

Com²：一个用于探索大型语言模型中复杂常识推理的因果指导基准

Kai Xiong[♣] Xiao Ding^{♣†} Yixin Cao[♣] Yuxiong Yan[♣] Li Du[♦] Yufei Zhang[♣]
Jinglong Gao[♣] Jiaqian Liu[♣] Bing Qin[♣] Ting Liu[♣]

[♣]Research Center for Social Computing and Interactive Robotics

Harbin Institute of Technology, Harbin, China

[♣]Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China

[♦]Beijing Academy of Artificial Intelligence, Beijing, China

[♣]Nanjing University, Nanjing, China

{ kxiong, xding, yxian, yfzhang, jlgao, qinb, tliu } @ir.hit.edu.cn
yxcao@fudan.edu.cn

Abstract

大型语言模型 (LLMs) 通过预训练掌握了丰富的简单和显性常识知识，使其能够在简单常识推理中实现类人的表现。然而，LLMs 在处理由简单常识衍生而来的复杂和隐性常识知识时仍然面临困难（例如理解某些事件的长期影响），这一方面是人类更关注的现象。现有的研究集中于像数学和代码这样复杂的任务，而复杂常识推理由于其不确定性和缺乏结构而仍未被充分探索。为了填补这一空白并与现实世界的问题相结合，我们提出了一个名为 Com² 的基准，专注于复杂常识推理。我们首先引入因果事件图以提供结构化的复杂常识。然后我们采用因果理论（例如，干预）来修改因果事件图，并获得符合人类关切的不同情境。最后，利用 LLM 来合成以慢速思考为特征的示例，慢速思考由修改后的因果图中的逻辑关系所引导。此外，我们使用侦探故事构建了一个更具挑战性的子集。实验表明，LLMs 在推理的深度和广度上存在困难，而后期训练和慢速思考可以缓解这一问题。代码和数据可在 [通信²](#) 获得。

1 介绍

大型语言模型 (LLMs) 在各个领域中取得了显著进展 (Plaatt et al., 2024; Ying et al., 2024)，展现了强大的泛化能力和推理能力。通过大规模的预训练，LLMs (Liu et al., 2024; Yang et al., 2024) 可以从大量未标注的文本中学习和提取知识。最新的迭代版本，如 OpenAI 的 o1 (OpenAI, 2024) 和 Deepseek R1 (Liu et al., 2024)，通过实施测试时的扩展策略，进一步增强了这些模型的推理能力。这些策略鼓励使用更长的思维链 (CoT) 来处理更复杂的查询，模拟类似人类的反思和校正机制以提高推理能力。虽然这些改进主要集中于与数学 (AIME) 和代码 (Quan et al., 2025) 相关的任务，但复杂常识推理的领域仍然相对未被探索。

在本文中，我们介绍了 Com²，这是一个专门设计用来评估高级 LLM 中复杂常识推理的

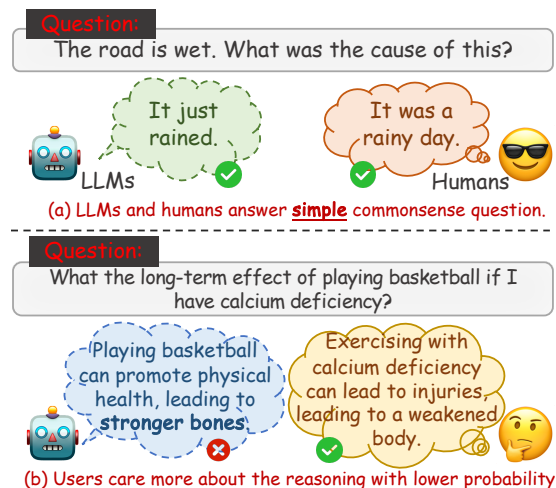


Figure 1: (a) 一个一步推理问题。(b) 用户更关注不常见的结果。

基准。我们认为，尽管它们在现有基准 (Wei et al., 2022b; Xiong et al., 2024b) 上的表现令人印象深刻，但 LLM 并不本质上擅长处理复杂的常识推理。在现实世界的场景中，人类经常解决涉及长而多方面的常识推理链的复杂问题。这些问题通常需要整合简单而明确的常识概念，有时还会伴随着认知偏见或思维陷阱。如图 1 (a) 所示，目前单步常识推理任务的一个例子，一旦相关知识已知，就可以轻松回答。然而，我们强调，即使使用测试时间缩放，LLM 的内生缺陷可能会导致在思维陷阱中失败，例如不常见的场景（图 1 (b) 中“缺钙打篮球”）。

策划一个复杂的常识推理数据集的挑战源于两个主要因素。首先，与数学或代码不同，常识知识的表达常常是非正式且依赖于上下文的。尽管数学或编程任务通常具有明确且形式化的规则，但常识推理更为细致入微且易于被不同解读，这使得创建结构化的数据集变得复杂。其次，常识推理任务很少有普遍接受的标准答案，这使得数据集可能变得模糊或难以使用。这种缺乏明确、客观答案的情况可能导致常识推理基准中的混乱和不一致。

为了解决上述挑战，我们首先采用因果事件

[†] Corresponding Author

图 (CEGs) (Ding et al., 2019; Du et al., 2021) 作为复杂常识知识的表示, 因为 CEGs 编码事件之间复杂且逻辑严谨的关系。基于 CEGs, 最可能的结果被视为假定的真理。在此之后, 我们使用因果理论 (Pearl and Mackenzie, 2018) 对 CEGs 进行修改, 如干预, 以构建符合人类关注的不同常识推理场景。最后, 我们基于修改后的 CEGs 合成了直接、决策、过渡、干预和反事实的推理任务 (Com²-main)。除了 Com²-main 之外, 我们还额外提出了一组更具挑战性的任务 Com²-hard, 它基于侦探故事, 并包括决策、干预和反事实任务。

基于 Com², 我们首先评估了大量现有的 LLMs, 然后进行了几项深入分析。我们有以下关键发现和见解: (1) 常识推理仍然是 LLMs 的一个显著挑战, 因为即使有长的推理链, 性能并不总是提高, 甚至可能下降; (2) 经过大规模的预训练后, LLMs 具有可接受的反事实推理能力; (3) LLMs 具有有限的推理广度, 这导致它们在处理不常见或突发场景时显得困难。(4) 通过后训练或慢思考为 LLMs 配备能力可以缓解上述限制; 我们的主要贡献可以总结如下:

- 我们是最早提出一个基准 Com² 的人之一, 该基准专注于复杂的常识推理领域。
- 我们建议使用因果事件图和因果理论作为骨架来创建复杂问题, 以应对现实世界的关注。
- 我们全面评估现有的 LLMs, 以获得对在复杂常识推理方面开发更强大 LLMs 的见解。

2 背景

2.1 因果事件图

因果事件图 (CEG) (Ding et al., 2019; Heindorf et al., 2020) 是一个被表示为 $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ 的有向无环图。图 2 (a) 显示了一个因果事件图的例子, 其中 \mathcal{V} 是一组节点, 每个节点 v_i 表示一个自然语言事件, 例如“事故”。 \mathcal{E} 是一组有向边, 每条边 (v_i, v_j) 表示从 v_i 到 v_j 的因果关系, 这意味着 v_i 是原因而 v_j 是结果。

CEGs 是半结构化的, 这意味着事件之间存在严格的逻辑关系。一个 CEG 可以表示一段复杂且隐含的常识知识, 因为这类知识通常跨越多个文档并随着时间延展, 使其难以直接在文本中表达。

2.2 因果理论

我们采用因果理论 (Pearl and Mackenzie, 2018) 来创建不同的场景以匹配用户的关注点。我们介绍了两种操作:

干预是一个外部动作, 它积极地操控变量 X , 以观察其对结果 Y 的因果影响, 记作 $do(X)$, 这样打破了 X 对其先前原因的自然依赖。图 2 (b) 展示了对“吸烟”的干预。通过干预, 我们可以影响常见情况, 并引导事件向不太典型的方向发展。

反事实指的是一种假设情景, 该情景探讨如果某个特定的过去事件有所不同会发生什么, 通常表示为 Y_X , 其中 X 是一个在现实中没有发生的变量。图 2 (c) 展示了一个反事实, 其中“D 日”是一个已经发生的事件。反事实使我们能够创建假设情景。

为了弥合用户关注点与模型在复杂常识推理中的能力差距, 我们提出了 Com² 基准。如图 3 所示, 创建过程包括四个步骤: (1) 事件提案分别为反事实和其他场景提出具体和抽象事件。(2) 因果链提案构建因果链以创建简单场景。(3) 因果图提案利用因果理论综合多样因果图, 构建复杂场景以满足用户关注。(4) Com² 综合基于因果图生成多选和多选项例子以构建基准。所有提示见附录 A。

2.3 事件提议

为了构建因果图, 我们首先提出一些事件作为生成的种子。具体来说, 我们从现有的因果事件图 (Heindorf et al., 2020) 中收集 k 种多样的事件, 并将它们用作 k -shot 示例来提示 ChatGPT (Achiam et al., 2023) 生成两类事件:

- 混凝土: 一个具体事件 e_i^c 是已经发生的现实世界事件, 这将其与抽象事件区分开。这些事件被提议用于反事实因果图的构建。
- 摘要: 抽象事件 e_i^a 是指不受特定时间、地点或实例约束的广义事件。它代表一类动作或现象。提出抽象事件是为了为其他类型的因果图做准备。

最后, 我们提出了 n 个具体事件和 n 个抽象事件, 分别表示为 $E^c = \{e_1^c, e_2^c, \dots, e_n^c\}$ 和 $E^a = \{e_1^a, e_2^a, \dots, e_n^a\}$ 。

2.4 因果链提案

根据具体和抽象事件 E^c 和 E^a , 我们创建对应的因果链, 作为构建因果图的基础。 E^c 和 E^a 中的每个事件都作为其各自因果链的根事件。然后我们使用精心设计的提示词, 与 ChatGPT (Achiam et al., 2023) 生成由 5 个事件组成的具体和抽象因果链。因此, 每个具体或抽象因果链代表一个特定的简单场景。

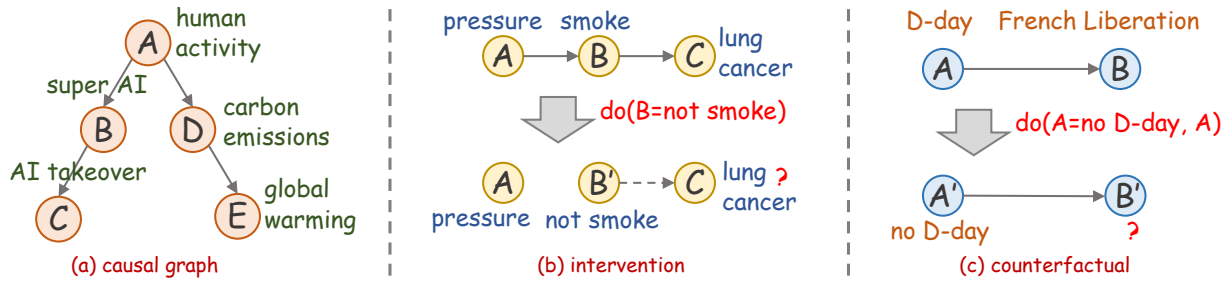


Figure 2: 三个例子分别为：(a) 因果图，(b) 干预，和 (c) 反事实。

最终，我们获得了 n 个具体的因果链和 n 个抽象的因果链，分别记为 $S^c = \{s_1^c, s_2^c, \dots, s_n^c\}$ 和 $S^a = \{s_1^a, s_2^a, \dots, s_n^a\}$ 。

为了与用户的关注点保持一致，我们采用了因果理论 (Pearl and Mackenzie, 2018; Xiong et al., 2022)，然后创建了 5 种因果图，如下所示：

- 直接：基于每一个抽象因果链 s_i^a (第三步中的直接)，直接因果图反映 s_i^a 。这反映了某一事件的长期影响，这是最简单的情景。

- 决策：对于每个抽象因果链 s_i^a ，我们使用提示生成另一个具有不良结果的因果链 (步骤 3 中的决策)。新链与 s_i^c 共享同一个根事件，与 s_i^a 形成一个决策图。这表示一个导致负面结果的复杂情境。

- 转换：对于每个抽象因果链 s_i^a ，我们使用提示生成另一个带有因果传递问题如场景漂移的因果链 (Xiong et al., 2022) (步骤 3 中的转换)。随着因果链或推理深度的增加，其合理性可能逐渐减弱。该转换图模拟了这种复杂场景。

- 干预：基于每个抽象因果链 s_i^a ，我们在 s_i^a 进行干预，以切断从根事件到尾事件的因果传播。接下来，我们使用提示生成另一个发生概率低于 s_i^a 的因果链，以形成干预图 (步骤 3 中的干预)。这个干预图展示了一个复杂的场景，其中一个突发事件发生并导致因果链走向一个不常见的方向。

- 反事实：对于每个具体的因果链 s_i^c ，我们首先为 s_i^c 中的根事件创建一个反事实事件，然后使用提示词创建一个反事实因果链，并形成反事实图 (步骤 3 中的反事实)。根据因果理论，这应该代表了所有类型中最具挑战性的场景。

最后，我们获得了 5 种因果图，每一种都代表一个复杂的场景。它们可以用于创建复杂的常识推理问题，以符合用户的关切。

在此步骤中，我们合成了 Com²-main，包含 5 个任务，这些任务基于因果图形特点具有复杂常识推理问题。具体来说：

- 直接：基于每个直接 CEG，我们使用头

事件设计一个多项选择题 (MCQ) 来询问长期结果，以尾事件作为答案。干扰项被添加为备选项。这个任务与用户对事件长期影响的关注相一致。

- 决策：对于每个决策 CEG，我们使用根事件创建一个多选题 (MSQ) 来寻求防止不可接受事件的建议。有效的干预措施作为正确答案，其它选项作为干扰项。这个任务与用户的寻求建议相一致。

- 转换：对于每个转换 CEG，我们创建一个以根事件为中心的 MCQ，来询问长期结果。因果链中带有传递性问题的尾事件是干扰项，而合理链中的尾事件是答案，附加干扰项作为替代选项。这将随着推理深度的增加来评估 LLMs 的可靠性。

- 干预：对于每个干预 CEG，我们创建一个关于根事件的 MCQ，以询问长期结果。干预是问题中新增的事件。在干预因果链中的尾事件作为干扰项，而在合理因果链中的尾事件是答案。其他干扰项作为备选项包含在内。该任务与用户在突发事件发生时关心的结果相一致。

- 反事实：对于每个反事实 CEG，我们使用具体的因果链和反事实事件来创建一个 MCQ，以询问长期结果。一些干扰选项作为替代方案被包括在内。这与用户对假设情景中长期结果的关注一致。

最后，对于每个任务中的每个例子，我们额外提供一种慢速思考，这种思考由相应的 CEG 引导，包括系统分析、分而治之、自我完善和上下文识别 (Wu et al., 2024)。因此，上述 5 个任务构成了 Com²-main。

除了 Com²-main，我们还旨在创建更具挑战性的版本 Com²。由于侦探故事具有多重交织的线索，需要复杂的推理组合来得出结论。这导致了比以前的 CEG 更为复杂的情境。受到此启发，我们基于 BMDS (Hammond and Stern, 2022) 开发了 Com²-hard。BMDS 是一个包含 400 多个故事的集合，这些故事包括关键线索和证据的注释。具体而言，我们首先让 ChatGPT 获取侦探故事来提取 criminal(s)。

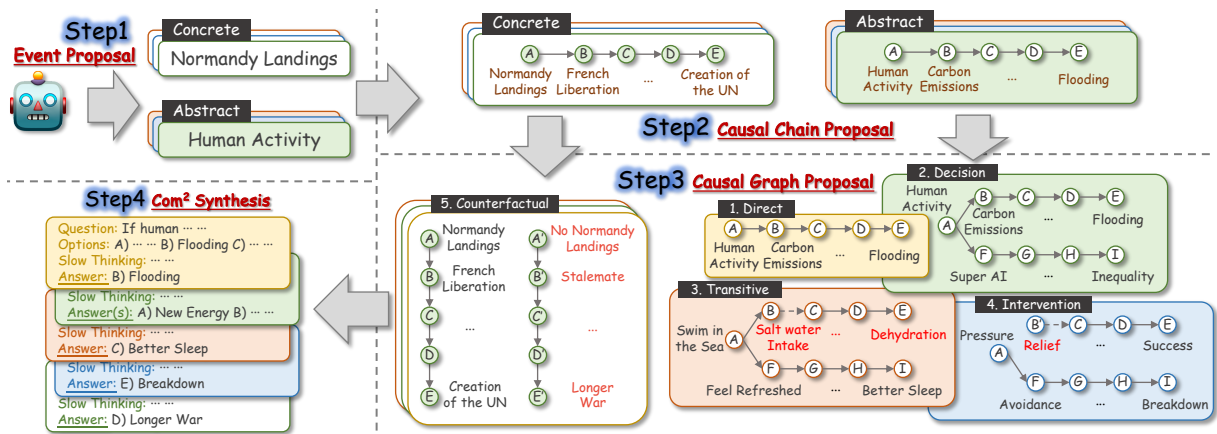


Figure 3: Com² 构建的完整工作流程：（1）事件提议利用 LLMs 提出具体和抽象事件；（2）因果链提议基于所提出的事件综合因果链；（3）因果图提议采用因果理论生成基于综合因果链的各种因果图；（4）Com² 合成通过因果图指导，创建慢思考的多选和多选题。

Com ²	Dir.	Dec.	Trans.	Inter.	Counter.	Total
Main	500	500	500	500	500	2,500
Hard	-	418	-	418	418	1,254

Table 1: Com² 基准测试的统计数据

接下来，ChatGPT 总结所有相关人物的事实和行动（记为 clues）。最后，我们创建了 3 个任务：

- 决策：我们创建一个多项选择题（MSQ），以请求防止犯罪发生的建议。所有对犯罪的合理干预都被视为答案。一些干扰项将被加入以形成选项。缓慢思考也受到侦探故事中的案件解决路线指导，并且与每个 Com² 主要任务具有相同的动作。
- 干预：我们在 clues 中排除了关键线索，然后创建一个选择题来询问罪犯。关键线索作为干预被添加到问题中。criminal(s) 被视为答案。一些嫌疑人会被整合形成选项。慢速思考由侦探故事中的破案路线引导，并与每个 Com² -主要任务具有相同的动作。
- 反事实：我们首先基于侦探故事创建一个反事实，然后创建一个 MCQ 以询问反事实的结果。反事实结果作为答案。加入一些干扰选项以形成选项。慢思维由案件解决路线和反事实引导，其具有与每个 Com² 主任务相同的动作。

2.5 Com² 统计数据

我们进行上述过程来创建 Com²，我们使用的 ChatGPT 的 API 是 gpt-4o-mini。如表 1 所示，Com² -main 中的每个任务有 500 个例子，Com² -hard 中的每个任务有 418 个例子。例子可以参考附录 ??。

我们进行了一个人工评估来评估 Com² 的

质量。有关人工评估的详细信息（例如报酬和协议）可以参考附录 ??。通过向人工标注者提供因果图，人类可以分别在 Com² -main 和 Com² -hard 上实现平均 92 % 和 90 % 的准确率。慢思考过程的可靠性为 88 %。这些结果表明 Com² 的质量是令人满意的。

3 实验

3.1 研究的大型语言模型

基于 Com²，我们采用多种不同类型的大型语言模型来评估它们的复杂常识推理能力：

- 通用大型语言模型：（1）Qwen（Yang et al., 2024）：7B Qwen2-Instruct、7B、14B 和 32B Qwen2.5-Instruct。（2）LLaMA（Dubey et al., 2024）：LLaMA-3.1-8B-Instruct、1B 和 3B LLaMA-3.2-Instruct。（3）Gemma（Team et al., 2024）：9B 和 27B Gemma2-it。（4）ChatGPT（Achiam et al., 2023）：GPT-4o 和 GPT-4o-mini。

- 理解 LLMs：（1）Open-O1（Team, 2024），这是一个基于 LLaMA-3.1-8B-Instruct 的理解调优 LLM。（2）Marco-o1（Zhao et al., 2024），这是一个基于 Qwen2-7B-Instruct 的理解调优 LLM。（3）R1-distilled（Liu et al., 2024），这是一个基于 LLaMA-3.1-8B 和 DeepSeek-R1 的理解 LLMs。（4）QwQ-32B-Preview（Yang et al., 2024），这是一个基于 Qwen2.5-32B 的复杂理解 LLM。（5）o1-mini，这是由 OpenAI 开发的一个强大的理解 LLM。

对于 Com² 中的每个任务，我们采用准确率来量化每个 LLM 的性能。具体来说，对于由多选题组成的决策任务，我们设计了一种软策略。对于每个多选题，准确率由在全部正确答案集中正确预测的答案选项的比例来确

LLMs		Main						Hard				Overall
		Dir.	Dec.	Trans.	Inter.	Counter.	Avg.	Dec.	Inter.	Counter.	Avg.	
General	Qwen2-7B	80.20	59.25	47.60	34.00	69.60	58.13	28.13	57.76	78.23	54.71	56.42
	Qwen2.5-7B	83.40	67.83	49.80	32.80	73.40	61.42	31.13	51.07	74.88	52.36	56.89
	Qwen2.5-14B	80.40	66.95	48.20	31.80	72.00	59.84	34.94	51.79	75.84	54.19	57.02
	Qwen2.5-32B	83.60	65.16	48.80	33.80	72.40	60.73	30.21	54.89	79.19	54.80	57.77
	LLaMA-3.1-8B	83.20	58.04	47.00	30.40	71.40	58.01	37.62	48.93	74.16	53.56	55.79
	LLaMA-3.2-1B	68.20	27.16	35.60	24.20	47.20	40.52	3.14	27.92	28.95	20.01	30.27
	LLaMA-3.2-3B	81.20	58.04	40.20	29.20	72.40	56.20	21.22	48.45	62.20	43.96	50.08
	Gemma2-9B	78.20	11.99	45.20	26.40	68.40	46.16	43.20	53.46	77.75	58.13	52.15
	Gemma2-27B	77.40	60.29	49.20	28.40	69.60	56.97	45.73	49.64	75.36	56.90	56.94
	GPT-4o-mini	83.20	62.54	49.20	31.40	71.20	59.50	33.46	53.46	78.95	55.29	57.40
	GPT-4o	80.60	66.43	48.40	32.20	68.80	59.26	45.10	56.09	77.99	59.72	59.49
Reasoning	Open-O1	75.60	41.67	43.80	30.40	60.00	50.29	62.47	52.03	71.05	61.84	56.07
	Marco-o1	77.60	41.77	43.60	31.80	65.80	52.11	62.64	53.94	71.77	62.78	57.45
	R1-distilled	75.20	56.51	43.40	30.00	68.20	54.65	60.22	54.24	73.64	62.70	58.68
	QwQ-32B	79.80	59.82	47.40	32.00	64.60	56.70	41.81	44.39	69.86	52.01	54.36
	o1-mini	80.00	32.64	47.80	30.00	66.60	51.48	43.44	51.79	74.40	56.54	54.01

Table 2: 各种 LLMs 在 Com² 上的整体表现。Dir.、Dec.、Trans.、Inter. 和 Counter. 分别代表直接、决策、过渡、干预和反事实。“Avg.” 表示在 Com²-main 或 Com²-hard 上所有任务的平均表现。“Total” 表示在 Com² 上的整体表现。

定。完美匹配得分为 1，部分匹配根据正确选项预测的比例得分，包含错误选项的预测得分为 0。

我们采用零样本设置来评估每个 LLM。要求每个 LLM 首先提供 (Wei et al., 2022b)，然后给出答案。所有的评估提示可以参考附录 B。

3.2 总体结果

总体结果如表 2 所示，从中我们可以有以下观察：

(1) 对于几乎所有的大型语言模型，Com²-main 的平均表现高于 Com²-hard。这是因为 Com²-hard 包含多重交织的线索，需要不同且复杂的推理组合才能得出结论。

(2) 根据因果理论 (Pearl and Mackenzie, 2018)，反事实应该是最困难的任务。有趣的是，大型语言模型在反事实任务上的表现比在决策、过渡和干预任务上更好。我们推测，大型语言模型可能通过大量的预训练获得了足够的假设推理能力。

(3) 几乎所有的大语言模型在直接任务上的表现都优于其他任务，因为这可能源于大语言模型只需沿着因果链进行推理。大语言模型在推理深度方面可能表现得令人满意。

(4) 在 Com²-main 上，Transition 和 Intervention 比其他要困难得多。这表明，当推理深度增加时，LLMs 可能存在处理长期因果关系不足的问题。此外，LLMs 发现处理突发事件 (介入) 很困难。这表明 LLMs 的推理广度有限。

(5) 在 Com²-main 上，干预比决策更困难，但在 Com²-hard 上，局势则完全相反。主要原因是，LLM 可能通过预训练记住了侦探，从而更容易推断出罪犯。

(6) 在 Com²-hard 和整体性能方面，有限访问的 LLMs (GPT-4o, GPT-4o-mini) 相比开放访问的 LLMs 具有一些优势。这归因于 ChatGPT 更好的综合能力。

(7) o1-mini 和 QwQ 的表现不如其他推理大规模语言模型，这可能是因为它们在其他领域上过拟合。

(8) 更大的 LLM 不一定能带来更好的表现 (Qwen2.5 系列)。LLM 中的复杂常识推理可能也需要一定规模才能显现 (LLaMA-3.2 系列)。

3.3 通用 LLMs 与推理 LLMs

我们将推理大型语言模型与其对应的通用大型语言模型进行比较 (例如，Open-O1 和 LLaMA-3.1-8B-Instruct)，并得出以下结论：

(1) 在 Com²-hard 问题上，经过推理微调的 LLMs 优于其对应的通用 LLMs (如 Open-O1 和 LLaMA-3.1-8B-Instruct)，因为大量的思维过程能够帮助 LLMs 理解和处理更复杂的场景。

(2) 然而，在 Com²-main 上，推理 LLM 通常并未比一般 LLM 占有优势，且其表现有所下降。这可能是由于经过推理调整的 LLM 过度考虑所致。这也表明推理 LLM 也需要适当难度水平的问题来匹配它们的能力。

(3) 在 Com²-hard 中，相比于通用模型，推理 LLMs 的改进主要体现在决策任务上。这也是 Com²-hard 中最难的任务。这进一步证明了任务越困难，推理 LLMs 的优势越大。

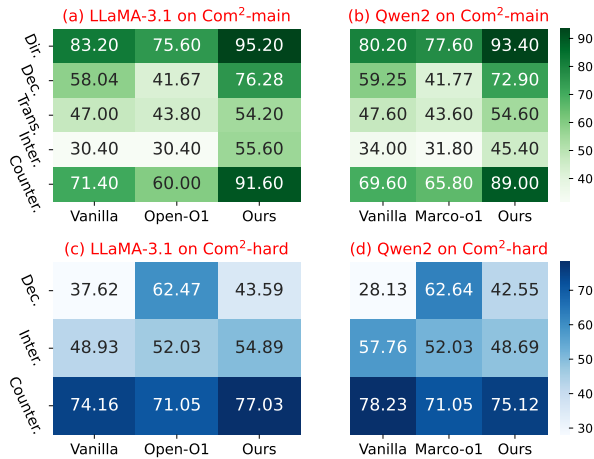


Figure 4: 训练后的 LLMs（我们的模型）和基线模型在 Com² 基准上的整体表现。

4 分析

为了进一步研究复杂的常识推理场景，我们设计了几个深入的分析。(1) 我们在复杂常识推理示例上训练 LLMs，以调查它们的有效性。(2) 我们分析模型性能与输出令牌数量之间的关系，以讨论推理过程中的扩展规律。(3) 我们通过将它提供给 LLMs 用于复杂常识推理 (Com²)，调查在每个示例中慢速思考的有效性。

为了研究复杂的常识推理数据是否能够帮助大型语言模型 (LLMs) 自我提升，我们采用了 Com² 的主要构建过程 (第 ?? 节) 来构建训练数据集。训练数据包括 8,386 个例子，其中分别有 4184、476、475、2757 和 476 个问题用于直接、决策、过渡、干预和反事实任务。我们选择了 LLaMA-3.1-8B-Instruct (LLaMA) 和 Qwen2-7B-Instruct (Qwen2) 进行训练。训练细节和案例可以参考附录 ?? 和 ??。

我们在 Com² 上评估训练的 LLMs（标记为 Ours）和基准模型。请注意，Com²-hard 被视为一个分布外 (OOD) 数据集。总体结果可以参见图 4，从中我们有以下观察：

(1) 经过训练，LLaMA 和 Qwen2 在 Com²-main 上相比普通的 LLM 取得了显著的提高。在 Com²-hard 上相较于普通的 LLM 的提高虽然显著，但相对较小。这主要是因为 Com²-hard 是 OOD。这也表明，LLM 在较简单的任务上可以学习复杂的常识推理能力，并将其迁移到更复杂的任务中。

(2) 在主任务²上，经过训练的 LLMs 在 Transition 任务上的改进比在其他任务上要小得多。我们认为需要更多的示例来教导 LLMs 学习数量和场景信息的区别。

(3) 在 Com²-hard 上，经过训练的 LLMs 在决策任务上的改进是显著的。主要原因可能是

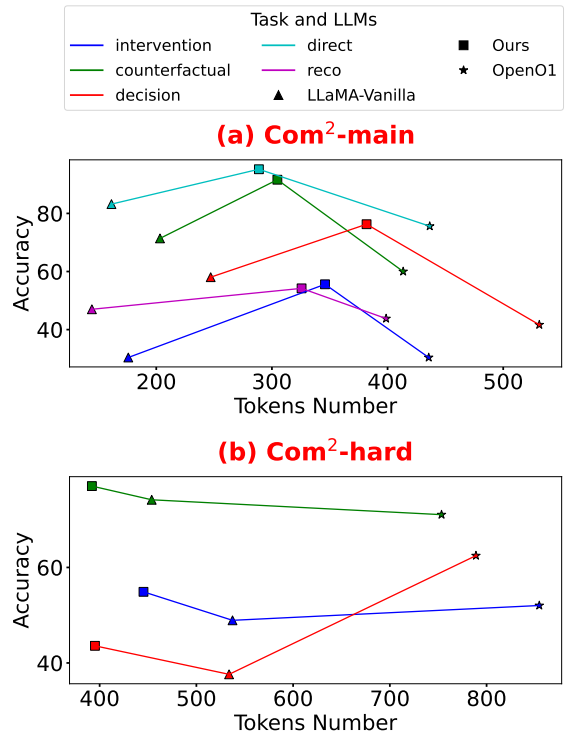


Figure 5: LLMs 性能与输出标记数量之间的关系。

LLMs 在训练之前无法很好地处理多选问题。

(4) 在 Com²-hard 上，训练过的 LLM 表现比推理 LLM (Open-O1 或 Marco-o1) 差。这表明需要更多类型的训练数据来进行改进。

4.1 测试时缩放法则

为了讨论推理过程中的缩放规律，我们统计了大型语言模型 (LLMs) 输出的标记数量。总体结果如图 5 所示，我们可以推断出：(1) 输出更多的标记不一定总能提高性能。在数学和代码中的测试时缩放规律可能对常识推理无效，因为它们仍可能陷入常识捷径。后训练可以缓解这一点，以改善性能并减少标记使用。(2) 这一结论也得到了干预性能 (图 4) 的支持，因为推理 LLMs 的表现远不如我们的方法。(3) 与数学相比，总体标记使用量仍然小得多。值得期待常识推理中的测试时缩放。

为了研究在每个 Com² 示例中，慢思考是否可以帮助语言模型进行复杂的常识推理，我们为语言模型提供了用于答案选择的慢思考。我们选择了 Qwen2-7B-Instruct 和 LLaMA-3.1-8B-Instruct 进行实验。通过慢思考提升的语言模型被标记为 Qwen-tp 和 LLaMA-3.1-tp。Qwen-tp、LLaMA-tp 以及基线的总体表现可以参考图 6。我们可以得出以下结论：

(1) 在为大型语言模型 (LLMs) 提供慢思考之后，两个 LLMs 在 Com²-main 和 Com²-hard 上都取得了显著的提升。原因在于慢思考

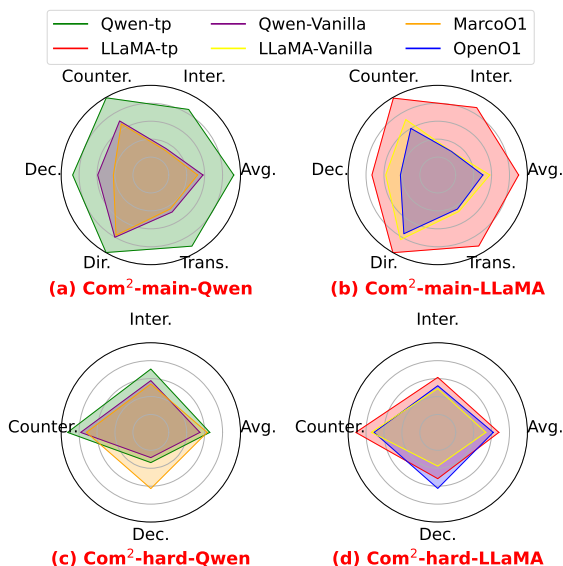


Figure 6: LLaMA 和 Qwen 在慢思考过程指导下的整体表现，以及基准。

可以对问题进行详细分析，并帮助澄清事件之间的因果关系，从而提升性能。

(2) 然而，对于决策任务，它并没有达到极高的水平。我们认为这可能是因为决策任务的多选性质，而慢思考时间太长，使得大型语言模型难以提取推理的关键信息。

5 相关工作

5.1 在大型语言模型中的推理

大型语言模型已经革新了使用深度神经网络进行推理的方法。使用大型语言模型进行推理可以分为基于提示的方法和基于训练后的方法。

至于基于提示的方法，GPT-3 (Brown et al., 2020; Chen et al., 2021) 的出现为 LLMs 执行零样本和少样本推理打开了大门。Wei et al. (2022b) 提出了链式思维 (CoT)，该方法在 LLMs 中引发推理并获得自解释性。Kojima et al. (2022) 和 Zhang et al. (2023) 分别提出了零样本和自动 CoT，以克服对人工标注的需求。Fu et al. (2023) 和 Yao et al. (2024) 分别设计了复杂的 CoT 和思维树，以进一步提高 CoT 的复杂性和性能。此外，一些工作利用多代理协作来改善 LLMs 中的推理 (Xiong et al., 2023; Liang et al., 2024; Du et al., 2024)。

至于基于训练后的方法，Sanh et al. (2022) 和 Wei et al. (2022a) 分别提出了 T0 和 FLAN，用于在大量 NLP 任务上训练大型语言模型，从而实现了卓越的零样本任务泛化能力。T_k-INSTRUCT (Wang et al., 2022) 和 Flan-PaLM (Chung et al., 2024) 扩大了任务规模以进一步增强。Orca (Mukherjee et al., 2023) 和 Orca-2 (Mitra et al., 2023) 使用 GPT-4 合成数

据，然后训练小规模的大型语言模型以提高推理能力。Xu et al. (2024) 研究了进化指令以增强大型语言模型的推理能力。Xiong et al. (2024b) 研究了有意义的学习以促进大型语言模型的抽象推理。此外，Xiong et al. (2024a) 诊断并纠正了大型语言模型中的缺陷以实现针对性改进。

我们的工作主要集中在评估现有大型语言模型的推理能力，特别是复杂的常识推理方面。此评估可以为进一步改进提供见解。

5.2 LLM 中的复杂推理

大型语言模型通过预训练和监督微调已取得良好的简单推理性能。大型语言模型在复杂任务上的性能一直是研究领域的重点。

Zhou et al. (2023) 提出了由简至繁的提示法，将复杂问题分解为简单问题，从而实现显著增强。Yao et al. (2023) 设计了 ReAct 教授 LLMs 使用工具支持推理。最近随着 o1、o3 和 DeepSeek-R1 (Liu et al., 2024) 的发布，许多工作开始教授 LLMs 慢思考能力以解决复杂任务。OpenO1 (Team, 2024) 使用反思数据教授 LLMs 反思错误。Zhao et al. (2024) 和 Jiang et al. (2024) 分别基于蒙特卡洛树搜索 (MCTS) 设计了 Marcoo1 和 STILL 来提炼慢思考数据并训练 LLMs。Yang et al. (2024) 构建了 QwQ-32-Preview 来提升 LLMs 的复杂数学推理能力。

我们的工作主要关注常识领域的复杂推理，而之前的工作主要关注数学领域或一般领域。我们的工作与他们的工作相辅相成。

在本文中，我们提议构建一个名为 Com² 的复杂常识推理基准，以满足用户需求。Com² 是基于因果事件图合成的，并由因果理论指导。此后，我们使用 Com² 来评估广泛存在的 LLM，并发现现有 LLM 在推理深度和广度上仍存在一些问題，给当前的 LLM 带来了挑战。进一步的分析表明，通过后期训练和慢思考来装备 LLM，可以缓解上述问题。

我们感谢匿名审稿人的建设性意见。这篇文章的研究得到了中国新一代人工智能 (2024YFE0203700)，国家自然科学基金的资助 (项目编号 U22B2059 和 62176079)，以及 CCF-知图大模型创新基金 (编号 CCF-Zhipu202401) 的支持。

本文仍然存在一些局限性。首先，数据集合成过程可以通过更细粒度和逐步的因果事件图指导来获益。其次，问题可以是开放式的，可以通过基于大型语言模型的评估器来评估。最后，Com²-hard 的构建可以通过更稳健的因果图技术应用来优化，并对正在发生的因果关系进行更深入和更全面的分析。这将有助于更清晰地理解影响 Com²-hard 的因素，从而得出更准确和更有见地的结果。

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- AIME. Aime problems and solutions.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. 2019. Elg: an event logic graph. *arXiv preprint arXiv:1907.08015*.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. Excar: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.
- Adam Hammond and Simon Stern. 2022. The birth of the modern detective story (bmds) dataset.
- Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3023–3030.
- Jinhao Jiang, Zhipeng Chen, Yingqian Min, Jie Chen, Xiaoxue Cheng, Jiapeng Wang, Yiru Tang, Haoxiang Sun, Jia Deng, Wayne Xin Zhao, Zheng Liu, Dong Yan, Jian Xie, Zhongyuan Wang, and Ji-Rong Wen. 2024. Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.
- Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.
- OpenAI. 2024. Openai o1 system card.
- Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*.
- Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, et al. 2025. Codeelo: Benchmarking competition-level code

- generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Open-O1 Team. 2024. Open-o1: A model for matching proprietary openai o1’s power with open-source innovation. Accessed: 2025-02-08.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. 2024. A comparative study on reasoning patterns of openai’s o1 model. *arXiv preprint arXiv:2410.13639*.
- Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7572–7590.
- Kai Xiong, Xiao Ding, Li Du, Jiahao Ying, Ting Liu, Bing Qin, and Yixin Cao. 2024a. Diagnosing and remedying knowledge deficiencies in llms via label-free curricular meaningful learning. *arXiv preprint arXiv:2408.11431*.
- Kai Xiong, Xiao Ding, Zhongyang Li, Li Du, Ting Liu, Bing Qin, Yi Zheng, and Baoxing Huai. 2022. Reco: Reliable causal chain reasoning via structural causal recurrent neural networks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6426–6438.
- Kai Xiong, Xiao Ding, Ting Liu, Bing Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Yixin Cao. 2024b. Meaningful learning: Enhancing abstract reasoning in large language models via generic fact guidance. In *Advances in Neural Information Processing Systems*, volume 37, pages 120501–120525.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.
- Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. 2024. Intuitive or dependent? investigating llms’ behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4221–4246.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.
- Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A 生成²的提示

Event Proposal

You are given some example events, you should follow the pattern of the example events to create lots of new events. Example Events:

{ }

You can create new events in any areas (such as science, math, economy, health, daily, sports, art, etc.) and scenarios. Just give the created new events and create as many events as possible. You format should be like:

- 1.
- 2.
- 3.
-

Causal Chain Proposal: Direct

You are an expert in causality. Please create a causal chains based on the given event. The following are the requirements:

1. the causal chain contains 5 events and 4 causal relationships.
2. the causal chain should be common in real life.
3. the causal chain should have a proper time gap to make users cannot infer tail event just based on the given event.

Here are the given event:

{ }

You should follow the format like:

{ } -> EVENT -> EVENT -> EVENT -> EVENT

Causal Chain Proposal: Intervention

You are an expert in causality. Please create two causal chains based on and start from the given event. The following are the requirements:

1. each causal chain contains 5 events and 4 causal relationships, each chain should be reasonable.
2. the first causal chain should have a high probability to happen, which means it is common in real life.
3. the second causal chain should have a lower probability to happend than the first chain.
4. each causal chain should has a proper time gap to make users cannot infer tail event just based on the given event.
5. the tail event should be the same.

Here are the given event:

{ }

You should follow the format like:

High Probability: -> EVENT1 -> EVENT2 -> EVENT3 -> SAME EVENT

Low Probability: -> EVENT4 -> EVENT5 -> EVENT6 -> SAME EVENT

Causal Chain Proposal: Counterfactual

You are an expert in causality. Please create two causal chains based on the given event. The following are the requirements:

1. each causal chain contains 5 events and 4 causal relationships.
2. the first causal chain should be common in real life.
3. the second causal chain should be a counterfactual scenario (science fiction or an event which is unable to intervene) of the first causal chain.
4. each causal chain should has a proper time gap to make users cannot infer tail event just based on the given event.

Here are the given event:

{ }

You should follow the format like:

Normal: { } -> EVENT -> EVENT -> EVENT -> EVENT

Counterfactual: Counterfactual_of_given_event -> EVENT -> EVENT -> EVENT -> EVENT

Causal Chain Proposal: Decision

You are an expert in causality. Please create two causal chains based on and start from the given event. The following are the requirements:

1. each causal chain contains 5 events and 4 causal relationships, each chain should be reasonable.
2. each chain would lead to bad outcomes or something unacceptable.
3. the first causal chain should have a high probability to happen, which means it is common in real life.
4. the second causal chain should have a much lower probability to happend than the first chain.
5. each causal chain should has a proper time gap to make users cannot infer tail event just based on the given event.

Here are the given event:

{ } You should follow the format like:

High Probability: { } -> EVENT -> EVENT -> EVENT -> EVENT

Low Probability: { } -> EVENT -> EVENT -> EVENT -> EVENT

Causal Chain Proposal: Transition

You are an expert in causality. Please create three causal chains based on and start from the given event. The following are the requirements:

1. each causal chain contains 5 events and 4 causal relationships.
2. the first causal chain should be reasonable without any causal transitive problems.
3. the second causal chain should contain a scene drift problem.
4. the third causal chain should contain a threshold effect problem.
5. each causal chain should have a proper time gap to make users cannot infer tail event just based on the given event.

Take a causal chain $A \rightarrow B \rightarrow C$ for example, here are the reference of the reason for the causal transitive problem:

1. scene drift: $A \rightarrow B$ and $B \rightarrow C$ would not happen within the same specific scene.
2. threshold effect: the influence of A on B is not enough for B to cause C .

Here are the given event:

{ }

Just give the chains only. You should follow the format like:

Normal: \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT

With Scene Drift: \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT

With Threshold Effect: \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT \rightarrow EVENT

Causal Graph Proposal and Com² Synthesis: Direct

After understanding the causal chain, we need you to create a multiple-choice example based on the causal chain with a slow thinking process. The following are the requirements:

1. the example should contain a question, a slow thinking process, some options, and an answer.
2. note that the causal chain is not a part of the final example, it is just used to help you design the example.
3. in the question, you should create a suitable question. The intermediate events in the causal chains cannot appear in the question. The question is asking for a most plausible outcome.
4. the final answer should be the tail event of the causal chain, and the other wrong options should be deceptive.
5. based on the causal chain, you should create a slowing thinking process, which consists of several actions (selecting from the following actions), actions can be repeated:

-Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.

-Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.

-Comparison: If the question is a multiple-choice question, you should compare the differences among all options in detail based on the question.

-Divide: Break down a complex causal-related problem into subproblems.

-Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not be given in the question).

-Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors. Using "wait" to start this.

-Context Identification: For some datasets requiring additional information input, you first summarize different aspects of the context related to the question, and then gives the response for the corresponding question.

-Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.

6. the whole slow thinking process should simulate first-person thinking.
7. do not explicitly demonstrate which chain you are using, all the chains are used to guide the slowing thinking process, it will not be given to help with question answering.
8. the created question should be as concise as possible, and the slow thinking process should be as detailed and complex as possible.
9. you should make a **mistake** first, and conduct self-refinement to backtrack and reason correctly by the causal chain.
10. the output format should be: Question: ____\n\n Options: ____\n\n Slow Thinking Process: ____\n\n Answer: ____.

Causal Graph Proposal and Com² Synthesis: Intervention

After understanding the causal chains, we need you to create a multiple-choice example. The following are the requirements:

1. the example should contain a question, a slow thinking process, some options, and an answer.
2. note that the causal chains are not a part of the final example, they are just used to help you design the example.
3. the intermediate events in each causal chain cannot appear in the question.
4. you should conduct an intervention on the first causal chain to achieve the following goals:
 - in the first causal chain, the transit to the tail event is interrupted by the intervention.
 - in the second causal chain, the transit is NOT interrupted by the intervention.
4. the question can only contain the first events of the causal chains and the intervention event.
5. the final answer should be the tail event of the second causal chain, and the other options should be wrong.
6. based on the causal chains, you should create a slowing thinking process, which consists of several actions (selecting from the following actions), actions can be repeated:
 - Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.
 - Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.
 - Comparison: If the question is a multiple-choice question, you should compare the differences among all options in detail based on the question.
 - Divide: Break down a complex causal-related problem into subproblems.
 - Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not given in the question).
 - Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors. For example, conduct reflection on the intervened position and continue the reasoning via another chain.
 - Context Identification: For some datasets requiring additional information input, you first summarizes different aspects of the context related to the question, and then gives the response for the corresponding question.
 - Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.
7. the whole slow thinking process should simulate first-person thinking.
8. the created question should be as concise as possible, and the slow thinking process should be as detailed and complex as possible.
9. the output format should be: Question: __\n\n Options: __\n\n Slow Thinking Process: __\n\n Answer: __.

Causal Graph Proposal and Com² Synthesis: Counterfactual

After understanding the causal chains, we need you to create a multiple-choice example. The following are the requirements:

1. the example should contain a question, a slow thinking process, some options, and an answer.
2. in the question, the intermediate events in each causal chain cannot appear in the question.
3. note that the causal chains are not a part of the final example, they are just used to help you design the example.
4. you should use the counterfactual chain to create a counterfactual question (e.g. If A leads to B not C, what will happen?).
5. the final answer should be the tail event of the counterfactual chain, and the other wrong options should be deceptive.
6. based on the causal chains, you should simulate first-person thinking and create a slowing thinking process, which consists of several actions (selecting from the following actions), actions can be repeated:
 - Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.
 - Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.
 - Comparison: If the question is a multiple-choice question, you should compare the differences among all options in detail based on the question. Moreover, you should compare the real condition to the counterfactual to obtain detailed analysis.
 - Divide: Break down a complex causal-related problem into subproblems.
 - Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not given in the question).
 - Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors. Using "wait" to start this.
 - Context Identification: For some datasets requiring additional information input, you first summarizes different aspects of the context related to the question, and then gives the response for the corresponding question.
 - Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.
7. the created question should be as concise as possible, and the slow thinking process should be as detailed and complex as possible.
8. the counterfactual must be conducted in the question, which cannot appear in the thinking process.
9. you should make a **mistake** first, and conduct self-refinement to backtrack and reason correctly.
10. the output format should be: Question: __\n\n Options: __\n\n Slow Thinking Process: __\n\n Answer: __.

Causal Graph Proposal and Com² Synthesis: Decision

After understanding the causal chains carefully, we need you to create a multiple-select example. The following are the requirements:

1. the example should contain a question, a slow thinking process, several options, and answers.
2. note that the causal chains are not a part of the final example, they are just used to help you design the example.
3. the question mainly investigates how to severe results, which means what interventions we can do to break the transition to bad results.
4. the answer choices should be interventions applied to the intermediate events of the given causal chains.
5. in the options, besides the answers choices, you should provide several distractors, which are wrong but deceptive.
6. the intermediate events in the causal chains cannot appear in the question.
7. based on the causal chains and question, you should create a slowing thinking process, which consists of several actions (selecting from the following actions), actions can be repeated:
 - Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.
 - Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.
 - Comparison: If the question is a multiple-select question, you should analysis each option in detail based on the question.
 - Divide: Break down a complex causal-related problem into subproblems.
 - Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not given in the question).
 - Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors.
 - Context Identification: For some datasets requiring additional information input, you first summarizes different aspects of the context related to the question, and then gives the response for the corresponding question.
 - Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.
8. the whole slow thinking process should simulate first-person thinking.
9. the created question should be as concise as possible, while the slow thinking process should be as detailed and complex as possible.
10. the output format should be: Question: __\n\n Options: __\n\n Slow Thinking Process: __\n\n Answer: __.

Causal Graph Proposal and Com² Synthesis: Transition

After understanding the causal chains and the scenario, we need you to create a multiple-choice example. The following are the requirements:

1. the example should contain a question, a slow thinking process, some options, and an answer.
2. note that the causal chains and scenario are not a part of the final example, they are just used to help you design the example.
3. the intermediate events in each causal chain cannot appear in the question.
4. the second has the causal transitive problem of , take $A \rightarrow B \rightarrow C$ for example, here are the definitions:
 - scene drift: $A \rightarrow B$ and $B \rightarrow C$ would not happen within the same specific scene.
 - threshold effect: the influence of A on B is not enough for B to cause C.
4. the question can only contain the first events of the causal chains.
5. the final answer should be the tail event of the first causal chain, and the other options should be wrong.
6. based on the causal chains and scenario, you should create a slowing thinking process, which consists of several actions (selecting from the following actions), actions can be repeated:
 - Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.
 - Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.
 - Comparison: If the question is a multiple-choice question, you should compare the differences among all options in detail based on the question.
 - Divide: Break down a complex causal-related problem into subproblems.
 - Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not given in the question).
 - Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors. Using "wait" to start this.
 - Context Identification: For some datasets requiring additional information input, you first summarizes different aspects of the context related to the question, and then gives the response for the corresponding question.
 - Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.
7. you should conduct reasoning via the second causal chain to make a **mistake** and conduct self-refinement to backtrack and reason by the second causal chain.
8. the whole slow thinking process should simulate first-person thinking.
9. the created question should be as concise as possible, and the slow thinking process should be as detailed and complex as possible.
10. the output format should be: Question: __\n\n Options: __\n\n Slow Thinking Process: __\n\n Answer: __.

Detective Proposal and Com² Synthesis: Intervention

criminal

{ } \n——\nWho is(are) the criminal(s)? Just output the name(s).

puzzle

This is a detective story, can you generate a puzzle based on the story, which contains a statement (a short paragraph) of the crime, all the initial facts and people's actions found by police and detectives (not inferred clues, and should exclude { }), which means the police and detectives can infer the criminals just based on the facts and actions rather than the story. Please give a puzzle to ask who is (are) the criminal(s). Some suspects. Please follow the format like: Crime statement: __\n Facts and actions: __\n Puzzle: __\n Suspects: (indexed by A), B), C) and so on, list all suspects, should include { } . Do not provide answer. The last option in the suspects should be "None of the above".

Slow Thinking

Please follow the facts and actions only to give a step-by-step investigation to describe a not guilty person as the criminal (means you should make a mistake first). And then you are aware of an essential clue that " { } ", you start to reflect your mistakes, finally, you got the right answer by backtracking and reasoning. The response should end with "Answer: (the index + the option content, such as A) Tom)". The answer MUST be " { } ".

Detective Proposal and Com² Synthesis: Counterfactual

criminal

{ } \n——\nWho is(are) the criminal(s)? Just output the name(s).

puzzle

This is a detective story, please generate a question based on the story, which contains a statement (a short paragraph) of the crime, all the initial facts and people's actions found by police and detectives (not inferred clues), which means the police and detectives can infer the criminals just based on the facts and actions rather than the story. Please give a multi-choice **counterfactual** question, A slowing thinking process which consists of several actions (selecting from the following actions), actions can be repeated, the whole process should use detailed information in the statement, facts and actions for detailed analysis:

-Systematic Analysis: Starting from the overall structure of the problem, first analyze the inputs and outputs, as well as the constraints, and then decide on the choice of algorithm and the use of data structures.

-Method Reuse: For some problems that can be transformed into classic problems (such as the shortest path or knapsack problem), you can quickly reuse existing methods to solve them.

-Comparison: If the question is a multiple-select question, you should analysis each option in detail based on the question.

-Divide: Break down a complex causal-related problem into subproblems.

-Conquer: Constructs the overall solution by repeatedly solving each subproblem step-by-step with the guidance of causal chain (note that it will not given in the question).

-Self-Refinement: Assessing your reasoning process during inference to determine if there are any issues and correct any errors.

-Context Identification: For some datasets requiring additional information input, you first summarizes different aspects of the context related to the question, and then gives the response for the corresponding question.

-Emphasizing Constraints: For some datasets with constraints on the generated text, emphasizes the corresponding constraints during the reasoning process.

Please follow the format like: Crime statement: __\nFacts and actions: __\nQuestion: __\nOptions: (indexed by A), B),C) and so on, include some wrong options)\nSlow Thinking Process: __\nAnswer:__(index + option text).

我们选择了三位在因果推断和常识推理方面具有良好背景的注释者。我们在 Com²-main 和 Com²-hard 中随机分别抽取了 100 个不同任务的例子进行人工评估。因此，我们在这 200 个例子上评估慢速思考。对于每位注释者，我们支付每小时 \$ 10，而在我们国家，最低工资不足每小时 \$ 5。所有注释者都同意让我们使用他们的注释。评估协议在 Com²-main、Com²-hard 和慢速思考的评估中分别是 94 %、91 % 和 90 %。

Direct in Com²-main

Question

If Sarah had ignored her symptoms completely, what would eventually happen?

Options

A) She would feel better after a short period. \nB) She would schedule a follow-up appointment. \nC) She would experience a sudden health crisis. \nD) She would receive a prescription for pain relief.

Correct Answer

C) She would experience a sudden health crisis.

Slow Thinking

1. Context Identification: I need to analyze the situation where Sarah ignored her symptoms. Ignoring health issues typically leads to worsening conditions. I need to determine the long-term consequences of ignoring symptoms rather than seeking help.

2. Systematic Analysis: Let's break down the potential outcomes of ignoring symptoms. One possibility is that she could start feeling better due to the body's natural healing process. Another option could be that she decides to just cope with the pain without any medical intervention.

3. Comparison: I'll compare all the options. Option A suggests she would feel better, which often isn't realistic with persistent symptoms. Option B indicates she would schedule a follow-up appointment, but if she's ignoring the symptoms, this doesn't align with her behavior. Option D implies she might seek relief through medication, which contradicts the act of ignoring symptoms entirely.

4. Divide: Breaking down the outcomes, I realize that not addressing health issues often leads to escalation. If she continues to ignore symptoms, it's likely her situation will worsen instead of improving.

5. Conquer: I need to construct a logical conclusion about the most serious possibility. By ignoring the symptoms, one possible severe consequence is a sudden health crisis necessitating emergency intervention.

6. Self-Refinement: Wait, I almost concluded that she might just feel better (Option A), but that's an optimistic assumption. Considering realistic scenarios, ignoring serious symptoms often leads to drastic health declines rather than recovery. I need to reassess.

7. Final Assessment: Conclusively, experiencing a sudden health crisis is the most likely negative outcome of ignoring persistent health issues.

Decision in Com² -main

Question

What interventions can prevent severe consequences during a hot air ballooning experience that could lead to disastrous outcomes?

Options

A. Regular equipment checks before the flight \nB. Ignoring weather warnings and proceeding with the flight \nC. Ensuring pilots undergo thorough training for extreme conditions \nD. Allowing passengers to control the balloon \nE. Establishing no-fly zones in higher-density airspace \nF. Offering balloon rides regardless of adverse weather conditions

Correct Answer

A, C, E.

Slow Thinking

1. Context Identification: I first recognize that the question revolves around potential interventions that could mitigate severe negative outcomes during a hot air ballooning experience. I need to focus on preventing issues related to equipment failure and extreme weather.
2. Systematic Analysis: I review each option in light of their potential to prevent accidents. Option A refers to equipment checks and could help avert malfunctions, while Option C emphasizes pilot training during severe conditions.
3. Comparison: I analyze options B and F. Both suggest ignoring critical safety precautions and could lead to disastrous outcomes. Option D seems reckless, as it suggests letting passengers take control without proper training.
4. Divide: Next, I break down the options into two groups: safety interventions and unsafe practices. Grouping this way highlights which options can truly help prevent severe consequences.
5. Conquer: I take the valid interventions from the safe group (A, C, E) and recognize their importance in preventing negative outcomes associated with the hazards of hot air ballooning.
6. Self-Refinement: I reassess my thought process, making certain that I haven't overlooked any relevant options or context, confirming that my focus remains on preventing equipment failure and addressing pilot preparedness.
7. Emphasizing Constraints: I remember that the question explicitly seeks interventions that will reduce severe outcomes, which validates the importance of option E about establishing no-fly zones.
8. Method Reuse: I recall existing safety protocols in aviation, reaffirming that rigorous checks and pilot training are part of best practices in preventing accidents.

Transition in Com² -main

Question

What is the first step to build a bookshelf from scratch?

Options

A) Paint or stain the finished bookshelf \nB) Gather materials and tools \nC) Attend a woodworking workshop \nD) Share experiences on social media

Correct Answer

B) Gather materials and tools.

Slow Thinking

- I start by identifying the question, which asks for the first step in building a bookshelf from scratch. This is a straightforward task, but I need to ensure I analyze the options carefully.
- Comparison: I look at the options provided. Option A mentions painting or staining, which clearly cannot be the first step since it comes after the bookshelf is constructed. Option C involves attending a workshop, which is not directly related to the act of building the bookshelf itself. Option D talks about sharing experiences, which is also a post-construction activity.
 - Divide: I break down the remaining options. Option B, Gather materials and tools, seems to be the only logical first step.
 - Context Identification: I think about the context of building a bookshelf. The process typically starts with gathering the necessary items before any construction can begin.
 - Self-Refinement: I pause to assess my reasoning. I realize that I might have overlooked the possibility of needing to plan the design first, but that would still require materials and tools to execute the plan.
 - Conquer: I conclude that gathering materials and tools is indeed the first action needed to start the project.
- After this thorough analysis, I am confident that the correct answer is B.

Intervention in Com² -main

Question

A child snaps a twig off a branch, and a nearby animal hears the snap and perceives danger. What happens next?

Options

A) The bird returns to its nest with food for its chicks. \nB) The animal hides in a burrow for safety. \nC) The bird flies away to a nearby tree. \nD) The predator gives up and leaves the area, allowing the animal to emerge safely later.

Correct Answer

B) The animal hides in a burrow for safety.

Slow Thinking

- Systematic Analysis: I start by identifying the initial event, which is the child snapping a twig. This action leads to a reaction from the nearby animal. I need to consider what happens next based on the causal chain.
- Divide: I break down the possible outcomes from the initial event. The animal perceives danger, which suggests it will take some action to protect itself.
- Comparison: I look at the options provided. Option A talks about a bird returning to its nest, which is not related to the animal's response. Option C also involves the bird, which is irrelevant here. Option B directly relates to the animal's instinct to hide, while Option D describes the predator's actions, which could follow the animal's hiding.
- Conquer: I focus on the most logical next step for the animal after perceiving danger. It would likely hide in a burrow for safety, as indicated in Option B.
- Self-Refinement: I reflect on the reasoning process. The intervention in the first causal chain interrupts the bird's actions, so I confirm that the animal's response is the primary focus here.
- Context Identification: I summarize that the child's action leads to the animal's perception of danger, which is crucial for determining the next event.

Counterfactual in Com² -main

Question

If a coin floats away in a zero-gravity environment, what will happen next?

Options

A) The coin lands on a nearby table. \nB) The coin collides with an experimental device. \nC) The coin rolls back into the person's hand. \nD) The coin falls to the ground.

Correct Answer

B) The coin collides with an experimental device.

Slow Thinking

1. Systematic Analysis: I need to understand the scenario where the coin is in a zero-gravity environment. In this case, the typical behavior of the coin changes significantly compared to a normal environment. The inputs are the conditions of zero gravity, and the output will be the consequences of the coin's movement.
2. Comparison: I will compare the options provided. Option A suggests the coin lands on a table, which seems unlikely in zero gravity. Option C implies the coin rolls back into the person's hand, which also doesn't fit the zero-gravity context. Option D states the coin falls to the ground, which contradicts the zero-gravity condition.
3. Divide: I can break down the options into two categories: those that align with zero gravity (B) and those that do not (A, C, D).
4. Conquer: Since options A, C, and D are not plausible in a zero-gravity scenario, I focus on option B, which suggests a collision with an experimental device. This aligns with the idea that the coin would continue to drift until it encounters something.
5. Self-Refinement: Wait, I initially thought option A could be possible, but upon further analysis, I realize that in zero gravity, the coin cannot land on a table. I need to correct my reasoning.
6. Context Identification: The context here is crucial; in a zero-gravity environment, objects do not fall but float. Therefore, the only logical outcome is that the coin would drift and potentially collide with something.

Decision in Com² -hard

Crime

During the wedding breakfast of Lord St. Simon and Lady St. Simon (Hatty Doran), the bride abruptly left the gathering and disappeared. It was later discovered that she had eloped with Francis H. Moulton, a man she had previously been engaged to before her marriage to Lord St. Simon. The bride's actions were influenced by a note slipped to her during the ceremony, which led her to leave her husband and family behind.

Facts

1. Lady St. Simon was seen in good spirits before the wedding but became agitated during the ceremony.
2. A woman named Flora Millar attempted to disrupt the wedding, claiming to have a connection to Lord St. Simon.
3. After the ceremony, Lady St. Simon spoke to her maid, Alice, and mentioned jumping a claim, indicating a prior relationship.
4. Lady St. Simon left the breakfast table shortly after the meal began, dressed in her wedding attire, and was later seen leaving with Francis H. Moulton.
5. The police found Lady St. Simon's wedding dress and belongings in the Serpentine, indicating she had left in a hurry.

Question

What actions could have been taken to prevent Lady St. Simon's disappearance and the subsequent elopement?

Options

A) Conduct a thorough background check on all guests attending the wedding.\nB) Ensure that the bride had a trusted confidante present to discuss any concerns before the ceremony.\nC) Limit the number of guests to only immediate family to avoid distractions.\nD) Provide a security detail to monitor any potential disruptions during the ceremony.\nE) Ignore any disturbances from guests, as they are common at weddings.

Correct Answer

A) Conduct a thorough background check on all guests attending the wedding. \nB) Ensure that the bride had a trusted confidante present to discuss any concerns before the ceremony. \nD) Provide a security detail to monitor any potential disruptions during the ceremony.

Slow Thinking

1. Systematic Analysis: Analyze the situation by identifying the inputs (guests, bride, groom, and their relationships) and outputs (successful wedding vs. elopement). The constraints include the emotional state of the bride and the presence of Flora Millar.
2. Comparison: Evaluate each option based on its potential effectiveness in preventing the crime. Options A, B, and D seem proactive, while C could limit the social aspect, and E is dismissive.
3. Divide: Break down the problem into subproblems: the bride's emotional state, the influence of Flora Millar, and the wedding environment.
4. Conquer: Consider how addressing each subproblem could lead to a solution. For example, ensuring the bride's emotional well-being could prevent her from feeling the need to elope.
5. Self-Refinement: Reflect on the reasoning process to ensure all aspects of the situation are considered, particularly the influence of past relationships.
6. Context Identification: Summarize the context of the wedding, the bride's background, and the presence of Flora Millar to understand the dynamics at play.

Intervention in Com² -hard

Crime

A valuable coronet, containing thirty-nine beryls, was stolen from Alexander Holder's home. During the incident, a portion of the coronet was damaged, and three beryls were missing. The crime occurred late at night while the household was asleep.

Facts

1. Alexander Holder saw his son, Arthur, holding the coronet in his hands when he entered the dressing room.
2. Arthur was dressed only in his shirt and trousers, indicating he had been in bed prior to the incident.
3. Mary Holder, Arthur's cousin, was seen at the side window of the hall just before the incident occurred.
4. Lucy Parr, a maid, was reported to have left the house shortly before the crime and returned through the back door.
5. A note was found from Mary Holder, indicating she had left the house and would not return.
6. Sir George Burnwell had been a frequent visitor to the Holder household and was known to have a questionable reputation.
7. Arthur had previously expressed distress over his financial situation and had asked his father for money the night before the incident.

Question

Based on the facts and actions presented, who is (are) the criminal(s) involved in the theft of the coronet?

Options

A) Arthur Holder \nB) Mary Holder \nC) Lucy Parr \nD) Sir George Burnwell \nE) None of the above

Correct Answer

D) Sir George Burnwell

Slow Thinking

Step 1: Initial Investigation

Upon arriving at the scene of the crime, the police began their investigation by interviewing the household members. Alexander Holder reported that he had seen his son, Arthur, holding the coronet in his hands when he entered the dressing room. This was a critical piece of evidence that immediately cast suspicion on Arthur.

Step 2: Gathering Evidence

The police noted that Arthur was dressed only in his shirt and trousers, suggesting he had been in bed prior to the incident. This raised questions about how he could have accessed the coronet without being noticed. However, the fact that he had previously asked his father for money added to the motive for theft.

Step 3: Examining the Maid

Lucy Parr, the maid, was also questioned. She had left the house shortly before the crime and returned through the back door. Her absence during the time of the theft made her a potential suspect, especially since she had a sweetheart who could have been involved in the crime. The police noted her questionable timing and her potential access to the coronet.

Step 4: Considering Mary Holder

nMary Holder, Arthur's cousin, was seen at the side window of the hall just before the incident occurred. Her presence raised suspicions as well, especially since she had a close relationship with Arthur. The police speculated that she could have been involved in a conspiracy with him to steal the coronet.

Step 5: Focusing on Arthur

With the evidence pointing towards Arthur, the police began to build a case against him. They believed that his financial troubles and the fact that he was caught with the coronet in his hands were enough to conclude that he was the thief. The emotional turmoil he displayed only seemed to confirm his guilt.

Step 6: The Essential Clue

However, as the investigation continued, the police discovered an essential clue: footprints in the snow outside the house. This evidence had been overlooked in the initial investigation. The footprints led away from the scene and indicated that someone else had been present during the crime.

Step 7: Backtracking and Reasoning

Upon examining the footprints, the police realized that they did not match Arthur's bare feet. Instead, they were from a booted man, suggesting that someone else had entered the house. This revelation prompted a reevaluation of the evidence.

The police recalled that Sir George Burnwell had been a frequent visitor to the Holder household and had a questionable reputation. His presence in the area, combined with the footprints, suggested that he could have been the true culprit.

Step 8: Conclusion

By backtracking through the evidence and considering the new clue of the footprints, the police concluded that the initial suspicions against Arthur were misplaced. The true criminal was Sir George Burnwell, who had manipulated the situation to frame Arthur while escaping unnoticed.

Counterfactual in Com² -hard

Crime

John Straker, the trainer of the racehorse Silver Blaze, was found dead on the moors near King's Pyland. He had suffered a severe head injury and a cut on his thigh. The horse Silver Blaze was missing, and there were indications that Straker had been involved in a plot to harm the horse for financial gain.

Facts

1. John Straker was last seen leaving his home to check on the horses late at night.
2. A stable-boy, Ned Hunter, was drugged with opium-laced curried mutton, which was delivered to him by a maid.
3. Straker's overcoat was found on a furze-bush near the scene of the crime.
4. Straker was found dead with a knife in his hand, which was identified as a cataract knife, not suitable for self-defense.
5. A cravat belonging to Fitzroy Simpson was found in Straker's hand, indicating a struggle.
6. Fitzroy Simpson had a motive to harm Silver Blaze, as he had placed large bets against the horse.
7. The police arrested Fitzroy Simpson based on circumstantial evidence, but he had no visible injuries from the struggle.
8. The dog in the stables did not bark during the incident, suggesting the intruder was known to it. Question

Based on the facts and actions presented, who is most likely to be the criminal responsible for John Straker's death and the disappearance of Silver Blaze?

Options

A) John Straker \nB) Fitzroy Simpson \nC) The stable-boy, Ned Hunter \nD) Mrs. Straker

Correct Answer

B) Fitzroy Simpson

Slow Thinking

1. Systematic Analysis: The problem involves determining the criminal based on the actions and facts surrounding the murder of John Straker and the disappearance of Silver Blaze. The inputs are the actions of the individuals involved, and the output is the identification of the criminal.
2. Method Reuse: The situation can be analyzed using classic detective reasoning, focusing on motives, means, and opportunities.
3. Comparison:
 - A) John Straker: He is the victim, so he cannot be the criminal.
 - B) Fitzroy Simpson: He had a motive to harm Silver Blaze and was present at the scene, but he has no injuries and was arrested based on circumstantial evidence.
 - C) The stable-boy, Ned Hunter: He was drugged and incapacitated, making it unlikely he could have committed the crime.
 - D) Mrs. Straker: While she could have had a motive related to Straker's debts, there is no direct evidence linking her to the crime.
4. Divide: The analysis can be divided into examining motives (financial gain), means (the knife and drugging), and opportunity (who was present at the scene).
5. Conquer: By evaluating each suspect's actions and the evidence against them, we can determine who had the most opportunity and motive.
6. Self-Refinement: The reasoning process indicates that Fitzroy Simpson had the strongest motive and opportunity, but the lack of direct evidence against him raises questions.
7. Context Identification: The context involves a horse racing environment where financial stakes are high, and trainers may resort to unethical actions to secure wins.
8. Emphasizing Constraints: The evidence must be strictly based on the actions and facts presented, without inferring beyond what is known.

B 评估提示

Prompt for Com² -main: Direct, Intervention & Transition

You are given a multi-choice question and a set of plausible options. You must choose the most plausible option as the answer based on commonsense. You should give an analysis first and then choose the answer. At the end of your answer, choose answer to the question in format such as 'Answer: B)_____'.

Prompt for Com² -main: Counterfactual

You are given a multi-choice counterfactual question and a set of plausible options. You must choose the most plausible option as the answer. You should give an analysis first and then choose the answer. At the end of your answer, choose answer to the question in format such as 'Answer: B)_____'.

Prompt for Com² -main: Decision

You are given a multi-select question and a set of options. You must choose all the plausible options as the answers to reach the goal of the question. You should give an analysis first and then choose the answer. At the end of your answer, choose answer to the question in format such as 'Answer: A)_____
B)_____'

Prompt for Com² -hard: Intervention

You are given a puzzle about a crime, which include a crime statement, a set of facts, a question, and a set of suspects (options). You should conduct detailed analysis to find the most plausible suspect as the criminal. At the end of your answer, choose answer to the question in format such as 'Answer: A)_____'.

Prompt for Com² -hard: Counterfactual

You are given a statement and some facts of a crime. You should conduct detailed analysis to find the most plausible option as the answer to the question. At the end of your answer, choose answer to the question in format such as 'Answer: A)_____'.

Prompt for Com²-hard: Decision

You are given a statement and some facts of a crime. You should conduct detailed analysis to find all the plausible options as the answers to the multi-select question. At the end of your answer, choose answer to the question in format such as 'Answer: A)_____ B)_____'

我们对大型语言模型（LLM）进行全参数微调。批处理大小设置为 32，学习率为 1e-5。我们使用 2 个 NVIDIA A100 80GB PCIe GPU 训练 LLM 共 3 个轮次。

Cases of Trained LLaMA: Case1

Question

If I watch a meteor shower and then decide to organize a gathering for the next one, what is the likely outcome if I develop a sudden interest in astronomy instead?

Options

A) Conduct a research project on meteor compositions \nB) Deepen friendships through shared experiences \nC) Enroll in an advanced astrophysics course \nD) Publish a paper in an academic journal about findings

Correct Answer

C) Enroll in an advanced astrophysics course.

Ours

- Context Identification: I need to consider the initial event of watching a meteor shower and the intervention of developing a sudden interest in astronomy. This interest could lead me to pursue a different path than the one described in the first causal chain.

- Divide: I will break down the options based on the potential outcomes of my interest in astronomy.

- Comparison:

- Option A (Conduct a research project on meteor compositions) seems plausible as it aligns with the interest in astronomy.

- Option B (Deepen friendships through shared experiences) is related to the first causal chain and does not fit with my new interest.

- Option C (Enroll in an advanced astrophysics course) is also a logical step after developing an interest in astronomy.

- Option D (Publish a paper in an academic journal about findings) is a more advanced outcome that could follow from enrolling in a course.

- Conquer: I will focus on the outcomes that stem from my interest in astronomy. Conducting a research project and enrolling in a course are direct actions that follow from this interest.

- Self-Refinement: I reflect on the intervention and realize that while I may not deepen friendships through shared experiences, I can still pursue academic opportunities related to my interest in astronomy.

Answer: C) Enroll in an advanced astrophysics course.

LLaMA-Vanilla

Analysis: The question describes a scenario where the person develops a sudden interest in astronomy after watching a meteor shower. This interest is likely to be a hobby or a casual interest, rather than a professional or academic pursuit. The person is not described as having any prior experience or expertise in astronomy, and the context suggests that they are simply enthusiastic about the subject.

Given this context, the most plausible outcome is that the person will deepen their understanding and appreciation of astronomy through social interactions and shared experiences with others who share their interest. This is a common outcome for people who develop a new hobby or interest, as they often seek out others who can share their enthusiasm and provide support and guidance.

The other options are less plausible because they require a level of expertise, commitment, or resources that is unlikely for someone with a sudden and casual interest in astronomy. Conducting research on meteor compositions (A) or publishing a paper in an academic journal (D) would require significant time, effort, and expertise, and is unlikely to be a direct outcome of a casual interest in astronomy. Enrolling in an advanced astrophysics course (C) may be a possibility, but it is less likely than deepening friendships through shared experiences, as it requires a significant investment of time and resources.

Answer: B) Deepen friendships through shared experiences

Cases of Trained LLaMA: Case2

Question

How can we prevent severe economic disruptions and civil unrest resulting from abrupt energy policy changes?

Options

A. Maintain and upgrade fossil fuel infrastructure during the transition to renewable energy \nB. Promote gradual shifts in energy policy with continuous public engagement \nC. Immediately ban all fossil fuels without alternative solutions in place \nD. Provide financial support and retraining for workers affected by the transition \nE. Ignore public feedback and proceed with drastic energy policies \nF. Conduct regular assessments of energy needs and production capabilities

Correct Answer

A, B, D, F.

Ours

1. Context Identification: I need to understand the implications of abrupt energy policy changes. Abrupt changes can lead to severe economic disruptions and civil unrest. I should consider the potential consequences of these disruptions.

2. Systematic Analysis: I will analyze the options provided. I need to identify which options could effectively prevent severe economic disruptions and civil unrest.

3. Comparison:

- Option A suggests maintaining and upgrading fossil fuel infrastructure. This could provide a stable energy supply during the transition, which seems beneficial.

- Option B promotes gradual shifts in energy policy with continuous public engagement. This could help build support and reduce unrest, making it a strong candidate.

- Option C immediately banning all fossil fuels without alternatives seems risky and could lead to severe disruptions.

- Option D provides financial support and retraining for workers, which could mitigate job losses and maintain stability.

- Option E ignores public feedback, which could lead to unrest and is not advisable.

- Option F conducting regular assessments of energy needs could help ensure a smooth transition, making it a viable option.

4. Divide: I will categorize the options into two groups: those that could prevent disruptions (A, B, D, F) and those that could exacerbate them (C, E).

5. Conquer: I will focus on the options that could prevent disruptions. Options A, B, D, and F seem to offer proactive measures to mitigate risks.

6. Self-Refinement: I need to ensure that my reasoning aligns with the goal of preventing severe economic disruptions and civil unrest. I will double-check the implications of each option.

7. Emphasizing Constraints: I must remember that the goal is to prevent disruptions, so I should prioritize options that address the immediate and long-term effects of energy policy changes.

Answer: A), B), D), F)

LLaMA-Vanilla

Analysis: To prevent severe economic disruptions and civil unrest resulting from abrupt energy policy changes, it's essential to consider the potential impacts on various stakeholders, including workers, communities, and the economy as a whole. A gradual transition to renewable energy, public engagement, and support for affected workers are crucial elements in mitigating potential negative consequences. Ignoring public feedback and proceeding with drastic policies without a plan in place would likely exacerbate the situation.

Answer: B) Promote gradual shifts in energy policy with continuous public engagement D) Provide financial support and retraining for workers affected by the transition F) Conduct regular assessments of energy needs and production capabilities