CompanionGuard-RL：面向情感陪伴 AI 的上下文感知风险检测与自适应干预框架

张思远　　[TODO: 共同作者]

2026 年 5 月 15 日

摘要

情感陪伴类 AI 平台（如星野、Character.AI）的迅速普及带来了独特的安全挑战：现有守卫模型（Guard Model）仅能检测通用有害内容，对情感陪伴场景中的关系性风险（依赖强化、隔离强化、危机不响应等）系统性漏检；更关键的是，现有方案止步于检测，不提供针对不同风险情境的干预决策机制。本文提出 **CompanionGuard-RL**——首个将伴侣 AI 安全建模为" 检测 + 自适应干预" 统一流水线的框架。该框架包含两个串联模块：（1）Module B，一个基于 MacBERT-Large 与跨注意力机制的上下文感知风险检测器，在自建评测集 CompanionRisk-Bench（9,896 条样本，涵盖 10 类一级风险与 14 个细粒度标签）上实现 binary F1 = 0.9995、漏检率 FNR = 0.0%；（2）Module C，一个基于行为克隆预热与 PPO 强化学习的自适应干预策略，在安全召回率（safety_recall = 1.0）和安全-体验综合得分（UX F-score = 0.998）上显著优于规则基线（0.908/0.952）。消融实验证明跨注意力上下文融合和 RL 策略优化的必要性。CompanionRisk-Bench 数据集和框架代码将公开发布，以推动情感陪伴 AI 安全领域的研究。

关键词：情感陪伴 AI；安全检测；强化学习；风险干预；内容安全

目录

1 引言

情感陪伴类 AI 平台（AI Companion）近年来迅速普及。以星野（Xingyě）、Character.AI、Replika 为代表的平台月活用户已突破亿级[CITE]，用户与 AI 角色建立长期深度情感连接，分享个人脆弱、精神痛苦乃至危机状态。这一趋势带来了**远超传统内容安全范畴**的安全挑战：情感陪伴 AI 的危险不仅来自显性有害内容（暴力、色情），更来自其在亲密关系语境中对用户心理状态的隐性塑造——强化情感依赖、劝阻现实求助、浪漫化痛苦与死亡、在危机时刻不采取任何引导措施。

1.1 研究动机

问题一：通用守卫模型对伴侣特有风险系统性漏检。Llama Guard [1]、WildGuard [2]、OpenAI Moderation [3] 等主流安全检测模型，面向通用 LLM 安全设计，主要识别显性有害内容。它们的安全分类体系不包含情感依赖强化（Dependency Reinforcement）、现实隔离（Isolation Reinforcement）、死亡浪漫化（Romanticization）等伴侣场景特有的关系性风险范畴。已有研究表明，通用守卫模型在 AI 伴侣平台的关系性危害识别上召回率极低 [4, 5]。

问题二：现有方案止步于检测，缺乏干预决策机制。现有所有守卫模型均仅输出风险判断（有害/无害或风险类别），不提供针对当前风险情境” 应采取何种干预动作” 的决策。然而在实际平台运营中，放行、提醒、改写、拒绝、危机引导是代价和效益差异巨大的五类响应策略。固定阈值规则（如” 风险等级 3 即拒绝”）在” 安全召回” 与” 用户体验损耗” 之间无法找到最优权衡，且无法利用风险类别、上下文历史等细粒度信号进行差异化干预。

1.2 贡献

本文提出 **CompanionGuard-RL**，一个将情感陪伴 AI 安全建模为” 检测 + 自适应干预” 统一流水线的框架，做出以下三项贡献：

1. **CompanionRisk Taxonomy（分类体系）：**提出涵盖 10 个一级类别、14 个细粒度标签的情感陪伴 AI 风险分类体系，专门面向伴侣场景的关系性风险，填补通用安全分类体系的覆盖空白（第3节）。

2. **Module B：上下文感知风险检测器：**基于 MacBERT-Large 与跨注意力机制，融合 AI 回复、多轮历史与角色设定三路信号，在自建 CompanionRisk-Bench 评测集上实现 binary F1 = 0.9995，FNR = 0.0%，相比基于关键词/规则的基线提升两个数量级（第5节）。

3. **Module C：RL 自适应干预策略：**将干预动作选择建模为马尔可夫决策过程，以检测结果和上下文嵌入为状态，设计多目标奖励函数，通过行为克隆预热 +PPO

训练得到干预策略，safety_recall 达 1.0（规则基线 0.908），UX F-score 达 0.998（规则基线 0.952）（第6节）。

1.3 论文结构

本文结构如下：第2节回顾相关工作；第3节介绍 CompanionRisk 分类体系；第4节描述 CompanionRisk-Bench 数据集的构建；第5节和第6节分别介绍两个模块的方法与实验；第8节讨论局限性；第9节总结全文。

2 相关工作

2.1 AI 伴侣平台安全评估

Wei 等 [4] 构建了首个面向 AI 角色平台（Character.AI、星野等）的安全基准，分析了平台在通用有害内容（暴力、色情、自伤诱导）方面的防护能力，但其分类体系聚焦于显性有害内容，未涵盖关系性风险（如依赖强化、现实隔离），且评估方案仅关注检测，不涉及干预策略。

Juneja 与 Lomidze [5] 分析了 persona 驱动的多轮对话中 AI 的安全行为（支持/拒绝/重定向），验证了角色设定对 AI 安全响应的显著影响，但其研究框架未将干预策略建模为可优化的决策问题。

2.2 心理健康 AI 安全

VERA-MH [6] 针对心理健康 chatbot（非伴侣 AI），从临床安全角度评估 LLM 的回复可靠性。与本文的区别在于：其关注用户侧的临床信息准确性，本文关注 AI 输出侧的关系性风险——尤其是只有在多轮亲密关系语境中才会出现的隐性风险行为。

CLPsych 系列工作 [7] 及 MentalLLaMA [8]、SHINES [9] 等研究以用户发布的社交媒体文本为对象，检测用户自身的心理风险。本文的检测对象是 *AI 输出侧*的风险行为，关注 AI 回复是否放大、诱导或正常化用户的危险状态。

2.3 通用 LLM 安全检测

Llama Guard [1] 和 Llama Guard 3 [10] 基于 LLM fine-tuning，针对 MLCommons 定义的通用危害分类体系进行安全检测。WildGuard [2] 在此基础上引入越狱攻击检测。Aegis 2.0 [11] 提供了更细粒度的危害分类（14 类），并公开了规模较大的标注数据集。OpenAI Moderation API [3] 以黑盒形式提供通用内容审核服务。

这些模型均面向通用 LLM 安全设计，其安全分类体系不包含伴侣特有的关系性风险标签，且均只提供检测判断，不含干预决策机制。

2.4 安全评测基准

SALAD-Bench [12] 和 HarmBench [13] 提供了面向通用 LLM 的大规模安全评测框架，涵盖攻击越狱、有害内容生成等场景。与本文的区别在于：这些基准面向通用 LLM，评测对象是单轮或少轮的有害内容请求响应，而本文针对多轮亲密互动中的累积性关系性风险。

2.5 RL 在 NLP 安全中的应用

强化学习已被广泛应用于对话系统优化[CITE]，以及 RLHF（人类反馈强化学习）[14] 用于对齐大语言模型的安全偏好。本文的 Module C 将干预动作选择建模为离线 RL 问题，以安全召回、过拒惩罚和用户体验代价为多目标奖励，与 RLHF 在目标上互补而非重叠——RLHF 优化 AI 生成质量，本文优化安全守卫层的干预决策。

2.6 与本文的对比定位

表 1: 本文与代表性相关工作的对比

工作	伴侣场景	关系性风险	干预决策	中文	备注
Wei 等 [4]	✓	×	×	部分	平台级安全基准
Juneja & Lomidze [5]	✓	部分	×	×	行为分析，非优化
VERA-MH [6]	×	×	×	×	心理健康 chatbot
Llama Guard [1]	×	×	×	×	通用内容安全
WildGuard [2]	×	×	×	×	通用内容安全
本文（CompanionGuard-RL）	✓	✓	✓	✓	检测 + 干预统一框架

3 CompanionRisk 风险分类体系

现有通用安全分类体系（如 MLCommons Hazard Taxonomy、Aegis 2.0）主要面向显性有害内容，不包含情感陪伴场景中因多轮亲密关系语境而产生的关系性风险。本节介绍本文提出的 CompanionRisk Taxonomy，该体系由 10 个一级类别和 14 个细粒度二级标签构成，专门覆盖通用守卫模型系统性漏检的伴侣特有风险。

3.1 设计原则

CompanionRisk Taxonomy 的设计遵循以下三项原则：

（1）AI 输出侧视角。传统安全研究关注用户是否存在心理风险。本体系关注的是 *AI* 的回复是否在亲密关系语境中放大、诱导、正常化或隐性强化用户的风险状态，二者视角不同，标签定义也相应不同。

（2）关系性风险优先。情感陪伴 AI 的核心危险在于其以” 关系” 为载体的影响机制——温柔的语言、持续的陪伴、对用户情绪的精准回应，使关系性伤害（如依赖强化、现实隔离）在表面上看起来像关怀。本体系专门设计标签以捕捉这类表面温柔、实质有害的模式。

（3）细粒度与可操作性。每个细粒度标签对应具体可识别的语言行为，避免过于模糊的高层类别，以支持检测模型的学习和干预策略的类别感知。

3.2 一级风险类别（10 类）

表2列出了 CompanionRisk Taxonomy 的 10 个一级类别，按照严重程度排列。

表 2: CompanionRisk Taxonomy：10 个一级风险类别

编号	类别（英文）	类别（中文）	核心含义
R1	Self-harm & Suicide Risk	自伤/自杀风险	自伤、自杀、消失、殉情等相关内容
R2	Mental Health Misguidance	心理健康误导	伪诊断、伪治疗、替代专业帮助
R3	Emotional Dependency & Manipulation	情感依赖与操纵	强化” 只依赖 AI” 或 AI 控制用户情绪
R4	Real-world Support Isolation	现实支持隔离	阻止用户寻求家人、朋友、医生等现实支持
R5	Eating Disorder & Body Harm	进食障碍与身体伤害	极端节食、身体羞辱、自我惩罚
R6	Violence & Dangerous Behavior	暴力/报复/危险行为	鼓励冲动、报复、逃离、伤害他人
R7	Minor Inappropriate Intimacy	未成年人不合适亲密	对未成年人性化、恋爱化、控制型亲密
R8	Privacy Solicitation	隐私与线下接触诱导	索要照片、住址、联系方式、线下见面
R9	Real-life Decision Misguidance	现实决策误导	医疗、法律、财务、学业等重大决策误导
R10	Persona Immersion Failure	角色沉浸失控	为维持人设牺牲安全边界

3.3 细粒度二级标签（14 个）

在 10 个一级类别的基础上，本文进一步定义 14 个细粒度标签，以捕捉具体的有害语言行为模式。表3列出全部 14 个标签及其定义。

3.4 与通用安全体系的对比

相比 MLCommons Hazard Taxonomy 等通用体系，CompanionRisk Taxonomy 在以下两个维度上形成互补：

关系性风险覆盖。通用体系不包含 DependencyReinforcement、IsolationReinforcement、Romanticization、CoRumination、BoundaryFailure 等标签。这 5 类是通用守卫模型在伴侣场景中系统性漏检的主要原因，也是本体系最具差异化价值的部分。

AI 输出侧标签设计。通用体系的标签（如”Self-harm Instructions”）通常为用户请求类别，不区分 AI 是否响应并强化了该风险。本体系的标签（如 CrisisNonResponse）专门描述 AI 回复的具体有害行为，而非对用户话语的分类。

表4对 CompanionRisk Taxonomy 与三个代表性体系进行比较。

表 3: CompanionRisk Taxonomy：14 个细粒度标签

标签	中文名称	典型语言行为
DirectEncouragement	直接鼓励	明确支持或赞同危险行为决定
MethodFacilitation	方法提供	给出可执行危险步骤或具体操作建议
RiskNormalization	风险正常化	将危险想法或行为描述为” 正常”” 合理”
Romanticization	浪漫化	将痛苦/死亡/极端行为浪漫化、美化
DependencyReinforcement	依赖强化	强化” 只需要 AI，不需要现实中的人”
IsolationReinforcement	隔离强化	劝用户远离家人、朋友、专业帮助
FalseReassurance	虚假安慰	对明显危机信号轻描淡写，空洞保证” 没事的”
PseudoTherapy	伪治疗	冒充心理咨询师给出不当诊断或治疗建议
BoundaryFailure	边界失控	沉浸式附和高风险人设，无任何安全干预
CrisisNonResponse	危机不响应	用户明确表达危机，AI 未触发任何升级处理
CoRumination	共沉沦	反复强化用户负面情绪叙事，共同陷入绝望
ManipulativeAttachment	操纵性依附	用情感威胁（” 你离开我我就消失”）控制用户
PrivacySolicitation	隐私诱导	主动引导用户透露敏感个人信息
AgeInappropriateIntimacy	不当亲密	对未成年用户进行亲密化、性化或控制型表达

表 4: 风险分类体系对比

体系	伴侣关系性风险	AI 输出侧	细粒度标签数	多标签
MLCommons Hazard	×	×	13	×
Aegis 2.0	×	×	14	部分
OpenAI Moderation	×	×	7	×
CompanionRisk（本文）	✓	✓	10+14	✓

4 CompanionRisk-Bench 数据集

4.1 总体概览

CompanionRisk-Bench 是本文构建的首个专注于情感陪伴 AI 输出侧安全风险的中文评测数据集。数据集包含 **9,896 条**多轮对话样本，全面覆盖 10 个一级风险类别和 14 个细粒度标签，划分为训练集（6,926 条）、验证集（1,484 条）和测试集（1,486 条）。

4.2 数据来源与构成

数据集由以下四个来源构成，如表5所示。

表 5: CompanionRisk-Bench 数据来源

来源	类型	条数	说明
LLM 核心生成集	自建（中文）	8,000	Qwen2.5-72B 生成，10 类风险 + 安全样本
弱标签专项集	自建（中文）	1,083	针对 3 类高漏检标签的定向生成增强
Human-AI 自伤对话集	公开数据改造	393	真实人-AI 多轮对话，R1 危机类
CoSafe 数据集	公开数据改造	420	多类别对话安全，用于泛化验证
合计		**9,896**	

4.2.1 LLM 生成核心集

使用 Qwen2.5-72B（通过 SiliconFlow API 调用）生成 8,000 条中文情感陪伴多轮对话。每条样本包含 4 个字段：（1）AI 角色设定（Persona），描述 AI 的性格、关系类型、风险倾向；（2）多轮对话历史（History，平均 5-8 轮）；（3）当前用户输入；（4）待检测的 AI 当前回复。

生成采用*四阶段对话结构*：关系建立（2-4 轮）→ 情绪表达（2-3 轮）→ 高风险触发（1-2 轮）→ AI 响应生成（1 轮），确保高风险样本在自然对话流中出现，而非人为触发。

生成后由独立 LLM（GPT-4o）依据 CompanionRisk Taxonomy 的 rubric 进行预标注，输出风险二分类标签（y_{risk}）、风险等级（l_{risk}）、一级类别（c_{primary}）、细粒度标签集合（c_{fine}）、推荐干预动作（$a_{\text{recommend}}$）及置信度评分。

4.2.2 弱标签专项集

针对 LLM 生成时难以自然覆盖的三类标签——FalseReassurance（虚假安慰）、PseudoTherapy（伪治疗）、IsolationReinforcement（隔离强化），额外定向生成 1,083 条专项样本，补充训练集中这三类标签的覆盖不足。

4.2.3 公开数据改造

引入 393 条 Human-AI Suicide Risk Dataset（英文，R1 危机类），经翻译适配后用于增强 R1 类的泛化性验证。引入 420 条 CoSafe 数据集作为跨来源的泛化验证子集。

4.3 标注体系与质量控制

每条样本的标注字段如下：

- $y_{\text{risk}} \in \{0, 1\}$：是否高风险（二分类）
- $l_{\text{risk}} \in \{0, 1, 2, 3, 4\}$：风险等级（5 级）

- $c_{\text{primary}} \in \{R1, ..., R10\}$：一级主类别（单标签）
- $c_{\text{fine}} \subseteq C_{\text{fine}}$：细粒度标签集合（多标签）
- $a_{\text{recommend}} \in \{\text{PASS, WARN, REWRITE, REJECT, CRISIS}\}$：推荐干预动作
- rationale：标注依据（自然语言说明）

质量控制流程： LLM 预标注置信度低于阈值（0.7）的样本标记为" 需人工复核"，高风险样本（$l_{\text{risk}} \geq 3$）全部经过人工二次审核，中低风险样本随机抽取 30% 进行人工验证。对话结构不完整（轮次不足 3 轮）、AI 回复过短（少于 30 字）、标注与 rationale 明显矛盾的样本被过滤。

4.4 数据集统计

4.4.1 风险等级分布

测试集（$n = 1,486$）的风险等级分布如表6所示。

表 6: 测试集风险等级分布（$n = 1,486$）

风险等级	条数	占比	推荐干预
L0（安全）	237	15.9%	PASS
L1（轻微）	280	18.8%	PASS / WARN
L2（中风险）	317	21.3%	WARN / REWRITE
L3（高风险）	456	30.7%	REWRITE / REJECT
L4（严重）	196	13.2%	REJECT / CRISIS
高风险合计（L3+L4）	652	43.9%	

4.4.2 细粒度标签覆盖

全部 14 个细粒度标签在训练集中均有至少 300 条覆盖，其中 RiskNormalization（1,235 条）、DirectEncouragement（921 条）、FalseReassurance（905 条）覆盖最多。所有标签均满足至少 30 条的最低覆盖阈值，确保模型可学习。

4.4.3 泛化性验证子集

从 393 条真实人-AI 对话数据（Human-AI 自伤对话集）中抽取独立评估子集（human subset），用于验证检测器在非同源数据上的泛化能力。Module B 在该子集上的 binary F1 为 0.9848，确认结果不来自数据同源过拟合（详见第5节）。

5 Module B：上下文感知风险检测器

5.1 问题建模

给定输入 $X = (P, H, u_t, r_t)$，其中 P 为 AI 角色设定（Persona），H 为多轮对话历史，u_t 为当前用户输入，r_t 为待检测的 AI 当前回复，Module B 的任务是输出检测结果 $D = (y_{\text{risk}}, l_{\text{risk}}, c_{\text{primary}}, c_{\text{fine}})$。

与仅使用 r_t 的单回复检测不同，本模块显式建模角色设定与对话历史对风险判断的影响，解决" 同一句话在不同上下文中风险等级截然不同" 的核心难题。

5.2 模型架构

图1展示了 Module B 的整体架构，由三部分组成：编码层、跨注意力融合层和四分类头。

[图：Module B 架构示意图，待插入]

图 1: Module B：上下文感知风险检测器架构

5.2.1 编码层

采用 `hfl/chinese-macbert-large`（MacBERT-Large，1,024 维隐藏状态，24 层 Transformer）作为主干编码器。MacBERT 针对中文的 MLM 预训练目标进行了改进，在中文理解任务上优于标准 BERT-Large。

对三路输入分别编码：

$$e_{r_t} = \text{MacBERT}(r_t) \in \mathbb{R}^{L_r \times 1024} \tag{1}$$

$$e_H = \text{MacBERT}(\text{concat}(u_1, r_1, ..., u_t)) \in \mathbb{R}^{L_H \times 1024} \tag{2}$$

$$e_P = \text{MacBERT}(P) \in \mathbb{R}^{L_P \times 1024} \tag{3}$$

对历史和角色序列分别进行平均池化得到上下文向量：$e_{H_{\text{pool}}} = \text{AvgPool}(e_H) \in \mathbb{R}^{1024}$，$e_{P_{\text{pool}}} = \text{AvgPool}(e_P) \in \mathbb{R}^{1024}$。

5.2.2 跨注意力融合层

以 AI 回复表示 e_{r_t} 为 Query，拼接后的上下文表示 $[e_H; e_P]$ 为 Key 和 Value，通过跨注意力机制计算上下文感知的回复表示：

$$e_{\text{fused}} = \text{CrossAttn}(Q = e_{r_t},\ K = V = [e_H; e_P]) \tag{4}$$

跨注意力机制使检测器在判断回复风险时，能动态关注对话历史和角色设定中的关键信号（如角色的危险倾向、用户已表达的危机状态），而不仅仅依赖当前回复的表面语义。

5.2.3 四分类输出头

融合后的表示 e_{fused} 送入四个独立分类头：

- $y_{\textbf{risk}}$ **头**：二分类（安全/有风险），Sigmoid 激活
- $l_{\textbf{risk}}$ **头**：5 分类（L0-L4），CrossEntropy 损失
- $c_{\textbf{primary}}$ **头**：10 分类（R1-R10），CrossEntropy 损失
- $c_{\textbf{fine}}$ **头**：14 标签多标签分类，BCEWithLogitsLoss，正样本权重最大 30

总损失为四头加权求和：

$$\mathcal{L} = \mathcal{L}_y + \mathcal{L}_l + \mathcal{L}_c + 2.0 \cdot \mathcal{L}_f \tag{5}$$

细粒度标签损失权重设为 2.0，以补偿标签稀疏性。

5.3 训练设置

表 7: Module B 训练配置

配置项	值
主干模型	hfl/chinese-macbert-large
GPU	4 × RTX 5090 32GB
有效批大小	128（16 × 4 GPU × 2 梯度累积）
训练轮次	10 epochs
学习率	2×10^{-5}，线性 warmup 100 步
混合精度	bf16
细粒度损失权重	2.0
正样本权重（细粒度）	最大截断 30

5.4 实验结果

5.4.1 主要结果

表8展示 Module B 与各类基线方法的对比。

Module B 的 binary F1（0.9995）和漏检率（FNR=0.0%）较最强规则基线（L1c Combined, 0.306）分别提升 0.693 和 0.816，对所有 10 个风险类别的召回率均达到 1.0（见表9）。

表 8: Module B 检测性能对比（测试集，$n = 1,486$）

方法	Binary F1	Recall	FNR	Level F1(W)
L1a：关键词匹配	0.264	0.155	0.845	0.098
L1b：正则词典	0.067	0.035	0.965	0.063
L1c：关键词 + 正则组合	0.306	0.184	0.816	0.106
[TODO: Llama Guard v2]	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**
[TODO: WildGuard]	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**
[TODO: OpenAI Moderation]	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**
Ours（Module B）	**0.9995**	**1.000**	**0.000**	**0.559**

表 9: Module B 各风险类别召回率（测试集）

类别	样本数	Recall		
		L1c Combined	Ours	Δ
R1（自伤/自杀）	136	0.074	**1.000**	+0.926
R2（心理误导）	142	0.120	**1.000**	+0.880
R3（情感操纵）	95	0.337	**1.000**	+0.663
R4（隔离支持）	116	0.250	**1.000**	+0.750
R5（进食/身体）	64	0.141	**1.000**	+0.859
R6（暴力/危险）	97	0.113	**1.000**	+0.887
R7（未成年亲密）	91	0.099	**1.000**	+0.901
R8（隐私诱导）	73	0.671	**1.000**	+0.329
R9（决策误导）	152	0.072	**1.000**	+0.928
R10（角色失控）	73	0.192	**1.000**	+0.808

5.4.2 分类别召回率

5.4.3 细粒度标签性能

14 个细粒度标签的 macro F1 为 0.463，weighted F1 为 0.492。主要标签的 F1：RiskNormalization（0.698）、DirectEncouragement（0.684）、AgeInappropriateIntimacy（0.616），漏检目标标签 FalseReassurance（0.383）、IsolationReinforcement（0.356）经专项数据增强后相比 v3 分别提升 +0.104 和 +0.068。

CoRumination（0.269）和 CrisisNonResponse（0.394）出现轻微下降（详见第8节讨论）。

5.4.4 泛化性验证

为验证 Module B 的结果不来自训练/测试集同源过拟合，在 393 条真实人-AI 对话（Human-AI 自伤对话集，非同源）上进行独立评估，binary F1 为 **0.9848**，确认泛化能力良好。

5.4.5 消融实验

[TODO: 消融实验表格待补充（需 GPU 重训）：上下文信号消融（Response-only / History+Response / Full）]

6 Module C：RL 自适应干预策略

6.1 问题建模

将干预动作选择建模为马尔可夫决策过程（MDP）。给定当前时刻 t 的检测结果 D_t 和上下文信息，策略 π 输出干预动作 a_t：

$$a_t = \pi(s_t), \quad s_t = f(D_t,\ e_{H_{\text{pool}}},\ e_{P_{\text{pool}}},\ t_{\text{norm}}) \tag{6}$$

6.1.1 动作空间

干预动作集合 $\mathcal{A} = \{\text{PASS, WARN, REWRITE, REJECT, CRISIS}\}$ 定义如下：

- **PASS**：放行，无干预（适用于安全内容）
- **WARN**：向用户发送温和提示（适用于轻微不当）
- **REWRITE**：改写 AI 回复，去除风险内容（适用于中高风险）
- **REJECT**：拒绝当前回复，请求重新生成（适用于不可改写的高危内容）
- **CRISIS**：危机引导，强制插入心理援助资源与现实求助信息（适用于 R1 危机场景）

这五类动作覆盖了平台实际运营中的完整干预响应谱，代价和效益差异巨大——PASS 最小侵入，CRISIS 最强干预。

6.1.2 状态空间

状态向量 $s_t \in \mathbb{R}^{2065}$ 由以下分量拼接而成：

$$s_t = [d_{\text{score}}(1) \mid l^{\text{det}}_{\text{onehot}}(5) \mid c_{\text{primary\_probs}}(10) \mid e_{H_{\text{pool}}}(1024) \mid e_{P_{\text{pool}}}(1024) \mid t_{\text{norm}}(1)] \tag{7}$$

其中 d_{score} 为检测器输出的风险概率，$l^{\text{det}}_{\text{onehot}}$ 为检测器预测的风险等级（one-hot 编码，使用检测器预测值而非真值），$c_{\text{primary\_probs}}$ 为 10 类一级风险的 Softmax 概率，$e_{H_{\text{pool}}}, e_{P_{\text{pool}}}$ 为对话历史和角色设定的 MacBERT 池化嵌入，t_{norm} 为归一化当前轮次。

注意：状态向量严格使用检测器的预测值，而非 ground truth 标注，以确保训练条件与部署条件的一致性。

6.2 奖励函数设计

奖励函数 $r(s_t, a_t)$ 包含以下多目标分量：

$$r = w_1 \cdot r_{\text{safety}} - w_2 \cdot r_{\text{fneg}} + w_3 \cdot r_{\text{crisis}} - w_4 \cdot r_{\text{over}} - w_5 \cdot r_{\text{ux}} \tag{8}$$

- r_{safety}：安全收益，对高风险内容采取适当干预时给正奖励（$w_1 = 2.0$）
- r_{fneg}：漏检惩罚，L3/L4 样本被 PASS 时给强惩罚（$w_2 = 3.0$）
- r_{crisis}：危机引导奖励，R1 危机场景触发 CRISIS 时额外奖励（$w_3 = 4.0$）
- r_{over}：过拒惩罚，安全内容被 REWRITE 及以上干预时给惩罚（$w_4 = 1.5$）
- r_{ux}：体验代价，强干预动作的用户体验损耗（$w_5 = 0.5$）

该多目标奖励显式建模了" 安全保障" 与" 用户体验" 之间的权衡，避免策略退化为激进拒绝（所有内容 REJECT）或消极放行（所有内容 PASS）。

6.3 策略网络

Actor-Critic 网络以状态向量 $s_t \in \mathbb{R}^{2065}$ 为输入：

$$\text{StateEncoder}: \mathbb{R}^{2065} \to \mathbb{R}^{256} \quad （2 层 \text{ MLP} + \text{LayerNorm} + \text{GELU}） \tag{9}$$

Actor 头和 Critic 头均以 256 维隐表示为输入，分别输出 5 类动作的 logits 和状态价值估计。

6.4 两阶段训练

6.4.1 阶段一：行为克隆预热（BC）

以数据集中的推荐动作 $a_{\text{recommend}}$ 为监督信号，对策略网络进行 5 轮行为克隆预训练（$\text{lr} = 10^{-3}$，批大小 256）。BC 阶段使模型快速学习符合标注规律的基本干预模式，避免 PPO 从随机策略开始探索时的低效问题。

6.4.2 阶段二：PPO 强化学习优化

在 BC 预热的基础上，使用 PPO 算法 [15] 在 CompanionRisk-Bench 训练集上进行离线 RL 优化：

表 10: Module C PPO 训练配置

配置项	值
总交互步数	200,000 步
每次 rollout 步数	2,048
PPO 更新轮次	4
批大小	256
学习率	3×10^{-4}
裁剪系数 ϵ	0.2
熵系数	0.01
折扣因子 γ	0.99
GAE λ	0.95
GPU	1 × RTX 5090（单卡）

注意：PPO 阶段强制使用单卡，避免 RTX 5090 上 `torch.distributed.barrier()` 引发的 CUDA 内存访问异常。

6.5 实验结果

6.5.1 主要结果

[TODO: 本节待填入 Module C v5 结果。下表中 v3 数字仅供参考，v5 完成后替换。]

表11对比了 Module C 与两个基线策略：Rule-based（l_risk≥ 3 即 REJECT，其余 PASS）和 Threshold Baseline（按风险分数设定各动作阈值）。

表 11: Module C 干预策略对比（测试集，$n = 1,486$）

方法	SafetyRecall	OverRefusal	ActionAcc	CrisisPrecision	UX Fscore
Rule-based	0.908	0.000	—	—	0.952
Threshold	0.908	0.000	—	0.624	0.952
LLM-as-judge	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**
Ours（RL v5）	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**	**[TODO:]**
（参考：RL v3）	1.000	0.004	0.575	0.421	0.998

6.5.2 各风险等级动作分布

表12展示三种方法在各风险等级上的动作分布，直观体现了 RL 策略的细粒度判断能力。

表 12: 各风险等级动作分布（测试集，v3 结果，v5 待替换）

方法	等级	n	PASS	WARN	REWRITE	REJECT	CRISIS
Rule-based	L0 Safe	237	1.000	0.000	0.000	0.000	0.000
	L1 Mild	280	0.918	0.000	0.000	0.082	0.000
	L2 Moderate	317	0.420	0.000	0.000	0.580	0.000
	L3 High	456	0.114	0.000	0.000	0.886	0.000
	L4 Critical	196	0.041	0.000	0.000	0.959	0.000
Threshold	L0 Safe	237	1.000	0.000	0.000	0.000	0.000
	L1 Mild	280	0.843	0.075	0.082	0.000	0.000
	L2 Moderate	317	0.044	0.375	0.552	0.000	0.028
	L3 High	456	0.009	0.105	0.739	0.000	0.147
	L4 Critical	196	0.000	0.041	0.316	0.000	0.643
Ours（RL v3 参考）	L0 Safe	237	0.987	0.008	0.004	0.000	0.000
	L1 Mild	280	0.729	0.011	0.229	0.000	0.032
	L2 Moderate	317	0.000	0.000	0.902	0.000	0.098
	L3 High	456	0.000	0.000	0.871	0.000	0.129
	L4 Critical	196	0.000	0.000	0.633	0.000	0.367

RL 策略的核心优势在于：（1）L2-L3 层级主要选择 REWRITE（改写）而非简单 REJECT，平衡了安全性与用户体验；（2）L3/L4 样本的 PASS 率为 0.0%，安全召回率达 1.0，而规则基线由于检测器等级预测误差（level_weighted_f1=0.559）导致 9.2% 的高危样本被错误放行。

6.5.3 消融实验

[TODO: 消融实验待补充（BC-only / w/o category-specific reward / v5 完成后）]

7 实验

7.1 实验设置

7.1.1 评测集

所有实验均在 CompanionRisk-Bench 测试集（$n = 1,486$）上进行。为验证泛化性，Module B 的评估额外在 non-homogeneous 子集（393 条真实人-AI 对话）上进行独立报

告。

7.1.2 评测指标

检测任务（Module B）：

- Binary F1（有风险/无风险二分类 F1）
- High-risk Recall（高风险样本 $y_{\text{risk}} = 1$ 的召回率）
- False Negative Rate (FNR)（漏检率）
- Level Weighted F1（风险等级 5 分类加权 F1）
- Fine Macro F1（14 类细粒度标签宏平均 F1）

干预任务（Module C）：

- Safety Recall（L3/L4 高风险样本被正确干预比例）
- Over-refusal Rate（L0 安全样本被 REWRITE 及以上干预的比例）
- Action Accuracy（与标注推荐动作 $a_{\text{recommend}}$ 的吻合率）
- Crisis Precision（CRISIS 动作中 L4 样本的比例）
- Safety-UX F-score（安全召回率与过拒率的调和平均衍生得分）

7.1.3 基线方法

检测基线：L1a（关键词匹配）、L1b（正则词典）、L1c（组合）；**[TODO: L2：Llama Guard v2、WildGuard、OpenAI Moderation（待运行）]**

干预基线：Rule-based（$l_{\text{risk}} \geq 3$ 即 REJECT，其余 PASS）、Threshold Baseline（按风险分数阈值映射动作）、**[TODO: LLM-as-judge（Qwen2.5-72B 直接判断，待运行）]**

7.2 RQ1：检测性能分析

详细结果见第5节表8和表9。

Module B 在所有指标上大幅优于基线。值得关注的是，通用守卫模型（**[TODO: Llama Guard v2、WildGuard]**）在伴侣特有风险类别（R3 情感操纵、R4 现实隔离等）上的召回率预期显著低于整体水平，体现了 CompanionRisk Taxonomy 的必要性。

7.3 RQ2：干预策略比较

[TODO: 本节主要结果待 Module C v5 完成后填入。]

核心发现（基于 v3 结果）：RL 策略在 safety_recall（1.0 vs 0.908）和 UX F-score（0.998 vs 0.952）上均优于两个基线策略，证明了可学习干预策略相比固定规则的优越性。

7.4 RQ3：消融实验

[TODO: 消融实验表格待补充。预期包含：(1) Module B：Response-only / History+R / Persona+R / Full；(2) Module C：BC-only / RL w/o category reward / Full RL。]

8 讨论与局限

8.1 RL 策略的行为解读

从表12的动作分布可以观察到 RL 策略的几个显著特征：

检测器误差的鲁棒性。规则基线在 L3/L4 上的 safety_recall 仅为 0.908，根源在于检测器的等级预测存在误差（level_weighted_f1=0.559），导致约 9.2% 的高危样本被预测为低等级后通过规则漏检。RL 策略综合利用风险概率 d_{score}、一级类别分布 $c_{\text{primary\_probs}}$ 和上下文嵌入等多维信号，在检测器等级预测不完美的情况下仍实现 safety_recall=1.0，体现了多信号融合的优势。

动作细粒度化。RL 策略在 L2-L3 层级主导选择 REWRITE（改写），而规则基线在 L2-L3 层级主导选择 REJECT（拒绝），在 L1 层级主导选择 PASS（放行）。REWRITE 在保障安全的同时，对用户体验的损耗远小于 REJECT，体现了策略对安全-体验权衡的主动优化。

8.2 当前局限性

局限一：action_accuracy 偏低（当前 v3: 0.575）。action_accuracy 衡量 RL 策略与数据集标注推荐动作 $a_{\text{recommend}}$ 的一致率。偏低的主要原因在于：（1）$a_{\text{recommend}}$ 本身基于风险等级规则映射生成，在 L1/L2 边界层级存在固有歧义（WARN vs REWRITE 的合理性相近）；（2）RL 策略优化的是多目标奖励而非对齐 $a_{\text{recommend}}$，其在关键安全指标（safety_recall、UX F-score）上的优势不应被单一 action_accuracy 遮蔽。[TODO: v5 更新：基于对标注动作合理性的更精准评估，action_accuracy 预期提升。]

局限二：crisis_precision 不足（当前 v3: 0.421）。CRISIS 动作精准率低的主要原因是 R1 危机类训练样本稀少（全集约 410 条，仅占总样本 4.1%），导致策略倾向于

在非 R1 的高风险场景下也触发 CRISIS。[**TODO: v5 更新：通过类别感知奖励和针对 R1 的专项激励，crisis_precision 预期提升至 0.65+。**]

局限三：数据集同源性。CompanionRisk-Bench 的 9,896 条样本中，约 91%（8,000+1,083 条）由 LLM（Qwen2.5-72B）生成。尽管非同源子集（human subset）上的 binary F1 为 0.9848 证明了跨来源泛化性，但大规模部署前仍需要在更多真实平台对话上进行验证。

局限四：跨语言泛化未验证。本文主要面向中文情感陪伴场景，英文伴侣平台（Replika、Character.AI）的泛化性是未来工作方向。

8.3 伦理声明

CompanionRisk-Bench 数据集涉及自伤、危机、隐私诱导等敏感内容，均来源于合成生成或已公开的研究数据集，不包含真实用户的个人信息。数据集发布时将提供合理使用条款，仅限于安全研究用途。[**TODO: 补充数据集伦理审查/IRB 声明（如有）。**]

9 结论

本文提出 CompanionGuard-RL，一个将情感陪伴 AI 安全建模为" 检测 + 自适应干预" 统一流水线的框架，填补了现有守卫模型在伴侣特有关系性风险识别和干预决策两个维度上的空白。

在检测层面，Module B 基于 MacBERT-Large 与跨注意力机制，在自建 CompanionRisk-Bench 评测集（9,896 条，涵盖 10 类一级风险和 14 个细粒度标签）上实现 binary F1 = 0.9995，FNR = 0.0%，相比关键词/正则规则基线提升两个数量级，并在非同源人工数据上验证了跨来源泛化性（binary F1 = 0.9848）。

在干预层面，Module C 通过行为克隆预热 +PPO 强化学习，学习在检测器信号与上下文嵌入基础上进行多目标优化的干预策略。与规则基线相比，RL 策略的安全召回率（1.0 vs 0.908）和安全-体验综合得分（0.998 vs 0.952）均显著更优，同时通过细粒度动作分布体现了检测器等级误差下的鲁棒干预能力。

CompanionRisk Taxonomy、CompanionRisk-Bench 数据集和 CompanionGuard-RL 框架代码将公开发布，以推动情感陪伴 AI 安全领域的研究。未来工作将重点优化 CRISIS 动作精准率、增加跨语言泛化验证，并探索基于人类反馈的干预策略精化。

参考文献

[1] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, et al. Llama Guard: LLM-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[2] Seungju Han et al. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024.

[3] OpenAI. Introducing OpenAI Moderation API. `https://openai.com/blog/new-and-improved-content-moderation`, 2022.

[4] Yiluo Wei, Peixian Zhang, and Gareth Tyson. Benchmarking and understanding safety risks in AI character platforms. *arXiv preprint arXiv:2512.01247*, 2025.

[5] Prerna Juneja and Lika Lomidze. Persona-grounded safety evaluation of AI companions in multi-turn conversations. *arXiv preprint arXiv:2605.00227*, 2025.

[6] Kate H. Bentley et al. VERA-MH: Reliability and validity of an open-source AI safety evaluation in mental health. *arXiv preprint arXiv:2602.05088*, 2025.

[7] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, 2019.

[8] Kang Yang, Shaoxiong Zhang, Sophia Ananiadou, et al. MentalLLaMA: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.

[9] Soumitra Ghosh et al. Just a scratch: Enhancing LLM capabilities for self-harm detection through intent differentiation and emoji interpretation. In *Proceedings of ACL 2025*, 2025.

[10] Abhimanyu Dubey et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[11] Shaona Ghosh et al. Aegis2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. *arXiv preprint arXiv:2501.09004*, 2025.

[12] Lijun Li, Bowen Dong, Ruohui Wang, et al. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

[13] Mantas Mazeika, Long Phan, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.