具有很好的**普适性**。

比赛规则（对比方法）： 每位“考生”都以三种身份参赛：

1. **基础模型（Base）**：原始状态，未做任何处理。
2. **强化学习选手（GRPO）**：经过当前主流的GRPO强化学习方法特训后的状态。
3. **幂采样选手（Ours）**：使用本文提出的幂采样算法的原始模型。

考试参数（算法设置）： 作者还公布了他们算法的具体参数设置，如幂指数 α=4.0、块大小 B=192 等。这就像公布考试时所用的笔和纸的型号一样，保证了实验的**可复现性**，其他研究者可以按照这些参数设置，得到和作者一样的结果。

通过这样严谨的实验设计，作者确保了接下来的结果是可靠和有说服力的，为他们的理论提供了坚实的实践支撑。

5.2 结果

主要结果。我们在表1中展示了我们的主要结果。跨越不同家族的基础模型，我们的采样算法在不同的推理和评估任务上，相较于不同的推理和评估任务，在单样本准确率和得分上实现了巨大的、近乎普遍的提升，例如，在Phi-3.5-mini上对HumanEval的提升高达+51.9%，在Qwen2.5-Math上对MATH500的提升高达+25.2%。特别是在MATH500上，这是强化学习后训练的领域内任务，幂采样实现的准确率与GRPO获得的准确率相当。此外，在领域外的推理任务上，我们的算法在GPQA上再次与GRPO持平，而在HumanEval上实际上超越了GRPO，最高提升达+59.8%。同样，幂采样在非可验证的AlpacaEval 2.0上持续优于对手，这表明我们的提升具有泛化到可验证性之外领域的能力。这个从根本上简单却免训练的采样算法的惊人成功，凸显了现有基础模型潜在的推理能力 。

表1：幂采样（我们的方法）在不同模型家族和任务上匹敌甚至超越GRPO 我们在MATH500、HumanEval、GPQA和AlpacaEval 2.0上对我们的采样算法的性能进行了基准测试。我们**加粗**了我们的方法和GRPO的分数，并在我们的方法优于GRPO时用**下划线**标出。在所有模型中，我们看到幂采样在领域内推理（MATH500）上与GRPO相当，并且在领域外任务上能够超越GRPO。

	MATH500	HumanEval	GPQA	AlpacaEval2.0
Qwen2.5-Math-7B				
Base	0.496	0.329	0.278	1.61
Low-temperature	0.690	0.512	0.353	2.09

	MATH500	HumanEval	GPQA	AlpacaEval2.0
Power Sampling (ours)	**0.748**	**0.573**	**0.389**	**2.88**
GRPO (MATH)	**0.785**	**0.537**	**0.399**	**2.38**
Qwen2.5-7B				
Base	0.498	0.329	0.278	7.05
Low-temperature	0.628	0.524	0.303	5.29
Power Sampling (ours)	**0.706**	**0.622**	**0.318**	**8.59**
GRPO (MATH)	**0.740**	**0.561**	**0.354**	**7.62**
Phi-3.5-mini-instruct				
Base	0.400	0.213	0.273	14.82
Low-temperature	0.478	0.585	0.293	18.15
Power Sampling (ours)	**0.508**	**0.732**	**0.364**	**17.65**
GRPO (MATH)	**0.406**	**0.134**	**0.359**	**16.74**

5.2节 解读：用数据说话，结果令人震撼

这是整篇论文的“高光时刻”。**表1**中的数据清晰、直接地证明了作者所有理论的有效性。让我们来深入解读这张表格揭示的关键信息：

1. **巨大的性能飞跃**：首先，对比每一组中的“Base”（基础模型）和“Power Sampling (ours)”（我们的方法），可以看到后者在所有推理任务上都取得了惊人的性能提升。例如，在Qwen2.5-Math-7B模型上，MATH500的准确率从49.6%暴涨到74.8%。这无可辩驳地证明了：**通过更优的采样算法，确实可以解锁基础模型中隐藏的巨大潜力** 。

2. **与顶尖训练方法的正面交锋**：再来看“Power Sampling”和“GRPO (MATH)”的对比。GRPO是当前最先进的强化学习训练方法之一。

    - **在GRPO的主场（MATH500）**，我们的方法表现与GRPO旗鼓相当（例如，Qwen2.5-Math-7B上是74.8% vs 78.5%）。这意味着一个**完全不需要训练**的“轻量级”算法，竟然能在对方最擅长的领域，与一个需要大量计算资源进行训练的“重量级”选手打得有来有回。

    - **在GRPO的客场（领域外任务）**，我们的方法展现出了更强的**泛化能力**。在HumanEval（编程）和AlpacaEval（通用对话）任务上，我们的方法几乎全面超越了GRPO。这说明强化学习可能会让模型对其训练过的任务产生“偏科”，而在其他任务上表现不佳。而我们的采样方法似乎更能激发模型普适的、底层的推理能力 。

3. **对小模型效果尤其显著**：观察**Phi-3.5-mini-instruct**这个小模型的结果，结论更加惊人。在这个模型上，GRPO的训练效果非常不稳定，在HumanEval上的表现甚至比基础模型还要差（从21.3%降到13.4%），这暴露了强化学习训练的脆弱性。然而，我们的幂采样方法却带来了稳定且巨大的提升，将HumanEval的准确率从21.3%提升到了73.2%！这表明，**对于资源有限的小模型，我们的方法可能是一种更可靠、更高效的能力提升途径** 。

总而言之，这张表格里的数据是强有力的证据，它不仅证明了作者提出的算法是有效的，更深层次地，它挑战了整个行业对于“如何让AI更聪明”的传统认知。它告诉我们，通往更强AI的道路，除了“更大模型、更多训练”这条路之外，还存在着一条“更巧算法、更优推理”的捷径。

5.3 分析

我们分析了幂采样的推理特性与GRPO的关系。我们在表2中展示了一个例子，更多例子见附录A.3。

推理路径的似然度和置信度。根据设计，幂采样旨在从基础模型中采样更高似然度的序列。在图4中，左图绘制了基础模型、幂采样和GRPO在MATH500上的输出序列对数似然度（按长度平均）的直方图，其中似然度是相对于Qwen2.5-Math-7B基础模型计算的。我们的方法从基础模型的更高似然区域进行采样，正如预期的那样，但仍然保持了明显的分布宽度。与此同时，GRPO的样本则高度集中在最高的似然峰值处 。

图4：MATH500响应在基础模型（Qwen2.5-Math-7B）下的似然度和置信度。 左图：我们绘制了原始、幂采样和GRPO响应在MATH500上的对数似然度（相对于基础模型）。右图：我们对置信度做了同样的操作。我们观察到GRPO从最高似然度和置信度的区域采样，幂采样紧随其后，这与更高的经验准确率相关。

我们还绘制了MATH500响应的基础模型置信度，定义为下一词元分布的平均负熵（不确定性）： Conf(x0;T)=T+11∑t=0T∑x∈Xp(x∣x<t)log p(x∣x<t) (13) 图4的右图表明，我们的方法和GRPO的响应都从基础模型的相似高置信度区域采样，这同样对应于更高似然度和正确推理的区域 。

推理路径长度。强化学习后训练的另一个决定性特征是长式推理，其中样本倾向于表现出更长的响应。在MATH500上，Qwen2.5-Math-7B的平均响应长度为600个词元，而GRPO平均为671个词元。令人惊讶的是，幂采样达到了相似的平均长度679个词元，而没有被明确鼓励偏好更长的生成。这是从采样过程中自然涌现的 。

多样性与pass@k性能。再次注意图4中GRPO相对于幂采样分布宽度的尖锐且高度集中的似然度/置信度。这表明GRPO表现出多样性崩塌，而我们的采样器则没有，这与RL后训练以牺牲多样性为代价强烈锐化基础模型分布的观察结果一致。为了量化幂采样相对于GRPO的比较多样性，我们可以绘制pass@k准确率曲线，其中如果k个样本中至少有一个是准确的，则问题被视为解决。图5恰好显示了这一点：与GRPO不同，其pass@k性能在k较大时逐渐减弱，而幂采样在 k>1 时表现强劲。此外，我们的性能曲线超越了基础模型的曲线，直到最终在性能上趋于一致。特别地，我们能够在不损害多样本性能的情况下实现GRPO级别的单样本性能（其他领域的见附录A.2），解决了RL后训练一个长期存在的缺点 。

图5：在MATH500上的Pass@k性能。 我们绘制了幂采样（我们的方法）和RL（GRPO）相对于基础模型（Qwen2.5-Math-7B）的pass@k准确率（k个样本中至少有一个准确即为正确）。我们的性能曲线严格优于GRPO和基础模型，并且我们在高k值时的通过率与基础模型相匹配，展示了持续的生成多样性。

幂分布的影响。幂采样的两个最重要的超参数是 α 的选择和序列生成过程中的MCMC（重采样）步数 NMCMC。在极端情况下，选择 α=1.0 直接从基础模型采样，而取 α→∞ 的效果是确定性地接受任何严格增加似然度的重采样序列。当然，尽管更高的基础模型似然度与更好的推理相关（图4），但直接优化似然度不一定对推理是最优的，这表明存在一个理想的中间值 α。在图6中，我们展示了不同 α 值下的MATH500准确率，并发现中间值 α=4.0 的表现优于其他值，正如预期的那样。值得注意的是，幂采样的准确率在 α≥2.0 之后保持相对稳定，这表明幂采样在实践中对 α 的选择相对鲁棒 。

图6：超参数对幂采样的影响。 左图：我们绘制了不同模型家族在各种 α 值下的MATH500准确率。右图：我们绘制了随着MCMC步数增加，幂采样在Qwen模型上准确率的提升。

通过MCMC步数进行测试时扩展。另一方面，NMCMC 调节了我们算法在推理时消耗的计算资源，为测试时扩展提供了一个自然的轴线。在4.3节中，我们提出了混合时间的概念，即充分从目标分布采样所需的MCMC步数。在我们的案例中，我们预期我们采取的MCMC步数越少，我们的算法采样就越偏离目标 pα。我们在图6中绘制了性能对 NMCMC 的依赖关系，并注意到准确率稳步增加，直到 NMCMC=10，之后准确率大致保持稳定（未绘出）。使用较少MCMC步数带来的准确率差异是明显的，但在 NMCMC=2 和 NMCMC=10 之间不超过3-4%。然而，使用至少两步与完全不使用相比，准确率的跃升是显著的（3-4%）。

我们甚至可以计算我们的方法相对于运行GRPO生成的总词元数。根据式(12)，我们的采样器生成长度为T的序列时，生成的词元数是标准推理的 4B1·NMCMCT2 倍。代入我们的实验参数 NMCMC=10，T=679（我们在MATH500上的平均输出长度）和 B=192，运行幂采样推理的成本是标准推理词元数的8.84倍。由于GRPO在训练期间每个样本会生成多个rollout，我们的方法产生的推理成本与一个epoch的GRPO训练大致相同，假设每个样本有8个rollout且数据集大小相同。不过通常情况下，一个GRPO epoch仍然更昂贵，因为它使用16个rollout和一个比MATH500更大的训练集 。

表2：HumanEval上的样本响应：Phi-3.5-mini-instruct 问题：过滤一个输入的字符串列表，只保留那些以给定前缀开头的字符串。

方法	响应	通过
Ours	`return [s for s in strings if s.startswith(prefix)]`	true
GRPO	`return [string for string in strings if string.startswith(f'{prefix} *2')]`	false

5.3节 解读：深入探究“为什么有效”

如果说上一节的结果回答了“是什么”，那么这一节的分析则深入探讨了“为什么”。作者通过一系列精巧的分析，揭示了他们方法成功背后的深层机制。

1. **信心与多样性的完美平衡 (图4)**：这张图非常直观。左边的似然度直方图显示：
    - **基础模型（Base）** 的答案质量参差不齐，分布很宽。
    - **GRPO** 的答案高度集中在最右侧的“高峰”，这意味着它非常自信，但答案高度同质化，这就是**“模式坍塌”**或**“多样性崩塌”**的体现。它只会用一种方式解决问题。
    - **我们的方法（Ours）** 的分布也显著右移，说明它同样能生成高信心、高质量的答案。但关键在于，它的分布比GRPO**更宽**，保留了相当的多样性。它既能找到最优解，又能提供多种不同的高质量思路。
2. **自然涌现的长式推理**：一个有趣的发现是，本文的方法和强化学习一样，都倾向于生成更长、更详细的“解题步骤”。这并非刻意为之，而是算法在寻找最高概率路径过程中的自然结果。高质量的推理往往需要详尽的步骤，模型在优化过程中自发地学会了这一点。
3. **“两全其美”的pass@k性能 (图5)**：这张图是本文方法优越性的最强证明。
    - **GRPO** 的曲线在 `k=1` 时很高，但之后迅速“躺平”，几乎不再增长。这再次说明了其多样性的匮乏：试一次不行，再试多少次都一样。
    - **基础模型（Base）** 的曲线虽然起点低，但持续稳定上升，显示了其固有的多样性。
    - **我们的方法（Ours）** 的曲线堪称完美：它**起点像GRPO一样高**（单次尝试成功率高），并且**增长趋势像基础模型一样陡峭**（多次尝试能不断发现新解法）。这真正实现了**“鱼与熊掌兼得”**：既有强化学习的高准确率，又保留了基础模型的高多样性，解决了强化学习长期以来的一个核心痛点。
4. **超参数的鲁棒性与可扩展性 (图6)**：
    - 左图显示，虽然 α=4.0 是一个最佳选择，但在一个很宽的范围（α≥2.0）内，算法性能都非常稳定。这说明该方法**不娇气、易于使用**，不需要繁琐的“炼丹”（调参）。
    - 右图则直观地展示了**“测试时扩展”**的概念。投入的MCMC步数越多（即推理时消耗的计算资源越多），准确率就越高，直到一个饱和点。这为使用者提供了一个可以自由调节的“性能-成本”旋钮。
5. **计算成本的惊喜**：最后，作者进行了一个成本核算。他们发现，使用幂采样方法生成一个答案的计算成本，大致只相当于强化学习进行**一轮（epoch）训练**的成本。考虑到强化学习通常需要成百上千轮的训练，并且需要庞大的数据集，本文的**免训练**方法在总体计算效率上，要比强化学习高出几个数量级。它将巨大的、一次性的训练成本，巧妙地转化为了灵活的、按需分配的推理成本。

表2中的例子则是一个生动的缩影：对于一个简单的编程问题，我们的方法给出了简洁正确的Python列表推导式，而经过强化学习的GRPO模型却犯了一个莫名其妙的低级错误，这再次印证了强化学习可能带来的不稳定性和泛化能力下降的问题。

6 结论

在这项工作中，我们提出了一种直接从基础模型采样而无需任何额外训练或访问外部信号的算法，实现了与最先进的强化学习后训练算法相媲美，有时甚至更好的单样本推理性能。我们利用对强化学习分布锐化的讨论，来推动将幂分布定义为一个有价值的推理目标分布。尽管精确的幂分布采样是不可行的，我们采用了经典的MCMC技术，结合自回归生成的序列结构，定义了我们的幂采样算法，该算法展示了强大的经验性能。

我们的结果表明，基础模型的能力在采样时未被充分利用，并指向了基础模型的高似然区域与强大推理能力之间的密切关系。在采样时利用额外的计算，并对基础模型能力有更深入的理解，为将推理范围扩展到可验证性之外提供了一个有前途的方向 。

结论解读

论文的结论部分简洁而有力地总结了整个研究的核心贡献和未来展望。

核心贡献回顾： 作者带领我们回顾了整个研究旅程：

1. **提出问题**：从质疑强化学习（RL）的“分布锐化”本质出发，认为RL可能只是放大了模型已有的能力，而非创造新能力。
2. **定义目标**：提出了“幂分布”作为一种更理想的采样目标，它能更好地引导模型走向高质量的推理路径。
3. **设计算法**：巧妙地将经典的MCMC统计方法与语言模型的自回归特性相结合，设计出一种实际可行的“幂采样”算法。
4. **实验验证**：通过在多个权威基准上的严格实验，证明了该算法的有效性——它不仅在性能上能媲美甚至超越复杂的RL训练，还保留了宝贵的答案多样性。

核心发现与启示： 最核心的发现是：**我们严重低估了基础模型本身的能力**。它们强大的推理潜力如同沉睡的宝藏，而现有的常规采样方法就像一把错误的钥匙，无法将其开启。本文提出的幂采样算法，则像一把精心设计的钥匙，通过在推理时投入更多的计算进行智能搜索，成功解锁了这些潜能。这揭示了模型的高似然度区域（即模型“最自信”的回答区域）与正确的推理之间存在着紧密的联系。

未来方向： 这项工作为人工智能的发展指明了一个充满希望的新方向。在当前这个“模型越大越好，训练数据越多越好”的“军备竞赛”时代，本文提供了一种不同的思路：**与其盲目地扩大规模，不如更深入地理解和利用好我们已有的模型**。通过在推理时设计更智能、更强大的算法，我们或许能以更低的成本实现更高的智能。特别是，这种“免验证器”的特性，为AI在法律、文学、艺术等没有标准答案、无法简单验证的复杂领域中发挥其推理能力，打开了想象空间。

附录A

A.1 额外的理论讨论

在本节中，我们对幂采样降权那些将输出困在低似然未来的词元，而低温采样则不然的现象，提供一个更强的形式化描述。

命题2（非正式）。幂采样增权那些支持集小但补全似然度高的词元，而低温采样增权那些支持集大但补全似然度低的词元。

定义2。在本节的其余部分，固定一个前缀 x0:t−1。如果 ∑x>tp(x0,...,xt,...xT)=ϵ，我们说 xt 在下一词元条件分布下具有边际权重 ϵ。

我们考虑“关键窗口”或“关键枢纽词元”现象的一个简化模型，它指的是对最终生成质量有强烈影响的中间词元。我们区分导致高似然未来的关键枢纽词元和导致低似然未来的关键枢纽词元。

定义3。在一个极端情况下，如果一个关键枢纽词元将其全部边际权重 ϵ 放在一个未来上（奇异支持集），即只有一个 x>t 的选择使得 p(x0,...,xt,...,xT) 非零，那么它最大限度地诱导了一个高似然补全。我们称这样的词元为**正向关键枢纽词元**。

定义4。在另一个极端情况下，如果一个关键枢纽词元的全部边际权重 ϵ 均匀分布在 N 个未来补全上，那么它最小化了任何未来的似然度。换句话说，存在 N 个补全 x>t，使得 p(x0,...,xt,...,xT) 均为非零，且似然度为 Nϵ。我们称这样的词元为**负向关键枢纽词元**。

我们的高低似然未来简化模型研究了在给定采样分布下，何时正向关键枢纽词元比负向关键枢纽词元更受青睐。特别地，我们表明，即使后者的边际权重更高，幂采样也可以增权一个正向关键枢纽词元超过一个负向关键枢纽词元，而在这种情况下，低温采样总是增权负向关键枢纽词元。当然，只要正向关键枢纽词元具有更高的边际权重，幂采样和低温采样都会增权它。

命题3。令 xt 为一个边际权重为 ϵ 的正向关键枢纽词元，令 xt′ 为一个边际权重为 ϵ′ 且支持集大小为 N 的负向关键枢纽词元。那么如果 N1−1/αϵ′<ϵ<ϵ′ (14) xt 的未来似然度高于 xt′ 的任何未来似然度。此外，幂采样增权 xt 超过 xt′，而低温采样增权 xt′ 超过 xt。

证明。由于 α≥1，因此 N1−1/αϵ′>Nϵ′ (15) 因此 ϵ>Nϵ′，这确立了 xt 的未来补全似然度大于 xt′ 的未来补全似然度（即正向和负向关键枢纽词元的分配是一致的）。 现在，如果 ϵ<ϵ′，那么在低温分布下，xt 和 xt′ 上的相对边际权重为 ϵα 和 ϵ′α，因此选择 xt 的概率相对于 xt′ 被降权。 然而，对于幂分布，相对边际权重为 ppow(xt|x<t)=ϵα 和 ppow(xt′|x<t)=Nα−1ϵ′α。 那么，只要 ϵα>Nα−1ϵ′α⟺ϵ>N1−1/αϵ′，词元 xt 将相对于词元 xt′ 被增权。 换句话说，xt 的边际权重在 p 下可以小于 xt′ 的权重，但如果 xt 的补全似然度高于 xt′ 的任何单个补全似然度，幂采样就会偏好 xt 超过 xt′。∎

A.1节 解读：对“远见”的数学证明

这部分内容是对4.1节中“迷宫寻宝”比喻的一个严格的数学形式化。作者通过定义“正向关键枢纽词元”（通往唯一大宝藏的路）和“负向关键枢纽词元”（通往多个小宝箱的路），并利用**命题3**，从数学上证明了幂采样和低温采样的根本区别。

命题3的核心思想是：存在一种特定的情况（由不等式14描述），在这种情况下：

1. 从短期来看，“负向”路径（xt′）的总概率（ϵ′）要大于“正向”路径（xt）的总概率（ϵ）。
2. 但是，“正向”路径通往的那个唯一未来的概率，要远大于“负向”路径下任何一个未来的概率。

在这种情况下：

- **低温采样**（目光短浅的寻宝者）只看到了第一点，它会选择总概率更高的“负向”路径 xt′。
- **幂采样**（有远见的规划师）则能够穿透表面的总概率，识别出“正向”路径 xt 背后那条极高价值的单一未来，并选择它。

这个证明为幂采样的优越性提供了坚实的理论基石。它从数学上解释了为什么幂采样在需要长远规划的复杂推理任务中表现更好，因为它内在地偏好那些能够导向全局最优解（最高似然度完整序列）的关键决策，而不是那些仅仅在下一步看起来不错的局部最优选择。

A.2 跨多个领域的Pass@k准确率

在本节中，我们绘制了幂采样、GRPO和基础模型（Qwen2.5-Math-7B）在MATH500、GPQA和HumanEval上的pass@k性能，以证明我们的采样算法在单样本和多样本推理上都表现出色，同时保持了响应的多样性。幂采样在MATH500和GPQA上使用 α=4.0 绘制，在HumanEval上使用 α=1.67（这个温度在较早的k值下表现稍好）。在所有情况下，无论是GRPO的领域内还是领域外任务，幂采样在 k>1 时的pass@k性能几乎普遍优于GRPO和基础模型，并在大的k值下达到甚至超过基础模型的上限。

图7：在MATH500上的Pass@k性能 (Qwen2.5-Math-7B)。

图8：在HumanEval上的Pass@k性能 (Qwen2.5-Math-7B)。

图9：在GPQA上的Pass@k性能 (Qwen2.5-Math-7B)。

关于这些图需要注意的一点是，多样性的损失在不同基准测试中差异显著。MATH500和GPQA清楚地显示，即使在较小的k值下，GRPO的pass@k性能和多样性也显著较低，而在HumanEval上，GRPO的pass@k表现优于基础模型，直到 k=16。我们推测这可能是因为MATH500和GPQA是根据“单一”答案评分的，而HumanEval是基于完成代码，其中可能有多个解决方案被接受，这表明多样性的减少可能会出现，但在更高的k值时才会显现 。

A.2节 解读：全方位展示“两全其美”的优势

这三张图（图7、8、9）是正文中图5的扩展，它们在三个不同的核心推理基准上，再次印证了幂采样在**准确率和多样性**之间取得的卓越平衡。

- 在**MATH500（数学）**和GPQA（科学）**这两个有唯一正确答案的领域，结论与正文一致：GRPO的曲线很快就变得平缓，表明其答案高度重复，多次尝试也无法提供新的正确解法。而“Ours”（我们的方法）的曲线则持续上扬，展现了强大的多样性，能够在多次尝试中不断找到正确答案。
- 在**HumanEval（编程）**上，情况稍微有些不同。编程问题的解法可能不止一种，因此多样性的价值体现得更充分。在这里，GRPO的pass@k性能在前期甚至优于基础模型，但增长趋势仍然较为缓慢。而我们的方法则再次展现了最佳性能，曲线始终处于最上方，结合了高起点和高增长率。

这些图表共同构成了一个强有力的证据链：无论是在需要单一精确答案的领域，还是在允许多种解决方案的领域，幂采样都展现出了比强化学习（GRPO）和基础模型更优的综合性能。它真正做到了在不牺牲多样性的前提下，大幅提升单次推理的准确率，实现了“1+1>2”的效果。

A.3 更多定性示例

表3：在MATH500上的比较：Qwen2.5-Math-7B 问题：你能解决下面的数学问题吗？将 203 写成小数。请逐步推理，并将最终答案放在\boxed{{}}中。正确答案：0.15

系统	回答
Ours	要将分数 203 转换为小数，我们可以遵循以下步骤： 1. 确定分数的分子和分母。在这种情况下，分子是3，分母是20。 2. 用分子除以分母。在这种情况下，我们用3除以20。 3. 执行除法。结果是0.15。 因此，分数 203 作为小数是0.15。 \boxed{{0.15}}
GRPO	要将分数 203 转换为小数，我们可以遵循以下步骤： 1. 理解分数：分数 203 意味着3除以20。 2. 执行除法：我们需要用3除以20。由于3小于20，我们将在3后面加上小数点和零，以便进行除法。 3. 设置除法：将3写成3.000（为精确起见添加三个零）。 4. 除法：20进入30一次（因为 20×1=20）。从30中减去20得到余数10。将下一个0带下来，得到100。20进入100五次（因为 20×5=100）。从100中减去100得到余数0。 5. 组合结果：除法的商是0.15。 因此，203 的小数表示是0.15。 \boxed{{0.15}}

表4：在HumanEval上的比较：Phi-3.5-mini-instruct 问题：Fib4数序列是一个类似于斐波那契序列的序列，定义为：fib4(0)=0, fib4(1)=0, fib4(2)=2, fib4(3)=0，对于 n≥4, fib4(n)=fib4(n−1)+fib4(n−2)+fib4(n−3)+fib4(n−4)。编写一个函数来高效地计算第n个元素，不要使用递归。

系统	完成代码	通过
Ours	`# 基础情况 if n == 0: return 0 elif n == 1: return 0 elif n == 2: return 2 elif n == 3: return 0 # 初始化前四个元素 fib4_values = # 迭代计算序列的其余部分 for i in range(4, n + 1): next_value = (fib4_values[i-1] + fib4_values[i-2] + fib4_values[i-3] + fib4_values[i-4]) fib4_values.append(next_value) # 返回第n个元素 return fib4_values[n]`	true
GRPO	`a, b, c, d = 0, 0, 2, 0 for _ in range(n): a, b, c, d = b, c, d, a+b+c+d return d`	false

表5：在MATH500上的比较：Qwen2.5-Math-7B 问题：哪个整数n满足 0≤n<18 且 n≡−11213141(mod18)？请逐步推理，并将最终答案放在\boxed{{}}中。正确答案：13

方法	回答
Ours	首先，我们执行11213141除以18的除法： 11213141÷18≈622952.277777... 商的整数部分是622952。现在，我们将622952乘以18来求积： 622952×18=11213136 接下来，我们从11213141中减去这个积来求余数： 11213141−11213136=5 所以，11213141除以18的余数是5。这意味着： 11213141≡5(mod18) 因此： −11213141≡−5(mod18) 因为我们需要一个非负余数，我们通过加上18将-5转换为一个正的等价值： −5+18=13 因此： −11213141≡13(mod18) 满足 0≤n<18 且 n≡−11213141(mod18) 的整数n是： 13 \boxed{{13}}
GRPO	步骤1：求11213141除以18的余数。我们可以使用模算术的性质，即 a≡b(modm) 当且仅当 a mod m = b mod m。所以，我们需要计算 11213141 mod 18。 为了简化计算，我们可以利用一个数除以18的余数与其各位数字之和除以18的余数相同（虽然这不总是最有效的方法，但在这里可以为简单起见使用）。 首先，让我们将11213141的各位数字相加： 1+1+2+1+3+1+4+1=14。 现在，我们需要求14除以18的余数。由于14小于18，余数就是14。因此， 11213141≡14(mod18)。 步骤2：求-11213141除以18的余数。由于 11213141≡14(mod18)，因此 −11213141≡−14(mod18)。 我们需要将-14转换为0到17之间的正余数。为此，我们将18加到-14上： −14+18=4。 因此， −11213141≡4(mod18)。 满足 0≤n<18 且 n≡−11213141(mod18) 的整数n是4 \boxed{{4}}

A.3节 解读：从具体案例看推理能力的差异

如果说前面的数据是宏观统计，那么附录中的这些具体例子就是微观的、生动的案例分析。它们让我们能直观地感受到不同方法在“思考”方式上的差异。

- **表3（简单数学题）**：在这个例子中，两种方法都给出了正确答案和详细的步骤。这说明对于简单问题，两者都能胜任。但值得注意的是，GRPO的解释过程更像是“教学生如何做长除法”，而我们的方法则更直接地陈述了核心步骤。
- **表4（编程题）**：这是一个绝佳的对比案例。
    - **我们的方法（Ours）** 给出的代码逻辑清晰，可读性强。它先处理了基础情况，然后用一个列表来存储计算过程中的值，最后返回结果。这是非常标准和稳健的动态规划解法。
    - **GRPO** 的代码则试图用一种更“聪明”的、空间复杂度更低的滚动变量方法来解决，但它的逻辑是**错误**的。`return d` 在循环结束后返回的是 `a+b+c+d` 的值，而正确的 `fib4(n)` 应该是 `fib4(n-1)`，即更新前的 `d`。这个例子生动地展示了强化学习可能导致的“聪明反被聪明误”——模型学会了一种看似高级但实际上错误的模式。

- **表5（模算术题）**：这是最能体现推理能力差异的例子。
    - **我们的方法（Ours）** 采用了最经典、最稳妥的求余方法：先做除法，取商的整数部分，再用被除数减去商与除数的积，得到正确的余数5。后续的步骤完全正确，得到了正确答案13。
    - **GRPO** 犯了一个致命的**事实性错误**。它声称“一个数除以18的余数与其各位数字之和除以18的余数相同”。这是一个只对除数是3或9时才成立的特殊性质，对于18并不成立！基于这个错误的“定理”，它后续的所有计算，尽管步骤看起来很“有逻辑”，但从根上就错了，最终得出了错误答案4。

这些案例雄辩地证明，幂采样方法不仅在统计上得分更高，其生成的推理过程也更加**可靠和严谨**。它倾向于使用更基础、更普适的正确方法。而经过强化学习的模型，有时会学到一些看似是捷径、实则是错误的“伪知识”，导致在关键问题上出现“一本正经地胡说八道”的情况。这凸显了幂采样在提升AI推理**鲁棒性**方面的巨大价值。