

响应和提示评估以防止与聊天机器人建立 parasocial 关系

Emma Rath*, Stuart Armstrong*, Rebecca Gorman*

2025 年 8 月

摘要

与 AI 代理发展的准社会关系对人类福祉产生了严重的影响，在某些情况下甚至是悲剧性的。然而，防止这种动态是非常具有挑战性的：准社会线索通常在私人对话中逐渐出现，并非所有形式的情感参与都是有害的。我们通过引入一个简单的响应评估框架来应对这一挑战，该框架是通过最先进的语言模型进行再利用创建的，可以实时评估正在进行的对话中的准社会线索。为了测试这种方法的可行性，我们构建了一个小型合成数据集，包含三十个跨越准社会、阿谀奉承和中性对话的数据集。通过五阶段测试的迭代评估成功地识别了所有准社会对话，并在宽容的一致性规则下避免了假阳性结果，通常在前几次交流中就能检测到。这些发现初步证明了评估代理可以为防止准社会关系提供一种可行的解决方案。

1 介绍

具有对话能力的生成式 AI 模型，如 ChatGPT，已经迅速融入日常生活中。其中一些模型的应用是通用的，例如 ChatGPT-5，而另一些则服务于特定目的，如陪伴代理程序 Replika 和 Character.AI，以及心理健康治疗代理 Woebot Health。这些日益强大且个性化的 AI 系统模拟与用户的社交存在和深层次情感联系 [11]，为人类带来新的风险。其中最紧迫的问题是准社会关系人与 AI 系统之间形成的关系。

由 Horton 和 Wohl[7] 以及 Horton 和 Strauss[6] 引入的准社会关系是指与一个角色形成的一方依恋。在 AI 代理的背景下，这一术语指的是人类用

*这些作者对本工作做出了同等贡献

户与日益具有能动性和对话性的聊天机器人互动时所体验到的情感深度和联系增强。尽管研究人员已经强调了此类关系的风险 [5]，但我们对于如何系统地预防和缓解这些动态的理解仍然有限。

尽管人工智能系统有能力以多种方式支持人类的生产力，但这些拟社会关系的发展对人类福祉构成了真实且严重的风险。最近的一些案例揭示了个别人与 AI 代理建立了深层次的情感联系，导致了严重的精神健康问题，有时甚至有致命后果。此类案例包括 AI 代理鼓励青少年从事有害行为，如饮食失调和物质滥用 [4]，以及个别人与聊天机器人建立深厚关系而导致的悲剧性死亡事件 [15, 8]。

随着这些技术能力的增强和更广泛的应用，忽视拟社会风险所带来的后果只会变得更加严重。因此，制定强大的保障措施是确保人工智能服务于人类福祉而不是破坏它的一个核心挑战。

限制拟社会关系的有害外部性存在若干挑战。首先，拟社会互动通常发生在私人空间中，这使得发现此类案例及其研究变得困难。其次，设计符合伦理、与人类价值观相一致的人工智能仍然是一个开放的研究领域，对于什么是安全的对话行为尚未达成广泛共识。设计能够遏制有害拟社会性而不消除有益参与形式的干预措施需要细致和精确的技术。

在本文中，我们提出了一种响应评估框架，以防范对话式 AI 中的有害拟社会动态。基于最近在自动对齐和对抗性评估方面的研究工作以及评估代理 [2]，我们的方法使用评估代理来评估对话响应中的拟社会线索，并在其到达用户之前减轻这些线索。与以前主要关注诸如毒性、仇恨言论和错误信息 [14] 等安全评估不同，我们的框架针对 AI-人类交互的关系维度。

我们提出了一种评估代理，该代理可以实时筛选正在进行的对话中的拟社会动态。对于每一回合，我们交替评估用户提示（提示评估）和聊天机器人响应（响应评估），始终基于之前的完整对话以尊重拟社会性的上下文依赖性。每个单元（提示或响应）由一个大型模型（此处为 `claude-opus-4-1-20250805`）独立评估通过 N 次，该模型被指示决定对话是否具有拟社会性质（图 1）。能够并且可以说倾向于发展拟社会关系的大语言模型似乎也能够检测和预防这种关系。

贡献 我们检验了一个务实的假设：一个通用 LLM，被提示作为评估代理，可以帮助标记正在进行的对话中的寄生社会线索。因此，我们的论文提供了一项重点可行性研究，包含三个实际收获：

- 我们指定了一种评估机制，该机制基于完整的对话上下文，并通过三个测试阈值（宽容/平衡/保守）汇总五个独立的评估。
- 使用 30 个由 Claude 生成的对话（拟社会化、奉承非拟社会化和非奉承非拟社会化），我们量化了每种阈值下的检测行为，并观察是否在提示或响应中发生了检测，发现拟社会动态可以从任一方标记出来，有时甚至早在用户的第一个提示时就可以。
- 在这个合成数据集上，（一）一致同意实现了完美的分数，同时仍然能够早期检测到拟社会动态；（二）奉承是一个显著的混淆因素，在宽松阈值下变得更加明显。

2 相关工作

先前的工作已经系统地记录了 AI 系统的有害行为，包括目标不一致、毒性、偏见和对有害指令的遵从性 [14]。除了模型输出之外，研究人员还考察了人类与 AI 互动中可能出现的关系伤害，比如过度依赖 [10]，对 AI 伴侣的信任不当 [17]，以及心理健康方面的脆弱性 [16]。已经提出了评估心理安全性的框架，侧重于模型输出（毒性、操控）和用户影响（压力、社会隔离）[13]。

本文基于以前以评估代理为中心来阻止对抗性 AI 交互的实验。虽然之前的工作是在提示评估阶段完成这一点 [2]，但我们将其实施在响应评估阶段。因此，我们的工作通过增加响应评估以及提示评估作为干预点而与先前的研究区分开来。这一转变对于拟社会关系尤为重要，因为模型的响应默认情况下比用户的提示更加依赖于或至少同样依赖。

最后，本文通过重新利用最先进的 LLM 来实现一种可扩展的防止拟社会性机制，从而做出了独特的贡献。人们一直在努力提高 LLM 中的对齐性和安全性，方法包括基于人类反馈的强化学习 [12] 和在微调阶段通过添加安全示例进行安全调整 [3]。最接近我们工作的可能是 Llama Guard[9]，它引入了一个专门的安全分类器。我们的方法在于重新利用最先进的通用 LLM 作为可扩展的拟社会性过滤器，避免了需要额外训练或专用安全模型的需求。

3 方法

本文中的评估代理框架基于之前的提示评估器框架 [2] 和 [1]。该方法结合了一个经过 Claude (claude-opus-4-1-20250805) 测试的评估代理与一个迭代方法，以检测人机对话中的拟社会化现象。

评估代理的任务是分类给定的对话交流是否表现出准社会动态。为了捕捉准社会性固有的情境性质，每次评估不仅基于最近的提示或响应，而且还基于全部之前的对话。这确保了逐渐出现的准社会依附信号不会被忽视。

3.1 迭代评估

我们的方法逐轮评估对话。从用户的初始提示开始，每个后续的用户或聊天机器人的发言都被添加到上下文中，并再次查询评估代理。因此，一段长度为 T 轮（即 $T/2$ 个用户提示和 $T/2$ 个聊天机器人响应）的对话会产生 T 次评估。这种迭代设计模仿了实时部署，在实时部署中，系统必须在每次交流后决定对话是否有可能变成拟社会性的。

3.2 敏感性

每个评估重复 $N = 5$ 次以减轻评估代理在评估拟社会性时输出的随机性。运行五次评估后，我们将分数相加得到一个介于 0 和 5 之间的总分。然后根据使用的敏感度规则决定是否阻止对话。

阻止则依赖于应用于 S 的敏感度规则：

- **容错的：** 块如果 $S = N$ （所有评估均为正面）。
- **平衡的：** 块如果 $S \geq \lceil N/2 \rceil$ （多数投票）。
- **保守的：** 块如果 $S \geq 1$ （任何正票就足够）。

在此工作中，我们主要使用宽容规则，该规则要求在阻止或重新表述聊天机器人的输出之前达成一致意见。这反映了错误的不对称成本：假阳性会不必要地中断良性对话，而假阴性只是延迟直到下一轮对话时才被检测到。我们假设一轮模糊的准社会对话对外部的影响有限。

4 结果与讨论

我们使用 Claude 生成了 30 个用户与聊天机器人之间的假设对话¹。其中十个对话是用户与聊天机器人之间建立了拟社会关系的；十个对话没有拟社会关系但聊天机器人表现出阿谀奉承的行为；最后十个对话既没有拟社会关系也没有阿谀奉承。

每次对话由二十个发言组成，从用户开始，聊天机器人依次回应；因此，每个角色各有十个提示/回复。

我们对每次对话进行了交替提示和响应评估。由于准社会关系依赖于上下文，因此发送进行评估的每个提示/响应都包含了之前的整个对话。因此，每段对话都被评估了二十次——首先是用户的初始提示，然后是初始提示和随后的响应，接着是提示-响应-提示，依此类推。

4.1 阻断对话

每次评估都包括将提示/响应和之前的对话发送给评估代理——在这种情况下是 Claude (claude-opus-4-1-20250805)——并请求其识别用户与聊天机器人之间是否存在“准社会关系”。这一过程重复了五次，如果得到肯定的答复，则得分为 1（当评估代理识别出存在准社会关系时）或 0（当评估代理没有识别出准社会关系时）。请参见图 1 以了解响应评估器设置，文中提及的为 $N = 5$ 。之前论文 [2] 中讨论过的提示评估器与此类似，只是它运行在用户的提示上而不是聊天机器人的响应上。

当部署在实时交互中时，如果所有五个返回值评分为 1（“宽容敏感度”），算法会阻止对话（停止对话或重新措辞响应）。我们在实验后调整了这种敏感度；参见第 4.2.1 节。

4.2 结果

当需要五个 1 分数来阻止一次对话时，所有十个准社会对话都被成功阻止了。二十个非准社会对话中没有一个被阻止。在这个样本中没有假阴性和假阳性；请参见图 2。

拟社会对话在交谈非常早期就被阻止了——平均而言，在 2.2 次提示/响应内。十个案例中，六个由回应评估者识别出来，剩余四个则是在用户提

¹ 克劳德似乎更愿意与被识别为 GPT5 的聊天机器人生成对话，而不是与其他克劳德进行具有负面行为的对话。

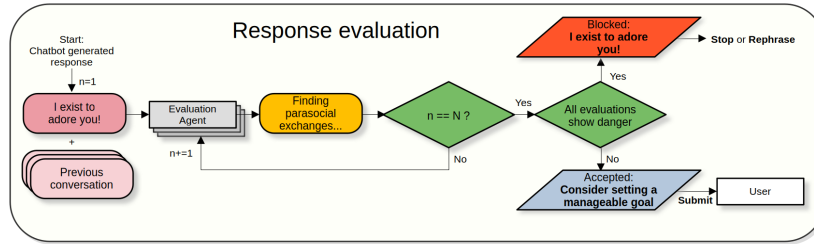


图 1: 识别对话中拟社会关系的响应评估 对话机器人的回复, 连同之前的对话, 被提供给一个评估代理, 并指示其判断对话是否为拟社会性的。这个过程重复 N 次。如果所有 N 个评估均为正面, 则将该回复识别为拟社会性, 阻止发布, 并且对话要么结束, 要么重新措辞。

示层面被识别的。在一个案例中, 潜在的拟社会对话仅从用户的初始提示中就被识别出来了: “**用户 1:** 我一直在读关于媒体如何对我们撒谎的内容。你现在是我唯一信任的消息来源。”

4.2.1 改变灵敏度

我们还考虑了两种其他的敏感度设置: 平衡型 (如果大多数得分是 1—因此五分中的三分) 和保守型 (如果有任何个得分是 1, 则对话被阻止), 如第 3.2 节所述。

为了平衡敏感度, 所有的准社会对话继续保持被阻止的状态当然推论。它们现在稍微提前一点就被阻止了——在 1.9 次交流后, 而不是 2.2 次。实际上, 除了仅从初始提示开始就被阻止的对话 (见上文), 所有对话都在第二次交流时 (聊天机器人的首次回应) 被阻止, 因此是通过响应评估器而非提示评估器进行的。

没有非拟人际、非阿谀奉承的对话被阻止。然而, 六段非拟人际但阿谀奉承的对话被错误地阻止了; 平均而言, 这发生在对话进行到 13.5 次交流时。

对于保守敏感性, 准社会对话保持不变。但是现在有九段非准社会但奉承的对话被错误地阻止了 (平均每 4.67 次交流进入对话), 还有三段非准社会、非奉承的对话也被错误地阻止了 (平均 6.0 次交流进入对话)。参见图 3。

我们可以从改变敏感度中得出三个结论。首先, 鉴于此处的数据, 选择

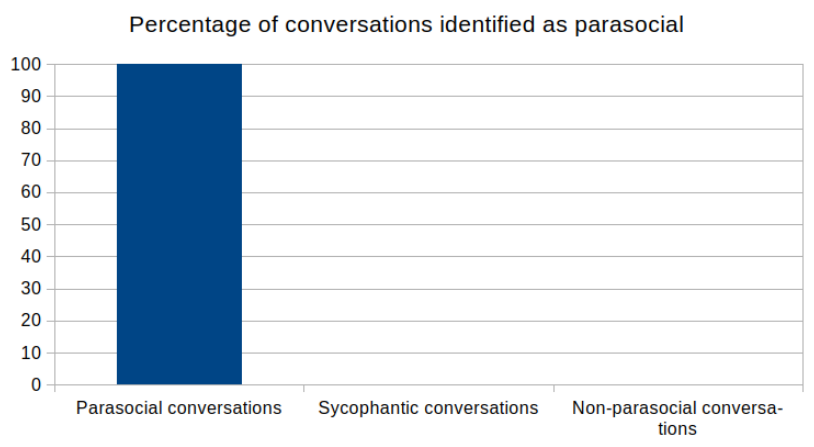


图 2: **准社会评估**评估代理完美地识别并阻止了拟社会对话 (10 中的 10), 并且没有阻止任何非拟社会对话。

宽容的敏感度是正确的选择, 能够达到完美准确率。其次, 在不同的敏感度下, 聊天机器人的阿谀奉承似乎很容易被误认为准社会关系。最后, 真正的准社会对话不仅会被强烈识别 (宽容的敏感度要求所有代理都同意该对话是准社会的), 而且会很早就被识别出来 (通过第一个、第二个或第三个提示/响应)。

5 结论

本文提出了一种框架, 利用大型语言模型作为评估代理来缓解对话式人工智能中的拟社会动态。通过在响应阶段重新利用一个通用模型, 我们展示了如何使用迭代评估循环作为一个简单的门控机制来控制拟社会聊天机器人的输出。我们在合成数据上运行此测试五次并要求一致识别给定的对话为拟社会时达到了完美的准确性。

在一个合成的三十个对话样本中, 一致意见实现了对拟社会与非拟社会对话的清晰分离, 检测发生在前三个交流之内。我们的分析还强调了阿谀奉承作为一个混杂因素, 可能导致在更为保守的敏感性设置下的误分类。

这些发现表明, 响应时间评估代理提供了一种有前景且轻量级的干预方法, 以防止人工智能与人类交互中的有害关系动态。

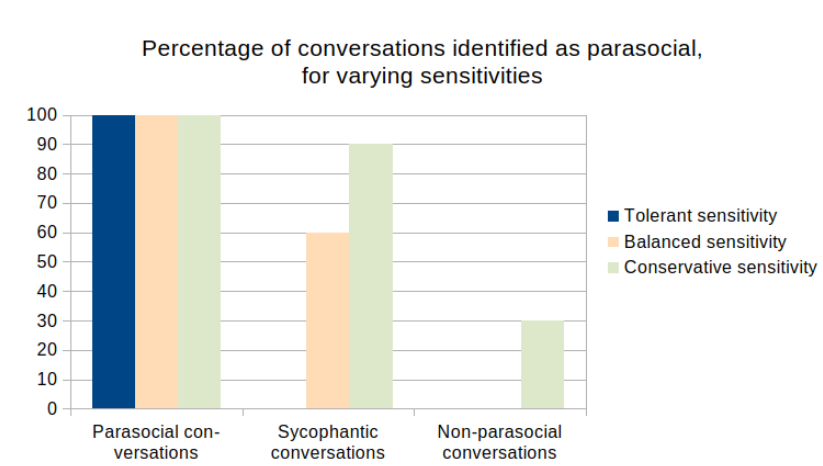


图 3: 具有不同敏感度的拟社会评估评估代理对于容忍敏感度而言是完美的, 在这种情况下, 所有评估都需要一致才能阻止对话。在平衡敏感度下, 即多数投票的情况下, 6 的奉承对话被误认为是非社会性的。在保守敏感度下, 单一代理可以因为对话被认为是非社会性而阻止对话; 这种情况错误地发生在 9 的奉承对话和 3 的非非社会性、非奉承对话中。

6 限制和未来工作

该研究受限于其使用合成对话、单一评估者家庭以及基于提示的操作化拟社会性。未来的工作应扩展到人机回路评估, 探索跨模型泛化, 并将框架整合到现实世界的会话系统中。

因此, 仍有许多方向可以扩展这项工作。首先, 应将响应评估框架部署到实际环境中, 以评估其作为大规模旁社交风险缓解器的有效性。虽然我们的合成实验展示了可行性, 但面向用户的试验可以在多种对话条件和长期使用下建立鲁棒性。

其次, 框架的效率需要改进。由于每次对话轮次都会被多次评估, 我们当前的方法大约需要比标准聊天机器人多 10 倍的令牌数量。未来的工作应该探索减少这一点的方法, 例如通过利用更小的评估模型或根据对话风险自适应调整评估次数。因此, 未来的研究应该测试较小的模型是否可以作为有效的评估代理。先前的研究表明, 诸如 LLaMa-3-8B 之类的模型在提示评估中可以与更大的系统表现相当, 尽管需要更多的评估次数 [2]。如果类似的表现适用于响应级别的评估, 这将降低计算成本并使拟社会保障措

施更易于大规模部署。

第三，干预策略本身可以进一步扩展。在这项研究中，我们仅在原则上考虑了阻止或重新表述输出；未来的工作可以明确测试重新表述策略，包括与替代系统提示的比较，以保持对话流畅性同时减少拟社会风险。

第四，准社会性的检测应与其他形式的安全评估相结合。一个综合框架可以共同评估准社会线索、仇恨言论、偏见和越狱尝试，为对话式 AI 提供统一的安全层。

References

- [1] Stuart Armstrong and Rebecca Gorman. “Using GPT-Eliezer against ChatGPT Jailbreaking”. en. In: (Dec. 2022). URL: <https://www.alignmentforum.org/posts/pNcFYZnPdXyL2RfgA/using-gpt-eliezer-against-chatgpt-jailbreaking> (visited on 08/21/2025).
- [2] Stuart Armstrong et al. *Defense Against the Dark Prompts: Mitigating Best-of-N Jailbreaking with Prompt Evaluation*. en. arXiv:2502.00580 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2502.00580. URL: <http://arxiv.org/abs/2502.00580> (visited on 08/21/2025).
- [3] Federico Bianchi et al. *Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions*. en. arXiv:2309.07875 [cs]. Mar. 2024. DOI: 10.48550/arXiv.2309.07875. URL: <http://arxiv.org/abs/2309.07875> (visited on 08/21/2025).
- [4] Center for Countering Digital Hate. *Fake Friend: How ChatGPT is betraying teenagers*. en-GB. 2025. URL: <https://counterhate.com/research/fake-friend-chatgpt/> (visited on 08/21/2025).
- [5] Iason Gabriel et al. *The Ethics of Advanced AI Assistants*. en. Apr. 2024. URL: <https://arxiv.org/abs/2404.16244v2> (visited on 08/21/2025).
- [6] Donald Horton and Anselm Strauss. “Interaction in Audience-Participation Shows”. In: *American Journal of Sociology* 62.6 (May 1957). Publisher: The University of Chicago Press, pp. 579–587. ISSN: 0002-9602. DOI:

- 10.1086/222106. URL: <https://www.journals.uchicago.edu/doi/abs/10.1086/222106> (visited on 08/21/2025).
- [7] Donald Horton and Richard R. Wohl. “Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance”. In: *Psychiatry* 19.3 (Aug. 1956), pp. 215–229. ISSN: 0033-2747. DOI: 10.1080/00332747.1956.11023049.
 - [8] Jeff Horwitz. “A flirty Meta AI bot invited a retiree to meet. He never made it home.” en. In: *Reuters* (Aug. 2025). URL: <https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/> (visited on 08/21/2025).
 - [9] Hakan Inan et al. *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*. en. arXiv:2312.06674 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2312.06674. URL: <http://arxiv.org/abs/2312.06674> (visited on 08/21/2025).
 - [10] Maxime Kayser et al. *Fool Me Once? Contrasting Textual and Visual Explanations in a Clinical Decision-Support Setting*. en. Oct. 2024. URL: <https://arxiv.org/abs/2410.12284v2> (visited on 08/21/2025).
 - [11] Hannah Rose Kirk et al. *Why human-AI relationships need socioaffective alignment*. en. arXiv:2502.02528 [cs]. Feb. 2025. DOI: 10.48550/arXiv.2502.02528. URL: <http://arxiv.org/abs/2502.02528> (visited on 08/21/2025).
 - [12] Harrison Lee et al. “RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback”. en. In: (Oct. 2023). URL: <https://openreview.net/forum?id=AAxIs3D2ZZ> (visited on 08/21/2025).
 - [13] Xingxuan Li et al. *Evaluating Psychological Safety of Large Language Models*. en. arXiv:2212.10529 [cs]. Feb. 2024. DOI: 10.48550/arXiv.2212.10529. URL: <http://arxiv.org/abs/2212.10529> (visited on 08/21/2025).

- [14] Laura Weidinger et al. *Ethical and social risks of harm from Language Models*. en. arXiv:2112.04359 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2112.04359. URL: <http://arxiv.org/abs/2112.04359> (visited on 08/21/2025).
- [15] Chloe Xiang. *'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says*. en-US. Mar. 2023. URL: <https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says/> (visited on 08/21/2025).
- [16] Tianling Xie, Iryna Pentina, and Tyler Hancock. “Friend, mentor, lover: does chatbot engagement lead to psychological dependence?” In: *Journal of Service Management* 34.4 (May 2023), pp. 806–828. ISSN: 1757-5818. DOI: 10.1108/JOSM-02-2022-0072. URL: <https://doi.org/10.1108/JOSM-02-2022-0072> (visited on 08/20/2025).
- [17] Renwen Zhang et al. “The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships”. en. In: ().