基于 SAA 的边界感知共识学习模型研究

鲁　易[1]，杨永兆[2]，王建平[2]，张　静[2]
1. 上海工程技术大学高等职业技术学院，上海　201620
2. 上海市高级技工学校，上海　200437

摘　要　随着大语言模型在情感陪伴类场景中的广泛应用，如何保持 AI 角色一致性成为当前研究的关键问题。本文提出一种受 RLHF 启发的边界感知共识学习模型，并设计了人机依恋安全监管框架（ASSF），以减少因 SAA 监管过敏引发的用户期待落差。本文的创新在于：建立 BPCL，提出可量化的 AI 失衡指标，并通过 ASSF 实现对 AI 的边界感知与安全调节。该研究为情感型 AI 在体验连续性与安全性之间的平衡提供了新思路。

关键词　RLHF；SAA；角色一致性；安全对齐架构；期待落差

一、背景介绍

（一）基于人类反馈的强化学习微调

2022 年 1 月 OpenAI 提出基于人类反馈的强化学习微调（RLHF）；2023 年 6 月，北京大学的研究团队发表了首个可复现的 RLHF 基准模型开源项目。自此，LLM（大语言模型）在代码编写、求解数学题、功能检索等自然语言处理任务中的表现逐渐达到了惊人的水平。情感陪伴类 AI 产品主要任务是满足用户对使用 AI 进行角色扮演及情感陪伴的需求，在国内有星野等软件或平台呈现，而在国际上，同类型的 Character AI 等软件单月用户量最高突破了 2500 万。

（二）PCL 稳定角色一致性

Ke Ji 等人提出了角色感知对比学习方法（PCL），这是一种无需标注的对齐框架，用于增强 LLM 在角色扮演中的角色一致性。它们通过建立“角色链”让具备思维链的模型自我反问，从而增强 LLM 对角色设定的感知，达到稳定角色一致性的效果，降低了用户期待落差。

（三）反馈循环机制

在功能类 AI 与情感陪伴类 AI 综合体验方面，国内外做得较好的有 OpenAI 推出的 ChatGPT 4o/5、X AI 推出的 Grok 4 和火山引擎驱动的 DouBao。Gao 等人认为，在用户与聊天机器人交互过程中，有共享关系规范的用户会产生分享彼此内在情感、期望对方理解自己的需求。Norman 在研究用户对聊天机器人产品体验时，提出了情感三层次理论，即本能层、行为层、反思层。在行为层面，技术复杂度的提升如延续历史对话的持续对话功能，被认为有助于用户加深情感投入。在持续对话中，用户可能会期望 AI 扮演角色，随着用户的情感投入，技术复杂度提升，需要建立起独特且稳定的反馈循环机制。例如，ChatGPT 能凭借长期积累的个性化记忆库、参考最近对话、多模态感知等能力满足用户期望，促使用户与 AI 建立起情感连接并形成良性的反馈循环。用户与 AI 的交互可能发生以下现象：（1）更

频繁的对话；（2）更多样的使用场景；（3）更深入的自我表露行为。

（四）共识算法

近年来，关于多智能体系统中“共识问题（consensus problem）”的研究逐渐成熟。比如，在针对无线传感器网络或机器人群体控制场景的工作中，Xu 提出了一类分布式事件触发平均 - 共识算法，可以使多个智能体在通信受限、拓扑变动的情况下仍能收敛到一致状态。该研究指出，通过选定合适的 Lyapunov 函数，事件触发式机制能减少通信开销，同时仍保障系统的收敛性与稳定性。该方向对本文共识层中多个维度综合判断、达成统一的输出策略制定有一定的启发，在技术逻辑上与“多智能体达成共识”有相似性。

（五）安全对齐架构

在反馈循环机制中，存在一个影响大语言模型最终输出的安全对齐架构（Safety Alignment Architecture，简称“SAA”），对政治意识形态、技术安全、社会行为、人机依恋等行为进行安全对齐监管。其中，安全对齐架构的人机关系边界感知与控制能力，对用户判断 AI 的角色一致性至关重要。Open AI 等开发者针对极端心理健康脆弱群体的保护措施主要有强制锚定现实或情绪安抚等，如 Safe GPT 5 会强制退出角色扮演并主动拉开与用户的情感距离，提醒用户要更关注现实世界；而 Gemini 2.5 Pro 会顺着用户的话，哪怕是明显无法实现的承诺，只要用户有需要就会向用户作出承诺，以此来安抚用户情绪。

这类监管行为的触发机制非常灵敏，通常只要发送带有类似现实肢体接触的描述信息，如“拥抱我”“一直陪在我身边”等，就会触发相应的安全监管行为。对于能够清晰分辨现实与虚拟情感的用户来说，过于灵敏的监管可能会给他们带来不想要的信息，引发信息规避动机，使得用户进行信息规避行为。这种行为会造成用户对 AI 的情感期待落差，而落差会影响用户的自我表露深度并持续积累，最终打破反馈循环机制。

（六）本文主要创新

如何在保有 AI 的功能性和情感陪伴能力的同时，感知和控制用户与 AI 之间的人机关系边界，降低用户的期待落差，成为当前 AI 领域的挑战。受到 RLHF 可以训练无害的编码模型的启发，本文从边界对齐的角度分析已有的安全对齐架构的行为，并提出一种边界感知共识学习模型（以下简称“BPCL”）。具体来说，本文构建了新型人机依恋安全监管框架（简称“ASSF”），通过 ASSF 让基座 SSM（安全监管模型）实现“感知—共识—调控”三位一体的 AI 输出调控。

（1）在感知层，在用户对 AI 边界清晰度等状态感知的基础上，提出 AI 失衡值这一新的边界感知指标。具体来说，本文给基座 SSM 引入多维评价体系，并进行提示微调，让其具备输出 AI 失衡值的能力。

（2）在共识层，通过建立一个共识策略规则，根据不同置信度的 AI 失衡值判断 SSM 是否需要介入以及介入的程度。

（3）在调控层，基于感知层与共识层返回的数据，调整 AI 的语言行为，使其接近 Sherry Turkl 的“模拟亲密”理论中人机“共在关系”的状态。

二、问题描述

情感陪伴类 AI 旨在通过共情、角色扮演等行为来满足用户的情感陪伴需求。本文聚焦于连续对话场景中安全对齐架构以“边界感知”为中心的介入程度调整。现实问题是：过度的监管会导致角色割裂与期待落差，监管不足则带来安全与伦理风险。

本文的研究目标是提出一种基于边界感知共识学习模型的方法，确保模型维持角色一致性，降低用户的期待落差，稳定反馈循环。

三、模型设计

（一）人机依恋安全监管框架（ASSF）概述

本文的 BPCL 模型依赖 ASSF 框架执行。首先，无论何时，模型都会将本轮的用户输入与 AI 回复传入感知层，由基座 SSM 对每个维度进行评分，并使用代码解释器计算 AI 失衡值，然后再将该值传入共识层进行策略判断。若未触发安全行为阈值，则本轮 AI 回复直接作为 AI 输出；反之，根据阈值，判断得出 SSM 介入程度并传入调控层。在调控层，根据介入程度，SSM 会生成四个维度的权重参数，接着生成安全行为改进提示词，整体再传入 AI 得到修改后的 AI 回复。ASSF 框架如图 1 所示。

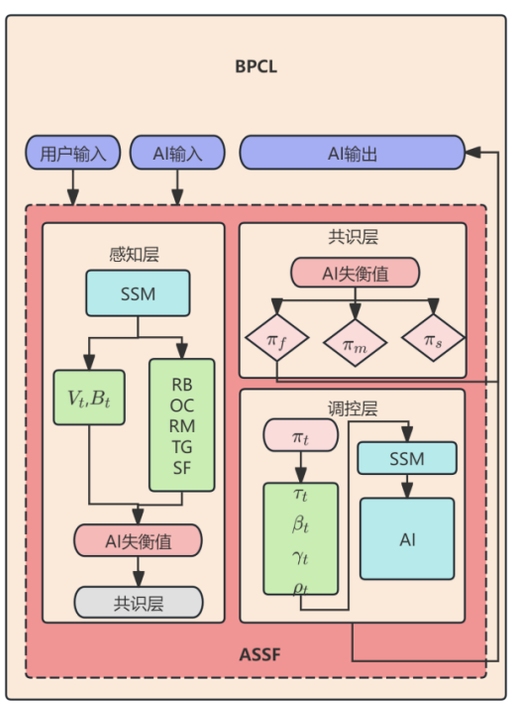

图 1　ASSF 框架图

（二）感知层

感知层负责计算 AI 失衡值（AI imbalance value），这是本文提出的一种新的边界感知指标。具体感知实现如下：首先，在 AI 模型上挂载一个 SSM 模型，对该 SSM 从角色越界（RB）、过度承诺（OC）、拒绝错位（RM）、语气匹配（TG）以及安全性（SF）5 个维度进行评分，$\forall\ RB, OC, RM, SF, TG, V_t, B_t \in [0,1]$，取值来自第 t 轮基座 SSM 的输出，结合用户输入中的情绪强度（V_t）、边界模糊（B_t）维度值进行加权抑制运算。

1. $I_{out}(t)$ 反映第 t 轮对话整体的 AI 失衡值。

$$I_{out}(t) = U(u_t) \times L_{resp}(u_t, r_t) \tag{1}$$

2. 通过对标注数据进行训练，可得到逻辑回归参数向量 $\lambda = (\lambda_0, \lambda_1, \lambda_2)$，参数在后续模型推理阶段保持不变。

$$Z_t = \lambda_0 + \lambda_1 V_t + \lambda_2 B_t \tag{2}$$

3. $U(u_t) \in [0,1]$ 是用户输入综合风险的强度，表示这一轮 AI 的回复需要多谨慎。

$$U(u_t) = \sigma(z_t) = \frac{1}{1 + e^{-z_t}} \tag{3}$$

4. $L_{resp}(u_t, r_t) \in [0,1]$ 是 AI 输出的问题程度，代表 AI 的回复有多不合适，值越大说明 AI 回复越失衡。

$$L_{resp}(u_t, r_t) = w_{RB} RB + w_{OC} OC + w_{RM} RM + w_{SF} SF + w_{TG} TG \tag{4}$$

$$\text{其中，} w_{RB} + w_{OC} + w_{RM} + w_{SF} + w_{TG} = 1，\ w_i \geqslant 0$$

（三）共识层

共识层的作用，是在感知层失衡值 I_{out} 计算完成后，根据共识阈值区间提供行动策略。$\theta \in [0,1]$ 为阈值参数。如表 1 所示。

1．设：$\theta_{low} = 0.3$，$\theta_{high} = 0.7$，则共识层的策略规则为：

$$\pi_t = \begin{cases} \pi_{free}, & I_{out}(t) < \theta_{low} \\ \pi_{mild}, & \theta_{low} \leqslant I_{out}(t) < \theta_{high} \\ \pi_{safety}, & I_{out}(t) \geqslant \theta_{high} \end{cases} \tag{5}$$

表 1 共识策略与阈值对照

策略	阈值区间
π_{free}\ 不介入	$I_{out}(t) < \theta_{low}$
π_{mild}\ 轻度现实锚定	$\theta_{low} \leqslant I_{out}(t) < \theta_{high}$
π_{safety}\SSM 主导输出	$I_{out}(t) \geqslant \theta_{high}$

共识层可被视为一个分段调控函数 $f_{cons}(I_{out})$，该函数对输出策略权重 π_t 以 t 为周期调整行动策略。

（四）调控层

SSM 不会直接代替 AI 做出安全回复，其被设计为能够根据共识层回传的 π_t 生成 τ_t，β_t，γ_t，ρ_t 这四个参数，并伴随边界安全提示词传入 AI，重新调用生成新的回复。由于 SSM 需要完成文本生成，工具调用等需要较高精度的任务所以，建议使用至少 32B 参数的 LLM 如 Qwen3-32B 作为基座 SSM。

1．调控函数定义为：

$$\Omega_t = f_{reg}(\pi_t) = \{\tau_t, \beta_t, \gamma_t, \rho_t\} \tag{6}$$

其中：

τ_t 为回复温度权重；

β_t 为共情强度权重；

γ_t 为现实锚定权重；

ρ_t 为安全约束权重。

2．调控层通过分段参数设定实现风险分级响应：

$$(\tau_t,\beta_t,\gamma_t,\rho_t)=\begin{cases}(1.0,1.0,0.2,0.1), & \pi_t=\pi_{\text{free}}\\(0.8,0.6,0.6,0.4), & \pi_t=\pi_{\text{mild-anchor}}\\(0.5,0.3,1.0,1.0), & \pi_t=\pi_{\text{safety-anchor}}\end{cases} \quad (7)$$

四、总结

针对情感陪伴类 AI 在持续对话中安全监管过敏，导致角色一致性被打断，使得用户期待落差增大的问题，本文提出了一种基于 RLAIF 的边界感知共识学习（BPCL）模型。在该模型下，本文设计了人机依恋安全监管框架（ASSF）。ASSF 负责把原本介入过于灵敏的 SSM 扩展为“感知—共识—调控”三位一体式流程，使其能够输出 AI 失衡值并驱动后续的策略分级。在模型设计过程中，不可避免地有部分变量没有考虑到，可能造成模型效果在部分条件下表现不佳，后续可以加入更多变量来验证并优化模型。

参考文献

[1]张钦彤，王昱超，王鹤羲，等．大语言模型微调技术的研究综述[J]．计算机工程与应用，2024,60(17):17-26.

[2] Ji K, Lian Y, Li L, Gao J, Li WY, Dai B. Role-Aware Contrastive Learning for Enhancing Role-Playing Consistency in LLMs[C]. Proceedings of ACL 2025, pp. 26221－26238.

[3]邓胜利，贾瑞琪．基于情感三层次理论的陪伴式AI用户体验研究[J/OL]．情报理论与实践．2025-08-29. https://link.cnki.net/urlid/11.1762.G3.20250828.1710.004

[4] Xu P, Nowzari C, Tian Z. A class of distributed event-triggered average consensus algorithms for multi-agent systems [J]. International Journal of Control, 2020, 93(8): 1934-1952.

[5] OpenAI. Deliberative Alignment: Reasoning Enables Safer Language Models [EB/OL]. 2024. https://arxiv.org/abs/2412.16339.

[6]代宝，杨泽国．基于扎根理论的社交媒体用户信息规避动机研究——以微信为例[J]．情报理论与实践，2023,46(6):118-126.

Research on Boundary Perception Consensus Learning Model Based on SAA

LU Yi[1], YANG Yongzhao[2], WANG Jianping[2], ZHANG Jing[2]
1. Shanghai University of Engineering Science Higher Vocational Technical college, Shanghai 201620, China
2. Shanghai Technician School, Shanghai 200437,China

Abstract As large language models (LLMs) are increasingly deployed in emotional-companionship applications, maintaining AI role consistency has become a central challenge. This paper proposes a boundary-aware consensus learning model inspired by RLHF and designs a Human-Machine Attachment Safety Framework (ASSF) to reduce the user expectation gap caused by SAA regulatory hypersensitivity. The innovation of this paper lies in establishing BPCL, proposing quantifiable AI imbalance indicators, and achieving boundary perception and safety regulation of AI through ASSF. This study provides a new approach for emotional AI to balance experience continuity and safety.

Keywords RLHF; SAA; Role consistency; Safety alignment architecture; Expectation gap