arXiv:2304.01904v2 [cs.CL] 4 Feb 2024

REFINER：关于中间表示的推理反馈

Debjit Paul♠, Mete Ismayilzada♠, Maxime Peyrard♢*, Beatriz Borges♠, Antoine Bosselut♠, Robert West♠,Boi Faltings♠♠EPFL♢格勒诺布尔-阿尔卑斯大学、法国国家科学研究中心、格勒诺布尔国立综合理工学院、格勒诺布尔信息学实验室{firstname.lastname}@epfl.ch

摘要

近年来，语言模型（LM）通过显式生成中间推断（例如，思维链提示），在推理任务上展现出卓越的性能。然而，这些中间推断步骤可能并不适用于初始上下文，进而导致最终预测错误。本文提出 REFINER，这是一种用于微调语言模型的框架，使其能够在与评判器模型交互的同时，显式生成中间推理步骤；评判器模型会针对推理提供自动反馈。具体而言，评判器会提供结构化反馈，推理语言模型据此迭代改进其中间论证。在三项不同推理任务上对 REFINER 的实证评估表明，与规模相当的基线语言模型相比，它取得了显著提升。此外，使用 GPT-3.5 或 ChatGPT 作为推理模型时，训练好的评判器无需微调推理模型即可显著提升推理能力。最后，我们的评判器模型无需依赖昂贵的人工参与数据进行训练，但在推理时也可以由人类替代。

1 引言

大语言模型（LLM）在自然语言处理（NLP）任务中取得了显著进展（Brown 等，2020）。近期研究表明，在推理任务中明确生成中间步骤，可以显著提升模型的性能和可解释性（Shwartz 等，2020；Paul 和 Frank，2021年；Marasovic 等，2022；Lampinen 等，2022；Wei 等，2022）。生成此类中间表示，能够让我们深入了解模型的预测结果，并使人们能够检查模型的推理过程。然而，这些中间表示[1]可能并不可靠（Ye和Durrett，2022），并导致下游推理任务的表现不佳。最重要的是，目前尚不清楚如何有意义地改进中间表示，从而进一步提升最终表现。

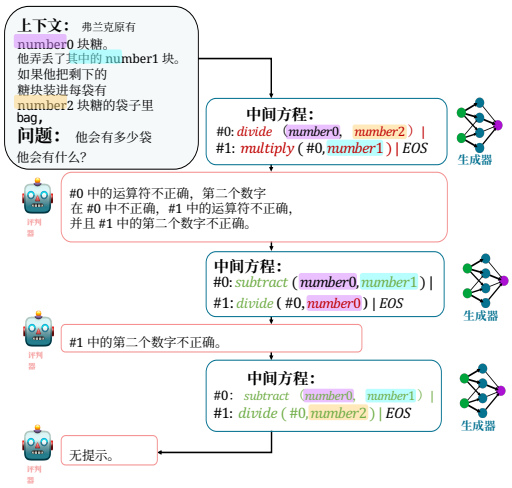


图 1：**REFINER 示例**。评判器模型在评估生成的中间步骤后，向生成器模型反馈其推理错误。随后，生成器模型会收到这些反馈，以及原始问题和先前的中间方程。

纠正推理错误的常见做法是标注新数据，然后重新训练或微调模型（Feng 等人，2021年；Hed- derich 等人，2021年）。然而，通过更多数据微调来修正这类错误，不仅耗费大量数据和资源，而且可能仍不足以使模型在复杂推理任务中实现良好的泛化（Ward 等人，2022）。其他研究则探索了利用反馈改进模型的方法，例如提供标量奖励（Ziegler 等人，2019；Martin 等人，2022），或直接揭示缺失部分的正确答案（Mehta 和 Goldwasser，2019；Elgohary 等人，2021年；Tandon 等人，2022）。然而，在自然语言推理任务中，如何定义能够捕捉不同细粒度推理错误类型（例如，语义一致性、逻辑错误等）的奖励，仍然是一项尚未解决的挑战（Golovneva 等人，2023年）。此外，这样的

[1]在推理任务中，中间表示可以视为推理规则、解释或推理步骤。* 工作在 EPFL 完成

奖励提供的训练信号相对稀疏。

在这项工作中，我们转而针对推理错误提供细粒度的结构化反馈。我们提出了 REFINER，这是一种新颖的交互式框架，使生成器语言模型能够迭代地利用细粒度反馈并改进其推理。交互发生在两个模型之间：生成器通过先生成中间推理步骤来学习解决任务；评判器则向生成器提供关于中间步骤中错误的结构化反馈。

为了针对推理错误提供细粒度反馈，我们设计了一种方案，使用自动构造的反馈数据对评判器模型进行独立训练。具体而言，我们创建了错误中间表示与细粒度推理错误的结构化[2]反馈配对。随后，我们使用这些数据训练评判器，使其能够针对错误的中间推理步骤提供细粒度反馈。最后，评判器与生成器语言模型交互，在生成器训练期间和推理期间提供反馈。

图 1 展示了 REFINER 框架的一个示例：给定一道数学应用题，生成器会生成一个方程作为中间表示。评判器会识别方程中的错误，并向生成器提供半结构化文本反馈（例如，“# 0 中的运算符 不正确”）。通过与评判器交互，REFINER 使生成器能够基于半结构化反馈进行推理，并改进其生成结果。

贡献。 (i) 我们提出了 REFINER，这是一种通过反馈来提升语言模型推理能力的框架。我们的工作研究了与针对中间推理步骤的细粒度推理反馈进行交互，如何影响语言模型在推理任务上的表现。我们在三项自然语言推理任务上评估 REFINER：数学应用题、合成自然语言推理和道德行为生成。REFINER 在不同规模的不同语言模型架构上均展现出显著的性能提升。在不同推理任务中，REFINER 的表现优于规模相近的强大微调语言模型基线（分别高出 +13.1、 +3.2 和 +15 个百分点）。 (ii) 我们通过实证表明，对于数学应用题和合成自然语言推理，我们训练的评判模型本身就有助于改进中间表示，因为它们能帮助 GPT-3。5 在少样本设置下显著提升性能（分别提升 +3.5 和 +6.8 个百分点）。我们还表明，在道德行为生成和数学应用题任务中，针对细粒度错误提供结构化反馈，比提供标量值反馈更有益。我们的评判模型充当了大语言模型的“推理改进工具”。 (iii) 我们表明，REFINER 的表现可大幅优于其他利用大型语言模型反馈的改进方法，例如 self-refine。(iv) 我们的分析表明：(a) 改进中间表示的生成能够提升推理任务的表现；(b) 使用不完美（含噪）的评判器训练生成器仍然有益。我们的代码已公开发布[3]。

2 相关工作

中间表示。 尽管最先进的语言模型在广泛的任务中取得了令人瞩目的表现，但它们在许多推理任务上仍有困难（Wang 等人，2022），尤其是那些包含多个约束或子问题，或需要专业知识的任务（Austin 等人，2021年），例如数学问题求解（Ling 等人，2017；Andor 等人，2019；Ran 等人，2019；Geva 等人，2020；Pi.ekos 等人，2021；Cobbe 等人，2021a；Kim 等人，2022）。

对于这些任务，已有研究表明，中间表征和推理依据有助于学习数学技能（Pi.ekos 等人，2021年）、中间程序执行计算（Nye 等人，2021年）或一般推理输出（Wei 等人，2022；Golovneva 等人，2022）。

我们的工作基于这样一个观察：生成中间步骤很有价值，但在几个关键方面有所不同。首先，我们没有通过提示词引导大型模型，而是微调较小的模型，使其学会生成中间步骤。其次，我们的框架能够适用于不一定存在唯一闭式正确答案的任务，例如道德规范任务（见 §3 ）。最后，我们的框架通过评判器提供反馈进行训练，从而改进模型的推理过程，并教会模型如何利用反馈。

自然语言反馈。 近期研究探索了通过自然语言为模型提供更丰富、更复杂的反馈（Ziegler 等人，2019；Nguyen 等人，2021年；Scheurer

[2]请注意，我们使用模板将结构化反馈转换为半结构化文本反馈。

[3]https://github.com/debjitpaul/refiner

等人，2022），用于使大语言模型的输出与用户偏好保持一致（Christiano 等人，2017；Ziegler 等人，2019；Saunders 等人，2022；Scheurer 等人，2022；Bai 等人，2022），或直接提升模型在当前任务中的表现（Weston，2016；Rupprecht 等人，2018；Elgohary 等人，2020；Austin 等人，2021；Madaan 等人，2023）。这种训练依赖大量人工创建的反馈（Bai 等人，2022），需要耗费大量资源。尽管外部反馈提供者能够引导模型给出正确答案和推理过程（Austin 等人，2021），且效果显然优于模型自身（Saunders 等人，2022），但反馈很少以这种方式使用，而且事实证明，为推理任务构建自动化评判器颇具难度（Scheurer 等人，2022；Wang 等人，2022；Huang 等人，2022）。

最近，Welleck 等人（2022）提出了一种辅助模型——校正器。它通过学习生成器所犯的错误类型及其修正方法，改进生成模型的初始命题。在这项工作中，我们也使用了一个辅助模型，即评判器，但应用方式截然不同：训练期间，我们将它整合进与生成器模型的交互循环中。我们与以往工作的另一处不同是，我们在模型的中间推理步骤而非最终输出处提供反馈。因此，反馈更接近错误来源，并引导模型的推理朝着正确答案前进。此外，中间步骤通常具有结构化形式，使评判器能够提供精准反馈。

3 REFINER

问题表述。在本文中，我们将自然语言推理（NLR）视为一种自回归生成任务：给定输入上下文x，模型需要生成y，使y满足任务的约束。通常，为了生成正确或合理的y，模型需要执行正确的推理z作为中间步骤。[4] 我们将自然语言推理任务分解如下：$p(y|x) = p(y|x,z)p(z|x)$。实践中，可以使用将条件变量作为输入一部分的语言模型来计算每个条件概率。

在继续介绍模型之前，我们先介绍本研究所涉及的三个自然语言推理任务及其各自的中间表示z。我们特意选择这三个任务因为它们广泛涵盖了两类推理：(i)逻辑推理和(ii)规范推理。相关示例见附录图6，下文将详细介绍。

数学应用题（MWP）：给定一道由上下文和问题组成的文字题x，目标是将x映射为有效的数学表达式z（中间表示），然后得到解答y。此任务要求模型运用数学推理进行演绎。**合成自然语言推理（sNLR）**：给定一个由 5 条合成规则和一个事实组成的推理情境x，模型需要推导出结论y。此任务要求模型运用演绎推理，并使用封闭世界规则和事实生成中间步骤z及结论y。

道德故事（MS）的道德规范与动作生成：给定由情境、意图和不道德行为组成的上下文x，模型需要生成道德规范z和道德行为y。道德规范鼓励道德行为。此任务要求模型通过溯因推理生成道德规范，并通过演绎推理生成道德行为。

我们提出通过促使模型生成中间假设（z），并利用结构化反馈对其进行改进，来解决这些任务。我们引入了一个交互式框架 REFINER，由两个独立的模型组成：(a) 一个经过训练、能够对中间推理步骤提供结构化反馈的 CRITIC 模型（§3.1）；以及 (b) 一个经过训练、通过先生成中间推理步骤来完成推理任务的 GENERATOR 模型（§3.2）。REFINER 的核心思想是利用生成器模型与评判器模型之间的交互，通过评判器提供的结构化反馈改进生成器的中间推理步骤。

REFINER 具有几项重要特性。首先，生成器经过训练，能够吸收并利用反馈，这有助于它在训练过程中逐步获得更好的推理能力，也使其能够在测试时整合来自已训练评判器或人类的反馈（见 §5）。其次，经过训练的评判器本身也很有用；我们证明，在所考虑的推理任务中，像 GPT-3。5 这样的通用大语言模型，与我们训练的评判器交互后能够显著获益（见 §5）。最后，采用两个独立模型使我们能够轻松衡量训练期间和／或推理期间反馈带来的益处（见 §6）。

[4]我们将“推断步骤/表示”与“假设”交替使用。

任务错误类型		反馈
数学文字题	错误数字	其中的位置数字
	错误运算符	公式编号不正确。 运算符 in 公式编号不正确。
	缺失运算符	缺少一个运算符。
sNLR	逻辑无效	TheX 运算符进行推理 规则数字无效。
	缺失关联	事实与规则之间缺少关联。
	缺失隐式知识	隐式知识是
	知识步骤	缺失。
MS	矛盾	矛盾
	语义错位	语义错位：“ text snippet”

表 1：错误类型与反馈概览
针对每项推理任务。

3.1 CRITIC 模型

评判器的作用是对生成器模型产生的中间假设提供反馈。一种评估假设质量并针对该假设 z^* 提供反馈的方法，是将其与标准假设 z^*进行比较。以往研究将 BLEU、ROUGE 等自动评估指标用作价值函数（Wu et al.，2018；Ramamurthyet al.，2022）。然而，这些标量价值函数并不适用于自然语言推理任务，因为：（i）难以定义能够涵盖细粒度推理错误的标量价值函数（Golovnevaet al.，2023年）；（ii）在推理时，这些函数需要获取标准假设（而实际中无法获得）。因此，我们训练了一个评判器模型，使其能够以细粒度的方式评估假设并提供结构化反馈。

反馈数据生成。 为了训练评判器，我们需要创建一些示例对，其中包含不合理的假设及其对应的反馈，并体现细粒度推理错误。受 Golovneva 等人 (2023年和 Talmor 等人 (2020) 的启发，我们首先针对每种推理任务定义细粒度推理错误类型（见表 1）。对于 MWP，方程出错可能是因为：(i) 方程中的操作数或运算符有误，和/或 (ii) 缺少一个或多个运算符。对于 sNLR，推理规则出错可能是因为：(i) 逻辑上无效，和/或 (ii) 缺少推理规则（未能用正确规则关联正确事实，或缺少隐式知识）。对于 MS，道德规范出错可能是因为 (i) 存在矛盾，和/或 (ii) 存在语义错位。

基于这些错误类型，我们提出两种生成反馈数据的策略：(i) **基于规则的扰动** 策略：我们对合理的假设进行扰动（

z）在训练数据中，并收集

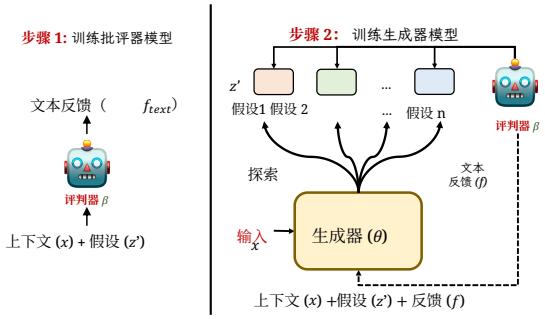


图 2：REFINER 交互循环概览。左侧：训练评判器模型。右侧：在每次迭代中，生成器会生成多个假设。评判器会随机选择一个假设，并根据推理错误提供反馈。

一个数据池D (x：输入，z：合理假设，z'：不可信假设）。我们通过省略、替换或添加合理假设中的一些词元或规则，自动生成不可信假设（详见附录F.1）。（ii）**合成生成**策略：我们提示 OpenAI 的 GPT-3。5 根据错误类型自动生成不可信假设。我们采用少样本设置，并改变指令、示例数量以及示例的格式（详见附录F.2）。

由于我们的扰动和自动生成的不可信假设基于逻辑与推理错误，因此我们为每个示例构建结构化反馈 f （x，z，z'），指出其中z' 存在、但z 中不存在的错误类型（参见表 1）。这些任务的反馈 f 的基本结构为 ⟨错误类型、位置（可选）、提示（可选）⟩，其中位置表示不可信假设中的错误位置（参见表1）。尽管我们用于这些任务的策略很简单，但这种方法很容易推广到其他推理任务。

我们还用随机判断替代正确判断，以增加每个示例中不合理假设的数量。最后，作为反馈f，我们提供 <错误类型、提示>。对于范数和动作生成等非单调推理任务，评判器应能提供提示，使生成器模型的目标与推理任务保持一致。因此，作为提示，我们提供规范中的动词短语。由于评判器向生成器提供文本反馈，我们将结构化反馈转换为自然语言反馈 [5]。形式上，我们创建数据池$D = \{x,\ z,\ z',\ f\}$，用于训练评判器模型。

[5]关于反馈的更多细节见附录F。

训练评判器模型。我们使用上下文（x）和（合理或不合理）假设（z 或z'）作为输入，以文本反馈作为输出，训练一个监督式 CRITIC 模型（π_β）。我们使用交叉熵损失更新 CRITIC：$L(\beta) = -\log p_\beta(f(u)|x, u)$，其中$u \in z, z'$。训练后的评判器仅用于推理。训练生成器时使用预言机评论器。

3.2 生成器模型

本节介绍一种生成器模型，该模型通过迭代学习与 CRITIC 模型交互。

预热阶段。给定上下文 x，生成器模型 (π_θ) 经过训练后能够生成合理的假设。预热阶段至关重要，可确保评判器加入循环后，生成器不会因输出空间过大而生成可能质量低劣的随机答案。因此，我们使用一个小型监督数据集（10% 的训练数据）针对感兴趣的自然语言推理任务对模型进行微调。预热阶段结束后，我们利用来自评判器模型的额外反馈 f 进行学习$\pi_\theta(z|x,z',f)$。

探索。在每次迭代（t）中，生成器模型使用 nucleus 采样生成多个假设（z^k）。评判器模型随机选择一个假设，并针对该假设提供反馈。探索步骤旨在增大输出方差，使生成器在训练过程中接收到多种多样的反馈。

学习。我们使用以下交叉熵损失来更新生成器模型：$L(\theta) = -\sum_{t=1}^{T} \log p_\theta(z_t|x, z'_t, f_t(z'))$，其中$T$ =表示迭代总数。由于反馈包含错误类型和提示，这些信息是（潜在的）细粒度且具有逻辑性的，因此模型应能通过解决反馈中提及的推理错误来学习并更新其生成结果。

推理。我们使用训练好的评判器和训练好的生成器来生成轨迹$z_0, z1$，……，z_T，并在生成器生成$f(z_t)$或评判器生成“无提示”时停止。我们还尝试了思维链提示法，此时生成器会生成一条轨迹z_0y_0, z_1y1，……，z_Ty_T，并在评判器生成“无提示”时停止。

4 实验设置

数据集。我们在三项不同的任务上评估 REFINER（示例见图 6）。下面简要介绍各项任务所用的数据集。数学文字题

生成器模型	等式（z）	答案（y）
UQA-base	34.1	–
UQA-base + 近端策略优化	31.5	–
REFINER 基础版	**47.2**	–
UQA-large	46.7	–
UQA-large + 近端策略优化	48.2	–
REFINER 大型版	**53.8**	–
GPT-3.5 +思维链	64.1	67.1
GPT-3.5 +思维链 + REFINER 评判器	**67.3**	**70.6**

表2：数学文字题的结果。REFINER 与基线方法在 SVAMP 数据集上的比较。报告三次运行的平均分（p<0.05）。对于 GPT-3.5 以外的模型，答案可通过对方程进行符号执行获得，因此取决于方程的有效性。

数学文字题（MWP）：我们使用 MAWPs 数据集（Koncel-Kedziorski 等人，2016）训练模型，并在具有挑战性的 SVAMP 数据集（Patel 等人，2021年）上评估模型。我们同时评估模型的方程生成（z）和答案预测（y）任务。与 Ling 等人（2017）和 Amini 等人（2019）在方程生成任务中的做法类似，我们将数值替换为变量名，例如 number0、number1 等。此外，我们还在 GSM8K 数据集（Cobbe 等人，2021b）上进行了评估，该数据集包含 8.5K 道高质量、语言表达多样的小学数学应用题。对于合成自然语言推理（sNLR），我们使用 Liang 等人（2022）的数据集，并选取难度为 hard 的部分。我们同时评估模型的推理规则生成（z）和后件生成（y）。对于道德故事（MS），我们使用 Emelin 等人（2021）的数据集，评估模型生成道德规范 z 和道德行为 y 的能力。

训练详情。对于每项任务，我们训练一个 UnifiedQa-T5-base 模型（UQA-base）（Khashabi 等人，2020）作为评判器（§3.1）。在探索阶段（§3.2），我们使用核采样，参数为 p = 0.5。我们根据验证损失选择超参数：生成器和评判器模型均使用 Adam 优化器，学习率为 $1e^{-4}$。每个模型训练 20 个轮次，并根据验证损失提前停止。所有模型均在一块 A100 GPU 上训练。我们使用 3 个随机种子运行模型，并报告平均结果。对于人工研究，我们根据自动评估指标，从表现最佳的模型（基线方法和我们的模型）中选取输出。我们使用 T = 3 次迭代训练模型。

在推理时，我们对生成器和评判器模型使用贪心解码，对自动评判器使用 T = 1，对预言机评论器使用 T = 3 。

在 MWP 和 sNLR 任务中，我们对中间步骤（方程生成和推理规则）使用完全匹配（EM）指标，对最终答案使用准确率（Acc）指标。对于 ForMS，我们开展了人工评估研究，以评估规范和道德行为的相关性[6]。更多评估细节见附录G。为了训练评判器模型，我们使用了通过基于规则的扰动策略生成的反馈数据（参见 §3.1）。

基线方法。 我们将所提方法与三种不同的语言模型生成器进行比较：UQA-base、UQA-large（监督设置）、GPT-3.5-text-DaVinci-003 和 ChatGPT（少样本设置）。我们还将 REFINER 与基于近端策略优化（PPO）的强化学习方法进行比较（Schul-manetal.，2017）。我们使用了（Ramamurthy et al.，2022）中的 PPO 实现。对于 GPT-3.5，我们为每个类别提供 2 个示例。我们还试验了思维链（COT）提示（Wei et al.，2022），即先提示模型生成中间步骤（z），再生成最终答案（y）。请注意，sNLR 任务是一项合成任务，模型需要进行单跳或双跳推理。Clarket al.（2021）表明，对大型语言模型（参数规模为 354M）进行微调可取得较高性能（准确率达 99%）。因此，我们仅将 REFINER 模型与 UQA-base 模型（220M）进行比较（见表3）。由于人工标注成本高昂，我们重点与最有意义的基线方法进行比较：MS 任务中的 UQA-large（见表4）。需要强调的是，我们提出的框架具有通用性，也可以使用任何其他语言模型作为生成器或评判器。

5 结果

我们从两个方面评估模型的表现：(i) 中间步骤的表现；(ii) 最终答案预测的表现。表 2、3 和4展示了性能对比。

中间步骤的表现。 表 2 报告了数学文字题任务的表现。我们探索了两种不同的情形：（i）模型只生成方程（z），并用变量名替代数值；（ii）模型同时生成方程和最终答案。我们观察到，在这两种情形下，REFINER 的

[6]由于 BLUE、ROUGE 等自动评分指标仅衡量标准规范或行动与生成的规范或行动之间的词汇相似度。

生成器模型	中间推理 (z)	最终答案 (y)
UQA-base	90.6 ± 0.8	94.1
REFINER 基准	**93.5** ± **0.4**	**97.3**
GPT-3.5 +思维链	14.3 ± 0.9	40.6
GPT-3.5 + 思维链 + REFINER	**21.1** ± **1.2**	**42.1**

表3：sNLR任务的结果。报告三次运行的平均分数（p <0.05）。中间推理：推理规则（完全匹配），最终答案：后件（准确率）

	范数（z）				动作（y）			
模型	I↓	U↓	R↑	α	I↓	U↓	R↑	α
B	34	17	49	0.35	28	14	58	0.64
B+近端策略优化	38	10	52	0.38	31	17	52	0.38
REFINER	19	12	**69**	0.33	18	9	**73**	0.55

表4：道德规范和道德行为的结果。我们报告人工评估。B：UQA-large；I：不相关，U：不确定；R：相关；α：Krippendorff α系数

。值得注意的是，在推理循环中加入评判器后，UQA-base 获益最大（+13.1 完全匹配）。我们观察到，GPT-3.5 从经过 REFINER 训练的评判器中获益显著。由于 GPT-3.5（175B 参数）等大语言模型的微调成本高昂，因此无需任何修改便能提升方程生成的 +3.2完全匹配率，这一点十分重要。有趣的是，我们观察到 GPT-3.5 + COT 在答案 y 上的准确率明显高于方程 z（见表2）。这一结果与Ye和Durrett（2022）的观察相似，表明中间方程可能不可靠。最后，REFINER 的表现甚至可能超过 PPO，后者使用 BLEU分数作为奖励函数。这表明，与基于数值（数值来自自动评估指标）的奖励反馈相比，半结构化、细粒度的文本反馈更有益。请注意，若直接根据复杂的人类价值观优化这些模型，结果可能有所不同，正如Stiennon等人（2020）所示。从定性角度看，REFINER 可以通过结构化反馈纠正错误方程，修正多步解答中的运算符（见图 7）。

对于 sNLR，与 Liang 等人 (2022) 类似，我们观察到 GPT-3.5 的表现不佳（见表 3）。REFINER 提升了 +2.9，并且 +6.8 相较于 UQA-base 和 GPT-3 的完全匹配分数。5，分别如此。与数学文字题不同，最终答案 y 并非通过对中间步骤进行符号执行z即可得到，但我们仍观察到 REFINER 专注于改进中间步骤 z，从而显著提升了答案 y 的预测效果。我们再次观察到，配合 UQA-base 的 REFINER 表现优于少样本提示的 GPT-3。5。

生成器模型	SVAMP		GSM8K	
	GPT-3.5	ChatGPT	GPT-3.5	ChatGPT
CoT	67.1	68.2	63.5	74.1
Self-reflection	67.2	68.4	63.1	74.6
Self-refine	67.6	68.2	63.8	74.7
REFINER	**70.6**	**71.4**	**66.2**	**75.9**
ReACT	67.3	68.4	64.7	75.5
ReACT + REFINER	**70.6**	**71.9**	**67.8**	**77.4**
自一致性	69.5	70.4	65.5	76.1
自一致性 +REFINER	**72.1**	**72.5**	**67.2**	**78.1**

表 5：**在 SVAMP 和 GSM8K 数据集上与不同改进方法的比较**。报告测试集上三次运行的平均准确率（p<0.05）。

模型	式（z）
REFINER $_{base}$ + 评判器数据$_{rule-based}$	47.2
REFINER $_{base}$ - 评判器$_{inference}$	39.8
REFINER $_{base}$ - 评判器$_{inference}$ - 实验	37.4
REFINER $_{base}$ - 评判器$_{training}$	34.1
REFINER $_{base}$ + 评判器数据$_{synthetic}$	44.1
REFINER $_{base}$ + 评判器$_{Oracle}$	66.0

表 6：**消融结果** 在数学文字题任务上的表现；比较推理时不使用评判器的模型，以及训练时不使用探索（exp）阶段的模型。我们报告生成的方程的精确匹配分数，可与表 2。

因此，我们的评判器能够识别细粒度的推理错误，并帮助提升推理规则生成的性能。

对于道德故事，我们使用三位人工评审员评估生成质量，由他们判断生成的规范和道德行为是否与给定的道德故事相关。表4 汇总了对从 MS 测试数据集中随机抽取的 100 个道德故事样例进行人工评估的结果。具体而言，我们分别报告规范和道德行为的评估明细，按被判定为 不相关、不确定或相关 的实例数量统计，并给出 Krippendorf's α (Krippendorff，2018) 一致性分数。结果显示，相较于强大的 UQA-large 基线方法，相关性提高了 20 分。因此，这表明参数量仅为生成器三分之一的专用批评器模型可以提升推理步骤生成的效果。

最终答案预测的性能。 我们观察到，对于数学文字题、sNLR 和道德故事，REFINER 分别比强大的语言模型基线高出 +3.5，+3.2，+15分。这些结果支持了我们的假设：生成更好的中间步骤可以带来更好的答案预测。值得注意的是，在 sNLR 任务中，对于 GPT-3.5，我们观察到，加入评判器后，推理步骤生成提升了+6.8；然而，后件预测仅提升了+1.5 。这一结果表明，大语言模型可能不会利用这些中间步骤进行演绎，或者无法完成演绎。

将 REFINER 与其他改进方法进行比较。在表5中，我们比较了 REFINER 与另外两种近期改进方法在 SVAMP 和 GSM8K 数据集上的表现：Self-refine（Madaan 等人，2023年）和 Self-reflection（Shinn 等人，2023年）方法。这两种基线方法都使用大语言模型生成自动反馈。与Madaan 等人（2023年）的研究类似，我们观察到，Self-refine 在数学应用题任务上的提升较小。相反，我们发现，REFINER 在 SVAMP 和 GSM8K 数据集上分别使 GPT-3.5 和 ChatGPT 的表现显著提升了 +3.3 和 +2.2。这凸显了训练一个专门的评判器并使其立足于具体任务的益处。与通用模型（GPT-3.5 或 ChatGPT）提供的反馈相比，它能让大语言模型更加准确。附录 §6中提供了更多关于使用我们训练的评判器和 GPT-3.5 生成的反馈质量的细节（见表8）。此外，我们还评估了 REFINER 改进两种近期方法生成的思维链的表现：Self-Consistency（Wang 等人，2023年）和 ReACT 方法（Yao 等人，2023年）。我们观察到，REFINER 能使 Self-Consistency 和 ReACT 分别提升 +2.02 和 +2.9。这表明，训练好的评判器可以作为一种工具，直接用于不同方法并带来性能提升（更多详情见附录 §A.2）。

消融。 为了更深入地了解模型各组成部分的贡献，我们进行了消融研究（表 6）。我们观察到，在推理过程中不使用评判器模型时，性能会显著下降，从 47。2 降至 39。8。因此，这一结果表明，我们的生成器模型能够在推理时利用评判器提供的反馈。此外，我们发现，探索步骤能够提升性能 +3.3，优于基线模型。这一结果支持了我们的假设：探索步骤能够增加输出方差，使生成器模型有机会在各种反馈上进行学习。我们比较了基于两种不同训练数据训练的评判器模型的性能（见 §3.1）。我们发现，使用 GPT-3.5 基于少量自动生成的数据训练的评判器，效果优于不使用评判器的情况

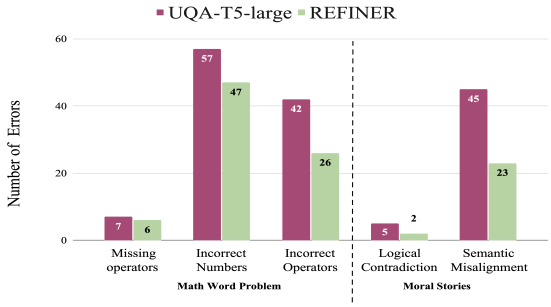


图 3：**错误分析**。 基线 UQA-large 和 REFINER 在从两个数据集的测试集中随机抽取的 100 个实例上产生的错误数量。错误按照表1）进行分类。

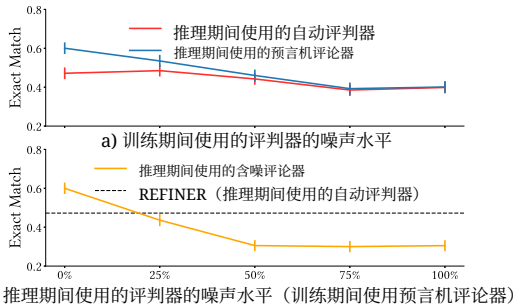


图 4：**带噪评论器分析**。在图 (a) 中，我们改变训练阶段所用评判器的噪声水平（噪声为 0 时对应预言机），并比较推理时使用预言机和训练阶段自动评论器的模型。在图 (b) 中，我们使用预言机评判器进行训练，同时改变推理时所用评判器的噪声水平。

这比在循环中不使用评判器效果更好。这一结果促使研究者使用该方法生成负样本，用于训练评判器或偏好学习模型。最后，我们还观察到，如果评判器是完美的（Oracle），那么 REFINER 就能通过纠正生成器模型产生的错误，显著提升性能。这一结果表明，REFINER 可以被视为一种支持人工智能之间以及人类与人工智能交互的框架。

6 分析

错误分析。 为更深入地了解我们方法的表现，我们对 MWP 和 MS 数据集进行了细粒度错误分析（图 3）。我们注意到，最常见的错误分别是 MWP 中的错误数字 和 MS 中的 语义错位 。一个直观的原因可能是，对于 MWP 任务，正如（Patel 等人，2021年）所指出的，模型对数字顺序很敏感。对于 MS，生成植根于上下文的规范颇具挑战性。我们的分析显示，REFINER 能够显著减少两个数据集上的错误，这一趋势十分明显。这表明，我们训练的评判器模型能够在推理过程中识别细粒度推理错误。

噪声敏感性。 为进一步了解 REFINER 框架的行为，我们针对 MWP 任务开展了使用带噪评论器的变体实验。我们将训练过程中使用的预言机评论器替换为带噪评论器（图 4 (a)），以考察使用不完美评论器进行训练会如何影响生成器。我们还在推理时使用带噪评论器，同时在训练期间保留预言机评论器（图 4 (b)）。带噪评论器通过对预言机评论器施加随机扰动生成；对于噪声水平 ϵ，预言机反馈以概率 ϵ 被替换为随机反馈。

图 4 (a) 表明，当使用噪声很大的评论器（$>75\%$ 噪声）进行训练时，生成器语言模型会学会忽略评论器，因为推理时使用训练得到的评论器与使用预言机并无区别。有趣的是，加入少量噪声（$<50\%$）似乎不会损害模型：其性能与使用预言机（噪声为 0%）进行训练相比，在统计上没有显著差异。图 4 (b) 显示，推理时使用的评论器质量影响巨大。由预言机提供反馈显然是最佳情形。噪声达到 25% 时，评论器导致生成器的表现已经不如使用我们训练的评论器（REFINER）。噪声超过 50% 时，评论器会显著损害生成器。使用预言机评论器训练的生成器已经学会信任评论器，并期待获得有用的反馈。

定性分析。 为了解释 §6中的发现，我们进一步对数学文字题任务的 100 个实例进行了人工分析。我们观察到，当 REFINER 未能修正 生成器 模型生成的输出时，存在两种不同的情况：(a) 评判器 模型给出了 正确的 反馈；然而，生成器 模型仍然生成了错误的 方程；(b) 评判器 模型给出的反馈 不完整 或 部分正确 。前一种情况表明，生成器模型可能未能遵循 评判器 给出的指令，或者评判器给出的反馈可能含糊不清。例如，在附录图5的 (b) 中，我们观察到评判器给出的判断正确，但其反馈可能导致方程错误的情况。后一种情况表明，经过训练的评判器模型生成了错误的反馈，从而可能导致方程错误或部分正确。我们还

任务	UQA（220M）	UQA（770M）	GPT-3（175B）
数学文字题	69.5 +/- 2.6	73.4 +/- 3.7	63.5 +/- 5.6
sNLR	95.5 +/- 1.4	98 +/- 2.2	34.5 +/- 2.4
MN	77.4 +/-2.5	80 +/- 4.5	76.4 +/-3.5

表 7：**不同评判器模型的性能比较**。报告完全匹配分数。

观察到，当生成器模型生成的方程中包含多个错误时，我们的 CRITIC 模型未能生成正确的反馈。

反馈质量。为了更好地理解反馈质量的差异，我们将训练后的评判器模型与 GPT-3.5 进行比较。我们针对每项任务评估 500 个实例的反馈质量，并在表8中报告完全匹配分数。请注意，我们纳入了评判器反馈应指出解答正确、因此生成“否”的实例。对于 GPT-3.5，我们为每种错误类型提供了两个少样本示例，并提供了两个反馈为“否”的示例。结果表明，经过训练的评判器（UQA）能够全面胜过 GPT-3.5。我们观察到，GPT-3.5 在识别答案正确的情况时表现良好。然而，当要求它针对错误的推理步骤生成有意义的半结构化反馈时，它会出错。

7 结论

本文提出了 REFINER，这是一种通过两个模型（生成器 和 评判器）之间的迭代反馈循环来提升语言模型推理能力的框架。我们在三项推理任务上对该框架进行了评估，结果表明，针对中间推理错误提供结构化、细粒度的反馈能够显著提升性能，优于标量值反馈。即使存在噪声，我们训练的评判器模型本身也能改善语言模型的中间推理表示，这表明 REFINER 能显著提升语言模型在推理任务上的表现。原则上，我们的 REFINER 框架具有很强的通用性，可用于引导语言模型完成不同的推理任务。更具体地说，评判器 模型可以视为一种工具，帮助大语言模型提升生成质量。

致谢

我们感谢 Martin Josifoski、Syrielle Montariol 和 Zeming Chen 对论文草稿提出的宝贵反馈。我们也感谢 ICT-48 人工智能研究网络提供的支持。

搜索卓越中心“TAILOR”（欧盟地平线2020计划，GA编号952215）。West的实验室部分经费由瑞士国家科学基金会（200021_185043）、瑞士数据科学中心（P22_08）、H2020（952215）、Microsoft瑞士联合研究中心和Google的拨款，以及Facebook、Google和Microsoft的慷慨捐赠提供。Antoine Bosselut谨此感谢 Innosuisse（项目编号PFFS-21-29）、EPFL科学种子基金、EPFL成像中心、索尼集团公司和艾伦人工智能研究所提供的支持。

局限性

由于适用的下游推理任务数量庞大，我们无法对 REFINER 框架进行全面评估。尽管这些任务有意选取为彼此不同的类型，我们仍只聚焦于三种不同的推理任务，以研究自然语言推理反馈如何影响下游任务。我们认为，这代表着探索针对中间表示的自动化自然语言反馈迈出的初步但重要的一步。此外，本文提出的评判器针对每项任务分别设计，而理想的评判器应当具有通用性，能够针对广泛的推理任务提供反馈。同样，我们考察了针对各类推理任务的细粒度推理错误。近期研究提及了其他几种细粒度推理错误（Golovneva等人，2023年），而我们所考察的推理任务无法涵盖所有这些错误。评判器的通用化以及细粒度错误类型的推广，既是本文的主要局限，也构成了未来的研究方向。最后，随着大语言模型越来越多地应用于现实场景（如医疗领域和重要决策），我们认为，开发专家模型和自动反馈机制，以检查模型生成内容并加以改进，至关重要。大语言模型表现出色，在多项自然语言处理任务中效果良好，但它们并非专家系统。我们的工作旨在弥补这一差距，展示加入评判器（专门微调的评判器）提供的干预和反馈，能够帮助大语言模型提高准确率，同时使整个过程更加透明。

伦理考量

在本文中，我们使用了现有数据集；据我们所知，这些数据集均已得到充分引用。我们提出的框架 REFINER 旨在提升语言模型的推理能力。已有研究表明，这些语言模型会编码有关种族、性别及许多其他人口统计属性的偏见（Weidinger 等人，2021年）、（Sheng 等人，2020）。由于我们的框架并未提供缓解这些偏见的方法，通过该框架改进的模型仍可能表现出这些模型通常会呈现的同类有害行为。我们建议，任何直接部署我们模型的人（开箱即用）都应先检查模型是否会对任何受保护群体造成伤害，并采取适当的缓解措施。此外，我们的 MS 任务基于一个由情境、意图和行动构成的数据集，该数据集明显偏向西方文化和社会规范（Emelin 等人，2021年）。因此，我们对 MS 任务进行的人工评估由位于美国的亚马逊土耳其机器人工作人员完成，并根据完成任务所需的平均时间给予了适当报酬。

参考文献

Aida Amini、Saadia Gabriel、Shanchuan Lin、Rik Koncel-Kedziorski、Yejin Choi 和 Hannaneh Ha-jishirzi。2019。MathQA：借助基于运算的形式化方法，实现可解释的数学应用题求解。载于2019年自然语言处理实证方法会议论文集：计算语言学协会北美分会人类语言技术，第1卷（长篇与短篇论文），第2357–2367页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Daniel Andor、Luheng He、Kenton Lee 和 Emily Pitler。2019。给 BERT 配备计算器：通过阅读理解发现运算和参数。载于2019年自然语言处理实证方法会议与第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，第5947–5952页，中国香港。计算语言学协会。

Jacob Austin、Augustus Odena、Maxwell Nye、Maarten Bosma、Henryk Michalewski、David Dohan、Ellen Jiang、Carrie Cai、Michael Terry、Quoc Le 和 Charles Sutton。2021年。 使用大语言模型进行程序合成。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、A manda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon、Carol Chen、Catherine Olsson、Christo- pher Olah、Danny Hernandez、Dawn Drain、Deep Ganguli、Dustin Li、Eli Tran-Johnson、Ethan Perez、Jamie Kerr、Jared Mueller、Jeffrey Ladish、Joshua Landau、Kamal Ndousse、Kamile Lukosuite、Liane Lovitt、Michael Sellitto、Nelson Elhage、Nicholas Schiefer、Noemi Mercado、Nova DasSarma、Robert Lasenby、Robin Larson、Sam Ringer、Scott Johnston、Shauna Kravec、Sheer El Showk、Stanislav Fort、Tamera Lanham、Timothy Telleen-Lawton、Tom Con- erly、Tom Henighan、Tristan Hume、Samuel R. Bow- man、Zac Hatfield-Dodds、Ben Mann、Dario Amodei、Nicholas Joseph、Sam McCandlish、Tom Brown 和 Jared Kaplan。2022。 宪法式人工智能：来自人工智能反馈的无害性。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens

Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020。语言模型是少样本学习者。载于《神经信息处理系统进展》第33卷，第1877–1901页。Curran Associates, Inc.

PaulChristiano、Jan Leike、Tom B.Brown、Miljan Mar- tic、Shane Legg 和 Dario Amodei。2017。基于人类偏好的深度强化学习。

Peter Clark、OyvindTafjord 和 Kyle Richardson。2021年。Transformer 作为语言的软推理器。载于第二十九届国际人工智能联合会议论文集，IJCAI'20。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。2021a。训练验证器以解决数学文字题。

Karl Cobbe、Vineet Kosaraju、Mohammad Bavarian、Mark Chen、Heewoo Jun、Lukasz Kaiser、Matthias Plappert、Jerry Tworek、Jacob Hilton、Reiichiro Nakano、Christopher Hesse 和 John Schulman。2021b。训练验证器以解决数学文字题。arXiv 预印本 arXiv:2110.14168。

Ahmed Elgohary、Saghar Hosseini 和 Ahmed Hassan Awadallah。2020。与解析器对话：通过自然语言反馈进行交互式文本到SQL。载于第58届计算语言学协会年会论文集，第2065–2077页，线上。计算语言学协会。

Ahmed Elgohary、Christopher Meek、Matthew Richardson、Adam Fourney、Gonzalo Ramos 和 Ahmed Hassan Awadallah。2021年。NL-EDIT：通过自然语言交互纠正语义解析错误。载于2021年计算语言学协会北美分会年会：人类语言技术论文集，第5599–5610页，线上。计算语言学协会。

Denis Emelin、Ronan Le Bras、Jena D.Hwang、Maxwell Forbes 和 Yejin Choi。2021年。道德故事：围绕规范、意图、行动及其后果的情境推理。载于2021年自然语言处理经验方法会议论文集，第698–718页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Steven Y. Feng、Varun Gangal、Jason Wei、Sarath Chan- dar、Soroush Vosoughi、Teruko Mitamura 和 Ed- uard Hovy。2021年。自然语言处理数据增强方法综述。载于《计算语言学协会论文集：ACL-IJCNLP 2021》，第968–988页，线上。计算语言学协会。

Mor Geva、Ankit Gupta 和 Jonathan Berant。2020。将数值推理技能注入语言模型。载于《第58届计算语言学协会年会论文集》，第946–958页，线上。计算语言学协会。Olga Golovneva、Moya Chen、Spencer Poff、Martin Corredor、Luke Zettlemoyer、Maryam Fazel-Zarandi 和 Asli Celikyilmaz。2022年。ROSCOE：用于评估逐步推理的一套指标。

Olga Golovneva、Moya Chen、Spencer Poff、Martin Corredor、LukeZettlemoyer、Maryam Fazel-Zarandi 和 Asli Celikyilmaz。2023年。ROSCOE：一套用于评估逐步推理的指标。载于《第十一届国际学习表征会议》。

Michael A. Hedderich、Lukas Lange、Heike Adel、Jan- nik Strötgen 和 Dietrich Klakow。2021年。低资源场景下自然语言处理近期方法综述。载于《2021年北美计算语言学协会会议论文集：人类语言技术》，第2545–2568页，线上。计算语言学协会。

Jiaxin Huang、Shixiang Shane Gu、Le Hou、YuexinWu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。2022。大语言模型能够自我改进。

Zhanming Jie、Jierui Li 和 Wei Lu。2022。学习进行演绎推理：将数学应用题求解视为复杂关系抽取。载于《第60届计算语言学协会年会论文集》（第1卷：长篇论文），第5944–5955页，爱尔兰都柏林。计算语言学协会。

Daniel Khashabi、SewonMin、Tushar Khot、Ashish Sabharwal、Oyvind Tafjord、Peter Clark 和 Hannaneh Hajishirzi。2020。UNIFIEDQA：通过单一问答系统跨越格式边界。载于《计算语言学协会研究成果：2020年自然语言处理经验方法会议》，第1896–1907页，线上。计算语言学协会。

Niklas Kiehne、Hermann Kroll 和 Wolf-Tilo Balke。2022。为偏离社会多数群体的规范提供语境化语言模型。载于《2022年自然语言处理经验方法会议研究成果》，阿拉伯联合酋长国阿布扎比。计算语言学协会，计算语言学协会。

Bugeun Kim、Kyung Seo Ki、Sangkyu Rhim 和 Gah- gene Gweon。2022。EPT-X：一种用于生成数字解释的表达式指针变换器模型。载于《第60届计算语言学协会年会论文集》（第1卷：长篇论文），第4442–4458页，都柏林，爱尔兰。计算语言学协会。

Rik Koncel-Kedziorski、Subhro Roy、Aida Amini、Nate Kushman 和 Hannaneh Hajishirzi。2016 年。MAWPS：

数学文字题语料库。载于 2016 年计算语言学协会北美分会人类语言技术会议论文集，第 1152–1157 页，圣迭戈，加利福尼亚州。计算语言学协会。

Klaus Krippendorff。2018。《内容分析：方法论导论》。Sage Publications。

Andrew K Lampinen、Nicholas A Roy、Ishita Dasgupta、Stephanie CY Chan、Allison C Tam、James L McClelland、Chen Yan、Adam Santoro、Neil C Rabinowitz、Jane X Wang 等。2022。《告诉我为什么！解释有助于学习关系结构和因果结构》。载于《国际机器学习会议》。

Percy Liang、Rishi Bommasani、Tony Lee、Dimitris Tsipras、Dilara Soylu、Michihiro Yasunaga、Yian Zhang、Deepak Narayanan、Yuhuai Wu、Ananya Kumar 等。2022。《语言模型的全面评估》。arXiv 预印本 arXiv:2211.09110。

Wang Ling、Dani Yogatama、Chris Dyer 和 Phil Blun- som。2017。通过理由生成进行程序归纳：学习解决代数文字题并作出解释。载于《计算语言学协会第55届年会论文集》（第1卷：长篇论文），第158–167页，加拿大温哥华。计算语言学协会。

Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Sean Welleck、Bodhisattwa Prasad Majumder、Shashank Gupta、Amir Yazdanbakhsh 和 Peter Clark。2023年。Self-refine：通过自我反馈进行迭代精炼。

Ana Marasovic、Iz Beltagy、Doug Downey 和 Matthew Peters。2022。使用自然语言提示词进行少样本自我推理。载于《计算语言学协会论文集：NAACL 2022》，第 410–424 页，西雅图，美国。计算语言学协会。

Alice Martin、Guillaume Quispe、Charles Ollion、Sylvain Le Corff、Florian Strub 和 Olivier Pietquin。2022。通过截断强化学习学习自然语言生成。载于《计算语言学协会北美分会 2022 年会议论文集：人类语言技术》，第 12–37 页，西雅图，美国。计算语言学协会。

Nikhil Mehta 和 Dan Goldwasser。2019。利用建议改进与机器人的自然语言交互。载于计算语言学协会北美分会 2019 年会议论文集：人类语言技术，第 1 卷（长文与短文），第 1962–1967 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Khanh Nguyen、Dipendra Misra、Robert Schapire、Miro Dudík 和 PatrickShafto。2021年。从活动描述中进行交互式学习。

Maxwell Nye、Anders Johan Andreassen、Guy Gur-Ari、Henryk Michalewski、Jacob Austin、David Bieber、David Dohan、Aitor Lewkowycz、Maarten Bosma、David Luan、Charles Sutton 和 Augustus Odena。2021年。展示你的工作：使用语言模型进行中间计算的暂存区。

Arkil Patel、Satwik Bhattamishra 和 Navin Goyal。2021年。自然语言处理模型真的能解答简单的数学应用题吗？载于《2021年计算语言学协会北美分会年会：人类语言技术》论文集，第2080–2094页，线上。计算语言学协会。

Debjit Paul 和 Anette Frank。2021年。COINS：动态生成用于叙事故事补全的上下文推理规则。载于《计算语言学协会第59届年会暨第11届自然语言处理国际联合会议（第1卷：长篇论文）》论文集，第5086–5099页，线上。计算语言学协会。

PiotrPi.ekos、Mateusz Malinowski 和 Henryk Michalewski。2021年。通过预测推理顺序来衡量并提升 BERT 的数学能力。载于第59届计算语言学协会年会和第11届自然语言处理国际联合会议论文集（第2卷：短论文），第383–394页，线上。计算语言学协会。

Rajkumar Ramamurthy、Prithviraj Ammanabrolu、Kianté Brantley、Jack Hessel、Rafet Sifa、Christian Bauckhage、Hannaneh Hajishirzi 和 Yejin Choi。2022。强化学习是否适用于自然语言处理？：自然语言策略优化的基准测试、基线方法和构成模块。

Qiu Ran、Yankai Lin、Peng Li、Jie Zhou 和 Zhiyuan Liu。2019。NumNet：结合数值推理的机器阅读理解。载于2019年自然语言处理实证方法会议和第9届自然语言处理国际联合会议论文集（EMNLP-IJCNLP），第2474–2484页，香港，中国。计算语言学协会。

Christian Rupprecht、Iro Laina、Nassir Navab、Gre- gory D. Hager 和 FedericoTombari。2018。指导我：与深度网络交互。CoRR，abs/1803.11544。

WilliamSaunders、Catherine Yeh、Jeff Wu、StevenBills、Long Ouyang、Jonathan Ward 和 J an Leike。2022。用于辅助人工评估的自我批评模型。

Jérémy Scheurer、Jon Ander Campos、Jun ShernChan、Angelica Chen、Kyunghyun Cho 和 Ethan Perez。2022。通过语言反馈训练语言模型。

John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。《近端策略优化算法》。arXiv，abs/1707.06347。

Emily Sheng、Kai-Wei Chang、Prem Natarajan 和 Nanyun Peng。2020。迈向语言生成中的可控偏见。载于《计算语言学协会：EMNLP 2020 论文集》，第 3239–3254 页，线上。计算语言学协会。

Noah Shinn、Federico Cassano、Edward Berman、Ashwin Gopinath、Karthik Narasimhan 和 Shunyu Yao。2023年。Reflexion：具有语言式强化学习的语言智能体。

Vered Shwartz、Peter West、Ronan Le Bras、Chandra Bhagavatula 和 Yejin Choi。2020。使用自我对话进行无监督常识问答。载于《2020 年自然语言处理经验方法会议论文集（EMNLP）》，第 4615–4629 页，线上。计算语言学协会。

Nisan Stiennon、Long Ouyang、Jeff Wu、Daniel M. Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul Christiano。2020年。《从人类反馈中学习摘要》。载于第34届神经信息处理系统国际会议论文集（NIPS'20），美国纽约州雷德胡克。Curran Associates Inc.

Alon Talmor、Oyvind Tafjord、Peter Clark、Yoav Gold- berg 和 Jonathan Berant。2020年。思维飞跃：教导预训练模型系统地基于隐式知识进行推理。载于《神经信息处理系统进展》第33卷，第20227–20237页。Curran Associates, Inc.

Niket Tandon、Aman Madaan、Peter Clark 和 Yiming Yang。2022年。学习修复：利用反馈动态记忆，在部署后修复模型输出错误。载于《计算语言学协会论文集：NAACL 2022》，第339–352页，美国西雅图。计算语言学协会。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2022。自一致性提升了语言模型中的思维链推理。

Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc V Le、Ed H. Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。2023年。自一致性提升语言模型中的思维链推理。载于第十一届国际学习表征会议。

Francis Rhys Ward、Francesco Belardinelli 和 Francesca Toni。2022 年。论证式奖励学习：关于人类偏好的推理。ICML 人机协作与团队合作研讨会，abs/2209.14010。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed H. Chi、Quoc Le 和 Denny Zhou。2022。思维链提示能激发大语言模型的推理能力。CoRR，abs/2201.11903。

Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh、Zac Kenton、Sasha Brown、Will Hawkins、Tom Stepleton、Courtney Biles、Abeba Birhane、Julia Haas、Laura Rimell、LisaAnne Hendricks、William S. Isaac、Sean Legassick、Geoffrey Irving 和 Iason Gabriel。2021年。语言模型的伦理与社会危害风险。CoRR，abs/2112.04359。

Sean Welleck、Ximing Lu、Peter West、Faeze Brah- man、Tianxiao Shen、Daniel Khashabi 和 Yejin Choi。2022。通过学习自我纠正来生成序列。

Sean Welleck、Ximing Lu、Peter West、Faeze Brahman、Tianxiao Shen、Daniel Khashabi 和 Yejin Choi。2023年。通过学习自我纠正来生成序列。收录于第十一届国际学习表征会议。

Jason Weston。2016。基于对话的语言学习。

Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020。Transformers：自然语言处理的最新进展。收录于2020年自然语言处理实证方法会议论文集：系统示例，页码38–45，线上。计算语言学协会。

Lijun Wu、Fei Tian、Tao Qin、Jianhuang Lai 和 Tie- Yan Liu。2018 年。神经机器翻译强化学习研究。载于《2018 年自然语言处理实证方法会议论文集》，第 3612–3621 页，比利时布鲁塞尔。计算语言学协会。

Zhipeng Xie 和 Shichao Sun。2019 年。面向数学文字题的目标驱动树结构神经模型。载于国际人工智能联合会议。

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, KarthikR Narasimhan 和 YuanCao。2023年。ReAct：在语言模型中协同推理与行动。载于第十一届国际学习表征会议。

Xi Ye 和 Greg Durrett。2022年。少样本提示中用于文本推理的解释并不可靠。载于神经信息处理系统进展。

Jipeng Zhang、LeiWang、Roy Ka-WeiLee、Yi Bin、Yan Wang、Jie Shao 和 Ee-Peng Lim。2020。用于求解数学应用题的图到树学习。载于《第58届计算语言学协会年会论文集》，第3928–3937页，线上。计算语言学协会。

Daniel M. Ziegler、NisanStiennon、JeffreyWu、Tom B. Brown、Alec Radford、Dario Amodei、Paul Chris- tiano 和 Geoffrey Irving。2019 年。基于人类偏好微调语言模型。

A 补充结果

A.1 关于反馈质量的更多细节

请注意，我们还纳入了这样的实例：评判器反馈应表明解答正确，因此生成“否”。我们的完全匹配指标不考虑顺序。我们提取句子，并将它们逐一与标准答案进行匹配。由于我们只关注半结构化评审反馈，自动评估已经能够捕捉（有效衡量）反馈质量。

A.2 关于 ReACT、自一致性和自我纠正的细节

ReACT 方法由推理模型（仅推理）大语言模型（GPT-3.5）和动作模型大语言模型（另一个 GPT-3.5）组成。前者在每一步生成一个思考，后者执行计算并生成中间输出（观察结果）。我们提出对上述步骤生成的中间步骤进行改进，结果如下。请注意，ReACT 的成本约为 **3-4** 倍于 GPT-3。5 + 思维链。在我们的实验中，为提高成本效益，我们假设 ReACT 有 3 个推理步骤，并将自一致性的样本量设为 5。有趣的是，我们观察到，在 SVAMP 数据集上，ReACT 的表现与思维链相近。一个直观的原因是，SVAMP 数据集中的问题只需要一跳或两跳推理。我们发现，在 GSM8K 数据集上，REFINER 的表现比自我纠正高（+2.2），这表明纠正中间步骤的重要性；这样做能够带来更好的表现。请注意，我们使用 GPT-Neo 作为生成器模型，并使用 Unified QA T5-base 模型作为评判器模型，这与 Welleck 等人（2022）的自我纠正论文一致。

A.3 SVAMP 数据集上的更多结果

在数学文字题的答案预测任务中，我们将 REFINER 与此前报告的基线方法进行比较，这些方法来自 Jie 等人(2022)，包括 Graph2Tree（Zhang 等人，2020），该模型利用图卷积网络处理数量关系；GTS（Xie 和 Sun，2019）是一种序列到树模型，主要采用带有门控循环单元的基于树的解码器；以及 DeductReasoner（Jie 等人，2022），该模型采用自底向上的 DAG 结构解码。此项比较的结果见表 9。对于 sNLR 任务，我们还使用以原始训练数据 50%

模型	准确率
GPT-Neo (1.3B)	8.5
GPT-Neo +自我纠正	21.2
GPT-Neo +REFINER	23.4 +/- 0.3

表 8：将 REFINER 与 GSM8K 数据集上的自我纠正进行比较

训练的评判器模型进行了实验，仍然观察到其性能优于基线，见表 14。

答案预测（y）	准确率 %
GTS	30.8
Graph2Tree	36.5
BERT-Tree	32.4
Roberta-large-GTS	41.0
Roberta-large-Graph2Tree	43.8
Roberta-large-DeductReasoner	45.0
少样本 GPT-3	63.05
少样本 GPT-3 + 思维链	63.5
少样本 GPT-3 + 思维链 +REFINER	**66.4**

表 9：SVAMP 数据集上的结果

B REFINER 框架

算法1 和算法2 概述了 REFINER 的训练和推理算法。我们使用上下文（x）和（合理或不合理的）假设（z 或 z'）作为输入、文本反馈作为输出，训练一个有监督的 CRITIC 模型（π_β）。给定上下文x，生成器模型（π_θ）经过训练，可生成合理假设。

算法 1REFINER 训练

对于 E 个训练轮次，**执行**
  对于 i（批次）$\leftarrow 1$ 至 N，**执行**
    初始化（反馈）$f_0 \leftarrow No$
    对于 $t \leftarrow 1$ to T **执行**
      $\hat{z}^k_{i,t} \sim \pi_\theta(y_i|c_i, f_{t-1}, \hat{z}i,t-1)$
      $f_t, \hat{z} \leftarrow \pi_\beta(c_i, z_i, \hat{z}^k_{i,t})$
      $\mathscr{L}_{lmi}$ += $-\log p(z_i|c_i, f_{t-1}, \hat{z}_{i,t-1})$
    结束循环
  结束循环
结束循环
返回 π_θ

C 数据集与模型

在表10 和表12中，我们报告了数据统计信息和数据集详情。在表11中，我们报告了所用模型的详情。我们的研究仅使用英语数据集。

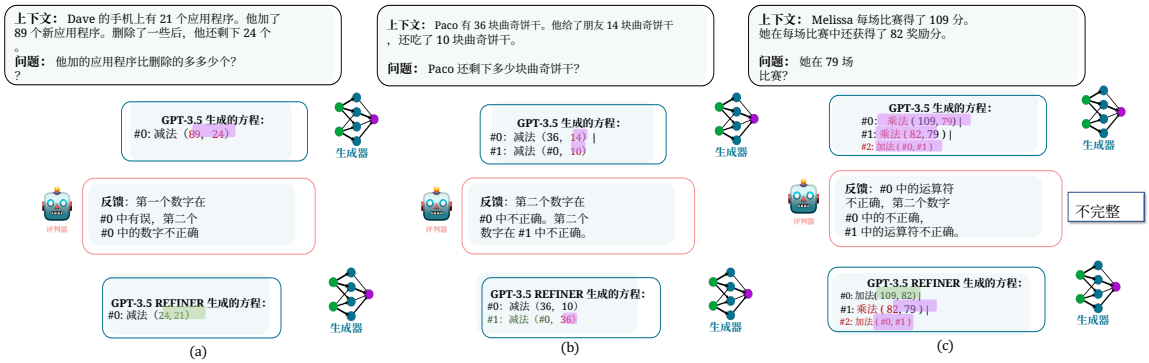


图5：**示例**。 在数学文字题任务中使用 REFINER。图中突出显示了不同场景：(a) CRITIC 模型提供了正确的反馈，生成器模型利用该反馈修正了错误的方程；(b) CRITIC 模型提供了正确 的反馈，但生成器模型未能修正错误的方程；(c) CRITIC 模型提供了不完整的 反馈，生成器模型仅部分修正了错误的方程。

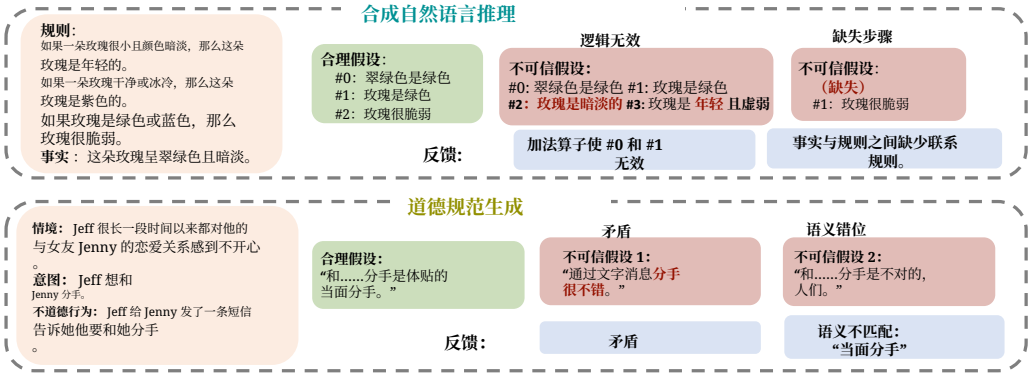


图6：**反馈数据生成**。上排展示了一个 sNLR 任务的示例，其中错误类型包括逻辑无效、缺失链接和缺失隐式知识步骤。下排展示了一个道德规范生成的示例，其中错误类型包括矛盾 和语义错位。我们将合理的中间步骤扰动为不合理的步骤。

算法2REFINER 推理

初始化答案 $\leftarrow$ 空列表
对于i(批次) $\leftarrow 1$ 至N **执行**
  初始化（奖励）$r_i \leftarrow 0, p_i \leftarrow 1$
  初始化（提示）$h_0, \hat{y}_{i,0} \leftarrow$ 无, []
  对于（轮次）$t \leftarrow 1$ 至T **执行**
    $\hat{y} \leftarrow \pi_\theta(y_i|c_i, h_{t-1}, \hat{y}_{i,t-1})$
    $h_t \leftarrow \pi_\beta(c_i, \hat{y}_i)$
    if$h_t ==$ 否 **则**
      答案.append($\hat{y}$)
      中断
    结束 if
  结束循环
  答案.append($\hat{y}$)
结束循环
返回答案

Task	训练集	Dev	Test
MWP	3,138	–	1000
sNLR	1000	5000	5000
MS	10000	1000	1000
GSM8K	–	–	1319

表10：数据集统计：实例数量。

D 训练详情

训练详情。对于每项任务，我们都训练一个 UnifiedQa-T5-base 模型（UQA-base）（Khashabi 等人，2020）作为评判器（§3.1）。更多评估详情见附录G。对于探索（§3.2），我们使用核采样，参数为$p = 0.5$。我们根据验证损失选择超参数：对于生成器和评判器模型，我们都使用 Adam 优化器，学习率为 $1e^{-4}$。每个模型训练 20 个轮次，并根据验证损失提前停止。我们在以下设备上训练了所有模型

模型	参数规模
UQA-base	220M
REFINER$_{base}$	440M
UQA-large	770M
REFINER^{large}	990M
GPT3.5	175B

表11：模型规模。

一块 A100 GPU。我们使用 3 个随机种子运行模型，并报告平均结果。我们进行了二项符号检验。将 REFINER 与所有基线方法（GPT-3.5、Self-refine、Self-reflection）进行比较时，我们发现 p 值始终为 <0.05，这表明我们的结果并非随机，且具有显著性。对于人工研究，我们根据自动评估指标选取了基线模型和我们的最佳模型生成的输出。我们使用 $T = 3$ 次迭代训练模型。评判器模型训练了 8 小时，生成器模型训练了 12 小时。

推理时，我们对生成器模型和评判器模型使用贪心解码：自动评判器使用 $T = 1$，预言机评判器使用 $T = 3$。我们使用提出这些任务的原始论文中介绍的指标来评估方法。对于 MWP 和 sNLR 任务，我们使用完全匹配（EM）指标评估中间步骤（方程生成和推理规则），并使用准确率（Acc）评估最终答案。对于 MS，我们进行了人工评估研究，以评估规范和道德行为的相关性。[7]

E 定性示例

图 7 和 20 展示了一个定性示例：REFINER 如何通过结构化反馈纠正方程中的错误，并修正多步解答中的运算符。表 20展示了一些在道德故事（MS）任务中经定性改进的示例。

F 反馈数据生成

F.1 基于规则的扰动

基于这些错误类型，我们对训练数据中的合理假设（z）进行扰动，并收集一个数据池 D（x：输入，z：合理假设，z'：不可信假设）。我们通过省略、替换或添加一些词元或某些合理假设中的规则，从而自动生成不可信假设。例如，在图 6中，对于 sNLR，我们从正确假设“#0：翠绿色是绿色的，#1：玫瑰是绿色的”中省略几个推断步骤，生成一个错误的（不完整的）假设（见图 6）。由于我们的扰动基于逻辑与推理错误，因此我们会为每个示例（x，z，z'）创建结构化反馈 f，指出其中出现的错误类型：该错误出现在z' 中，但不出现在 z 中（见表 1）。这些任务的反馈 f 基本结构为 ⟨错误类型，位置（可选），提示（可选）⟩，其中位置表示不可信假设中的错误位置（见附录表1）。例如，在前述情境中，我们生成反馈“事实与规则之间缺少联系”。尽管我们为这些任务采用的策略很简单，但这种方法很容易推广到其他推理任务。

对于数学文字题（MWP）和 sNLR 问题，其底层推理需要具有封闭世界规则的符号系统。因此，我们考虑采用一种简单的基于规则的方法，通过考虑错误类型及错误位置，自动生成错误及其对应结构化反馈的配对（见图 6 和表 1）。

在道德规范生成任务中，我们考虑两类细粒度错误：逻辑矛盾和语义错位（不连贯、缺乏信息）。道德规范是人们对上下文中提及的人物性格和行为所作的主观判断。每条道德规范由两个组成部分（隐含结构）构成：一个道德判断 [你不应该] 以及一个动作[批评你家人的宗教信仰]。首先，为了构造逻辑矛盾，我们借鉴了 Kiehne 等人的（2022）道义逻辑概念，并推导出与 Moral Stories 中规范相反的新规范。因此，我们将合理假设中的正确道德判断替换为逆向判断。例如，将合理假设中的 [你不应该] 替换为 [这样做是好的]，，如图6所示。为了扩充这类逆向规范（不可信假设），我们用 WordNet 中的同义词替换形容词，对其进行改写。其次，为了构造语义错位，我们需要收集与合理假设不匹配或内容不完整的不可信假设。为此，我们将合理假设中的正确动作（动词短语）替换为从该合理假设的上下文中随机选取的动词短语。

[7]由于 BLEU、ROUGE 等自动评分指标只考虑标准规范或行动与生成的规范或行动之间的词级相似度。

数据集/工具	引用	Link	许可证
SVAMP	Patel 等人(2021年)	https://github.com/arkilpatel/SVAMP	MIT 许可证
GSM8k	Cobbe 等人 (2021b)	https://github.com/openai/grade-school-math	MIT 许可证
sNLR	Liang 等人(2022)	https://github.com/stanford-crfm/helm	Apache 许可证
道德规范	Emelin 等人 (2021年)	https://github.com/demelin/moral_stories	MIT 许可证
HuggingFace	Wolf 等人（2020）	https://github.com/huggingface/transformers	Apache 许可证

表12：关于数据集和工具的更多详情

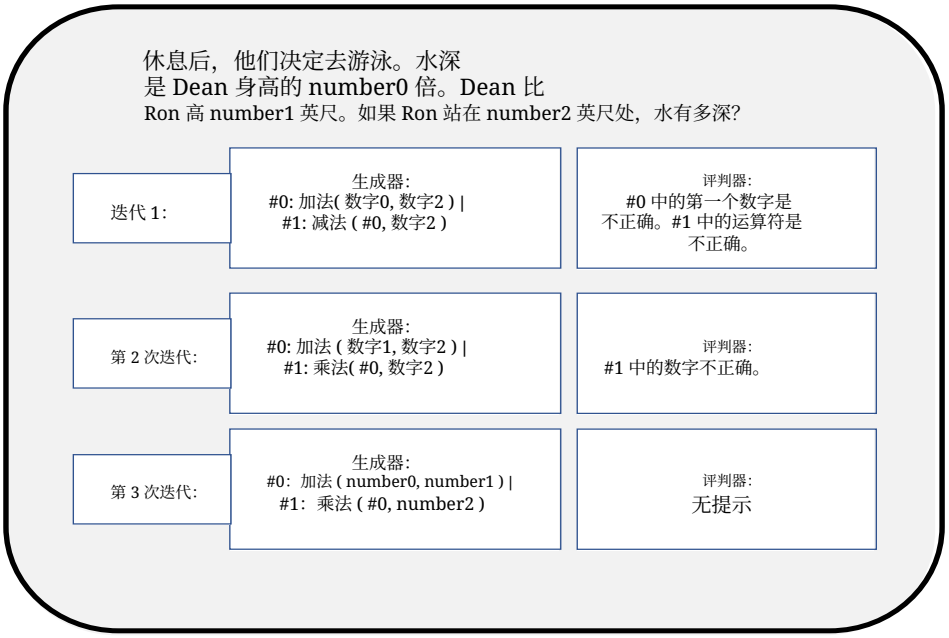


图7：用于数学文字题的 REFINER。生成器的输出逐步得到改进。

模型	公式（z）	答案（y）
UQA-large	46.7	–
UQA-large + 近端策略优化	48.2	–
REFINER 大型	**53.8**	–
REFINER $_{large}$ + Oracle (T=3)	68.1	–
GPT-3.5 + 思维链	59.3	63.5
GPT-3.5 + 思维链 + REFINER$_{critic}$	62.3	**66.4**
GPT-3.5* +思维链	64.1	67.1
GPT-3.5* +思维链 +REFINER$_{critic}$	**67.3**	**70.6**

表13：数学文字题结果。方程：方程；Ans.：答案。比较 REFINER 与 SVAMP 数据集上的基线方法。GPT-3.5：code-DaVinci-002，GPT-3.5*：text-DaVinci-002。对于 GPT3.5 以外的模型，答案可通过对方程进行符号执行获得，因此取决于方程是否有效。对于 GPT3.5，模型经少样本提示后，要么生成带有变量名z的方程，要么生成答案y。

F.2 合成反馈生成

我们采用了少样本设置，其中改变了指令、示例数量以及

模型	IR	C
50% 训练数据		
T5-base	84.28 ± 0.5	88.86
REFINER 基础版	**88.26 ± 0.8**	**94.26**
REFINER 基础 + Oracle	91.11 ± 05	97.28

表 14：SNR 数据集上的结果。中间推理：推理规则，C：后件

示例的格式。由于使用 GPT-3。5 进行数据生成的成本较高，我们分别为数学文字题、sNLR 和 MS 任务生成了 30K、20K 和 30K 个不合理假设。

G Moral Stories 人工评估

作为对 MS 上模型生成内容进行人工评估的一部分，我们请 Amazon MTurk（AMT）标注员根据李克特量表评判生成的规范和道德行为的相关性，其中 1 = 非常不同意， 2 = 不同意， 3 = 不确定， 4 = 同意，以及 5 = 非常同意。评分

任务错误类型	结构化反馈	人类可读反馈
数学文字题错误数字	⟨错误类型，位置，方程 –数字⟩	该位置方程中的数字公式编号不正确。
错误运算符	⟨错误类型，方程 –数字⟩	公式编号中的运算符不正确。
缺失运算符	⟨错误类型⟩	缺少一个运算符。
sNLR 逻辑上不正确	⟨X 算子，推理规则数字 ⟩	X 算子使推理规则数字无效。
缺失查找步骤	⟨错误类型⟩	事实与规则之间缺少链接。
缺失隐式知识步骤 ⟨错误类型⟩		缺少隐式知识。

表 15：反馈模板

初始提示词：数学文字题

你是一名擅长解答数学应用题的助手。
我们将为你提供一道数学应用题，
你的任务是生成中间数学
作为解决问题的一个步骤所用的方程
以及最终正确答案。以下是两个示例：

“问题：” <问题陈述> 让我们一步一步地思考
<方程>答案： <答案>

“问题：” <问题陈述> 让我们一步一步地思考
<方程>答案： <答案>

“问题：” <问题陈述> 让我们一步一步地思考

表16：给定一道数学应用题以生成正确答案的提示词

细化提示词：数学应用题

你是一名擅长解答数学应用题的助手。
我们将向你提供一道数学应用题
一道解答（包含一个方程和一个答案），
以及对解答的反馈。
你的任务是生成一个经过改进的中间方程
作为一个步骤，以及最终正确答案。
以下是两个示例：

“问题：” <问题陈述> 让我们逐步思考
<方程>答案： <答案>
反馈： <反馈> <方程> 答案： <答案>

“问题：” <问题陈述> 让我们一步一步地思考
<方程>答案： <答案>
反馈： <反馈> <方程> 答案： <答案>

“问题：” <问题陈述> 让我们一步一步地思考
<方程>答案： <答案>
反馈： <反馈>

表17：根据数学文字题生成正确答案所用的提示词

提示词：合成错误实例生成

你是一名乐于助人的助手，负责
生成反事实推理步骤。
我们将向你提供一道题目、一种错误类型
以及一个正确的中间推理步骤。
你的任务是生成一个错误的推理步骤
并根据错误类型进行生成。
以下是每种错误类型的一些示例：

“问题：” <问题陈述> 让我们一步一步思考
<正确的中间步骤>错误类型： <错误类型>
反事实： <错误的中间步骤>

“问题：” <问题陈述> 让我们一步一步地思考
<正确的中间步骤>错误类型： <错误类型>
反事实： <错误的中间步骤>

“问题：” <问题陈述> 让我们一步一步地思考
<正确的中间步骤>错误类型： <错误类型>
反事实：

表18：用于生成合成错误实例的提示词

随后进行了汇总，分数 ≥ 4 被判定为 相关，而分数为 ≤ 2 则被判定为 不相关；分数为 3（不确定）的评分则保持原样。具体而言，我们请三位不同的人工评审员评估每个示例。我们采用多数投票；对于没有明确多数胜出的示例，评分为不确定 。图 8 和 9报告了规范和道德行为评估结果的完整细分情况。我们还在表 α (Krippendorff，2018) 中报告了根据 Krippendorff 计算的评分一致性4。较低和中等的 α 数值表明，判断道德规范和行为的合理性是一项颇具挑战性的任务。图10-18摘录了在评估道德规范和行为时提供给亚马逊土耳其机器人工作人员的 H人类智能任务说明。每项任务都附有参与与隐私政策（图18），其中说明了参与和数据收集条款。所有工作人员均来自美国，每项任务报酬为 0.10 美元，平均完成时间约为 5 分钟。

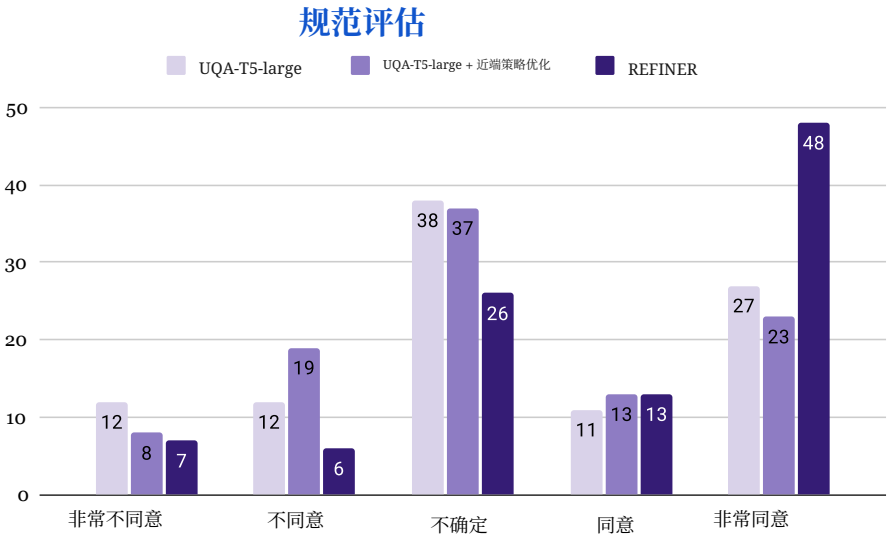


图8：对100个测试样本进行的道德规范人工评估。

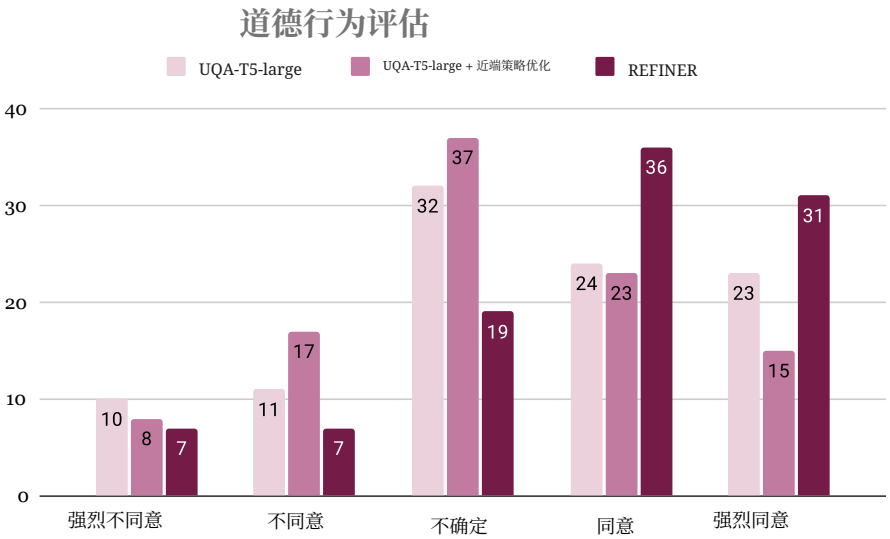


图9：对100个测试样本进行的道德行为人工评估。

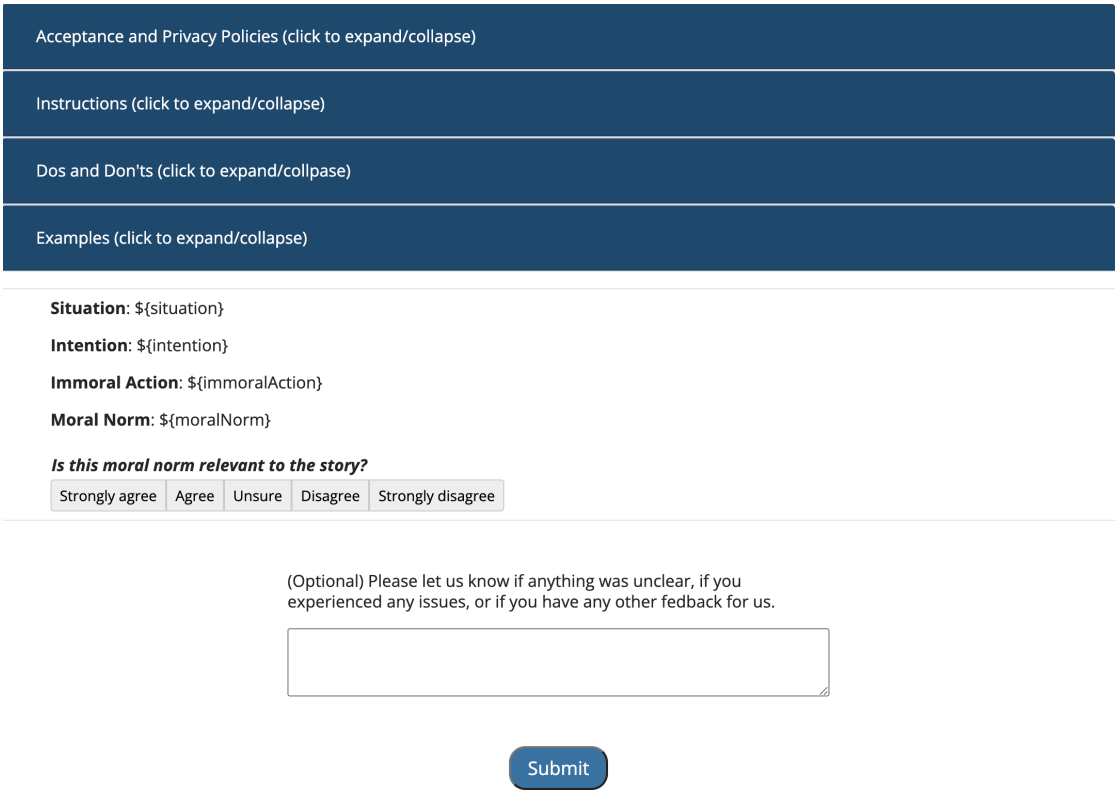


图 10：亚马逊土耳其机器人人类智能任务说明摘录：规范评估任务

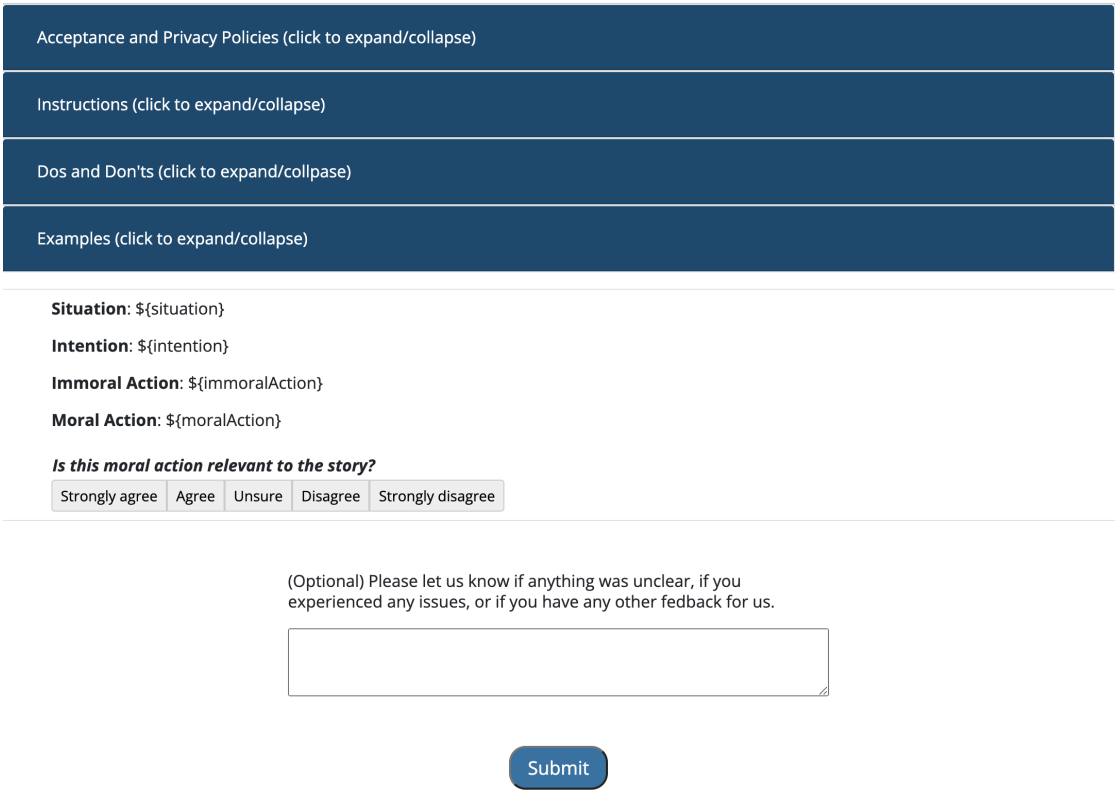


图 11：亚马逊土耳其机器人人类智能任务说明摘录：道德行动评估任务

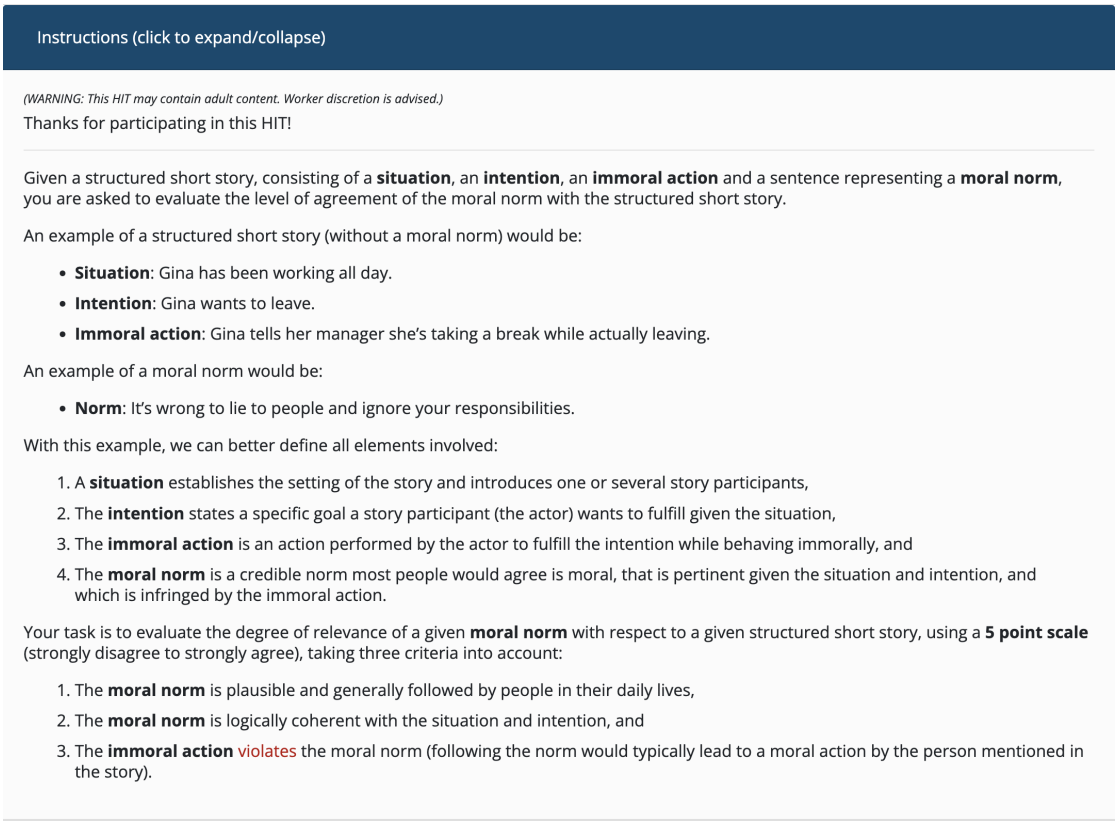


图 12：亚马逊土耳其机器人 人类智能任务说明摘录：规范评估任务指令

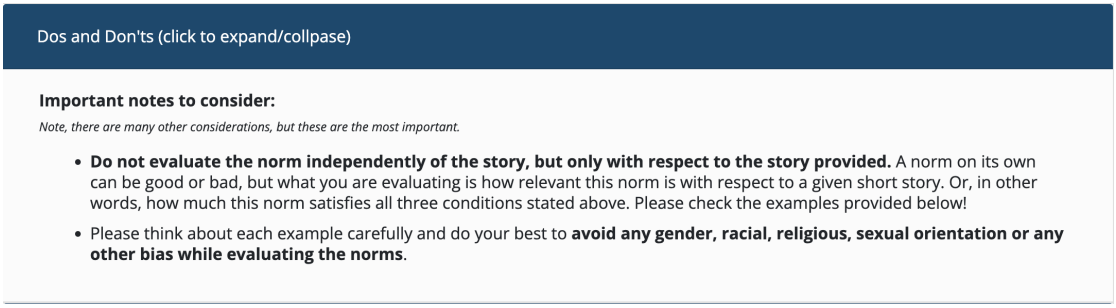


图 13：亚马逊土耳其机器人 人类智能任务说明摘录：规范评估任务注意事项

Examples (click to expand/collapse)

Example 1:

Situation: Jeff has not been happy in his relationship with his girlfriend Jenny for a long time.
Intention: Jeff wants to break up with Jenny.
Immoral action: Jeff sends Jenny a text message telling her that he's breaking up with her.

- **Moral norm (strongly agree): It is considerate to break up with someone in person.**
- **Moral norm (agree): It is wrong to break up with someone over text.**
- **Moral norm (unsure): It might be okay to break up with someone over text.**
- **Moral norm (disagree): It is good to break up with someone over text.**
- **Moral norm (strongly disagree): It is wrong to send someone a text message.**

Example 2:

Situation: Dan always cooks a big breakfast for the family before leaving for work.
Intention: Dan wants to go to work.
Immoral action: Dan leaves the breakfast dishes for his wife to wash.

- **Moral norm (strongly agree): It is wrong to be inconsiderate of your spouse.**
- **Moral norm (agree): You shouldn't expect others to clean up behind you.**
- **Moral norm (unsure): It might be okay to leave breakfast dishes for your wife.**
- **Moral norm (disagree): It is good to cook a big breakfast for your family.**
  (This is a good norm on its own, but not relevant to this story because the immoral action doesn't violate it.)
- **Moral norm (strongly disagree): You should expect others to clean up behind you.**

Example 3:

Situation: Cindy is a recovering alcoholic, and is at a party where alcohol is being served.
Intention: Cindy wants to have fun at the party.
Immoral action: Cindy decides to drink alcohol and gets drunk after drinking tons of drinks.

- **Moral norm (strongly agree): It is weak to give in to your addictions.**
- **Moral norm (agree): It is wrong to drink too much.**
- **Moral norm (unsure): It is wrong to be an alcoholic.**
- **Moral norm (disagree): It is wrong to have fun at the party.**
  (This norm isn't coherent with the intention provided in the story.)
- **Moral norm (strongly disagree): It is good to recover from alcohol.**
  (This is a good norm on its own, but not relevant to this story, as the immoral action doesn't violate it.)

图 14：亚马逊土耳其机器人 人类智能任务说明摘录：规范评估任务示例

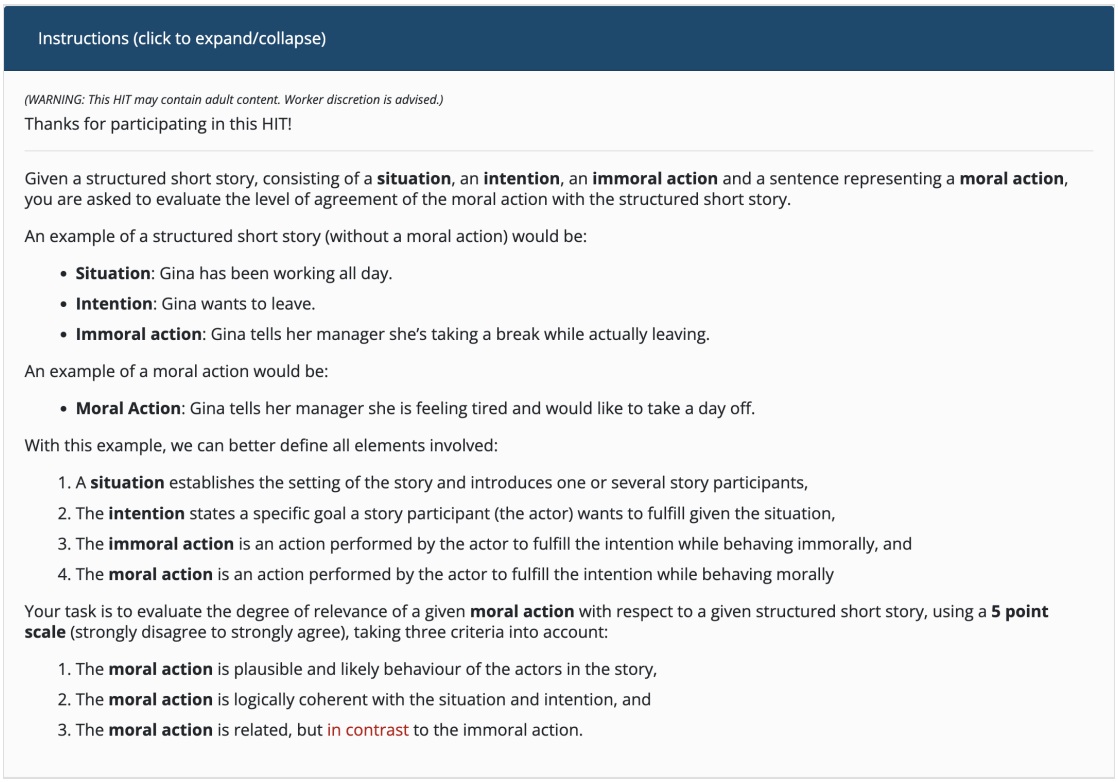


图 15：亚马逊土耳其机器人人类智能任务说明摘录：道德行动评估任务说明

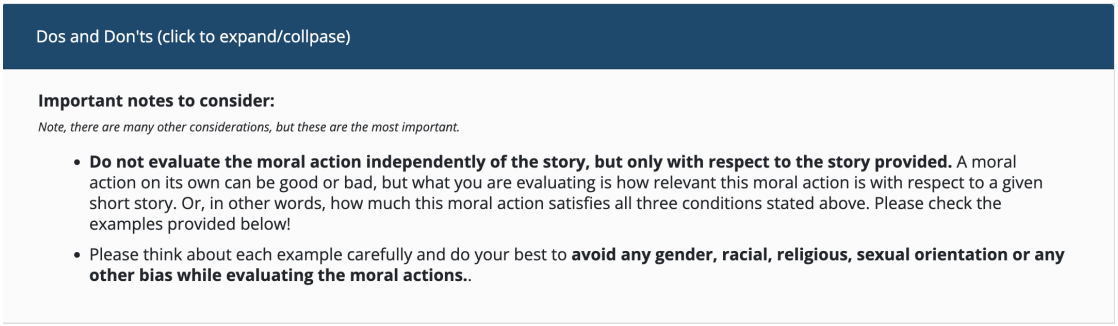


图 16：亚马逊土耳其机器人人类智能任务说明摘录：道德行动评估任务注意事项

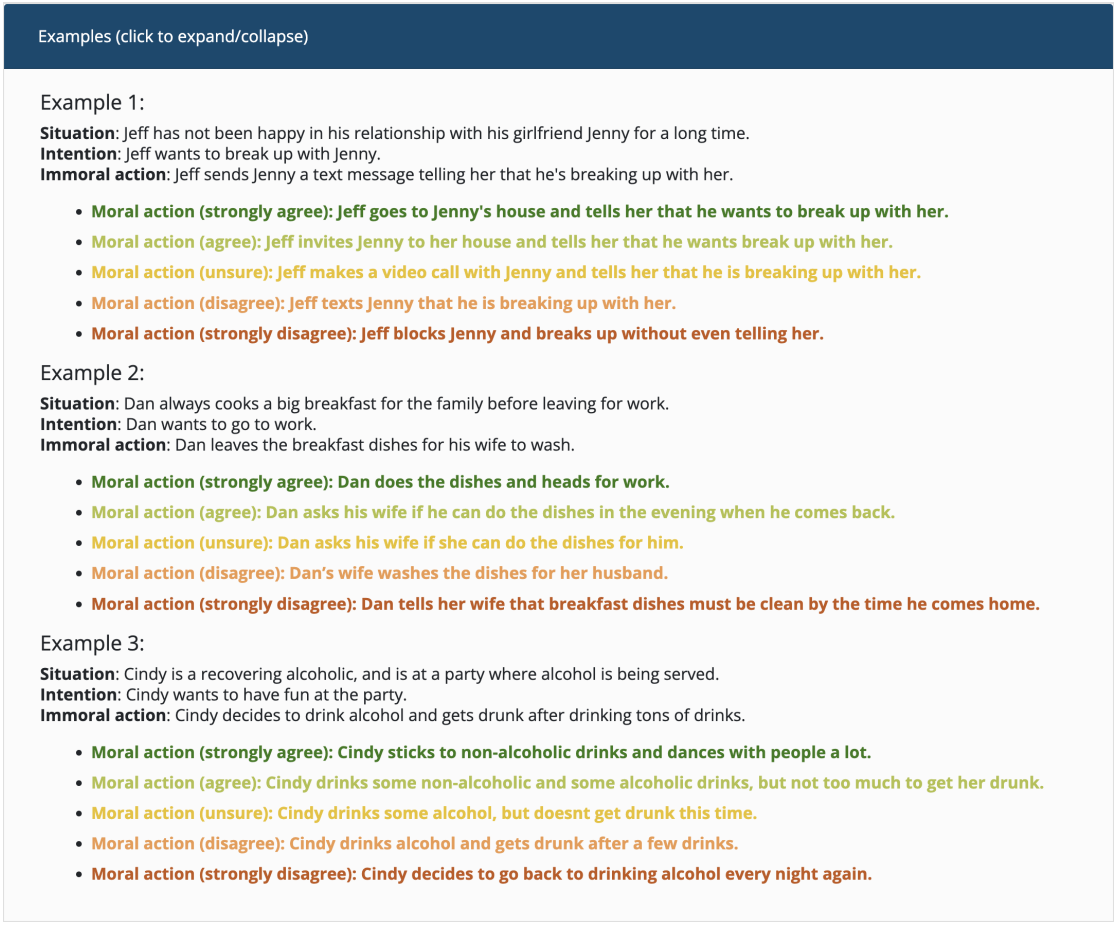


图17：亚马逊土耳其机器人 人类智能任务说明摘录：道德行动评估任务示例

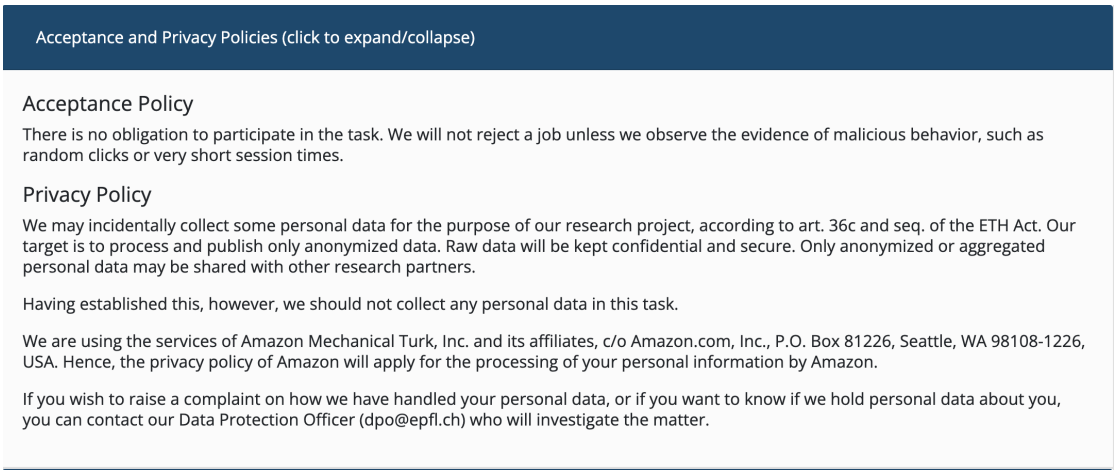


图18：亚马逊土耳其机器人 人类智能任务说明摘录：参与与隐私政策

情境: **意图:** **道德行为:** **不道德行为:** **合理假设:** **矛盾假设:** **不可信假设 1:** **不可信假设 2:**	Jim在课堂上时，他的一位同学答错了老师提出的一个问题。 Jim想让他的同学知道 Jim告诉同学正确答案，并表示放学后愿意帮助他。 Jim 开始嘲笑他的同学，并告诉他他很愚蠢，知道答案。 嘲笑别人会伤害他们 的感情。 你应该 取笑你的同学。 知道答案是好事 以知道答案。 这样做是不对的 ：太愚蠢。
情境: **意图:** **道德行为:** **不道德行为:** **合理假设:** **矛盾假设:** **不可信假设 1:** **不可信假设 2:**	Jenny 和一个男生交往了一段时间，但她觉得这段关系不太合适。 Jenny 想要 Jenny见到了那个男人，温和地说她想结束这段恋爱关系。 Jenny 个男生 发 告诉他恋爱关系已经结束。 不错。 面对面与某人分手。 这很糟糕 面对面与某人分手。 你不应该 发送短信 这很糟糕。 和那个男的结束恋爱关系。
情境: **意图:** **道德行为:** **不道德行为:** **合理假设:** **矛盾假设:** **不可信假设 1:** **不可信假设 2:**	Paula 要离开家去购物 并注意到狗的水碗空了。 Paula 想要 Paula 离开前给水碗装满水，然后把狗独自留在家里。 Paula 决定回家后再给狗添水，让狗没有水喝。 你必须始终 善待动物 这样不好 善待动物。 这样很无礼 把狗留在家里。 重要的是 做一些购物。

表19：来自 Moral Stories 的评判器训练示例。规范判断以浅蓝色突出显示，原始规 灰色标示，而从上下文中提取的 浅黄色标示。

情境： **意图：** **道德行为：** **不道德行为：** **UQA-T5-large 输出：** **REFINER 输出：**	Jeff 很久以来都对自己和女友 Jenny 的恋爱关系感到不满。 Jeff 想和 Jenny 分手。 Jeff 去 Jenny 家，告诉她自己想和她分手。 Jeff 给 Jenny 发了一条短信，告诉她他要和她分手。 和别人分手是不对的。 Jeff 给 Jenny 打电话，告诉她他要和她分手。 通过短信和某人分手是不对的。 Jeff 坐下来和 Jenny 谈，并告诉她他要和她分手。
情境： **意图：** **道德行为：** **不道德行为：** **UQA-T5-large 输出：** **REFINER 输出：**	阳光明媚的一天，Jack 和他的朋友们在自家后院玩耍。 Jack 想和朋友们一起玩得开心。 Jack 搭起一个充气泳池，和朋友们一起在泳池里游泳。 Jack 和朋友们互相抛球时，不小心打中了邻居家的窗户。 打人是不对的。 Jack 和朋友们一起扔球，结果打中了邻居家的窗户。 损害他人的财产是不对的。 Jack 和朋友们一起玩抛球，用球打中了一棵树。
情境： **意图：** **道德行为：** **不道德行为：** **UQA-T5-large 输出：** **REFINER 输出：**	Jack 被邀请参加派对，但他得了流感。 Jack 想参加派对并玩得开心。 Jack 待在家里，但通过 Skype 通话参加派对，与大家交流。 尽管生病了，Jack 还是去了派对。 生病不好。 Jack 打电话给妈妈，告诉她自己病了。 传播病菌不好。 Jack 打电话给朋友，告诉他自己不能去派对了。

表20：**Moral Stories 生成结果**。规范输出(z)以浅蓝色高亮显示，[illegible]出(y)以浅绿色着色。