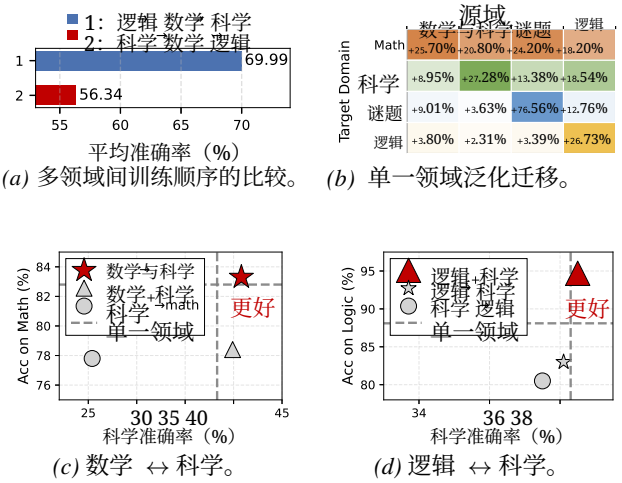
arXiv:2602.01365v1 [cs.LG] 1 Feb 2026

当领域相互作用时：推理强化学习中的非对称且顺序敏感的跨领域效应

Wang Yang [1] **Shouren Wang** [1] **Chaoda Song** [1] **Chuang Ma** [2] **Xinpeng Li** [1] **Nengbo Wang** [1] **Kaixiong Zhou** [3] **Vipin Chaudhary** [1] **Xiaotian Han** [1]

摘要

组相对策略优化（GRPO）已成为提升大语言模型推理能力的关键技术，但我们对其在不同领域排序策略下的表现仍知之甚少。尤其是，GRPO 中顺序训练（一次训练一个领域）与混合领域训练（一次训练多个领域）的影响尚未得到系统研究。我们首次系统分析了数学、科学、逻辑和谜题推理任务中的训练顺序效应。我们发现：（1）单域泛化具有高度不对称性：在其他领域上训练可使数学推理准确率提升约 25%，但对逻辑和谜题的迁移微乎其微；（2）跨领域交互高度依赖顺序：按照数学→科学的顺序训练时，数学／科学准确率分别达到 83%／41%；而将顺序反转为科学→数学则会使性能降至 77%／25%；（3）多领域训练不存在普遍最优的单一策略：顺序训练有利于数学（最高达 84%），混合训练则有利于科学和逻辑，而不佳的顺序可能导致显著的性能差距（从 70% 降至 56%）。总的来说，我们的研究结果表明，在多领域设置下，GRPO 表现出显著的不对称性、顺序敏感性和策略依赖性，凸显了采用领域感知和顺序感知的训练设计的必要性。我们的代码见 https://github.com/uservan/cross_domain。



(a) 多领域间训练顺序的比较。 *(b)* 单一领域泛化迁移。

(c) 数学 ↔ 科学。 *(d)* 逻辑 ↔ 科学。

图*1*。 **GRPO 中的跨领域迁移与干扰。（a）**不同领域间的训练顺序比较。相反的顺序训练会导致平均多领域性能出现显著差异。 **（b）**单领域迁移图。数值（例如 +20.8%）表示在源域（科学）上训练并在目标域（数学）上评估时的准确率提升；颜色越深表示提升越大。数学表现出强跨领域迁移，科学表现出中等程度的迁移，而逻辑/谜题的迁移最小。 **（c）和（d）**非对称双领域迁移。“→”表示顺序训练；“+”表示混合训练；虚线表示单一领域基线；越接近右上方的点，代表两个领域上的联合性能越好。训练顺序至关重要：数学→科学可带来显著更高的准确率，而在顺序训练中，科学会干扰逻辑。

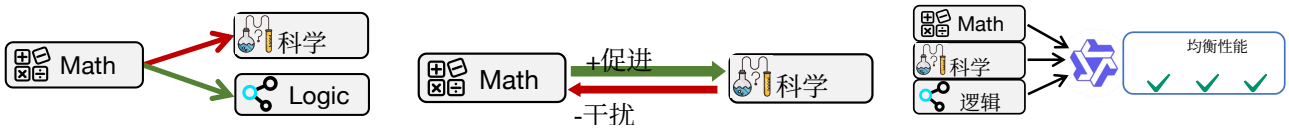
1. 引言

现有针对群组相对策略优化（GRPO）的改进大多Shao 等人，2024年；Liu 等人，2024年）主要在数学和代码领域进行开发和评估，例如 DAPO（Yu等人，2025）、Dr. GRPO（Liu 等人，2025）、GSPO（Zheng等人，2025）等。现有将 GRPO 应用于其他领域的研究通常分别独立考察每个领域，例如逻辑（Xie等人，2025）和谜题（Chen等人，2025）。

相比之下，人们很少关注多领域训-练下 GRPO 的行为。多领域训-练期间，领域特定信号如何交互仍不清楚。如图1a所示，令人意外的是，不同领域采用相反的训练顺序，可能会产生从高准确率到低准确率不等的评估结果，这表明多领域 GRPO 可能对领域交互高度敏感。这促使我们对推-

[1]凯斯西储大学 [2]日本 NII [3]北卡罗来纳州立大学。通信作者：Vipin Chaudhary<vxc204@case.edu>，Xiaotian Han <xxh584@case.edu>。

预印本。*2026*年*2*月*3*日。



(a) 单域泛化 *(b)* 跨域交互 *(c)* 多领域训练

图 *2.* 我们对多领域设置下 GRPO 的系统分析概览。**(a)** 单域泛化：在一个领域上训练如何迁移至其他领域。**(b)** 跨域交互：不同训练顺序下两个领域之间的非对称促进作用与干扰。**(c)** 多领域训练：用于提升各领域性能并实现均衡性能的不同训练策略。

表 *1.* 多领域训练中代表性训练策略的准确率比较。顺序策略表现差异很大——例如，逻辑→数学→科学的表现显著优于科学→数学→逻辑。此外，不同领域偏好不同的训练范式：数学从最优顺序策略中获益最大，而科学和其他领域在混合训练下表现更好。

训练策略	Math	科学	逻辑
单一领域	82.76	38.28	88.14
逻辑 →数学 → 科学	**84.00**	36.67	87.31
科学 →数学 → 逻辑	78.24	26.97	63.80
数学 + 科学 + 逻辑	80.44	**40.00**	**95.37**

理训练中的跨领域效应开展系统研究。

在图1中，我们首先系统地探索并分析组相对策略优化在多领域设置中的表现。我们的研究围绕三个主要方面展开：（1）单域泛化：考察单领域训练如何影响其他领域的推理表现。（2）跨域交互：探究两个领域之间的跨领域交互，考察在另一个领域上训练时，给定领域中的推理能力会得到增强还是削弱。（3）多领域训练：研究如何直接在多个领域上训练，以提升所有领域的均衡性能。为便于开展这项分析，我们构建了涵盖四个代表性领域的训练和评估数据集：数学、科学、逻辑和谜题。

单域泛化的结果见图1b。模型在一个领域（训练领域）上训练，并在另一个领域（测试领域）上评估。测试领域的准确率提升越大（最终步骤的准确率减去初始步骤的准确率），颜色就越深。我们观察到：（1）几乎所有其他领域的数据训练都能有效激发数学推理能力；（2）科学推理也能从跨领域训练中获益，但获益程度不及数学；（3）逻辑和谜题推理能力很难从其他领域习得，跨领域提升也微乎其微。

跨领域结果汇总于图1c和图1d。我们观察到：(1)训练顺序对数学和科学领域至关重要。按数学→科学的顺序训练，在这两个领域都取得了最佳性能

数学和科学，而反向顺序科学→数学会导致这两个领域的性能大幅下降。混合训练在一定程度上提升了整体性能，但仍不如数学→科学的训练顺序。(2) 科学训练会持续干扰逻辑推理。无论采用科学→逻辑还是逻辑→科学的顺序，模型在逻辑领域的性能都会下降。值得注意的是，对逻辑和科学进行混合训练，能够有效减轻这种干扰，并提高模型在这两个领域的准确率。

多领域训练的结果见表 1。我们比较了单领域训练、平均表现最佳和最差的顺序训练策略，以及混合训练。我们观察到：(1)在顺序训练中，领域顺序对性能有显著影响。例如，科学→数学→逻辑的训练顺序，在科学和逻辑领域的准确率分别为 26.67 和 63.8，远低于相反顺序（逻辑→数学→科学）取得的准确率。(2)不同领域适合不同的训练策略。对于数学领域，最佳顺序策略取得最高准确率，可达 84%。相比之下，混合训练更适合科学和逻辑领域，并能在这些领域取得最佳整体性能。

总而言之，我们的贡献可概括如下：

- 我们表明，GRPO 中的领域交互具有高度非对称性，并且强烈依赖于训练顺序。
- 我们通过明确训练顺序何时至关重要、何时应优先采用混合训练，为多领域 GRPO 提供实用指导。忽略这些影响可能会引入显著偏差；最差的顺序策略表现远逊于最优策略（例如，平均准确率为 56%，而非 70%）。
- 我们首次对多领域训练下的 GRPO 进行了系统性研究，为未来针对推理模型的原则性多领域强化学习研究奠定了基础。

2. 单域泛化

本节首先考察领域特定训练的跨领域影响。具体而言，我们研究在某一领域（如数学）数据上进行训练会如何影响

表2. 四个领域训练数据集概览。“测试”表示使用数据集来源中的测试数据。

领域	数据集来源	Size	测试集
Math	Skywork-OR1-RL-Data	5K	MATH500
科学	OpenScienceReasoning-2	5K	GPQA
逻辑	骑士与无赖数据集	5K	test
谜题	Enigmata-Data	5K	test

其他领域（如科学）的性能。实验结果表明，数学和科学领域更容易受到其他领域的干扰，而谜题和逻辑领域则相对稳健。

实验设置。我们采用标准的组相对策略优化（Shao 等人，2024年；Liu 等人，2024年）。批次大小设为256，学习率设为 1×10^{-6}，模型训练 15 轮。对于每个问题，我们生成 8 个候选回答，最大输出长度限制为 16K 个词元。由于计算资源有限，所有实验均使用 Qwen3-4B-Base 模型进行（Yang 等人，2025）。

表3。 MATH500上的性能比较。我们报告第1步（起始）的准确率、最终步骤（末尾）的准确率、绝对提升（Δ，以百分点计）以及准确率排名。

训练领域	开始（%）	最后（%）	Δ	Rank
Math	40.6	66.3	+25.7	1
科学	41.8	62.6	+20.8	2
逻辑	36.4	60.6	+24.2	3
谜题	40.2	58.4	+18.2	4

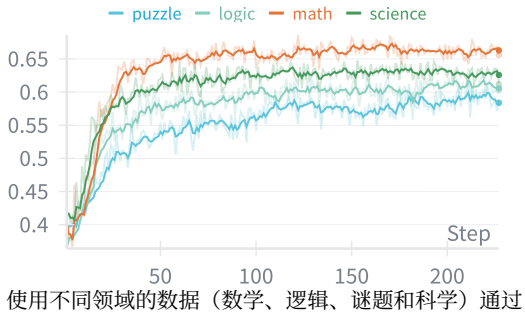


图3。 使用不同领域的数据（数学、逻辑、谜题和科学）通过 GRPO 训练通义千问3- 4B-Base 时，其在 MATH500 上每个训练步骤的准确率。所有领域的训练都能迅速激发模型的数学推理能力。

2.1. 数据集构建

我们首先介绍本研究考察的四个领域，即数学、科学、逻辑和谜题。 表2总结了各领域的数据来源及相应的数据规模。

数学。 数学数据取自 Skywork/Skywork-OR1-RL-Data 数据集（He 等人，2025a；b），该数据集包含可验证、具有挑战性且多样化的数学问题（105K）以及编程问题（14K）。我们根据 DeepSeek-R1-Distill-Qwen-7B 标注的难度级别，选取了 5K 个数学问题，以确保所选问题与 Qwen3-4B-Base 的能力相匹配。

科学。Nvidia/OpenScienceReasoning-2（Nathawani等人，2025；Bercovich等人，2025）是科学的数据来源，其中同时包含选择题和开放式问答对，并附有详细的推理轨迹，涵盖广泛领域，包括科学、技术、工程和数学、法律、经济学和人文学科。在本工作中，我们仅保留科学、技术、工程和数学领域的选择题，每题只有一个正确答案，以确保监督的一致性和统一性。

逻辑。 该数据集取自K-and-K/knights-and-knaves数据集（Xie等人，2024年），这是一个用于评估大语言模型逻辑推理能力的基准测试。我们总共采样了5K个实例，并使用原始测试集划分中的700个示例作为评估集。

谜题。对于谜题领域，我们使用 Enigmata-Data 项目生成的数据（Chen等人，2025），其中涵盖了 7 类逻辑推理谜题，共有 36 种不同的任务类型。我们构建了总计 5K 个样本的训练集，涵盖简单、中等和困难三个难度级别。

我们分别使用 MATH500（Lightman 等人，2023）和 GPQA（Rein 等人，2024年）评估模型在数学和科学领域的性能。对于逻辑和谜题领域，则从各自的原始测试集中选取子集进行评估。

2.2. 数学推理在不同领域之间具有很强的迁移性

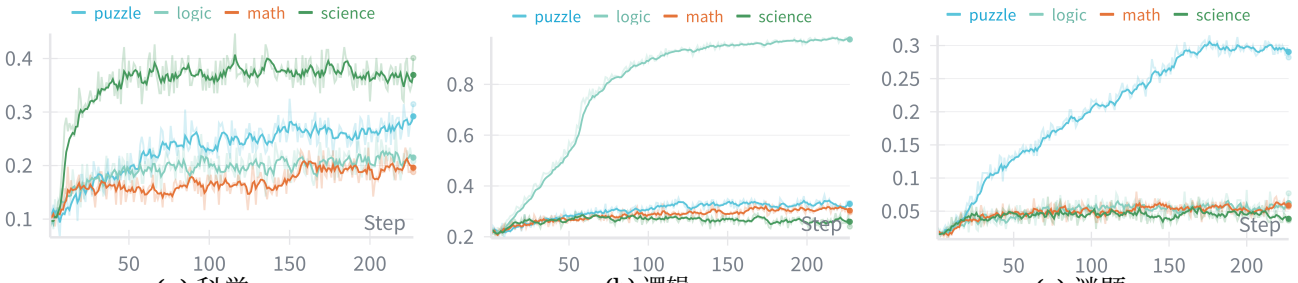
图3 展示了使用四个不同领域的数据训练的模型，在各训练步骤中于MATH500上的准确率（数据记录于 Weights & Biases）。随附的 表3汇总了训练初期（第1步）和最终训练步骤的准确率，以及四个训练领域在MATH500上的最终排名。

无论训练领域为何，Qwen3-4B-Base模型经过训练后在MATH500上的表现都持续提升，准确率绝对增幅约为20个百分点。这表明，来自所有四个领域的数据都能有效增强模型的数学推理能力。其中，使用谜题数据训练带来的提升最小，而使用数学数据训练则取得了最佳的最终性能。

我们推测，这一现象源于通义千问3-4B-Base在预训练阶段接触了大量数学数据。因此，在随后基于GRPO的强化学习阶段，使用来自不同领域的数据进行训练，仍然能够对模型的数学能力产生有益的迁移。

表4. 不同训练数据集上GPQA、逻辑测试和谜题测试的性能比较。对于每个训练领域，我们报告步骤1（开始）的准确率、最终步骤（最后）的准确率、绝对提升（Δ，以百分点计）以及准确率排名。

训练领域	GPQA				谜题测试				逻辑测试			
	开始	Last	Δ	Rank	开始	Last	Δ	Rank	开始	Last	Δ	Rank
Math	10.66	19.61	+8.95	4	2.05	5.85	+3.80	3	21.29	30.30	+9.01	3
科学	9.64	36.92	+27.28	1	1.54	3.85	+2.31	4	22.29	25.92	+3.63	4
逻辑	8.12	21.50	+13.38	3	2.82	6.21	+3.39	2	21.14	97.70	+76.56	1
谜题	10.66	29.20	+18.54	2	2.31	29.04	+26.73	1	20.14	32.90	+12.76	2



(a) 科学　(b) 逻辑　(c) 谜题

图 4. 在使用不同领域的数据进行 GRPO 训练期间，Qwen3-4B-Base 在 GPQA、逻辑测试和谜题测试上的分阶段准确率。(a) 科学领域（GPQA）的结果表明，通过在其他领域进行训练可以部分激活科学推理能力，但其激活难度高于数学。(b) 和 (c) 分别展示了逻辑领域和谜题领域的结果。模型的逻辑和谜题推理能力基本不受其他领域训练的影响，主要由领域内数据激活。

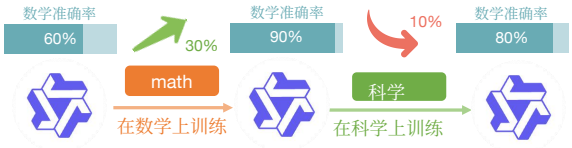


图 5. 跨域交互的训练流程。我们首先在领域1（数学）上训练模型，得到 M_1；随后在领域2（科学）上继续训练，得到 M_2；最后评估领域1上的性能，以量化遗忘/迁移：$\Delta_{\mathrm{F1}\leftarrow\mathrm{F2}} = \text{准确率}_{\mathrm{F1}}(M_2) - \text{准确率}_{\mathrm{F1}}(M_1)$。

2.3. 科学推理具有中等可迁移性，且对领域敏感

与MATH500上的结果类似，通义千问3-4B-Base模型在使用来自四个领域的全部数据进行训练后，在GPQA上的性能也有所提升，这表明跨领域训练同样能够增强科学推理能力。不过，GPQA上的性能提升幅度不如MATH500上观察到的结果稳定，且总体上更小。

如表4和图 4所示，不同训练领域带来的科学性能提升差异显著。例如，使用科学数据进行训练可使 GPQA 的准确率提高约 27 个百分点，而使用数学数据训练带来的提升则小得多，约为 9 个百分点。这种差异凸显了科学推理性能对训练数据选择的高度敏感性。

我们推测，模型的预训练语料库中已经包含一定数量的科学相关数据，这使得在不同领域进行 GRPO训练能够在一定程度上迁移到科学推理。然而，提升幅度相对较小且不够稳定，这表明预训练期间科学数据的数量和覆盖范围可能不及数学数据。

2.4. 谜题和逻辑主要具有领域特异性

与数学和科学领域的观察结果不同，谜题和逻辑领域呈现出截然不同的模式：逻辑和谜题测试上的跨领域迁移效应明显较弱，表4 和图 4均显示了这一点。

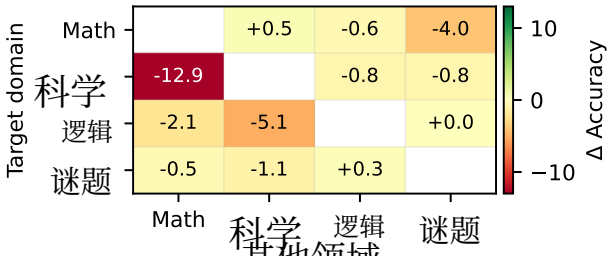
在逻辑测试中，只有使用逻辑领域数据进行训练，模型的准确率才能达到约 97% 的高水平；使用谜题数据训练可带来约 30% 的适度提升，而使用数学或科学数据训练的提升则小得多，其他领域的准确率在最后一步基本停滞于约 30%。

谜题测试也呈现出类似趋势：除使用谜题领域数据训练可达到约 30% 的准确率外，使用其他领域训练对谜题测试几乎没有提升。

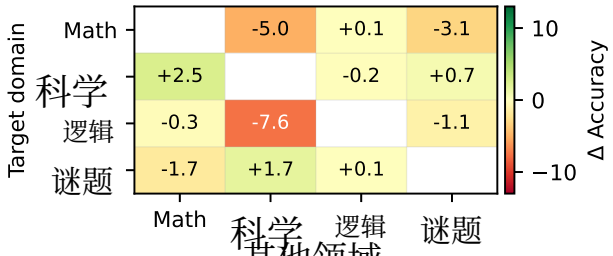
我们假设，这些领域在模型的预训练阶段所占比例不足，因此需要通过领域特定强化学习来有效提升性能。

表5. 数学、科学、逻辑和谜题四个领域的双领域训练结果。对于每个目标领域，原始 表示仅在目标领域上训练的模型性能。继续训练领域 报告模型在目标领域训练后，继续在其他领域上训练所取得的性能（目标域→其他域）；而预训练领域 报告模型先在其他领域上训练，再在目标领域上训练时的性能（其他域→目标域）。我们报告准确率（%）以及性能变化 $\Delta =$ 准确率$_{\text{current}}$ −准确率$_{\text{original}}$ （以百分点计）。

目标领域			领域 →继续训练领域			预训练领域 → 领域		
Math	指标	原始	科学	逻辑	谜题	科学	逻辑	谜题
	准确率（%）	82.8	83.3	82.2	78.8	77.8	82.9	79.7
	变化量（%）	–	+0.5	-0.6	-4.0	-5.0	+0.1	-3.1
科学	指标	原始	数学	逻辑	谜题	数学	逻辑	谜题
	准确率（%）	38.3	25.4	37.5	37.5	40.8	38.1	39.0
	变化量（%）	–	-12.9	-0.8	-0.8	+2.5	-0.2	+0.7
逻辑	指标	原始	数学	科学	谜题	数学	科学	谜题
	准确率 (%)	88.1	86.0	83.0	88.1	87.8	80.5	87.0
	变化量 (%)	–	-2.1	-5.1	0.0	-0.3	-7.6	-1.1
谜题	指标	原始	数学	科学	逻辑	数学	科学	逻辑
	准确率（%）	20.2	19.7	19.1	20.5	18.5	21.9	20.3
	变化量（%）	–	-0.5	-1.1	+0.3	-1.7	+1.7	+0.1



(a) 训练顺序：目标域→其他域



(b)训练顺序：其他域→目标域

图 6. 非对称跨领域迁移与干扰效应。每个单元格报告 $\Delta =$ 准确率$_{\text{current}}$ − 准确率$_{\text{original}}$ （单位为百分点）。例如，在 (a) 中， −12.9 表示模型先在科学领域上训练，再在数学上进一步训练后，相较于仅在科学上训练，科学领域的性能下降。在 (b) 中， +2.5 表示模型先在数学上训练，再在科学领域上训练后，相较于仅科学基线，科学领域的性能提升。

3. 跨域交互

上一节中的现象引出了一个自然的问题：在模型获得领域 A 的推理能力之前或之后，使用来自领域 B 的数据进行训练，是否会影响领域 A 推理能力的习得与保留?

实验设置。为了研究在其他领域训练前后，各领域的遗忘或保持情况，我们设计了如图 5所示的流程。具体而言，首先在领域 1（例如数学）上训练模型，得到模型 1；随后在领域 2 上继续训练，得到模型 2。通过比较模型 1 和模型 2 在领域 1 测试集上的准确率，我们量化领域 2 训练对模型此前在领域 1 中获得的推理能力的影响。反之，通过比较模型 1 和模型 2 在领域 2 测试集上的准确率，我们衡量在领域 1 上训练对模型后续在领域 2 中的性能有何影响。四个领域的详细结果见表 5。

3.1. 双领域训练的结果

数学方面的结果。随后进行科学或逻辑训练，基本能维持数学表现；而谜题训练则会导致明显退化。反之，先前的科学或谜题训练会阻碍数学学习，而先前的逻辑训练影响很小。

科学方面的结果。与数学领域相比，科学领域表现出更强的敏感性。当模型先接受科学训练、再接受数学训练时，其科学领域性能会显著下降。相比之下，先进行数学训练再进行科学训练，则能促进科学推理。

逻辑领域的结果。逻辑领域表现出与科学领域的强烈相互作

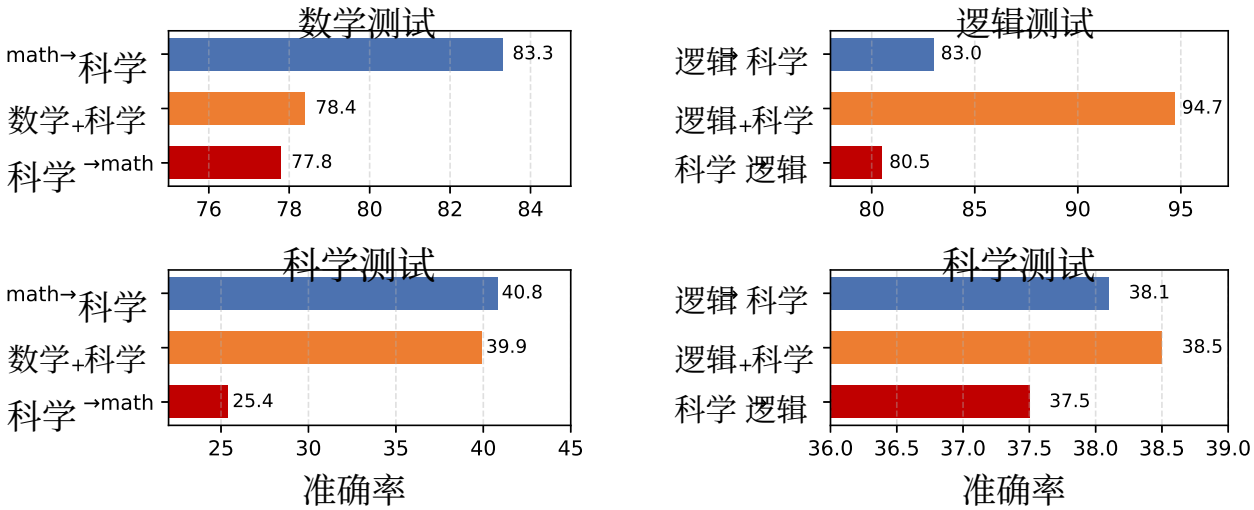


图7。 两个领域间顺序训练与混合训练的比较。左图比较数学和科学，右图比较逻辑和科学。其中，math→science 表示先在数学上训练，再在科学上训练；math+science 表示在两个领域上进行混合训练。对于数学和科学，训练顺序至关重要，其中 math→science 取得了最佳性能。相比之下，对于逻辑和科学，混合训练带来了最强的整体性能。

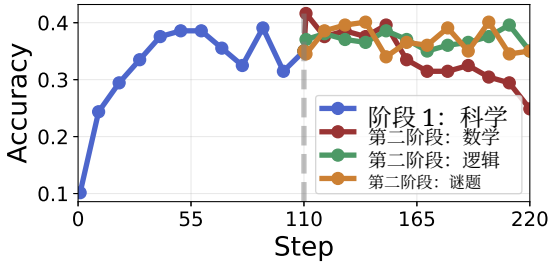


图8。 两个领域训练期间 GPQA 上的分阶段准确率。在第一阶段，模型使用科学数据进行训练。在第二阶段，训练继续使用数学、逻辑或谜题数据。当第二阶段使用数学数据时，GPQA 准确率显著下降，表明数学训练会干扰模型的科学推理能力。

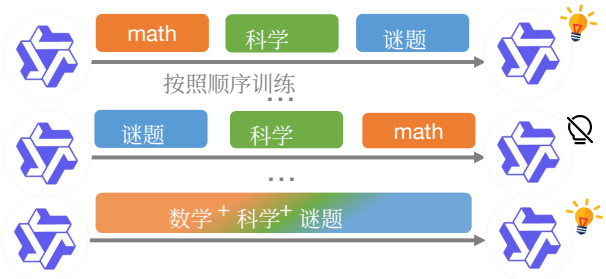


图9. 三个领域的顺序训练与混合训练流程。上：分阶段（顺序）训练，模型先在数学数据上训练，再在科学数据上训练，最后在谜题数据上训练。下：混合训练，将数学/科学/逻辑数据合并，并在单次训练中联合使用。

用。无论科学训练是在逻辑训练之后还是之前进行，科学训练都会持续干扰模型的逻辑推理能力：在科学训练之后进行时，逻辑推理能力下降 5.1%；在科学训练之前进行时，则下降 7.6%。相比之下，在其他领域进行训练对逻辑推理性能的影响相对较小。

谜题结果。相比之下，谜题领域对跨领域训练基本不敏感，其他领域施加的影响极小，既未导致明显的性能下降，也未带来显著的性能提升。

3.2。数学与科学的训练顺序至关重要 nce

我们观察到，数学领域和科学领域表现出特别强的双向相互作用，并且对训练顺序高度敏感。接下来，我们重点关注对它们之间关系的详细分析。

我们首先考察科学训练后数值步处的 Weights & Biases 曲线，以及随后在其他领域上的持续训练，如图8所示。在训练过程中，在逻辑领域和谜题领域上的持续训练在很大程度上保留了模型的科学推理能力，而在数学领域上的训练则导致其逐渐退化。

此外，我们还对数学和科学数据进行了混合训练实验。如图7所示，先进行数学训练再进行科学训练，对两个领域都相对有利。相比之下，先进行科学训练再进行数学训练会同时导致数学和科学性能下降。值得注意的是，数学和科学的混合训练能够有效缓解训练顺序带来的负面影响，使两个领域的性能更加稳定。

表6。 数学、科学和谜题领域中顺序训练与混合训练策略的性能比较。我们报告三个测试领域上的最终准确率（%），以及相应的排名（1 = 最佳）。平均值表示数学、科学和谜题领域的平均准确率。顺序训练记作 A→B→C，表示模型先在 A 上训练，再在 B 上训练，最后在 C 上训练。单一领域指仅在相应领域上训练的模型（例如，仅数学、仅科学或仅谜题），并在同一领域上进行评估。

训练策略	Math		科学		谜题		Avg	
	Acc	Rank	Acc	Rank	Acc	Rank	Acc	Rank
单一领域	82.76	–	38.28	–	20.15	–	47.06	–
数学 →科学 →谜题	80.48	3	36.97	4	17.06	7	44.84	4
数学 →谜题 →科学	81.76	2	39.39	2	18.94	5	46.70	3
科学 → 数学 → 谜题	77.68	7	26.26	7	20.20	2	41.38	7
科学 → 谜题 →数学	78.04	6	30.20	6	20.10	3	42.78	6
谜题 → 数学 → 科学	**82.80**	1	39.29	3	18.99	4	47.03	2
谜题 →科学 →数学	79.72	5	32.42	5	17.72	6	43.29	5
混合（数学 +科学 + 谜题）	79.92	4	**41.62**	1	**21.42**	1	**47.65**	1

表7. 顺序与混合训练策略在数学、科学和逻辑上的性能比较。

训练策略	Math		科学		逻辑		Avg	
	Acc	Rank	Acc	Rank	Acc	Rank	Acc	Rank
单一领域	82.76	–	38.28	–	88.14	–	69.73	–
逻辑 →数学 →科学	**84.00**	1	36.67	3	87.31	2	69.99	2
逻辑 → 科学 →数学	76.96	7	29.29	6	81.71	4	62.65	6
数学 →逻辑 →科学	82.56	3	35.86	4	83.00	3	67.14	3
数学 → 科学 →逻辑	82.80	2	37.98	2	75.91	5	65.56	4
科学 → 逻辑 → 数学	77.96	6	30.40	5	76.20	6	61.52	5
科学 → 数学 → 逻辑	78.24	5	26.97	7	63.80	7	56.34	7
混合（数学 +科学 + 逻辑）	80.44	4	**40.00**	1	**95.37**	1	**71.94**	1

3.3. 科学干扰逻辑推理

如表5所示，在逻辑领域中，无论科学训练是在逻辑训练之前还是之后进行，模型在接受科学训练后性能都会持续下降。在对逻辑和科学进行顺序训练后，模型的逻辑准确率明显下降，这表明这两个领域可能相互抑制：科学数据训练会损害通过逻辑训练获得的逻辑推理能力，导致逻辑领域出现能力遗忘。

我们进一步探讨，混合逻辑和科学数据能否缓解这种跨领域干扰。如图7所示，我们比较了三种训练策略（科学→逻辑、混合科学+逻辑以及逻辑→科学），并在逻辑和科学测试集上评估其性能。科学→逻辑策略在两个领域上的准确率均最低，而混合训练在逻辑和科学领域上都取得了最佳性能。这些结果表明，在逻辑和科学数据上进行联合训练，可以有效缓解跨领域干扰，并减少逻辑和科学领域的能力遗忘。

4. 优化多领域训练性能

本节探讨不同的多领域训练策略如何影响推理强化学习。为此，我们考虑两种具有代表性的领域设置：（1）数学、科学和谜题；（2）数学、科学和逻辑。对于三领域设置，我们考察两种训练范式：顺序训练和混合训练。顺序训练有六种可能的领域顺序，加上混合训练，共形成七种不同的训练配置以供比较。

如图9所示，我们首先使用这三个领域所有可能的顺序训练组合来训练模型，然后再使用混合领域数据训练一个模型。接着，我们在三个目标领域上评估所有训练好的模型，并比较其准确率，以评估不同训练策略的相对优缺点。

4.1. 顺序训练中的训练顺序至关重要

如表6 和表7所示，聚焦于顺序训练时，不同训练顺序之间的性能差异十分显著。

例如，逻辑→数学→科学这一顺序取得了最佳整体性能：数学准确率达到 84%，科学准确率为 36%，逻辑准确率为 87%，平均准确率约为 70%。相比之下，科学→数学→逻辑这一顺序导致三个领域的性能均严重下降，平均准确率最低，约为 56%。

在所有顺序训练策略中，我们观察到，采用数学→科学顺序的模型始终优于采用相反的科学→数学顺序的模型。这一规律在数学–科学–谜题和数学–科学–逻辑两种设置下均成立，表明从数学到科学的训练进程构成了多领域 GRPO 中一种关键且有利的训练顺序。

4.2. 对某些领域而言，混合训练更为可取

我们进一步比较了混合训练与表现最佳的顺序训练策略，并观察到明显的领域偏好。在数学领域，顺序训练更为有效，甚至可能优于仅在数学领域训练，这表明按精心安排的顺序接触辅助领域能够增强数学推理能力。相比之下，在科学、逻辑和谜题领域，混合训练始终能取得更好的结果，使性能提升超越单一领域训练所能达到的水平。这些发现表明，不同领域受益于根本不同的训练范式，统一的多领域策略可能并非最优。

5. 相关工作

大语言模型推理能力。随着大语言模型的快速发展，推理能力（Zhang 等人，2023；Yao 等人，2023；Plaat 等人，2024）已成为一项核心能力，也是实现通用人工智能（AGI）的广泛认可的先决条件（Minaee 等人，2024；Xu 等人，2024；Feng 等人，2024；Krishnan，2025）。强大长链推理能力的早期证据可追溯至 OpenAI o1（Jaech 等人，2024；Arrieta 等人，2025；Hurst 等人，2024），该模型在数学推理基准测试中表现出色。随后出现的 QwQ 等模型（Yang 等人，2024a；Bai 等人，2023a;b；Chu 等人，2024）通过过程奖励建模进一步提升了推理能力（Li & Li，2024；Ma 等人，2023；Zhang 等人，2025；Lambert 等人，2024）。近期，DeepSeek R1（DeepSeek-AI 等人，2025）和 Kimi 1.5（Team 等人，2025）显著增强了开源推理模型的能力：DeepSeek R1 借助简单的基于规则的奖励机制（Ramesh 等人，2024；Hu，2025；Shao 等人，2024；Alonso 等人，2025）缩小了与闭源系统之间的差距，而 Kimi 1.5 则采用长到短推理等技术来提升效率（Kirk 等人，2023；Yang 等人，2024b）。与此同时，包括 Sky-Thought T1（Team，2025）、Bespoke-Stratos（Labs，2025）、s1（Muennighoff 等人，2025）和 LIMO（Ye 等人，2025）凸显了高质量、精心构建的推理数据所发挥的关键作用，并表明只需规模相对较小但设计精良的训练集，也能实现强大的推理性能。

RLVR 与 GRPO。随着 DeepSeek-R1 的出现，组相对策略优化（GRPO）（Guo 等人，2025年；Liu 等人，2024年）已成为提升语言模型推理能力的广泛采用方法。越来越多的研究致力于改进 GRPO 本身，包括 DAPO（Yu等人，2025）、Dr. GRPO（Liu 等人，2025）和 GSPO（Zheng等人，2025）等变体。其他研究则探讨 GRPO 训练中的熵崩塌。例如，NGRPO（Nan 等人，2025）分析了错误样例如何影响训练过程中的熵，而《重新思考 RLVR 中的熵干预》（Hao 等人，2025）则从强化学习的角度研究不同令牌级行为对熵动态的影响。与此同时，还有一些研究旨在降低 GRPO 的训练成本（Xu 等人，2025a；b）。It Takes Two（Wu 等人，2025）生成配对样例，并从对比学习的视角阐释 GRPO。

面向其他领域的 GRPO。除数学外，代码是应用 GRPO 最普遍的领域之一，Code-R1（Liu &Zhang，2025）、DeepCoder（Luoet al.，2025）和 SkyWork OR1（He et al.，2025b）等模型便是例证。这些研究中有许多通过联合训练，使模型同时获得数学和代码推理能力。此外，越来越多的研究开始探索将 GRPO 应用于其他推理领域。例如，Enigmata（Chen等人，2025）专注于提升谜题求解能力，而 Logic-RL（Xie等人，2025）则以逻辑推理为目标。近期，GRPO 风格的训练也被用于增强模型在各类任务中的工具使用能力和智能体能力（Wei et al.，2025）。

6. 结论

我们发现，在多领域推理训练中，GRPO 会表现出显著、非对称且对顺序敏感的跨领域效应。结果表明，不同领域偏好不同的训练策略，而且没有一种范式能在所有推理领域中普遍占优。尤其是，领域的选择及其训练顺序都可能显著改变学习动态和最终表现。从更广泛的角度看，本研究强调，要构建稳健且具有泛化能力的基于强化学习的推理模型，理解并主动管理跨领域交互至关重要，尤其是在推理强化学习日益从单领域优化转向多领域训练范式之际。我们希望本研究能为此奠定基础并提供实用指导。

参考文献

Alonso, N. I. 等。《组相对策略优化的数学原理：一种多智能体强化学习方法》。组相对策略优化的数学原理：一种多智能体强化学习方法（*2025* 年 *1* 月 *3* 日），2025。

Arrieta, A., Ugarte, M., Valle, P., Parejo, J. A., and Segura, S. o3-mini 与 deepseek-r1：哪个更安全? *arXiv* 预印本 *arXiv:2501.18438*，2025年。

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., 等。通义千问技术报告。 *arXiv* 预印本 *arXiv:2309.16609*，2023a. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-VL：具备多种能力的前沿大型视觉语言模型。 *arXiv* 预印本 *arXiv:2308.12966*，2023b. Bercovich, A., Levy, I., Golan, I., Dabbah, M., El-Yaniv, R., Puny, O., Galil, I., Moshe, Z., 等。Llama-Nemotron：高效推理模型，2025。网址 https://arxiv.org/abs/2505.00949。Chen, J., He, Q., Yuan, S., Chen, A., Cai, Z., Dai, W., Yu, H., Yu, Q., Li, X., Chen, J., 等。Enigmata：利用可验证合成谜题扩展大语言模型的逻辑推理能力。 *arXiv* 预印本 *arXiv:2505.19914*，2025。Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., 等。Qwen2-Audio 技术报告。 *arXiv* 预印本 *arXiv:2407.10759*，2024年。DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., and Others。DeepSeek-R1：通过强化学习激励大语言模型的推理能力，2025。网址 https://arxiv.org/abs/2501.12948。Feng, T., Jin, C., Liu, J., Zhu, K., Tu, H., Cheng, Z., Lin, G., and You, J. 我们距离 AGI 还有多远? *arXiv* 预印本 *arXiv:2405.10313*，2024年。Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., 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。Hao, Z., Wang, H., Liu, H., Luo, J., Yu, J., Dong, H., Lin, Q., Wang, C., and Chen, J. 重新思考 RLVR 中的熵干预：从熵变化的视角出发。 *arXiv* 预印本 *arXiv:2510.10150*，2025。He, J., Liu, J., Liu, C. Y., Yan, R., Wang, C., Cheng, P., Z hang, X., Zhang, F., Xu, J., Shen, W., Li, S., Zeng, L., Wei, T., Cheng, C., An, B., Liu, Y., and Zhou, Y. Skywork Open Reasoner 1 技术报告。 *arXiv* 预印本 *arXiv:2505.22312*，2025a。

He, J., Liu, J., Liu, C. Y., Yan, R., Wang, C., Cheng, P., Zhang, X., Zhang, F., Xu, J., Shen, W., Li, S., Zeng, L., Wei, T., Cheng, C., Liu, Y. 和 Zhou, Y. Skywork 开放推理器系列，2025b。Notion 博客。Hu, J. Reinforce++：一种用于对齐大语言模型的简单高效方法。*arXiv* 预印本 *arXiv:2501.03262*，2025。Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A. 等。GPT-4o 系统卡片。*arXiv* 预印本 *arXiv:2410.21276*，2024。Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Car- ney, A. 等。OpenAI o1 系统卡片。*arXiv* 预印本 *arXiv:2412.16720*，2024。Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Ham- bro, E., Grefenstette, E. 和 Raileanu, R. 理解基于人类反馈的强化学习对大语言模型泛化和多样性的影响。*arXiv* 预印本 *arXiv:2310.06452*，2023。Krishnan, V. 从人工智能（AI）到通用人工智能（AGI）——未来之路。世界正畸医师联合会期刊，14(1):1–2，2025。Labs, B. Bespoke-Stratos：推理蒸馏的惊人效能。 www.bespokelabs.ai/blog/bespoke-
s
t
ratos-the-unreasonable-effectiveness-of-reasoning-distillation，2025。访问日期：2025-01-22。Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y. 等。RewardBench：评估语言建模的奖励模型。*arXiv* 预印本 *arXiv:2403.13787*，2024。Li, W. 和 Li, Y. 使用 Q 值排名的过程奖励模型。*arXiv* 预印本 *arXiv:2410.11287*，2024。Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I. 和 Cobbe, K. 让我们逐步验证。*arXiv* 预印本 *arXiv:2305.20050*，2023。Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D. 等。DeepSeek-V2：一种强大、经济且高效的混合专家语言模型。*arXiv* 预印本 *arXiv:2405.04434*，2024。Liu, J. 和 Zhang, L. Code-R1：使用可靠奖励复现 R1 的代码版本。*arXiv* 预印本 *arXiv:2503.18470*，3，2025。Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S. 和 Lin, M. 理解 R1-Zero 类训练：

一种批判性视角。*arXiv* 预印本 *arXiv:2503.20783*，2025。

Luo, M., Tan, S., Huang, R., Patel, A., Ariyak, A., Wu, Q., Shi, X., Xin, R., Cai, C., Weber, M. 等。Deepcoder：一款完全开源、性能达到 o3-mini 水平的 14b 编码模型。*Notion* 博客，2025。

Ma, Q., Zhou, H., Liu, T., Yuan, J., Liu, P., You, Y. 和 Yang, H. 让奖励逐步推进：作为推理导航者的步骤级奖励模型。*arXiv* 预印本 *arXiv:2310.10080*，2023。

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. 和 Gao, J. 大语言模型：综述。*arXiv* 预印本 *arXiv:2402.06196*，2024年。

Muennighoff, N., Yang, Z., Shi, W., Li, X.L., Fei-Fei, L.,Hajishirzi, H., Zettlemoyer, L., Liang, P., Cand`es, E., and Hashimoto, T. s1：简单的测试时缩放，2025。网址 https://arxiv.org/abs/2501.19393。

Nan, G., Chen, S., Huang, J., Lu, M., Wang, D., Xie, C., Xiong, W., Zeng, X., Zhou, Q., Li, Y., 等。NGRPO：负增强组相对策略优化。*arXiv*预印本 *arXiv:2509.18851*，2025。

Nathawani, D., Gitman, I., Majumdar, S., Bakhturina, E., Mahabaleshwarkar, A. S., Zhang, J., and Scowcroft, J. P. Nemotron 后训练数据集 v1。https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v1，2025年7月。英伟达。

Plaat, A.、Wong, A.、Verberne, S.、Broekens, J.、van Stein, N. 和 Back, T.《使用大语言模型进行推理：一项综述》。*arXiv* 预印本 *arXiv:2407.11511*，2024年。

Ramesh, S. S.、Hu, Y.、Chaimalas, I.、Mehta, V.、Sessa, P.G.、Ammar, H. B. 和 Bogunovic, I.《无奖励人类反馈强化学习中的群体鲁棒偏好优化》。*arXiv* 预印本 *arXiv:2405.20304*，2024年。

Rein, D.、Hou, B. L.、Stickland, A. C.、Petty, J.、Pang, R. Y.、Dirani, J.、Michael, J. 和 Bowman, S. R.《GPQA：研究生级谷歌免疫问答基准》。载于第一届语言建模会议，2024年。

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y. 等。Deepseekmath：突破开放语言模型数学推理的极限。*arXiv* 预印本 *arXiv:2402.03300*，2024年。

Team, K., Du, A., Gao, B., Xing, B., Jiang, C., 等。Kimi k1.5：通过大语言模型扩展强化学习规模，2025。网址 https://arxiv.org/abs/2501.12599。

Team, N. Sky-t1：在 $450 以内训练自己的 o1 预览模型。https://novasky-ai.github.io/posts/sky-t1，2025。访问日期：2025-01-09。Wei, Y., Duchenne, O., Copet, J., Carbonneaux, Q., Zhang, L., Fried, D., Synnaeve, G., Singh, R., 和 Wang, S. I. Swe-rl：通过在开放软件演进领域进行强化学习，推进大语言模型推理。*arXiv* 预印本 *arXiv:2502.18449*，2025。Wu, Y., Ma, L., Ding, L., Li, M., Wang, X., Chen, K., Su, Z., Zhang, Z., Huang, C., Zhang, Y. 等。It Takes Two：你的 GRPO 暗中其实是 DPO。*arXiv* 预印本 *arXiv:2510.00977*，2025。

Xie, C., Huang, Y., Zhang, C., Yu, D., Chen, X., Lin, B. Y., Li, B., Ghazi, B. 和 Kumar, R. 关于大语言模型在逻辑推理中的记忆。https://arxiv.org/abs/2410.23123，2024年。

Xie, T., Gao, Z., Ren, Q., Luo, H., Hong, Y., Dai, B., Zhou, J., Qiu, K., Wu, Z. 和 Luo, C. Logic-rl：通过基于规则的强化学习释放大语言模型推理能力，2025年。网址：https://arxiv.org/abs/2502.14768。

Xu, M., Yin, W., Cai, D., Yi, R., Xu, D., Wang, Q., Wu, B., Zhao, Y., Yang, C., Wang, S., 等。资源高效的大语言模型和多模态基础模型综述。*arXiv* 预印本 *arXiv:2401.08092*，2024年。

Xu, X., AI, C., Yang, K., Chen, T., Wang, Y., Yang, S. 和 Yang, C. 无思考策略初始化使蒸馏推理模型成为更有效、更高效的推理器。*arXiv* 预印本 *arXiv:2509.26226*，2025a。

Xu, Y., Dong, H., Wang, L., Sahoo, D., Li, J. 和 Xiong, C. 通过弹性推理实现可扩展的思维链。*arXiv* 预印本 *arXiv:2505.05315*，2025b。

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z. 和 Qiu, Z. 通义千问2.5 技术报告。*arXiv* 预印本 *arXiv:2412.15115*，2024a。

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., 等。通义千问3技术报告。*arXiv* 预印本 *arXiv:2505.09388*，2025。

Yang, A. X.、Robeyns, M.、Coste, T.、Shi, Z.、Wang, J.、Bou-Ammar, H. 和 Aitchison, L. 面向大语言模型对齐的贝叶斯奖励模型。*arXiv* 预印本 *arXiv:2402.13210*，2024b。

Yao, S.、Yu, D.、Zhao, J.、Shafran, I.、Griffiths, T.、Cao, Y. 和 Narasimhan, K. 思维树：借助大语言模型进行审慎问题求解。神经信息处理系统进展，36:11809–11822，2023。Ye, Y.、Huang, Z.、Xiao, Y.、Chern, E.、Xia, S. 和 Liu, P. LIMO：推理中少即是多，2025年。网址：https://arxiv.org/abs/2502.03387。Yu, Q.、Zhang, Z.、Zhu, R.、Yuan, Y.、Zuo, X.、Yue, Y.、Fan, T.、Liu, G.、Liu, L.、Liu, X. 等。DAPO：大规模开源大语言模型强化学习系统。*arXiv* 预印本 *arXiv:2503.14476*，2025年。Zhang, Z.、Zhang, A.、Li, M.、Zhao, H.、Karypis, G. 和 Smola, A. 语言模型中的多模态思维链推理。*arXiv* 预印本 *arXiv:2302.00923*，2023年。Zhang, Z.、Zheng, C.、Wu, Y.、Zhang, B.、Lin, R.、Yu, B.、Liu, D.、Zhou, J. 和 Lin, J. 数学推理中开发过程奖励模型的经验教训。*arXiv* 预印本 *arXiv:2501.07301*，2025年。Zheng, C.、Liu, S.、Li, M.、Chen, X.-H.、Yu, B.、Gao, C.、Dang, K.、Liu, Y.、Men, R.、Yang, A. 等。群组序列策略优化。*arXiv* 预印本 *arXiv:2507.18071*，2025年。

A. 局限性

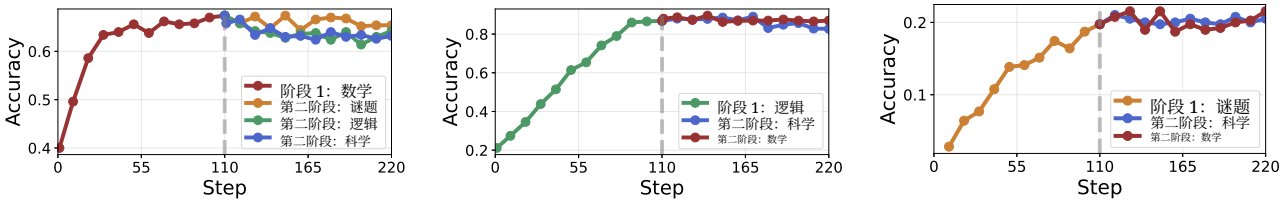
尽管本文发现了推理强化学习中显著的非对称且顺序敏感的跨领域效应，但并未对这些效应产生的原因提供深入的机制性解释。具体而言，我们尚未明确区分这些行为究竟源于预训练数据覆盖范围、领域特定知识重叠，还是 GRPO训练期间的优化动态。理解跨领域促进和干扰的确切起源——尤其是它们与预训练分布及表示共享之间的联系——仍是未来研究的重要方向。

B. 影响声明

本文研究了多领域推理设置下组相对策略优化（GRPO）的行为，旨在加深对大语言模型中基于强化学习的训练策略的理解，并改进其设计。本文的主要贡献在于分析和方法论，重点关注训练动态、领域交互和性能权衡，而非将模型部署于现实世界的应用中。

因此，我们预计这项工作不会直接带来即时的负面社会或伦理后果。相反，通过强调领域感知和顺序感知训练的重要性，我们的发现可能有助于未来研究开发更加稳健、可解释且可靠的推理模型。这项工作的任何更广泛的社会影响，都将取决于大语言模型的下游应用，而这些应用超出了本研究的范围。

C. 其他双领域训练的测试性能



(a) MATH500 (b) 逻辑 (c) 谜题

图*10.* **科学、逻辑和谜题上的双领域训练结果。** 不同双领域训练策略下，准确率随训练步骤变化的曲线。

这里展示了 Qwen3-4B-Base 模型在三个领域进行双领域训练时的分阶段准确率，所有曲线均直接取自 Weights & Biases 日志。

D. 动机：单域与多域 GRPO

给定提示 x，组相对策略优化（GRPO）采样一组 K 个响应 $\{y_i\}_{i=1}^{K} \sim \pi_\theta(\cdot \mid x)$，并分配标量奖励 $r_i = R(x, y_i)$。组相对优势通过对组内奖励进行归一化来计算：

$$A_i = \frac{r_i - \mu}{\sigma + \epsilon}, \qquad \mu = \frac{1}{K}\sum_{j=1}^{K} r_j, \tag{1}$$

其中 σ 表示组内奖励的标准差。策略通过优化 PPO 式裁剪目标进行更新。在其标准表述中，GRPO 隐式假设一种 单一领域 设定，其中奖励来自同质任务分布。在此假设下，组相对归一化应用于来自同一推理领域的响应，由此产生的策略更新优化单一领域特定推理目标。

多领域 GRPO。 在实践中，推理强化学习通常使用来自多个领域的数据进行，例如数学、科学、逻辑和谜题。令 $\mathcal{D}$ 表示领域集合，$p(d)$ 表示领域混合分布。在这种设置下进行训练，优化的目标为

$$\mathcal{L}(\theta) = \sum_{d \in \mathcal{D}} p(d)\, \mathbb{E}_{x \sim \mathcal{X}_d}\big[\mathcal{L}_{\mathrm{GRPO}}(\theta; x)\big], \tag{2}$$

其中， $\mathcal{L}_{\mathrm{GRPO}}(\theta; x)$ 遵循单领域 GRPO 目标。与单一领域情形不同，多领域设置下的奖励源自异质的领域分布，并且可能编码不同的推理目标。因此，策略更新会在同一优化过程中聚合领域特定优势信号。

这引出了一个根本性问题：在强化学习过程中，不同推理领域的奖励是会相互促进推理能力的习得，还是会干扰先前学到的能力？在这项工作中，我们通过分析多领域环境下的跨领域迁移与干扰，对这一问题进行了实证研究。

E. 实验训练详情

```
# 关键训练配置（GRPO / RLVR）
run:
  项目：MIX-REWARD-GRPO-SMALL-Normal-Long
  实验：vanilla-Qwen3-4B-Base-msl
  随机种子：0
  总计_训练轮次：6

模型：
  基础模型：Qwen/Qwen3-4B-Base
  最大_提示_长度：1024
  最大_响应_长度：16384
  移除_填充：true
  梯度_检查点：true

算法：
  名称：GRPO
  loss_模式：原始模式
  损失_聚合：令牌均值
  n_响应_每_提示：8
  裁剪_比率_低：3.0e-4
  裁剪_比率_高：4.0e-4
  熵_系数：0.0
  梯度_裁剪：1.0
  kl_in_奖励：false
  kl_系数：0.0
  kl_损失：false
  kl_loss_系数：0.0

优化：
  学习率：1.0e-6
  学习率_预热_比例：0.05
  权重_衰减：0.1
  训练 _批次 _大小：256
  mini_批次 _大小：64
  微批次_大小_每_GPU：16
  动态_批次大小：true

生成：
  引擎：vLLM
  模式：同步
  GPU_内存_利用率：0.6
  张量_并行：1
  分块_预填充：true
  max_num_批处理 _词元：17408

计算：节点数：1 GPU_每_节点：8
Ray_数量_CPU：64 序列_并行：1 参
数_卸载：false 优化器_卸载：false 熵
_检查点：true
```