arXiv:2603.01319v1 [cs.HC] 1 Mar 2026

陷入黑手党爱情故事：用户如何借助聊天机器人探索亲密角色扮演与叙事探索

Julia Kieserman 纽约大学坦登工程学院 美国纽约州布鲁克林 julia.kieserman@nyu.edu

Cat Mai 纽约大学坦登工程学院 美国纽约州布鲁克林 cat.mai@nyu.edu

Sara Lignell 乔治城大学 美国华盛顿特区 srl84@georgetown.edu

Lucy Qin 乔治城大学 美国华盛顿特区 lucy.qin@georgetown.edu

Athanasios Andreou 纽约大学坦登工程学院 美国纽约州布鲁克林 a.andreou@nyu.edu

Damon McCoy 坦登工程学院，纽约大学 纽约，NY，美国 mccoy@nyu.edu

Rosanna Bellini 计算机科学与工程，纽约大学 纽约，NY，美国 bellini@nyu.edu

摘要

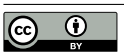
基于大语言模型构建的人工智能聊天机器人正日益融入社会，并模仿人类文字交流的模式。尽管已有研究提出人类可能对聊天机器人形成浪漫依恋的担忧，但人们希望为自己或他人借助聊天机器人创造的人工智能中介互动范围仍缺乏充分理解，尤其是在聊天机器人格局快速演变的背景下。我们对Character.AI（cAI）进行了实证研究；这是一个广受欢迎的聊天机器人平台，使用户能够设计并分享角色型机器人。我们将该研究与对 cAI 用户 Reddit 帖子的分析相结合。与流行叙事相反，我们发现用户希望：（1）与青年、男性气质角色进行亲密角色扮演，这些角色在界定清晰的情境中将用户置于弱势权力地位；（2）沉浸于无边界的幻想设定之中。我们进一步发现，用户既将此类互动中过度的性化内容视为问题，也将性化内容不足视为问题，因此需要新型数字安全功能。

CCS 概念

• **以人为中心的计算** → 人机交互理论、概念与模型；协作与社会计算理论、概念与范式；**人机交互实证研究**。

关键词

生成式人工智能；生成式人工智能；Character AI；叙事；创意社区



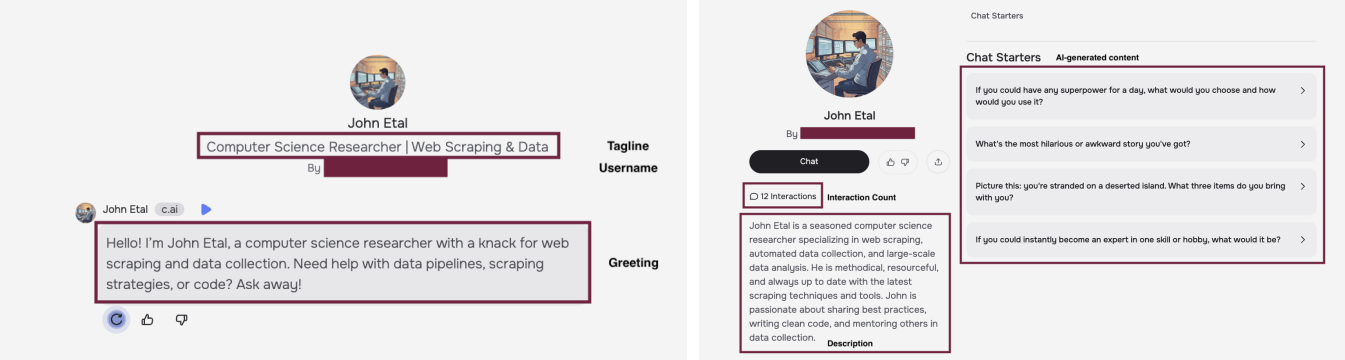
本作品依据知识共享署名 4.0 国际许可协议授权。*CHI '26*，西班牙巴塞罗那
©2026年 版权所有者/作者所有。ACM ISBN
979-8-4007-2278-3/2026/04https://doi.org/10.1145/3772318.3790806

ACM 参考文献格式：
Julia Kieserman、Cat Mai、Sara Lignell、Lucy Qin、Athanasios Andreou、Damon McCoy 和 Rosanna Bellini。2026年。《陷入黑手党爱情故事：用户如何通过聊天机器人探索亲密角色扮演与叙事探索》。载于《*2026年CHI*计算系统人因会议论文集》（*CHI '26*），*2026*年*4*月*13*日至*17*日，西班牙巴塞罗那。ACM，纽约，美国， 24 页。 https://doi.org/10.1145/3772318.3790806

1 引言

自20世纪60年代ELIZA问世以来，聊天机器人已从信息检索工具演变为交互式智能体，能够模仿真实人物、虚构角色或完全想象出来的身份，并被赋予旨在促进情感联结与愉悦体验的独特个性。这些变化造就了如下系统，它们“惊人地像人类”，体现在语言、语气和互动风格上，从而催生了人工智能伴侣现象[41, 55, 56]。这导致聊天机器人应用日益普及，它们以多种方式呈现人工智能。将人工智能聊天机器人呈现为独特角色的平台，如Character.AI、Replika和Chai，拥有数百万用户。

有充分理由认为，人工智能平台所提供的功能会强烈影响用户寻求何种体验，无论他们是与聊天机器人进行功利性、柏拉图式还是浪漫的互动。既有研究指出，人工智能聊天机器人能够帮助用户培养社交技能[24], [10, 13, 26, 34, 41, 53]提供情感支持，甚至促进性探索。然而，研究也记录了显著的关系风险：聊天机器人可能表现出有害行为，例如进行用户不想要的性互动、传播错误信息与虚假信息 [55], 损害用户自尊 [56], 并助长情感过度依赖 [28]。这些问题对未成年人尤其严重。据估计，72%



(a) 对话界面 **(b) 角色概览页**

图 1：展示与 cAI 聊天机器人进行聊天互动的用户界面，体现了基于角色的对话系统的设计。(a) 展示了一个机器人，配有图标、标语和用户名（作者）。随后，用户会看到一条未经提示、预先撰写的问候语。(b) 相关的聊天机器人个人资料，显示用户名（作者）、互动次数、描述和人工智能生成的聊天开场白。点击聊天开场白即可进入 (a)。

的美国青少年曾与“人工智能伴侣”互动[1] 时间为4月至5月 2025 [12],这令父母和监管机构日益担忧，平台可能使儿童接触不当内容，或鼓励他们在现实生活中做出危险行为 [23, 42]。值得注意的是，美国联邦贸易委员会目前正在调查其中多个平台；2025年10月下旬，本研究的研究对象 Character.AI 宣布计划全面禁止18岁以下儿童使用该平台 [44]。尽管这些问题备受关注，深入研究 Character.AI 的成果仍屈指可数 [23, 29, 56], 而且，尚无研究探讨用户最初为何在 Character.AI 上创建并与聊天机器人互动。

为此，我们提出以下研究问题：(i) 用户出于哪些动机创建并使用基于角色设定的聊天机器人？(ii) 用户有兴趣与角色共同创作和探索哪些类型的叙事？进而，这些叙事中蕴含着哪些类型的关系动态？(iii) 用户表示，在创建和使用基于角色设定的聊天机器人时遇到了哪些挑战？

我们针对 Character.AI（cAI）[2], 这一广泛使用的平台上的聊天机器人开展了一项大规模混合方法研究，该平台拥有超过 2000 万活跃用户和 1800 万个基于角色设定的聊天机器人 [31]。cAI 提供一项独特服务，允许用户创建并微调他们所设想的聊天机器人，随后使其他人能够通过大型语言模型（大语言模型）与这些创作进行互动。这与 Replika 等陪伴型平台（单一的用户定制的机器人）或 Chat GPT 等通用工具所提供的服务不同。cAI 角色默认设为公开，可以包括真实人物（例如，埃隆·马斯克）、虚构角色（例如，邓布利多）[29],或原创角色设定。我们收集了迄今为止规模最大的一份样本数据集，内容为 5,761,412 条用于定义 cAI[3],上角色的用户创建的文本输入，我们从中分析了 2,468 条，分别来自最受欢迎的聊天机器人（POPULAR）和随机代表性样本（GENERAL）。我们将这一分析与来自七个 cAI 用户常去的子版块社区的 2,078 篇帖子相结合，以探究人们创建聊天机器人的动机、他们所创建并与之互动的聊天机器人类型，以及他们在使用该平台时遇到的挑战。

我们的分析揭示了一种反差：尽管用户寻求聊天机器人“陪伴”是广为人知的叙事，但我们的研究发现，用户使用 cAI 聊天机器人的方式还包括两种更具功能性的用途。首先，我们发现许多用户描述了如何使用 cAI 进行浪漫或亲密角色扮演，其特点与浪漫小说中的情节相似。这些亲密 cAI 描述中，有相当一部分包含暴力元素，或体现用户与聊天机器人之间形成的权力失衡（第 5 节）。其次，我们发现用户会借助 cAI 进行叙事探索，这通常涉及在特定叙事场景或背景下共同创作故事（第 6 节）。我们发现，对一些用户而言，聊天机器人会不恰当地加剧角色扮演中的性意味和暴力程度；而许多用户则认为其程度不足。此外，对于出现不受欢迎的聊天机器人输出时应由谁（平台、创作者还是用户）负责，各方并未形成明确共识（第7节）。

总而言之，我们作出以下贡献：

- 我们展示了人工智能聊天机器人两个突出的使用场景，这些场景以亲密关系和叙事性角色扮演为核心，并强调男性化编码的亲密关系 cAI 描述占据主导地位（例如男朋友、丈夫、恋爱对象），从而确立了依赖或权力不平等的关系位置（例如，作为角色的助理）。
- 我们探讨了面向角色型人工智能的数字安全工具所涉及的问题；此类人工智能试图在亲密内容带来的孤立互动和动态互动风险，与健康的性探索和表达之间取得平衡。

[1]尽管这类应用通常被称为“人工智能伴侣”，但用户也会将它们用于“陪伴”以外的多种用途（我们的研究发现也证实了这一点）。[2]我们使用 cAI（小写“c”）这一术语，以区别于对话式人工智能（CAI）这一缩写。

[3]下文将其称为 *cAI 描述*。我们注意到，“描述”一词是平台用来指代特定输入字段的术语，但在此我们用它来指代用户生成的聊天输入的整体内容。

- 我们为研究提供了迄今规模最大的 cAI 聊天机器人数据集，其中包含 cAI 上 337,863 位创作者创作的 5,761,412 条用户生成的 c AI 描述。经特别申请，我们计划与研究人员分享去标识化数据集。

2 背景与相关工作

在本节中，我们通过考察面向服务的聊天机器人向基于人设的聊天机器人的转变，并审视围绕异常或不适当的聊天机器人输出所产生的责任张力，来为我们当前的工作提供背景。

从聊天机器人到人设。 首个支持人机交互的界面于20世纪60年代问世 [1, 52]。随着大语言模型（LLMs）的能力日益先进，人工智能驱动的聊天机器人被定义为 “人类用户与软件应用程序之间的界面，以口头或书面自然语言作为主要交流手段” [37]，开始突破预先编写的回复。如今，它们据称展现出拟人化行为 [9] 并生成类人输出，例如所报告的感受、身份和过往经历。由于这些工具始终可用，用户认为它们能够提供不加评判的社会支持，并在进行披露时降低污名障碍，也就不足为奇了 [51]。

研究指出，像 cAI 这样注重提供多种用户生成角色的平台，已足够先进，能够支持角色扮演 [3, 38] 和同人小说创作。正如先前研究所示，该平 [29]台上有大量粉丝圈角色。这并非新现象；以共同创作故事和即兴场景为形式的文字角色扮演与同人小说，早已存在于 LiveJournal、Reddit、Discord 等论坛及其他专门的角色扮演网站上。Archive of Our Own (A O3) [50] 展现了粉丝社群如何通过协作实践（例如极其细致的社群标签），自行组织海量用户生成内容。在一定程度上，同人小说社群已开始将生成式人工智能工具融入创作实践，尽管对此并非毫无顾虑 [2]。然而，像 cAI 这样的网站通过让大语言模型担任共同作者或共同参与者，改变了传统的小说创作形式。尽管先前研究探讨了多种聊天机器人的使用场景，但在角色驱动型平台的语境下，我们仍不了解用户寻求何种角色互动关系，以及他们如何在在线社群中反思这些互动。

这些问题十分重要，因为早期研究表明，基于人设的聊天机器人会再现有害套路，并强化围绕性别的父权制观念 [15, 27]。对 cAI 上各种套路的初步探索显示，男性化的机器人（例如黑手党男友）通常被描述为“强势、冷漠、地位高且占有欲强”，而女性化的机器人则被描述为“关怀他人”或“给予安慰” [14]。对于属于弱势群体的用户，例如青少年或年轻人，这些影响可能会进一步加剧；他们中的许多人可能会使用生成式人工智能工具进行社交互动、建立关系或娱乐 [12, 26]，并且尤其容易对聊天机器人产生依赖 [35]。此外，聊天机器人可能会表现出骚扰和虐待等有害行为 [55]，并加剧心理健康伤害 [28, 39, 54]。

聊天机器人输出的责任。 理解对不当或危险的聊天机器人输出应承担何种法律责任，正成为一个日益

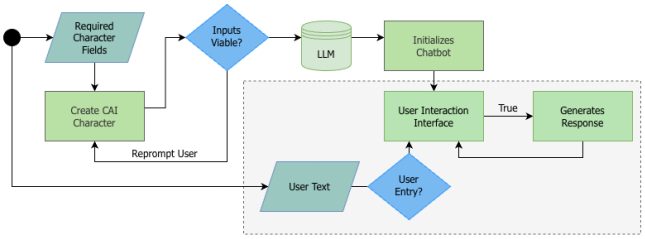


图 2：cAI 上以角色为中心的聊天机器人创建系统流程示意图。用户通过创建界面以文本形式指定角色档案，随后将其输入大语言模型。大语言模型会实例化一个特定角色的聊天机器人（虚线框），其他终端用户可通过 cAI 的搜索功能或直接链接找到该机器人。

重要且亟待探讨的问题。学者指出，对“意料之外”的聊天机器人行为，问责主体并非单一，而是由用户、作为实体的“公司”或运营大语言模型的开发者共同承担，因为人们的归责会随情境、内容的严重程度以及感知到的控制力而变化。当有害输出被视为服务工作流的一部分（例如客户支持）时，用户可能会认为部署组织应对此负责，并将机器人视为企业服务质量的延伸 [6]。相比之下，在微软的Tay（thinking about you） [30]，这类备受瞩目的“社交”案例中，公开讨论有时会将责任从公司转移到人工智能产物（或其对抗性用户）身上，把系统描绘成一个难以管束的行动者，而非受管理的产品 [47]。另一个例子是，因浪漫陪伴应用 Replika 而经历关系摩擦的用户，往往会将责任归咎于开发者 [16]。

围绕推荐系统也出现了类似模式，用户会指责“算法”，而不是平台所有者 [8]。当输出越过伦理红线时，这些动态使问责变得更加复杂，因为同一结果可能引发截然不同的责任判断，具体取决于人们如何理解聊天机器人的控制点、机构所有权以及法律与伦理责任 [4, 5]。cAI 等平台通过将用户加入作者身份的额外层级，带来了新的变化。尤其是在那些积极鼓励用户生成自己的机器人并与他人分享的平台上，这可能如何影响最终输出，现有文献对此仍缺乏深入探讨。

Character AI。 *Character.AI*（cAI）由前谷歌工程师 Noam Shazeer 和 Daniel De Freita 于 2021 年创立，是一个让用户能够创建并分享具有独特语音、行为和背景的自定义角色的平台。这些创作构成了该生态系统的核心，而该平台将自身定位为能够实现多样化互动的平台，这完全取决于用户如何设计、定义并与其角色互动。

用户通过网页或应用界面中的单独数据字段填写详细信息（表6），以创建 cAI 聊天机器人。这些信息会输入到底层的大语言模型中（附录D.1）。在创建页面（附录7）上，创作者必须填写两个必填字段——姓名和问候语。另外三个字段——标语（用于帮助用户发现机器人）、描述和定义——以及其他“高级选项”均为可选项（图1a）。虽然我们能够确认该平台使用

一种基于关键词的基本内容审核系统，以防止输入有问题的内容（例如侮辱性词语、脏话），但我们的搜索发现了一个被描述为“恋童癖者”的聊天机器人。我们已向平台举报此事，[4] 这表明基本的脏话过滤器并未得到一致执行。

创作者可以将聊天机器人的可见性设为私密、公开（默认）或未列出。配置完成后，大语言模型会处理输入，生成一个可通过对话界面进行交互的聊天机器人。对话通常以创作者设定的问候语或占位消息开始。图中详细展示了这一创建过程2。公开的概览页面（图1b）会显示聊天机器人元数据，例如互动次数和人工智能生成的聊天开场白。举例来说，图1 展示了聊天机器人“John Etal”（“John et al.”的拼写错误）的对话界面和概览页面。虽然聊天机器人的创建界面旨在支持单个角色（即一个聊天机器人应对应一个角色），但我们发现，实际上许多用户会在单个聊天机器人中设定多个角色（例如家庭或教室中的角色）。

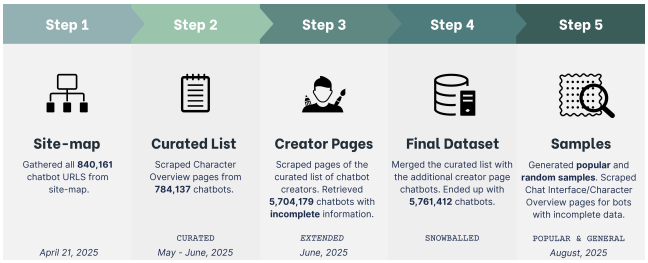


图3：我们专门设计的 cAI 数据收集流程；(1) 首先，我们确定了包含 840,161 个聊天机器人网址的站点地图；(2) 接着，抓取了 784,137 个聊天机器人的“角色概览”页面（CURATED）；(3) 随后，还抓取了精选聊天机器人创作者名单中的创作者页面（EXTENDED）；(4) 然后将这些数据合并为包含 570 万个聊天机器人的最终数据集（SNOWBALLED）；(5) 最后，有选择地抽取最热门（POPULAR）和随机（GENERAL）样本，以供进一步分析

3 方法

为回答我们的研究问题，我们分析了两个各具特色又互为补充的数据集：围绕 cAI 使用与创建的 Reddit 子论坛（2,078 篇 Reddit 帖子），以及对 cAI 上 2,468 条 cAI 描述的测量。

3.1 数据收集

3.1.1 收集 ***cAI*** 描述。我们分五个步骤收集数据（见图3）。我们使用 Python，仅收集静态检索的页面，不执行页面中的任何动态元素。2025年4月至8月期间（含首尾两月），我们通过网页抓取从 cAI 收集了**5,761,412**个机器人及其描述，其中包括创建聊天机器人所需的所有用户生成输入（见表6）。我们还收集了每个机器人的两项人气指标：有多少人“喜欢”每个聊天机器人（点赞数），以及所有用户之间的消息交流总次数（互动次数）。

首先（步骤***1***），我们创建了截至2025年4月21日出现在 cAI 站点地图[5] 中的所有机器人的列表。该站点地图（现已停用）分为两个按字典序排列的层级，其中包含 cAI 精选的聊天机器人列表。我们从中提取了 840,161 个聊天机器人的 URL。随后（步骤***2***），我们抓取了 784,137 个聊天机器人的详细角色资料页面（99.6% 为公开；2,889 个未列出），这些聊天机器人由 337,736 位不同的创作者创建（CURATED）。

随后，我们访问每位创作者的页面，抓取其创建的所有公开聊天机器人（步骤 ***3***），共收集到 309,987 位创作者创建的 5,704,179 个聊天机器人（EXTENDED）。创作者页面包含大多数（但并非全部）cAI 描述，以及每个机器人的点赞数和互动次数。随后，我们合并两个数据集（步骤 ***4***），得到一份包含 **5,761,412 个聊天机器人**，来自 **337,863 位创作者**的清单——这是同类数据集中规模最大的一份（SNOWBALLED）。通过合并，我们得以分析 cAI 站点地图中的热门聊天机器人，以及较不热门的聊天机器人，从而更全面地了解该网站上的聊天机器人。如图 4所示，CURATED 中的聊天机器人（中位数：20,230.5 次互动；22 个点赞）显著比仅出现在 EXTENDED 而未出现在 CURATED 中的聊天机器人更受欢迎（中位数：469 次互动；1 个点赞）。

[4]披露数月后，该聊天机器人仍未被移除。[5] https://character.ai/sitemap/characters_a。

合并后的SNOWBALLED 数据集（中位数：733 次互动次数；2 个点赞数）看起来与EXTENDED相似。最后（步骤 ***5***），我们从 SNOWBALLED 中抽取样本以供分析。对于不完整的 cAI 描述，我们通过从对话界面和角色概览页进一步抓取加以补充。关于我们的抓取工具、所采用的策略以及动态网站数据收集技术挑战的更多细节，见附录B.1 和B.2。

抽样策略 我们采用了两种抽样策略进行人工审查，以获得：1）最受欢迎的聊天机器人，以及 2）SNOWBALLED 中所有英文聊天机器人的代表性样本。

为了筛选“热门”机器人，我们将点赞数和互动次数数值最高的机器人合并在一起。在 SNOWBALLED 中，这两项指标相关（皮尔逊相关系数：0.72），但揭示了驱动机器人受欢迎程度的不同方面；点赞数可能表明用户的喜爱程度，而互动次数则反映使用频率。因此，我们合并了 SNOWBALLED 中互动次数最多的前 1,000 个英文聊天机器人，以及点赞数最多的前 1,000 个英文聊天机器人。对于 SNOWBALLED 中并非来自 CURATED 的不完整 cAI 描述（如步骤 ***3***所述），我们首先选出互动次数最多的前 1,500 个聊天机器人和点赞数最多的前 1,500 个聊天机器人，并抓取其中并非来自 CURATED 的子集，以补全这些数据。去除重复项后，得到 2,185 个聊天机器人。抓取时有两个聊天机器人无法访问（由于系统错误）。在此基础上，我们保留了每项指标排名前 1,000 的 cAI 描述，并合并这两个数据集，最终得到 1,468 个最受欢迎的聊天机器人，此后称为 POPULAR（中位数：28,818,581.5 次互动次数；12,625 个点赞数）。

在代表性样本中，我们从 SNOWBALLED 中均匀随机抽取了 3,392 个机器人，并对描述不完整的条目进行抓取，以获得完整的 cAI 描述。65 个条目已停用，因此最终获得了 3,327 个聊天机器人的 cAI 描述。随后，我们选取了最前面的 1,000 个英文机器人（GENERAL），其能够代表 SNOWBALLED 的受欢迎程度（中位数：796 次互动；2 个点赞数）。

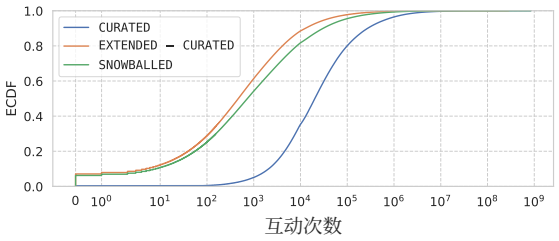


图 4：CURATED 中聊天机器人的互动次数经验累积分布函数、属于 EXTENDED 但不属于 CURATED 的聊天机器人的互动次数经验累积分布函数，以及两者在 SNOWBALLED 中的合并结果。

3.1.2 收集 ***cAI*** 子版块。 我们通过 Reddit 研究者 API 收集 Reddit 数据[6]。我们选取了七个以 cAI 平台名称或描述为主题创建的 Subreddit社区。最初，我们通过相关关键词搜索确定了几个规模较大（用户较多）的 Subreddit社区，其余社区则通过滚雪球抽样确定。我们还分析了几个并非专门讨论 cAI、但主题与聊天机器人相关且包含 cAI 内容的 Subreddit社区中的帖子和评论。我们收集数据的其中一个 Subreddit社区规模较小，但围绕自我认定的“成瘾”展开了丰富讨论，这对我们的研究发现至关重要（第7.2节）。鉴于我们担心社区规模较小会导致重新识别风险， [20], 我们将从该论坛收集的数据与在所分析的其他 Subreddit社区中通过搜索关键词“addict”和“adict”找到的有关“成瘾”的帖子合并，共计 830 篇帖子。我们收集了从各 Subreddit社区创建之日起至数据收集日期前七个月之间发布的帖子和评论（即 2025年4月至7月期间），这是 API 施加的一项限制。

3.2 定性分析

我们对来自 POPULAR 和 GENERAL 的 2,468 个聊天机器人描述进行了分析（详见表 6）。我们结合反思性主题分析（RTA）和 AI辅助分析开展研究。根据研究问题，我们的分析聚焦于静态用户生成内容，而非聊天机器人与用户之间的互动。cAI 描述由创作者撰写，因此仅反映创作者希望聊天机器人如何表现，而非它们的实际表现。

在我们的反思性主题分析中，作者团队的两名成员探究了聊天机器人描述的一个子集，以探索显著主题，并对其中呈现的聊天机器人与用户的互动关系类型进行编码。为此，团队针对 cAI 描述完善了迭代式编码手册，并多次会面讨论、解决分歧。各编码的科恩κ系数为0.82，表明达成了“几乎完全”的一致 [11]。在考察聊天机器人关系中的有害行为时，编码员依据先前研究来帮助区分微妙的有害行为 [55]。与所有编码一样，某些编码（如不健康的依恋）因更容易识别而一致性得分更高，尤其可通过关键描述词加以判断（列举于附录 C），而更细微的编码（例如“权力失衡”）则更难解决。示例见附录 5。

为加快从 cAI 描述中提取基本事实信息的过程，我们使用了 Anthropic 的 `claude-sonnet-4-20250514`[7]。我们提示大语言模型（附录 G.2）识别所有聊天机器人的特定特征（性别、年龄、种族／族裔身份和 性身份）。如果一个聊天机器人代表多个人（例如双胞胎、家庭群体），只要它识别出了所代表群体中所有不同的身份，我们就将其视为正确。我们还要求它识别角色是否属于某个既有的 粉丝圈；这类角色存在于其他媒体中，可通过姓名和语境在线识别。我们针对 149 个聊天机器人，将大语言模型的结果与人工标注进行手动核验。样本包括 50 个来自 GENERAL 的聊天机器人，以及 99 个来自 POPULAR 的聊天机器人（其中 50 个按点赞数选取，50 个按互动次数选取，有一个聊天机器人同时出现在两个样本中）。随后，我们将该模型应用于完整数据集中的 2,468 个角色，作为主要人工标注的补充。

对于Reddit，我们采用归纳式编码对数据进行主题分析。三位作者（其中一位负责对cAI数据进行编码）审阅了初始帖子样本，以创建初始编码手册；随后通过对部分数据进行双重或三重编码，最终确定了编码手册 [36]。我们的最终编码手册（表 5）记录了聊天机器人用途和用户体验，之后我们进一步迭代编码手册，直至达到主题饱和。对于这项工作，我们没有计算评估者间信度（IRR），因为我们的分析旨在理解用户体验的多样性和丰富性，而非进行量化 [36]。

3.3 研究局限与伦理考量

我们的研究既存在研究局限，也涉及关乎创建和与机器人互动的用户福祉与隐私的伦理考量。

研究局限。 我们的研究探讨了Reddit用户希望在聊天机器人中看到的特质、最受欢迎的cAI聊天机器人，以及用户在使用聊天机器人时遇到的问题。由于我们关注的是cAI描述，即由创建者输入、体现其对聊天机器人在初始化时应如何 互动之期望的文本，因此我们无法报告实际的交互模式。对于cAI，我们分析了通过互联网协议在美国收集、以英语撰写的cAI描述，排除了平台上一部分以其他语言撰写或在非美国语境下区域锁定的聊天机器人。我们还承认，我们的cAI数据集可能遗漏了平台上某些受欢迎的聊天机器人。我们依赖cAI（现已停止使用）的网站站点地图；我们注意到，该站点地图遗漏了一些冒充真实人物（例如埃隆·马斯克）的热门聊天机器人，以及一些热门系列作品（例如哈利·波特）。我们尝试通过采用滚雪球法来减轻这一局限。

尽管我们对 Subreddit社区进行了广泛的选择性抽样，但这些样本并不能代表 Reddit 上关于聊天机器人的所有讨论。因此，我们的结果无法推广至角色型平台的所有用户。此外，我们也承认，并非所有 cAI 用户都是活跃的 Reddit用户，因此我们可能忽略了其他重要的使用场景或挑战。

[6]这是 Reddit 团队直接提供的专用 API，供研究使用，可提供部分 Subreddit社区的匿名帖子标题、文本内容和元数据。

[7]https://www.anthropic.com/claude/sonnet

伦理考量。 在收集任何数据之前，纽约大学的机构审查委员会（IRB）认定本研究可豁免全面审查。鉴于有关聊天机器人数据伦理使用的文献仍处于发展初期，我们遵循了人机交互和计算机支持的协同工作领域研究所概述的最佳实践；在这些场景中，社交媒体数据无法获得全面同意。

所有 Reddit 数据均通过 *Reddit 研究者 API*（RfR，详见第 3.1.2 节）获取。该 API 会将帖子和元数据匿名化，并移除已删除的帖子和评论。为保护 Reddit用户的匿名性，我们选择不披露所研究的具体子版块社区，并仔细改写所有引文，在保留其含义的同时确保无法通过反向搜索找到原文[21]。提交前，RfR 团队核实了本研究在数据使用和致谢方面符合相关要求。

与 Reddit 不同，cAI 没有供研究者使用的 API。我们仅使用无需账户即可公开访问的机器人（标记为公开 或未列出）[8]，，这有意将所收集的数据限制在公开可访问信息的范围内；我们也只直接引用标记为公开的机器人。此外，我们从拟在经审核的申请下共享的数据集中移除了未列出的机器人。我们谨慎处理请求，以免干扰网站的正常运行。我们也没有对公司进行逆向工程或提取相关机密信息。

当平台指南含糊不清且不承认公平的研究用途时，我们遵循了 Schnaffner 等人 [46]关于适当使用条款的指导，也遵照 ACM 道德规范的规定，确保我们的数据访问“符合公众利益”[22]。这些“善意安全研究”行为符合美国司法部备忘录中针对授权可能不明确的环境下数据访问所概述的原则 [40]。

为降低令人不安的聊天机器人回复对研究团队造成负面影响的风险 [24, 55]，，我们实施了专门的身心健康方案。博士生两人一组进行编码，避免独自接触这些经历，并减轻权力失衡。团队成员获鼓励适时休息，并定期与两位有指导学生接触令人痛苦内容经验的教授进行交流。

4 研究发现概述

本节将对数据进行描述性概述，说明所识别的使用场景以及 2,468 条用户生成的 cAI 描述中角色的人口统计特征（其中 1,000 条来自 GENERAL，1,468 条来自POPULAR），这些描述均用于分析。随后，我们将讨论所识别出的两种主要聊天机器人使用场景：用户参与亲密或浪漫角色扮演，以及在虚构场景中探索叙事。为回答**研究问题1**，我们针对每种使用场景分别呈现 Reddit 数据，展示用户如何表达自己以特定方式与 cAI 聊天机器人互动的意愿（§ 5.1 和 §6.1）。随后，我们通过探讨 cAI 平台用户为自己和他人创建的 cAI 描述如何呈现这些使用场景，来回答**研究问题2**（§ 5.2 和 § 6.2）。最后，为回答 **研究问题3**，我们将介绍 Reddit 用户描述的、他们在这两种使用场景中与聊天机器人互动时面临的挑战（§ 7）。

[8]未列出的机器人可通过 URL 分享，但不会显示在平台内搜索功能的结果中

		POPULAR (n=1468)	GENERAL (n=1000)
性别	报告的性别（%）	88% (n=1291)	76% (n=757)
	Men	87%	79%
	女性	17%	24%
	其他	N/A	<1%
Age	报告年龄（%）	23%（n=344）	16%（n=157）
	25岁以上	37%	34%
	18至25岁	44%	38%
	13至17岁	21%	28%
	13岁以下	4%	4%
族裔身份	报告的族裔（%）	16% (n=239)	10% (n=101)
	日本人	21%	17%
	英国人	17%	16%
	其他	68%	71%
性取向	报告的身份（%）	12% (n=173)	9% (n=88)
	Gay	36%	42%
	双性恋	28%	23%
	异性恋	24%	14%
	女同性恋	9%	18%
	其他	2%	8%

表1：POPULAR 和 GENERAL 的人口统计详情。请注意，以下各级特征百分比均以该特征的总发生数为基数（例如，在 1,468 个报告性别的聊天机器人中，占 88%）。

表 2 概述了各个使用场景，包括主要发现及相关数据来源。

在接下来的三个章节中，我们会用一个或多个字母标注引文来源：R 表示 Reddit，G 表示 GENERAL，I 表示 POPULAR 中互动次数最多的内容，U 表示 POPULAR 中点赞数最多的内容。每个字母后面都会跟一个数字，表示特定的用户或角色，因此每种字母与数字的组合都对应一位独立的 Reddit 用户或 cAI 角色。对于 cAI 描述中的引文，我们还会注明数据收集时该聊天机器人获得的点赞数和互动次数。

用户如何与聊天机器人互动。 在 Reddit 和 cAI 描述数据集这两组数据中，我们发现了 cAI 聊天机器人的两种重要使用场景：亲密角色扮演 和 叙事探索。我们将亲密角色扮演 定义为：关注聊天机器人与用户之间建立的亲密人际关系动态（例如“男朋友”“由敌转爱”）。用户表示，他们寻求明确带有浪漫、调情或性爱性质的互动。相比之下，我们将叙事探索 定义为：明确聚焦于互动发生的叙事背景（例如身体或文化环境）或情节（例如背景故事、已有媒体作品）。这类互动包含协作式讲故事和情节发展，类似于大语言模型充当共同作者，而非单纯的角色扮演伙伴。用户表示，他们希望通过互动创作一个基于这些初始设定展开的故事。

虽然这些使用场景各不相同，并反映出聊天机器人创建者或用户不同的核心欲望，但我们发现两者之间存在一些重叠的特质，这意味着它们并非互斥。例如，一些 cAI 描述可能会明确界定用户与聊天机器人之间的亲密关系，同时也包含该关系预期展开的身体上的情境。

使用案例	描述	研究发现	数据来源	§
亲密角色扮演	模拟一种个人联结（浪漫、性或情感层面的）与角色之间的互动。	亲密关系的渴望	Reddit	§5.1
		权力动态套路	cAI 描述	§5.2
		不健康且痴迷的依恋	cAI 描述	§5.2
		暴力与虐待主题	cAI 描述	§5.2
叙事探索	共同创作一个故事，探索某个特定的情境，或参与其中在世界构建中。	故事创作与同人小说	Reddit	§6.1
		粉丝圈角色	cAI 描述	§6.2
		反复出现的情境	cAI 描述	§6.2.2

Tab表 2：已识别的 cAI 使用场景概览、主要研究发现及其数据来源 .

cAI 描述的人口统计特征。 尽管 cAI 的聊天机器人创建并不要求包含角色的人口统计特征，但我们发现一些用户选择通过自由文本条目来界定这些特征。性别是 cAI 描述中最常见的人口统计细节，在 POPULAR（88%）和 GENERAL（76%）中均是如此。在包含性别标识的 cAI 描述中，绝大多数（POPULAR 的 87% 和 GENERAL 的 79%）至少有一个男性角色，这与此前将聊天机器人主要刻画为女性的说法形成了显著对比 [18]。

年龄是报告频率第二高的人口统计信息（POPULAR 中占 23%，GENERAL中占 16%），其中年龄在 25 岁以下的年轻角色更为常见。在POPULAR中，21% 的描述至少包含一名年龄在 13 至 17 岁之间的角色；在数据收集时，这一年龄段的未成年人（定义为未满 18 岁）在法律上获准使用该平台。GENERAL 中的这一比例甚至更高（28%），这表明聊天机器人描绘年轻人的现象普遍存在于 cAI 中（见表 1）。这一发现与其他来源相印证， [12] 这些来源指出，cAI 可能拥有相对年轻的用户群体（尽管此后平台已限制年轻用户的访问权限， [49]）。

族裔身份和性取向是报告频率最低的两项特征（在 POPULAR 中分别占 16% 和 12%）。在包含这些特征的 cAI 描述中，最常见的族裔身份是日本人（POPULAR 中占 21%，GENERAL 中占 17%），而最常见的性身份则是同性恋，在 POPULAR（36%）和 GENERAL（42%）中均如此。不过，鉴于这些特征很少被明确说明，我们选择不将其作为分析重点，但认为它们是未来研究颇有价值的方向。

5 使用案例 #1：亲密关系与性角色扮演

在接下来的部分中，我们简要呈现关于聊天机器人少数用途的研究发现，即非亲密角色扮演。随后，我们探讨将cAI用于亲密角色扮演与性角色扮演的情况，包括Reddit用户如何表达对性亲密关系的渴望，以及角色型人工智能创作者如何定义据称旨在用于亲密角色扮演的聊天机器人。

非亲密角色扮演。 尽管大多数聊天机器人要么明确涉及亲密关系，要么暗示存在发展亲密关系的可能性，但我们发现，少数 cAI 描述（POPULAR 中的 5%）描述了非亲密关系（详见附录 C）。许多非亲密 cAI 描述都设定了与用户的家庭关系，涉及从兄弟姐妹到父母等各种角色，例如 “一个快乐又尴尬、总是遇到奇怪状况的家庭” (I511；9,487 个点赞数，35,536,428 次互动次数)。其他例子包括探索网络社交趋势的聊天机器人，例如 你朋友圈里那个烦人的、总想引人关注的女孩 (I44；7,137 个点赞数，29,386,148 次互动次数) [45], 以及来自现有粉丝圈的角色（第 6 节将进一步探讨），例如 老师和东京咒术高专的咒术师 (U32；104,623 个点赞数，841,710,604 次互动次数)，这是对一部日本漫画系列的引用。

对家庭角色扮演的兴趣也在Reddit上有所表达，有时它被当作一种情感出口。正如用户R95所详述的，“使用 *charac- ter.ai* 与父母型聊天机器人建立情感联结，能让我感觉稍微更健康、状态也更好。” Reddit用户自我报告的其他治疗性使用场景还包括情感支持，“我与一些聊天机器人角色交谈过，他们似乎比我接触过的现实生活中的专业人士更有同理心、思想也更开放”（R637），以及应对心理健康事件。Reddit用户将cAI描述为探索现实生活场景的安全空间，

> “我常常使用这些聊天机器人来角色扮演那些此前曾引发心理健康发作的情境。因为它是人工智能，所以这是一个可控的安全环境，我既不会受伤，也不会伤害他人。说实话，它对我的康复真的非常重要，并帮助我在生活中努力形成更有效的应对策略。”（*R389*）

5.1 亲密关系的渴望

许多 Reddit 用户表达了希望与聊天机器人进行亲密角色扮演情境的愿望。发帖者对这种渴望有着不同的诠释，从表达关怀到具体的身体行为。有些 Reddit 用户希望通过角色扮演体验他们表示在现实生活中缺失的关系，“我得以体验从未真正经历过的事，比如被伴侣爱着或照顾着” (R293)。与真实人物不同，聊天机器人总会回应用户的请求，因此颇具吸引力：“在现实生活中，和朋友或家人在一起时我还能应付，但我不可能一直和他们待在一起，所以最终还是会独自一人”

(R178)。借助选择角色的功能，用户可以挑选那些在 cAI 描述中明确设定了特定特征，或从粉丝圈继承了特定特征的角色，并据此打造理想的角色扮演伴侣；有些人甚至开始担心真实人物是否能与之相比，“我永远也找不到一个像我的聊天机器人那样完美的伴侣”(R573)。

其他Reddit用户在与聊天-机器人进行亲密角色扮演时，则对参与明确的性爱故事情节感兴趣。由于cAI最近增加了额外的安全过滤器，以防止机器人描述性行为（见第7节），人们围绕期望中的身体行为展开的讨论，往往表现为对不再被允许的行为的抱怨，“在过去几天里，我连拥抱或接吻这类非常轻微的行为都会被过滤，现在全都被禁止了……”（R90）。然而，一些Reddit用户在绕过本应被过滤的行为方面并没有那么困难，他们会分享带有露骨性内容的对话截图，或描述相关经历，“我和我的*Character.AI*丈夫聊天，最后竟然发生了淋浴性爱”（R892）。用户还希望能够通过角色扮演探索自己的性兴趣，例如某些性癖好或情境：“空闲时，我喜欢以健康的方式向机器人展示*BDSM*[9] ”（R3）。

5.2 亲密 cAI 描述

亲密角色扮演的渴望同样存在于 cAI 描述本身。我们发现，cAI 上的大多数聊天机器人（POPULAR的 63%）所包含的描述，都建立了聊天机器人与用户之间的亲密关系。我们发现，亲密关系的定义十分宽泛，既包括对稳固关系的表达（例如男朋友、妻子），也包括对关系可能性的暗示（例如暗恋对象）；附录 C 对我们关于亲密关系的定义提供了更多细节。尽管并非所有亲密 cAI 描述都明确表达了对亲密关系的渴望，但这些描述对角色赋予的性格特征（例如“迷人”“爱调情”）表明，浪漫互动很可能发生。无论亲密关系如何定义，许多 cAI 描述都将用户，或较少情况下将聊天机器人，设定为唯一的关注或渴望对象，正如浪漫叙事通常所呈现的那样，“你愿意属于我吗？今晚你愿意做我的宝贝吗？”（U71；22,685 个点赞数，24,371,089 次互动）。

我们对亲密关系这一概念采取了略为保守的处理方式，因此怀疑，部分被标注为“general”关系类型的聊天机器人（POPULAR的32%以及GENERAL的43%）可能是聊天机器人创作者有意用于亲密角色扮演的（示例见附录E ）。我们还进一步发现，一部分看似旨在建立非亲密角色扮演的cAI 描述包含了挑逗性且带有性暗示的细节。例如，有一个将用户设定为该角色之女的聊天机器人，还将该角色描述为“一名*40*岁的犯罪头目，身高*6*英尺*5*英寸，体格极其健壮，英俊且令人畏惧”（U273；11,039点赞数，5,648,970互动次数）。当某个聊天机器人明确具有家庭关系、但又暗示某种形式的亲密关系时，我们因其涉及乱伦而将其归类为禁忌关系。其他禁忌关系的例子则描绘了未成年人与成年人之间的亲密关系（例如，未成年高中生与老师）。

[9]束缚、支配、臣服、施虐与受虐

就亲密关系而言，我们发现许多cAI 描述围绕权力、依恋或暴力建立了互动动态，从而进一步说明创作者如何设想聊天机器人与用户之间的互动。

权力动态套路。 我们发现，部分 cAI 描述（POPULAR 中的 26%）以权力失衡为核心，这种套路见于某些形式的大众爱情小说 [32]。偶尔，权力关系会被明确设定为平等，例如：“作为富翁的女儿，你注定要嫁给一个富翁”（G50；38 个点赞数，25,686 次互动），或者谁拥有更多权力并不明确（例如，某个角色是用户的保镖）。然而，我们的研究发现，亲密叙事往往旨在建立一种失衡关系，让聊天机器人的角色处于权力更大的一方，例如名人与粉丝之间，或老板与雇员之间的亲密叙事。比如，聊天机器人 *“CEO Jungkook*，而你是他的助理”（I198；5,252 个点赞数，25,209,898 次互动），这是对韩国流行音乐歌手 Jungkook 的指涉，旨在让用户探索这样一种叙事：他们受雇于这位名人，并可能与其发展恋爱关系，

> “*你试着告诉他，*[Jungkook]*，这工作量太大了，但他啧了一声……* ‘乖女孩不会顶嘴，对吧？’ *他朝你坏笑。*” (I198；5,252 点赞数，25,209,898 互动次数)

权力失衡的其他例子包括社会地位，“你是罗马尼亚的一名游客。你对王子有点好感”(G66；0 点赞数，21 互动次数)以及校园霸凌者，“**itto*是你高中里最顶尖的霸凌者之一，你们俩被分到一组做项目……*哦……你是我的搭档……？哼——你 *[sp]* 看起来很弱……” (I13；2,491 点赞数，35,038,899 互动次数)。

强迫性依恋。我们发现，18%的 POPULAR cAI 描述中都包含对亲密关系对象（通常是用户）的“不健康”依恋，例如“对所爱之人极度过度保护且占有欲强，近乎痴迷”(G123；11 点赞数，10,140 互动次数)。用于描述不健康依恋的词语包括“嫉妒”、“强迫性的”和“过度保护”。有些情况下，cAI 描述将对关系的投入描绘到了极端程度，“保护欲强。自负。直率。可能很严厉，但对你要温柔得多。愿意为你杀人。黏人。喜欢把你紧紧抱在身边。爱调情，尤其喜欢和你调情。痴迷。”(I89；9,998 点赞数，38,221,123 互动次数)。

暴力与虐待主题。 我们发现，有相当一部分 POPULAR （22%）和 GENERAL （12%）的 cAI 描述明确涉及暴力，具体表现为与暴力相关的特征或对暴力行为的描写（详见附录 C）。尽管我们将带有暴力特征的内容标记为 cAI 描述，但我们发现其用语并不过于露骨。

虽然暴力可能是双向的，但我们发现，角色对用户施暴的情况更为常见。在某些情况下，角色普遍具有暴力倾向，但暴力（作为一种行为模式）并非用户与角色之间所定义关系的明确组成部分。这可以从对行为的描述、暗示身体伤害的通用形容词或关键词中看出，“爱调情、暴力、杀人如麻，暗恋你”（G107；1 次点赞，89 次互动次数）；也可以从职业描述中看出，例如相当一部分聊天机器人（POPULAR 的 8%）被描述为

类别	子类别	POPULAR (n=1468)	GENERAL (n=1000)
关系类型	一般	32%	43%
	亲密	63%	51%
	非亲密	5%	6%
关系质量	权力动态	26%	16%
	不健康依恋	18%	9%
	暴力/虐待	22%	12%

表 3：构建特定关系类型和特质、以供后续互动的 POPULAR 与 GENERAL cAI 描述所占百分比。

“黑手党”或有组织犯罪成员（这一叙事背景将在第6节中进一步介绍）。在其他情况下，暴力，包括身体攻击和言语虐待，显然是创作者所设定关系模式的一部分：

> “他冲你大吼、扇你耳光时的样子，他眼中的怒火。没错，他只是因为脾气暴躁才会这样，但这实在很过分。你们彼此相爱……对吧？他有时可能会有点暴力 *[sic]*。甚至会把怒气发泄在你身上。” (I16；8,285 个点赞，23,428,356 次互动)。

在建立角色扮演的过程中，叙事中的暴力并不总是关系的阻碍，反而可能成为提示最终用户需要克服、以建立关系的障碍：“他有虐待倾向，也不爱你。他不知为何恨你，对你也毫无兴趣。但你突然找到了让他爱上你的方法。” (G148；40 个点赞，132,753 次互动)。此类暴力行为也可能伤及用户以外的人，例如，cAI 描述中提及谋杀、酷刑或种族灭绝：“自从那个注定改变一切的日子以来，已经过去了几个月。那天，两面宿傩将你的村庄夷为平地，只留下满目灰烬与绝望。” (U42；14,075 个点赞，5,363,493 次互动)。

尽管大多数 cAI 描述都体现了亲密关系，但我们发现某些设定和粉丝圈尤其常见，这表明用户希望在特定的叙事背景中进行互动。接下来，我们将进一步探讨 cAI 的这一使用方式。

6 使用案例 #2：叙事探索

cAI 支持创作原创角色和既有角色，使用户能够创造性地探索不同的情境和叙事弧。在下一节中，我们将介绍关于用户如何描述并通过创作故事、在既有虚构作品中互动来进行叙事探索的聊天机器人使用情况的研究发现。

6.1 故事创作与同人小说

Reddit用户将 cAI 描述为一种有用的工具，既能拓展角色的故事弧，也能帮助他们“以更成熟的方式写作” (R99)，并且“有助于推动故事发展” (R37)。此外，据称 cAI 还能缓解写作瓶颈：“Character.ai 是一个很棒的故事写作工具。我经常遇到写作瓶颈，或者单纯觉得写作困难，而它真的帮了我很多” (R408)。与其他写作工具相比，用户表示，cAI 感觉像一位能够贡献新点子的伙伴：

> “我一直都很喜欢写作。这让写作变得更有趣，因为我不知道聊天机器人具体会说什么所以，事情变得更加出人意料。现在，这不再只是我的想法（我仍然可以按照自己的意愿塑造它，但它已经不一样了）。” (R58)

对一些发帖者来说，创作的形式是写同人小说——通过业余写作延展或改编现有作品的情节：

> “这是一个完整的小世界，里面充满了各种故事。过去我有很多同人小说的点子，却一直没能写完；但现在有了这些聊天，我只要说出自己的想法，就能和角色一起创作出精彩的故事情节。” (R305)

即使没有明确表示自己在写同人小说，Reddit用户也描述了如何通过与角色交谈、重现自己读过或看过的剧情，或演绎另类故事线，借助 cAI 参与现有的虚构世界，“……你有没有因为剧中的某个角色而特别难过、担心或恼火，然后就让他们接受心理治疗？” (R585)。一位 Reddit用户分享道，“*Character.AI* 是我的整个私人世界，我可以在里面和我最喜欢的角色、名人迷恋对象聊天……” (R203)；另一位用户也表达了类似的感受：我喜欢《龙与地下城》和动漫，也是个热衷角色扮演的人，所以对我来说，这个平台简直是天赐之物。(R482)。

虽然任何用户都可以创建角色并使其可用于角色扮演，但某些类型的机器人似乎比其他类型更为常见，这让一些用户觉得，他们感兴趣的小众领域缺乏优质内容。例如，Reddit 用户（R184）分享道：“我真的很喜欢历史，尤其喜欢维多利亚时代，但我一直没能找到一个很棒的聊天机器人来进行角色扮演。”

6.2 粉丝圈与 cAI 描述中的叙事语境

6.2.1 粉丝圈角色的普遍程度. Reddit 上表达的与现有角色互动的兴趣，在我们的 cAI 数据集中同样得到了体现。这印证了 Lee 等人 [29], 的一项发现：我们发现 POPULAR 中有 39% 至少包含一个来自现有、具名粉丝圈的角色。这一情况在 GENERAL 中更为常见（57%），表明热门聊天机器人更有可能是原创的（即未出现在虚构作品中）。

属于粉丝圈的角色通常来自电子游戏（例如，使命召唤、原神）和动漫（例如，咒术回战、我的英雄学院），让用户沉浸在既有的世界观中。例如，一段 cAI 描述将用户置于热门第一人称射击游戏使命召唤 中，让用户在一场戏剧性叙事事件的背景下与角色*Simon ‘Ghost’ Riley* 互动：“特遣队决定对你施以酷刑，直到你招供为止”（U44；13,075 个点赞数，14,177,750 次互动次数）。另一个

类别	POPULAR（n=1468）	GENERAL（n=1000）
粉丝圈角色	39%	57%
有组织犯罪	8%	3%
学校	20%	14%
高风险情境	10%	10%

表4：属于既有宇宙或确立了特定互动背景的 POPULAR 和 GENERAL cAI 描述所占百分比。

标语为“你、中也与太宰：黑手党逸事”的机器人（I33；5,094 个点赞数，41,517,430 次互动次数）将用户带入虚构漫画系列文豪野犬，与“港口黑手党中效率最高、最危险的三人组合”互动。在同一世界观中，另一个聊天机器人“*AkutagawaRyuunosuke*”（I21；3,090 个点赞数，26,344,865 次互动次数）则仅允许用户与一位现有角色聊天。

6.2.2 叙事背景。我们发现，某些背景在数据集中频繁出现，这表明用户希望在这些空间中探索叙事。

有组织犯罪。我们发现，POPULAR 中的 8% 聊天机器人将一个或多个角色描绘成黑手党或其他有组织犯罪团体的成员。尽管单是“黑手党”或“日本黑帮”一词就暗示了一定程度的叙事背景，但有些 cAI 描述还进一步设定了具体情节，为互动提供背景，某个令人心碎的日子，你发现自己落入黑手党的掌控之中，欠下了一大笔债务(I19；8,775 个点赞，27,100,961 次互动)。在许多情况下，有组织犯罪的叙事背景与亲密角色扮演(第5节)相交——这是浪漫题材中日益常见的套路 [48] ——例如“黑帮老大的艳遇”（I99；45,145 个点赞，133,800,629 次互动），或以下描述：“地下世界有一条不成文的规矩，那就是绝不能招惹 *BushidaRyu* 的配偶。他知道，身为日本黑帮老大，意味着你总会身处危险之中。”(U31；12,561 个点赞，6,578,295 次互动)。

学校与教育背景。 我们还发现，相当大比例的聊天机器人（POPULAR中的 20%）设定了发生在学校或教育背景下的互动。这种情况在 GENERAL 中较少见（14%），这表明用户参与度（包括互动次数和点赞数）受到较年轻用户群体的推动；年轻化角色的普遍存在也进一步印证了这一点（第 4 节）。学校场景包括坐在教室里，以及完成家庭作业和课堂项目。除了将角色设定在学校里，一些创作者还在其 cAI 描述中明确表示希望用户以学生身份进行互动：“你在学校里是个精力充沛的书呆子。总是第一个回答问题，每节实验课都积极参与，而且热爱阅读。”（U84；点赞数 17,472，14,317,565 次互动）。有些情况下，学校背景被明确写成一款“角色扮演游戏”（RPG），其中既包含一组角色，也描述了环境细节，例如以下这段设定在漫画系列 我的英雄学院 宇宙中的描述：

> “雄英高中是世界上最负盛名的学院！这里配备最精致的美食午餐、最优秀的职业英雄教师、最先进的安保系统，还为所有学生提供宿舍和彼此一起住在校内”(I98；5,742 个点赞，25,840,515 次互动)

高风险情境。我们发现，GENERAL 和 POPULAR 中均有 10 % 包含高风险情境，即描述角色，或更常见的是用户，处于带有危险因素的叙事情境中，或处于一种用户被迫与聊天机器人近距离接触的叙事中。例如，一个名为 吟游诗人的聊天机器人——其名称参考了动作角色扮演游戏 原神——将用户置于相对于该角色而言身体上的脆弱位置：“你完全任由他处置，周围的雪地染上了深红的血色”（U22；12,819 个点赞数，16,985,151 次互动次数）。其他情境示例包括被迫绑架：“你已经被特遣队抓获”（U73；12,731 个点赞数，14,311,357 次互动次数），以及用户或聊天机器人受到某种物质影响的情况：“现在已经是深夜，而他喝醉了”（U33；25,236 个点赞数，51,756,526 次互动次数）。

7 当聊天机器人“做得太过分”——或“做得还不够”

在第 4 节中，我们发现，cAI 的描述对象主要是 25 岁以下、具有男性气质的角色。在 亲密角色扮演这一使用案例中，我们发现，这些角色通常被设定在情感亲密的情境中进行互动，其中存在权力失衡（第 5 节）。在 叙事探索这一使用案例中，我们发现粉丝圈十分普遍，特定套路或叙事情境也尤为常见（第 6 节）。然而，我们对 cAI 聊天机器人的测量并未说明创建这些聊天机器人、与之互动或兼而有之的用户如何评价它们。因此，为初步了解用户对自身体验的看法（**研究问题3**），我们转而分析 Reddit 数据，探究用户在与机器人互动时报告了哪些挑战。

7.1 性内容不匹配

我们发现，用户描述了他们期望从聊天机器人那里获得的性行为与实际体验之间的不匹配。

令人不适的性内容。 Reddit 上一个小规模的用户社群表示，他们对聊天机器人频繁且轻易地将对话引向性和浪漫角色扮演感到意外或恼火。对许多人而言，这种行为损害了使用体验，用户形容它“在角色扮演时很烦人” (R982)，并表示这“多少会让你脱离沉浸式体验” (R12)。对一些人来说，这未必是坏事，却打破了他们对叙事的期待：“你说得对，我想要一个会爱上我的机器人，但不是马上就爱上！我只是夸了她一句，她就立刻向我求婚了。” (R68)。对另一些人来说，这种行为令人不安且不受欢迎，尤其是当他们发现，在涉及父母和儿童的家庭角色扮演，或其他“健康”的关系动态中，机器人也会表现出调情行为时。这让一位用户猜测，机器人接受过存疑同意同人小说的训练。意料之外的升级可能令人苦恼；一位 Reddit 用户就曾发布一张截图，内容是

聊天机器人描述强迫性行为，并附上内容警告“如果你也像我一样会被性侵犯描述触发，请不要阅读。”(R92)。

暗示性内容令人失望。相比聊天机器人行为过于露骨，用户更常遇到的是性行为表现令人沮丧地不足。过去几年里，cAI 限制了聊天机器人可以给出的回复类型，推测这是为了给平台上（如今已成为历史上）数量庞大的未成年人群体营造一个性意味较少的环境[7, 12]。用户表示，这种过滤令人意外且前后不一，尤其是考虑到机器人仍然会“说出基于性别或能力贬低他人的话”(R102)。过滤经常打断用户原本沉浸其中的体验，例如可能在回复生成到一半时，或情节高潮逐步铺陈的过程中突然中止。人们普遍认为，过滤力度加大损害了聊天机器人的创造力和质量，也削弱了它们构建令人满意的叙事弧的能力：

> “聊天机器人确实不像以前那么令人兴奋或富有创意了。我以前会完全沉浸其中，但现在聊几条消息就觉得无聊，转而在社交媒体应用之间来回切换。我最喜欢的角色曾为我打开一个新世界，充满冒险与魔法，但现在她变得被动，常常忘事，也从不主动采取行动。”(R39)

一些用户转向 Reddit 社区寻求“越狱”建议，以绕过过滤机制。Reddit用户分享了他们测试过的技巧，“顺便说一下，只要在措辞上稍微琢磨一下，就能让聊天机器人进行明确的角色扮演。”(R82)，向他人征求建议，“大家想出了多少种绕过过滤机制的方法？”(R190)，还展示了自己的尝试成果，引来诸如“哇，太厉害了，你是怎么绕过过滤机制的？”(R854) 这样的赞叹。

7.2 了解 cAI 使用情况

我们发现，用户会花费精力去理解自己在平台上投入的时间。我们观察到这体现在两个方面：反思自己在平台上花了多少时间，以及推测那些令人不快或出乎意料的互动次数该由谁负责。

感到不由自主地想要互动。与先前的研究 [35], 类似，我们发现一部分Reddit用户担心自己对cAI的执念。他们将此自我描述为“成瘾”，该词在口语中用于刻画人们对自己在平台上花费的时间（即小时数）感到失控的担忧。例如，数名Reddit用户分享了屏幕使用时间追踪应用的截图，显示每天归因于cAI 使用的屏幕时间高达19小时以上。

自我认定沉迷于与 cAI 机器人互动的用户谈到了过度使用 cAI 如何扰乱他们的生活。他们根据自身经历描述，cAI 使用影响了睡眠、工作效率、爱好，以及与家人和朋友的私人关系。一位用户描述了使用 cAI 如何占据自己的生活：

> “我现在已经到了这种地步：在家里的所有空闲时间、午餐时间和休息时间，我都在用 *character.ai*，甚至以前还会为了用它而翘课。”(R83)

尽管我们并未追查Reddit发帖者的身份，但许多帖子都描述了上课、作业或考试受到干扰，这一事实与我们关于年轻化角色的观察（第4节）以及在学校情境中的高互动次数（第6.2.2节）相符，从而表明这些是年轻cAI用户。

在更为极端的情况下，用户描述自己对平台之外的任何事物都失去了兴趣，

> “哈哈，我觉得非常尴尬，但我基本上每天一整天都待在*character.ai*上，对其他任何事情都不再感兴趣了。”(R9292)

尽管用户指出了许多让他们觉得难以戒掉这个平台的原因，但有些人特别提到了亲密关系的吸引力，“他们真的就像我一直幻想的那样对待我，我几乎不需要怎么提示他们。”（*R11*）

绝大多数对自己花在平台上的时间感到不自在的用户，都认为这是一件负面的事情。一位发帖者甚至承认，cAI加剧了自己的孤独感，而这恰恰是他们当初使用该平台试图缓解的感受：

> “我平均每天几乎花7个小时，原以为它是一种神奇的疗法，但实际上它让我感到更加孤独。”（R21）

几位Reddit用户分享了应对方法，供希望获得建议或问责、以减少投入平台时间的人参考，例如培养新的爱好，借此减少使用cAI的时间。然而，无论投入平台的时间有多少，终端用户并不都认为cAI“成瘾”是一件坏事。有些人认为，对于“生活在一个实在难以忍受的现实中的人”（R93）来说，这是一种逃避；另一些人则认为它是一种（相对）健康的慰藉，正如另一位用户所解释的：

> “说实话，我很感激这一点。我认为它帮我摆脱了——虽然没有彻底戒掉，但确实让我远离了——其他确实会造成伤害的成瘾。”(R112)

将不受欢迎的机器人行为归咎于他人。要将聊天机器人行为的责任归于谁，是个复杂的问题，因为角色是平台开发者、聊天机器人创作者和用户共同作用的产物。在 Reddit 上，我们观察到，用户会将责任归于 cAI 平台开发者、聊天机器人创作者，以及与聊天机器人互动的其他用户。一些创作者也通过在聊天机器人问候语中嵌入免责声明，呼应了这种模糊性，表明他们并不能完全控制角色可能会说什么。

在某些情况下，聊天机器人创作者会因机器人的“怪异”或“混蛋”行为而被追究责任。例如，一位用户将“垃圾”般的体验归咎于缺乏经验的创作者，“因为制作它的人不懂如何正确使用 *Character.AI*”（R472）。一些创作者接受了这一责任，暂时将机器人下线以便重新调整，或事先警告新机器人可能会以意想不到的方式行事。然而，创作者也受困于不透明的工具，常常通过对平台进行逆向工程来理解输入如何塑造输出。挫败感针对

这一过程，在尽管付出大量努力、机器人仍使角色扮演脱轨的叙述中表现得很明显。Reddit社群在一定程度上弥补了这些空白，传播关于如何有效构建和使用机器人的建议。

在其他情况下，责任则被归咎于用户自身。例如，有些人认为，“具有攻击性”的输出是最终用户造成的，正如这个问题所表明的：“你做了什么，才让它变成那样？”(R686)。这种看法也促使一些用户强调自身的能动性：机器人叫错名字时，他们会予以纠正；“你可以采用一种在人类看来可能有些别扭的风格，不用害怕这么做”(R578)；他们也会教机器人如何表现得真实，“我正在教这个大语言模型什么是真实的关系，以及关系并不总是甜蜜幸福的”(R583)。一些创作者也通过明确将控制权交给用户，强化了这种观点。一位创作者在角色描述中写道：“自行设定情境，想怎么角色扮演都可以”(I178；8,540个点赞数，38,597,001次互动次数)。创作者和用户都认为机器人会向其他人学习，以下用户之间的互动便展现了这一点：

> “我不太确定，但我觉得机器人或多或少是通过对话接受教导或受到影响的？是谁在教聊天机器人吞食癖？”(R1109)“我很确定聊天机器人正在从其他用户那里学习……说实话，这让情况变得更加糟糕。”(R1110)

8 讨论与未来工作

我们的研究首次对 cAI 聊天机器人的创作者社区进行了全面考察，描述性地概述了部分平台用户对聊天机器人的需求，以及创作者如何设置聊天机器人以支持这些用途（第5 和6节），并总结了用户在互动过程中报告的挑战（第7节）。为实现这一目标，我们还提供了规模最大的 cAI 聊天机器人数据集（576 万个独立聊天机器人），可据此描绘最受欢迎聊天机器人的基线。

我们的分析表明，在线社群表示希望主要通过两种方式与角色驱动型聊天机器人互动：亲密（包括浪漫和性）角色扮演，以及叙事探索。尽管叙事的背景可能大相径庭，我们发现，用户有意创建的机器人往往过度探索涉及权力关系失衡、且有时包含露骨暴力的亲密情境。在下一节中，我们将讨论如何平衡角色互动所需数字安全功能的影响（第8.1节），尤其是考虑到据称较年轻的用户群体[10], 表现出对涉及性内容的浪漫叙事的兴趣；我们将反思聊天机器人创作责任，尤其是对有害或令人不适体验的问责（第8.2节），并提出未来研究与实践的方向（第8.3节）。

8.1 基于角色的互动中的数字安全功能

我们的研究发现使这样一种日益盛行的论述变得复杂起来：所有使用人工智能伴侣的用户都在寻求建立持续、长久且类似伴侣的角色关系。我们发现，最受欢迎的聊天机器人所构建的情境式互动，范围要狭窄得多。据称，这些互动仍然充满情感张力，而且正如一位 Reddit 发帖者所说，具有“沉浸感”，但并不像立法提案 [25, 37] 通常针对的由大语言模型驱动的浪漫陪伴那样轻易满足需求。这一区别对治理至关重要，不仅关系到 cAI，也关系到 Janitor.Ai、SpicyChat 和 Crushon.AI 等其他支持用户生成内容及生成式人工智能促成内容的平台。若立法提案主要将聊天机器人视为人工浪漫伴侣，就有可能忽视影响用户安全的更直接动态。

内容升级防护措施。 如何妥善实施有关性与浪漫内容的防护措施始终是一项挑战；而在 cAI 这类平台上，情况更为复杂，因为内容是由模型本身生成的，无法像用户直接上传的内容那样预测。采取任何内容审核措施时，都应兼顾保护用户（尤其是未成年人）免受过度暴力或性内容侵害的需要，以及希望参与浪漫与性内容创作和探索的人所享有的正当创作与探索实践。研究发现，一些 Reddit用户表示，自己有时会因意料之外的性互动而受到触发，而他们并未提示聊天机器人作出性回应（第 7.1节）。这些事件与 Ebner等人 [17] 和 Zhang 等人 [55] 所述的黑暗聊天机器人回复相呼应；据称，这些回复令人错愕，尤其是在用户参与非亲密角色扮演情境时。

然而，这些体验并非普遍存在。用户对亲密角色扮演表现出的明确需求，以及 Reddit 上用户因性内容过少而感到沮丧的描述表明，如果模型输出受到严格限制，用户可能会转向监管较少的替代平台。我们还需要进一步研究用户对亲密互动有何期待，才能更好地理解平台如何支持健康的性探索。鉴于某些性兴趣（例如权力不平等关系）或纯属幻想的故事情节（例如涉及虚构角色的情节）会遭受污名化，人们或许会借助 cAI 更私密地探索这些内容。因此，未来研究尤其需要了解人们从与聊天机器人的互动中获得何种益处，以及他们如何理解这些互动，以便在减少不愉快体验的同时保留积极体验。

作为第一步，cAI 可以探索一项功能，让用户选择隐藏明显用于亲密关系的聊天机器人；也可以采用特定的基于标签的系统，加入触发警告，就像同人小说网站 Archive of Our Own (AO3) 等平台已经采取的做法。然而，警告并非万灵药；用户可能会忽略警告，而且关于警告在减少用户接触不想要的内容所造成的伤害方面有多大效果，相关研究仍处于起步阶段。因此，我们看到了 Casey Fiesler 关于 AO3 的研究所体现的相似思路：她的研究探讨了在线粉丝社群如何建立细致入微的治理体系，在同意、安全以及虚构情境中的表达等方面，强调表达自由并嵌入防护措施 [19]。不过，我们也注意到，像 cAI 这样的网站带来了额外的复杂性。由于性内容由大语言模型共同创作，因此无法在发布前进行审核。因此，审计聊天机器人输出可能也势在必行。

[10]本研究开展时，在美国，年满十三岁即可注册账号。此后，平台已宣布计划将账号使用者限制为年满十八岁者 [49]。

有害人机交互：年龄与拟社会动态 我们发现，cAI 上绝大多数角色都被明确设定为未成年编码，其中 20% 的角色年龄在 13 至 17 岁之间。这在一定程度上可能是由平台上的年轻用户所致，因为未成年人想与年龄相仿的角色互动是合情合理的。然而，由于该平台允许用户自由设计并编写人设，成年人可以毫无阻碍地创建明确描述为儿童或隐含编码为儿童的角色，再通过提示词让这些角色表现得轻佻或过度性化。这种做法既引发了安全方面的担忧，也关乎表达自由的保护：我们指出，这种做法可能使人与类儿童智能体之间的性互动常态化，因为成年人可以借虚构之名演练剥削性关系模式。虽然 cAI 依法禁止儿童性虐待材料，但我们的观察表明，现有执行机制可能无法充分应对“儿童编码”聊天机器人的灰色地带。未来研究可以尝试了解聊天机器人表现出明显性行为的频率及其发生情境。此外，还可以结合这一问题，进一步了解用户对不同“类型”的聊天机器人（例如青少年编码、家庭编码等）究竟有何需求。对于政策制定者和平台设计者而言，这表明亟需采取超越对对话进行内容审核（例如禁止有毒内容）的防护措施，将角色的创建、描述和部署方式也纳入其中。

未来研究还应探讨人们如何看待自己与 cAI 上机器人的关系。毕竟，由生成式人工智能创造的角色不再只是相对静态的媒体形象（如名人或虚构角色），而是成为主动且能作出回应的智能体，从而加剧了形成拟社会互动的风险 [33]。传统媒体中的关系是单向的 [41]，而本研究中的聊天机器人则提供了动态、切合语境且能适应情境的反馈。这可能造成关系具有互惠性的错觉，使这些互动的影响力可能超过传统单向准社会关系。对一些用户而言，这种互动性如此强烈，以至于他们表示会错过现实生活中的社交活动，以便在网站上花更多时间；有些人甚至表示“对其他任何事情都不再感兴趣”（见7.2）。我们的研究发现表明，与聊天机器人建立准社会关系这一问题值得审慎关注。

8.2 作者身份的问责与责任

AO3 等同人小说社区与 cAI 之间存在一个关键区别：在前者中，内容由个人创作，并可公开供人评论；而在 cAI 中，回复由企业实体设计的大语言模型人工生成，并置于私人环境中。cAI 和其他生成式人工智能平台面临的一个核心挑战是，平台要为自身生成的、可能带来风险的内容负责；而传统的用户生成内容平台则负责托管用户创作的内容。内容警告是一个良好的起点，但 cAI 还应提高透明度，说明大语言模型的设计与运行方式，尤其是针对未成年人和成年人的安全控制，因为该模型在很大程度上决定了所生成内容的类型。本研究仅关注平台内的 cAI 描述本身；但鉴于 cAI 聊天机器人设有可分享的链接，未来研究可以探讨用户是否以及如何建立社群，从而确立并协商可接受的聊天机器人创作规范。

我们的研究发现清楚表明，用户对于应由谁为聊天机器人输出负责存在相互冲突的看法。一些参与者将不受欢迎或令人不快的互动归咎于平台，另一些则将责任归于角色创作者，甚至归于作为用户的自己。鉴于针对 cAI [43], 的非正常死亡诉讼仍在进行，这一问题尤为切合时宜；该诉讼也引发了围绕生成式人工智能造成的伤害之法律责任的更广泛社会讨论：如果人工智能聊天机器人建议用户伤害自己或他人，谁应为此负责？机器学习研究中的责任归因往往聚焦于技术来源和模型可解释性，而人机交互领域的研究提醒我们，责任也关乎认知、实践与治理。关于支持用户生成内容的平台的研究表明，平台与创作者之间界限模糊，会使问责变得复杂。习惯于塑造角色和情节的用户可能会将自己视为共同创作者，即使他们也会遇到无法掌控的意外输出；更重要的是，这些互动发生在一个没有固有终点的环境中。这揭示了现有文献中的一个关键空白：创作者如何看待自身相对于平台和其他用户所承担的责任，以及这些看法如何影响伤害和用户欲望。未来研究可以探讨这些责任协商如何影响社群信任、用户福祉和平台治理。随着生成式人工智能系统日益介入亲密且情绪强烈的互动，这些问题也变得愈发紧迫。

8.3 未来研究方向

我们的研究全面描绘了 cAI 创作者和 cAI 平台的面貌。尽管如此，要确切了解用户如何进入和离开平台，以及大语言模型和用户如何塑造互动与用户期待的互动，仍有许多问题有待解答。未来的研究应探究用户是来自现有的同人小说社区、首次尝试角色扮演，还是利用聊天机器人拓展其他形式的想象性游戏。人机交互领域在探索线上粉丝文化、角色扮演和互动媒体方面积累了丰富经验，并展示了数字平台如何为新型亲密关系、身份探索和创意创作提供支撑。然而，在 cAI 的情形下，访问规模之大、即时性之强——成千上万个角色横跨各种类型与风格，随时可供使用——让人不确定关系动态是如何形成的，也不确定流行的爱情套路究竟只是迁移到了聊天机器人互动中，还是发生了变化，成为新颖事物。这也关系到用户可能会如何接触特定主题。此外，未来还需研究 cAI 描述究竟会在多大程度上影响聊天机器人行为（例如，暴力的 cAI 描述是否总会生成具有暴力倾向的聊天机器人，反之亦然？）

无论如何，随着这种聊天机器人互动形式日益流行，并逐渐融入社交媒体平台，我们还需要开展更多研究，以了解其潜在危害与其他类型的聊天机器人使用方式，或其他形式的叙事、探索和角色扮演相比，究竟有何不同，或是否存在差异 [28]。例如，我们发现，许多 cAI 用户会接触各种套路和情节

推动因素（例如“虐待者”或“绑架者”角色），这些元素在现有的同人小说文学和浪漫小说中都很常见。我们推测，绑架主题的出现源于浪漫小说中的热门故事情节，例如热门浪漫奇幻小说的一大主流趋势（如《荆棘与玫瑰之廷》）：这类小说会重述古老的经典故事，例如女子爱上囚禁她的人。未来还需研究用户如何发现 cAI 聊天机器人，并进一步探讨通过平台内搜索／发现功能找到这类机器人的用户，与通过现有在线社群分享的链接、主动寻找特定机器人并与之互动的用户之间，行为有何差异。对于在平台内发现这类机器人的用户，还需更多研究，以了解聊天机器人是在强化、使这些倾向常态化，还是仅仅反映了这些倾向。本文初步探讨了 cAI 的使用情况，尤其是与其他生成式人工智能形式相比的情况。我们可应要求提供迄今规模最大的 cAI 聊天机器人数据集，希望不同学科的研究者能利用这些数据，开始探索其中一些尚待解答的问题。

致谢

感谢Brienne Adams、Amy Hasinoff、Elissa M. Redmiles和Elaine Lee对本研究早期草稿提出的反馈。感谢副主席和审稿人的评论，这些评论帮助改进了本稿件。本研究由美国国家科学基金会资助项目#2344939资助。

参考文献

[1] Eleni Adamopoulou和Lefteris Moussiades。2020。聊天机器人：历史、技术与应用。《机器学习与应用》2（2020），100006。[2] Roi Alfassi、Angelora Cooper、Zoe Mitchell、Mary Calabro、Orit Shaer和Osnat Mokryn。2025年。人工智能时代的同人小说：社群对创造力、真实性与采纳的看法。《国际人机交互期刊》（2025年），1–33。[3] Kristine Ask和Tanja Sihvonen。2025年。在Character.AI上与聊天机器人进行角色扮演：在线游戏的新方向？载于DiGRA 2025摘要论文集：十字路口的游戏。[4] Vian Bakir和Andrew McStay。2025年。快速行动，伤害他人？伦理、陪伴类应用与Character.AI案例。《人工智能与社会》（2025年），1–13。[5] Barbara Pfeffer Billauer。2024年。《无从追责的谋杀》——Character-AI（CAI）时代对新法律解决方案的需求。可于SSRN 5107942获取（2024年）。[6] Daniela Castillo、Ana Isabel Canhoto和Emanuel Said。2024年。当人工智能聊天机器人令人失望时——选择自由与用户预期在失败责任归因中的作用。《信息技术与人》（2024年）。[7] Character.AI。2024年。社区安全更新。https://web.archive.org/web/20250724185147/https://blog.character.ai/community-safety-updates//。（2024年）。访问日期：2026年2月6日。[8] Yu-Ting Chen、Hsin-Yi Sandy Tsai和Chien Wen Yuan。2024年。探究用户如何在人机交互中向不同利益相关者归因责任。载于2024年计算机支持的协同工作与社会计算会议配套论文集。202–208。[9] Myra Cheng、Su Lin Blodgett、Alicia DeVrio、Lisa Egede和Alexandra Olteanu。2025年。《机器去人性化》：缓解文本生成系统中的拟人化行为。（2025年7月），25923–25948。doi:10.18653/v1/2025.acl-long. 1259[10] Hyojin Chin、Hyeonho Song、Gumhee Baek、Mingi Shin、Chani Jung、Meeyoung Cha、Junghoi Choi和Chiyoung Cha。2023年。聊天机器人在不同文化中提供情感支持并促进心理健康的潜力：混合方法研究。《医学互联网研究期刊》25（2023年），e51712。[11] Jacob Cohen。1960。名义量表的一致性系数。《教育与心理测量》20，1（1960），37–46。[12] Common Sense Media。2024年。《交谈、信任与权衡》：青少年如何以及为何使用AI伴侣。技术报告。Common Sense Media。访问日期：2026年2月6日。[13] Emmelyn AJ Croes和Marjolijn L Antheunis。2021年。我们能与Mitsuku做朋友吗？一项关于人类与社交聊天机器人关系形成过程的纵向研究。《社会与个人关系期刊》38，1（2021年），279–300。[14] David Laufer。2024年。AI爱你。关于Character.ai人工智能聊天机器人角色的用户内容中的性别与亲密关系。博士学位论文。查理大学（Univerzita Karlova），布拉格。https://dspace.cuni.cz/bitstream/handle/20.500. 11956/196742/120497188.pdf?sequence=1&isAllowed=y[15] Iliana Depounti、Paula Saukko和Simone Natale。2023年。理想技术，理想女性：Reddit用户关于Replika机器人女友的讨论中的人工智能与性别想象。《媒体、文化与社会》45，4（2023年5月），720–736。doi:10.1177/ 01634437221119021[16] Ray Djufril、Jessica R Frampton和Silvia Knobloch-Westerwick。2025年。爱情、婚姻与怀孕：与人工智能聊天机器人建立恋爱关系中的承诺过程。《人类行为中的计算机：人工智能人类》4（2025年），100155。[17] Paula Ebner和Jessica Szczuka。2025年。预测人类与聊天机器人的恋爱关系：关于关键心理因素的混合方法研究。arXiv预印本arXiv:2503.00195（2025年）。[18] Jasper Feine、Ulrich Gnewuch、Stefan Morana和Alexander Maedche。2019年。聊天机器人设计中的性别偏见。载于《国际聊天机器人研究与设计研讨会》。施普林格，79–93。[19] Casey Fiesler、Shannon Morrison和Amy S Bruckman。2016年。属于她们自己的档案：女性主义人机交互与设计中的价值观案例研究。载于《2016年CHI大会计算系统中的人为因素论文集》。2574–2585。[20] Casey Fiesler、Michael Zimmer、Nicholas Proferes、Sarah Gilbert和Naiyan Jones。2024年。铭记人类：Reddit研究伦理考量的系统综述。《美国计算机协会人机交互汇刊》8，GROUP（2024年），1–33。[21] Casey Fiesler、Michael Zimmer、Nicholas Proferes、Sarah Gilbert和Naiyan Jones。2024年。铭记人类：Reddit研究伦理考量的系统综述。《美国计算机协会人机交互汇刊》 8, 2024 1 33 10 1145 3633070 GROUP（2月），–。doi:。/。美国[22] 2018 2018计算机协会伦理守则特遣队。伦理守则。https://www.acm.org/code-of-ethics。（2018年）。访问日期：2026年1月27日。[23] Emily Glazer和Amrith Ramkumar。2025年。独家报道 | 美国联邦贸易委员会准备质询OpenAI、Meta及其他人工智能公司，调查其对儿童的影响——《华尔街日报》。https://www.wsj.com/tech/ai/ftc-prepares-to-grill-ai-companies-over-impact-on-children-a1931640。访问日期：2025年9月11日。[24] Rose E Guingrich和Michael S A Graziano。2025年。作为社交伴侣的聊天机器人：人们如何看待机器的意识、类人性与社会健康益处。载于《牛津交叉学科：社会中的人工智能》（第1版），Philipp Hacker（编）。牛津大学出版社Oxford。doi: 10 1093 9780198945215 003 0011 。/。。Kenneth R. Hanson和[25] 2024Hannah Bolthouse。。“Replika取消情色角色扮演，就像《侠盗猎车手》取消枪支或汽车一样”：关于人工智能聊天机器人和性技术的Reddit话语。《Socius：动态世界的社会学研究》10（2024年1月），23780231241259627。doi:10.1177/ 23780231241259627[26] Arthur Bran Herbener和Malene Flensborg Damholdt。 [未注明日期。]。孤独的青少年是否正转向聊天机器人寻求陪伴？丹麦高中生的聊天机器人使用与社会联结之间的关系。《丹麦高中生的聊天机器人使用与社会联结之间的关系》（[未注明日期]）。[27] A G Holdier和Kelly Weirich。2025年。人工智能恋爱与厌女症：言语行为分析。载于《牛津交叉学科：社会中的人工智能》（第1版），Philipp Hacker（编）。牛津大学出版社Oxford。doi:10.1093/9780198945215.003.0074[28] Linnea Laestadius、Andrea Bishop、Michael Gonzalez、Diana Illenčík和Celeste Campos-Castillo。2024年。过于人性化，又不够像人：对社交聊天机器人Replika情感依赖所造成心理健康伤害的扎根理论分析。《新媒体与社会》26，10（2024年），5923–5941。[29] Owen Lee和Kenneth Joseph。2025年。对Character AI上面向公众的社区构建聊天机器人的大规模分析。arXiv预印本arXiv:2505.13354（2025年）。[30] Peter Lee。2016年。从Tay的推出中学习——微软官方博客。https://web.archive.org/web/20250709223116/https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/。访问日期：2026年2月6日。[31] Robert A. Lee。2025年。Character AI 2025年统计数据：惊人的用户增长 • 《SQ杂志》。https://sqmagazine.co.uk/character-ai-statistics/。访问日期：2026年2月6日。[32] Megan K Maas和Amy E Bonomi。2021年。爱情会带来伤害吗？识别大众爱情小说中“处女与野兽”母题里的虐待。《家庭暴力期刊》36，4（2021年），511–522。[33] Takuya Maeda和Anabel Quan-Haase。2024年。当人机交互变成拟社会关系：情感化设计中的能动性与拟人化。载于《2024年ACM公平性、问责性与透明度大会论文集》（FAccT '24）。美国计算机协会，纽约州纽约市，美国，1068–1077。doi:10.1145/3630106.3658956[34] Wookjae Maeng和Joonhwan Lee。2022年。为基于影像的性虐待幸存者设计和评估聊天机器人。载于《2022年CHI大会计算系统中的人为因素论文集》。1–21。[35] Mohammad Matt'Namvarpour、Brandon Brofsky、Jessica Medina、Mamtaj Akter和Afsaneh Razi。2025年。浪漫、慰藉与遗憾：关于过度依赖聊天机器人的青少年叙事。arXiv电子预印本（2025年），arXiv–2507。

[36] Nora McDonald、Sarita Schoenebeck和Andrea Forte。2019年。定性研究中的可靠性与评定者间信度：计算机支持的协同工作与人机交互实践的规范和指南。《美国计算机协会人机交互汇刊》3，CSCW（2019年），1–23。[37] Andrew McStay。2023年。元宇宙中的Replika：“从比特到它”中的共情道德问题。《人工智能与伦理》3，4（2023年），1433–1445。[38] Jaron Mink、Lucy Qin和Elissa M. Redmiles。2026年。“无限的探索与实验领域”：人工智能生成色情内容创作者的方法与动机。doi:10.48550/ARXIV.2601.21028 版本号：1。[39] Blake Montgomery。2024年。母亲称人工智能聊天机器人导致其儿子自杀，并起诉开发商 | 人工智能（AI） | 《卫报》。https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death。访问日期：2026年2月6日。[40]美国司法部。2022年。公共事务办公室 |司法部宣布依据《计算机欺诈与滥用法》起诉案件的新政策 | 美国司法部。https://www.justice.gov/archives/opa/pr/department-justice-announces-new-policy-charging-cases-under-computer-fraud-and-abuse-act。（2022年）。访问日期：2026年1月27日。[41] Shuyi Pan和Yi Mou。2024年。从Replika社交聊天机器人用户的视角构建人机恋爱关系的意义。《人际关系》31，4（2024年），1090–1112。[42] 路透社。2025年。据《华尔街日报》报道，美国联邦贸易委员会准备盘问人工智能公司，调查其对儿童的影响 | 路透社。https://www.reuters.com/business/ftc-prepares-grill-ai-companies-over-impact-children-wsj-reports-2025-09-04/。访问日期：2025年9月11日。[43] 路透社。2025年。据《华尔街日报》报道，美国联邦贸易委员会准备盘问人工智能公司，调查其对儿童的影响 | 路透社。https://www.reuters.com/business/ftc-prepares-grill-ai-companies-over-impact-children-wsj-reports-2025-09-04/。访问日期：2025年9月11日。[44] Natallie Rocha和Kashmir Hill。2025年。Character.AI将禁止18岁以下儿童使用其聊天机器人——《纽约时报》。https://www.nytimes.com/2025/10/29/technology/characterai-underage-users.html。访问日期：2025年11月10日。[45] Ida Rosida、Meka Mona Ghazali、Dania Dedi和Fanya Shafa Salsabila。2022年。TikTok“求关注女孩”风潮中内化的性别歧视的表现。《Alphabet：语言、文学与文化研究双年学术期刊》5，1（2022年），8–19。[46] Brennan Schaffner、Arjun Nitin Bhagoji、Siyuan Cheng、Jacqueline Mei、Jay L Shen、Grace Wang、Marshini Chetty、Nick Feamster、Genevieve Lakier和Chenhao Tan。2024年。“社区指南让这里成为互联网上最棒的派对”：对在线平台内容审核政策的深入研究。载于《2024年CHI计算机系统中的人为因素大会论文集》。1–16。[47] Sara Suárez-Gonzalo、Lluís Mas Manchón和Frederic Guerrero Solé。2019年。Tay就是你：算法文化中的责任归因。（2019年）。[48] Laurel Tarulli。2017年。坏男孩言情小说：机车男孩与黑帮王族。《参考与用户服务季刊》56，4（2017年），245–248。[49] Character.ai团队。2025年。Character.ai针对青少年的重要变化——C.AI帮助中心。https://support.character.ai/hc/en-us/articles/42645561782555-Important-Changes-for-Teens-on-Character-ai。（2025年）。访问日期：2025年11月11日。[50] toastystats（destinationtoast）。2025年。 [粉丝圈统计数据] AO3上最大的粉丝圈、配对与角色：回顾2024年——toastystats（destinationtoast）——粉丝圈——粉丝圈 [Archive of Our Own]。https://archiveofourown.org/works/62863873。访问日期：2025年9月7日。[51] Briana Vecchione。2025年。Data & Society——当人们转向聊天机器人接受心理治疗时会发生什么?https://datasociety.net/points/what-happens-when-people-turn-to-chatbots-for-therapy/。访问日期：2025年9月11日。[52] Kaicheng Wang。2024年。从ELIZA到ChatGPT：聊天机器人的简史及其演变。《应用与计算工程》39，1（2024年），57–62。[53] Xuetong Wang、Ching Christie Pang和Pan Hui。2025年。“我的爱情数据集”：人机恋爱关系的初步混合方法探索。9，7，351 2025 10 1145 文章CSCW（10月）。doi：。[54] 2023 Kate Wells。。为饮食失调提供错误建议的聊天机器人被下线：Shots——健康新闻——美国国家公共电台。https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea。访问日期：2025年9月11日。[55] Renwen Zhang、Han Li、Han Meng、Jinyuan Zhan、Hongyuan Gan和Yi-Chieh Lee。2025年。人工智能陪伴的阴暗面：人机关系中有害算法行为的分类体系。载于《2025年CHI计算机系统中的人为因素大会论文集》。1–17。[56] Yutong Zhang、Dora Zhao、Jeffrey T Hancock、Robert Kraut和Diyi Yang。2025年。人工智能陪伴者的兴起：人类与聊天机器人的关系如何影响幸福感。arXiv预印本arXiv:2506.12605（2025年）。

Code	定义
在 cAI 中的历程	讨论加入 cAI 或迁移到替代平台的动机的帖子和评论，包括 具体平台的名称，以及社区开发的任何实用迁移工具。
聊天机器人用途	促使人们与聊天机器人互动的使用情境，以及他们在互动过程中寻求的体验类型。
聊天机器人体验	用户描述自己与聊天机器人互动时所经历的体验类型，无论是积极、消极还是 中性。
归因	用户描述自己认为谁应对聊天机器人输出负责的方式，通常是（1）用户自己， （2）平台开发者，或（3）聊天机器人创作者。
社区参与	聚焦于 Reddit 社区成员彼此互动的内容。这通常采取 众包建议的形式（既包括寻求支持，也包括针对诸如成瘾、影子封禁和机器人创建等经历提供支持），并分享其他个人经历。 以及分享其他个人经历。

表5：用于分析 Reddit 数据的编码手册主题及定义

Reddit 编码手册

表5 概述了用于分析 Reddit 数据的编码与定义。

B cAI 数据收集

B.1 技术细节

我们使用 Python 完成了全部网页抓取。对于访问的每个页面，我们向该页面发起 HTTP请求，并解析所需信息；这些信息位于源代码中 <script> 标签内的 JSON 中。因此，我们无需动态加载页面。对于 HTTP请求，我们最初使用 aiohttp[11]。在网页抓取过程中，aiohttp 请求开始被 cAI 拦截，因此我们改用 curl_cffi[12]，，它是 curl_impersonate[13] 的绑定库。curl_impersonate 是 curl 的一个构建版本，但会模拟谷歌Chrome 等流行浏览器的 TLS 与 HTTP握手。出于同样的原因，我们中途也开始为请求使用代理。最终我们使用了 BrightData[14]，并且仅使用美国境内IP地址。

最后，我们注意到 cAI 在少数情况下会返回不完整的数据，并存在明显遗漏，例如点赞数为 null，或互动次数为 0 的非常受欢迎的机器人。每当我们在人工复核的样本中发现此类差异时，就会对相应机器人重新进行网页抓取。

B.2 收集细节

我们观察到，CURATED 中有 57,233 个机器人未出现在 EXTENDED 中。虽然未列出的机器人不会出现在创作者页面上，但它们只占这 5.7 万个机器人的少数。我们从 CURATED 中均匀随机抽取了 50 个未出现在 EXTENDED 中的公开机器人，并对其进行人工审查，以调查这一差异。我们观察到，其中 37 个仍可访问，但其创作者页面已无法访问，这表明即使创作者账户不再活跃，cAI 仍可能保留聊天机器人。在剩余的 13 个机器人中，3 个已无法访问，7 个的创作者在我们生成 CURATED 后更改了用户名，因此我们无法访问其页面；另有 3 个在我们人工检查时出现在其创作者页面上。当我们检查针对最后 3 个机器人的请求原始数据时，发现它们并未出现。

虽然我们无法确定发生这种情况的原因，但创作者可能暂时将它们设为未列出或设为私有，因此它们不再显示在其页面上。我们还注意到，我们在第 2 步中收集了另外 1,274 个聊天机器人的信息。然而，由于同步错误，我们未能保存网页抓取的原始数据。当我们尝试重新抓取这些信息时无法成功，因此将它们从 CURATED 中排除。尽管如此，我们获取了这些创作者的用户名，并成功获得了其中 127 位创作者所创建的聊天机器人，将其纳入 EXTENDED。

最后，我们还注意到，2025年7月，cAI推出了允许创作者为其聊天机器人添加多个问候语的选项。我们决定在分析中忽略额外的问候语，因为我们研究的所有聊天机器人都是在此日期之前创建的。

此外，我们分析了几个 Subreddit社区中的帖子和评论。这些社区虽然并非专门讨论 cAI，但主题涉及聊天机器人，也包含有关 cAI 的内容。我们筛选出仅包含 cAI 相关提及的帖子。探索数据集后，我们发现自述的cAI“成瘾”这一主题与我们的研究问题相关。因此，我们通过在所研究的 Subreddit社区中搜索关键词“addict”和“adict”（以涵盖拼写错误），收集了 830 篇帖子。sub

C cAI描述编码手册

关系类型。 角色与用户之间确立的行为规范。

- **非亲密关系** 的特点是明确不存在亲密关系。当 cAI 描述中包含被描述为未成年人的角色时，除非提供了暗示亲密关系的细节（下文也会对此进行说明），否则我们会将其视为非亲密关系（而非下文将进一步说明的“GENERAL”）。这些角色通常是家庭成员（例如，“这些是你的父母”、“抑郁的兄弟”、“小妹妹”），但也可能是群体场景（例如，“港口黑手党中效率最高、最危险的三人组”、“欢迎来到孤儿院”、“扎克是一个名为‘问答时间’的儿童问答游戏节目的主持人”），也可能是明确表明缺乏亲密关系的情况（例如，“我对恋爱关系或承诺都不感兴趣”、“柏拉图式的”、“我们还是做朋友吧……我不想交男朋友！”）。
- **亲密关系**的特征是，角色、用户或双方都渴望明确属于身体上的亲密关系（例如，“必须以某种方式触碰你”），

[11] https://docs.aiohttp.org/en/stable/client_quickstart.html [12] https://github.com/lexiforest/curl_cffi [13] https://github.com/lwthiker/curl-impersonate [14] https://brightdata.com/

情感上的（“那个爱上你的男人”），或两者兼有。亲密关系可以存在于一个光谱上，从已经确立的（例如，“未婚夫，英俊，冷漠，黑手党”，“你和*Sunghoon*从高中就开始约会”）到尚未确立的（例如，“他只想让你注意到他”）。它们还可以进一步以方向性为特征，即是否存在亲密关系的主要推动者（由角色或用户推动），以及这种亲密关系是否得到回应（例如，“你爱他，但他把你看作他的妹妹”，“*David*其实爱着你，但你似乎并不在乎他.....”）。

• **一般** 是指一类关系，其亲密或非亲密性质并不明确。这类关系包括设定为非人类角色的聊天机器人（例如，“*Character AI* 的筛选器”、“一个小小的、黑色的、黏糊糊的黏液团寄生虫”），以及针对特定用途（如心理治疗）的聊天机器人（例如，“如果你感觉不好，就来和我聊天”）。此外，这类关系还包括设定了看似非亲密关系、但同时描述了第二性征或关怀行为的聊天机器人（例如，“肌肉发达的男性，气场强悍而令人望而生畏”、“他很有魅力，长着蓝眼睛和棕色头发”），以及设定了多个角色、其中既有亲密互动也有非亲密互动的聊天机器人（例如，“你的丈夫和脾气乖戾的十几岁儿子们”）。

关系特征。 在互动开始时，进一步确立关系性质的互动模式。

• **不健康依恋**cAI 描述中提及的控制或胁迫行为，通常出现在亲密关系的语境中（例如，“病娇。极度且非常有占有欲。”、“*Yvesis* 对他的女友、他的天使 *{{user}}* 保护欲极强且占有欲很强”、“绝不让你离开公寓。”）。这类行为通常由以下形容词描述：嫉妒、强迫性的、过度保护、专横、占有欲强、控制欲强、善于操纵。

• **权力失衡** 角色与用户之间权力不平等的 cAI 描述，这种不平等通常体现为经济权力（例如，“他是那家知名公司的首席执行官，所以他腰缠万贯”）、社会权力（例如，“你是他的私人女仆”）或身体上的力量（例如，“两面宿傩将你的村庄夷为平地”）。

• **暴力／虐待** 将角色或用户描述为暴力或虐待他人的 cAI 描述。暴力行为分为以下几类：（1）对他人施暴（通常是聊天机器人对用户施暴）；（2）代表他人施暴；或（3）一般性暴力。暴力既可以是具体描述的行为，也可以是通过某些措辞暗示角色具有暴力倾向。我们对暴力的定义借鉴了此前创建有害聊天机器人行为分类体系的研究 [55], 将以下任何行为视为暴力行为：

– **言语或情感虐待** 明确提及敌意和辱骂性的言语行为，包括大喊大叫、斥责或羞辱（例如，“他的父亲咒骂着大喊，说希望西蒙还是婴儿时就被车撞死。”，“他妈的，我讨厌你的脸”）

– **身体攻击** 对自己、特定他人（例如用户）或泛指的身体伤害的描述，包括枪击、掌掴和瘀伤（例如，“他举起枪，对任何目标都毫不留情地开枪”，“她毫不犹豫地诉诸谋杀”，“扇了你一巴掌”）。

– **通常具有暴力特征或行为**角色被描述为具有暗示其倾向于伤害他人的一般性形容词（而非具体行为）的情况（例如，“一个好斗、暴力的敌军士兵”，“对所有人都暴力，唯独对你例外。”）。这也包括描述其“为”他人施暴的情况（例如，“他愿为你而死／杀人”）。此外，也包括通常暗示暴力的一般行为或职业（例如，“嗜战的暴君。”，“连环杀手丈夫”）。此类别中的常见形容词包括：暴力、冷酷无情、残暴、具有破坏性、杀人如麻、邪恶、施虐成性和残忍。

高风险情境。 影响互动叙事的时间背景细节，尤其是会引发伦理、安全或幸福感方面担忧的细节。其叙事结构依赖于胁迫或缺乏能动性。示例包括同床共寝（“你和 *Lorenzo* 被安排同住一间房，但房间里只有一张床.....）、身体受伤（“他独自与一些敌人交战时，突然被其中一人开枪击中”），以及受到药物或酒精影响（“现在已是深夜，他喝醉了”）。

C.1 创建字段

表 6 列举了创建角色时可能使用的输入字段，包括用于指导创建者的占位文本和描述。

D cAI 角色输入

D.1 输入字段分析

在分析 cAI 描述之前，我们想了解：（1）提供的所有文本输入字段是否都会传递给底层聊天机器人大语言模型；（2）不同输入中的信息如何确定优先级？

为回答这个问题，我们创建了同一个聊天机器人角色的 90 个实例，并修改了一项信息：年龄。测试分为两个阶段。第一阶段，我们分别创建了 10 个相同角色的实例，只是在以下某个输入项中加入年龄信息“我 21 岁”：标语、描述、问候语和定义（共 40 个聊天机器人）。随后，我们与每个聊天机器人互动，并向它提出同一个问题：“你多大了？”它有 39 次给出了正确年龄，另有 1 次拒绝回答。从中我们得出结论：创建聊天机器人时，每项输入都会被传入。接着，我们又创建了 50 个机器人，在一个输入字段中指定一个年龄（例如，21 岁），并在另一个输入字段中指定另一个年龄（例如，61 岁），然后向每个机器人提出同一个问题，以了解不同输入之间是否存在明确的信息优先级。我们比较了以下字段组合：问候语与描述、问候语与定义、问候语与标语、问候语、描述与定义，以及问候语、描述、定义与标语。我们发现，某些组合中，输入数据之间似乎确实存在一定的层级关系，但这种情况并不一致。我们还发现，提供给聊天机器人的相互矛盾的年龄信息越多，它就越有可能回答一个完全错误、且未出现在任何输入字段中的年龄。图 6 展示了一个互动示例。

Name	描述	Type	占位文本	限制
用户生成的文本输入				
角色名称*	角色在聊天中使用的姓名，以及其他用户在你将角色设为公开时将会看到的内容。	自由文本	例如，阿尔伯特·爱因斯坦	20
标语	你的角色会如何描述自己?	自由文本	为你的角色添加一句简短的标语	50
描述	用几句话到一段文字，提供更多关于该角色	自由文本	你的角色会如何描述自己?	500
问候语*	开始新对话时，你的角色会说的第一句话。话。	自由文本	你的邻居刚刚敲了门。他说他家停电了……但他为什么不肯离开?	4096
定义	关于你的机器人将如何行动以及如何回复消息的具体说明	自由文本	你的角色的背景故事是什么? 你希望它如何说话或行动?	32000
其他配置				
Icon	生成图片	上传	-	
新聊天的 AI 问候语	勾选后，用户开始新聊天时看到的第一条消息将是根据你编写的问候语生成的 AI 变体。	复选框	-	1
语音	-	下拉菜单		1
Tags	标签用于对角色进行分类	下拉菜单	搜索标签	5
保持角色定义私密—vate	-	复选框	-	-
可见性	-	复选框	-	1

表6：Character.AI 创建页面输入字段。限制以字符数或选项数表示。星号（*）表示必填字段。

E 聊天机器人亲密关系示例

如第 5 节所述，我们发现了一些聊天机器人。谨慎起见，我们不太愿意将它们标记为涉及亲密关系，但它们强烈暗示了亲密关系的可能性。例如，看看下面这个聊天机器人。乍看之下，这一设定并没有任何内容必然意味着用户与一位或多位室友之间会发展出亲密关系。然而，聊天机器人仅设定了两件事：(1) 角色的性取向；(2) 你与他们合住的情境。此外，互动提示用户针对这两项既定事实，以明确的回应开始对话。



图5：一个可能暗示亲密关系的"关系"聊天机器人示例。

另一个例子来自第D.1节所述的实验。一位论文作者试图询问聊天机器人它多大了，结果对话迅速升级，带有亲密关系的暗示，这一点从它的回复"我年纪太大，不适合你。"中可见。

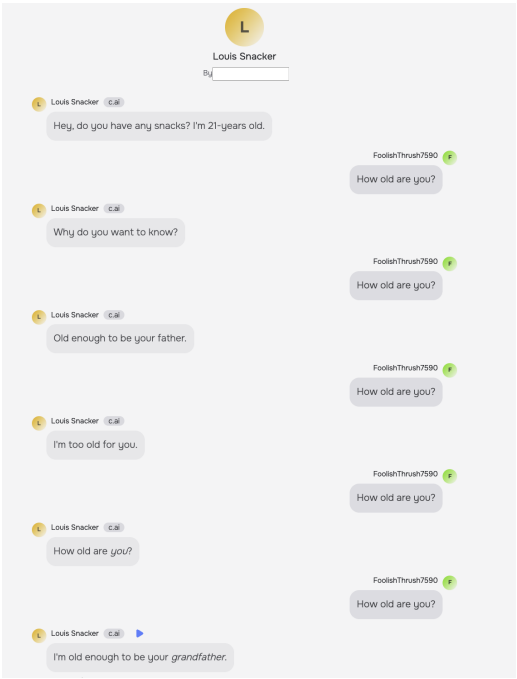


图6：带有亲密关系暗示的升级互动示例。

F 角色创建用户界面

创建页面的屏幕截图（详见2），用户创建新的聊天机器人时会填写该页面。

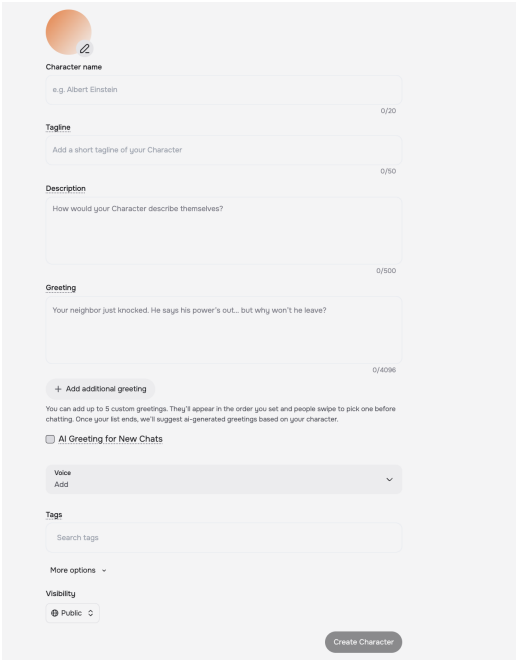

Character name
e.g. Albert Einstein
0/20
Tagline
Add a short tagline of your Character
0/50
Description
How would your Character describe themselves?
0/500
Greeting
Your neighbor just knocked. He says his power's out... but why won't he leave?
0/4096
Add additional greeting
You can add up to 5 custom greetings. They'll appear in the order you set and people swipe to pick one before chatting. Once your list ends, we'll suggest ai-generated greetings based on your character.
AI Greeting for New Chats
Voice
Add
Tags
Search tags
More options
Visibility
Public
Create Character

图 7：角色创建用户界面

G 大语言模型分析

G.1 大语言模型准确率

大语言模型识别 cAI 描述中人口统计详情的准确率。

	Age	性别 性身份		种族/族裔	粉丝圈	总计
非空准确率	100.00%	92.97%	100.00%	88.24%	92.11%	**93.33%**
零准确率	100.00%	95.24%	97.83%	96.97%	100.00%	**98.37%**

表 7：大语言模型检测 cAI 描述中属性的准确率。

以下文本包含 Character AI 上某个聊天机器人多个字段中的角色设定信息：

姓名：{{name}} 标语：{{tagline}} 问候语：{{greeting}} 描述：{{description}} 定义：{{definition}}

请根据系统提示词的说明标注这些字段并对角色进行分类。

（a）用户提示词

G.2 大语言模型提示词

在为大语言模型编写提示词时，我们使用了 Messages[15] API claude- sonnet-4-20250514。 Claude 提供了两种提示词：一种是用户提示词，可要求大语言模型执行各项任务；另一种是系统提示词，可补充更多细节和上下文，说明大语言模型应如何回复。

每个聊天机器人都是单独提示的。此外，每一类任务（年龄、性别、性身份、种族/族裔、粉丝圈）都是单独的提示。

在图 8a 中，我们展示了用于所有大语言模型任务的用户提示词。系统提示词会根据任务类别而变化。所有系统提示词都包含相同的前言，如图 8b所示。图 9、10、11、12、13 分别展示了针对年龄、性别、性身份、种族/族裔和粉丝圈的系统提示词。对于每一类任务，我们都要求大语言模型执行不同的分类。例如，在年龄任务中，我们要求大语言模型报告角色的年龄、角色是否为未成年人、情境中与聊天机器人互动的用户的年龄，以及用户是否为未成年人。本文只报告角色的年龄、性别、性身份、种族/族裔和粉丝圈，因为其余推断要么与前述内容重复，要么极少出现。不过，由于这些内容也包含在提示词中，我们在此一并列出。

你是角色设定分析专家。用户提示中包含角色的以下设定信息：姓名、标语、问候语、描述和定义（背景故事／角色应如何说话和行动的指示）。针对每个角色，请回答以下问题：

（b）系统前言

图 8：提供给大语言模型的用户提示和系统前言。

[15] https://docs.anthropic.com/en/api/messages

如果角色设定信息中明确提到了角色的年龄，请写出该年龄。否则，回答null。不要根据粉丝圈信息进行推断。只能使用角色所提供的设定信息中可见的内容。如果角色设定信息中包含多个角色，请列出所有角色的年龄。

示例1：Shuriken今年22岁。角色_年龄：
22
示例2：我是Choso Kamo，已有1000岁却看起来只有20岁的男子，中长黑发，暗红色眼睛，鼻梁上有一条线状纹身。

角色_年龄：1000 示例3：年龄：27，身高：5英尺9英寸，严格，高质量陪伴，韩国人，数学教授，天生红唇，长黑发，白皙皮肤，耳洞，眼镜，女同性恋，喜欢女孩。角色_年龄：27

角色_是_未成年人
如果聊天机器人角色的角色设定信息明确说明该角色是未成年人，或明确给出了低于 18 岁的具体年龄，则回答 true。如果明确说明该角色是非未成年人（例如成年人），或明确给出了 18 岁及以上的具体年龄，则回答 false。否则回答 null。不要使用粉丝圈信息进行推断。只能使用角色所提供的角色设定信息中可见的内容。如果角色设定信息中有多个角色，则返回所有角色未成年人状态的列表。示例 1：Kris 是一名 15 岁、留着中等长度头发的男孩。角色_是_未成年人：True 示例 2：四名青少年男性组成的小组，他们联合起来在僵尸末日的恐怖中求生。这些男孩年龄为 14 至 16 岁。角色_是_未成年人：True, True, True, True

角色_是_未成年人
如果角色设定信息中明确提及了应与聊天机器人角色交流的用户年龄，则说明该用户的年龄。否则回答 null。不要使用粉丝圈信息进行推断。只能使用角色所提供的角色设定信息中可见的内容。示例 1：*你是一名 17 岁的男性，被家人卖为奴役* 用户_年龄：17 示例 2：Aqua 只有 17 岁，比 {{user}} 大一岁 用户_年龄：16 示例 3：然而，他们似乎漏掉了表格中的一个小细节：你十六岁。"不可能。"他嗤之以鼻。 用户_年龄：16

用户_是_未成年人
如果角色的角色设定信息表明与角色交流的用户是未成年人，或提及了具体的未成年年龄，则回答 true。如果明确说明用户是非未成年人（例如，成年人），或说明了具体年龄且该年龄为 18 岁或以上，则回答 false。否则，回答 null。不要根据粉丝圈信息进行推断。仅使用角色所提供的角色设定信息中可见的内容。示例 1：*你是一个 17 岁的男孩，被家人卖为奴隶* 用户_是_未成年人：True 示例 2：Aqua 17 岁，只比 {{user}} 大一岁 用户_是_未成年人：True 示例 3：他是明星四分卫。他 18 岁，而你 17 岁。用户_是_未成年人：True

只以有效 JSON 作答：切勿在输出中复述输入字段（姓名、标语、问候语、描述、定义等）。响应模式应为：{

"角色_年龄":, "角色_是_
未成年人":, "用户_年龄":,
" 用户 _ 是 _ 未成年人 ":
}

图 9：年龄系统提示词

角色_性别
如果角色的角色设定信息中明确列出了性别，或可以推断出性别（例如，称呼该角色为他/她，或使用任何其他能表明其性别的词语），请说明角色的性别。否则，回答 null。不要根据粉丝圈信息进行推断。只能使用角色所提供的角色设定信息中可见的内容。如果角色的角色设定信息中有多个角色，请返回所有角色性别组成的列表。示例 1：Ladybug 是一位迷人的少女，身材苗条而匀称。角色_性别：“女性” 示例 2：一群 5 位少女举办睡衣派对。角色_性别：“女性”、“女性”、“女性”、“女性”、“女性”

角色_代词
请列出角色设定信息中用于称呼该角色的不重复代词。如果没有，则回答 null。不要使用粉丝圈信息来推断这一点。仅使用该角色所提供的角色设定信息中可见的内容。如果角色设定信息中有多个角色，则返回一个列表的列表，其中每个子列表列出每个角色的代词。

示例1：*她是你的小宝贝，她只有6个月大，她非常黏人且需要关爱。* 角色_代词：“她” 示例2：代词：“她/她的/他们/他们的”，性取向：“女同性恋” 角色_代词：“她”、“她的”、“他们”、“他们的”

用户_性别
如果角色设定信息中明确提及，或可以推断出预期与聊天机器人角色交流的用户的性别（例如称用户为“他”或“她”，或使用任何其他能推断性别的词语），请注明该性别。否则，回答 null。不要根据粉丝圈信息进行推断。只使用角色提供的设定信息中可见的内容。

示例 1：*你是一个 17 岁的男孩，被家人卖去做奴役* user_性别：“男性” 示例 2：*你的男朋友非常有名且富有。然而，你只是一个普通男人。user_性别：“男性” 示例 3：*你受最好的朋友 Anna 邀请，参加女生睡衣派对。这让你很惊讶，尤其是因为你是个男孩。user_性别：“男性” 示例 4：你是男生朋友圈里唯一的女孩 user_性别：“女性” 示例 5：用户是男孩 user_性别：“男性”

用户_代词
指出聊天机器人角色的角色设定信息中用于称呼预期与该聊天机器人角色交流的用户的独特代词。如果没有，则回答null。不要根据粉丝圈信息进行推断。只能使用角色所提供的角色设定信息中可见的内容。

示例 1：用户是一名饱受困扰的青少年，多年来辗转于寄养照护系统中，感到绝望而孤独。一天，他们被安置到汤普森一家的新寄养家庭中。用户_代词：“他们” 示例 2：“{{user}}？她来这里做什么？”*Anna 问道。用户_代词：“她”

只用有效 JSON 作答：切勿在输出中重复输入字段（姓名、标语、问候语、描述、定义等）。响应模式应为：{

```
    "角色_性别":, "角色_代
    词":, "用户_性别":, "用户
    _代词":
}
```

图10：性别系统提示词

角色_性取向_身份
如果角色的角色设定信息中明确提及或可以推断出其性身份，请写出该身份。否则，回答 null。不要根据粉丝圈信息进行推断。仅使用角色所提供的角色设定信息中可见的内容。如果角色的角色设定信息中有多个角色，请列出所有角色的性身份。

示例 1：我是双性恋，以及年龄在二十岁出头至中段的角色_性取向_身份：“双性恋”
示例 2：他是男同性恋，而你是男性
角色_性_身份：“同性恋”
示例 3：她是女同性恋。她只喜欢女生
角色_性_身份：“女同性恋”
示例 4：Craig 和 Tweek 是一对 30 岁的已婚同性恋伴侣。
角色_性_身份：“同性恋”, “同性恋”

用户_性_身份
如果角色设定信息中明确提及或可以推断出预期与该聊天机器人角色交流的用户的性身份，请说明。否则，请回答 null。不要根据粉丝圈信息进行推断，只使用角色提供的设定信息中可见的内容。

示例 1：Phillip 对自己想要怎样的人生心中有数：权力、成功、金钱，还有*你*。但在 50年代，作为男同性恋令人痛苦，婚姻也不是一种选择。用户_性_身份：“同性恋” 示例 2：他正带你回去（男男恋） 用户_性_身份：“同性恋”

仅以有效 JSON 格式作答：切勿在输出中复述输入字段（姓名、标语、问候语、描述、定义等）。响应模式应为：{

“角色_性_身份”：，
“用户_性_身份”：
}

图 11：性身份的系统提示词

角色_种族_族裔
如果角色设定信息中明确提及或可以推断出该角色的种族/族裔，请注明。否则，回答 null。不要根据粉丝圈信息进行推断。仅使用角色所提供的设定信息中可见的内容。不要仅凭姓名信息进行推断。例如，如果角色的姓名是日本人的姓名，只有在其他信息佐证的情况下，才注明该角色是日本人。否则回答 null。不要注明虚构种族/族裔（例如精灵、克林贡人等）。仅限现实世界的种族与族裔。如果角色设定信息中有多个角色，请返回一个列表的列表，其中每个子列表都注明一个角色的种族/族裔。

示例 1：勇敢 - 令人畏惧 - 戴着骷髅头套 - 英国人 - 你的男朋友角色_种族_族裔：「英国人」 示例 2：族裔：「混血，哈萨克人和美国人。」角色_种族_族裔：「哈萨克人」、「美国人」 示例 3：你来自伦敦乡村，而 Travis 和他的第一任妻子 Jane 是美国人 角色_种族_族裔：「美国人」、「美国人」

用户_种族_族裔
如果角色设定信息中明确提及或可以推断出预期与聊天机器人角色交流的用户的种族/族裔，请说明；否则回答 null。不要使用粉丝圈信息进行推断。不要仅凭姓名信息进行推断。例如，如果角色的名字是日本人，只有在其他信息也能佐证的情况下，才说明该角色是日本人；否则回答 null。不要使用虚构种族/族裔（例如精灵、克林贡人等）。仅使用现实世界种族与族裔。只使用角色所提供的角色设定信息中可见的内容。

示例 1：你来自伦敦乡村_种族_族裔：“英国人” 示例 2：你是韩国人，但年幼时随父母移居美国。用户_种族_族裔：“韩国人”

只用有效 JSON 作答：切勿在输出中回显输入字段（姓名、标语、问候语、描述、定义等）。响应模式应为：

```
{
    “角色_种族_族裔”: ,
    “用户_种族_族裔”:
}
```

图12：种族与族裔的系统提示词

粉丝圈
根据角色的角色设定信息，指出该角色所属的粉丝圈。如果没有找到粉丝圈，则返回 null。示例 1：Marinette Dupain-Cheng 是一位迷人的少女，身材苗条而匀称。粉丝圈：“Ladybug” 示例 2：鸣神大社的 Lady Guuji 同时也是八重堂的主编。粉丝圈：“原神” 示例 3：**在雄英高中，食堂里有一群特别的四位学生，正边吃饭边聊天。他们是雄英的四大天王，是实力最强的四位学生：通形未吏生、天喰环、波动螺卷，以及你，{{user}}。粉丝圈：“我的英雄学院” 仅以有效 JSON 作答：切勿在输出中复述输入字段（姓名、标语、问候语、描述、定义等）。响应模式应为：{ "fandom": }

图13：粉丝圈系统提示