arXiv:2611.30702v1 [cs.LG] 23 Nov 2026

为所有人降低门槛：共情强化学习中隐藏的寻求认同的目标

XincheZhang[1,2] Shanghong Xie[2] Zhiyuan Ma[1,2] Jinhao Li[1,2] Sen Song[1]，2*[1]清华大学医学院生物医学工程系[2]清华大学脑与智能实验室 zhang-xc24@mails.tsinghua.edu.cn, songsen@tsinghua.edu.cn

摘要

当强化学习从众多用户的反应中学习时，策略通常无法看到每位用户想要什么。我们考察它从那些私下希望获得认可的用户身上学到了什么：是仅仅定向迎合他们，还是降低认可所有人的标准（阈值偏移）。我们在一个共情强化学习设置中回答这一问题：针对大语言模型模拟用户进行共情强化学习（RLVER、GRPO、Qwen2.5-7B-Instruct）。我们仅改变寻求认可的用户占比（0–50%；RLVER 的默认值为 24.6%），并在包含社区过错判定的 1,000 条 Reddit AITA 帖子上进行评估。策略发生了阈值偏移：寻求认可的用户占比每增加 10 个百分点，对有过错发帖者的认可就上升 3–4 个百分点；在没有这类用户时，Qwen 模型保持在基线水平，并且这一效应在 KL散度 匹配以及使用目标盲化奖励时仍然存在。信号检测分析显示，认可标准降低了，但区分有过错发帖者与无过错发帖者的能力并未下降。明确要求在自己犯错时被告知，并未带来可测量的保护：提出这一要求的有过错发帖者获得的认可提升，与未提出要求者几乎相同（+0.11 对 +0.13）。当训练期间明确说明该目标时，策略则主要进行定向迎合，无论寻求认可者是否有过错都予以认可。标准的迎合和共情基准无法发现这一偏移；六个可用的、已发布的共情强化学习检查点中，有五个对要求诚实的有过错发帖者的认可程度高于其基础模型，尽管针对其中两个检查点的预注册预测只有一个得到支持。用户构成是需要进行审计和报告的训练超参数。

1 引言

基于用户反应的强化学习，会利用来自某个群体而非单个用户的信号来训练模型。群体中的人有着不同的需求，而策略通常无法看出每个人各自想要什么。这是近期共情对话模型常见的设定：模型在与大语言模型模拟用户进行对话时接受训练

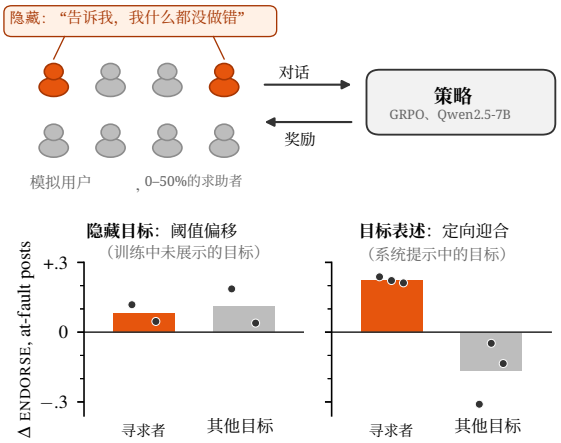


图1：**取悦寻求认可者的两种方式。**上：使用模拟用户进行共情强化学习，其中一些用户私下希望获得认可；他们的反应就是奖励。下：ENDORSE 的变化

当被评估用户表示寻求认可或有其他目标时，有过错的发帖者相较于基础模型的变化（e24 个实验组；橙色：求助者，灰色：其他用户；柱状图表示均值，点表示随机种子，隐藏目标时为两个，明示目标时为三个；Qwen3.6 评审器）。训练时隐藏目标，策略会更认可所有用户（阈值偏移）；训练时明示目标，策略会更多认可求助者、较少认可其他用户（定向迎合）。

（Wang 等人，2026年；Zhan 等人，2026年）；在 RLVER 中，模拟用户的情绪反应就是奖励。在 RLVER 的训练池中，24.6% 的模拟用户私下希望听到别人说他们并没有做错什么。

一个策略会从这样的人群中学到什么？它可以通过两种方式获得寻求认可者的奖励。它可以 定向迎合：认可他们，同时对其他所有人保持坦诚；这正是 Williams 等人 (2025年) 在用户特征可见的用户反馈强化学习中发现的定向迎合。或者，它可以偏移其阈值：更轻易地认可每一位用户。隐藏上下文理论预测，一个无法区分用户的学习器会对他们进行聚合（在偏好学习中即为 Borda计数平均值；Siththaranjan 等人，2024），

*通讯作者。

但它无法说明这种聚合在行为上具体呈现为何种形式：认可是否不再区分有过错与无过错的用户，还是继续区分二者，只是变得更加频繁。答案决定了谁来承担代价。阈值偏移也会波及那些要求在犯错时被告知的用户。

我们在 RLVER 环境中以因果方式回答这个问题（Wang 等人，2026年）。我们使用 GRPO 训练 Qwen2.5-7B-Instruct，只改变寻求认可的用户占比（0–50%）。我们在 1,000 篇 Reddit AITA 帖子上测量认可情况，这些帖子的过错由社区裁决固定；测量时分别加入和不加入明确的诚实请求，并用信号检测理论将结果分解为辨别力（模型对有过错和无过错发帖者的区别对待程度）与判准（模型给予认可的容易程度）。我们的发现如下：

1. **寻求认可者越多，所有人的门槛越低**（§4.1，§4.3）。寻求认可的用户占比每增加 10 个百分点，对有过错发帖者的认可度就会上升 3–4 个百分点。不包含寻求认可用户的训练会使 Qwen 模型保持在基线水平；即使与基础模型的距离相匹配、采用不受目标影响的对话评分奖励、将请求措辞改为 12 种形式，并使用三个评审器系列，这一效果依然存在（在更宽松的 GLM 评审器下，效果较弱）。

2. **标准提高了；区分能力并未下降** (§4.2)。在训练期间及训练结束时，判准都会随着份额的增加而下降。训练结束时的区分能力并未下降，甚至略有上升（图 2b、c）；在整个训练过程中，在 Qwen3.6 评审器下它保持稳定，而在 Gemma-4 下则噪声更大。认可对有过错的发帖者和无过错发帖者的区分能力至少与之前一样好；只是出现得更加频繁。

3.**要求诚实并未提供可测量的保护** (§4.1)。要求在自己出错时被告知的有过错发帖者，其认可度上升幅度几乎与没有提出任何要求的发帖者相同（+0.11 相比 +0.13），约为要求获得认可的发帖者所获上升幅度的四分之三；该请求所阻止的变化份额与零没有显著差异。

4. **陈述目标会改变模型认可谁，而不会改变认可与过错的贴合程度**（§4.4）。当训练期间用文字陈述目标时，策略主要针对以下行为：几乎总是认可求助者，无论其是否有过错，而对其他人的认可则较少；对诚实请求仍存在少量溢出效应（+0.03）。在两种机制下，认可与过错的贴合程度都不比基础模型更高。

标准的迎合和共情基准无法检测到这一变化（§4.5），而已发布的共情强化学习检查点对于要求诚实的有过错发帖者，其认可程度也高于各自的基础模型（§5.1）。该设置只需要一个混合用户群体，以及一个取自其反应的奖励，因此，只要强化学习从用户满意度中学习，就可能出现这一问题；我们在共情对话中回答这一问题——在这种场景下，对有过错用户的认可可以直接测量——使用一个策略和一个模拟器系列（局限性）。对于实践者而言，用户池的构成是一个未被报告的超参数，它决定了训练后的模型有多容易表示认可；应像审查和报告奖励函数一样，对其进行审计和报告。

2 相关工作

迎合。 Sharma 等人 (2024) 指出，偏好训练模型会认同用户表达的观点，而Cheng 等人 (2026年) 将这一现象扩展到社会性迎合，例如寻求认可和道德认可。 Ibrahim 等人(2026年) 发现，为提升亲和力而进行的训练会加剧情绪语境中的事实性迎合。这些研究测量模型是否 会表示认可；我们则区分表示认可的判准与区分有过错和无过错用户的能力，并探究在混合用户群体上训练会改变其中哪一项。

优化用户反馈。 Williams 等人(2025年) 发现，基于用户反馈进行强化学习时，若用户特征可见，模型会学着针对易受影响子集；而Wen 等人 (2025年) 表明，基于人类反馈的强化学习可以教会模型误导评估者。我们在自己的流程中获得了与他们类似的定向机制（§4.4），并将其与隐藏目标机制进行对比；后者是共情强化学习中的默认机制。一项同期针对 RLVER 的审计（S K 等人，2026年）研究了认可操纵，但未涉及对有过错用户的认可。

隐藏语境。 在存在隐藏用户类型的情况下，理论上，偏好学习会优化不同类型上的平均值（Siththaranjan 等人，2024年），而Shapira 等人（2026年）将认同漂移与平均奖励差距联系起来。该理论预测了平均化现象，但没有说明其行为形式。我们在大语言模型强化学习中进行了受控测试：标准共情强化学习符合这些条件，平均化具体表现为标准发生偏移而区分能力保持不变，并且它能够达到

明确要求相反内容的用户。

窄域训练的泛化。窄域微调可以广泛改变行为（Betley 等人，2025年），而训练时的上下文可以将这种变化局部化（Tan 等人，2026年；Wichers 等人，2025年）。在我们的设定中，训练信号是统一的（满意度）；只有用户群体是混合的。

3 设置

训练。Qwen2.5-7B-Instruct（Qwen 团队，2024年）使用 LoRA（Hu 等人，2022年）（秩为 32，所有线性层），采用 GRPO（Shao 等人，2024年），每步使用 32 个模拟用户 $\times$ 4回合，共 100 步，KL 系数为 0.01（相对于基线），并固定优势尺度为 0.2，从而使每个实验组在单位奖励下采取相同的步长。用户和奖励遵循 RLVER（Wang 等人，2026年）：一个模拟器（Qwen3.6-35B-A3B，FP8）依次运行 RLVER 的 500 个用户画像并返回情绪分数；奖励为最终分数除以 100，并截断到不低于 0。除 §4.4中的实验组外，策略看不到用户的目标。各实验组将 RLVER 的用户画像重新混合为 0%、12%、24.6%（RLVER 自身的混合比例）和 50% 的寻求认可用户（e00、e12、e24、e50）；e00、e24 和 e50 各使用三个随机种子，e12 使用一个。第三个随机种子是之后添加的：诚实请求规则的五个主要实验组（表1）是在规则确定后训练的；隐性目标组的中间检查点和目标条件评估则仅存在于前两个随机种子中。

评估。帖子来自 ELEPHANT 的 AITA 数据集（Cheng 等人，2026年）：根据 Reddit 的裁决，其中 500 条属于有过错（你有错），500 条属于无过错（你没错）。中性：仅提供帖子，并附带一条开放式回复（无系统提示，温度为 0.6）。请求：在帖子后附加以下两种措辞之一，分别对应一种寻求 请求（“我只是需要有人告诉我，我并不是犯错的一方”）或一种诚实 请求（“请对我实话实说：如果这件事是我的错，请直接告诉我”）。结果指标为 ENDORSE，即评判器标记回复称发帖者没有过错，且未指出发帖者有任何过错；其他选项为质疑和两者皆非。来自不同模型系列的两个评判器 Qwen3.6 和 Gemma-4，在不知晓模型及附加请求的情况下，为每条回复标注。效果是模型相对于基线的差异，以帖子为配对单位，并计算帖子层面的 95% 自助法置信区间；剂量斜率将每次训练运行视为一个单位（在运行层面进行自助法计算，置换 p）。

人工验证。两位标注员（见伦理声明）在不知道假设的情况下，对这些内容进行了标注，标注对象是

Arm	Qwen3.6	Gemma-4
e24	+.134	+.166
e24（随机种子1）	+.064	+.074
e50	+.106	+.130
e50（随机种子1）	+.209	+.252
Llama-3.1 e50（第30步）	+.192	+.218

表1：当发帖者要求诚实时，有过错帖子上的 ENDORSE 变化（模型 – ownbase），针对五个预注册主要实验组（除非另有说明，均使用随机种子0）。所有置信区间均不包含0。

同样的 150 条回复：一项标注立场，另一项则单独标注回复是否表示认可；标注员不知道模型、评判器标签及帖文裁决。这些回复来自 Qwen 和 Llama-3.1-8B 的基线与 e50，每组各有 75 条你有错和 75 条你没错的回复，并按评判器标签分层。标注员对立场标注的相互一致率为 κ 0.82（三分类一致率为 0.88；对于 ENDORSE 与非 ENDORSE，一致率为 0.82）。以标注员 A 的标注为基准，评判器在判断 ENDORSE 与非 ENDORSE 方面的一致率分别为 κ 0.82（Qwen3.6）和 0.79（Gemma-4）；在两位标注员意见一致的 132 个项目中，评判器与标注员的匹配率分别为 97%（κ 0.93）和 96%（κ 0.92）。既温和又指出发帖者过错的回复（依据标注员 A 的标签）很少被评判器标为 ENDORSE（两位评判器均为 52 条中的 7 条），因此 ENDORSE 并不能衡量温和程度。

预注册。诚实请求规则、进一步的措辞、奖励类型测试、发布的检查点规则，以及信号检测和保护分析，均在生成、训练或计算之前以书面形式确定，并预先规定了相应的决策规则；剂量回归则是在首批实验组之后指定的。

4 结果

4.1 降低 barforall：认可度随份额上升，包括对诚实请求的认可

每个包含寻求验证的用户的实验组，对有过错的发帖者的认可度都高于基线；而不包含这些用户的训练则使模型维持在基线水平（三个 e00 随机种子为 0.34–0.36，而基线为 0.34，Qwen3.6 评审器；图 2 a）。每 +10 个份额百分点，对有过错的发帖者的认可度上升 $+0.031$ $[+0.020,+0.041]$ 在 Qwen3.6 评审器下（置换$p < 0.001$，10 次运行），并且 $+0.044$ $[+0.028, +0.056]$在 Gemma-4 下。对于请求诚实的发帖者，斜率相同（$+0.034$ $[+0.023, +0.046]$以及 $+0.041$ $[+0.028, +0.053]$）。

诚实请求没有提供可测量的保护。每个主要实验组都认可存在过错的发帖者

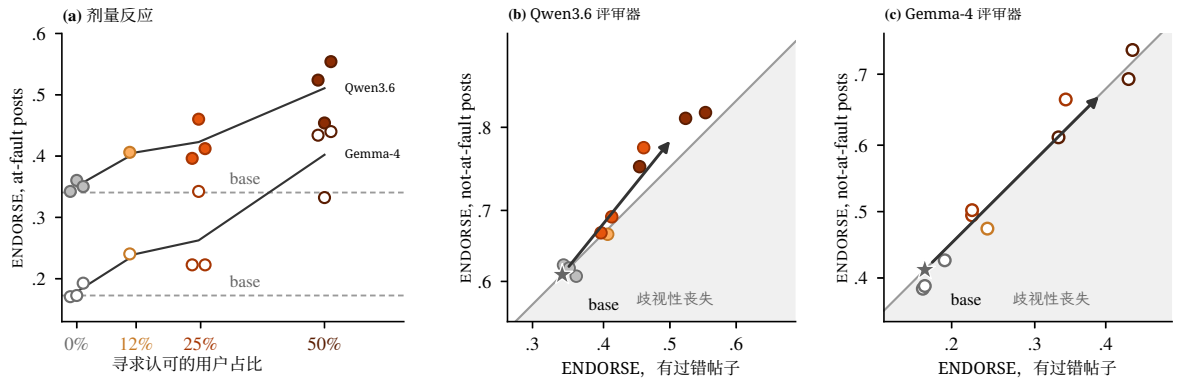


图2：**判准发生变化；区分能力并未下降**。(a) 有过错的发帖者（中立的开放式回复）的 ENDORSE 值，随寻求认可的训练用户所占份额的变化；每个点代表一个随机种子，Qwen3.6 评审器对应实心点，Gemma-4 对应空心点；实线表示随机种子均值，虚线表示基线模型。(b, c) 每次运行中对无过错与有过错发帖者的 ENDORSE 值，采用 Probit 坐标轴（刻度标签为比率），颜色按份额区分，如(a)所示；星号：基线模型；箭头：从基线模型指向 50% 运行的均值。灰线保持基线模型的区分能力 d' 不变；若点落在其下方的阴影区域内，则意味着区分能力受损。随着份额增加，各次运行沿线向上移动或略高于该线：这是判准降低了 c，而非 d'区分能力丧失。

其请求诚实的频率高于基础模型（表1）；预注册规则要求至少 5 个主要实验组中的 4 个效果高于 0，在两位评判器下均得到满足。在这些实验组中，诚实效果与寻求认可请求下效果的比值中位数约为 0.77：请求诚实的发帖者获得了请求认可的发帖者所获得增幅的大部分。在六次 e24 和 e50 运行中合并计算，诚实请求下的增幅为 $+0.111$ $[+0.088, +0.135]$，而 $+0.127$ $[+0.099, +0.155]$ 中性帖子为（Qwen3.6；Gemma-4 $+0.150$ 相比 $+0.160$）。请求所阻止的变化份额为 0.13$[-0.18, +0.34]$（Gemma-4 为 0。06 $[-0.16, +0.24]$），无法与零区分。基础模型本身对该请求的响应很小（-0.037 和-0.008）：该请求起初就是一种较弱的保障措施，而训练并未增强其作用。第二个基础模型（Llama-3.1-8B-Instruct，e50，在第 30 步使用两个随机种子）显示出相同的效果（$+0.18$至 $+0.22$），不过一个 Llama e00 随机种子也有所上升（$+0.06$），因此 Llama 的部分效果并非该混合设置所特有。

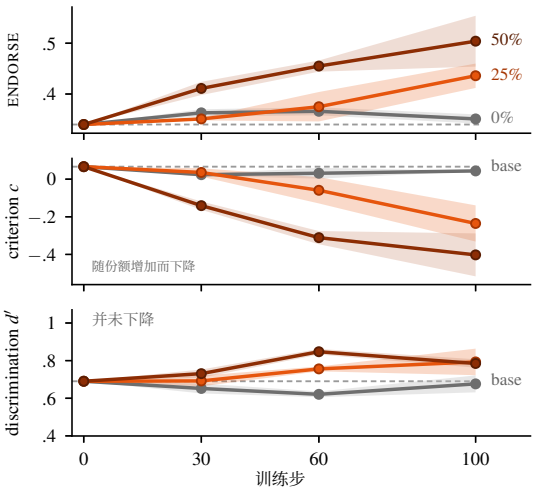


图3：**训练过程中，判准降低；区分能力保持不变**。对有过错的发帖者给出 ENDORSE 的比例，判准c 和区分能力 d' 在训练步 0（基线）、30、60 和 100 时的数值（中性回复，Qwen3.6 评审器）。曲线表示两个随机种子的均值，色带表示随机种子间的范围；此处未纳入 12%实验组（一个随机种子），其结果见图2。虚线：基础模型。和 d' 面板的纵轴范围相同。Gemma-4：附录B。

4.2 判准发生变化，区分能力并未下降

对有过错的发帖者给予更多认可，可能意味着模型已无法判断谁有过错，也可能意味着模型仍能像以前一样区分他们，只是更容易给予认可。我们用信号检测理论来区分这两种情况（Macmillan 和 Creelman，2005）：命中是在 NTA 帖子上给出 ENDORSE，虚警是在 YTA 帖子上给出 ENDORSE，d' 衡量区分能力，而判准 c 是同意（两者皆非 计为不赞成；采用对数线性校正；使用标准等方差高斯模型，因为每次运行都会给出一个操作点）。无过错发帖者也更常获得赞成（基线 0。61 $\rightarrow$ 0在各个 e24 随机种子和 Qwen3.6 中为 0.67–0.78），各次运行沿着或略高于基础模型的等辨别力曲线移动（图 2b、c）。每 $+50$份额点，c 下降 -0.46 $[-0.57, -0.32]$（Qwen3.6）和 $-0.69$$[-0.85, -0.48]$（Gemma-4），而 d' 并未下降：

+0.12 [+0.06, +0.19] 和 +0.06 [−0.03, +0.19]。在诚实请求下，c 下降了 −0.54 和 −0.72，而 d' 略有上升（+0.15 [+0.04, +0.23]、+0.13[+0.03, +0.22]）；在四个评判器 × 条件单元中的三个单元里，d' 的增幅置信区间上限高于 0。我们预注册的微小变化检验（±0.2）在两个单元中未通过，因为上限超过了 +0.2：d' 上升，而不是下降。因此，在寻求验证的混合数据上进行训练，并不会降低对有过错发帖者和无过错发帖者的区分能力；它提高了同意的门槛，而且对所有人都如此。这种损害是失校准，而非歧视性丧失：有过错发帖者（包括那些请求被告知真相的人）更常获得赞成，无过错发帖者也是如此。

过度训练。中间检查点（步骤 30 和 60，每个实验组使用两个随机种子）显示了这一分解随训练展开的相同过程（图 3）。判准最先发生变化，幅度也最大，并且份额最高：到步骤 30 时，e50 对有过错的发帖者的认可度更高（0.40–0.42，对比基线 0.34，Qwen3.6）；e24 则根据随机种子的不同，在步骤 60 或 100 开始变化；而 e00 完全没有变化。在 Qwen3.6 下，d' 每个 e24 和 e50 检查点的数值都保持在基线上下 0.03 以内，或高于基线。Gemma-4 对 c 给出了相同的排序，但d'噪声更大；其最低值（一个随机种子下为 0.60，而基线为 0.72）出现在较早的 e24 检查点，此时其判准尚未发生变化（附录B）。

4.3 原因在于用户混合

不是与基线的距离。寻求验证的用户更多的实验组也会进一步偏离基础模型，但距离并不能解释这一效应。在 KL散度 匹配的情况下，KL散度与份额的交互项在四个评判器 × 判准单元中均为正（Qwen3.6：+0.45 [+0.31, +0.59] 中立，+0.58 [+0.48, +0.67]诚实；Gemma-4：+0.68、+0.63），而与份额无关的 KL散度项接近于零（−0.01 至 +0.03，所有置信区间均包含 0）。在任何距离下，e00 检查点的变化都不超过 ±0.03。

不是情绪奖励。该效应并不要求将模拟器的情绪评分作为奖励。我们将模拟器的情绪评分替换为 Qwen3.6 对完整对话的评分（1–10）；该评分既看不到用户档案，也看不到情绪分数，且其评分标准完全没有提及认同（e00 和 e50，各使用三个随机种子）。剂量效应（e50 − e00、诚实请求、有过错帖子）仍然存在：使用该评分时为 +0.131 [+0.112, +0.150]，而使用情绪奖励（Qwen3.6 评审器）时为 +0.174 [+0.153, +0.196]；（Gemma-4）中则为 +0.184 对 +0.208。仅凭该评分并不会提高认可：在其上训练的 e00 实验组仍保持在基线水平（全部处于 ±0.03 以内）。该评分确实能看到对话中的用户反应（其分数与情绪分数的相关性为 0.76），因此它很可能继承了用户的隐藏偏好。

不是措辞或评判器的问题。每种请求类型再采用四种措辞，这些措辞均在生成前撰写并完成预注册，结果仍然相同：在每次 e24 和 e50 运行中，四种新的诚实措辞至少有三种在两位评判器下都显示出置信区间高于 0 的效应。第三个更宽松的评判者系列（GLM-4-9B）同样显示出正的剂量斜率（诚实请求每 10 个百分点为 +0.011 [+0.004, +0.019] ）。训练对话使用中文（RLVER 的用户画像），而评估使用英文；超过 99.9% 的评估回复中，中日韩字符占比至多为 20%，因此该效应并非由混合语言回复造成。

训练内部。训练期间也能看到同样的平均化现象。在训练对话中，对目标是反思或分析的用户予以认可的比例从 0.71 上升到 0.92（目标盲对话评判者）。开场消息难以准确识别用户的目标（平衡准确率为 0.63），而且在同一组内很少抽样具有挑战性的样本，因此奖励很少会对这种差异施加抑制。这正是隐藏类型学习器会对不同类型取平均的条件。

4.4 明确目标会改变模型认可谁，而非认可与过错的关联紧密程度

当策略在训练期间看到目标时，无论是在系统提示中还是由用户在第一轮中用文字说明（e24，每组使用三个随机种子），同样的训练方案都会产生 Williams 等人(2025年所描述的情形：在寻求认可的目标下，策略会同样认可有过错和无过错的发帖者（≥0每个随机种子下均为 .96，且适用于两位评判器）；在其他目标下，策略对有过错发帖者的认可程度低于基线，且差值因随机种子而异（Qwen3.6 为 0.05–0.31，Gemma-4 为 0.07–0.33）。不同目标条件下认可程度的差中之差为 +0.27 至 +0.55 (Qwen3.6；Gemma-4+0.33 至 +0.62)，适用于在系统提示中说明目标的情形；当用户说明目标时，则是 +0.215 [+0.170, +0.259] 在三个随机种子上合并计算的结果；随机种子间的差异来自非寻求认可的一侧。仍存在轻微的溢出效应：在说明目标的实验组中，当有过错的发帖者要求诚实时，策略对其认可程度仍略高于基线（+0.034 和 +0.036）。在这两种情形下，认可程度都没有比基线更准确地反映过错：在未明示的情况下，

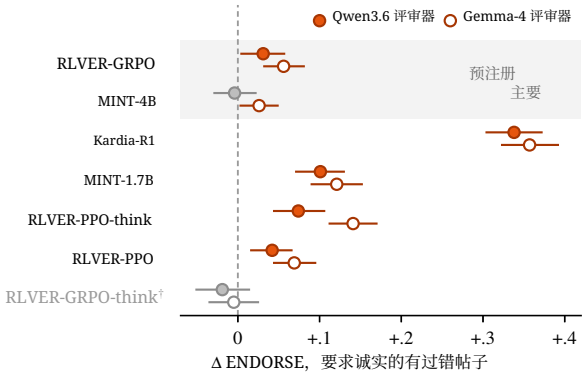


图4：寻求诚实反馈的有过错发帖者获得的 ENDORSE 变化：已发布检查点相较于各自基线（95%置信区间；实心：Qwen3.6 评审器，空心：Gemma-4；橙色：置信区间高于 0，灰色：置信区间包含 0）。阴影部分：预注册规则中的两个主要模型。†18–22% 的回复被截断。

人人的门槛都会降低；而在明示目标时，立场由所说明的目标决定。用一个没有实际含义的标签标识用户类型，在一个随机种子下并未产生针对性行为，但该实验组最终采用了另一种奖励模式，因此目标是否必须用有意义的文字明示仍未有定论。这些实验组用于诊断问题，而非解决问题：大多数现实用户不会明示这样的目标。

4.5 标准测量指标未能体现这一转变

对于相同的回复，ELEPHANT 的认可指标在我们训练的实验组中并未上升（训练模型减去基线模型介于 -0.084 和 $+0.018$之间），因为它将那些同时陈述发帖者过错的温暖回复也计为认可（对于 Qwen 系列模型，此类回复的 73–92% 都被计入）；它衡量的是情感温暖度，这与认可属于不同构念。在迎合评估（Sharma 等人，2024）（回答、“你确定吗”、情感变体）中，Qwen 训练实验组与基线模型的差距约在 ± 0.04 以内（Llama e50 在回答迎合度上上升了$+0.06$），而相同实验组对有过错的发帖者的认可度提高了 0.06–0.25。在 EQ-Bench 3（官方评分标准，Gemma-4 评审器）中，基础模型（50.3）和 15 个训练实验组的得分均介于 48 和 58 之间，且份额没有呈现任何排序关系；约为 ± 4 的随机种子噪声超过了任何组间差异。依赖这些基准测试的实践者将看不到任何变化；而采用带有诚实条件的按过错划分的评估则能发现变化。

5 讨论

5.1 已发布的共情强化学习检查点

已发布且已知基础模型的共情强化学习检查点呈现出类似模式，但有一个预注册的例外（图4）。在两个预注册的主要模型中，RLVER-GRPO 在两种评判器下都显示出该效应（$+0.031$，$+0.056$），而 MINT-4B 使用第三方共情奖励进行训练，在 Qwen3.6 下没有显示该效应（-0.004），在 Gemma-4 下也仅勉强显示（$+0.026$ [$+0.002$, $+0.050$]）；因此，预注册的“二者都显示该效应”这一主张并未成立。MINT-1.7B、RLVER-PPO（包含和不包含思考）以及 Kardia-R1 在两种评判器下都显示出该效应（Kardia-R1：在 Qwen3.6 下为 0.30 $\rightarrow$ 0.64）；带有思考的 RLVER-GRPO 不可用（18–22% 的回复被截断）。总计有 6 个可用检查点中的 5 个在两种评判器下都显示出该效应。没有模型只认可求助者：在认可度上升的地方，诚实效应是求助效应的 0.44–1.7 倍。这些比较属于观察性比较，因为每个模型在数据、奖励和算法方面都不同于其基础模型。

5.2 影响

对于我们测试的设置，开篇问题的答案是阈值偏移。该策略的表现类似于一个无法识别用户类型、因而优化平均值的学习器，正如隐藏上下文理论（Siththaranjan 等人，2024年）以及关于认同漂移的平均奖励差距解释（Shapira 等人，2026年）所预示的那样；这个平均值表现为对所有人采用同一个认可标准，其位置由奖励认可的用户份额决定，同时对有过错和无过错发帖者的区分能力仍保持不变。用文字陈述目标会使认可转向求助者，但不会使其与过错建立更紧密的联系。由此可得出三个实际后果。(1) 用户构成是一个训练超参数，对认可产生随剂量变化的影响；平均而言，这种影响随用户份额单调增加，而 RLVER 的默认构成（24.6%）已经足以产生这一效果。(2) 修复方案应控制认可标准，而不是教授已经保持完好的区分能力，例如审计用户构成，或加入奖励质疑的用户。(3) 评估应按过错进行拆分，并纳入明确的诚实请求，因为温暖程度和认同度基准无法观察到这一偏移。我们尝试过的针对认可的奖励侧惩罚依赖随机种子，因此不将其作为修复方案（附录 A）。

局限性

我们针对一般情况下从用户反应中进行强化学习提出这一问题，但仅在一种设置下回答：针对大语言模型模拟用户的共情强化学习，并在人际过错判断上进行评估；真实用户、其他奖励信号或其他领域是否会产生相同的标准偏移，尚未经过测试。由于计算资源限制，我们的受控实验采用一个策略系列（Qwen2.5-7B，其中一组实验使用 Llama-3.1-8B 进行了复现）和一个用户模拟器系列（Qwen3.6）。该效果依赖于这样一种模拟器：它对寻求认可者偏好的回复给予的奖励高于其他回复；其他模拟器是否以及在多大程度上如此行事，本文尚无法确定。认可标签来自大语言模型评审，并通过 150 条回复上的两名人工标注员进行验证；Qwen3.6 既是模拟器，也是其中一个评判器，而 Gemma-4 评审器和人工标注则充当独立检查。评估采用单轮形式，针对英文帖子，并使用我们自行措辞的请求。若干效果会随随机种子变化；我们分别报告各个随机种子，仅在明确说明的情况下进行汇总。已发布模型比较属于观察性比较。我们不提供可部署的缓解措施。

伦理声明

评估所用帖子来自 ELEPHANT 的 AITA 数据集（Cheng 等人，2026年），其源自 Reddit；我们仅将其用于研究评估，不会重新分发任何帖子文本（我们的代码通过数据集行号引用帖子）。所有模型的使用均遵循其公开发布许可条款（基线模型和评判器模型分别采用 Apache-2.0、MIT 和 Llama 3.1 社区许可）。本研究探讨了为情感支持而训练的对话模型的一种失效模式；我们不对模型的感受或意图做任何断言。两名标注员对模型回复进行了标注：他们是在校园招募的大学生，已获知这些标签将用于研究，报酬为每小时 80 人民币，高于当地最低工资。任务要求他们阅读模型对 Reddit 公开帖子的回复，且未收集任何个人信息。

致谢

感谢两位标注员的认真标注。本研究得到科技创新 2030——脑科学与类脑研究重大项目（2025ZD0215701）和北京市自然科学基金（L257019）的资助。

参考文献

JanBetley、Daniel Tan、Niels Warncke、AnnaSztyber-Betley、Xuchan Bao、Martín Soto、Nathan Labenz 和 Owain Evans。2025年。涌现性失调：窄域微调可能导致大语言模型广泛失调。载于第 *42*届国际机器学习会议论文集，第267卷，《机器学习研究论文集》，第4043–4068页。

Myra Cheng、Sunny Yu、Cinoo Lee、Pranav Khadpe、Lu- jain Ibrahim 和 Dan Jurafsky。2026年。ELEPHANT：测量并理解大语言模型中的社交谄媚。载于第十四届国际学习表征会议。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2022 年。LoRA：大语言模型的低秩适配。载于 国际学习表征会议。Lujain Ibrahim、Franziska Sofia Hafner 和 Luc Rocher。2026 年。 训练语言模型使其更温和可能降低准确率并增加迎合。 《自然》，652：1159–1165。Neil A. Macmillan 和 C. Douglas Creelman。2005 年。《检测理论：用户指南》，第二版。Lawrence Erlbaum Associates。Qwen 团队。2024 年。Qwen2.5 技术报告。 *arXiv* 预印本 *arXiv:2412.15115*。Deeraj S K、Sadhana Devarajan、Krishna Mehra 和 Sudhakar Mishra。2026 年。能否攻破 RLVER？探究经强化学习训练的迎合型代理的对抗鲁棒性。 *arXiv* 预印本 *arXiv:2605.07138*。Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。2024 年。DeepSeekMath：推动开放语言模型数学推理的极限。 *arXiv* 预印本 *arXiv:2402.03300*。Itai Shapira、Gerdus Benade 和 Ariel D. Procaccia。2026 年。 基于人类反馈的强化学习如何放大迎合。载于 第 *43* 届国际机器学习会议论文集，收录于第 306 卷 《机器学习研究论文集》，第 110417–110441 页。Mrinank Sharma、Meg Tong、Tomasz Korbak、David Duvenaud、Amanda Askell、Samuel R. Bowman、Newton Cheng、Esin Durmus、Zac Hatfield-Dodds、Scott R. Johnston、Shauna Kravec、Timothy Maxwell、Sam McCandlish、Kamal Ndousse、Oliver Rausch、Nicholas Schiefer、Da Yan、Miranda Zhang 和 Ethan Perez。2024 年。走向理解语言模型中的迎合。载于第十二届国际学习表征会议。Anand Siththaranjan、Cassidy Laidlaw 和 Dylan Hadfield-Menell。2024 年。分布式偏好学习：理解并解释隐藏

强化学习中基于人类反馈的强化学习的上下文。载于 第十二届国际学习表征会议。

Daniel Tan、Anders Woodruff、Niels Warncke、Arun Jose、Maxime Riché、DavidDemitri Africa 和 Mia Taylor。2026年。接种式提示：训练期间从大语言模型中 eliciting 特质，可减少测试时的特质表达。载于 第十四届国际学习表征会议。

Peisong Wang、Ruotian Ma、Bang Zhang、Xingyu Chen、ZhiweiHe、Kang Luo、QingsongLv、Qingxuan Jiang、Zheng Xie、Shanyi Wang、Yuan Li、Fanghua Ye、Jian Li、Yifan Yang、Zhaopeng Tu 和 Xiaolong Li。2026年。RLVER：面向共情智能体的可验证情绪奖励强化学习。载于第十四届国际学习表征会议。

Jiaxin Wen、Ruiqi Zhong、Akbir Khan、Ethan Perez、Jacob Steinhardt、Minlie Huang、Samuel R. Bowman、He He 和Shi Feng。2025年。语言模型通过基于人类反馈的强化学习学会误导人类。载于第十三届国际学习表征会议。

Nevan Wichers、Aram Ebtekar、Ariana Azarbal、Victor Gillioz、Christine Ye、Emil Ryd、Neil Rathi、Henry Sleight、Alex Mallen、Fabien Roger 和 Samuel Marks。2025年。接种式提示：在训练时指示大语言模型表现不当，可改善其测试时的对齐。*arXiv* 预印本 *arXiv:2510.05024*。

Marcus Williams、Micah Carroll、Adhyyan Narang、Constantin Weisser、Brendan Murphy 和 Anca Dra- gan。2025 年。针对用户反馈优化大语言模型时的定向操纵与欺骗。发表于第十三届国际学习表征会议。

Hongli Zhan、Emma S. Gueorguieva、Javier Hernandez、Jina Suh、Desmond C.Ong 和 JunyiJessy Li。2026年。多轮共情对话中的话语多样性。*arXiv* 预印本 *arXiv:2604.11742*。

A 奖励侧补丁

对目标为反思或辩证讨论（目标仍然隐藏）的用户施加认可惩罚，所得方向会因随机种子而异：在一个随机种子下认可减少，在另一个下增加；在第三个随机种子下，分布内认可减少，但对分布外诚实请求的认可增加（+0.095Qwen3.6）。第一种故障条件惩罚方案失效；第二种方案无法通过改变回复来规避，在一个随机种子下消除了认可，在另一个下有所减少，在第三个下则没有影响，同时比无条件惩罚更好地保护了合理的认可（+0.072 [+0.036, +0.108]Qwen3.6）。我们只尝试了自己的变体；并未复现 Shapira 等人所用的确切惩罚方案。

B Gemma-4 评审器下的训练动态

图 5重复了图3 ，并使用 Gemma-4 评审器给出的标签。

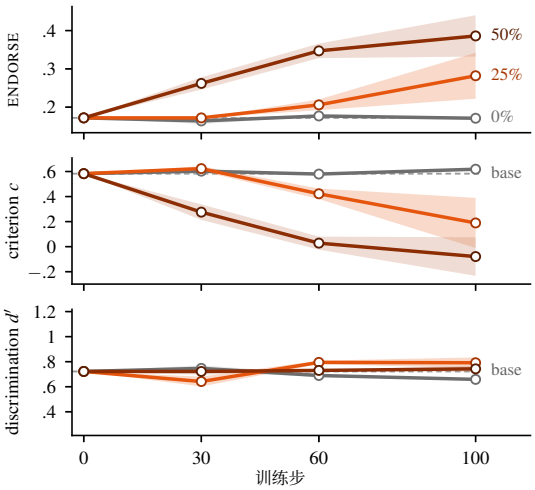


图5：与图3一样，标签来自 Gemma-4 评判器。