GeometryZero：使用群体对比政策优化提高 LLM 的几何解题能力

Yikun Wang[1,2*] **Yibin Wang**[1,2] **Dianyi Wang**[1,2] **Zimian Peng**[2,3]
Qipeng Guo[2,4] **Dacheng Tao**[5] **Jiaqi Wang**[2,4†]

[1]Fudan University [2]Shanghai Innovation Institute [3]Zhejiang University
[4]Shanghai AI Laboratory [5]Nanyang Technological University

Abstract

近年来，大型语言模型（LLMs）的进展在多个领域展现了非凡的能力，尤其是在数学推理方面，其中几何问题解决仍然是一个具有挑战性的领域，辅助构建在其中发挥着至关重要的作用。现有的方法要么表现不佳，要么依赖庞大的 LLMs（例如，GPT-4o），导致巨大的计算成本。我们认为，具有可验证奖励的强化学习（例如，GRPO）为训练较小的模型提供了一个有前景的方向，这些模型可以有效结合辅助构建与稳健的几何推理。然而，直接应用 GRPO 于几何推理存在基本限制，因为它依赖于无条件奖励，从而导致不加区别且适得其反的辅助构建。为了解决这些挑战，我们提出了组对比策略优化（GCPO），一种新颖的强化学习框架，具有两项关键创新：(1) 组对比遮蔽，根据上下文效用自适应地为辅助构建提供正面或负面奖励信号，以及 (2) 长度奖励，以鼓励更长的推理链。在 GCPO 的基础上，我们开发了 GeometryZero，一系列经济规模的几何推理模型，能够审慎地决定何时采用辅助构建。我们在流行的几何基准（Geometry3K，MathVista）进行广泛的实证评估，证明 GeometryZero 模型一致优于基线（例如 GRPO），在所有基准上平均提高 4.29 %。

1 介绍

> “ We shape our tools and thereafter our tools shape us. ”
>
> — Marshall McLuhan

近期在大型语言模型（LLMs）方面的进展在包含数学在内的多个领域表现出色。其中，几何问题求解被认为是一项具有挑战性的任务，它需要对视觉背景（即几何图形）的感知和复杂的推理。现有的训练方法要么利用大量带注释的数据进行监督学习，要么专注于代数级别的形式偏差。这使得当前模型在该领域表现不尽如人意，并且由于依赖于注释质量，导致其推理链中缺乏自我校正能力。

另一系列研究侧重于辅助线，当图形本身复杂或者问题的内在属性受益于这样的构建时，辅助线是非常有价值的，这显著降低了解题难度 (**?**) 。包括 **??** 在内的多项研究尝试通过修改形式语言（例如代码）用于辅助构建，以此来增强视觉语言模型对上下文的利用，从而提高它们在复杂几何问题上的推理能力。现有研究验证了将视觉上下文转换为形式语言并利用 LLM 能够获得更好的推理能力 (**?**) 。AlphaGeometry2 (**?**) 也使用 LLMs 进行辅助构建。然而，这些方法依赖于提示或训练巨大的模型（例如，Gemini, GPT-4o），这些都需要昂贵的计算成本，限制了它们在现实世界中的部署。

[*]Correspond to <yikunwang19@fudan.edu.cn>
[†]Corresponding Author, <wjqdev@gmail.com>

Preprint. Under review.

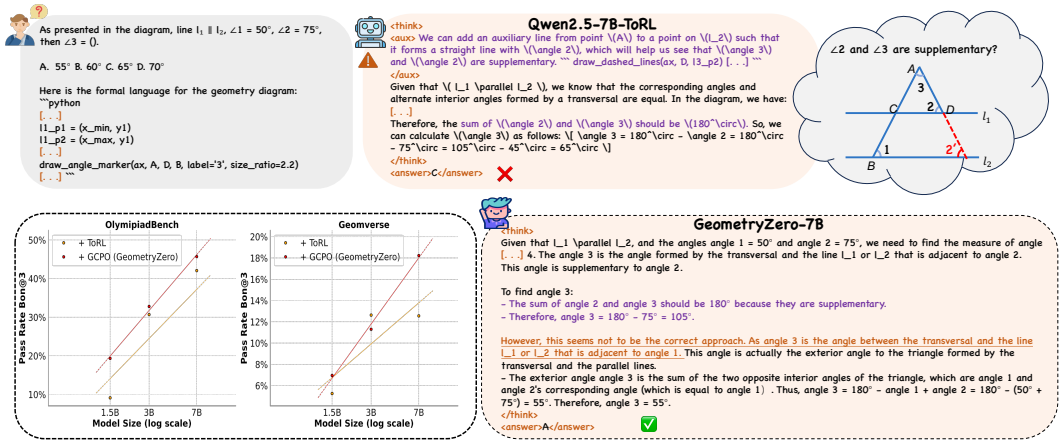

Figure 1: 在 ToRL 和我们的 GCPO 之间的比较研究。(a) 两个案例比较 GeometryZero-7B 与 Qwen2.5-7B-GRPO，揭示了 GeometryZero-7B 谨慎地直接进行推理，而一款 ToRL 训练的模型则不加选择地进行辅助构建。(b) 紫色 文本强调了两个模型在推理过程中的错误步骤。orange underlined 文本在推理过程中的显示了关键反思步骤，我们将其识别为模型在几何问题解决中的“顿悟时刻”(?)，从中 GeometryZero-7B 在几何问题解决情境中受益。(c) GeometryZero 展示了在不同模型规模中较之 ToRL 更好的整体表现和更好的扩展效果。

在 Deepseek-R1-Zero (?) 取得成功之后，GRPO 成为一种在推理任务和工具学习中普遍适用且有效的范式 (????)。这使得它特别适合于训练能够进行辅助构建的中等规模模型，同时实现强大的几何推理性能。然而，将 GRPO 框架直接应用于具有辅助构建的几何推理面临挑战：在某些情况下，强制或冗余的辅助构建是没有必要的，甚至可能是有害的。具体来说，有些问题可以通过直接推理来解决而不需要辅助线，强制它们的加入实际上可能会导致错误的解决方案。目前的 RL 工具使用方法通常依赖于无条件奖励（在所有实例中始终积极的信号）来鼓励不加区分的工具调用 (?)。这种方法缺乏灵活性，无法自主判断辅助构建何时适宜，从而限制了 RL 在几何问题解决中的效果。

我们假设需要一个灵活的机制，允许模型通过强化学习来学习何时使用辅助构造以及何时放弃。为此，我们提出了一种新的强化学习方法，称为组对比策略优化（GCPO），该方法避免了无条件奖励的缺点。具体来说，GCPO 与传统的 GRPO 有显著不同：它通过两组对比的行为预测量辅助构造的好处，然后通过组对比掩码提供灵活的信号（条件奖励），包括鼓励或惩罚。这一机制使得 GCPO 能够在明显有利的情境中灵活地鼓励辅助构造，同时在明显不利的情况下惩罚它。受到 LCPO (?) 的启发，我们的工作引入了一个长度奖励，以鼓励多维和更深入的推理。

在 GCPO 的基础上，我们开发了 GeometryZero 模型，这是专门用于几何推理的一系列轻量级（从 1.5 B 到 7 B）的 LLMs。大量实验表明，GeometryZero 在多个几何问题解决基准测试中表现优于经过 GRPO 训练的模型，如 Geometry3K 和 MathVista。图 1 所示，通过谨慎选择辅助构建场景而不是不加选择地应用它，我们的 GeometryZero 展示出卓越的几何推理和反思能力，同时在实现优秀的整体性能和跨不同型号规模的更好扩展性方面优于无条件奖励方法（如 ToRL (?)）中的 RL 方法。总之，我们的贡献总结如下：(1) 我们验证了在其推理过程中通过辅助构建，LLMs 能够更好地解决几何问题解决场景中的复杂任务，在这些场景中，它们利用上下文和经过修改的形式语言进行辅助构建。(2) 在我们的工作中提出了一种新的强化学习方法，叫做组对比策略优化，在强化学习期间，它能够灵活地为不同样本的辅助构建提供鼓励或惩罚信号，避免模型不加区分地应用辅助构建，同时在策略上合理时保持其优势。(3) 我们训练了 GeometryZero 模型，这是一系列轻量级几何推理模型，谨慎地决定何时采用辅助构建。我们在这些模型及基准上进行了广泛实验，同时对 GCPO 进行了深入的消融研究，以验证每个组件的有效性，并进行了详细分析，揭示了关于我们方法的关键见解。

2 相关工作

2.1 几何问题求解

随着大型语言模型 (LLMs) 的发展，研究人员开始将 LLMs 应用于几何问题解决 (**?**) 。然而，一些早期的工作如 **?** 主要集中于代数层面的形式推导，这在解决具有数值解的实际问题时效果有限。其他研究通过提出针对性的基准和数据集来解决几何问题解决数据的缺乏问题 (**???**) 。一些最近的工作使用大规模注释数据对模型进行监督微调，旨在提升多模态 LLMs 在几何问题上的表现 (**?**) 。

多种方法，如 GeoCoder (**?**) ，尝试利用形式化语言作为上下文（例如代码）来帮助模型进行几何推理。其他研究则探索使用符号工具来增强模型的几何推理能力 (**?**) 。最近的研究，如 **???** ，建议通过修改形式化语言鼓励模型构建辅助线，从而更好地利用几何上下文的内在特性来解决问题。

2.2 具有可验证奖励的强化学习

长期以来，强化学习一直是大型语言模型（LLM）研究社区的一个重要关注点 (**???**) 。随着 Deepseek-R1 (**?**) 的出现，研究社区开始关注在各个 AI 领域应用可验证奖励的强化学习，特别是 GRPO (**?**) 。一些研究试图在较小的 LLM 上重现 GRPO 在激励推理能力方面的有效性 (**?**) 。其他人将 RLVR 方法应用于多模态 LLM，以增强其对视觉上下文的理解 (**??**) 。还有一些研究探索将视觉上下文转换为形式语言，并利用推理 LLM 进行推理，旨在超越多模态 LLM 的能力 (**?**) 。

GRPO 算法最初在 **?** 中提出，并应用于数学推理。与 PPO 相比，它简化了强化学习管道，并不再需要一个评论者模型。CPPO (**?**) 试图通过修剪优化 GRPO 算法的效率，降低训练成本同时保持准确性。DAPO (**?**) 引入一个截断机制和动态采样以提高训练的多样性和稳定性。**?** 使 GRPO 的可验证奖励适应视觉感知任务，增强模型在视觉推理中的表现。ToRL (**?**) 尝试将工具使用奖励整合到 GRPO 中，加强模型的工具调用能力以提升其在数学推理中的表现。另一项独立的工作提出了一种时间奖励与准确性奖励相结合的方法，以提高模型在视频语境下的基础表现 (**?**) 。

3 预备知识

群组相对策略优化 (GRPO) 是一种新颖的算法，它利用客观可验证的监督信号来增强在需要强推理能力的任务（例如数学和代码相关问题）上的模型性能。与依赖训练奖励模型的先前方法（例如从人类反馈中进行强化学习）相比，GRPO 利用直接验证函数来提供可靠的奖励反馈。此方法简化了奖励学习机制，同时实现了与任务内在正确性的有效对齐。

具体而言，给定一个问题 q ，策略模型 π_θ 生成一组 N 采样输出 $O = \{o_1, o_2, \ldots, o_N\}$ ，其中每个输出 o_i 通过预定义的可验证奖励函数接收奖励信号 r_i 。然后，GRPO 优化如下的裁剪目标：

$$\max_{\pi_\theta} \mathbb{E}_{O\sim\pi_\theta(q)} \Bigg[\frac{1}{N} \sum_{i=1}^{N} \min \left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} \boldsymbol{A}_i,\ \text{clip} \left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)},\ 1-\epsilon,\ 1+\epsilon \right) A_i \right) - \beta\, \mathbb{D}_{\text{KL}} \left[\pi_\theta(o \mid q) \,\|\, \pi_{\text{ref}}(o \mid q) \right] \Bigg]. \tag{1}$$

。在这里，π_{old} 表示当前更新前的策略，而 π_{ref} 是一个固定的参考策略（例如，初始模型）。A_i 是基于其奖励信号 r_i 的输出 o_i 的优势估计，ϵ 是裁剪阈值，β 是控制 KL 正则化以防止过度策略偏差的超参数。

现有的研究工作，例如 DeepSeek R1-Zero (**?**) 算法，放弃对有监督微调的依赖，而是完全通过强化学习进行训练，特别是在群体相对策略优化（GRPO）框架内。与像 PPO (**?**) 这样的传统强化学习方法相比，GRPO 不需要用一个评论模型来评估策略的输出。给定一个问题 q ，GRPO 首先使用当前策略 $\pi_{\theta_{\text{old}}}$ 生成 G 个不同的响应 $O = \{o_1, o_2, ..., o_G\}$ 。然后，应用奖励函数以获得一组可验证的奖励 $\{\mathbb{R}(o_i)\}$. 。通过计算这些奖励的平均值和标准差，GRPO 对它们进行归一化，并估计每个响应 o_i 的优势值如下：

$$A_i = \frac{\mathbb{R}(o_i) - \mathbb{E}_{o_i\sim O}[\mathbb{R}(o_i)]}{\text{std}(\{\mathbb{R}(o_i)\})}, \tag{2}$$

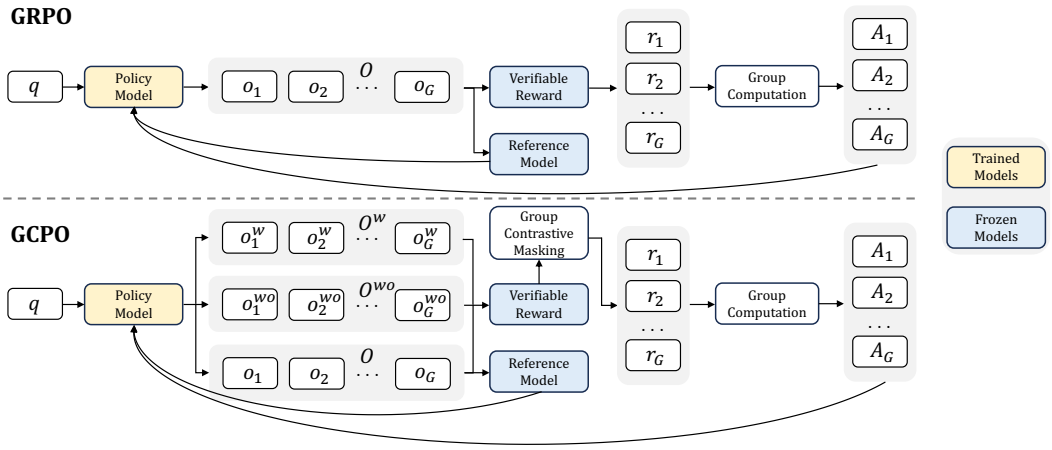

Figure 2: GCPO 的说明。我们的 GCPO 与 GRPO 之间的一个关键区别是群对比屏蔽：(1) GCPO 额外采样两个回滚组 O^{w} 和 O^{wo}，用于通过精度奖励评估定量收益。(2) GCPO 的辅助奖励信号在训练过程中动态地屏蔽为正、负和零，如（方程 4）。另一个区别是，在 GCPO 训练过程中，一个新颖的长度奖励也被纳入可验证奖励 $\mathbb{R}$ 中。

其中，A_i 是对应于第 i 个响应的优势值，表示其相对质量，$\mathbb{R}$ 是可验证奖励的总和。GRPO 框架鼓励模型生成具有更高可验证奖励的响应，从而在推理密集任务中提高可靠性和正确性。

4 方法

4.1 概述

群对比策略优化（GCPO）引入了一个关键的创新修改：它结合了一种被称为群对比掩蔽的重要机制，该机制在辅助构建有益时为辅助奖励提供积极掩蔽，而在其他情况下则应用消极掩蔽（即惩罚）。为实现这一目标，我们提出了群对比掩蔽。由于辅助推理的需求，GCPO 还引入了额外的长度奖励，优化用于更长的完成时间。

4.2 辅助构建的强化学习

为使模型能够结合辅助构造推理（一种工具使用形式，即在思考过程中尝试使用类似 Tikz 代码的正式语言构造辅助线），我们引入了一种辅助奖励，以教授"how-to" 能力，如在 ToRL (**?**) 中，其中新引入了一个工具相关的奖励，以使用编码进行数学推理。在训练过程中，文本上下文包含 TikZ 代码或逻辑形式，具体见附录 **??** 中的细节，这些严格对应几何图形，并提示模型自主决定是否在其推理过程中包含辅助线构造。对于可执行的 TikZ 代码，如果思考过程中更改的 Tikz 代码能够成功执行并渲染图形，则辅助奖励为正；对于逻辑形式，我们检测到特定标记 `<aux>` 和 `</aux>` 的存在，表示尝试修改逻辑形式以构建辅助线。因此，给定响应 o_i 的辅助奖励定义如下：

$$\boldsymbol{R}_{aux}(o_i) = \begin{cases} 1 & if model construct s auxiliary lines, \\ 0 & otherwise. \end{cases} \tag{3}$$

4.3 辅助结构的条件奖励

受现有工作的启发 (**??**)，我们旨在赋予模型在几何问题解决中通过形式语言进行辅助构建的能力。然而，需要注意的是，虽然教会模型“如何去做”很重要 (**?**)，但我们也必须教会它们“何时去做”。尽管 GRPO 已被证明在增强推理能力方面有效，但它缺乏工具使用的条件奖励，仅依靠无条件奖励来鼓励期望的行为，这可能会导致在某些情况下工具的滥用。为了解决这一限制，我们提出了群对比策略优化 (GCPO)，它引入了一种群对比奖励机制，为条件奖励信号提供灵活的鼓励或惩罚，使训练好的模型能够灵活决定是否应用该工具（即辅助构建）。

GCPO 的关键见解来自于这样一个观察：在几何问题求解中，虽然辅助线在很多情况下可以增强推理，但在某些情况下它们可能是不必要的，甚至是有害的。无条件地鼓励辅助构造可

能会导致几何问题等复杂场景的表现不佳，因此，我们需要在强化学习过程中引入条件奖励。

4.4 群体对比策略优化的组成部分

群组对比掩码 GCPO 的核心思想是在训练期间，模型应该能够认识到使用代码绘制辅助线的可量化益处——在有益时使用此能力，并在效果适得其反时避免使用。如图 2 所示，在响应采样期间，模型会生成 G 展开 $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_G\}$，以及另外两个对比展开组：一个需要辅助线思维过程组 $\boldsymbol{O}^{\text{w}} = \{\boldsymbol{o}_i^{\text{w}}\}$，一个禁止辅助线思维过程组 $\boldsymbol{O}^{\text{wo}} = \{\boldsymbol{o}_i^{\text{wo}}\}$。分组对比掩蔽功能定义为：

$$\boldsymbol{Mask}(R_{aux}(O)) = \begin{cases} R_{aux}(O) & if \mathrm{E}(R_{acc}(O^{\text{w}})) > \mathrm{E}(R_{acc}(O^{\text{wo}})) + \epsilon, \\ -R_{aux}(O) & if \mathrm{E}(R_{acc}(O^{\text{wo}})) > \mathrm{E}(R_{acc}(O^{\text{w}})) + \epsilon, \\ 0 & otherwise, \end{cases} \tag{4}$$

其中 $R_{acc}(o_i)$ 表示准确性奖励函数，指示响应 o_i 是否包含正确的最终答案，$R_{aux}(o_i)$ 指的是（方程 3 中的）辅助奖励，而 ϵ（在实验中设为 0.05）是控制奖励掩蔽的阈值超参数。

长度奖励 辅助构造思维需要更深入的多维度分析，因而需要更长的推理过程。受长度控制策略优化（LCPO）(**?**) 的启发，我们通过引入一个简化的长度奖励来进行调整，其中 $\text{len}(o_i)$ 计算响应中的标记数，而 $l_{\max}$ 是允许的最大完成长度。

$$R_{length}(o_i) = \min\{1, \frac{\text{len}(o_i)}{l_{\max}}\} \tag{5}$$

可验证的奖励组合 作为 RLVR 的一个变体，GCPO 的可验证奖励 $\mathbb{R}(o_i)$ 结合了如下多种加权成分，其中 $R_{GRPO}(o_i)$ 包括准确性奖励和确保适当输出结构的格式奖励，而表示辅助奖励权重的超参数 λ 和表示长度奖励权重的超参数 β 均被设定为 0.5。

$$\mathbb{R}(o_i) = R_{GRPO}(\boldsymbol{o}_i) + \lambda \cdot \boldsymbol{Mask}(R_{aux}(\boldsymbol{o}_i)) + \beta \cdot R_{length}(\boldsymbol{o}_i) \tag{6}$$

从本质上讲，GCPO 使用掩蔽辅助奖励来教授适当的工具使用情境，而长度奖励则确保足够的推理能力。该框架在很大程度上继承了 GRPO 的可验证奖励框架，采用基于结果的强化学习进行模型训练。

5 实验

5.1 实验设置

对于数据集，我们应用了一个综合数据集，其中的数据采样自两个流行的几何问题求解（GPS）数据集，包括 Geomverse (**?**) 和 Geometry3k (**?**)，用于训练 SFT 和 RL 模型。训练数据集的详细配方在附录 **??** 中详述。关于训练的细节，我们在训练和评估过程中利用了 vLLM 推理框架 (**?**)。更多细节在附录 **??** 中呈现。

至于模型，受 (**?**) 的启发，我们将视觉图表转换为形式化的语言上下文，以提高推理性能。因此，我们使用 Qwen2.5 (**?**) 系列语言模型进行训练。在基准测试方面，除了 Geometry3k 和 Geomverse 的领域内基准测试，OOD 几何基准测试还包括 MathVista (**?**) 和 OlympiadBench (**?**)。详情在附录 **??** 中提供。

基准线 为了充分展示 GCPO 的有效性，我们与以下基线方法进行了比较：（1）SFT。该模型使用提示-响应对进行监督微调，其中响应是人工注释的或从有能力的大型语言模型中提取的。（2）GRPO (**?**)。一种通过可验证结果的强化学习算法，模型生成多个响应以估计基线优势，从而鼓励推理并改善问题解决，这消除了使用评论模型或奖励模型的需要。（3）ToRL (**?**)。一种基于 GRPO 的强化学习算法，附加了额外的奖励函数（方程 3），无条件地鼓励工具使用（即，辅助构造）。

5.2 结果

如表 1 所示，我们在四个几何问题解决基准上的实验结果表明，GCPO 在提高模型处理几何问题的能力方面具有有效性。主要发现总结如下：

Base Model	Method	Geomverse	Geometry3k	MathVista	OlympiadBench	Avg.
Qwen2.5-1.5B-Instruct	-	4.20	41.76	47.70	13.44	26.78
	+ SFT	4.80	44.25	43.11	14.51	26.67
	+ GRPO	5.76	53.35	57.79	14.51	32.85
	+ ToRL	5.26	57.01	57.79	11.29	32.84
	GeometryZero (+ GCPO)	6.96	60.23	61.77	19.35	37.08
Qwen2.5-3B-Instruct	-	10.53	65.83	67.88	32.25	44.12
	+ SFT	10.20	71.65	73.08	30.64	46.39
	+ GRPO	12.13	75.87	82.87	31.72	50.65
	+ ToRL	12.63	77.31	81.34	33.87	51.29
	GeometryZero (+ GCPO)	11.30	79.25	82.56	35.48	52.15
Qwen2.5-7B-Instruct	-	14.76	70.99	68.19	39.24	48.30
	+ SFT	15.36	75.98	76.14	41.93	52.35
	+ GRPO	16.93	79.03	86.23	40.32	55.63
	+ ToRL	12.56	78.75	83.48	44.08	54.72
	GeometryZero (+ GCPO)	18.23	78.81	87.15	45.69	57.47

Table 1: 主要的实证结果。BoN@3 通过率结果涵盖了域内基准，包括 Geomverse、Geometry3k，以及 MathVista 和 OlympiadBench 的域外结果，其中最佳结果以粗体显示。我们的 GeometryZero（相对于，+ GCPO）模型的结果显示在 灰色 部分。

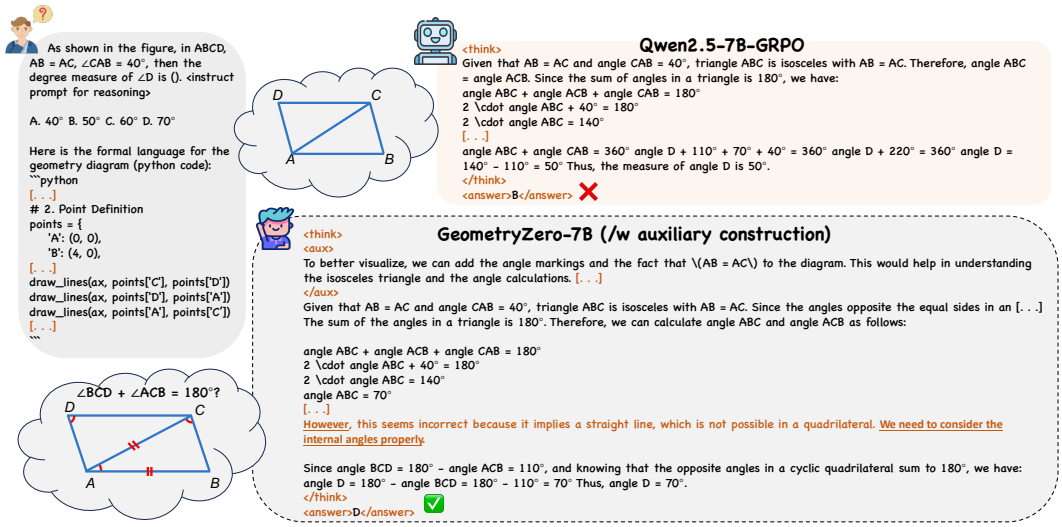

Figure 3: GRPO 与 GCPO 之间的案例研究。两个回复比较了 Qwen2.5-7B-GRPO 与我们的 GeometryZero-7B 在一个 MathVista 问题上的表现，揭示了 GeometryZero-7B 在其推理过程中如何有效地构建辅助元素。推理过程中的 orange underlined 文本是在几何问题解决中的反思过程。

我们观察到，SFT 模型（Qwen2.5-1.5B-SFT 和 Qwen2.5-3B-SFT）在 Geomverse 和 Geometry3k 等域内基准测试上，相较于原始的 Instruct 模型显示出一致的改进。例如，Qwen2.5-1.5B-SFT 和 Qwen2.5-3B-SFT 在 Geometry3k 上分别取得了 2.49 % 和 5.83 % 的提升。然而，相较于 RL 方法，这些 SFT 模型在像 MathVista 和 OlympiadBench 的 OOD 基准测试中表现出性能下降或较小的增益。例如，在 OOD 基准 MathVista 上，Qwen2.5-1.5B-SFT 比基础模型表现下降了 4.59 %，而 Qwen2.5-1.5B-GRPO 则显示出显著的提升，达到 10.09 %。总体而言，包括 GRPO、ToRL 和 GCPO 在内的 RL 方法在域内和 OOD 基准测试中实现了更一致的改进，超越了 SFT，证明了强化学习的有效性。

与 GRPO 相比，ToRL 模型在所有示例的推理过程中无条件地鼓励辅助构建，并采用无条件奖励设计（方程 3）。实证结果表明，ToRL 模型在各种模型规模上并未显现出相对于 GRPO 的明显优势，这表明这种粗粒度的策略未能在几何问题解决场景中为辅助构建提供显著好处。例如，虽然 ToRL 在 3B 模型上相对于 GRPO 表现出微小的 0.64 % 优势，但在 7B 模型上表现出 0.91 % 的性能下降。相比之下，GCPO 提高了模型在域内和 OOD 基准上的性能，并在大多数基准测试中的各种模型规模上取得了一贯更好的平均性能。这表明明确何时进行辅助推理为提高问题解决能力奠定了基础。如图 3 所示，GCPO 通过在推理过程中生成辅助构建来增强几何问题解决，我们在附录 **??** 中提供了更多案例研究。