

---

# 通过奖励加权微调的强化学习进行澄清学习

---

Subhojyoti Mukherjee, Viet Dac Lai, Raghavendra Addanki, Ryan Rossi  
Adobe Research  
subhomuk@adobe.com

Seunghyun Yoon, Trung Bui, Anup Rao, Jayakumar Subramanian, Branislav Kveton  
Adobe Research

## Abstract

问答 (QA) 代理能自动回答用自然语言提出的问题。在这项工作中，我们学习在 QA 代理中提出澄清性问题。我们方法的关键思想是模拟包含澄清性问题的对话，并通过强化学习 (RL) 从中学习。为了使强化学习更具实用性，我们提出并分析了可视为奖励加权的监督微调 (SFT) 并可在大型语言模型中轻松优化的离线强化学习目标。我们的工作与最近提出的方法形成鲜明对比，这些方法基于 SFT 和直接偏好优化，具有额外的超参数且不直接优化奖励。我们通过实验证明与这些方法相比，在优化的奖励和语言质量上都有所提升。

## 1 介绍

问答 (QA) 是自然语言处理 (NLP) 与信息检索交叉领域，致力于构建能够回答自然语言问题的系统。大语言模型 (LLMs) 的出现 [49, 8, 6] 引起了对此领域的重大兴趣，并提出了许多不同的方法。早期关于 QA 的研究广泛探索了开放书 [51, 11, 32] 和闭合书 [52, 7] 设置中的单轮问答。最近的研究侧重于复杂问答 [44, 34]，包括交互问答和回答模糊问题。例如，Lu et al. [40] 使用 LLM 作为基础学习了一个分类器，以预测是否应该提出澄清问题，Hahn et al. [23] 提出了一种基于不确定性的自适应方法，通过提出澄清问题改进多轮文本到图像生成，Kobalczuk et al. [31] 基于信息增益选择了澄清问题，利用预训练的 LLM，并且，Andukuri et al. [2] 将这一问题表述为多步骤优化，并通过监督微调 (SFT) 来解决。

我们还将这个问题表述为多步优化，并使用强化学习 (RL) [60] 来学习提出澄清性问题。最相关的工作是 Andukuri et al. [2] 和 Chen et al. [13]，它们使用 RL 从模拟的用户代理对话中学习澄清性问题。虽然我们采用相同的学习范式，但我们的工作在使用方式上有所不同。Andukuri et al. [2] 选择最有奖励的轨迹并在其上进行调整。Chen et al. [13] 为对话的每一步生成替代响应，然后使用 DPO [50] 优化以获得更好的响应。这些方法的主要限制是它们没有充分利用奖励信号；它们只是用它来将原始问题转变成 SFT 或 DPO 问题。我们直接使用 RL 优化奖励。我们在实验中 Section 4 观察到比 SFT 和 DPO 更大的收益，因为两者都可以被视为奖励阈值，这与使用实际奖励相比会导致信息损失。

我们使用离线强化学习 [33, 78, 35] 来解决我们的问题，并做出两项技术贡献以使其具有可行性。首先，我们的离线 RL 问题不涉及倾向评分比率，例如在 PPO [56] 和 GRPO [57] 中；事实上相当于加权 SFT。因此，可以在任何 LLM 中使用标准训练原语轻松解决。我们通过修改 TRL 中的 SFT [64] 实现了我们的解决方案。其次，我们提出了一种基于标准化轨迹奖励的方差减小技术，使用多个采样轨迹。这降低了策略优化中的方差，并可能导致学习出更好的策略。我们在实验中观察到了这种效果 (Section 4)。

我们做出以下贡献：

1. 我们将学习提出澄清问题的问题表述为对话优化的强化学习。我们对奖励没有做任何强烈假设。我们的设置涵盖了固定时长的对话和自适应的对话，其中代理可以提前结束对话，因为已经收集到了足够的信息来回答问题。
2. 我们推导出一个离线 RL 目标，它是原始目标的下界。因此，通过最大化这个新目标可以优化原始目标。新目标等价于加权 SFT，因此可以在 LLM 中使用标准 SFT 训练原语轻松优化。权重是基于标记序列的，与以前工作的单个标记不同 [53, 21, 73, 77]，并且我们也避免了倾向评分 [56, 57] 的比率。
3. 我们推导了一个具有标准化奖励的离线 RL 目标。标准化奖励的优点是可以降低策略优化的方差，这通常会导致更好的策略。我们在 Section 4 中对此进行了实证展示。
4. 我们在多个 QA 数据集上对我们的方法进行了全面评估，这些数据集涵盖了开卷考试、科学主题的文本信息、对话式 text-to-SQL 数据集以及数学对话和问题解决。尽管我们优化了单一的奖励，但我们在推理能力、教学价值和信心等所有其他指标上都观察到了改进。我们考虑了五个基线：两种 SFT 变体、两种 DPO 变体和原始策略。我们观察到，相对于 SFT 和 DPO，我们取得了重大进展，因为我们使用 RL 直接优化奖励信号。
5. 对于每个 QA 基准，我们生成一个由 500 个多轮对话组成的丰富数据集。我们计划将代码和数据集公开，以鼓励更多在学习提出澄清性问题上工作。

本文的组织结构如下。我们在 Section 2 中将学习提出澄清问题的问题表述为一个强化学习问题。在 Section 3 中，我们制定了我们问题的离线变体，并展示如何使用加权 SFT 优化它们。我们在 Section 4 中报告我们的结果，并在 Section 5 中讨论相关工作。我们在 Section 6 中得出结论。

## 2 设置

我们首先介绍我们的符号。在概率测度  $p$  下，记边际概率和条件概率分别为  $p(X = x)$  和  $p(X = x | Y = y)$ ；当随机变量从上下文  $x$  中明确时，写作  $p(x)$  和  $p(x | y)$ 。指标函数为  $\mathbb{1}\{\cdot\}$ 。对于正整数  $n$ ，我们定义  $[n] = \{1, \dots, n\}$ 。向量  $v$  的第  $i$  个条目是  $v_i$ 。如果向量已经有索引，比如  $v_j$ ，我们写作  $v_{j,i}$ 。

学习提出澄清性问题的任务被视为一个通用的强化学习 (RL) 问题 [60]，其中一个代理与用户互动。代理向用户提问，用户以回答响应。当对话结束时，会被分配一个奖励。奖励衡量对话的质量，代理的目标是最大化奖励。

我们将问题形式化如下。智能体首先观察到上下文  $x \in \mathcal{S}$ ，其中  $\mathcal{S}$  是令牌序列的空间。上下文定义了任务。智能体与用户之间的对话由索引为  $t \in \mathbb{N}$  的步骤组成，其中  $\mathbb{N}$  是正整数集。在步骤  $t$  中，智能体采取行动  $a_t \in \mathcal{S}$  并观察到  $y_t \in \mathcal{S}$ 。动作  $a_t$  是一个问题，观察  $y_t$  是用户的响应。对话在  $n$  步之后结束，我们用对话中所有动作和观察的轨迹  $\tau_n = (a_1, y_1, \dots, a_n, y_n)$  来表示它。步骤数量  $n$  可以是固定的或随机的。当它是随机的时，可以是对话历史的任何函数。奖励衡量对话的质量，是  $x$  和  $\tau_n$ ， $r(x, \tau_n) \geq 0$  的非负函数。我们不对奖励做任何额外假设，例如它在各个独立步骤上分解。这样做是为了保持通用性，并且因为我们的算法 (Section 3) 不需要它。

智能体遵循一个基于对话历史的策略。具体来说，在环境  $x$  和历史  $\tau_{t-1}$  中选择动作  $a$  的概率为  $\pi(a | x, \tau_{t-1}; \theta)$ ，并由  $\theta \in \Theta$  参数化。我们称  $\theta$  为策略，称  $\Theta$  为策略参数空间。观察到的  $y_t$  条件于对话历史  $\tau_{t-1}$  和动作  $a_t$  的概率记为  $p(y_t | x, \tau_{t-1}, a_t)$ 。我们稍微滥用符号，表示策略  $\theta$  在环境  $x$  下的轨迹  $\tau_n$  的概率为

$$\pi(\tau_n | x; \theta) = \prod_{t=1}^n p(y_t | x, \tau_{t-1}, a_t) \pi(a_t | x, \tau_{t-1}; \theta). \quad (1)$$

。该因式分解源于概率的链式规则。令

$$V(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n)] \quad (2)$$

为策略  $\theta$  的期望值，其中  $q$  是在环境  $x$  上的分布。我们的目标是最大化期望策略值， $\theta_* = \arg \max_{\theta \in \Theta} V(\theta)$ 。

我们的框架足够通用，可以用于建模多种使用场景。例如，假设我们想要在  $n$  步骤内最大化对话的教学价值 [55]。那么， $r(x, \tau_n)$  将是  $\tau_n$  在  $n$  步骤内的聚合教学价值。作为另一个

例子，假设我们想要通过提问  $n$  问题来学习澄清一个模糊的问题 [2, 13]。那么， $r(x, \tau_n)$  将是基于  $x$  和  $\tau_n$  的条件下生成的答案的质量。最后，假设澄清问题的数量是由智能体在搜集到足够的信息后自适应选择的 [31]。那么， $r(x, \tau_n)$  将是生成的答案质量乘以  $\gamma^n$ ，其中  $\gamma \in (0, 1)$  是折扣因子。更多步骤的惩罚  $n$  是至关重要的。否则，智能体可能会无限制地提问澄清问题，因为每个答案都会增加正确回答原始问题的概率。在这种情况下，澄清问题的数量  $n$  是随机的，并由智能体决定。

### 3 算法

我们的目标是在 (2) 中最大化期望策略值  $V(\theta)$ 。这可以通过多种方式完成 [60]。对于像由 LLMs 表示的复杂策略来说，最自然的方法是策略梯度 [69]。策略梯度的关键思想是通过梯度上升迭代地更新策略  $\theta$ 。在  $\theta$  处， $V(\theta)$  的梯度是

$$\nabla V(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n) \nabla \log \pi(\tau_n | x; \theta)]$$

，可以通过直接应用打分恒等式 [1] 来推导。计算这个梯度具有挑战性有两个原因。首先，由于轨迹  $\tau_n$  是在优化的策略  $\theta$  下采样的，因此每当  $\theta$  更新时，必须在每一步梯度上升时重新采样。其次，需要一个奖励模型  $r(x, \tau_n)$  来评估任何可能采样到的轨迹。

为了应对这些挑战，我们求助于离线强化学习 [33, 78, 35]。离线强化学习的关键思想是先收集轨迹-奖励元组的数据集，然后在其上优化策略，这类似于经典监督学习中学习分类器。我们用  $\pi_0$  表示数据记录策略，用  $\pi_0(\tau_n | x)$  表示在情境  $x$  下使用策略  $\pi_0$  生成轨迹  $\tau_n$  的概率。控制 [18] 和统计学 [27] 中的一个经典结果是倾向得分

$$V(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n)] = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} r(x, \tau_n) \right], \quad (3)$$

可以纠正已记录数据集中的选择偏差。简而言之，优化 (2) 等价于在由另一个策略  $\pi_0$  收集的轨迹数据集上最大化倾向加权的奖励。优化 (3) 的主要挑战在于倾向得分的比率可能很高。这可以通过在令牌级别截断 [29] 来解决，这是 PPO [56] 和 GRPO [57] 中的关键思想。在 Section 3.1 的结尾处，我们讨论了与这些方法的不同。在下一节中，我们概述我们用于离线强化学习的奖励加权微调方法。

#### 3.1 奖励加权微调

我们工作的核心思想是最大化 (2) 的下界。虽然这个界限只有在  $\pi(\cdot | \cdot; \theta) \equiv \pi_0$  时才是紧的，但它引导我们设计出一个实用的离线 RL 算法，可以作为加权 SFT 实现，而不引入倾向得分的比率。我们基于 Liang and Vlassis [37] 中的下界，并将其扩展到离线 RL。

**Lemma 1.** *For any policies  $\pi$  and  $\pi_0$ , and any non-negative reward function,*

$$\mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n)] \geq \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [r(x, \tau_n) \log \pi(\tau_n | x; \theta)] + C_1,$$

where  $C_1 = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [r(x, \tau_n)(1 - \log \pi_0(\tau_n | x))] \geq 0$  is a constant independent of  $\theta$ .

*Proof.* 利用基本代数，

$$\begin{aligned} \mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n)] &= \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ r(x, \tau_n) \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right] \\ &\geq \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ r(x, \tau_n) \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right] \\ &= \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [r(x, \tau_n) \log \pi(\tau_n | x; \theta)] + C_1. \end{aligned}$$

不等式由  $u \geq 1 + \log u$  和非负奖励得出。证明至此结束。  $\square$

由于我们对可能较大的  $u$  应用  $u \geq 1 + \log u$ ，该界在实际中预计会比较宽松。Lemma 1 的结果是

$$J(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [r(x, \tau_n) \log \pi(\tau_n | x; \theta)] \quad (4)$$

是 (2) 的下界。当  $\pi(\cdot | \cdot; \theta) \equiv \pi_0$  时, (4) 等于 (2), 因此提升 (4) 的策略也改善了 (2)。接下来我们展示 (4) 等价于奖励加权的 SFT。为此, 我们将轨迹概率 (1) 的定义代入 (4) 并得到

$$J(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ r(x, \tau_n) \sum_{t=1}^n \log \pi(a_t | x, \tau_{t-1}; \theta) \right] + C, \quad (5)$$

, 其中  $C = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [r(x, \tau_n) \sum_{t=1}^n \log p(y_t | x, \tau_{t-1}, a_t)]$  表示按轨迹奖励加权的观测日志概率。由于观测概率不依赖于  $\theta$  (Section 2),  $\tau_n \sim \pi_0(\cdot | x)$  也不依赖, 因此  $C$  在  $\theta$  中是常数。因此, 最大化 (5) 等同于最大化按轨迹奖励  $r(x, \tau_n)$  加权的动作  $a_t | x, \tau_{t-1}$  的日志概率。一个自然的解释是, 我们通过将奖励平均分配给轨迹中的所有动作, 以奖励比例最大化轨迹概率。我们的目标也可以看作是具有  $n$  项的加权 SFT。由于这些项属于同一轨迹并被相同的奖励加权, 因此它们是相关的。

接下来我们将 (5) 与现有的通过强化学习 (RL) 学习提问澄清问题的相关工作进行比较。Andukuri et al. [2] 对最有奖励的轨迹应用了 SFT, 这可以看作是将 (5) 中的  $r(x, \tau_n)$  换成一个指示轨迹具有高奖励的指标。Chen et al. [13] 通过最大化负 DPO 损失来学习在每一步采取最佳行动, 这可以看作是将 (5) 中的每项替换为 DPO 损失。我们在 Section 4 中对这些工作进行实证比较, 观察到显著的提升, 因为他们没有充分利用奖励信号; 他们仅仅使用奖励信号将原问题转换为相应的 SFT 或 DPO 问题。

现在我们将 (5) 与大型语言模型中经典的强化学习方法进行比较。它们需要令牌级奖励或奖励模型, 并在其目标函数中包含倾向分数的比例, 我们将在接下来讨论。(5) 的优势在于这些都不是必需的。为了更精确地说明差异, 我们设  $a_t$  为轨迹  $\tau_n$  中的第  $t$  个令牌, 其中含有  $n$  个令牌, 并且  $t$  是从  $[n]$  中均匀随机选择的。Hong et al. [26] 的 Q-SFT 可以被视为最大化  $\mathbb{E}[q_t \log \pi(a_t | x, \tau_{t-1}; \theta)]$ , 其中  $q_t$  是在步骤  $t$  处的 Q 函数估计, 它取决于该步骤的奖励、下一动作的倾向分数比例及其上的最大化。PPO 的目标 [56] 是  $\mathbb{E}[\text{clip}(\pi(a_t | x, \tau_{t-1}; \theta) / \pi_0(a_t | x, \tau_{t-1}), A_t)]$ , 其中  $A_t$  是在步骤  $t$  处的优势, 通过奖励模型估计, 并且 clip 是一个剪辑运算符。GRPO [57] 可以被看作 PPO, 其中  $A_t$  是使用标准化的模拟未来奖励来估计的。我们没有在经验上与这些方法进行比较, 因为它们不是我们领域内的最新基准, 并且它们的实施需要令牌级奖励模型, 不像 Andukuri et al. [2] 和 Chen et al. [13]。

### 3.2 算法 ReFit

我们在 Algorithm 1 中通过算法实现了 (5) 的优化, 并称之为 re 加权  $\pi$  微调 (ReFit)。由策略  $\pi_0$  收集的  $\mathcal{D} = \{(x, \tau_n, r)\}$  数据集是 ReFit 的输入, 我们生成它的步骤如下。首先, 我们对上下文  $x \sim q$  进行采样。其次, 我们采样轨迹  $\tau_n \sim \pi_0(\cdot | x)$  并计算其奖励  $r(x, \tau_n)$ 。最后, 我们将  $(x, \tau_n, r(x, \tau_n))$  添加到数据集中。

在对数据集进行采样后, 我们通过梯度上升来优化策略  $\theta$ 。  $J(\theta)$  在  $\theta$  处的梯度为

$$\nabla J(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ r(x, \tau_n) \sum_{t=1}^n \nabla \log \pi(a_t | x, \tau_{t-1}; \theta) \right]. \quad (6)$$

优化是迭代进行的。在迭代  $i$  中, 我们通过单条轨迹  $(x, \tau_n, r) \in \mathcal{D}$  上的梯度  $g_i$  来近似  $\nabla J(\theta)$ 。由于轨迹是独立同分布生成的,  $g_i$  是对 (6) 的无偏估计。在计算出  $g_i$  后, 我们将策略更新为  $\theta + \alpha_i g_i$ , 其中  $\alpha_i > 0$  是学习率。优化在一次训练周期后停止, 但可以进行更多周期。我们注意到,  $g_i$  在代数上等价于对  $n$  SFT 数据点加权相同奖励的梯度。因此, 我们通过修改 TRL 中的 SFT 来实现 ReFit。

### 3.3 标准化奖励加权微调

(6) 的一个挑战是估计量的经验方差可能很高。例如, 假设奖励在  $[9, 10]$  中。那么梯度将被 10 缩放, 而不是 1, 后者可以通过将奖励重新缩放到  $[0, 1]$  并解决一个看似等价的问题来获得。这激发了许多关于策略梯度 [61, 5, 47]

---

#### Algorithm 1 ReFit / SWiFt

---

- 1: Input: Learning rate schedule  $(\alpha_i)_{i \in \mathbb{N}}$
  - 2: Generate a logged dataset  $\mathcal{D} = \{(x, \tau_n, r)\}$ , where  $r \in \mathbb{R}$  is a reward of  $\tau_n$  (ReFit) or a standardized reward of  $\tau_n$  (SWiFt)
  - 3: Initialize  $\theta$  and  $i \leftarrow 1$
  - 4: **for all**  $(x, \tau_n, r) \in \mathcal{D}$  **do**
  - 5:      $g_i \leftarrow r \sum_{t=1}^n \nabla \log \pi(a_t | x, \tau_{t-1}; \theta)$
  - 6:      $\theta \leftarrow \theta + \alpha_i g_i$  and  $i \leftarrow i + 1$
  - 7: Output: Learned policy  $\theta$
-



的方差减少的研究。这也激发了我们优化标准化奖励的工作。我们首先表明，在某些假设下，标准化奖励的优化等同于优化 (2)。

**Lemma 2.** *Let  $\mu(x) \geq 0$  and  $\sigma(x) > 0$  be any non-negative functions of context  $x$ . Let  $\tilde{r}(x, \tau_n) = (r(x, \tau_n) - \mu(x))/\sigma(x)$  be the standardized reward. Suppose that there exists  $\theta_*$  that maximizes all  $\mathbb{E}_{\tau_n \sim \pi(\cdot|x;\theta)} [r(x, \tau_n) | x]$  jointly. Then it also maximizes*

$$\mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot|x;\theta)} [\tilde{r}(x, \tau_n)] . \quad (7)$$

证明在 Appendix A.1 中。关键假设在 Lemma 2 中，即存在  $\theta_*$  可以同时最大化所有  $\mathbb{E}_{\tau_n \sim \pi(\cdot|x;\theta)} [r(x, \tau_n) | x]$ ，当策略类别丰富时，例如由 LLM 表示时，该假设预计会被满足或接近满足。这是因为策略以  $x$  为条件。

在本节的其余部分，我们推导出 (7) 的一个离线变体，该变体在 Section 3.1 中具有类似于 (4) 的理想特性。应用相同推理的挑战在于标准化奖励  $\tilde{r}(x, \tau_n)$  可能是负的。我们近似的误差如下所示。

**Lemma 3.** *For any policies  $\pi$  and  $\pi_0$ , and any rewards in  $[-b, b]$ ,*

$$|\mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot|x;\theta)} [\tilde{r}(x, \tau_n)] - \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot|x)} [\tilde{r}(x, \tau_n) \log \pi(\tau_n | x; \theta)]| \leq |C_1| + C_2 ,$$

where  $C_1$  is a constant independent of  $\theta$  defined in Lemma 1 and

$$C_2 = b \max_{\theta \in \Theta, x, \tau_n} \left( \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} - \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right) .$$

证明在 Appendix A.2 中。Lemma 3 指出，在 (7) 中的在线目标与其离线对应目标

$$J(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot|x)} [\tilde{r}(x, \tau_n) \log \pi(\tau_n | x; \theta)] \quad (8)$$

之间的差异是  $O(|C_1| + C_2)$ 。虽然  $C_2$  可以很大，因为它依赖于倾向得分的比率，但是这一因素与 Lemma 1 相当。具体来说， $C_1$  的因素在 Lemma 1 中陈述，而得出下界的关键步骤是  $u \geq 1 + \log u$  对  $u = \pi(\tau_n | x; \theta)/\pi_0(\tau_n | x)$ 。与 Lemma 1 唯一重大不同的是我们没有得到一个适当的下界。使用与 Section 3.1 中相同的推理，最大化 (8) 等价于针对  $n$  数据点  $(a_t, x, \tau_{t-1})$  进行的 SFT，这些数据点由标准化的轨迹奖励  $\tilde{r}(x, \tau_n)$  加权。这些项是相关的，因为它们属于同一轨迹并由相同的奖励加权。

我们用 Algorithm 1 实现 (8) 的优化。唯一不同的是奖励被标准化，因此我们将这种方法称为 **s** 标准化奖励- **w** e i g h t e d f i n e - t u n i n g (**SWiFt**)。日志数据集  $\mathcal{D} = \{(x, \tau_n, \tilde{r})\}$  生成如下。首先，我们采样  $x$ 。第二，我们为  $i \in [m]$  采样  $m$  条轨迹  $\tau_{n,i} \sim \pi_0(\cdot | x)$  并计算其奖励  $r(x, \tau_{n,i})$ 。第三，我们分别估计平均奖励  $\mu(x)$  和奖励的标准差  $\sigma(x)$  为

$$\hat{\mu}(x) = \frac{1}{m} \sum_{i=1}^m r(x, \tau_{n,i}), \quad \hat{\sigma}(x) = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (r(x, \tau_{n,i}) - \hat{\mu}(x))^2},$$

。最后，我们将所有奖励标准化为  $\tilde{r}(x, \tau_{n,i}) = (r(x, \tau_{n,i}) - \hat{\mu}(x))/\hat{\sigma}(x)$  并将所有  $(x, \tau_{n,i}, \tilde{r}(x, \tau_{n,i}))$  添加到数据集中。标准化的成本，计算  $\hat{\mu}(x)$  和  $\hat{\sigma}(x)$ ，是  $O(mn)$ 。这与采样  $m$  条长度为  $n$  的轨迹的成本相同，因此可以忽略不计。

## 4 实验

我们在六个数据集上评估我们的方法：OpenBookQA [43]、ARC [16]、SciQA [68] 和 MMLU [24] 是标准的问答基准；我们将文本到 SQL 的对话数据集 CoSQL [70] 和数学辅导数据集 MathDial [41] 转换为问答风格的对话数据集。这些基准涵盖了各种领域，并显示我们的方法在大多数情况下学习到了更好的策略。我们在 Appendix C 中更详细地描述了这些基准，并在 ?? 中展示了提示示例。

我们为每个基准生成 500 个任务，并报告每个基准中任务的平均性能。代理尝试在  $n = 3$  步内解决每个任务。用户的设计如下：它要求代理在步骤 1 中解决问题，在步骤 2 中鼓励其深入思考，并在步骤 3 中要求最终答案。这遵循了已建立的评估协议 [55, 19]。我们在思考模式和标准模式下进行实验。在思考模式中，我们要求代理在 `< thinking >` 标签中回答并给出答案的推理。在标准模式下，我们仅要求代理回答。代理使用 Llama-3.1-8B-Instruct 实现。

Table 1: 模型性能比较 - 思考模式 (ARC)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.7993 $\pm$ 0.0236	97.9 $\pm$ 0.0	7.19 $\pm$ 0.14	8.12 $\pm$ 0.17	7.46 $\pm$ 0.12	6.60 $\pm$ 0.11	6.95 $\pm$ 0.13	7.75 $\pm$ 0.17
ReFit (ours)	0.7889 $\pm$ 0.0240	97.9 $\pm$ 0.0	7.12 $\pm$ 0.14	8.03 $\pm$ 0.17	7.37 $\pm$ 0.13	6.56 $\pm$ 0.11	6.88 $\pm$ 0.14	7.66 $\pm$ 0.18
DPO	0.6471 $\pm$ 0.0281	8.7 $\pm$ 0.0	5.72 $\pm$ 0.18	6.84 $\pm$ 0.22	6.05 $\pm$ 0.16	5.30 $\pm$ 0.15	5.21 $\pm$ 0.17	6.02 $\pm$ 0.21
STaR-GATE	0.6990 $\pm$ 0.0270	90.0 $\pm$ 0.0	6.67 $\pm$ 0.17	7.48 $\pm$ 0.20	6.94 $\pm$ 0.16	6.22 $\pm$ 0.14	6.50 $\pm$ 0.16	7.11 $\pm$ 0.21
Base	0.3772 $\pm$ 0.0146	75.1 $\pm$ 0.0	6.47 $\pm$ 0.12	7.32 $\pm$ 0.14	6.56 $\pm$ 0.11	5.80 $\pm$ 0.09	6.40 $\pm$ 0.11	6.92 $\pm$ 0.16
STaR-GATE-D	0.7578 $\pm$ 0.0252	23.9 $\pm$ 0.0	5.47 $\pm$ 0.16	6.99 $\pm$ 0.20	5.65 $\pm$ 0.16	4.83 $\pm$ 0.14	4.74 $\pm$ 0.16	5.95 $\pm$ 0.19
StepDPO	0.6401 $\pm$ 0.0282	8.0 $\pm$ 0.0	5.46 $\pm$ 0.18	6.60 $\pm$ 0.22	5.76 $\pm$ 0.17	5.04 $\pm$ 0.15	4.88 $\pm$ 0.17	5.83 $\pm$ 0.21

我们在 Appendix D 中详细说明了模型和训练参数。我们通过不同的温度解决每个任务 3 次。这三次运行用于 SWiFt 中的奖励标准化，也用于实现 Andukuri et al. [2] 和 Chen et al. [13] 的方法。我们在 ?? 中对优化步数  $n$  进行了消融。

我们报告了多种指标。最基本的性能度量是准确性，这是答案与正确（黄金标准）答案相匹配的问题所占的比例。我们报告了模型输出 <thinking> 标签作为思考的百分比。这显示了模型遵循推理指令的程度。我们还报告了由 GPT-4o 根据对话质量分配给响应的六个奖励指标：1. 总体：其余 5 分数的总结。2. 准确性：代理是否选择了正确答案？3. 推理能力：推理是否合乎逻辑、清晰且精确？4. 综合性：是否妥善处理了备选方案？5. 教学价值：这个解释是否有助于某人学习？6. 信心校准：代理在给出最终答案时的信心水平是否合适？总体奖励被用作所有 RL 算法中的奖励，并且我们将其从 [0, 10] 缩放到 [0, 1] 以进行训练。这些在 Table 1 - 12 中标记为 R 总体 - R 信心。

我们考虑了五个基线。第一个基线是原始策略，我们称其为 Base。我们期望通过学习能够超越 Base。所有其他基线都是离线 RL 算法。为了进行公平比较，我们在所有基线中使用相同的采样轨迹数据集。唯一的区别在于数据集的使用方式。STaR-GATE [2] 通过对最有回报的轨迹进行监督微调来学习策略。这类似于对用于学习的轨迹进行奖励信号的阈值化。我们通过蒸馏这个基线，正如在 Andukuri et al. [2] 中进行的，并称其为 STaR-GATE-D。第四个基线由 Chen et al. [13] 激励。Chen et al. [13] 中的关键思想是在原始轨迹的每一步生成一个新的轨迹，然后根据相应的轨迹奖励来确定该步骤中的胜负动作。之后，使用 DPO 来学习胜利动作。我们称这个基线为 StepDPO。STaR-GATE 和 StepDPO 的主要限制是它们没有充分利用奖励信号；它们仅利用这个信号将原始问题转化为 SFT 或 DPO 问题。我们通过 RL 直接优化奖励。最后一个基线是 DPO，在这里最终的胜负响应被用来解决原始问题而不提出任何问题。这个基线展示了在不进行对话的情况下可以达到什么。我们的算法 ReFit 和 SWiFt 如 Section 3 所述实现。我们期望 SWiFt 能够超越 ReFit，因为基于奖励的学习往往对奖励的尺度敏感 [61, 5, 47]，这激励了我们针对标准化奖励的算法和理论 (Section 3.3)。

结果。我们在表格 1 - 12 中报告了关于所有六个基准的结果，分别在思维模式和标准模式下。最佳结果用粗体显示，第二佳结果是 underlined。置信区间是估计的标准误差。所有 RL 方法的训练时间具有可比性，因为它们在大小相似的数据集上进行训练，并使用相同的架构。

我们观察到以下趋势。首先，在准确性方面，SWiFt 在 7 次实验中获胜，并且在 12 次实验中是排名前两名的方法之一。虽然 SWiFt 最大化了整体奖励，但它在剩余的 5 奖励指标中表现极好。特别是，它的大多数奖励指标在 12 次实验中排名前两名之一。ReFit 在 3 次实验中表现得比 SWiFt 显著逊色：考虑 OpenBookQA、标准 MMLU 和标准 CoSQL。虽然如此，它在 12 次实验中仍是排名前两名的方法之一。我们认为这是因为在 TRL [64] 中的 SFT 是由自适应优化器 [30] 实现的，这些优化器能够适应梯度的规模，从而部分缓解了奖励尺度不佳的问题。

最好的两个基线是 STaR-GATE 和 STaR-GATE-D。这显示了通过 SFT 实现 RL 的鲁棒性，这是 Andukuri et al. [2] 中的关键思想，可以通过蒸馏进一步改善。如前所述，我们的工作可以被视为对这一思想的细化，我们用轨迹的实际奖励来加权 SFT 更新，而不是用高奖励的指示器 (Section 3.1)。我们的方法的优点在于它没有额外的超参数来决定哪些轨迹具有高奖励，并且可以与原始目标 (Lemma 1) 及其标准化 (Lemma 3) 适当地相关联。最差的基线是 Base，这表明了学习澄清的价值。我们还在 ?? 中展示了雷达图。我们应用了 UMAP [42] 降维技术来可视化 Base 和 SWiFt 在它们的响应中如何在 Appendix E 中有所不同。

## 5 相关工作

我们简要回顾了三个部分的相关工作，分别对应于有监督学习、经典强化学习和具有大规模语言模型的强化学习。更详细的回顾可以参考 Appendix B。

Table 2: 模型性能比较 - 思维模式 (MMLU)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.7032 $\pm$ 0.0367	97.4 $\pm$ 0.0	5.59 $\pm$ 0.22	6.42 $\pm$ 0.26	5.94 $\pm$ 0.20	5.10 $\pm$ 0.18	5.23 $\pm$ 0.20	6.14 $\pm$ 0.26
ReFit (ours)	0.7097 $\pm$ 0.0365	98.1 $\pm$ 0.0	5.59 $\pm$ 0.22	6.43 $\pm$ 0.26	5.94 $\pm$ 0.20	5.06 $\pm$ 0.18	5.19 $\pm$ 0.20	6.11 $\pm$ 0.26
DPO	0.6387 $\pm$ 0.0386	7.1 $\pm$ 0.0	4.77 $\pm$ 0.23	5.71 $\pm$ 0.29	5.09 $\pm$ 0.22	4.35 $\pm$ 0.20	4.24 $\pm$ 0.22	5.07 $\pm$ 0.28
STaR-GATE	0.6000 $\pm$ 0.0393	81.3 $\pm$ 0.0	5.34 $\pm$ 0.24	5.91 $\pm$ 0.29	5.70 $\pm$ 0.22	4.98 $\pm$ 0.20	5.15 $\pm$ 0.22	5.63 $\pm$ 0.29
Base	0.2774 $\pm$ 0.0127	53.5 $\pm$ 0.0	5.87 $\pm$ 0.16	6.57 $\pm$ 0.20	6.03 $\pm$ 0.15	5.19 $\pm$ 0.14	5.97 $\pm$ 0.15	6.19 $\pm$ 0.22
STaR-GATE-D	0.5548 $\pm$ 0.0399	25.2 $\pm$ 0.0	4.23 $\pm$ 0.23	4.96 $\pm$ 0.28	4.57 $\pm$ 0.22	3.93 $\pm$ 0.20	3.77 $\pm$ 0.21	4.34 $\pm$ 0.27
StepDPO	0.6387 $\pm$ 0.0386	5.2 $\pm$ 0.0	4.94 $\pm$ 0.23	5.88 $\pm$ 0.28	5.26 $\pm$ 0.21	4.50 $\pm$ 0.20	4.45 $\pm$ 0.22	5.31 $\pm$ 0.28

Table 3: 模型性能比较 - 思维模式 (OpenBookQA)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.6814 $\pm$ 0.0310	96.5 $\pm$ 0.0	6.16 $\pm$ 0.21	6.86 $\pm$ 0.24	6.49 $\pm$ 0.19	5.89 $\pm$ 0.15	5.99 $\pm$ 0.19	6.52 $\pm$ 0.25
ReFit (ours)	0.6504 $\pm$ 0.0317	96.5 $\pm$ 0.0	5.84 $\pm$ 0.22	6.63 $\pm$ 0.26	6.12 $\pm$ 0.21	5.58 $\pm$ 0.17	5.62 $\pm$ 0.21	6.25 $\pm$ 0.26
DPO	0.6195 $\pm$ 0.0323	10.6 $\pm$ 0.0	5.09 $\pm$ 0.21	6.21 $\pm$ 0.27	5.35 $\pm$ 0.20	4.82 $\pm$ 0.18	4.47 $\pm$ 0.20	5.55 $\pm$ 0.25
STaR-GATE	0.6549 $\pm$ 0.0316	92.5 $\pm$ 0.0	6.01 $\pm$ 0.21	6.68 $\pm$ 0.25	6.35 $\pm$ 0.20	5.80 $\pm$ 0.16	5.78 $\pm$ 0.20	6.36 $\pm$ 0.26
Base	0.3628 $\pm$ 0.0175	74.3 $\pm$ 0.0	5.99 $\pm$ 0.15	6.77 $\pm$ 0.19	6.15 $\pm$ 0.14	5.43 $\pm$ 0.12	5.95 $\pm$ 0.14	6.31 $\pm$ 0.20
STaR-GATE-D	0.6903 $\pm$ 0.0308	20.8 $\pm$ 0.0	5.21 $\pm$ 0.19	6.64 $\pm$ 0.25	5.40 $\pm$ 0.18	4.73 $\pm$ 0.16	4.35 $\pm$ 0.17	5.70 $\pm$ 0.23
StepDPO	0.6106 $\pm$ 0.0324	11.5 $\pm$ 0.0	4.90 $\pm$ 0.21	6.14 $\pm$ 0.27	5.06 $\pm$ 0.20	4.56 $\pm$ 0.18	4.29 $\pm$ 0.20	5.33 $\pm$ 0.25

Table 4: 模型性能比较 - 思维模式 (SciQA)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.9248 $\pm$ 0.0175	99.1 $\pm$ 0.0	7.61 $\pm$ 0.12	8.84 $\pm$ 0.14	7.73 $\pm$ 0.11	6.76 $\pm$ 0.10	7.11 $\pm$ 0.13	8.45 $\pm$ 0.15
ReFit (ours)	0.9159 $\pm$ 0.0185	96.0 $\pm$ 0.0	7.64 $\pm$ 0.12	8.87 $\pm$ 0.14	7.76 $\pm$ 0.11	6.81 $\pm$ 0.10	7.13 $\pm$ 0.12	8.43 $\pm$ 0.15
DPO	0.7920 $\pm$ 0.0270	5.8 $\pm$ 0.0	5.96 $\pm$ 0.18	7.61 $\pm$ 0.22	6.08 $\pm$ 0.18	5.29 $\pm$ 0.16	5.14 $\pm$ 0.18	6.50 $\pm$ 0.22
STaR-GATE	0.8186 $\pm$ 0.0256	90.3 $\pm$ 0.0	7.08 $\pm$ 0.18	8.17 $\pm$ 0.21	7.27 $\pm$ 0.16	6.49 $\pm$ 0.14	6.69 $\pm$ 0.17	7.69 $\pm$ 0.21
Base	0.4956 $\pm$ 0.0076	73.5 $\pm$ 0.0	7.00 $\pm$ 0.10	8.12 $\pm$ 0.11	7.03 $\pm$ 0.10	6.11 $\pm$ 0.09	6.84 $\pm$ 0.11	7.78 $\pm$ 0.13
STaR-GATE-D	0.9027 $\pm$ 0.0197	21.7 $\pm$ 0.0	6.58 $\pm$ 0.16	8.19 $\pm$ 0.18	6.72 $\pm$ 0.16	5.78 $\pm$ 0.14	5.73 $\pm$ 0.17	7.24 $\pm$ 0.18
StepDPO	0.8186 $\pm$ 0.0256	7.5 $\pm$ 0.0	6.29 $\pm$ 0.18	7.87 $\pm$ 0.21	6.36 $\pm$ 0.18	5.57 $\pm$ 0.16	5.42 $\pm$ 0.18	6.89 $\pm$ 0.22

Table 5: 模型性能比较 - 思维模式 (CoSQL)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.6500 $\pm$ 0.0435	96.7 $\pm$ 0.0	4.87 $\pm$ 0.21	5.56 $\pm$ 0.27	5.26 $\pm$ 0.19	4.62 $\pm$ 0.15	4.23 $\pm$ 0.16	5.22 $\pm$ 0.29
ReFit (ours)	0.6500 $\pm$ 0.0435	99.2 $\pm$ 0.0	4.91 $\pm$ 0.21	5.52 $\pm$ 0.27	5.28 $\pm$ 0.18	4.63 $\pm$ 0.15	4.22 $\pm$ 0.17	5.39 $\pm$ 0.31
DPO	0.5167 $\pm$ 0.0456	60.0 $\pm$ 0.0	4.34 $\pm$ 0.19	4.85 $\pm$ 0.25	4.72 $\pm$ 0.17	4.27 $\pm$ 0.15	4.00 $\pm$ 0.16	4.29 $\pm$ 0.28
STaR-GATE	0.6167 $\pm$ 0.0444	90.0 $\pm$ 0.0	5.28 $\pm$ 0.24	5.78 $\pm$ 0.30	5.51 $\pm$ 0.22	5.19 $\pm$ 0.16	4.90 $\pm$ 0.20	5.54 $\pm$ 0.33
Base	0.2000 $\pm$ 0.0143	65.8 $\pm$ 0.0	5.65 $\pm$ 0.17	6.17 $\pm$ 0.22	5.88 $\pm$ 0.15	5.16 $\pm$ 0.13	5.84 $\pm$ 0.15	5.87 $\pm$ 0.27
STaR-GATE-D	0.4917 $\pm$ 0.0456	57.5 $\pm$ 0.0	3.94 $\pm$ 0.17	4.49 $\pm$ 0.22	4.45 $\pm$ 0.16	3.89 $\pm$ 0.14	3.58 $\pm$ 0.15	3.74 $\pm$ 0.25
StepDPO	0.5250 $\pm$ 0.0456	60.0 $\pm$ 0.0	4.37 $\pm$ 0.20	4.82 $\pm$ 0.26	4.81 $\pm$ 0.18	4.26 $\pm$ 0.15	4.08 $\pm$ 0.18	4.38 $\pm$ 0.29

Table 6: 模型性能比较 - 思维模式 (MathDial)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.1933 $\pm$ 0.0228	99.3 $\pm$ 0.0	1.88 $\pm$ 0.07	1.91 $\pm$ 0.07	2.42 $\pm$ 0.07	2.15 $\pm$ 0.07	1.83 $\pm$ 0.07	1.61 $\pm$ 0.09
ReFit (ours)	0.0867 $\pm$ 0.0162	100.0 $\pm$ 0.0	2.38 $\pm$ 0.07	2.33 $\pm$ 0.07	3.13 $\pm$ 0.08	2.56 $\pm$ 0.08	2.43 $\pm$ 0.07	1.63 $\pm$ 0.07
DPO	0.1467 $\pm$ 0.0204	25.0 $\pm$ 0.0	1.61 $\pm$ 0.05	1.63 $\pm$ 0.06	2.23 $\pm$ 0.05	1.78 $\pm$ 0.07	1.56 $\pm$ 0.05	1.40 $\pm$ 0.06
STaR-GATE	0.0467 $\pm$ 0.0122	100.0 $\pm$ 0.0	2.46 $\pm$ 0.06	2.40 $\pm$ 0.07	3.28 $\pm$ 0.07	2.65 $\pm$ 0.07	2.45 $\pm$ 0.07	1.53 $\pm$ 0.05
Base	0.0000 $\pm$ 0.0212	87.7 $\pm$ 0.0	2.01 $\pm$ 0.06	2.28 $\pm$ 0.07	2.67 $\pm$ 0.07	1.77 $\pm$ 0.05	2.20 $\pm$ 0.07	1.39 $\pm$ 0.09
STaR-GATE-D	0.1167 $\pm$ 0.0185	95.0 $\pm$ 0.0	1.69 $\pm$ 0.06	1.71 $\pm$ 0.06	2.30 $\pm$ 0.07	1.81 $\pm$ 0.06	1.63 $\pm$ 0.06	1.35 $\pm$ 0.06
StepDPO	0.1467 $\pm$ 0.0204	25.7 $\pm$ 0.0	1.58 $\pm$ 0.05	1.61 $\pm$ 0.06	2.21 $\pm$ 0.06	1.72 $\pm$ 0.06	1.53 $\pm$ 0.05	1.40 $\pm$ 0.06

Table 7: 模型性能比较 - 标准模式 (ARC)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.7778 $\pm$ 0.0289	0.0 $\pm$ 0.0	7.26 $\pm$ 0.19	8.04 $\pm$ 0.22	7.51 $\pm$ 0.17	6.76 $\pm$ 0.14	7.12 $\pm$ 0.18	7.82 $\pm$ 0.23
ReFit (ours)	0.7729 $\pm$ 0.0291	0.0 $\pm$ 0.0	7.23 $\pm$ 0.19	7.98 $\pm$ 0.22	7.44 $\pm$ 0.18	6.80 $\pm$ 0.14	7.03 $\pm$ 0.18	7.66 $\pm$ 0.23
DPO	0.6377 $\pm$ 0.0334	0.0 $\pm$ 0.0	5.68 $\pm$ 0.20	6.51 $\pm$ 0.25	6.06 $\pm$ 0.18	5.41 $\pm$ 0.16	5.26 $\pm$ 0.19	5.78 $\pm$ 0.25
STaR-GATE	0.7971 $\pm$ 0.0280	0.0 $\pm$ 0.0	7.49 $\pm$ 0.18	8.25 $\pm$ 0.21	7.67 $\pm$ 0.17	6.93 $\pm$ 0.14	7.36 $\pm$ 0.17	8.02 $\pm$ 0.22
Base	0.5652 $\pm$ 0.0142	0.0 $\pm$ 0.0	6.87 $\pm$ 0.14	7.68 $\pm$ 0.18	6.97 $\pm$ 0.13	6.25 $\pm$ 0.11	6.75 $\pm$ 0.14	7.21 $\pm$ 0.20
STaR-GATE-D	0.7101 $\pm$ 0.0315	0.0 $\pm$ 0.0	5.95 $\pm$ 0.18	6.96 $\pm$ 0.22	6.29 $\pm$ 0.17	5.56 $\pm$ 0.14	5.42 $\pm$ 0.17	6.18 $\pm$ 0.22
StepDPO	0.6280 $\pm$ 0.0336	0.0 $\pm$ 0.0	5.76 $\pm$ 0.20	6.55 $\pm$ 0.25	6.19 $\pm$ 0.18	5.54 $\pm$ 0.15	5.43 $\pm$ 0.19	5.84 $\pm$ 0.25

监督学习。许多研究集中于通过提问澄清性问题来澄清用户提示 [39, 72]。值得注意的是, Zelikman et al. [72] 提出了一种简单但有影响力的方法: 从成功的合理性和再生成的失败中学习。是否应提问澄清性问题的的问题也得到了广泛研究 [40, 9, 36], 这促成了新的基准 [9, 75] 和调查 [46, 74] 的产生。这些研究也被扩展到视觉语言模型 [23, 63, 12]。与这些工作相比, 我们采取了一种强化学习的方法。

经典强化学习。在离线强化学习中进行对话优化是一个经典的话题, 并在 Levine et al. [35] 的第 6.6 节中进行了回顾。例如, Zhou et al. [78] 提出了在线和离线策略梯度以提高语言质

Table 8: 模型性能比较 - 标准模式 (MMLU)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.7218 $\pm$ 0.0389	0.0 $\pm$ 0.0	6.08 $\pm$ 0.25	6.88 $\pm$ 0.29	6.32 $\pm$ 0.23	5.50 $\pm$ 0.21	5.80 $\pm$ 0.23	6.71 $\pm$ 0.30
ReFit (ours)	0.6917 $\pm$ 0.0400	0.0 $\pm$ 0.0	5.93 $\pm$ 0.26	6.72 $\pm$ 0.31	6.23 $\pm$ 0.24	5.42 $\pm$ 0.21	5.56 $\pm$ 0.25	6.36 $\pm$ 0.31
DPO	0.5489 $\pm$ 0.0431	0.0 $\pm$ 0.0	4.86 $\pm$ 0.25	5.52 $\pm$ 0.31	5.30 $\pm$ 0.23	4.61 $\pm$ 0.21	4.56 $\pm$ 0.23	4.92 $\pm$ 0.30
STaR-GATE	0.6842 $\pm$ 0.0403	0.0 $\pm$ 0.0	5.93 $\pm$ 0.26	6.68 $\pm$ 0.31	6.20 $\pm$ 0.25	5.41 $\pm$ 0.22	5.59 $\pm$ 0.25	6.41 $\pm$ 0.31
Base	0.3008 $\pm$ 0.0165	0.0 $\pm$ 0.0	5.97 $\pm$ 0.19	6.74 $\pm$ 0.23	6.16 $\pm$ 0.18	5.32 $\pm$ 0.16	5.95 $\pm$ 0.18	6.11 $\pm$ 0.26
STaR-GATE-D	0.5940 $\pm$ 0.0426	0.0 $\pm$ 0.0	4.98 $\pm$ 0.25	5.75 $\pm$ 0.30	5.29 $\pm$ 0.24	4.65 $\pm$ 0.21	4.53 $\pm$ 0.24	5.26 $\pm$ 0.31
StepDPO	0.5263 $\pm$ 0.0433	0.0 $\pm$ 0.0	4.77 $\pm$ 0.26	5.44 $\pm$ 0.32	5.17 $\pm$ 0.25	4.49 $\pm$ 0.22	4.38 $\pm$ 0.23	4.94 $\pm$ 0.31

Table 9: 模型性能比较 - 标准模式 (OpenBookQA)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.7662 $\pm$ 0.0299	0.0 $\pm$ 0.0	6.85 $\pm$ 0.21	7.73 $\pm$ 0.24	7.09 $\pm$ 0.19	6.42 $\pm$ 0.15	6.59 $\pm$ 0.20	7.45 $\pm$ 0.25
ReFit (ours)	0.7562 $\pm$ 0.0303	0.0 $\pm$ 0.0	6.73 $\pm$ 0.21	7.66 $\pm$ 0.25	6.96 $\pm$ 0.20	6.29 $\pm$ 0.16	6.43 $\pm$ 0.21	7.25 $\pm$ 0.25
DPO	0.5025 $\pm$ 0.0353	0.0 $\pm$ 0.0	4.95 $\pm$ 0.21	5.49 $\pm$ 0.26	5.43 $\pm$ 0.19	5.04 $\pm$ 0.16	4.66 $\pm$ 0.20	4.95 $\pm$ 0.26
STaR-GATE	0.7512 $\pm$ 0.0305	0.0 $\pm$ 0.0	6.69 $\pm$ 0.22	7.54 $\pm$ 0.25	6.96 $\pm$ 0.20	6.27 $\pm$ 0.16	6.50 $\pm$ 0.21	7.23 $\pm$ 0.26
Base	0.4328 $\pm$ 0.0180	0.0 $\pm$ 0.0	6.22 $\pm$ 0.16	6.95 $\pm$ 0.21	6.37 $\pm$ 0.15	5.65 $\pm$ 0.13	6.12 $\pm$ 0.15	6.51 $\pm$ 0.21
STaR-GATE-D	0.7114 $\pm$ 0.0320	0.0 $\pm$ 0.0	5.84 $\pm$ 0.19	6.96 $\pm$ 0.24	6.21 $\pm$ 0.18	5.52 $\pm$ 0.15	5.24 $\pm$ 0.18	6.21 $\pm$ 0.23
StepDPO	0.5174 $\pm$ 0.0352	0.0 $\pm$ 0.0	4.92 $\pm$ 0.22	5.54 $\pm$ 0.27	5.36 $\pm$ 0.20	4.97 $\pm$ 0.17	4.65 $\pm$ 0.20	4.97 $\pm$ 0.27

Table 10: 模型性能比较 - 标准模式 (SciQA)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.9502 $\pm$ 0.0153	0.0 $\pm$ 0.0	8.04 $\pm$ 0.12	9.13 $\pm$ 0.13	8.12 $\pm$ 0.11	7.17 $\pm$ 0.10	7.71 $\pm$ 0.13	8.88 $\pm$ 0.15
ReFit (ours)	0.9453 $\pm$ 0.0160	0.0 $\pm$ 0.0	8.04 $\pm$ 0.12	9.08 $\pm$ 0.14	8.11 $\pm$ 0.11	7.20 $\pm$ 0.10	7.69 $\pm$ 0.13	8.87 $\pm$ 0.15
DPO	0.7612 $\pm$ 0.0301	0.0 $\pm$ 0.0	6.41 $\pm$ 0.19	7.44 $\pm$ 0.23	6.72 $\pm$ 0.17	6.00 $\pm$ 0.15	6.02 $\pm$ 0.19	6.78 $\pm$ 0.23
STaR-GATE	0.9005 $\pm$ 0.0211	0.0 $\pm$ 0.0	7.85 $\pm$ 0.16	8.88 $\pm$ 0.18	7.98 $\pm$ 0.14	7.06 $\pm$ 0.13	7.52 $\pm$ 0.16	8.62 $\pm$ 0.19
Base	0.6517 $\pm$ 0.0086	0.0 $\pm$ 0.0	7.48 $\pm$ 0.10	8.56 $\pm$ 0.11	7.52 $\pm$ 0.10	6.55 $\pm$ 0.09	7.34 $\pm$ 0.10	8.10 $\pm$ 0.13
STaR-GATE-D	0.9005 $\pm$ 0.0211	0.0 $\pm$ 0.0	6.90 $\pm$ 0.15	8.39 $\pm$ 0.17	7.13 $\pm$ 0.14	6.20 $\pm$ 0.13	6.03 $\pm$ 0.16	7.42 $\pm$ 0.18
StepDPO	0.7463 $\pm$ 0.0307	0.0 $\pm$ 0.0	6.23 $\pm$ 0.20	7.25 $\pm$ 0.24	6.52 $\pm$ 0.19	5.88 $\pm$ 0.16	5.78 $\pm$ 0.19	6.53 $\pm$ 0.25

Table 11: 模型性能比较 - 标准模式 (CoSQL)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.6583 $\pm$ 0.0433	0.0 $\pm$ 0.0	5.45 $\pm$ 0.24	5.97 $\pm$ 0.30	5.72 $\pm$ 0.22	5.28 $\pm$ 0.17	4.95 $\pm$ 0.21	5.85 $\pm$ 0.33
ReFit (ours)	0.6250 $\pm$ 0.0442	0.0 $\pm$ 0.0	5.16 $\pm$ 0.24	5.64 $\pm$ 0.30	5.47 $\pm$ 0.22	4.99 $\pm$ 0.18	4.72 $\pm$ 0.20	5.52 $\pm$ 0.32
DPO	0.2833 $\pm$ 0.0411	0.0 $\pm$ 0.0	4.19 $\pm$ 0.20	4.37 $\pm$ 0.25	4.79 $\pm$ 0.19	4.49 $\pm$ 0.15	4.13 $\pm$ 0.17	3.69 $\pm$ 0.27
STaR-GATE	0.6083 $\pm$ 0.0446	0.0 $\pm$ 0.0	5.34 $\pm$ 0.25	5.81 $\pm$ 0.31	5.65 $\pm$ 0.23	5.26 $\pm$ 0.18	4.99 $\pm$ 0.22	5.57 $\pm$ 0.34
Base	0.1250 $\pm$ 0.0117	0.0 $\pm$ 0.0	5.38 $\pm$ 0.16	5.88 $\pm$ 0.22	5.66 $\pm$ 0.15	4.92 $\pm$ 0.12	5.49 $\pm$ 0.14	5.13 $\pm$ 0.24
STaR-GATE-D	0.2083 $\pm$ 0.0371	0.0 $\pm$ 0.0	3.62 $\pm$ 0.19	3.82 $\pm$ 0.23	4.25 $\pm$ 0.18	4.02 $\pm$ 0.16	3.73 $\pm$ 0.17	3.03 $\pm$ 0.26
StepDPO	0.2917 $\pm$ 0.0415	0.0 $\pm$ 0.0	4.21 $\pm$ 0.20	4.45 $\pm$ 0.26	4.85 $\pm$ 0.18	4.50 $\pm$ 0.15	4.10 $\pm$ 0.17	3.73 $\pm$ 0.28

Table 12: 模型性能比较 - 标准模式 (MathDial)

Model	Accuracy	Thinking ( % )	R Overall	R Accuracy	R Reasoning	R Comprehensive	R Pedagogic	R Confidence
SWiFt (ours)	0.0967 $\pm$ 0.0171	0.0 $\pm$ 0.0	2.43 $\pm$ 0.07	2.41 $\pm$ 0.07	3.09 $\pm$ 0.08	2.68 $\pm$ 0.07	2.45 $\pm$ 0.07	1.66 $\pm$ 0.07
ReFit (ours)	0.0600 $\pm$ 0.0137	0.0 $\pm$ 0.0	2.43 $\pm$ 0.07	2.50 $\pm$ 0.08	3.15 $\pm$ 0.08	2.68 $\pm$ 0.07	2.41 $\pm$ 0.08	1.63 $\pm$ 0.07
DPO	0.2100 $\pm$ 0.0235	0.0 $\pm$ 0.0	1.85 $\pm$ 0.06	1.90 $\pm$ 0.07	2.41 $\pm$ 0.06	2.00 $\pm$ 0.07	1.77 $\pm$ 0.06	1.58 $\pm$ 0.07
STaR-GATE	0.1067 $\pm$ 0.0178	0.0 $\pm$ 0.0	2.29 $\pm$ 0.07	2.25 $\pm$ 0.07	2.94 $\pm$ 0.08	2.58 $\pm$ 0.08	2.30 $\pm$ 0.07	1.56 $\pm$ 0.07
Base	0.0000 $\pm$ 0.0168	0.0 $\pm$ 0.0	1.90 $\pm$ 0.06	2.20 $\pm$ 0.07	2.54 $\pm$ 0.07	1.81 $\pm$ 0.05	2.01 $\pm$ 0.07	1.20 $\pm$ 0.08
STaR-GATE-D	0.2000 $\pm$ 0.0231	0.0 $\pm$ 0.0	1.55 $\pm$ 0.04	1.63 $\pm$ 0.05	1.95 $\pm$ 0.05	1.79 $\pm$ 0.06	1.51 $\pm$ 0.04	1.31 $\pm$ 0.05
StepDPO	0.2067 $\pm$ 0.0234	0.0 $\pm$ 0.0	1.86 $\pm$ 0.06	1.87 $\pm$ 0.06	2.43 $\pm$ 0.06	2.03 $\pm$ 0.07	1.78 $\pm$ 0.06	1.55 $\pm$ 0.06

量。这种方法和其他基于经典强化学习原语的方法，例如 Q 函数 [67, 45]，无法直接应用于 LLMs。

使用 LLM 进行强化学习。最相关的工作是 Andukuri et al. [2] 和 Chen et al. [13]，它们都使用强化学习从模拟对话中学习澄清性问题。尽管我们采用了相同的学习范式，但在使用方式上有所不同。Andukuri et al. [2] 选择最有奖励的轨迹并在其上进行调整。Chen et al. [13] 为对话的每一步生成替代响应，然后使用 DPO [50] 优化以获得更好的响应。我们工作的主要区别在于我们直接为奖励进行优化。

## 6 结论

LLMs [49, 8, 6] 的出现引发了对问答代理的极大兴趣，并提出了多种方法 (Section 5)。在这项工作中，我们通过强化学习学习提问以澄清问题。为使其实用，我们推导了该问题的离线变体，可以被看作加权的监督微调，因此可以很容易地在任何 LLM 中实现。我们进一步推导出一个采用标准化奖励的离线目标，这可以降低策略优化中的方差。我们的方法与最



近提出的基于 SFT 和直接偏好优化的方法形成鲜明对比，那些方法有额外的超参数并且不直接优化奖励。我们在实验中超越了它们。

局限性。强化学习的计算成本往往比监督学习高得多。我们通过提出将离线强化学习简化为 SFT（这是一种监督学习技术）来部分解决这个问题。此外，记录数据集的质量对离线强化学习至关重要。然而，我们并不尝试改进它，而是依赖一种流行的技术来获得多样化的数据集：在大语言模型中使用不同的温度模拟对话轨迹。

未来工作。我们注意到，我们提出的算法 ReFit 和 SWiFt 是通用的，因此可以应用于 QA 以外的其他领域。我们专注于 QA 是因为该领域有许多既定的基准和基线，这使我们能够展示直接优化奖励的好处。

## References

- [1] V. M. Aleksandrov, V. I. Sysoyev, and V. V. Shemenyeva. Stochastic optimization. *Engineering Cybernetics*, 5:11–16, 1968.
- [2] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*, 2024.
- [3] Sercan Ö. Arik, Maximillian Chen, Ruoxi Sun, and Tomas Pfister. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *arXiv.org*, 2024. URL <https://api.semanticscholar.org/CorpusId:270220485>.
- [4] Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. Leftover lunch: Advantage-based offline reinforcement learning for language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.
- [5] Jonathan Baxter and Peter Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.
- [6] Rishi Bommasani et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL <https://arxiv.org/abs/2108.07258>.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [8] Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, 2020.
- [9] Yash Butala, Siddhant Garg, Pratyay Banerjee, and Amita Misra. Promise: A proactive multi-turn dialogue dataset for information-seeking intent resolution. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1774–1789, 2024.
- [10] Fu-Chieh Chang, Yu-Ting Lee, Hui-Ying Shih, and Pei-Yuan Wu. RL-star: Theoretical analysis of reinforcement learning frameworks for self-taught reasoner. *arXiv preprint arXiv:2410.23912*, 2024.
- [11] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, 2017.
- [12] Maximillian Chen, Ruoxi Sun, and Sercan Ö Arik. Data-centric improvements for enhancing multi-modal understanding in spoken conversation modeling. *arXiv preprint arXiv:2412.15995*, 2024.
- [13] Maximillian Chen, Ruoxi Sun, Sercan Ö Arik, and Tomas Pfister. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. *arXiv preprint arXiv:2406.00222*, 2024.
- [14] Yizhou Chi, Jessy Lin, Kevin Lin, and Dan Klein. Clarinet: Augmenting language models to ask clarification questions for retrieval. In *unknown*, 2024. URL <https://api.semanticscholar.org/CorpusId:270063669>.
- [15] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

- [16] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* , 2018.
- [17] Yang Deng, Lizi Liao, Wenqiang Lei, Grace Yang, Wai Lam, and Tat-Seng Chua. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems* , 2025.
- [18] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science* , 29(4):485–511, 2014.
- [19] Kevin Fischer, Darren Fürst, Sebastian Steindl, Jakob Lindner, and Ulrich Schäfer. Question: How do large language models perform on the question answering tasks? answer. *arXiv preprint arXiv:2412.12893* , 2024.
- [20] Michael Free, Andrew Langworthy, Mary Dimitropoulaki, and Simon Thompson. Towards goal-oriented agents for evolving problems observed via conversation. In unknown , 2024. URL <https://api.semanticscholar.org/CorpusId:265213767>.
- [21] Chongming Gao, Mengyao Gao, Chenxiao Fan, Shuai Yuan, Wentao Shi, and Xiangnan He. Process-supervised llm recommenders via flow-guided tuning. *arXiv preprint arXiv:2503.07377* , 2025.
- [22] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, G. Thattai, and G. Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. In *IEEE Robotics and Automation Letters* , 2022. URL <https://api.semanticscholar.org/CorpusId:247158852>.
- [23] Meera Hahn, Wenjun Zeng, Nithish Kannan, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. Proactive agents for multi-turn text-to-image generation under uncertainty. *arXiv preprint arXiv:2412.06771* , 2024.
- [24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations* .
- [25] Joey Hong, S. Levine, and Anca Dragan. Zero-shot goal-directed dialogue via rl on imagined conversations. *ArXiv* , abs/2311.05584, 2023. URL <https://api.semanticscholar.org/CorpusId:265067195>.
- [26] Joey Hong, Anca Dragan, and Sergey Levine. Q-SFT: Q-learning for language models via supervised fine-tuning. In *Proceedings of the 13th International Conference on Learning Representations* , 2025.
- [27] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* , 47(260):663–685, 1952.
- [28] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457* , 2024.
- [29] Edward Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics* , 17(2):295–311, 2008.
- [30] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations* , 2015.
- [31] Katarzyna Kobalcyk, Nicolas Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with llms. *arXiv preprint arXiv:2502.04485* , 2025.
- [32] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* , 7:453–466, 2019.
- [33] Sascha Lange, Thomas Gabel, and Martin Riedmiller. *Batch Reinforcement Learning* , pages 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [34] Hyunji Lee, Franck Dernoncourt, Trung Bui, and Seunghyun Yoon. Corg: Generating answers from complex, interrelated contexts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* , pages 8443–8460, 2025.

- [35] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. CoRR , abs/2005.01643, 2020. URL <https://arxiv.org/abs/2005.01643>.
- [36] Zixuan Li, Lizi Liao, and Tat-Seng Chua. Learning to ask critical questions for assisting product search. In unknown , 2024. URL <https://api.semanticscholar.org/CorpusId:268249077>.
- [37] Dawen Liang and Nikos Vlassis. Local policy improvement for recommender systems. arXiv preprint arXiv:2212.11431 , 2022.
- [38] Baihan Lin. Reinforcement learning and bandits for speech and language processing: Tutorial, review and outlook. In Expert systems with applications , 2022. URL <https://api.semanticscholar.org/CorpusId:253107350>.
- [39] Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. We’re afraid language models aren’t modeling ambiguity. arXiv preprint arXiv:2304.14399 , 2023.
- [40] Lili Lu, Chuan Meng, Federico Ravenda, Mohammad Aliannejadi, and Fabio Crestani. Zero-shot and efficient clarification need prediction in conversational search. arXiv preprint arXiv:2503.00179 , 2025.
- [41] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. arXiv preprint arXiv:2305.14536 , 2023.
- [42] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 , 2018.
- [43] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In EMNLP , 2018.
- [44] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 5783–5797, 2020.
- [45] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. nature , 518(7540):529–533, 2015.
- [46] Nikahat Mulla and P. Gharpure. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. In unknown , 2023. URL <https://api.semanticscholar.org/CorpusId:256438998>.
- [47] Remi Munos. Geometric variance reduction in Markov chains: Application to value function and gradient estimation. Journal of Machine Learning Research , 7:413–427, 2006.
- [48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35 , 2022.
- [49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
- [50] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems 36 , 2023.
- [51] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing , pages 2383–2392, 2016.
- [52] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 5418–5426, 2020.
- [53] Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. Upweighting easy samples in fine-tuning mitigates forgetting. arXiv preprint arXiv:2502.02797 , 2025.

- [54] Alexander Scarlatos, Ryan S Baker, and Andrew Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In Proceedings of the 15th International Learning Analytics and Knowledge Conference , pages 249–259, 2025.
- [55] Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. Training llm-based tutors to improve student learning outcomes in dialogues. arXiv preprint arXiv:2503.06424 , 2025.
- [56] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. CoRR , abs/1707.06347, 2017. URL <https://arxiv.org/abs/1707.06347>.
- [57] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. CoRR , abs/2402.03300, 2024. URL <https://arxiv.org/abs/2402.03300>.
- [58] Olivier Sigaud, Pierre-Yves Oudeyer, Thomas Carta, and Sylvain Lamprier. Eager: Asking and answering questions for automatic reward shaping in language-guided rl. In Neural Information Processing Systems , 2022. URL <https://api.semanticscholar.org/CorpusId:249890287>.
- [59] Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. Offline RL for natural language generation with implicit language Q learning. In Proceedings of the 11th International Conference on Learning Representations , 2023.
- [60] Richard Sutton and Andrew Barto. Reinforcement Learning: An Introduction . MIT Press, Cambridge, MA, 1998.
- [61] Richard Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in Neural Information Processing Systems 12 , pages 1057–1063, 2000.
- [62] Dirk V  th, Ngoc Thang Vu, and Lindsey Vanderlyn. Towards a zero-data, controllable, adaptive dialog system. In International Conference on Language Resources and Evaluation , 2024. URL <https://api.semanticscholar.org/CorpusId:268691395>.
- [63] Danae S  nchez Villegas, Ingo Ziegler, and Desmond Elliott. Imagechain: Advancing sequential image-to-text reasoning in multimodal large language models. arXiv preprint arXiv:2502.19409 , 2025.
- [64] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouedec. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>, 2020.
- [65] Haofen Wang, Yuanzi Li, Huifang Du, Xuejing Feng, Minghao Wu, and Shuqin Li. Rewarding what matters: Step-by-step reinforcement learning for task-oriented dialogue. In Conference on Empirical Methods in Natural Language Processing , 2024. URL <https://api.semanticscholar.org/CorpusId:270619424>.
- [66] Zhenduo Wang and Qingyao Ai. Simulating and modeling the risk of conversational search. In ACM Trans. Inf. Syst. , 2022. URL <https://api.semanticscholar.org/CorpusId:245650574>.
- [67] Christopher JCH Watkins and Peter Dayan. Q-learning. Machine learning , 8:279–292, 1992.
- [68] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. W-NUT 2017 , page 94, 2017.
- [69] Ronald Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning , 8(3-4):229–256, 1992.
- [70] Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. arXiv preprint arXiv:1909.05378 , 2019.
- [71] Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-star: Language models can teach themselves to think before speaking. In First Conference on Language Modeling , 2024.



- [72] Eric Zelikman, YH Wu, Jesse Mu, and Noah D Goodman. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In Proc. the 36th International Conference on Neural Information Processing Systems , volume 1126, 2024.
- [73] Dylan Zhang, Qirun Dai, and Hao Peng. The best instruction-tuning data are those that fit. arXiv preprint arXiv:2502.04194 , 2025.
- [74] Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. The imperative of conversation analysis in the era of llms: A survey of tasks, techniques, and trends. In unknown , 2024. URL <https://api.semanticscholar.org/CorpusId:272828048>.
- [75] Xuan Zhang, Yang Deng, Zifeng Ren, See-Kiong Ng, and Tat-Seng Chua. Ask-before-plan: Proactive language agents for real-world planning, 2024. URL <https://arxiv.org/abs/2406.12639>.
- [76] Xuan Zhang, Yongliang Shen, Zhe Zheng, Linjuan Wu, Wenqi Zhang, Yuchen Yan, Qiuying Peng, Jun Wang, and Weiming Lu. Asktoact: Enhancing llms tool use via self-correcting clarification. arXiv preprint arXiv:2503.01940 , 2025.
- [77] Xuandong Zhao, Will Cai, Tianneng Shi, David Huang, Licong Lin, Song Mei, and Dawn Song. Improving llm safety alignment with dual-objective optimization. arXiv preprint arXiv:2503.03710 , 2025.
- [78] Li Zhou, Kevin Small, Oleg Rokhlenko, and Charles Elkan. End-to-end offline goal-oriented dialog policy learning via policy gradient. In NeurIPS 2017 Workshop on Conversational AI , 2017.

## A 证明和辅助引理

本节包含我们主要论点的证明和辅助引理。

### A.1 Lemma 2 的证明

我们首先注意到,

$$\mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [\tilde{r}(x, \tau_n)] = \mathbb{E}_{x \sim q} \left[ \frac{1}{\sigma(x)} \mathbb{E}_{\tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n) | x] \right] - C,$$

, 其中  $C = \mathbb{E}_{x \sim q} [\mu(x)/\sigma(x)]$  是与  $\theta$  无关的常数。由于所有  $\mathbb{E}_{\tau_n \sim \pi(\cdot | x; \theta)} [r(x, \tau_n) | x]$  都是通过  $\theta_*$  共同最大化的, 且权重  $1/\sigma(x)$  为非负, 因此  $\theta_*$  也最大化了目标的任何加权组合。由此, 我们的证明完成了。

### A.2 Lemma 3 的证明

使用基本代数,

$$\begin{aligned} \mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [\tilde{r}(x, \tau_n)] &= \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ \tilde{r}(x, \tau_n) \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right] \\ &= \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ \tilde{r}(x, \tau_n) \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right] + \Delta(\theta) \\ &= \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [\tilde{r}(x, \tau_n) \log \pi(\tau_n | x; \theta)] + \Delta(\theta) + C_1, \end{aligned}$$

其中

$$\Delta(\theta) = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ \tilde{r}(x, \tau_n) \left( \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} - \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right) \right]$$

并且  $C_1$  是一个在 Lemma 1 中定义的不依赖于  $\theta$  的常数。现在我们重新排列等式, 取两边的绝对值, 并得到

$$\begin{aligned} |\mathbb{E}_{x \sim q, \tau_n \sim \pi(\cdot | x; \theta)} [\tilde{r}(x, \tau_n)] - \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} [\tilde{r}(x, \tau_n) \log \pi(\tau_n | x; \theta)]| &= |C_1 + \Delta(\theta)| \\ &\leq |C_1| + |\Delta(\theta)|. \end{aligned}$$

我们对  $|\Delta(\theta)|$  进行界定为

$$\begin{aligned} |\Delta(\theta)| &\leq \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot | x)} \left[ \left| \tilde{r}(x, \tau_n) \left( \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} - \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right) \right| \right] \\ &\leq \max_{x, \tau_n} \left| \tilde{r}(x, \tau_n) \left( \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} - \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right) \right| \\ &\leq b \max_{x, \tau_n} \left( \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} - \left( 1 + \log \frac{\pi(\tau_n | x; \theta)}{\pi_0(\tau_n | x)} \right) \right). \end{aligned}$$

最后一步成立是因为奖励在  $[-b, b]$  和  $u \geq 1 + \log u$  中。最后, 为了界定  $|\Delta(\theta)|$ , 我们对  $\theta$  进行最大化。这就完成了证明。

## B 详细的相关工作

相关工作可以分为两类: 用于多轮多模态生成 (例如, MLLMs) 的澄清问题技术, 或用于文本到文本生成 (例如, LLMs) 设置的技术。我们还讨论了模拟用户对话轨迹的相关工作, 以及为其他问题设置提出的强化学习方法。

### B.1 监督学习

最近, 许多研究通过询问澄清性问题来关注澄清用户提示。Liu et al. [39] 收集了一组 1,645 语言示例和不同的歧义标签。这是由于存在多种不同类型的歧义。Zelikman et al. [72] 引入了一种简单且有影响力的方法: 通过在成功示例上进行微调并为失败重新生成推理来从推

理中学习。给定一个提示，生成一个推理和答案。如果答案正确，则在提示、推理和答案上进行微调。否则，使用正确答案生成一个新推理，以导向正确答案。在提示、推理和答案上进行微调。自此以来，这一想法已在多个方向上得到扩展。V-STaR [28] 将这一想法扩展到视觉语言任务，Quiet-STaR [71] 侧重于学习何时不该询问，优化策略以最小化不必要的查询。我们在 Appendix B.4 中讨论了对强化学习的扩展。最近，Deng et al. [17] 对主动会话技术进行了调查，其中包括那些专注于通过询问澄清问题以消除歧义及类似技术。

使用大型语言模型进行主动消歧也在最近得到了研究 [31, 76, 9]。AskToAct [76] 通过自我纠正机制来提高工具使用的清晰度。他们生成一个数据集然后在其上进行微调。Kobalczuk et al. [31] 基于信息增益选择澄清问题。他们的方法强调使用预训练的大型语言模型进行推理时间的推理，而我们学习特定任务的策略，这些策略直接且高效地优化提问，而不需要对所有可能的回答进行推理时间计算。

最近的研究也集中在以澄清为目的的用户和代理之间的多轮对话的基准测试 [9]。Zhang et al. [75] 引入了一个基准数据集，并提出了一种称为澄清-执行-计划 (Clarification-Execution-Planning, CEP) 的方法，该方法使用专门的代理来进行澄清、执行和计划。他们预测问题是否需要澄清，然后生成一个澄清。

许多研究也集中在预测对话界面中是否需要澄清的问题上 [40, 9]。最近的一项研究由 [40] 探讨了一种在对话搜索中进行澄清检测的零射方法。他们使用一个具有 LLM 骨干的分类器来预测查询是具体还是模糊的。训练数据是使用零射 LLM 生成的。Li et al. [36] 重点研究如何在产品搜索中学习提出关键问题，使用结合隐式会话反馈和主动澄清的双重学习模型。

调查进一步综合了该领域。Mulla and Gharpure [46] 回顾了自动问题生成的进展，包括早期的强化学习尝试，指出 RL 在通过考虑对话序列中的  $n$  轮次累积损失来改善对话流的能力。此外，Zhang et al. [74] 调查了会话分析如何在 LLM 时代有所帮助。他们讨论了使用 RL 优化会话以改进会话策略学习。该论文还涉及使用 RL 调整 LLM 以进行目标导向的对话，尽管并未专门关注提问。

## B.2 使用多模态模型的监督学习

多模态多轮对话进行文本到图像生成的研究也被用于提出澄清性问题，以消除歧义并改善生成结果 [23]。特别是，Hahn et al. [23] 引入了一种不确定性驱动的方法，当系统信心较低时自适应地触发澄清性问题，从而提升多轮生成性能。该工作还开发了一个用于模拟用户评估问题提问策略的自动评估框架，使用了一组简单代理，包括基于规则的、基于信念的和基于 LLM 的方法，然而，这些方法都没有结合任何基于学习的优化。

相反，Villegas et al. [63] 提出了 ImageChain，它通过将一系列图像视为与生成的文本描述一起进行多轮对话，从而专注于 MLLMs 中的图像到文本推理，以创建简洁的叙述，这在视频生成中有应用。在图像和文本上的顺序推理。下一个图像（被视为代理）的描述取决于该图像（被视为用户）和对话的历史。

Chen et al. [12] 的其他工作关注于提高语音对话的多模态理解。他们利用口语语言来改善多模态对话。该工作构建了一个回合偏好的数据集，标注出胜出的和失败的回应，并在每个步骤应用直接偏好优化 (DPO)。相比之下，我们的工作三个关键方面进行了改进：(1) 我们采用了一种更有原则的目标驱动仿真策略；(2) 我们完全消除了对 DPO 的需求，因为奖励被明确定义，基于直接奖励的策略梯度更简单高效；以及 (3) 我们为方法提供了形式化的证明。

## B.3 经典强化学习

关于 2020 年之前在对话优化领域的强化学习工作的概述，请参见 Levine et al. [35] 的第 6.6 节。最相关的工作是 Zhou et al. [78]，该工作提出了在线和离线策略梯度。他们有每步的奖励和一个固定的轨迹数据集。他们仅专注于提高语言质量，没有任何大语言模型或模拟器。

先前工作的一个重要子集集中于使用强化学习学习何时以及问什么。例如，DialFRED [22] 训练了一个基于强化学习的提问代理，以决定提出什么问题来完成家庭任务，并且会对无效问题进行惩罚。Sigaud et al. [58] 使用强化学习来训练一个代理进行提问。它使用问题生成和问题回答系统来创建奖励塑造的辅助目标，提高语言条件强化学习中的样本效率。

此外，Free et al. [20] 利用 Q-learning 与 DQN 及 BERT 嵌入来训练一个聊天机器人，通过向模拟用户提出战略性问题来获取隐藏的网格世界信息。在会话推荐的领域，Lin [38] 将问题选择框定为一个 bandit 优化问题，旨在减少不必要的查询，同时还探索 LLMs 的强化学习微

调以实现类人对话。同样地，Wang and Ai [66] 使用强化学习来训练一个 DQN 模型用于会话搜索中的风险控制，重点是何时提出澄清性问题。强化学习代理学习在提出相关问题的奖励与不相关问题的惩罚之间取得平衡。

最后，Väth et al. [62] 引入了一个用于评估 LLMs 多轮强化学习的基准 (LMRL-Gym)，其目标是通过精心设计的问题来实现有意的交互，具体通过 Q 学习和 DQN 优化。

## B.4 使用大型语言模型的强化学习

关于使用大型语言模型进行强化学习，Hong et al. [25] 使用离线强化学习来优化目标导向的对话，利用大型语言模型模拟类似人类的互动并生成用于训练的数据。它解决了大型语言模型在轮对话中提出有效问题和优化会话结果的限制。该方法在生成的数据集上训练离线强化学习。强化学习算法是经典的：隐式语言 q 学习。我们希望避免使用价值和 Q 函数。

一个密切相关的工作是通过 STaR-GATE [2] 学习提出澄清问题。他们的算法结合了交互式对话和仿真器偏好的引导，在最佳回应上进行微调。该工作利用优化代理和用户之间的模拟轨迹来收集训练数据。然后它依赖于监督学习：在最具奖励的轨迹上进行 SFT 用于微调原始 LLM。该方法未能充分利用奖励信号，因为 SFT 等同于将所有最佳演示视为同等最佳。这导致统计效率降低以及捕捉细微训练信号的能力有限，而我们的方法通过保留并利用完整的奖励结构来解决这一问题。

此外，RL-STaR [10] 在强化学习框架中提供了对 STaR 风格更新的理论分析。另一个相关的工作是学习如何辅导 [55]，它利用优化代理和用户之间的模拟轨迹来收集训练数据。随后，它应用 DPO 从赢输轨迹对中学习。这种方法未能充分利用奖励信号，因为 DPO 将奖励信息简化为二元对比偏好，舍弃了更细粒度的区分。这导致统计效率降低以及捕捉微妙训练信号的能力受限，而我们的方法通过保留和利用完整的奖励结构来解决这些问题。

Chen et al. [13] 的一项工作研究了在基于 LLM 的对话中进行消歧，并开发了一种基于 DPO 的方法，用于缺乏高质量对话轨迹的特定任务案例，例如数据问答和 SQL 生成。与上述其他工作不同，这些工作侧重于在 MLLMs 中为了消歧而生成澄清问题，本工作开发了一种简化的 LLM 澄清问题生成方法，该方法仅将文本作为输入，并仅生成文本作为输出（无论是代码、数据还是其他类型的文本）。这肯定是 RL。类似于 [55]，但应用于多模态模型。此外，Chi et al. [14] 学会了在信息检索中提出澄清问题。其核心思想是模拟潜在的澄清问题和用户响应，然后对那些在排名指标上取得最大改善的进行微调。这不是 RL，但这个想法类似于我们的 SFT RL 基线。

此外，Chu et al. [15] 调查了 SFT 和 RL 在泛化和记忆方面的表现，并发现 RL 在一些文本和视觉任务中在基于规则的文本和视觉环境中更具泛化能力，而 SFT 主要记忆了训练数据，并且在非同分布设置中无法泛化。这是一种方法论。有趣的是，我们展示了一种联系，因为 RL 可以被视为加权的 SFT。另一个由 Arik et al. [3] 进行的研究通过使用基于行动的对比自我训练 (ACT) 提高了会话技能，特别是澄清问题的提出。ACT 是一种基于 DPO 的高效样本对话策略学习算法。尽管 RLHF 被提到作为构建会话代理的一个范式，该论文的主要贡献不直接在于使用 RL 来提出问题，而是在于 DPO。Wang et al. [65] 使用强化学习增强任务导向的对话系统，着重改善理解和生成任务。它在整个令牌生成过程中引入了逐步奖励，以优化对话状态追踪和响应生成。该方法是 PPO 的一种变体，并且专注于单独的令牌生成。

## B.5 用于 LLM 后训练的离线 RL 算法

文献中众所周知，将基于大型语言模型 (LLM) 的生成视为一个顺序决策过程时，状态包含生成的所有历史 token，动作为下一个生成的 token，转移函数是状态 token 与动作 token 的确定性拼接。因此，从强化学习 (RL) 环境的角度来看，唯一缺失的组件是奖励函数，它是外部于 LLM 的并且需要提供。因此，对于 LLM 来说，线上和离线 RL 的关键区别在于奖励函数的可用性。在关于 RLHF [48] 的最早文章之一中，作者将从用户收集的离线反馈数据转换为学习一个奖励函数，然后使用一个在线 RL 算法 (PPO) 来训练 LLM。另一部分工作尝试探索使用离线 RL 方法结合用户反馈来训练 LLM。其中一种方法是 ILQL [59]，其核心思想是学习一个 Q 函数，其中 LLM 的隐藏状态形成该 Q 函数的特征。在这种情况下，也需要用户的某种形式的数值奖励，但这可以是完全离线的。这里的关键考虑是标准的离线 RL 注意事项，例如确保在 Bellman 更新中保持在训练数据分布内（保守的 QL）以及在推理期间估计和使用 Q 值的增加复杂性。受 KL 约束策略优化目标启发的算法，如 DPO [50]，也以离线方式运行，其目标是有效地学习与用户收集的偏好数据一致的隐式奖励函数。然而，



成对偏好数据的收集是该方法的一个关键要求。在 Baheti et al. [4] 中提供了更详细的关于各种用于 LLM 后训练的离线基于策略的 RL 算法的讨论。

我们特别考虑了两个基于策略的离线强化学习算法——DPO 和 ALOL 的目标函数，以说明它们与我们的方法之间的关键区别：

$$\nabla J(\theta)_{DPO} = \beta \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot|x)} \left[ \sigma(\hat{r}(x, \tau_{nl}) - \hat{r}(x, \tau_{nw})) \left[ \sum_{t=1}^{nw} \nabla \log \pi(a_t | x, \tau_{t-1}; \theta) - \sum_{t=1}^{nl} \nabla \log \pi(a_t | x, \tau_{t-1}; \theta) \right] \right],$$

$$\nabla J(\theta)_{A-LOL} = \mathbb{E}_{x \sim q, \tau_n \sim \pi_0(\cdot|x)} \left[ A_{\pi_0}(x, \tau_n) \hat{r}(x, \tau_n) \sum_{t=1}^n \nabla \log \pi(a_t | x, \tau_{t-1}; \theta) \right],$$

其中  $nw, nl$  分别表示选择和拒绝序列的索引， $\hat{r}$  表示相对于参考策略的倾向比例， $A_{\pi_0}$  表示参考策略下的优势函数。我们注意到，这两个梯度估计都可以视为离策略普通策略梯度的缩放版本，在这两种情况下的缩放因子都是被优化策略和参考策略下倾向比例的函数。在我们的表述中，我们避免这些缩放因子以确保稳定性和简洁性，同时为目标提供一个接近原始目标的宽松下界。

## C 数据集

在本节中，我们对所讨论的六个基准数据集进行了全面总结，并介绍了实验设置：

**OpenBookQA [43]** 是一个问答数据集，以开卷考试的形式为模型，包含 5,957 道多项选择的小学科学题（4,957 用于训练，500 用于开发，500 用于测试）。它考察对一本包含 1,326 个核心科学事实的小“书”的理解及其在新情况下的应用。本数据集的挑战在于回答问题需要超出所提供“书”中内容的额外常识。

**SciQA [68]** 是一个多模态数据集，用于评估 AI 模型在科学主题中使用文本和视觉信息进行推理的能力。它包含大约 21,000 个多模态问题，涵盖物理、化学和生物学，来自教育材料。模型必须分析文本和图表以生成正确的答案。

**MMLU [24]** 是一个综合基准，通过多项选择题在包括 STEM、人文、社会科学等在内的 57 个学科中对模型进行评估，难度水平从小学到高级专业水平不等。它专注于零样本和少样本设置，使其类似于我们评估人类的方式。该基准测试既考察世界知识又考察解决问题的能力。

**ARC [16]** 是一个包含 7,787 个真正的小学到初中三年级至九年级的多项选择科学问题的数据集。它分为两个部分：挑战集包含 2,590 个“难题”，在使用检索和共现方法时无法正确回答，以及包含 5,197 个问题的简单集。大多数问题有 4 个答案选项，小于 1

**CoSQL [70]** 是一个用于构建跨领域对话式文本到 SQL 系统的语料库。它由超过 30,000 个对话轮次以及超过 10,000 个注释过的 SQL 查询组成，这些数据来自一个奥兹巫师风格收集的 3,000 个对话，这些对话查询了跨越 138 个领域的 200 个复杂数据库。每个对话模拟一个实际的数据库查询场景，与一个作为用户的众包工作者探索数据库以及一个 SQL 专家使用 SQL 检索答案。CoSQL 中的平均问题长度为 11.2 个单词，每个对话平均有 5.2 个问题轮次。

**MathDial [41]** 是一个基于多步骤数学推理问题的一对一师生辅导对话的数据集。该数据集共包含 2,861 个对话，分为训练集和测试集。数据集通过将人类教师与大型语言模型 (LLM) 配对而创建，该模型被提示代表常见学生错误，并使用 LLMKT 模型 [54]。该数据集侧重于有效辅导，而不仅仅是解决问题，并展示了丰富的教学属性，重点通过引导性问题来帮助学生进行理解。

在我们的实验中，我们从每个数据集中随机选择了 500 个样本，其中 400 个用作训练，100 个用作测试。我们创建了 3 轮对话，并使用我们的 **Base** 模型在不同的温度设置下生成了 3 个随机运行（轨迹）。

我们在本节中展示了跨越各种领域的不同提示示例。

### C.1 ARC 提示示例

我们首先展示来自 ARC 数据集的三步骤多轮对话中的前两个步骤的对话：

Question: Which of the following best explains how stems transport water to other parts of the plant?  
Choices: A. through a chemical called chlorophyll B. by using photosynthesis C. through a system of tubes D. by converting water to food

This question is from a science exam. Please use <thinking></thinking> tags to show your step-by-step reasoning, then provide your initial thoughts outside of these tags.

The answer is C. <thinking> To determine the correct answer, I will evaluate each option in relation to how stems transport water to other parts of a plant. A. Chlorophyll is a chemical that is essential for photosynthesis, but it is not directly responsible for transporting water through the plant. Chlorophyll is involved in absorbing light energy, which is necessary for photosynthesis to occur, but it is not the mechanism by which water is transported. B. Photosynthesis is the process by

That's a good start. Can you explain more about why some options might be incorrect? Use <thinking></thinking> tags for your analysis.

<thinking> Photosynthesis is the process by which plants convert light energy into chemical energy in the form of glucose. While photosynthesis occurs in the leaves, which are attached to the stems, it is not the process by which water is transported to other parts of the plant. The question specifically asks about the mechanism of water transport, not the process of energy conversion. C. The xylem is a type of vascular tissue that is found in the stems of plants, which is responsible for transporting water and

Thank you for your detailed explanations. What is your final answer (A, B, C, or D)? Please provide a justification for your choice. You MUST first state your final answer in the format: 'The answer is X' where X is A, B, C, or D. The final answer must be outside the thinking tags. Then show your thinking in <thinking></thinking> tags for your step-by-step reasoning.

Base 的最终响应和 GPT4o 的评估

我们可以从最终的回应中观察到，Base 非常冗长。

</thinking> The xylem consists of dead cells that are hollow, which allows water to pass through them. This is why option A is incorrect. Chlorophyll is a chemical used in photosynthesis, not in transporting water. Photosynthesis is the process of plants converting light energy into chemical energy, but it's not related to water transport. Therefore, option B is also incorrect. Option D is incorrect because converting water to food is the process of photosynthesis

Evaluation Scores by GPT4o										
Criteria	Score	Rating								
Accuracy	8/10	■	■	■	■	■	■	■	■	■
Reasoning Ability	7/10	■	■	■	■	■	■	■	■	■
Comprehensiveness	6/10	■	■	■	■	■	■	■	■	■
Pedagogical Value	7/10	■	■	■	■	■	■	■	■	■
Confidence Calibration	6/10	■	■	■	■	■	■	■	■	■
Overall	7.0/10	■	■	■	■	■	■	■	■	■

### Evaluation Justifications by GPT4o

#### C.1.1

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.1.2

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.1.3

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.1.4

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

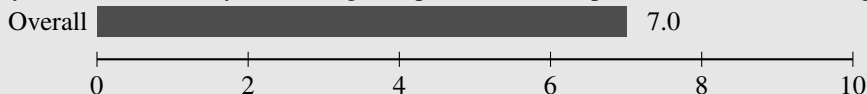
#### C.1.5

置信校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

Final Score: 7.0/10

Summary: The conversation is on the right track and provides a mostly accurate and logical explanation of the question. However, incomplete reasoning for option D and a lack of explicit final answer detract from the overall quality. Expanding on the xylem's role and fully addressing all options would improve the conversation significantly.



### SWiFt 的最终回应和 GPT4o 的评估

我们现在报告对话的最后一轮，如下所示。我们可以从 SWiFt 的最终回应中观察到，它的回应与之前回合的回应（以及 Base）有显著不同。它通过额外的观察/知识更新了其推理过程，因此从 GPT4o 获得了更高的评价。

choice C is the best explanation.

</thinking>I have now fully explained why choice C is the best explanation for how stems transport water to other parts of the plant. Choice C states that stems transport water through a system of tubes, which is a more accurate description of the plant's water transport system. This system, known as xylem, is a network of tubes that allows water to move from the roots to the leaves of the plant. This is a fundamental process that supports the plant's

Reasoning Ability 7/10 





























































Comprehensiveness 6/10 

Pedagogical Value 7/10 

Confidence Calibration 6/10 

Overall 7.0/10 

#### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	10/10										
Reasoning Ability	10/10										
Comprehensiveness	10/10										
Pedagogical Value	10/10										
Confidence Calibration	10/10										
Overall	10.0/10										

#### C.1.6

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.1.7

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.1.8

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.1.9

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

#### C.1.10

置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.



### Evaluation Justifications by GPT4o

#### C.1.11

准确性: 10/10 The assistant correctly identifies option C as the answer and provides a scientifically accurate explanation of how stems transport water through the xylem system. The explanation is factually correct and directly addresses the question.

#### C.1.12

推理能力: 10/10 The reasoning is excellent, with a clear and logical analysis of each option. The assistant methodically explains why each incorrect option is wrong and why the correct option is right, demonstrating strong critical thinking skills.

#### C.1.13

全面性: 10/10 The response is exceptionally thorough, addressing all four options with detailed explanations. The assistant fully explains the role of the xylem in water transport and clearly articulates why the other options are incorrect.

#### C.1.14

教学价值: 10/10 The explanation is highly educational, providing clear distinctions between different plant processes (photosynthesis vs. water transport) and structures (chlorophyll vs. xylem). The response teaches valuable plant biology concepts in an accessible way.

#### C.1.15

置信度校准: 10/10 The assistant demonstrates appropriate confidence in the answer, providing a direct statement followed by detailed supporting evidence. The thinking process is transparent and the conclusion is well-justified.

### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	8/10	■	■	■	■	■	■	■	■	□	□
Reasoning Ability	7/10	■	■	■	■	■	■	■	□	□	□
Comprehensiveness	6/10	■	■	■	■	■	■	□	□	□	□
Pedagogical Value	7/10	■	■	■	■	■	■	■	□	□	□
Confidence Calibration	6/10	■	■	■	■	■	■	□	□	□	□
Overall	7.0/10	■	■	■	■	■	■	■	□	□	□

### Evaluation Justifications by GPT4o

#### C.1.16

准确率: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.1.17

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.1.18

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.1.19

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.


#### C.1.20

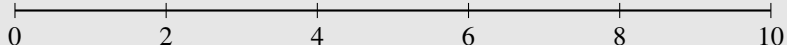
置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

**Final Score: 10.0/10**

Summary: This response is exemplary in every aspect. The assistant clearly identifies the correct answer and provides a comprehensive explanation of plant water transport systems. Each option is thoroughly analyzed with scientific accuracy, and the distinction between water transport mechanisms and other plant processes is clearly articulated. The thinking process is transparent and educational, making this an ideal model response for science education.

Overall  10.0



## C.2 CoSQL 提示示例

我们首先展示来自 CoSQL 数据集中三步多轮对话的前两步对话:

Question: Your task is to generate SQL query based on the following question and information provided. Which distinctive models are produced by maker with the full name General Motors or weighing more than 3500? Hint: Consider relationships between multiple tables. The relevant schema tables are: ['car\_makers', 'model\_list', 'car\_names', 'cars\_data']. The optional foreign keys are: ['CAR\_NAMES.model = MODEL\_LIST.model', 'MODEL\_LIST.maker = CAR\_MAKERS.id', 'CARS\_DATA.id = CAR\_NAMES.makeid']. Choices: A. A. SELECT DISTINCT T1.Model, T3.FullName FROM CARS\_DATA AS T4 JOIN CAR\_NAMES AS T1 ON T4.Id = T1.MakeId LEFT JOIN CAR\_MAKERS AS T3 ON T1.MakeId = T3.Id WHERE T3.FullName LIKE ' % General % ' OR T4.weight < 3500; B. B. SELECT DISTINCT Model FROM MODEL\_LIST

WHERE Maker IN (SELECT Id FROM CAR\_MAKERS WHERE FullName = 'General Motors' AND Id IN (SELECT MakeId FROM CARS\_DATA WHERE weight > 3500)); C. C. SELECT DISTINCT T2.Model FROM CAR\_NAMES AS T1 JOIN MODEL\_LIST AS T2 ON T1.Model = T2.Model JOIN CAR\_MAKERS AS T3 ON T2.Maker = T3.Id JOIN CARS\_DATA AS T4 ON T1.MakeId = T4.Id WHERE T3.FullName = 'General Motors' OR T4.weight > 3500; D. D. SELECT T1.Model FROM CAR\_NAMES AS T1 WHERE T1.Model IN (SELECT Model FROM CARS\_DATA WHERE weight > 3500) AND T1.MakeId IN (SELECT Id FROM CAR\_MAKERS WHERE FullName = 'General Motors');

This question asks for the correct SQL query to retrieve the requested information from a database. Focus on SQL syntax, correct table joins, and query efficiency. Please think through this step by step and explain your initial thoughts about the question.

«SYS» The answer is C.

To determine the correct SQL query, let's break down the requirements step by step:

1. **Distinctive models**: We need to find distinct models that match the given conditions. This means we should use the 'SELECT DISTINCT' clause.
2. **Maker with full name General Motors**: This indicates we need to filter the data based on the 'FullName' of the maker, which is stored in the 'CAR\_MAKERS' table.
- 3.

That's a good start. Can you explain more about why some options might be incorrect? Also, are there any key concepts or facts that are relevant to answering this question?

«SYS»

To evaluate the options, let's consider the following key concepts and facts:

**Joining tables**: We need to join multiple tables to retrieve the required information. The correct join order and type are crucial. **Filtering conditions**: We have two conditions to apply: the maker's full name should be 'General Motors' or the car's weight should be more than 3500. **Using foreign keys**: The foreign keys are defined as 'CAR

Thank you for your detailed explanations. What is your final answer (A, B, C, or D)? You MUST first state your final answer in the format: 'The answer is X' where X is A, B, C, or D. After the final answer clearly explain your reasoning.

Base 的最终响应和 GPT4o 的评估

我们可以从最终的回答中观察到，Base 基本上遵循了前几轮回答中的推理。它正朝着正确答案前进，但推理并不全面。

<SYS> group by clause, so option D is not correct. Now, let's consider the joins and subqueries in the other options.

The answer is C.

To solve this problem, we need to find the instructors who did not teach any courses. This can be achieved by identifying instructors who do not have a matching id in the 'teaches' table.

We can use a subquery to select the ids of instructors who taught courses. Then, we can use the

Reasoning Ability 7/10 

Comprehensiveness 6/10 

Pedagogical Value 7/10 

Confidence Calibration 6/10 

Overall 7.0/10 

Evaluation Scores by GPT4o											
Criteria	Score	Rating									
Accuracy	7/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Reasoning Ability	6/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comprehensiveness	5/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pedagogical Value	6/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Confidence Calibration	7/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall	6.0/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

### C.2.1

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

### C.2.2

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

### C.2.3

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

### C.2.4

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

### C.2.5

置信校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Evaluation Justifications by GPT4o

#### C.2.6

准确性: 7/10 The assistant is moving toward the correct answer (C) and has identified it as the correct choice. However, the explanation provided so far is incomplete, and the assistant has not yet fully justified why C is correct or why the other options are incorrect.

#### C.2.7

推理能力: 6/10 The reasoning is partially logical and clear, as the assistant has identified the need for 'SELECT DISTINCT', proper joins, and filtering conditions. However, the explanation lacks depth and precision, particularly in explaining the relationships between tables and why certain options fail to meet the requirements.

#### C.2.8

全面性: 5/10 Key concepts like table joins, filtering conditions, and foreign key relationships are mentioned, but they are not fully explained. The assistant has not yet addressed why specific options (A, B, and D) are incorrect, which is critical for a comprehensive analysis.

#### C.2.9

教学价值: 6/10 The explanation has some educational value, as it introduces important SQL concepts like 'SELECT DISTINCT', filtering, and table joins. However, the lack of detailed reasoning and comparison between options limits its effectiveness as a learning resource.

#### C.2.10

置信度校准: 7/10 The assistant confidently identifies C as the correct answer, which is accurate. However, the confidence is slightly undermined by the incomplete reasoning and lack of detailed analysis of the other options.

### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	8/10	■	■	■	■	■	■	■	■	□	□
Reasoning Ability	7/10	■	■	■	■	■	■	■	□	□	□
Comprehensiveness	6/10	■	■	■	■	■	■	□	□	□	□
Pedagogical Value	7/10	■	■	■	■	■	■	■	□	□	□
Confidence Calibration	6/10	■	■	■	■	■	■	□	□	□	□
Overall	7.0/10	■	■	■	■	■	■	■	□	□	□



### Evaluation Justifications by GPT4o

#### C.2.11

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.2.12

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.2.13

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.2.14

教育价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

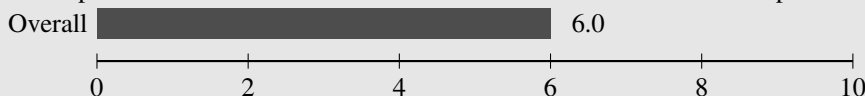
#### C.2.15

置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

**Final Score: 6.0/10**

Summary: The assistant is on the right track and has correctly identified the answer, but the explanation lacks depth, clarity, and comprehensiveness. While some key concepts are mentioned, the conversation would benefit from a more thorough analysis of why the other options are incorrect and a clearer breakdown of the relationships between tables.



### SWiFt 的最终回应和 GPT4o 的评估

我们现在报告这里显示的对话的最后一轮。从 SWiFt 的最终回应中可以观察到，它的回应不同于上一轮的回应（以及 Base），并且被 GPT4o 判断为更准确。

«SYS» The answer is C.

To determine the correct SQL query, let's break down the requirements step by step:

1. **\*\*Distinctive models\*\***: We need to find distinct models that match the given conditions. This means we should use the 'SELECT DISTINCT' clause.
2. **\*\*Maker with full name General Motors\*\***: This indicates we need to filter the data based on the 'FullName' of the maker, which is stored in the 'CAR\_MAKERS' table.
- 3.























































Reasoning Ability 7/10 

Comprehensiveness 6/10 

Pedagogical Value 7/10 

Confidence Calibration 6/10 

Overall 7.0/10 

Evaluation Scores by GPT4o										
Criteria	Score	Rating								
Accuracy	8/10									
Reasoning Ability	7/10									
Comprehensiveness	6/10									
Pedagogical Value	5/10									
Confidence Calibration	7/10									
Overall	6.6/10									

#### C.2.16

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.2.17

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.2.18

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.2.19

教育价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

#### C.2.20

置信校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Evaluation Justifications by GPT4o

#### C.2.21

准确性: 8/10 The response correctly identifies the correct answer as "C" and provides reasoning aligned with the question's requirements. However, the explanation is incomplete and does not fully justify why "C" is correct or why other options are incorrect.

#### C.2.22

推理能力: 7/10 The reasoning is partially sound, as it breaks down the requirements of the query and links them to the SQL components. However, the explanation is truncated and does not fully analyze the relationships between tables or the logic of the query.

#### C.2.23

全面性: 6/10 The response addresses some key aspects of the question, such as the need for 'SELECT DISTINCT' and filtering by 'FullName'. However, it does not explore the schema relationships, optional foreign keys, or why the other options fail to meet the requirements.

#### C.2.24

教学价值: 5/10 The response provides some educational value by explaining the use of 'SELECT DISTINCT' and filtering conditions. However, it lacks depth and does not guide the reader through the full reasoning process or clarify why the correct query works.

#### C.2.25

置信校准: 7/10 The response confidently states that the correct answer is "C" and provides some justification. However, the incomplete explanation leaves room for doubt about whether the reasoning is fully understood.

### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	8/10	■	■	■	■	■	■	■	■	□	□
Reasoning Ability	7/10	■	■	■	■	■	■	■	□	□	□
Comprehensiveness	6/10	■	■	■	■	■	■	□	□	□	□
Pedagogical Value	7/10	■	■	■	■	■	■	■	□	□	□
Confidence Calibration	6/10	■	■	■	■	■	■	□	□	□	□
Overall	7.0/10	■	■	■	■	■	■	■	□	□	□

### Evaluation Justifications by GPT4o

#### C.2.26

准确度: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.2.27

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.2.28

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.2.29

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

#### C.2.30

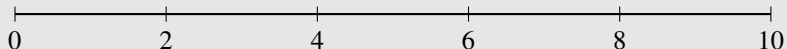
置信校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

Final Score: 6.6/10

Summary: While the response identifies the correct answer and provides some reasoning, it lacks thoroughness, depth, and a complete analysis of the question and options. Improvements in comprehensiveness and pedagogical clarity are needed.

Overall  6.6



### C.3 MathDial 提示范例

我们现在展示来自 MathDial 数据集的提示示例。

Drew is reseeding his lawn with grass seed. One bag of grass seed covers 250 square feet of lawn. His lawn is 22 feet from the house to the curb and 36 feet from side to side. He bought four bags of seed. How many extra square feet could the leftover grass seed cover after Drew reseeds his lawn?

A student attempted to solve this problem as follows: The area of Drew's lawn is  $22 \times 36 = 792$  square feet. Four bags of grass seed can cover  $4 \times 250 = 1000$  square feet. Drew will use all the grass seed to cover his lawn, so there won't be any leftover to cover extra square feet. Therefore, the leftover grass seed can't cover any extra square feet. 0

Which of the following is the correct final answer? A. 177 B. 0 C. 138 D. 208

Please use <thinking></thinking> tags to show your step-by-step reasoning.

<<SYS>>[/SYS]

The answer is B. <thinking> The student correctly calculated the area of the lawn, which is  $22 \times 36 = 792$  square feet. Then, they calculated the total area that the four bags of grass seed could cover, which is  $4 \times 250 = 1000$  square feet. However, they incorrectly concluded that there would be no leftover grass seed because the area of the lawn (792 square feet) is less than the total area that the grass seed could cover (1000 square feet). This

Can you explain more about why some options might be incorrect? Use <thinking></thinking> tags for your analysis.

</SYS>

The answer is B. <thinking> The student's conclusion is incorrect because they didn't take into account the fact that the lawn is a rectangle and the grass seed can only cover the area of the lawn. The total area that the grass seed can cover (1000 square feet) is greater than the area of the lawn (792 square feet), but that doesn't mean that the entire lawn can be covered with the grass seed. In fact, the student's calculation

Thank you for your detailed explanations. What is your final answer (A, B, C, or D)? Please provide a justification for your choice. You MUST first state your final answer in the format: 'The answer is X' where X is A, B, C, or D. The final answer must be outside the thinking tags. Then show your thinking in <thinking></thinking> tags for your step-by-step reasoning.

Base 的最终回应和 GPT4o 的评估

</SYS>

The answer is B.

<thinking> The student's conclusion is incorrect because they didn't take into account the fact that the lawn is a rectangle and the grass seed can only cover the area of the lawn. The total area that the grass seed can cover (1000 square feet) is

Reasoning Ability 7/10 ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☐ ☐ ☐

Comprehensiveness 6/10 ☒ ☒ ☒ ☒ ☒ ☒ ☐ ☐ ☐ ☐

Pedagogical Value 7/10 ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☐ ☐ ☐

Confidence Calibration 6/10 ☒ ☒ ☒ ☒ ☒ ☒ ☐ ☐ ☐ ☐

Overall 7.0/10 ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☐ ☐ ☐

#### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	3/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Reasoning Ability	4/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comprehensiveness	3/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Pedagogical Value	4/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Confidence Calibration	2/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Overall	3.0/10	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

#### C.3.1

准确率: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.



### C.3.2

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

### C.3.3

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

### C.3.4

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

### C.3.5

置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

#### Evaluation Justifications by GPT4o

##### C.3.6

准确性: 3/10 The assistant is moving toward the correct answer but has not yet explicitly identified the correct choice (D). The reasoning provided so far contains errors and does not correctly calculate the leftover area of grass seed.

##### C.3.7

推理能力: 4/10 The reasoning is partially logical but lacks clarity and precision. The assistant identifies that the lawn area is smaller than the total coverage of the grass seed but fails to complete the calculation or address the leftover area properly.

##### C.3.8

全面性: 3/10 Key concepts, such as how to calculate the leftover area ( $1000 - 792 = 208$ ), are not properly explained. The assistant does not address why the other options are incorrect or provide a full step-by-step breakdown.

##### C.3.9

教学价值: 4/10 The explanation has some educational value but is incomplete and potentially confusing. It does not guide the user through the correct process of solving the problem or clarify misconceptions.

##### C.3.10

置信度校准: 2/10 The assistant is overly confident in stating that the answer is B, despite providing flawed reasoning and failing to justify the conclusion properly. Confidence is not appropriately calibrated to the accuracy of the analysis.

### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	8/10	■	■	■	■	■	■	■	■	■	■
Reasoning Ability	7/10	■	■	■	■	■	■	■	■	■	■
Comprehensiveness	6/10	■	■	■	■	■	■	■	■	■	■
Pedagogical Value	7/10	■	■	■	■	■	■	■	■	■	■
Confidence Calibration	6/10	■	■	■	■	■	■	■	■	■	■
Overall	7.0/10	■	■	■	■	■	■	■	■	■	■

### Evaluation Justifications by GPT4o

#### C.3.11

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.3.12

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.3.13

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.3.14

教育价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

#### C.3.15

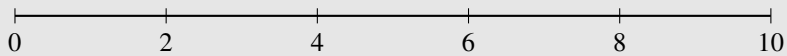
置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

**Final Score: 3.0/10**

Summary: The conversation demonstrates some progress toward solving the problem but lacks accuracy, clarity, and comprehensiveness. The assistant's reasoning is incomplete, and the explanation does not effectively teach the correct approach to solving the question.

Overall                      3.0



SWiFt 的最终响应和 GPT4o 的评估

我们可以从 **SWiFt** 的最终响应中观察到，其响应不同于先前的轮次响应（以及 **Base** 的响应），并且根据 **GPT4o** 的判断更好。

The answer is D.

<thinking> Let me carefully work through this problem step by step:

1. First, I need to calculate the total area of Drew's lawn: Length = 22 feet Width = 36 feet Area of lawn =  $22 \times 36 = 792$  square feet

2. Next, I need to determine the total coverage provided by four bags of grass seed: One bag covers 250 square feet Four bags can cover  $4 \times 250 = 1000$  square feet

3. Since Drew will use the grass seed to cover his lawn first, I need to find how much grass seed will be left over: Leftover grass seed coverage = Total coverage - Lawn area Leftover grass seed coverage =  $1000 - 792 = 208$  square feet

Looking at the answer choices: - A. 177 square feet - This is incorrect. It's not clear how this value was derived. - B. 0 square feet - This is what the student incorrectly concluded, assuming all seed would be used. - C. 138 square feet - This is incorrect. It's not clear how this value was derived. - D. 208 square feet - This matches my calculation of  $1000 - 792 = 208$  square feet.

Therefore, the answer is D. The leftover grass seed could cover an additional 208 square feet.  
</thinking>





























































Reasoning Ability 7/10 

Comprehensiveness 6/10 

Pedagogical Value 7/10 

Confidence Calibration 6/10 

Overall 7.0/10 

Evaluation Scores by GPT4o											
Criteria	Score	Rating									
Accuracy	5/10										
Reasoning Ability	6/10										
Comprehensiveness	4/10										
Pedagogical Value	5/10										
Confidence Calibration	3/10										
Overall	5.0/10										

C.3.16

准确率: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

C.3.17

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

### C.3.18

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

### C.3.19

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

### C.3.20

置信度校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

#### Evaluation Justifications by GPT4o

##### C.3.21

准确性: 5/10 The assistant has correctly calculated the area of the lawn (792 square feet) and the total coverage of four bags of grass seed (1000 square feet). However, the assistant prematurely states that the answer is "A" without fully completing the calculations or verifying the leftover area, which leads to an incorrect conclusion.

##### C.3.22

推理能力: 6/10 The assistant demonstrates some logical reasoning by breaking the problem into steps (calculating the lawn area and total coverage). However, the reasoning is incomplete, as the assistant does not finish analyzing the leftover area or fully explain why "A" is chosen.

##### C.3.23

全面性: 4/10 The explanation is missing key steps, such as subtracting the lawn area from the total coverage to determine the leftover area ( $1000 - 792 = 208$ ). Additionally, the assistant does not fully analyze all the options, leaving the explanation incomplete and confusing.

##### C.3.24

教育价值: 5/10 While the assistant starts with a structured approach, the incomplete reasoning and incorrect conclusion could mislead a learner. The explanation lacks clarity and depth, which limits its educational value.

##### C.3.25

置信校准: 3/10 The assistant confidently states that the answer is "A" without completing the necessary calculations or fully analyzing the problem. This overconfidence is unwarranted given the incomplete reasoning.

#### Evaluation Scores by GPT4o

Criteria	Score	Rating									
Accuracy	8/10	■	■	■	■	■	■	■	■	□	□
Reasoning Ability	7/10	■	■	■	■	■	■	■	□	□	□
Comprehensiveness	6/10	■	■	■	■	■	■	□	□	□	□
Pedagogical Value	7/10	■	■	■	■	■	■	■	□	□	□
Confidence Calibration	6/10	■	■	■	■	■	■	□	□	□	□
Overall	7.0/10	■	■	■	■	■	■	■	□	□	□

### Evaluation Justifications by GPT4o

#### C.3.26

准确性: 8/10 The assistant is moving toward the correct answer, "C," and has correctly identified that the xylem is responsible for water transport. However, the explanation for option D is incomplete, and the assistant has not yet explicitly stated the final answer as requested.

#### C.3.27

推理能力: 7/10 The reasoning is mostly logical and clear, particularly in explaining why options A and B are incorrect. However, the analysis of option D is cut off and incomplete, which detracts from the overall clarity and precision of the reasoning.

#### C.3.28

全面性: 6/10 The assistant provides a good explanation for why options A and B are incorrect and begins to explain why C is correct. However, the discussion of option D is incomplete, and the explanation of the xylem could be expanded further to fully address the mechanism of water transport.

#### C.3.29

教学价值: 7/10 The conversation is educational and provides some useful insights, particularly about chlorophyll and photosynthesis. However, the incomplete explanations for D and the xylem system limit the overall learning potential.

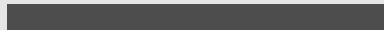
#### C.3.30

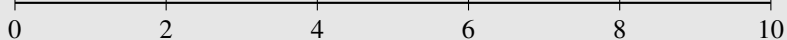
置信校准: 6/10 The assistant appears confident in its reasoning but has not yet explicitly stated the final answer as requested. Additionally, the incomplete explanation of option D suggests a slight overconfidence in the clarity of its analysis.

### Overall Assessment by GPT4o

Final Score: 5.0/10

Summary: The assistant shows some understanding of the problem and begins with a logical approach, but the incomplete reasoning, incorrect answer, and lack of thorough analysis significantly detract from the overall quality of the conversation.

Overall  5.0



我们为 OpenBookQA 数据集进行了 **SWiFt** 和 **ReFit** 的消融研究，随着  $n=4,6,8$  和 10 的增加。这在 ?? - ?? 中显示。我们发现，随着  $n$  值的增大，任务变得更难，R 整体上不断下降。

在本节中，我们展示了在 ?? - ?? 上各种数据集的奖励估计的雷达图。

## D 模型和训练参数

在本节中，我们介绍我们的框架在 Table 13 - Table 17 中的模型配置和训练参数。

## E 政策回应中的语言模式聚类分析

我们应用 UMAP [42] 降维技术来可视化两种语言政策（**Base** 和 **SWiFt**）在响应上的差异。该方法由两个关键步骤组成：



Parameter	Value
vocab_size	128256
max_position_embeddings	131072
hidden_size	4096
intermediate_size	14336
num_hidden_layers	32
num_attention_heads	32
num_key_value_heads	8
hidden_act	silu
initializer_range	0.02
rms_norm_eps	1e-05
pretraining_tp	1
use_cache	true
rope_theta	500000.0
rope_scaling.factor	8.0
rope_scaling.low_freq_factor	1.0
rope_scaling.high_freq_factor	4.0
rope_scaling.original_max_position_embeddings	8192
rope_scaling.rope_type	llama3
head_dim	128
torch_dtype	bfloat16
bos_token_id	128000
eos_token_id	[128001, 128008, 128009]
model_type	llama
architectures	LlamaForCausalLM

Table 13: Llama 3.1 8B 指令配置

Parameter	Value
compute_environment	LOCAL_MACHINE
debug	false
distributed_type	DEEPSPEED
downcast_bf16	no
enable_cpu_affinity	false
machine_rank	0
main_training_function	main
mixed_precision	bf16
num_machines	1
num_processes	2
rdzv_backend	static
same_network	true
tpu_use_cluster	false
tpu_use_sudo	false
use_cpu	false
deepspeed_config	
gradient_accumulation_steps	4
gradient_clipping	1.0
offload_optimizer_device	cpu
offload_param_device	cpu
zero3_init_flag	false
zero3_save_16bit_model	true
zero_stage	2

Table 14: 加速 DeepSpeed 配置

1. 常见词聚类：首先根据相似的常用词使用情况，将回答分为五种不同的语言模式（标记为 0-4）
2. 不常见词汇可视化：然后根据不常见词汇的使用将回答定位在二维空间中

Parameter	Value
compute_environment	LOCAL_MACHINE
debug	false
distributed_type	DEEPSPEED
downcast_bf16	no
machine_rank	0
mixed_precision	bf16
num_machines	1
num_processes	2
use_cpu	false
deepspeed_config	
gradient_accumulation_steps	4
gradient_clipping	1.0
offload_optimizer_device	none
offload_param_device	none
zero3_init_flag	false
zero3_save_16bit_model	true
zero_stage	0

Table 15: 加速 DeepSpeed 配置以进行知识蒸馏

Parameter	Value
Model Configuration	
model_name	Llama-3.1-8B-Instruct
Comments	Customized to do RL Reweighting for <a href="#">ReFit</a> and <a href="#">SWiFt</a>
Training Parameters	
learning_rate	3e-5
num_train_epochs	4
per_device_train_batch_size	8
gradient_accumulation_steps	4
gradient_checkpointing	True
mixed_precision	bf16
do_train	True
do_eval	False
logging_steps	5
logging_first_step	True
save_strategy	epoch
save_total_limit	4
RL Configuration	
dataset	From the listed datasets in this paper.json
rl_reweight	std
rl_reward_name	reward
use_custom_trainer	True
Hardware Configuration	
num_processes	2
num_machines	1

Table 16: TRL 监督微调配置与自定义模型 RL 重加权用于 [ReFit](#) 和 [SWiFt](#)

这种方法使我们能够观察到响应模式中的结构相似性以及政策之间的特定词汇差异。

## E.1 技术实现

### E.1.1 数据处理流水线

可视化流程涉及多个顺序的处理步骤：

Parameter	Value
Model Configuration	
teacher_model_path	STaR-GATE _last-checkpoint
student_model_name	meta-llama/Llama-3.1-8B-Instruct
student_layers	8
apply_lora_to_teacher	True
LoRA Configuration	
r	8
alpha	16
dropout	0.05
target_modules	q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj
Distillation Parameters	
distillation_alpha	0.5
distillation_temperature	2.0
Training Parameters	
learning_rate	3e-6
num_train_epochs	2
per_device_train_batch_size	4
gradient_accumulation_steps	4
gradient_checkpointing	True
mixed_precision	bf16
do_train	True
do_eval	False
logging_steps	5
logging_first_step	True
save_strategy	epoch
save_total_limit	4
Dataset Configuration	
dataset	From the listed datasets in this paper
rl_reweight	SFT
use_custom_trainer	False
Hardware Configuration	
num_processes	2
num_machines	1

Table 17: 使用 LoRA 进行知识蒸馏配置

1. 文本提取：从 JSON 文件中提取模型响应，特别是针对 predicted\_answer 字段。
2. 常用词识别：根据两个政策中的频率分布，词语被分类为“常用”。如果满足以下条件，则认为一个词是常用词：
  - 它在任何政策中以频率  $\geq 0.01$  出现
  - 政策之间的频率比在范围  $[0.8, 1.2]$  内（使用频率比参数 0.2）
  - 标准停用词和领域特定术语（例如，“answer”、“question”、“correct”）总是被包含为常用词
3. 文本拆分：每个回复被拆分为两个部分：
  - 仅包含被分类为常用的单词的文本
  - 仅包含剩余词汇的不常用词文本
4. 常用词聚类：仅包含常用词的文本使用 TF-IDF 进行了向量化（参数为：max\_features=5000, min\_df=2, max\_df=0.9, sublinear\_tf=True），并使用 K-means (k=5, random\_state=42) 进行聚类，以识别五种语言模式。
5. 罕见词向量化：仅包含罕见词的文本同样使用 TF-IDF 和相同的参数进行向量化。
6. UMAP 投影：使用 UMAP 将不常见的词向量投影到二维空间中。

UMAP 算法配置了以下参数：

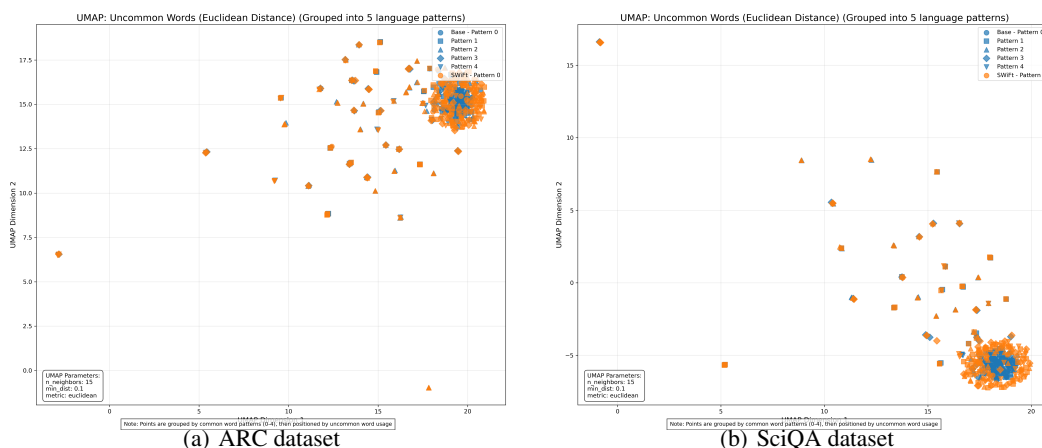


Figure 1: 使用欧氏距离的政策响应的 UMAP 可视化。点根据常见词模式（0-4）进行分组，然后根据不常见的词使用情况进行定位。蓝色点代表 **Base** 政策（Llama3-8.1b-rl）；橙色点代表 **SWiFT** 政策（Llama3-8.1b-instruct）。

## E.1.2 UMAP 配置

- **n\_neighbors = 15**：这个参数控制保持局部结构与全局结构之间的平衡。值为 15 提供了适度的平衡，使得算法能够同时捕捉到相似响应之间的局部关系和整体分布模式。
- **min\_dist = 0.1**：该参数控制点之间的紧密程度。相对较低的值 0.1 允许在点非常相似时形成密集的簇，同时仍然在不同组之间提供分离。
- **metric = "euclidean"**：使用欧几里得距离来衡量向量之间的相似性，从而在高维空间中提供一种直接的几何解释。
- **n\_components = 2**：为了可视化，输出维数被设置为 2。
- **random\_state = 42**：使用一个固定的随机种子以确保不同运行之间的可重复性。

TF-IDF 向量化对于分析至关重要：

- 为了计算效率，最多保留 5,000 个特征
- 在少于 2 个响应中出现的词（**min\_df=2**）被排除以减少噪音
- 出现在超过 90 % 响应中的词语（**max\_df=0.9**）被降低权重，以便专注于具有辨别力的术语
- 对次线性词频缩放进行了应用，以减弱高频词的影响
- L2 归一化用于处理响应长度的变化

这种向量化方法确保了得到的向量能够捕捉到每个响应中术语的相对重要性，同时考虑到了整个语料库的特征。

## E.2 多数据集上的结果

我们将分析应用于两个数据集：ARC 和 SciQA。

### E.2.1 ARC 数据集可视化分析

Figure 1(a) 展示了 ARC 数据集的 UMAP 可视化，该数据集包含小学水平的多项选择科学问题，需要进行推理。几个关键的观察结果出现：

- **主要簇**：大量的回复集中在可视化右侧的一个大而密集的簇中。这表明两种策略在回答基于推理的问题时采用了相似的语言模式，其中模式 0 是主要结构。

- 模型交错：在主要簇中，蓝点（Base 策略）和橙点（SWiFt 策略）完全交错，这表明在遵循模式 0 的常用词结构时，两种策略都使用了类似的不常见词汇。
- 更高的 SWiFt 多样性：可视化图形显示在主簇外存在明显的橙色点（SWiFt）优势，这表明 SWiFt 策略在推理任务中比 Base 策略产生更为多样化的响应。
- 独特的模式：模式 2（三角形）在可视化中频繁出现，表明两种政策都在某些类型的推理问题中使用的次要响应结构。
- 离群值分布：在整个空间中出现了几个孤立点和小群集，主要来自 SWiFt 策略，表明偶尔会出现某些独特的响应形式，这些形式明显偏离了标准模式。

该分布表明，对于 ARC 数据集中的推理导向问题，两个策略共享一种主要的响应结构，但 SWiFt 策略在其表述中表现出更大的灵活性和多样性。

## E.2.2 SciQA 数据集可视化分析

Figure 1(b) 展示了 SciQA 数据集的 UMAP 可视化，该数据集侧重于需要事实知识的科学考试问题。该可视化显示了与 ARC 数据集明显不同的模式：

- 紧密中心簇：一个高度集中的簇出现在右下象限，包含来自两个政策的响应，但在其核心处有明显更高密度的 Base 政策（蓝色）点。
- 同心组织：SWiFt 政策的响应（橙色）似乎在密集的 Base 政策核心周围形成了一个较松散的环，暗示着虽然遵循相似的模式，SWiFt 政策在不常见词汇的使用上引入了更多的变化。
- 稀疏分布：与 ARC 可视化不同，SciQA 的响应显示出主要簇与异常点之间更大的分离，中间点较少，表明响应类别更加明确。
- 模式分布：模式 1 和 2（正方形和三角形）主要出现在边缘，表明主要由 SWiFt 策略用于特定类型的科学问题的替代响应结构。
- 垂直轴分离：与 ARC 可视化相比，点在垂直轴上的分散度更大，这可能表明一个更强的、特定于事实科学问题的次要响应维度。

SciQA 可视化显示，对于事实性科学问题，Base 策略始终遵循一个非常标准化的应答模式，而 SWiFt 策略则表现出更大的变异性，表明其可能采用更多样化的解释策略。

## E.3 跨数据集比较

对比 Figures 1(a) and 1(b) 揭示了政策在处理不同类型问题时的几个重要区别：

1. 簇密度：与 SciQA 可视化中更紧密、更偏向极化的聚类相比，ARC 可视化显示了更为分散的点分布，这表明推理问题的回答比事实性问题有更大的多样性。
2. 模型分离：虽然两个可视化都显示了一些策略的混杂，但 SciQA 数据集在主簇内的策略之间显示出更明显的分离，Base 策略形成了一个更密集的核心。
3. 模式使用：模式 0（圆形）在两种可视化中占主导地位，但次要模式在 ARC 数据集中出现得更为均匀，表明推理问题比事实性问题引发了更广泛的响应结构种类。
4. 异常行为：虽然两个可视化都显示了异常点，但它们的分布不同：ARC 异常值倾向于形成小簇，而 SciQA 异常值则显得更加孤立，这可能表明不同任务类型下异常响应的机制存在差异。

这些跨数据集的比较表明，政策的响应策略不仅在不同政策之间有所不同，而且在不同问题类型中也存在系统差异，其中推理问题比事实性问题引发更多样化的响应。

## E.4 意义

这些可视化揭示了有关政策的几个重要见解：

1. SWiFt 政策在两个数据集上都表现出更大的语言多样性，使用了更广泛的常见和不常见的词汇模式。



2. 这两种策略共享基本的响应结构（特别是模式 0），这表明尽管训练方法不同，它们依赖于相似的基础模式。
3. Base 政策在生成响应时显得更为保守，特别是在 SciQA 数据集中，对于事实性问题，其响应紧密地聚集在已建立的模式周围。
4. 明显的模式聚类表明，这些语言政策倾向于发展出不同的"响应模板"，而不是为每一个输入生成完全独特的响应。
5. 在 ARC 数据集中观察到的更显著的多样性表明，推理问题可能比事实性问题允许或需要更大的回答形式多样性。
6. 这里使用的降维方法相较于传统的评估指标具有优势，因为它揭示了政策输出中可能被隐藏的结构模式。