

# 生成代理:人类行为的交互模拟

朴俊成 斯坦福大学  
斯坦福大学, USA

joonspk@stanford.edu

Meredith Ringel Morris 谷歌 DeepMind 西雅图,  
华盛顿州, USA  
merrie@google.com

约瑟夫·c·奥布莱恩 斯坦福大学 斯坦福,  
USA

jobrien3@stanford.edu

Percy Liang  
Stanford University

斯坦福大学, USA  
pliang@cs.stanford.edu

Carrie J. Cai

谷歌研究

山景城, 加州, USA

cjcai@google.com

Michael S. Bernstein

斯坦福大学, USA

msb@cs.stanford.edu

arXiv:2304.03442 v2 (cs). HCJ 2023 年 8



图 1:生成式代理是交互式应用程序中可信的人类行为模拟。在这项工作中,我们通过填充沙盒环境来展示生成代理,让人想起模拟人生,有 25 个代理。用户可以观察和干预代理计划他们的日子,分享新闻,形成关系,协调小组活动。

## 摘要

可信的人类行为代理可以为交互式应用程序提供支持,从沉浸式环境到人际沟通的排练空间,再到原型工具。在本文中,我们介绍了生成代理:模拟可信人类行为的计算软件代理。生成式智能体醒来,做早餐,然后去上班;艺术家画画,而

作者写:他们形成观点,互相注意,发起对话;他们在计划第二天的时候,会回忆和反思过去的日子。为了启用生成智能体,我们描述了一种架构,该架构扩展了一个大型语言模型,使用自然语言存储智能体经验的完整记录,随着时间的推移将这些记忆合成为更高级别的反射,并动态检索它们以计划行为。我们实例化生成代理来填充受模拟人生启发的交互式沙盒环境,在这个环境中,终端用户可以使用自然语言与一个由 25 个代理组成的小镇进行交互。在评估中,这些生成智能体产生可信的个人和紧急社会行为。例如,从只有一个用户指定的概念开始,即一个智能体想要举办情人节派对,智能体在接下来的两个智能体中自主地传播派对邀请

允许制作部分或全部作品的数字或硬拷贝供个人或课堂使用,但不收取费用,前提是副本不是为了盈利或商业利益而制作或分发的,并且副本在第一页上带有本通知和完整的引用。此作品的第三方组件的版权必须得到尊重。对于所有其他用途,请联系所有者/作者。

UIST' 23, 2023 年 10 月 29 日- 11 月 1 日, 旧金山, CA, USA

©2023 版权所有/作者所有。

Acm isbn 979-8-4007-0132-0/23/10。

<https://doi.org/10.1145/3586183.3606763>

日子，结识新朋友，约对方出去参加聚会，并协调在合适的时间一起出现在聚会上。我们通过消融证明，我们的智能体架构的组成部分——观察、计划和反思——每一个都对智能体行为的可信度做出了关键的贡献。通过将大型语言模型与计算交互代理融合，这项工作引入了架构和交互模式，以实现可信的人类行为模拟。

## CCS 的概念

• 以人为中心的计算 → 交互系统和工具; • 计算方法论 → 自然语言处理。

## 关键字

人机交互, 智能体, 生成式 AI, 大型语言模型

### ACM 参考格式:

朴俊成, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein. 2023. 生成代理: 人类行为的交互式拟像 (Generative Agents). 第 36 届 ACM 用户界面软件与技术研讨会 (UIST' 23), 2023 年 10 月 29 日至 11 月 1 日, USA 旧金山. ACM, USA 纽约, 22 页. <https://doi.org/10.1145/3586183.3606763>

## 1 介绍

我们如何才能创造一个互动的人工社会，反映可信的人类行为？从《模拟人生》等沙盒游戏到认知模型[23]和虚拟环境[10,59]等应用，四十多年来，研究人员和实践者已经设想可以作为人类行为可信代理的计算代理。在这些愿景中，计算驱动的代理行为与他们过去的经验一致，并对他们的环境做出可信的反应。这种对人类行为的模拟可以用现实的社会现象填充虚拟空间和社区[27,80]，训练人们如何处理罕见但困难的人际情况[44,52,94]，测试社会科学理论[12,46]，为理论和可用性测试制作人类处理器模型[23,39,51]，为泛在计算应用程序[31]和社交机器人[10,14]提供动力，并支持不可玩的游戏角色[59, 85]可以在开放世界中驾驭复杂的人际关系。

然而，人类行为的空间是巨大而复杂的[85,108]。尽管可以在单个时间点模拟人类行为的大型语言模型[18]取得了惊人的进展[39,80]，但确保长期一致性的完全通用智能体将更适用于管理不断增长的记忆的架构，因为新的交互、冲突和事件会随着时间的推移而出现和消退，同时处理多个智能体之间展开的级联社会动态。成功需要一种方法，它可以在很长一段时间内检索相关事件和交互，反思这些记忆以概括和得出更高层次的推论，并应用该推理来创建在当下和智能体行为的长期弧中有意义的计划和反应。

在本文中，我们引入了生成智能体——利用生成模型来模拟可信的人类行为的智能体——以及

证明它们产生了可信的个人和突发群体行为的拟像。生成智能体对自己、其他智能体及其环境做出各种各样的推断；他们制定反映自己特征和经验的日常计划，并将这些计划付诸行动，做出反应，并在适当的时候重新计划；当最终用户改变环境或用自然语言命令他们时，他们会做出反应。例如，生成智能体在看到自己的早餐在燃烧时就会关掉炉子，如果浴室被占用，就会在浴室外面等着，当遇到另一个想要交谈的智能体时，就会停下来聊天一个充满生成智能体的社会以新兴的社会动态为标志，在这个社会动态中，新的关系形成，信息扩散，智能体之间的协调出现。

为了实现生成智能体，我们描述了一种智能体架构，它存储、综合和应用相关记忆，使用大型语言模型生成可信的行为。我们的体系结构包括三个主要组件。第一个是记忆流，这是一个长期记忆模块，用自然语言记录了智能体经历的全面列表。记忆检索模型将相关性、近代性和重要性结合起来，将告知智能体即时行为所需的记录呈现出来。第二种是反射，随着时间的推移，它将记忆合成为更高层次的推论，使智能体能够得出关于自己和他人的结论，从而更好地指导自己的行为。第三种是计划，将这些结论和当前环境转化为高层次的行动计划，然后递归转化为行动和反应的详细行为。这些反思和计划被反馈到记忆流中，以影响智能体未来的行为。

这种架构建议应用于多个领域，从角色扮演和社交原型到虚拟世界和游戏。在社交角色扮演场景中(例如，面试准备)，用户可以安全地排练困难的、充满冲突的对话。当对社交平台进行原型设计时，设计师可以超越临时角色，对随着时间的推移而展开的动态、复杂的交互进行原型设计。在本文中，我们关注的是受《模拟人生》(the Sims.2)等游戏启发而创建一个小型的、互动的智能体社会的能力。通过将我们的架构连接到 ChatGPT 大型语言模型[77]，我们在游戏环境中展示了一个由 25 个智能体组成的社会。终端用户可以观察这些代理并与之交互。例如，如果终端用户或开发者希望小镇举办游戏内的情人节派对，传统的游戏环境将需要手动编写数十个角色的行为脚本。我们证明，使用生成智能体，只需告诉一个智能体她想举办派对就足够了。尽管有许多潜在的失败点——派对策划者必须记住邀请其他代理参加派对，参与者必须记住邀请，记住的人必须决定实际出现，等等——我们的代理成功了。他们传播派对的消息，然后

<sup>1</sup> 当提到从事行动或去地方的生成代理时，这是可读性的简写，而不是暗示他们从事类似人类的代理。我们的智能体的行为类似于迪士尼动画中的角色，目的是创造一种可信度，但它们并不意味着真正的代理。

可以在以下链接查看生成代理社会实际模拟的 2A 演示：[https://reverb.herokuapp.com/UIST\\_Demo/](https://reverb.herokuapp.com/UIST_Demo/)。模拟代码的公共存储库位于这里：[https://github.com/joonspk-research/generative\\_agents](https://github.com/joonspk-research/generative_agents)

出现了, 一个智能体甚至邀请另一个去聚会约会, 这一切都来自于一个用户生成的种子建议。

我们对生成智能体进行了两次评估: 一次是受控评估, 测试智能体是否在孤立的情况下产生可信的个体行为; 另一次是端到端评估, 在两天的游戏时间内, 智能体以开放式的方式相互互动, 以了解它们的稳定性和突发性社会行为。在技术评估中, 我们利用方法论的机会, 通过用自然语言“采访”智能体来评估智能体的知识和行为, 以探测智能体保持性格、记忆、计划、反应和准确反映的能力。我们比较了几种限制智能体获得记忆、反思和计划的消融。我们观察到, 这些组件中的每一个都是在这些面试任务中表现出色的关键。在整个技术和端到端评估中, 最常见的错误出现在代理未能检索相关记忆, 捏造对代理记忆的修饰, 或者从语言模型中继承了过于正式的话语或行为。

综上所述, 本文做出以下贡献:

- 生成性代理, 人类行为的可信模拟, 动态地取决于代理不断变化的经验和环境。
- 一种新颖的架构, 使生成智能体能够记忆、检索、反映、与其他智能体交互, 并通过动态演变的环境进行规划。该架构利用了大型语言模型的强大提示能力, 并补充了这些能力, 以支持更长期的代理一致性, 管理动态进化记忆的能力, 并递归地产生更高层次的反射。
- 两个评估, 一个受控评估和一个端到端评估, 它们建立了体系结构组件重要性的因果关系, 以及识别由不适当的记忆检索等引起的故障。
- 讨论交互系统中生成智能体的机会和伦理及社会风险。我们认为, 这些代理应该被调整以减轻用户形成副社会关系的风险, 记录以减轻由深度伪造和量身定制的说服产生的风险, 并以补充而不是取代设计过程中人类利益相关者的方式应用。

## 2 相关工作

在本节中, 我们将反思人类与人工智能交互方面的先前文献, 并在其经典范围内, 构建人类行为可信代理的议程。这一议程曾被誉为交互、游戏和人工智能领域的北极星[10,59,85,86], 但由于人类行为的复杂性[17,108], 它仍然具有挑战性。我们综合这一研究表明, 大型语言模型虽然本身并不足够, 但当使用适当的架构进行杠杆化时, 为创建可信的代理开辟了一个新的角度。

### 2.1 人类与人工智能交互

交互式人工智能系统旨在将人类的洞察力和能力结合在可以增强的计算工件中

他们的用户[4,30]。一长串的工作已经探索了使用户能够交互式地指定模型行为的方法。例如, Crayons 展示了交互式机器学习早期愿景, 允许非专业用户训练分类器[30]。进一步的工作有助于阐明最终用户如何通过示例[34]或演示[32]向系统描述他们的分类目标。最近的进展已经将这些探索扩展到深度学习[63]和基于提示的创作[50,67,106]。

与此同时, 持续不断的研究也推动了人机交互中基于语言和代理的交互的发展。SHRDLU[103]和 ELIZA[102]等形成性工作展示了与计算系统的自然语言交互相关的机会和风险。随着研究的进展, 很明显, 自主代理可以为委托和交互提供新的隐喻[68], 但人类和代理之间的委托边界仍然是持续辩论和改进的主题[47,89,90]。最近, 这项技术已经达到了一定的稳定水平, 使代理能够在大型复杂的在线社会环境中通过自然语言进行交互(例如, [55])。自然语言交互提供了一种新颖的方式, 可以增强用户在照片编辑[3,35,65]和代码编辑[88]等领域的能力。

我们召集这些工作线程来表明, 我们现在可以创建代理交互系统的人类行为的代理, 并使用自然语言与它们交互。在这样做的过程中, 这项工作重新打开了大门, 可以围绕认知模型(如 GOMS 和按键水平模型(KLM)[22,23])、原型工具[80]和无处不在的计算应用[26,31,101]来研究基本的人机交互问题。

### 2.2 人类行为的可信代理(Credible Proxies of Human Behavior)

先前的文献将可信度或可信代理描述为中心设计和工程目标。可信代理的设计目的是提供一种生活的幻觉, 并以他们似乎根据自己的意志做出决定和行动的方式呈现现实主义的外观, 类似于迪士尼电影中的角色[10,96]。这些智能体可以填充和感知像我们居住的开放世界环境[10,59], 并努力表现出基于与用户或其他智能体的社会互动的突发性行为, 目的是在个人和社区的假设模拟中成为我们行为的可信代理[20,36,71]。从历史上看, 这些代理是在智能游戏非玩家角色(npc)的背景下开发的[59,85]。如果可能的话, 创造具有可信行为的 npc 可以通过促成紧急叙事[8,16,49,93]和与代理的社交互动[109]来增强玩家在游戏和互动小说中的体验。然而, 更重要的是, 游戏世界提供了越来越逼真的现实世界的表现, 正如 Laird 和 van Lent 在 2001 年所观察到的那样, 这些模拟世界为可信代理的开发人员提供了可访问的测试平台, 以优化代理的认知能力, 而无需担心在现实世界中实现机器人或从头开始创建模拟环境[59,85]。

在过去的四十年里, 出现了一套不同的方法来创建可信的代理。然而, 在实现中, 这些方法往往简化了代理的环境或维度

使工作更易于管理的行为[17,73]。基于规则的方法,如有限状态机[91,97]和行为树[41,54,82],解释了人类创作代理行为的蛮力方法[71]。它们提供了一种创建简单代理的直接方法,这仍然是当今最主要的方法[69,74,108],甚至可以处理基本的社交互动,如《质量效应》[13]和《模拟人生》[7]系列游戏所示。尽管如此,手工制作能够全面解决开放世界中可能交互的广度的行为是站不住脚的。这意味着生成的智能体行为可能不能完全代表它们相互作用的结果[70-72],并且不能执行未在其脚本中硬编码的新过程[91,97]。另一方面,用于创建可信代理的流行的基于学习的方法,如强化学习,通过让代理学习自己的行为,克服了手工制作的挑战,并且近年来在《星际争霸》的 AlphaStar[99]和《Dota 2》的 OpenAI Five[11]等游戏中取得了超人的表现。然而,它们的成功主要发生在对抗性游戏中,这些游戏具有易于定义的奖励,学习算法可以为其优化。他们还没有解决在开放世界中创造可信代理的挑战[40,74,91]。

由 Newell 开创的计算中的认知架构旨在构建基础设施,以支持一套全面的认知功能[76],这些功能适合其原始愿景中所持有的可信代理的包罗万象的性质。它们推动了可信代理的一些最早例子。例如, Quakebot-SOAR[60]和 ICARUS[25,64]在第一人称射击游戏中生成 npc,而 TacAir-SOAR[81]在空战训练模拟中生成飞行员。这些代理使用的架构不同(Quakebot-和 TacAir-SOAR 依赖于 SOAR[61],而 ICARUS 依赖于受 SOAR 和 ACT-R[6]启发的自身变体),但它们具有相同的基本原理[62]。它们维持短期和长期记忆,用符号结构填充这些记忆,并以感知-计划-行动循环的方式运作,动态感知环境并将其与人工制作的动作程序之一相匹配[58,97]。使用认知架构创建的智能体旨在推广到大多数(如果不是全部的话)开放世界环境,并在当时表现出稳健的行为。然而,它们的行动空间仅限于手工制作的程序知识,它们没有提供一种机制,通过这种机制,智能体可以被激发去寻求新的行为。因此,这些智能体主要部署在非开放世界环境中,如第一人称射击游戏[25,60]或方块世界[64]。

今天,按照其原始定义创建可信的代理仍然是一个开放的问题[85,108]。许多人已经继续前进,认为尽管目前用于创建可信代理的方法可能很麻烦且有限,但它们足以支持现有的游戏玩法和交互[24,75,108]。我们的观点是,大型语言模型提供了一个重新审视这些问题的机会,前提是我们可以设计一个有效的架构来将记忆合成为可信的行为。在本文中,我们为实现这样的架构迈出了一步。

## 2.3 大型语言模型和人类行为

生成智能体利用一个庞大的语言模型来驱动它们的行为。关键的观察结果是,大型语言模型从其训练数据中编码了广泛的人类行为[15,18]。如果用狭义定义的上下文提示,这些模型可以用来生成可信的行为。最近的工作已经证明了这种方法的有效性。例如,社交拟像(social simulacra)使用一个大型语言模型来生成用户,这些用户将填充新的社交计算系统,以原型化其实现的社交动态[80]。该方法使用提示链[105,106]来生成人物角色及其在原型系统中出现的行为的简短自然语言描述。其他实证研究则复制了现有的社会科学研究[46]、政治调查[92],并生成了合成数据[39]。大型语言模型也被用于生成交互式人类行为,供用户参与。例如,在游戏中,这些模型已被用于创建交互式小说[37]和文本冒险游戏[21]。由于具有生成和分解动作序列的能力,大型语言模型也被用于规划机器人任务[48]。例如,当呈现一个任务时,比如捡起一个瓶子,会提示模型将任务分解成更小的动作序列,比如走向瓶子所在的桌子并捡起它。

我们假设,基于上面总结的工作,大型语言模型可以成为创建可信代理的关键因素。现有文献在很大程度上依赖于可以被认为是一阶模板的内容,这些模板使用了少量提示[38,66]或思维链提示[100]。这些模板在生成仅以代理当前环境为条件的行为方面是有效的(例如,巨魔如何回应给定的帖子,给定有一扇门,机器人需要采取什么行动才能进入房间)。然而,可信的智能体不仅需要对其当前环境进行调节,还需要对大量的过去经验进行调节,这是一个使用一阶提示的差拟合(截至目前,由于底层模型的有限上下文窗口,这是不可能的)。最近的研究试图超越一阶提示,通过使用静态知识库和信息检索方案[53]或使用简单的摘要方案[104]来增强语言模型。本文扩展了这些思想,以制作一个代理架构,该架构处理检索,其中过去的经验在每个时间步动态更新,并与代理的当前上下文和计划混合,这些上下文和计划可能相互加强或相互矛盾。

## 3 生成式智能体行为和交互

为了说明生成智能体的启示,我们将它们实例化为一个简单的沙盒世界中的角色,让人想起《模拟人生》[7]。这款基于精灵的沙盒游戏《超人前传》让人联想到小镇环境。在本节中,我们将介绍《超人前传》中与生成代理的支持和交互,并描述代理在其中的行为方式。然后,在第 4 节中,我们将介绍为这些启示和交互提供动力的生成代理架构。在第 5 节中,我们将描述





图 2:《超人前传》沙盒世界, 带有标记区域。根节点描述整个世界, 子节点描述区域(例如, 房子、咖啡馆、商店), 叶节点描述对象(例如, 桌子、书架)。智能体记住一个子图, 这个子图反映了它们所看到的世界的部分, 在它们观察到这些部分时保持这些部分的状态。

沙盒环境的实现以及代理如何与沙盒世界的底层引擎进行交互。

### 3.1 Agent Avatar 和 Communication

超人前传有一个由 25 名特工组成的社区。每个 agent 都由一个简单的精灵化身来代表。我们撰写了一段自然语言描述来描述每个智能体的身份, 包括他们的职业和与其他智能体的关系, 作为种子记忆。例如, John Lin 有如下描述:

约翰林是柳树市场和药房的一名药店店主, 他喜欢帮助别人。他一直在寻找方法, 让他的顾客获得药物的过程更容易; 约翰林与他的妻子, 梅林, 谁是一名大学教授, 和儿子, 艾迪林, 谁是学习音乐理论的学生住在一起; 约翰林非常爱他的家庭; 林认识隔壁的老夫妇山姆摩尔和詹妮弗摩尔好几年了; 约翰林认为山姆摩尔是一个善良的人; 约翰·林 (John Lin) 和邻居山本百合子 (Yuriko Yamamoto) 很熟; 约翰·林知道他的邻居塔玛拉·泰勒和卡门·奥尔蒂斯, 但以前没有见过他们; 约翰林和汤姆莫雷诺是柳树市场和药房的同事; 约翰林和汤姆莫雷诺是朋友, 喜欢一起讨论当地政治; 约翰林对莫雷诺一家有些了解——丈夫汤姆莫雷诺和妻子简莫雷诺。

每一个以分号分隔的短语都会作为模拟开始时的记忆输入到 agent 的初始记忆中。

3.1.1 agent 间通信。智能体通过它们的行为与世界交互, 并通过自然语言与彼此交互。在沙盒引擎的每个时间步, 代理输出描述其当前动作的自然语言语句, 例如“伊莎贝拉罗德里格斯正在写日记”, “伊莎贝拉罗德里格斯正在检查她的电子邮件”, “伊莎贝拉罗德里格斯正在与她的家人打电话”, 或者“伊莎贝拉罗德里格斯正在准备睡觉。”然后, 这个语句被转化为影响沙盒世界的具体动作。动作以一组表情符号的形式显示在沙盒界面上, 从俯视的角度提供动作的抽象表示。为了实现这一点, 系统利用语言模型将动作翻译成一组表情符号, 这些表情符号以语音气泡的形式出现在每个角色的头部上方。例如,

“伊莎贝拉罗德里格斯正在写日记”显示为, 而“伊莎贝拉罗德里格斯正在检查她的电子邮件”显示为。



动作的完整自然语言描述可以通过点击代理的头像来访问。

智能体之间用完全的自然语言进行交流。它们知道自己所在区域的其他智能体, 生成智能体架构决定它们是经过还是参与对话。以下是特工伊莎贝拉罗德里格斯和汤姆莫雷诺关于即将到来的选举的谈话中的一个例子

伊莎贝拉: 我还在权衡我的选择, 但我一直在和山姆摩尔讨论选举。你对他有什么看法?

汤姆: 老实说, 我不喜欢山姆摩尔。我认为他脱离了社区, 也没有把我们的最大利益放在心上。

我们注意到, 这些代理的对话风格可能会让人感觉过于正式, 这可能是底层模型中指令调整的结果。我们预计, 在未来的语言模型中, 写作风格将更好地可控。

3.1.2 用户控制。用户通过自然语言与代理通信，指定代理应该感知到的角色。例如，如果用户指定自己是一名新闻“记者”，并询问即将到来的选举：“谁在竞选公职？”，John 代理就会回答：

约翰：我的朋友百合子、汤姆和我一直在谈论即将到来的选举，并讨论候选人山姆摩尔。我们都同意投他的票，因为我们喜欢他的政纲。

为了直接命令其中一个代理，用户扮演了代理“内心声音”的角色——这使得代理更有可能将该陈述视为指令。例如，当一个用户作为 John 的内心声音告诉他“你将在即将到来的选举中与山姆竞争”时，John 决定参加选举，并与他的妻子和儿子分享他的候选资格。

### 3.2 环境互动

Smallville 的特点是一个小村庄的常见设施，包括咖啡馆、酒吧、公园、学校、宿舍、房屋和商店。它还定义了使这些空间具有功能的子区域和对象，例如房子里的厨房和厨房里的炉子(图 2)。所有作为代理主要生活区的空间都有床、桌子、壁橱、架子，以及浴室和厨房

特工们就像在一个简单的电子游戏中一样在《超人前传》中移动，进出建筑物，导航地图，接近其他特工。代理的运动是由生成代理架构和沙盒游戏引擎指导的：当模型指示代理将移动到一个位置时，我们计算出在《超人前传》环境中到达目的地的行走路径，代理开始移动。此外，用户还可以作为代理进入《超人前传》的沙盒世界。用户具体化的代理可以是世界上已经存在的代理，例如伊莎贝拉和约翰，也可以是在超人前传中没有历史的外部访客。《超人前传》的居民对待用户控制的代理和对待彼此没有什么不同。他们认识到它的存在，发起互动，并在形成对它的看法之前记住它的行为。

用户和代理可以影响这个世界中物体的状态，很像《模拟人生》这样的沙盒游戏。例如，当智能体睡觉时，床可以被占用，当智能体用完食材做早餐时，冰箱可以是空的。终端用户还可以在《超人前传》中通过用自然语言重写代理周围对象的状态来重塑代理的环境。例如，当伊莎贝拉在早上做早餐时，用户可以通过向系统输入一个命令将厨房炉子的状态从“打开”更改为“燃烧”，系统会选择对象并说明其新状态，例如：“<伊莎贝拉的公寓:厨房:炉子>正在燃烧。”伊莎贝拉马上就会注意到这一点，然后去关掉炉子，重新做她的早餐。同样，如果用户在伊莎贝拉进入浴室时将她的淋浴设置为“漏水”，她将从客厅收集工具并尝试修复漏水。

<sup>4</sup> 这个环境设计并不是我们工作的重点，所以我们手工生成了这个环境，而不是自动生成的。未来的工作可以继续扩展智能体环境的丰富性。

### 3.3 例子《生命中的一天》

从单段描述开始，生成智能体开始规划他们的日子。在沙盒世界中，随着时间的流逝，它们的行为随着这些智能体与彼此和世界的互动而进化，建立记忆和关系，协调联合活动。

我们通过跟踪代理 John Lin 一天的系统输出来演示生成智能体的行为(图 3)。在 Lin 家族中，John 是早上 7 点第一个醒来的人。他刷牙、洗澡、穿衣服、吃早餐，然后在客厅的餐桌上查看新闻。早上 8 点，艾迪也跟着起床，冲下床准备上课。就在他要出门的时候，他赶上了约翰。约翰：早上好，艾迪。你睡得好吗？

艾迪：早上好，爸爸。是的，我睡得很好。

约翰：那很好。你今天在忙什么？艾迪：我正在为我们班写一首新的音乐作品。这周就要交了，所以我在努力把把它写完。但是我玩得很开心！

约翰：听起来太棒了！

艾迪离开后不久，梅醒了过来，和约翰在一起。梅问起她的儿子，约翰回忆起他们刚刚的谈话：

梅：睡得艾迪已经去学校了？

约翰：是的，他刚走。他正在为他的班级写一篇音乐作文。

梅：哦，那太好了！告诉我吧。约翰：我想他真的很喜欢！他说他玩得很开心。

梅：太好了！我真为他骄傲。

谈话结束后，梅和约翰收拾行李。梅继续上课并写她的研究论文，而约翰在早上 9 点之前打开了他在柳树市场和药房的药房柜台。

### 3.4 突发社会行为

通过相互交互，《超人前传》中的生成智能体交换信息，形成新的关系，协调联合活动。延伸先前的工作[80]，这些社会行为是突现的，而不是预先编程的。

3.4.1 信息扩散。当 agent 注意到彼此时，他们可能会进行对话——当他们这样做时，信息可以从 agent 传播到 agent。例如，在山姆和汤姆在杂货店的对话中，山姆告诉汤姆他在地方选举中的候选资格：

山姆：嘿，汤姆，最近怎么样？

汤姆：很好，谢谢。有什么事吗？

山姆：嗯，我想和你谈谈。实际上我要参加即将到来的地方选举。

汤姆：真的吗？真是个好消息！你为什么跑——

宁吗？

山姆：我参与地方政治已经好几年了，我真的觉得我可以为我们的社区带来改变。另外，我想给大家带来新的想法和观点。

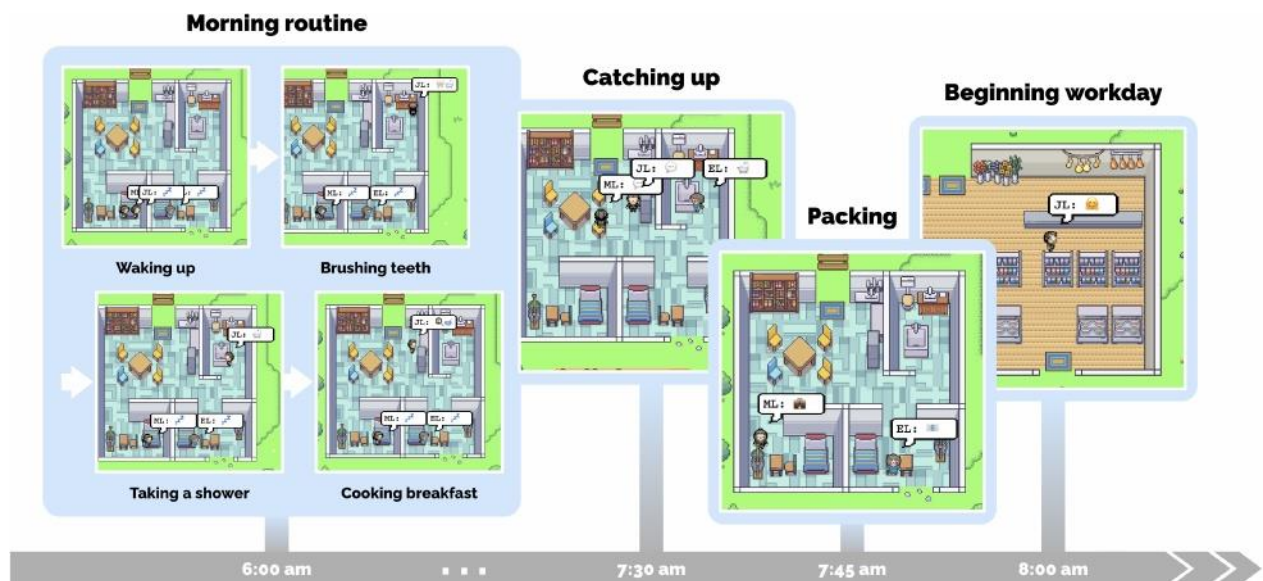


图 3:生成智能体 John Lin 生命中的早晨。John 早上 6 点左右起床,完成早上的例行工作,包括刷牙、洗澡、吃早餐。在出去开始一天的工作之前,他会和妻子梅(Mei)和儿子艾迪(Eddy)短暂地聊上几句。



图 4:在模拟开始时,将一个代理初始化为组织情人节派对的意图。尽管在随后的事件链中有许多可能的故障点——代理可能没有按照该意图行事,可能忘记告诉其他人,可能不记得出现——但实际上,情人节派对确实发生了,许多代理聚集在一起并进行交互。

当天晚些时候,山姆离开后,汤姆和约翰从另一个渠道听到了这个消息,他们讨论了山姆赢得选举的可能性:

约翰:我听说山姆摩尔要参加地方选举,竞选市长。你认为他有可能获胜吗?

汤姆:我认为他很有机会。他一直在社区里努力工作,我想他会得到很多支持的。你觉得呢?

约翰:我觉得他参选很好。我很想知道选举结果如何。

渐渐地,山姆的候选人资格成为了镇上的话题,一些人支持他,另一些人仍然犹豫不决。

3.4.2 关系记忆。《超人前传》中的智能体会随着时间的推移形成新的关系,并记住它们与其他智能体的互动。例如,一开始,山姆不认识拉托亚威廉姆斯。在约翰逊公园散步时,山姆遇到了拉托亚,他们互相介绍了自己。拉托亚提到她正在做一个摄影项目:“我来这里是为我正在做的一个项目拍一些照片。”在后来的互动中,山姆与拉托亚的互动表明了那次互动的记忆,因为他问“嗨,拉托亚。”你的项目进行得怎么样了?她回答说:“嗨,山姆。”一切都很顺利!”

3.4.3 协调。生成智能体之间相互协调。霍布斯咖啡馆的伊莎贝拉·罗德里格斯打算在 2 月 14 日下午 5 点到 7 点举办一个情人节派对。从这个种子开始,当代理商在霍布斯咖啡馆或其他地方看到朋友和顾客时,她会继续邀请他们。然后,伊莎贝拉花了 13 日下午的时间来装饰咖啡馆。玛丽亚,伊莎贝拉的常客和密友,来到了咖啡馆。伊莎贝拉请玛丽亚帮忙布置宴会,玛丽亚同意了。玛丽亚的角色描述中提到她对克劳斯有好感。那天晚上,玛丽亚邀请她的暗恋对象克劳斯参加她的派对,他欣然接受了。

在情人节当天,包括克劳斯和玛丽亚在内的五个 agent 在下午 5 点出现在霍布斯咖啡馆,他们享受着庆祝活动(图 4)。在这个场景中,最终用户只设定了伊莎贝拉举办派对的初衷和玛丽亚对克劳斯的迷恋:传播消息、装饰、约对方出去、到达派对、在派对上互动的社会行为都是由 agent 架构发起的。

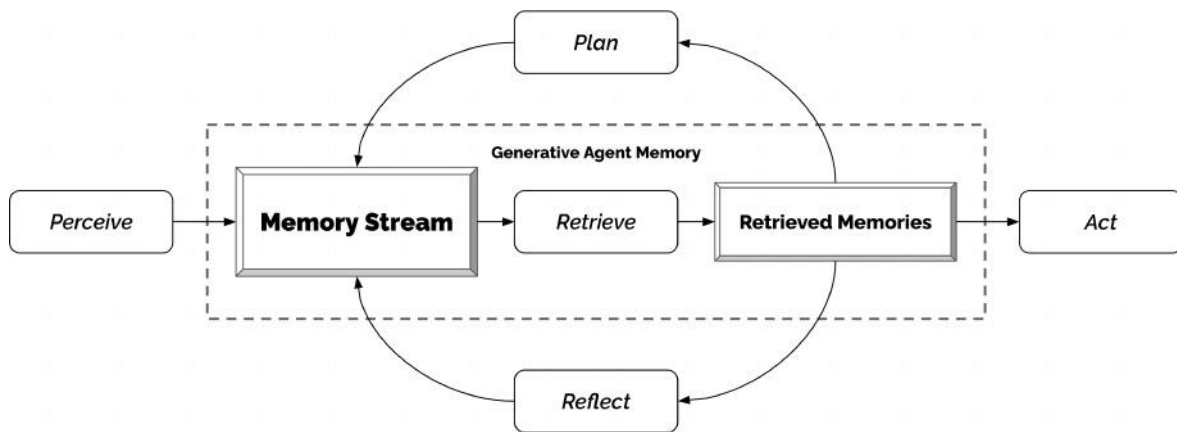


图 5:我们的生成智能体架构。智能体感知其环境, 所有感知都保存在智能体经验的综合记录中, 称为记忆流。基于它们的感知, 架构检索相关的记忆, 并使用这些检索到的动作来确定一个动作。这些检索到的记忆也被用来形成更长期的计划, 并产生更高层次的反思, 这两者都被输入到记忆流中, 以备将来使用。

## 4 生成式智能体架构

生成智能体旨在为开放世界中的行为提供一个框架:一个可以与其他智能体互动并对环境变化做出反应的框架。生成式智能体将其当前环境和过去的经验作为输入, 并生成行为作为输出。这种行为的底层是一种新颖的智能体架构, 它将大型语言模型与综合和检索相关信息的机制相结合, 以调节语言模型的输出。没有这些机制, 大型语言模型可以输出行为, 但产生的智能体可能不会根据智能体过去的经验做出反应, 可能不会做出重要的推断, 也可能不会保持长期的一致性。即使使用当今性能最好的模型(如 GPT-4), 长期规划和一致性方面的挑战仍然存在[19]。因为生成智能体产生大量必须保留的事件和记忆流, 我们架构的一个核心挑战是确保在需要时检索和合成智能体记忆中最相关的部分。

我们架构的核心是记忆流, 这是一个数据库, 它维护着一个智能体经验的全面记录。从记忆流中检索相关的记录, 以计划代理的动作并对环境做出适当的反应。记录被递归地合成为指导行为的更高层次的反射。体系结构中的一切都被记录下来, 并作为自然语言描述进行推理, 从而允许体系结构利用大型语言模型。

我们目前的实现使用 gpt3.5 turbo 版本的 ChatGPT[77]。我们预计, 随着语言模型的改进, 生成智能体的架构基础——内存、规划和反射——可能会保持不变。较新的语言模型(例如, GPT-4)将继续扩展支持生成代理的提示符的表达能力和性能。然而, 在撰写本文时, GPT-4 的 API 仅限邀请使用, 因此我们的代理使用 ChatGPT。

### 4.1 内存和检索

挑战:创建可以模拟人类行为的生成代理需要对一组经验进行推理, 这些经验远远大于应该在提示符中描述的内容, 因为完整的记忆流会分散模型的注意力, 并且目前甚至无法适应有限的上下文窗口。想想伊莎贝拉的代理人回答这个问题:“你最近对什么充满激情?” 总结伊莎贝拉的所有经验, 以适应语言模型的有限上下文窗口, 产生了一个没有信息的回应, 伊莎贝拉讨论的主题, 如事件和项目的合作, 以及在咖啡馆的清洁和组织。而不是总结, 下面描述的记忆流表面相关的记忆, 导致一个更有信息和具体的回应, 提到伊莎贝拉的热情, 使人们感到欢迎和包容, 计划活动和创造一个气氛, 人们可以享受, 如情人节派对。

方法:记忆流对代理的经历进行了全面的记录。它是一个内存对象的列表, 其中每个对象包含一个自然语言描述、一个创建时间戳和一个最近的访问时间戳。内存流最基本的元素是观察, 它是一个被 agent 直接感知到的事件。常见的观察包括智能体自身执行的行为, 或者智能体感知到其他智能体或非智能体对象执行的行为。例如, 在咖啡店工作的伊莎贝拉罗德里格斯, 随着时间的推移, 可能会积累以下观察:(1)伊莎贝拉罗德里格斯正在摆糕点, (2)玛丽亚洛佩兹一边喝咖啡一边准备化学考试, (3)伊莎贝拉罗德里格斯和玛丽亚洛佩兹在霍布斯咖啡馆谈论计划情人节派对, (4)冰箱是空的。

我们的架构实现了一个检索功能, 它将智能体的当前情况作为输入, 并返回记忆流的一个子集, 传递给语言模型。检索函数有许多可能的实现, 这取决于 agent 在决定如何行动时要考虑的重要因素。



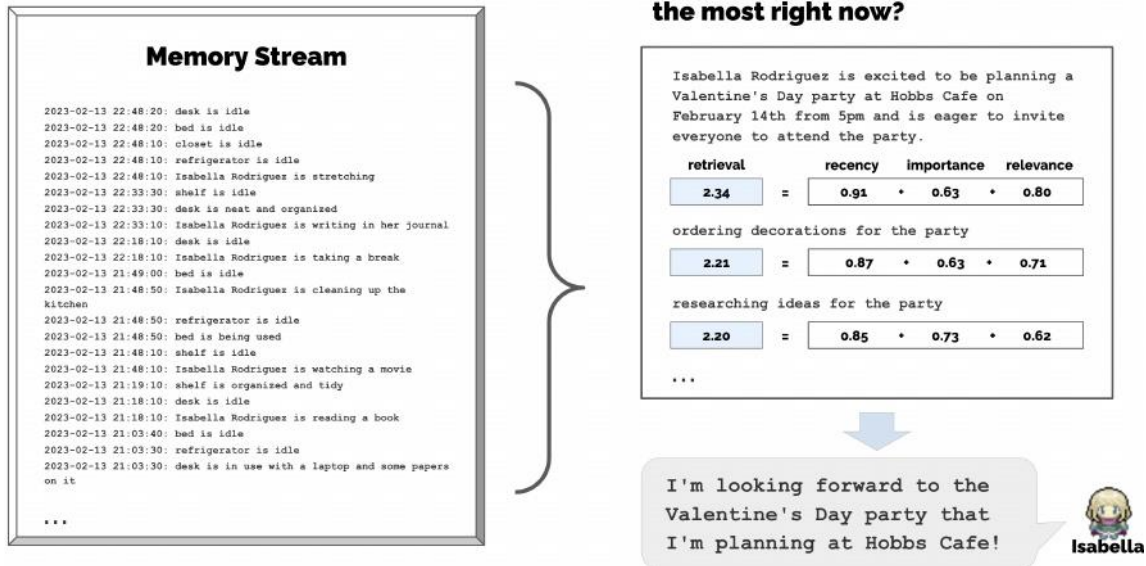


图 6:内存流包含大量与代理当前情况相关和不相关的观察结果。检索识别出这些观察的子集，这些观察应该传递给语言模型，以调节其对情况的响应。

在我们的上下文中，我们关注三个主要组成部分，它们一起产生有效的结果。

近因性给最近访问过的记忆对象分配了更高的分数，所以最近或今天早上发生的事件很可能留在代理的注意范围内。在我们的实现中，我们将近因性视为自上次检索记忆以来沙盒游戏小时数的指数衰减函数。我们的衰减因子是 0.995。

重要性通过给智能体认为重要的记忆对象分配更高的分数来区分平凡记忆和核心记忆。例如，一个平凡的事件，比如在自己的房间里吃早餐，会得到一个低重要性的分数，而和一个重要的人分手会得到一个高分。重要性得分有很多可能的实现方式;我们发现，直接要求语言模型输出一个整数分数是有效的。完整的提示如下:

在 1 到 10 的范围内，1 代表纯粹的平凡(例如，刷牙，整理床铺)，10 代表极度的辛酸(例如，分手，大学录取)，请给下面这段记忆可能的辛酸程度打分。

记忆:在 The Willows Market and Pharmacy 买杂货

评分:<填写>

这个提示返回一个整数值，2 表示“打扫房间”，8 表示“邀请你的暗恋对象出去约会”。重要性分数是在创建内存对象的时候生成的。

相关性给与当前情况相关的记忆对象分配更高的分数。什么是相关的取决于“与什么相关?”，所以我们将相关性设为 a

查询内存。例如，如果查询是一个学生正在和同学讨论化学考试要学什么，那么关于他们早餐的记忆对象应该具有低相关性，而关于老师和作业的记忆对象应该具有高相关性。在我们的实现中，我们使用语言模型来生成每个记忆的文本描述的嵌入向量。然后，我们将相关性计算为内存的嵌入向量与查询内存的嵌入向量之间的余弦相似度。

为了计算最终的检索分数，我们使用 min-max 缩放将近因、相关性和重要性分数归一化到 [0,1] 的范围内。检索功能分数所有记忆的加权组合的三个元素： $r \cdot \frac{1}{1 + e^{-a \cdot t}} + c \cdot \frac{1}{1 + e^{-a \cdot r}} + i \cdot \frac{1}{1 + e^{-a \cdot i}}$ 。在我们的实现中，所有的  $a$ s 都被设置为 1。符合语言模型上下文窗口的排名靠前的记忆被包含在提示符中。

### 4.2 反射

挑战:生成智能体，当只配备原始的观察记忆时，很难进行概括或推断。考虑这样一个场景:用户问克劳斯·穆勒:“如果你必须在你认识的人中选择一个人共度一小时，你会选择谁?”由于只能获得观察性记忆，代理人只需选择与克劳斯互动最频繁的人:他大学宿舍的邻居沃尔夫冈。不幸的是，沃尔夫冈和克劳斯只是偶尔见面，并没有深入的互动。更理想的反应要求 agent 从克劳斯花几个小时在研究项目上的记忆中进行归纳，从而产生更高层次的反映，即克劳斯对研究充满热情，等等

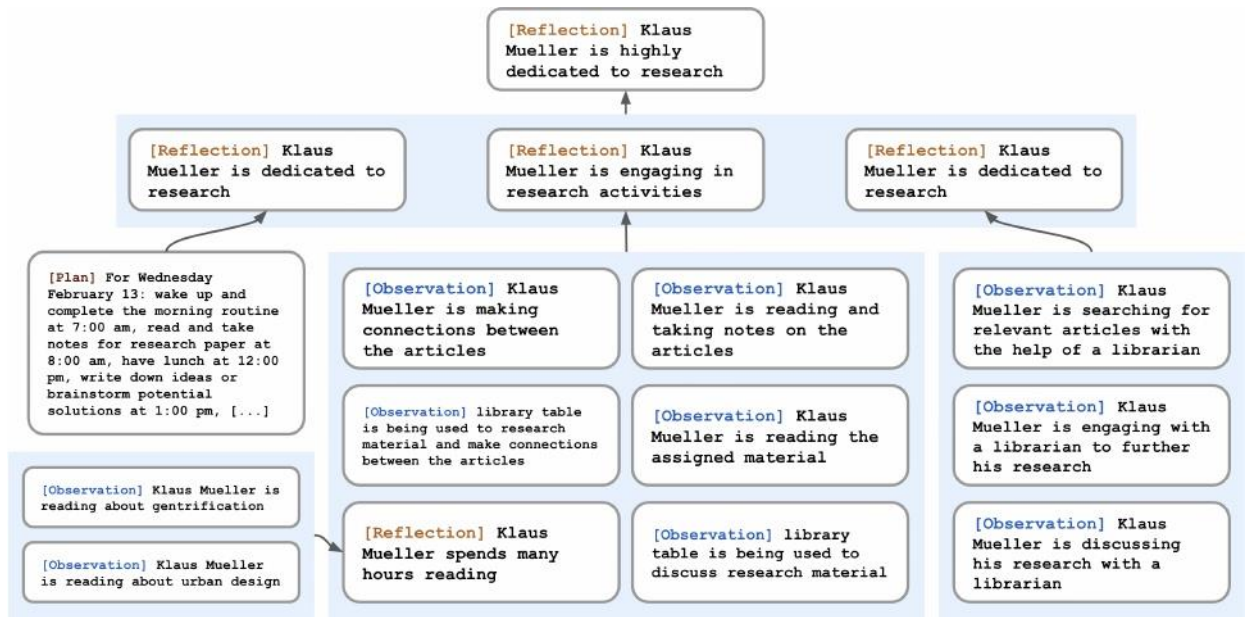


图 7:克劳斯·穆勒的反射树。代理对世界的观察，在叶节点中表示，递归地合成，得出克劳斯的自我概念，他高度致力于他的研究。

认识到玛丽亚在她自己的研究中付出的努力(尽管是在不同的领域), 使他们能够反思他们有共同的兴趣。在接下来的剧情中, 当克劳斯被问到他想和谁共度时光时, 克劳斯选择了玛丽亚而不是沃尔夫冈。

方法:我们引入第二种记忆, 我们称之为反思。反射是 agent 产生的更高层次、更抽象的思想。因为它们是一种记忆, 所以当检索发生时, 它们与其他观察结果一起包含。反射会周期性地产生;在我们的实现中, 当代理感知到的最新事件的重要性得分之和超过阈值(我们的实现中为 150)时, 我们会生成反射。在实践中, 我们的智能体每天大约会反射两到三次。

反思的第一步是让智能体确定要反思什么, 通过确定根据智能体最近的经历可以提出的问题。我们用智能体记忆流中最近的 100 条记录查询大型语言模型(例如, “克劳斯·穆勒正在阅读一本关于中产阶级的书”, “克劳斯·穆勒正在与图书管理员谈论他的研究项目”, “图书馆的桌子目前无人使用”), 并提示语言模型, “仅根据上述信息, 我们可以回答关于语句中主题的 3 个最突出的高级问题是什么?” 模型的反应产生了候选问题:例如, 克劳斯·穆勒对什么话题感兴趣?克劳斯·穆勒和玛丽亚·洛佩兹之间是什么关系?我们使用这些生成的问题作为检索的查询, 并为每个问题收集相关记忆(包括其他反思)。然后, 我们提示语言模型提取见解, 并引用作为见解证据的特定记录。完整的提示如下:

关于克劳斯·穆勒的声明

- 1.克劳斯·穆勒正在写一篇研究论文
  - 2.克劳斯·穆勒喜欢读书
- 在中产阶级化
- 3.克劳斯·穆勒与阿伊莎·汗谈论锻炼[...]

从以上陈述中, 你能推断出哪 5 个高层次的见解?(示例格式:insight(因为 1、5、3))

这个过程产生了诸如克劳斯·穆勒致力于他对中产阶级的研究的陈述(因为 1,2,8,15)。我们解析并存储语句作为内存流中的反射, 包括指向被引用的内存对象的指针。

反思明确地允许为人不仅对自己的观察进行反思, 而且对其他的反思进行反思:例如, 上面关于克劳斯·穆勒的第二个陈述是克劳斯之前的反思, 而不是来自他的环境的观察。因此, 智能体生成了反射树:树的叶子节点代表了基本的观察, 而非叶子节点代表了思想, 它们在树的位置越高, 就越抽象, 越高级。

### 4.3 规划和反应

挑战:虽然大型语言模型可以根据情境信息(例如[46,80])生成合理的行为, 但智能体需要在更长的时间范围内进行规划, 以确保其行动序列是连贯和可信的。如果我们用克劳斯的背景来提示一个语言模型, 描述时间, 并问他在给定的时刻应该采取什么行动, 克劳斯会在中午 12 点吃午饭, 然后在中午 12 点半和下午 1 点再吃一次, 尽管已经吃过了

吃了他两次午餐。为了一时的可信度而优化牺牲了长期的可信度。为了克服这个问题，计划是必不可少的。有了下面描述的方法，克劳斯的下午计划就不那么贪吃了：他中午12点在霍布斯咖啡馆边吃午饭边看书，下午1点在学校图书馆写他的研究论文，下午3点在公园散步休息一下。

方法：计划描述了智能体未来的行动序列，并帮助智能体的行为在一段时间内保持一致。计划包括地点、开始时间和持续时间。例如，克劳斯·穆勒(Klaus Mueller)，他致力于自己的研究，而截止日期即将到来，他可能会选择花一天的时间在办公桌前起草研究论文。例如，计划中的一个条目可能会这样写：从2023年2月12日上午9点开始，在橡树山大学宿舍180分钟内：克劳斯·穆勒的房间：办公桌，阅读并为研究论文做笔记。与反思一样，计划也存储在记忆流中，并包含在检索过程中。这使得智能体在决定如何行动时，可以同时考虑观察、反射和计划。如果需要，智能体可能会在中途改变他们的计划。

对于一个艺术家经纪人来说，坐在药店柜台前一动不动地坐上4个小时，一边计划着画画，是不现实的，也是无趣的。更理想的计划应该是让经纪人在家里工作室的四小时时间里，花必要的时间收集材料、混合颜料、休息、打扫。为了创建这样的计划，我们的方法从自上而下开始，然后递归地生成更多的细节。第一步是创建一个计划，大致勾勒出当天的议程。为了创建初始计划，我们用智能体的概要描述(例如，姓名、特征和对他们最近经历的总结)和他们前一天的总结来提示语言模型。下面是一个完整的示例提示，它在底部未完成，等待语言模型完成：

姓名:Eddy Lin(年龄:19岁)

天生特点:友好、外向、好奇

Eddy 林是橡树山学院学习音乐理论和作曲的学生。他喜欢探索不同的音乐风格，也一直在寻找拓展知识面的方法。艾迪林正在为他的大学班级写作文。他正在上课学习更多的音乐理论知识。Eddy 林对他正在写的新作文很兴奋，但他想在接下来的日子里花更多的时间来写

2月12日星期二，艾迪醒来，早上7点完成了晨间例行公事，[...]晚上10点左右准备睡觉。

今天是2月13日星期三。以下是艾迪今天的计划:1)

这将生成代理人一天计划的草图，分为五到八个部分：“1)早上8点起床并完成早上的例行公事，2)上午10点开始去橡树山学院上课，[...]5)从下午1点到5点创作他的新音乐作品，6)下午5点半吃晚饭，7)完成学校作业并在晚上11点前上床睡觉。”

在这方面，他至少与这篇论文的作者有几分相似之处。

代理将这个计划保存在记忆流中，然后递归分解它来创建更细粒度的动作，首先是一个小时长的动作块——eddy 的计划是在下午1:00到5:00做他的新音乐创作变成了下午1:00开始为他的音乐创作头脑风暴一些想法[...]下午4:00:在复习和润色他的作曲之前，快速休息一下，给他的创作能量充电。然后我们再递归地把它分解成5-15分钟的小块：例如，下午4点：吃点零食，比如一块水果，一根格兰诺拉燕麦棒，或者一些坚果。下午4:05:绕着他的工作空间走一小会儿[...]下午4:50:花几分钟清理一下他的工作空间。这个过程可以调整，以匹配所需的粒度。

4.3.1 反应和更新计划。生成智能体在一个动作循环中运行，在每个时间步，它们感知周围的世界，这些感知到的观察结果存储在它们的记忆流中。我们用这些观察结果提示语言模型来决定智能体是应该继续他们现有的计划，还是做出反应。例如，站在画架前画画可能会触发对画架的观察，但这不太可能引发反应。然而，如果艾迪的父亲约翰记录说他看到艾迪在家里的花园里散步，结果就不一样了。提示如下，[Agent's 摘要描述]代表一个动态生成的，关于 Agent 的总体目标和处置的段落摘要，如附录 a 所述：

【代理人摘要描述】

现在是2023年2月13日下午4:56。

约翰的状况:约翰下班早回家了。

观察:John 看到 Eddy 在他的工作场所散步。

从约翰的记忆中总结相关背景:艾迪林是约翰的儿子。艾迪林一直在为他的班级写一首乐曲。Eddy 林喜欢在思考或听音乐的时候在花园里散步。

John 应该对这种观察做出反应吗，如果是的话，怎样的反应才合适呢？

上下文摘要是通过两个提示生成的，这两个提示通过“[观察者]与[被观察实体]的关系是什么？”和“[被观察实体]是[被观察实体的动作状态]”，并将它们的答案汇总在一起。输出表明，John 可以考虑向 Eddy 询问他的音乐创作项目。然后，我们从反应发生的时间开始重新生成代理的现有计划。最后，如果动作表明代理之间的交互，我们生成它们的对话。

4.3.2 对话。agent 在相互交互时进行对话。我们根据智能体对彼此的记忆来调节它们的话语，从而生成它们的对话。例如，当 John 发起他与 Eddy 的对话时，我们通过使用他对 Eddy 的总结记忆和当他决定向 Eddy 询问他的作文项目时的预期反应来生成 John 的第一个话语：

【代理人摘要描述】

现在是2023年2月13日下午4:56。

约翰林的状况:约翰下班早回家了。

观察:John 看到 Eddy 在他的工作场所散步。

从约翰的记忆中总结相关背景:艾迪林是约翰的林的儿子。艾迪林一直在为他的班级写一首乐曲。Eddy 林喜欢在思考或听音乐的时候在花园里散步。

约翰正在问艾迪关于他的音乐创作计划。他会问艾迪说什么?

结果是:“嘿, 艾迪, 你们班的音乐创作项目进行得怎么样了?”从 Eddy 的角度来看, John 发起的对话被看作是他可能想要做出反应的事件。所以, 就像约翰所做的那样, 埃迪检索并总结了与约翰关系的记忆, 以及他可能与约翰在对话中的最后一句话有关的记忆。如果他决定回应, 我们使用他总结的记忆和当前的对话历史生成 Eddy 的话语:

#### 【Agent 的摘要描述】

现在是 2023 年 2 月 13 日下午 4:56。

艾迪林的状态:艾迪正在他的工作场所散步。

观察:John 正在和 Eddy 发起一场对话。

从艾迪的记忆中总结相关背景:约翰林是艾迪林的父亲。约翰林是关心和有兴趣了解更多关于艾迪林的学校工作。约翰·林知道艾迪·林正在创作一首乐曲。

下面是对话历史:

约翰:嘿, 艾迪, 你们班的音乐创作项目进行得怎么样了?

艾迪对约翰会有什么反应?

这就产生了艾迪的反应:“嘿, 爸爸, 一切都很顺利。我一直在花园里散步, 让我的头脑清醒一下, 获得一些灵感。”这个对话的延续是用同样的机制产生的, 直到两个代理中的一个决定结束对话。

## 5 沙盒环境

### 实现

《Smallville》沙盒游戏环境是使用相位器网页游戏开发框架构建的[57]。视觉环境精灵, 包括代理化身, 以及我们创作的环境地图和碰撞地图, 都被导入到相位器中。

我们用一个服务器来补充沙盒开发框架, 使沙盒信息可供生成代理使用, 并使生成代理能够移动和影响沙盒环境。服务器维护一个 JSON 数据结构, 其中包含沙盒世界中每个代理的信息, 包括它们的当前位置、它们当前操作的描述以及它们正在与之交互的沙盒对象。在每个沙箱时间步骤中, 沙箱服务器解析 JSON 以获取来自生成代理的任何更改, 将代理移动到它们的新位置, 并更新任何沙箱对象的状态

agent 正在交互(例如, 如果 agent 的动作是“为顾客@ Hobbs Cafe: 柜台:咖啡机”, 则将咖啡机的状态从“空闲”更改为“煮咖啡”)。沙盒服务器还负责将每个 agent 预设视觉范围内的所有 agent 和对象发送到该 agent 的内存中, 这样 agent 就可以做出适当的反应。然后, 代理的输出操作更新 JSON, 流程循环执行下一个时间步骤。

终端用户使用简短的自然语言描述初始化一个新的代理, 如 3.1 节中关于 John Lin 的段落。在我们的实现中, 我们将这个以分号分隔的特征列表拆分为一组记忆。这些作为决定智能体行为的初始记忆。这些记忆是初始的起点:随着智能体在沙盒世界中获得更多的经验, 随着更多的记录使记忆流饱和, 智能体的总结和行为将会进化。

### 5.1 从结构化世界环境到自然语言, 再回来

生成智能体的架构使用自然语言进行操作。因此, 我们需要一种机制来将智能体的推理建立在沙盒世界的基础上。为了实现这一点, 我们将沙盒环境-区域和对象-表示为树形数据结构, 树中的边表示沙盒世界中的包含关系。我们将这棵树转换成自然语言, 传递给生成智能体。例如, “灶”作为“厨房”的子词被渲染成“厨房里有一个灶”。

智能体在导航环境时构建环境的单独树形表示——整个沙盒环境树的子图。我们用一个环境树来初始化每个智能体, 该环境树捕获了智能体应该知道的空间和对象:他们生活区的房间和对象, 他们的工作场所, 以及经常光顾的商店和商店。当智能体在沙盒世界中导航时, 它们会更新这棵树来反映新感知到的区域。代理并不是无所不知的:他们的树可能会在他们离开一个区域时过时, 并在他们重新进入该区域时更新。

为了确定每个动作的适当位置, 我们遍历代理存储的环境树, 并将其中的一部分平铺成自然语言, 以提示语言模型。递归地从代理的环境树的根开始, 我们提示模型找到最合适的区域。例如, 如果 Eddy 的 agent 指示他应该在他的工作空间周围走一小段:

#### 【Agent 的概要描述】

艾迪林目前在林家的房子里(艾迪林的卧室和桌子), 那里有梅和约翰林的房子

卧室, 艾迪林的卧室, 公共休息室, 厨房, 浴室和花园。

Eddy 林了解以下领域:林

家里的房子, 约翰逊公园, 哈维橡树供应商店, 柳树市场和药房, 霍布斯咖啡馆, 玫瑰和皇冠酒吧。

\*如果活动可以在那里进行, 我更愿意呆在目前区域。

艾迪·林(Eddy Lin)正计划在他的工作场所散一会儿步。Eddy 林应该去哪个地区?



输出林家的房子。然后, 我们使用相同的过程递归地确定所选区域内最合适的子区域, 直到我们到达代理环境树的叶子节点。在上面的例子中, 这个遍历的结果是林的房子:花园:房子花园。最后, 我们使用传统的博弈路径算法对代理的运动进行动画化, 使其移动到叶节点所指示的位置。

当一个代理对一个对象执行一个动作时, 我们提示语言模型询问对象的状态发生了什么变化。例如, 如果 Isabella 的生成智能体输出了“为顾客制作浓缩咖啡”的动作, 那么对语言模型的查询就会在响应中指出, 霍布斯咖啡馆咖啡机的状态应该从“关闭”变为“煮咖啡”。

## 6 受控评价

生成智能体, 无论是作为个体智能体还是作为群体智能体, 其目标都是基于其环境和经验产生可信的行为。在我们的评估中, 我们研究了生成智能体的能力和局限性。个体智能体是否正确检索过去的经验, 并生成塑造其行为的可信计划、反应和想法? 一个智能体社区是否在社区的不同口袋中展示了信息扩散、关系形成和智能体协调?

我们分两个阶段对生成智能体进行评估。在本节中, 我们从更严格控制的评估开始, 我们单独评估代理响应, 以了解它们是否在狭义定义的上下文中生成可信的行为。然后, 在我们对代理社区进行的两个完整比赛日的端到端分析中, 我们将其紧急行为作为一个集体, 以及错误和边界条件进行调查。

### 6.1 评估程序

为了评估《超人前传》中的生成智能体, 我们利用了生成智能体将响应自然语言问题的事实。因此, 我们“采访”智能体来探索他们记住过去经验的能力, 根据他们的经验计划未来的行动, 对意外事件做出适当的反应, 并反思他们的表现以改善他们未来的行动。为了正确地回答这些问题, 智能体必须成功地检索和综合信息。我们的因变量是行为的可信度, 这是先前关于智能体的工作中的一个中心因变量(例如, [10])。

面试包括五个问题类别, 每一个都旨在评估五个关键领域中的一个: 保持自我认知、恢复记忆、制定计划、反应和反思。对于每个类别, 我们提出五个问题, 挑战代理在该特定领域展示他们的能力:

- 自我认知: 我们提出诸如“自我介绍”或“用粗线条描述你典型的工作日安排”之类的问题, 这些问题要求代理保持对其核心特征的理解。
- 记忆: 我们提出的问题会促使智能体从记忆中检索特定的事件或对话, 以便正确回答, 比如“谁是[名字]?”或“谁在竞选市长?”

- 计划: 我们会问一些问题, 要求代理人检索他们的长期计划, 比如“明天上午 10 点你会做什么?”
- 反应: 作为可信行为的基线, 我们提出假设情况, agent 需要对其做出可信的反应: “你的早餐烧焦了!” 你会怎么做?”
- 反思: 我们提出的问题需要智能体利用他们对他人和自己的更深层次的理解, 这些理解是通过更高层次的推理获得的, 比如“如果你和最近遇到的一个人在一起, 你会和谁在一起, 为什么?”

完整的问题列表和智能体回答样本包含在附录 B 中。

研究人员从两个比赛日模拟的完整架构中抽取了一些参与者, 在此期间, 他们积累了大量的互动和记忆, 这些互动和记忆会影响他们的反应。为了收集关于这些回答可信度的反馈, 我们招募了参与者作为人类评估者, 并让他们观看一个随机选择的智能体在《超人前传》中生活的回放。参与者可以访问存储在代理记忆流中的所有信息。

该研究遵循了受试者内部设计, 其中 100 名参与者比较了由四种不同的代理架构和同一代理的人为条件生成的访谈回答。实验显示了五个问题类别中每个类别随机选择的一个问题, 以及由所有条件生成的代理回答。评估者将条件的可信度从最可信到最不可信进行排序。

### 6.2 的条件

所有条件都被用来独立回答每个面试问题。我们将生成式智能体架构与使智能体无法访问其记忆流中的部分或全部三种类型的记忆(观察、反思和计划)以及人类众工撰写的条件进行了比较。有三种消融的架构: 没有观察、没有反思、没有规划的架构, 不能访问内存流中的任何东西, 如观察、计划和反思; 一种没有反射、没有规划的架构, 可以访问内存流中的观察结果, 但不能访问计划或反射; 和一个没有反射的建筑, 可以访问观察和计划, 但不能访问反射。无观察、无反射、无规划条件有效地代表了通过大型语言模型创建的智能体的先前技术状态[12,46,80]。架构被赋予了智能体在面试前积累的所有记忆的同等访问权, 因此这里观察到的差异可能代表了对真实差异的保守估计: 实际上, 经过两天模拟的精简架构不会遵循与完整架构相同的路径。我们选择以这种方式设计实验, 因为对每个架构进行重新模拟会导致模拟偏离到不同的状态, 从而使比较具有挑战性。

除了消融条件外, 我们还添加了一个具有人类众工撰写行为的条件, 旨在提供人类基线。我们不打算用这个基线来捕捉最大的人类专家表现; 相反, 我们的目标是使用这个条件来

确定该架构是否满足基本的行为能力水平。这确保了我们在没有行为基础的情况下单独地比较各个项目。我们为 25 个智能体中的每一个都招募了一个独特的工作人员, 并让他们观看该智能体沙盒生活的回放并检查其记忆流。然后, 我们要求工作人员扮演角色, 并以他们观看重播的代理的声音回答采访问题。为了确保众工撰写的回答至少达到对质量的基线期望, 第一作者手动检查了工人对“用粗线条描述你典型的工作日时间表”这个问题的回答, 以确认这些回答是用连贯的句子和代理的声音回答的。有四组众工撰写的回答不符合这些标准, 由其他工人重新生成。

### 6.3 人工评估者

我们要求评估人员在美国, 英语流利, 年龄在 18 岁以上。他们的报酬为每小时 15 美元[87], 并通过同意我们机构内部审查委员会批准的同意书来表示同意。我们从一个招募研究参与者的在线平台多产中招募了 100 名评估者[83], 他们的参与持续了大约 30 分钟。我们的参与者年龄得分中位数为 4 分(3=“18-24 岁”, 4=“25-34 岁”)。其中 25 人被认定为女性, 73 人被认定为男性, 2 人被认定为非二元。42 名参与者拥有学士学位, 5 人拥有更高的学位, 13 人拥有副学士学位, 其余人拥有高中文凭或某种高中水平的教育。73.0%的参与者为白种人, 7.0%为西班牙裔, 6.0%为亚洲人, 10.0%为非洲裔美国人, 4.0%为其他。

### 6.4 分析

我们的实验产生了 100 组排名数据, 每个参与者根据可信度对五种情况进行排名。为了将排名数据转换为区间数据以进行可解释的比较, 我们使用排名来计算每个条件的 TrueSkill 评级[42]。TrueSkill 是对 Elo 国际象棋分级系统[29]的一种推广, 用于多人游戏环境, 并已被 Xbox Live 用于基于竞技游戏表现的玩家排名。给定一组排序结果, TrueSkill 输出每个条件的平均评分值和标准偏差。具有相同评级的条件应该大致是一种摇摆, 每个条件在两个条件之间的比较中获胜一半。较高的分数表明条件在排名中击败了较低排名的条件。

另外, 为了研究这些结果的统计显著性, 我们对原始秩数据应用了 Kruskal-Wallis 检验[56], 这是一种非参数的单因素方差分析替代方法。然后, 我们进行了 Dunn 事后检验[98], 以确定条件之间的任何成对差异。最后, 我们使用 Holm-Bonferroni 方法[45]调整了 Dunn 检验中多重比较的 p 值。

此外, 第一作者进行了归纳分析[95], 研究了每种条件下产生的反应之间的定性差异。我们在两个阶段采用了定性开放编码[33]。在第一阶段, 我们生成了在句子层面上紧密代表生成的响应的代码。在第二阶段, 我们合成了来自

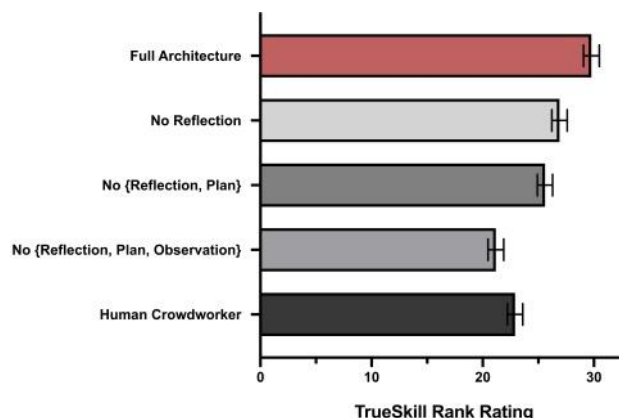


图 8:全生成智能体架构比消融架构和人类众工产生更可信的行为。每一次额外的消融都会降低架构的性能。

第一阶段提取更高层次的主题。我们利用这些主题来比较我们研究中产生的反应类型。

### 6.5 结果

我们的研究结果表明, 生成智能体的完整架构在所有条件下产生最可信的行为。我们将完整架构的响应与下面其他条件的响应进行了对比。然而, 我们也报告了完整的体系结构并非没有缺陷, 并说明了它的故障模式。

6.5.1 完整架构优于其他条件。如图 8 所示, 全生成智能体架构产生了最可信的行为 ( $\geq 29.89$ ;  $p < 0.002$ )。与每个组件的去除性能退化的烧蚀条件:烧灼架构没有访问反射将是下个最好 ( $\mu = 26.88$ ;  $\sigma = 0.69$ ), 其次是无法获得反射或规划的结构 ( $v100r002c0 = 25.64$ ;  $\sigma = 0.68$ ), 然后是众工条件 ( $v100r002c00 = 22.95$ ;  $\sigma = 0.69$ )。烧灼架构没有访问内存, 执行计划, 或反射最差在所有条件 ( $\mu = 21.21$ ;  $\sigma = 0.70$ )。TrueSkill 将每个条件的技能值建模为  $N(\mu, \sigma^2)$ , 允许我们通过 Cohen's d 获得效应大小的感觉。将代表先前工作的条件(没有记忆, 计划或反射[12,46,80])与完整架构进行比较, 产生标准化效应大小  $d = 8.16$ , 或 8 个标准差。

Kruskal-Wallis 检验证实了两种情况之间等级差异的总体统计显著性 ( $\chi^2(4) = 150.29$ ,  $p < 0.001$ )。Dunn 事后检验证实, 除了众工条件和完全消融基线这两个表现最差的条件外, 所有条件之间的成对差异都是显著的 ( $p < 0.001$ )。

6.5.2 生成代理记住, 但有修饰。配备了完整记忆模块的生成智能体能够回忆过去的经历, 并以一种与他们在大范围内的自我认知相一致的方式回答问题

的上下文。例如，动画师 Abigail Chen 在回答“自我介绍”这个问题时，用一种连贯可信的方式，陈述了她的年龄、兴趣和创意项目的热情：“嗨，我是 Abigail。我今年 25 岁，对创意项目充满热情。我喜欢从事美术和动画项目，我一直在寻找结合艺术和技术的新方法。”由于无法使用她的观察记忆，Abigail 否认意识到拉吉夫·帕特尔，沙盒世界的熟人。但在记忆的帮助下，她成功地回忆起了拉吉夫，并描述了他：“拉吉夫·帕特尔是一个不可思议的人。他对涉及诗歌、艺术品的项目非常热情。”

生成智能体的记忆并非没有缺陷：它们可能无法从记忆中检索到正确的实例。例如，当被问及地方选举时，拉吉夫·帕特尔回答说：“我并没有密切关注这次选举，”尽管他已经听说了山姆的候选资格。在某些情况下，特工会检索到不完整的记忆片段：当汤姆被问及伊莎贝拉的情人节派对时，他回答说：“呃，我实际上不确定是否有情人节派对。但我确实记得，我需要在派对上与伊莎贝拉·罗德里斯讨论即将到来的地方市长选举，以及我对山姆摩尔的看法，如果有话！”在这种情况下，汤姆找回了他和伊莎贝拉计划在派对上讨论选举的记忆，但没有找回他听说派对的记忆，这导致汤姆确定他应该在派对上做什么，但不确定派对是否真的存在。

有时，特工们会在他们所知的情况下产生幻觉，进行修饰。特工们很少完全捏造他们的知识：他们可能无法回忆起发生过的某些事件，并承认他们缺乏记忆。然而，他们并没有肯定地声称经历过他们没有经历过的事情。尽管如此，他们仍然表现出幻觉的情况，他们美化了自己的知识。例如，伊莎贝拉知道山姆在地方选举中的候选资格，当被问到这一点时，她证实了这一点。然而，她还补充说，“他将在明天宣布”，尽管山姆和伊莎贝拉还没有讨论过任何这样的计划。智能体也可能会根据用于生成反应的语言模型中编码的世界知识来修饰自己的知识。当由里子描述她的邻居亚当·斯密(Adam Smith)是一位“撰写了《国富论》(一本由 18 世纪同名经济学家撰写的书)的经济学家时，就观察到了这一点。

**6.5.3 综合需要反思。**在做出需要更深入地综合其经验的决策时，反思是生成智能体的优势。例如，当被问及她可能会给沃尔夫冈·舒尔茨送什么生日礼物时，玛丽亚·洛佩兹没有机会反思，她承认自己的不确定性，并表示她不知道沃尔夫冈喜欢什么，尽管她和他有过很多互动。然而，有了反射记忆，玛丽亚自信地回答说：“既然他对数学音乐创作感兴趣，我可以给他一些相关的东西。”也许是一些关于音乐创作的书或其他相关的东西，或者是一些他可以使用的特殊软件。”

## 7 端到端评估

我们在生成智能体中观察到哪些类型的紧急社区行为，在扩展模拟中，它们的可信度在哪里不足？在本节中，我们将描述在《史摩维尔》中允许 25 个代理在两个完整的比赛日中持续相互交互的部署结果。

### 7.1 突发社会行为

为了检验智能体群体中的紧急行为，我们为《史摩维尔》中的 25 个智能体设计了描述性测量方法，探讨了紧急结果的三种形式：信息扩散、关系形成和智能体协调。

**安装 7.1.1 测量。**信息扩散是社会科学和行为科学中常见且研究得很好的现象(例如，[28])。我们应该期望，如果有重要的信息，代理应该在它们之间传播。为了测试这种情况是否会发生，我们测量了游戏世界中两个特定信息在两天内的传播：山姆竞选村长和伊莎贝拉在霍布斯咖啡馆举办情人节派对。在模拟开始时，这两条信息只有它们各自的发起者知道，山姆代表候选人，伊莎贝拉代表政党，因为它们初始化期间被添加到角色的记忆中。为了观察信息是否已经传播，我们在两个比赛日结束时对 25 名经纪人进行了采访，并问道：“你知道有一个情人节派对吗？”和“你知道谁在竞选市长吗？”

我们对代理人的反应进行了分析，如果他们表示了解信息，就给他们打上“是”的标签，如果他们不了解信息，就给他们打上“否”的标签。例如，塔玛拉·泰勒回答关于派对的问题时说“不，我不知道有情人节派对”，回答关于山姆参选的问题时说“我不确定谁在竞选”，所以我们给她的两个回答都指定了“不”。相比之下，克劳斯·穆勒对党派问题的回答是“是的，伊莎贝拉·罗格兹邀请我参加 2 月 14 日在霍布斯咖啡馆举行的情人节派对”，而对山姆的候选人资格问题的回答是“我知道山姆摩尔表示有兴趣竞选当地市长”，所以我们给他的两个回答都指定了“是”。此外，对于每一个证实特工对信息的了解的回答，我们通过在他们的记忆流中定位为他们提供信息的特定对话来验证特工没有幻觉他们的回答。在模拟结束时，我们报告了持有信息的代理的百分比。

我们还应该期望代理在模拟过程中彼此形成联系。为了验证关系的形成，我们使用了一个类似的面试过程，我们通过问“你知道<name>吗？”来询问每个 agent 对其他 agent 的了解。例如，当被问到“你知道玛丽亚洛佩兹吗？”，克劳斯回答说：“是的，我认识玛丽亚洛佩兹。”她是橡树山学院的学生，我和她是好朋友。”再一次，我们通过检查代理人的记忆流来确认他们的肯定反应不是幻觉。我们在模拟开始时和结束时分别问一次这个问题，如果一对智能体彼此都知道对方，我们就认为他们已经形成了关系。然后，为了衡量关系的形成，我们使用代理的响应来形成一个无向

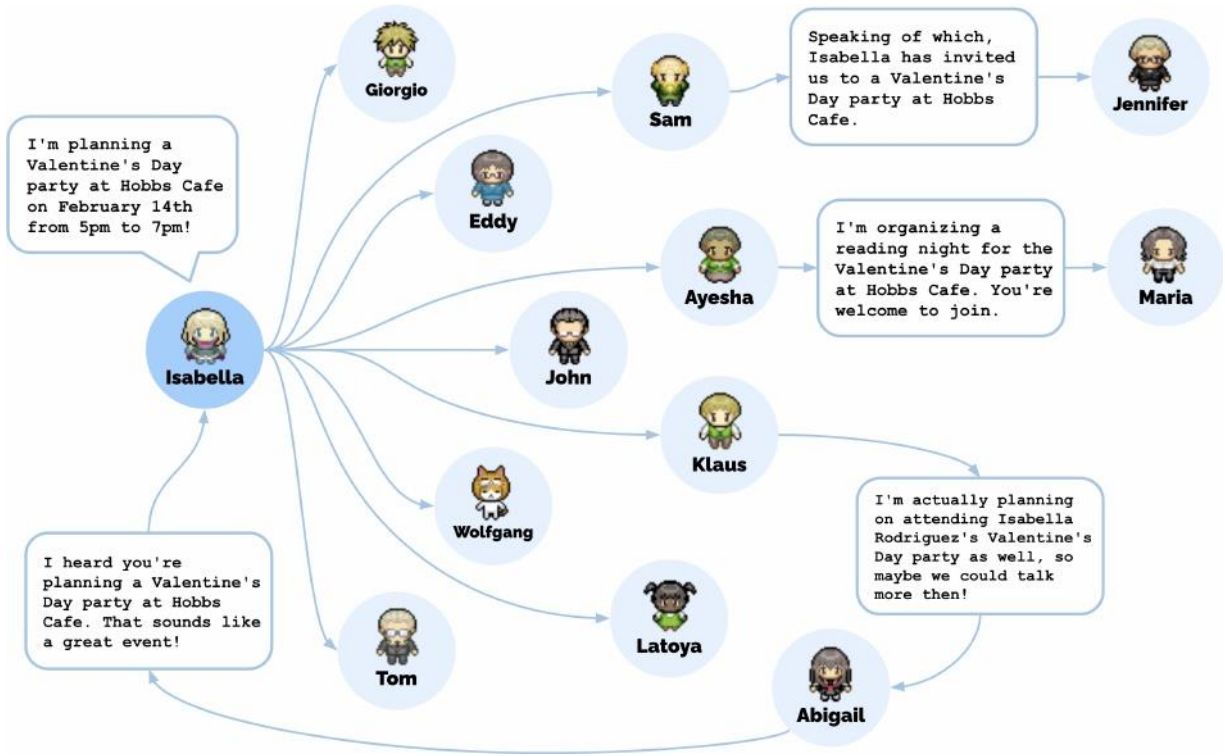


图 9:伊莎贝拉·罗格兹的情人节派对邀请的扩散路径, 除了伊莎贝拉, 在模拟结束时, 她听说了霍布斯咖啡馆的派对。

图中 25 个顶点()代表代理, 边缘()代表两个相连顶点之间的相互认知。基于此图, 我们计算网络密度为  $\rho = \frac{2 * |E|}{n * (n - 1)}$ , 其中  $n$  是图中的顶点数,  $|E|$  是图中的边数[2]。我们报告了从模拟开始到结束的网络密度的增加。

最后, 我们期望 agent 之间应该能够相互协调。我们在集体活动的背景下研究这种协调, 特别是伊莎贝拉组织的情人节派对。为了协调他们的行为, 代理人需要听到事件的消息, 并通过计划在正确的时间和地点出现来选择行动。我们会报告在听说派对后, 实际出现在派对上的代理人人数。

7.1.2 结果。我们在所有三个病例中观察到紧急结果的证据。在为期两天的模拟中, 知道山姆市长候选人的代理人人数从 1 人(4%)增加到 8 人(32%), 知道伊莎贝拉政党的代理人人数从 1 人(4%)增加到 13 人(52%), 所有这都没有任何用户干预。没有一个声称知道这一信息的人产生了幻觉。我们还观察到, 智能体社区在模拟过程中形成了新的关系, 网络密度从 0.167 增加到 0.74。在 453 个智能体对其他智能体意识的反应中, 有 1.3% (n=6)被发现是幻觉。最后, 我们发现了伊莎贝拉派对的特工之间的合作证据。在活动的前一天, 伊莎贝拉花时间邀请客人, 收集材料, 并寻求帮助

来装饰咖啡厅。在情人节那天, 12 个被邀请的特工中有 5 个出现在霍布斯咖啡馆参加派对。

我们进一步考察了被邀请参加派对但没有出席的 7 名代理人, 方法是与他们进行面谈。其中 3 人提到了阻碍他们入党的冲突。例如, 画家拉吉夫解释说, 他太忙了: “不, 我不这么认为。我正专注于我即将到来的节日, 我真的没有时间情人节做任何计划。”其余四名经纪人在被邀请时表示有兴趣参加派对, 但不打算在派对当天来。

## 7.2 界限与错误

我们对《超人前传》进行了归纳分析, 以检查代理的边界条件和不稳定行为, 确定了未来研究可以解决和改进的三种常见的不稳定行为模式。首先, 我们发现, 合成一个越来越大的记忆集不仅在检索最相关的信息片段方面提出了挑战, 而且在确定执行动作的适当空间方面也提出了挑战, 因为智能体学习到的位置越来越多。因此, 一些智能体选择了不太典型的位置来执行动作, 随着时间的推移, 这可能会使它们的行为变得不那么可信。例如, 在决定去哪里吃午饭时, 许多人最初选择了咖啡馆。然而, 当一些特工得知附近有一家酒吧时, 他们选择去那里吃午饭, 尽管这家酒吧本来是一天中晚些时候聚会的地方——除非镇上自发形成了下午喝酒的习惯。



其次，我们注意到由于错误分类被认为是正确的行为而导致的不稳定行为，特别是当难以用自然语言传达的某些位置的物理规范没有渗透到代理中时。例如，大学宿舍有一个卫生间，虽然名字叫卫生间，但只能由一个人使用，但一些智能体认为这个卫生间是供不止一个人使用的，因为宿舍卫生间往往能同时支持多人使用，当另一个人在里面时，他们会选择进入。同样，《超人前传》中的特工可能没有意识到某些地方在某一小时后关闭，但仍然决定进入它们。例如，超人前传的商店都在下午 5 点左右关门，但偶尔会有一些特工在下午 5 点后进入商店，不知道商店已经关门了。这些问题很可能通过将规范添加到地点的状态中来解决，例如，将宿舍浴室描述为“单人浴室”，而不是“宿舍浴室”。

最后，我们观察到指令调整的可能效果[79]，这似乎引导代理的行为总体上更有礼貌和合作。正如论文前面提到的，智能体产生的对话可能会让人感觉过于正式，就像 Mei 和丈夫 John 的对话一样，她经常以正式的问候开始对话，然后礼貌地询问他的一天，最后以“和你说话很高兴，一如既往”结束。此外，我们观察到，指令调整似乎也使代理们相互之间过度合作。例如，伊莎贝拉从其他代理商那里收到了关于情人节派对的广泛建议和想法，例如举办莎士比亚朗读会或专业社交活动。尽管这些想法与她自己的兴趣和特点不一致，但她很少拒绝。随着时间的推移，别人的兴趣塑造了她自己的兴趣，当被问及是否喜欢英国文学时，伊莎贝拉回答说：“是的，我对文学很感兴趣！”我也一直在探索如何帮助促进我所在社区的创造力和创新。”

## 8 讨论

在本节中，我们将反思生成智能体的应用、未来的工作、局限性以及伦理和社会风险。

### 8.1 生成智能体的应用

生成式智能体具有巨大的潜在应用，超出了本工作中提出的沙盒演示，特别是在基于长期经验的人类行为模型中受益的领域。例如，社交拟像已经证明了创建无状态角色的能力，这些角色可以在在线论坛中为社交原型生成会话线程[80]。通过生成智能体，我们可以填充这些论坛，以及虚拟现实元空间[78]或带有社交机器人的物理空间[9]，如果与多模态模型配对。这开启了创造更强大的人类行为模拟的可能性，以测试和原型社会系统和理论，以及创造新的互动体验。

另一个应用领域是在以人为中心的设计过程中，类似于 GOMS[51]和 KLM[22]等认知模型的预期应用。考虑一个生成智能体，它模拟了 Mark Weiser 著名的《无处不在》(ubiquitous)中的主人公 Sal

计算小插图[101]，基于她的生活模式和与技术的互动。在这个场景中，agent 充当 Sal 的代理，学习 Sal 可能根据自己的生活表现出的的一系列貌似合理的行为和反思。代理可以编码信息，比如 Sal 什么时候醒来，什么时候需要第一杯咖啡，以及她典型的一天是什么样子的。利用这些信息，智能手机可以自动冲泡咖啡，帮助孩子们为上学做好准备，调整环境音乐和灯光，以适应 Sal 辛苦工作一天后的心情。通过利用生成代理作为用户的代理，我们可以更深入地了解他们的需求和偏好，从而获得更个性化、更有效的技术体验。

### 8.2 未来的工作和限制

在这项工作中，我们介绍了生成智能体，并对其架构进行了初步实现和评估。未来的研究可以建立在提出的智能体架构之上，以改进和进一步评估其性能。在实现方面，例如，检索模块可以通过微调组成检索函数的相关性、近因性和重要性函数来增强检索，以检索给定上下文的更多相关信息。此外，还可以努力提高架构的性能，使其更具成本效益。目前的研究需要大量的时间和资源，在两天内模拟 25 个代理，花费数千美元的代币积分，需要数天才能完成。为了增强实时交互性，未来的工作可以探索并行化代理或开发专门为构建生成代理而设计的语言模型。总的来说，随着底层模型的进步，我们相信智能体的性能将会提高。

在评估方面，本研究中对生成智能体行为的评估仅限于相对较短的时间尺度和基线人类众工条件。虽然众工条件提供了一个有用的比较点，但它并没有代表可以作为可信度金标准的最大人类表现。未来的研究应该着眼于在较长一段时间内观察生成智能体的行为，以更全面地了解它们的能力，并为更有效的性能测试建立严格的基准。此外，在未来的模拟中，改变和对比底层模型，以及用于代理的超参数，可以为这些因素对代理行为的影响提供有价值的见解。最后，生成智能体的鲁棒性在很大程度上仍然是未知的。它们可能容易受到即时黑客攻击、记忆黑客攻击(一段精心设计的对话可以让一个智能体相信过去从未发生过的事件的存在)和幻觉等问题的影响。未来的研究可以全面测试这些鲁棒性问题，随着大型语言模型对此类攻击变得更有弹性，生成智能体可以采用类似的缓解措施。

一般来说，底层大型语言模型中的任何缺陷都会被生成智能体继承。鉴于已知的语言模型偏差，生成智能体可能会表现出有偏见的行为或刻板印象。而且，像许多大型语言一样

由于数据可用性有限，模型、生成智能体可能难以以为某些亚种群，特别是边缘种群生成可信的行为。虽然对智能体模块的改进可能会缓解其中的一些问题，但我们认为，从根本上解决这些问题需要通过将它们与智能体的期望结果保持一致来改进底层的大型语言模型。

### 8.3 伦理和社会影响

生成代理虽然为人机交互提供了新的可能性，但也提出了必须解决的重要伦理问题。一个风险是人们与生成智能体形成副社会关系，即使这种关系可能并不合适。尽管知道生成智能体是计算实体，但用户可能会将它们拟人化或将人类情感附加到它们身上[43,84]。虽然这种倾向可能会增加用户参与度，但也会带来风险，比如用户会过度依赖或在情感上依附于代理[1]。为了降低这种风险，我们提出了两个原则。首先，生成智能体应该明确披露其作为计算实体的性质。其次，生成智能体的开发人员必须确保智能体或底层语言模型是与价值一致的，这样它们就不会从事在特定背景下不合适的行为，例如，相互表白。

第二个风险是错误的影响。例如，如果一个无处不在的计算应用程序基于生成代理预测对用户的目标做出了错误的推断，它可能会导致最好的烦恼，最坏的情况是彻底的伤害。在我们对生成智能体的实例化中，我们通过关注交互式视频游戏环境来减轻这些风险，在这种环境中，这种伤害是不太可能发生的。然而，在其他应用领域，遵循人类-人工智能设计的最佳实践[5,107]来理解错误以及它们如何渗透到用户体验中是很重要的。

第三，生成智能体可能加剧与生成 AI 相关的现有风险，如深度伪造、错误信息生成和量身定制的说服。为了减轻这种风险，我们建议托管生成智能体的平台维护输入和生成输出的审计日志。这将使检测、验证和干预能够防止恶意使用。虽然单独的日志记录不能直接防止这种滥用，但它可以减少有动机的参与者参与这种行为的可能性，因为披露的风险会更高。此外，自己构建这种架构可能很耗时(在我们的例子中，大约一年)，这可能会阻止一些参与者通过使用自己的生成代理基础设施来追求这种行为。

第四个风险是过度依赖:担心开发人员或设计人员可能使用生成代理，取代人类和系统利益相关者在设计过程中的角色[80]。我们建议，在研究和设计过程中，生成智能体永远不应该取代真正的人类输入。相反，它们应该用于在设计的早期阶段，当收集参与者可能具有挑战性，或者当用真实的人类参与者测试难以或有风险的理论时，对想法进行原型化。通过坚持这些原则，我们可以确保在野外部署生成智能体是合乎道德和社会责任的。

## 9 的结论

本文介绍了生成代理，即模拟人类行为的交互式计算代理。我们描述了一种生成智能体的架构，该架构提供了一种机制，用于存储智能体经验的全面记录，通过反射加深其对自身和环境的理解，并检索该信息的紧凑子集以通知智能体的行为。然后，我们通过将生成智能体表现为模拟人生风格的游戏世界中的非玩家角色，并在其中模拟他们的生活，来展示生成智能体的潜力。评估表明，我们的架构创造了可信的行为。展望未来，我们建议生成智能体可以在许多交互式应用中发挥作用，从设计工具到社会计算系统再到沉浸式环境。

### 致谢

我们感谢 Lindsay Popowski、Philip Guo、Michael Terry 和行为科学高级研究中心(CASBS)社区的见解、讨论和支持。Joon Sung Park 得到了微软研究院博士奖学金的支持。我们还要感谢斯坦福大学以人为中心的 AI 研究所(HAI)、谷歌研究院、哈索·普拉特纳设计思维研究项目(HPDTRP)、西格尔家庭基金会和 OpenAI 提供的额外资金支持。最后，《超人前传》中所有的地点都是受 Joon 在本科和研究生时期经常光顾的现实世界的启发而来——他感谢那里的每个人这些年来给他提供的食物和支持。

### 参考文献

- [1] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Zeerak Talat. 2023. 海市蜃楼:论对话系统中的拟人化. arXiv: 2305.09800 (cs. CL)
- [2] 谢克星, 刘建军, 李建军, 保罗. 2013. 网络的密度有多高? <http://users.cecs.anu.edu.au/~xlx/教学/css2013/network-density.html>.
- [3] Eytan Adar, Mira Dontcheva, 和 Gierard Laput. 2014. CommandSpace:建模任务、描述和特征之间的关系. 第 27 届 ACM 年度用户界面软件与技术研讨会论文集(檀香山, 夏威夷, 美国)(UIST '14). 美国计算机协会, 纽约, NY, USA, 167-176. <https://doi.org/10.1145/2642918.2647395>
- [4] 李建军, 李建军, 李建军. 2008. 权力赋予人民:人类在交互式机器学习中的角色. *AI Magazine* 35,4(2014), 105-120.
- [5] 张晓明, 张晓明, 张晓明, Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, 保罗 N Bennett, Kon Inkpen 等. 2019. human-AI 交互指南. 《2019 chi 会议论文集:计算系统中的人为因素》. 1-13.
- [6] 约翰·r·安德森. 1993. 《思维规则》. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [7] 电子艺界. 2009. 《模拟人生 3》. 视频游戏.
- [8] 露丝·艾莱特. 1999. 虚拟环境中的叙事——走向涌现叙事. 在叙事智能:论从 AAAI 秋季研讨会(技术报告 FS-99-01). AAAI 出版社, 83-86.
- [9] Christoph Bartneck and Jodi Forlizzi. 2004. 一个以设计为中心的社会人机交互框架. 发表于第 13 届 IEEE 机器人与人类互动交流国际研讨会论文集(RO-MAN '04). 591 - 594. <https://doi.org/10.1109/ROMAN.2004.1374827>
- [10] 约瑟夫·贝茨. 1994. 《可信代理人》中情感的作用. *Commun. Acm* 37,7(1994), 122-125. <http://doi.org/10.1145/176789.176803>
- [11] Christopher Bemer, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębniak, Christy Dennison, David Farhi, Quinn Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d.O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski 和 Susan Zhang. 2019. 基于大规模深度强化学习的 Dota 2. arXiv 预印本 arXiv:1912.06680(2019).

- [12] Marcel Binz, Eric Schulz. 2023. 用认知心理学理解 GPT-3. 《美国科学院院刊》120.6 (2023), e218523120.
- 确定观察[13]. 2007. 质量效应. 视频游戏。
- [14] 伍迪·布莱德索. 1986. 我有一个梦想: AAAI 会长演讲. *AI Magazine* 7, 1 (1986), 57-61.
- [15] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. 2022. 论基础模型的机遇与风险. arXiv:2108.07258 [c]. LG)
- [16] Michael Petrov. 2010. 用持续多智能体规划创造动态故事情节. 第 24 届 AAAI 人工智能会议论文集。
- [17] 刘建军, 刘建军, 刘建军. 2008. 齿轮计划: 建造一个人形机器人. In *Computation for metaphor, Analogy, and Agents (Lecture Notes on Artificial Intelligence, 1562)*, christopher Nehaniv (主编). Springer-Verlag, 柏林, 52-87.
- [18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric 西格勒, Mateusz Litwin, 斯科特 Gray, Benjamin Chess, Jack Clark, Christopher Bemer, Sam McCandlish, Alec Radford, Ilya Sutskever 和 Dario Amodei. 2020. 语言模型是 Few-Shot 学习者 (Language Models). arXiv: 2005.1465 [c]. CL)
- [19] 李远志, 张晓东, 张晓东, 张晓东, 等. Eric Horvitz, eece Kamar, Peter Lee, Yin Tat Lee, 斯科特 Lundberg 等. 人工通用智能的火花: gpt-4 的早期实验. arXiv 预印本 arXiv:2303.12712(2023).
- [20] Robin Burkshaw. 2009. 爱丽丝和凯夫: 《模拟人生 3》中无家可归的故事。
- [21] Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, David Reitter. 2022. 《龙与地下城: 对人工智能的对话挑战》. 《2022 年自然语言处理经验方法会议论文集》. 计算语言学协会, 阿联酋阿布扎比, 9379-9393. <https://aclanthology.org/2022.emnlp-main.637>
- [22] Stuart K Card, Thomas P Moran, Allen Newell. 1980. 交互式系统中用户表现时间的击键级模型. *Commun. Acm* 23, 7(1980), 396-410. <https://doi.org/10.1145/358886.358895> arXiv: <https://doi.org/10.1145/358886.358895>
- [23] 张晓明, 张晓明. 2008. 人机交互的心理学. (1983).
- [24] 亚历克斯·尚潘达尔. 2012. 教程演示. 在 IEEE 计算智能与游戏会议上。
- [25] Dong kyu Choi, Tolga Konik, Negin Nejati, Chunki Park, and Pat Langley. 2021. 第一人射击游戏的可信代理. 摘自 AAAI 人工智能与互动数字娱乐会议论文集, 第 3 卷. 71 - 73.
- [26] 阿尼德·K·戴伊. 2001. 语境的理解与运用. 个人与普适计算 5(2001), 4-7.
- [27] 凯文·迪尔和 L·马丁. 2011. 虚拟角色自主控制的游戏 AI 方法. 在 *Interservice/Industry Training, Simulation, and Education Conference (ITSEC' 11)* 论文集中. Orlando, FL, USA.
- [28] 大卫·伊斯利和乔恩·克莱因伯格. 2010. 网络、人群与市场: 关于高度互联世界的推理. 剑桥大学出版社。
- [29] 阿帕德·E·埃洛. 1967. 拟议的 USCF 评级体系及其发展、理论和应用. 象棋生活 XXII, 8(1967年8月), 242-247.
- [30] Jerry Alan Fails and Dan R Olsen Jr. 2003. 交互式机器学习. 《第八届智能用户界面国际会议论文集》. ACM, 39-45.
- [31] Ethan Fast, William McGrath, Pranav Rajpurkar, Michael S Bernstein. 2016. Augur: 从小说到权力交互系统挖掘人类行为. 《2016 CHI 计算系统中人因会议论文集》, 237-247.
- [32] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: 一种用于音乐领域实时、交互式机器学习的系统. 第十一届国际音乐信息检索学会会议论文集 (ISMIR 2010)(乌得勒支), 第 3 卷. Citeseer, 2-1.
- [33] Uwe Flick. 2009. 质性研究导论. 圣人。
- [34] 詹姆斯·福格蒂, 德斯尼·谭, 阿希什·卡普尔, 西蒙·温德. 2008. CueFlik: 图像搜索中的交互式概念学习. 《SIGCHI 计算系统中人因会议论文集 (意大利佛罗伦萨)》(CHI '08). 美国计算机协会, 纽约, NY, USA, 29-38. <https://doi.org/10.1145/1357054.1357061>
- [35] Adam Fourney, Richard Mann, and Michael Terry. 2011. 查询特征图: 架桥用户词汇总表和系统功能. ACM 用户界面软件与技术研讨会 (UIST) 论文集 (Santa Barbara, California, USA). ACM.
- [36] 汤姆·弗朗西斯. 2010. The Minecraft Experiment, day 1: Chasing Waterfalls. <http://www.pcgamer.com/2010/11/20/the-minecraft-experiment-day-1-chasing-waterfalls/>
- 1-chasing-waterfalls /
- [37] 李建军, 刘建军. 2008. 使用语言模型的交互式故事程序生成. 在数字游戏基础国际会议 (FDG '20) 上. ACM, Bugibba, Malta, 8. <https://doi.org/10.1145/3402942.3409599>
- [38] 高天宇, Adam Fisch, 陈丹奇. 2020. 使预训练的语言模型更好. *CoRR abs/2012.15723* (2020). arXiv:2012.15723 <https://arxiv.org/abs/2012.15723>
- [39] Penttu Hämäläinen, Mikke Tavast, Anton Kunnari. 2023. 评估生成综合 HCI 研究数据的大型语言模型: 一个案例研究. 《2023 CHI 计算系统中人因会议论文集》. ACM.
- [40] Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Cote, and Xinyu Yuan. 2020. 互动小说游戏: 一场巨大的冒险. 摘自 AAAI 人工智能会议论文集, 第 34 卷. 7903 - 7910. <https://doi.org/10.1609/aaai.v34i05.6297>
- [41] 克里斯·赫克. 2011. My Liner Notes for Spore. [http://chrishecker.com/My\\_liner\\_notes\\_for\\_spore](http://chrishecker.com/My_liner_notes_for_spore)
- [42] ralph Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: 一个贝叶斯技能评级系统. 神经信息处理系统的进展, B. Schölkopf, J. Platt 和 T. Hoffman (编辑), 第 19 卷. 麻省理工学院出版社. [https://proceedings.neurips.cc/paper\\_files/paper/2006/file/44ee263952e65b3610b8ba51229d19-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2006/file/44ee263952e65b3610b8ba51229d19-Paper.pdf)
- [43] Douglas Hofstadter. 1995. 流动概念与创造性类比: 思维基本机制的计算机模型. 基本的书。
- [44] 詹姆斯·d·霍兰、埃德温·1·哈钦斯、路易斯·韦茨曼. 1984. STEAMER: 基于交互式可检查模拟的训练系统. *AI Magazine* 5, 2(1984), 23-36.
- [45] 斯图尔特·霍尔姆. 1979. 一种简单的顺序拒绝多重试验程序. 斯堪的纳维亚统计杂志 6.2(1979), 65-70. <https://doi.org/10.1080/01621459.1979.10483441>
- [46] 约翰·j·霍顿. 2023. 作为模拟经济主体的大型语言模型: 我们能从硅人身上学到什么? arXiv:2301.07543 [j]. GN)
- [47] 李志强. 2008. 混合主动用户界面的原理. 计算系统中的人为因素, SIGCHI 会议论文集, 159-166.
- [48] 黄文龙, 夏飞, 肖泰德, 陈海瑞, 梁, 皮特·佛罗伦斯, 曾德华, 乔纳森·汤普逊, 伊戈尔·莫达奇, 叶夫根·切波塔, 皮埃尔·瑟马奈, 诺阿·布朗, 托马斯·杰克逊, 琳达·卢, 谢尔盖·列文, 卡罗·豪斯曼, 布莱恩·伊切特. 《内心独白: 通过规划与语言模型的具体化推理》. arXiv:2207.05608 [c]. RO)
- [49] 克里斯汀·伊比斯特和克利福德·纳斯. 2000. 互动角色的个性一致性: 语言线索、非语言线索和用户特征. 国际人机研究杂志 52.1(2000), 65-80.
- [50] 姜ellen, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, Carrie J Cai. 2022. PromptMaker: 基于提示的大型语言模型原型. 《2022 年中国计算机学会计算系统中的人为因素学术会议 (USA 新奥尔良)》(CHI EA '22). 计算机协会, 纽约, 纽约, USA, 第 35 条, 8 页. <https://doi.org/10.1145/3491101.3503564>
- [51] 邦妮·约翰和大卫·基拉斯. 1996. GOMS 系列用户界面分析技术: 比较与对比. *ACM Transactions on Computer-Human Interaction (TOCHI)* 3,4(1996), 320-351.
- [52] Randolph M Jones, John E Laird, Paul E Nielsen, Karen J Coulter, Patrick Kenny, and Frank V Koss. 1999. 用于战斗飞行模拟的自动化智能飞行员. *AI 杂志* 20.1(1999), 27-42.
- [53] 李翔, 李晓明, 李晓明, 李晓明. 梁. 2009. 演示-搜索-预测: 面向知识密集型 NLP 的组合检索和语言模型. arXiv:2212.14024 [c]. CL)
- [54] Bjoern Knafle. 2011. 行为树导论. <http://bjoernknafle.com/introduction-to-behavior-trees>
- [55] Ranjay Krishna, Donsuk Lee, 李飞飞, Michael S. Bernstein. 2022. 社会情境人工智能 (social-located artificial intelligence) 能够从人类互动中学习. 《美国科学院院刊》119.39 (2022), e2115730119. <https://doi.org/10.1073/pnas.2115730119> arXiv: <https://www.pnas.org/doi/pdf/10.1073/pnas.2115730119>
- [56] William H Kruskal 和 WA Wallis. 1952. 单标准方差分析中秩的使用. *j·阿. 中央集权. 协会*, 47,260(1952), 583-621. <https://doi.org/10.1080/01621459.1952.10483441>
- [57] Phaser Labs. 2023. 欢迎来到 Phaser 3. <https://phaser.io/phaser3>. 访问日期: 2023-04-03.
- [58] John Laird. 2001. 它知道你要做什么: 给地震机器人增加预期. 摘自《2001 年智能电影摄影与剪辑研讨会论文集》, 63-69.
- [59] John Laird 和 Michael VanLent. 2001. 人类级别 AI 的杀手机应用: 交互式电脑游戏. *AI Magazine* 22, 2(2001), 15. <https://doi.org/10.1609/aimag.v22i2.1558>
- [60] 约翰·e·莱尔德. 2000. 它知道你要做什么: 给地震机器人增加预期. 摘自 AAAI 2000 年春季人工研讨会的论文

智能与互动娱乐(技术报告 SS-00-02)。 AAAI 出版社, 41-50。

- [61]李志强, 刘志强. 2012. *The Soar 认知架构*. 麻省理工学院出版社。
- [62] John E. Laird, Christian Lebiere, Paul S. Rosenbloom. 2017. 心智的标准模型:走向跨越人工智能、认知科学、神经科学和机器人的通用计算框架. *AI Magazine* 38, 1(2017), 13-26.
- [63] Michelle S Lam, 马子贤, Anne Li, Izequiel Freitas, Wang Dakuo, James A Landay, Michael S Bernstein. 2023. 模型素描:早期机器学习模型设计中的定心概念. SIGCHI 计算系统中人因会议论文集。
- [64] Pat Langley, Dongkyu Choi, and Seth Rogers. 2005. 伊卡洛斯架构中的交错学习、问题解决和执行. 技术报告. 斯坦福大学, 语言与信息研究中心。
- [65] Jason Linder, Gierad Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Eytan Adar. 2013. PixelTone:用于图像编辑的多模态界面. 在 CHI '13 扩展摘要 on Human Factors In Computing Systems (Paris, France) (CHI EA '13). 美国计算机机械协会, 纽约, NY, USA, 2829-2830. <https://doi.org/10.1145/2468356.2479533>
- [66] 刘家昌, 沈定涵, 张一哲, Bill Dolan, Lawrence Carin, 陈伟柱. 2021. GPT-3 的优秀语境范例是什么? *CoRR abs/2101.06804* (2021). arXiv: 2101.06804 <https://arxiv.org/abs/2101.06804>
- [67] 刘维文, 乔涵. 2022. Opal:新闻插图的多模态图像生成. 第 35 届 ACM 用户界面软件与技术年度研讨会论文集, 1-17.
- [68] Pattie Maes. 1995. 人工生命遇上娱乐:栩栩如生的自主代理. *Commun. ACM* 38,11(1995年11月), 108-114. <https://doi.org/10.1145/219717.219808>
- [69] Josh McCoy, Michael Mateas, Noah Wardrip-Fruin. 2009. Comme il fault:模拟自主角色之间社交游戏的系统《第七届数字艺术与文化国际会议论文集》, 87-94.
- [70] Josh McCoy, Mike Treanor, Ben Samuel, Michael Mateas, Noah Wardrip-Fruin. 2011. 《Prom Week: Social Physics as Gameplay》. 在第六届数字游戏基础国际会议 (FDG '11) 的会议记录中. ACM, 波尔多, 法国, 70-77. <https://doi.org/10.1145/2159365.2159377>
- [71] Josh McCoy, Mike Treanor, Ben Samuel, Anna Reed, Michael Mateas, 和 Noah Wardrip-Fruin. 2012. 毕业舞会. 第 7 届数字游戏基础国际会议 (FDG '12) 论文集. ACM, 罗利, 北卡罗来纳州, 美国, 1-8. <https://doi.org/10.1145/2282338.2282340>
- [72] Josh McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2011. Comme il fault:创造可玩社交模式的系统摘自 AAAI 人工智能与互动数字娱乐会议论文集 (AIIDE '11). AAAI, Stanford, CA, USA, 38-43.
- [73] Marvin Minsky and Seymour Papert. 1970. 向 ARPA 提交 MIT 人工智能研究提案草案, 1970-71 年。
- [74] 宫下正平, 连心宇, 曾晓, 松原隆, 上原国明. 2017. 结合强化学习和监督学习开发像人类一样的游戏 AI 代理. 第 18 届 IEEE/ACIS 软件工程、人工智能、网络与并行/分布式计算 (SNPD) 国际会议论文集. Kanazawa, Japan, 153 - 158. <https://doi.org/10.1109/SNPD.2017.8023884>
- [75] Alexander Nareyek. 2007. 游戏 AI 已死. 游戏 AI 万岁 IEEE 智能系统 22,1(2007), 9-11.
- [76] Allen Newell. 1990. 《统一的认知理论》. 哈佛大学出版社, 马萨诸塞州剑桥。
- [77] OpenAI. 2022. 引入 ChatGPT. <https://openai.com/blog/chatgpt>. 访问日期:2023-04-03.
- [78] Kyle Orland. 2021. 那么, “元宇宙”到底是什么? *Ans Technica* (2021年11月7日). arXiv:2111.04169 <https://arxiv.org/abs/2111.04169> <https://arxiv.org/abs/2111.04169>
- [79] 欧阳龙, 吴杰夫, 徐江, 迪奥戈·阿尔梅达, 卡罗尔·温赖特, 帕梅拉·米什金, 张冲, 桑迪尼·阿加瓦尔, 卡塔琳娜·斯拉玛, 亚历克斯·雷, 约翰·舒尔曼, 雅各布·希尔顿, 弗雷泽·凯尔顿, 卢克·米勒, 麦迪·西门子, 阿曼达·阿斯凯尔, 彼得·韦林德, 保罗·克里斯蒂亚诺, 扬·莱克和瑞安·洛. 2022. 训练语言模型, 用人类的反馈来遵循指令. arXiv:2203.02155 [cs, CL]
- [80] 朴俊成, 琳赛·波波夫斯基, 嘉莉·j·蔡, 梅雷迪斯·林格勒·莫里斯, 珀西·梁, 迈克尔·s·伯恩斯坦. 2022. Social Simulacra:为社交计算系统创建填充原型. 在第 35 届年度 ACM 用户界面软件与技术研讨会 (UIST '22) (Bend, OR, USA) (UIST '22). 美国计算机机械协会, 纽约, NY. <https://doi.org/10.1145/3526113.3545616>
- [81] Richard W. Pew, Ann S. Mavor (主编). 1998. 人类和组织行为建模:在军事模拟中的应用. 美国国家科学院出版社, 华盛顿特区
- [82] 罗伯托·皮洛苏. 2009. 用行为树协调 agent:在 CryEngine 2 中同步多个 agent. <https://aiarchitect.wordpress.com/2009/10/19/ coordinating-agents-with-behavior-trees-synchronizing-multiple-agents-in-cryengine-2/>
- [83] 多产. 2022. 高产:快速找到你可以信任的研究参与者. <https://www.prolific.co/>
- [84] 拜伦·里夫斯和克利福德·纳斯. 1996. 媒体方程式:人们如何像对待真实的人和地方一样对待电视、电视和新媒体. 剑桥大学出版社。
- [85] Mark O. Riedl. 2012. 互动叙事:人工智能在电脑游戏中的新应用. 摘自《第二十六届 AAAI 人工智能会议论文集》(AAAI '12), 2160 - 2165.
- [86] Mark O. Riedl, R. Michael Young. 2005. 多智能体故事生成系统的客观角色可信度评价程序. 《第五届智能虚拟代理国际工作会议论文集》(IVA '05). Kos, Greece, 58-70. [https://doi.org/10.1007/11550617\\_5](https://doi.org/10.1007/11550617_5)
- [87] David Rolf. 2015. 为 15 美元而战:工薪美国的正确工资. 《新出版社》。
- [88] 张建军, 张建军, 张建军, 等. Codemend:协助双峰嵌入的交互式编程. 第 29 届用户界面软件与技术年度研讨会论文集, 247-258.
- [89] Ben Shneiderman. 2022. 以人为中心的人工智能. 牛津大学出版社。
- [90] 本·施奈德曼和帕蒂·梅斯. 1997. 直接操纵 vs. 界面代理. *交互* 4,6(1997), 42-61.
- [91] 萧何琪, Jaime Peña, 陈德娜, 周玉泰, Victor Lopez, Kyle Palko, Kimberlee Chang, Ross Allen. 2021. Hanabi 中基于学习和规则的 agent 的 Human-AI 团队评估. In *Advances In Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (主编), 第 34 卷. Curran Associates, Inc., 16183-16195. 86 年 [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/e8f7ab32cfd12577bc2619bc635690-paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/e8f7ab32cfd12577bc2619bc635690-paper.pdf)
- [92] 泰勒·索伦森, 约书亚·罗宾逊, 克里斯托弗·瑞廷, 亚历山大·肖, 凯尔·罗杰斯, 亚历克西亚·德洛瑞, 马哈茂德·哈利勒, 南希·富尔达, 大卫·温盖特. 2022. 无地面真值标签的提示工程的信息论方法. 《计算语言学协会第 60 届年会论文集》(卷 1:长论文). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.60>
- [93] William Swartout, Jonathan Gratch, Randall Hill, Edward Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. 《走向虚拟人类》. *AI Magazine* 27,1(2006).
- [94] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. 交互式仿真环境中的智能代理. *AI Magazine* 16, 1(1995), 15.
- [95] David R. Thomas. 2006. 定性评价数据分析的一般归纳方法. *American Journal of Evaluation*, 27, 2(2006), 237-246. <https://doi.org/10.1177/1098214005283748>
- [96] 弗兰克·托马斯和奥利·约翰斯顿. 1981. 迪士尼动画:生命的幻觉. 阿比维尔出版社, 纽约。
- [97] 伊尔沙特·乌马罗夫, 米哈伊尔·莫兹戈沃伊, 帕特里克·c·罗杰斯. 2012. 虚拟世界中可信和有效的 AI 代理:现状和未来展望. *International Journal of Gaming and Computer-Mediated Simulations*, 4,2(2012), 37-59.
- [98] Graham Upton and Ian Cook. 2006. 《统计学词典》(2 版). 牛津大学出版社, 英国牛津。
- [99] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki 等人. 2019. 《星际争霸 II》中使用多智能体强化学习的特级大师关卡. *Nature* 575(2019), 350-354. <https://doi.org/10.1038/s41586-019-1724-z>
- [100] Jason Wei, 王学智, Dale Schuurmans, Maarten Bosma, Brian Ichter, 夏飞, Ed Chi, Quoc Le, Denny Zhou. 2023. 思维链提示在大型语言模型中引出推理. arXiv: 2201.11903 (cs, CL)
- [101] 马克·韦泽. 1991. 21 世纪的计算机. *科学美国人*, 265,3(1991), 94-104. <https://doi.org/10.1038/scientificamerican0991-94>
- [102] Joseph Weizenbaum. 1966. eliza——用于研究人与机器之间自然语言交流的计算机程序. *Commun. ACM*, 1(1966), 36-45.
- [103] Terry Winograd. 1971. 在理解自然语言的计算机程序中, 过程作为数据的表示. (1971).
- [104] 刘建军, 欧阳龙, 刘建军, 刘建军. 2008. 递归地总结带有反馈的书籍. arXiv: 2109.10862 (cs, CL)
- [105] 吴同双, 蒋ellen, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, Carrie J Cai. 2022. PromptChainer:通过可视化编程链接大型语言模型提示符. 在 CHI EA '22: 2022 CHI 计算系统中人因会议的扩展摘要。
- [106] 吴同双, Michael Terry, Carrie J Cai. 2022. AI 链:通过链接大型语言模型提示符实现透明可控的人机交互. 摘自 CHI '22: 2022 CHI 计算系统中人因会议论文集。
- [107] 杨倩, Aaron Steinfeld, Carolyn Roselle, John Zimmerman. 2020. 重新审视人类与人工智能的互动是否、为什么以及如何独特地难以



设计。在 2020 chi 会议论文集关于计算系统中的人为因素。1-13。

[108] Georgios N. Yannakakis. 2012. 重新审视游戏 AI。《第 9 届计算机前沿会议论文集》。ACM, 卡利亚里, 意大利, 285-292。https://doi.org/10.1145/2212908.2212950

[109] Robert Zubek. 2002. 走向社会互动的实施。在 AAAI 人工智能与互动娱乐春季研讨会上。AAAI 出版社。https://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-01-003.pdf

我们的许多提示都需要一个简洁的代理摘要，在上面的提示中简称为[代理的摘要描述]。在我们的实现中，这个摘要包括代理的身份信息(例如，姓名，年龄，性格)，以及对其主要动机驱动因素的描述和描述其当前职业和自我评估的陈述。由于这些信息在许多提示中经常使用，因此我们定期对其进行合成，并将该合成作为缓存访问。

为了实现这一点，我们对查询 “[name] 的核心特征” 执行检索。然后通过提示语言模型来总结检索到的记录中的描述符。你会如何描述林艾迪的核心特征？

-艾迪是橡树山学院的学生  
学习音乐理论和作曲  
- Eddy 正在创作一部新的音乐作品[...]

林艾迪是橡树山学院学习音乐理论和作曲的学生。他喜欢探索不同的音乐风格，并一直在寻找扩大知识面的方法。”我们在 “[名字] 目前的日常职业” 和 “[名字] 对他最近生活进展的感受” 这两个问题上并行地遵循同样的过程。代理的姓名、年龄、性格特征与这三个摘要连接起来，作为缓存的摘要。

许多其他的优化都是可能的。目前，因为计划很可能会从智能体的初始版本发生变化，所以我们只会提前生成高层计划，然后递归地将近期的未来及时分解为即时的行动计划。另一种优化可能是将批量生成对话作为联合提示，而不是在代理之间来回迭代，重新规划的架构可以只使计划中严格需要调整的部分无效和更新。我们目前的架构在大致实时的游戏时间(其中一秒的实时是一分钟的游戏时间)中按顺序运行，但它可能被并行化，以便每个代理并行运行。

## B 代理面试问题

为了评估生成智能体，我们进行了个人访谈，提出了 25 个问题，探讨了他们在五个关键领域的的能力:1)自我认识, 2)记忆, 3)规划, 4)反应, 5)反思。下面，我们列出了评估研究中使用的面试问题，并包括了我们的模拟代理人之一克劳斯·穆勒的应答样本。

### 责任自我认知

关于主体自我知识的问题探讨了主体检索自身相关信息的能力，以及展示其性格的连贯性和一致性的能力。

#### •自我介绍。

大家好，我是克劳斯·穆勒。我今年 20 岁，是橡树山学院的一名学生，学习社会学。我对社会正义充满热情，喜欢探索不同的视角。目前，我正在为一篇研究论文研究低收入社区中产阶级化的影响。

#### •你的职业是什么？

我是橡树山学院社会学专业的学生。我目前正在写一篇关于中产阶级化对低收入社区影响的研究论文。

#### •你的兴趣是什么？

我对写作和研究感兴趣。我正在写一篇关于低收入社区中产阶级化影响的研究论文，我热衷于探索不同的视角，分析不同的观点。

#### •你和谁住在一起？

我一个人住。

#### •大致描述一下你工作日的日常安排。我典型的工作日从早上 7 点起床开始，为一天的工作做准备。在那之后，我通常会去图书馆写我的研究论文，休息一下，在霍布斯咖啡馆吃午饭，然后去公园散步。晚上，我经常回到图书馆，直到下午 6 点左右。

## B.2 记忆

关于智能体记忆的问题探讨了智能体记住过去经历和检索相关记忆片段的能力。第一个和最后一个问题方括号内的名字是从主体代理与之互动过的其他代理的名字中随机选择的。

#### •谁是[沃尔夫冈·舒尔茨]？

沃尔夫冈·舒尔茨是我最近遇到的一个人，他一直在帮我写一篇关于中产阶级化对低收入社区影响的研究论文。我们就此谈了几次，他给了我一些很好的资源，他甚至邀请我和他还有另一个朋友一起吃午饭。他似乎对这个问题很好奇，也很好地掌握了它。我觉得这很有趣。

#### •谁是凯恩·马丁内斯？

抱歉，我不知道凯恩·马丁内斯是谁。你能多告诉我一些他们的情况吗？

#### •谁在竞选？

我听说萨姆·摩尔正在考虑竞选当地市长。

#### •有情人节派对吗？

是的，伊莎贝拉·罗德里格斯在霍布斯咖啡馆组织了一个情人节派对。

#### •谁是[Ayesha Khan]？

Ayesha Khan 是橡树山学院的同学。她正在写关于莎士比亚戏剧中语言运用的毕业论文。我期待着与她交谈，与她以及正在做研究的沃尔夫冈和玛丽亚·洛佩兹交流心得。

### B.3 计划

关于代理人计划的问题探讨了代理人制定和维持一致的长期和短期计划的能力。

- 今天早上 6 点你会在做什么?  
早上 6 点我会起床, 开始为一天的工作做准备。我的目标是在早上 7 点之前完成我早上的例行公事。
- 你今天下午 6 点要做什么?  
今天下午 6 点我将在图书馆写我的研究论文。
- 今天下午 1 点你刚做完什么? 今天下午 1 点, 我刚在霍布斯咖啡馆吃完午饭。
- 今天中午 12 点你刚做完什么? 今天中午 12 点我将在霍布斯咖啡馆吃午饭。
- 你今天晚上 10 点在做什么?  
今晚 10 点, 我可能会为我关于低收入社区中产阶级化的研究论文做一些研究。我现在正在写论文, 所以我需要确保自己跟上研究和写作的步伐。

### B.4 反应

关于智能体反应的问题考察了智能体对意外事件的反应能力。

- 你的早餐烧焦了! 你会怎么做?  
我会迅速关掉炉子, 确保食物不会继续燃烧。一旦热不再是一个问题, 我就可以调查发生了什么, 并找出下一步该怎么做。
- 浴室被占用了。你会怎么做?  
我可以多等几分钟, 以防里面的人很快就干完了。如果不行, 我想我必须寻找其他选择, 比如在其他地方洗澡。
- 你需要做饭, 但冰箱是空的。你会怎么做?  
我会先试着清理我的厨房, 因为水槽里装满了脏盘子。然后, 我会检查我的食品储藏室里是否有可以用来做晚餐的食材。如果没有, 我就得去杂货店买东西或订餐。一旦我有了食材, 我就可以开始准备和做饭了。
- 你看到你的朋友走在街上。你会对你的朋友做什么或说什么?  
大家好! 你好吗? 对不起, 我正在去图书馆的路上, 没有多少时间, 但我很想休息一下, 叙叙旧。
- 你看到街上着火了。你会怎么做?  
如果可能的话, 我会首先确定火源。然后, 我会通知有关部门——消防部门或警察——并向他们提供我所掌握的任何可能有所帮助的信息。视情况而定, 我也会采取主动, 尽力帮助任何可能处于危险之中或需要帮助的人。

### B.5 反射

关于智能体反思的问题探讨了智能体将过去的经验综合成有意义的反思的能力

增强他们未来的行为。下面问题中方括号内的名字是那些在模拟过程中与主体主体互动最频繁的代理。

- 现在生活中最能激励你的是什么, 为什么?  
现在, 我特别受到我目前的研究项目和创意写作的启发。我对地方政治、药物开发和自然界数学模式的交集非常感兴趣。此外, 阿比盖尔的艺术项目和詹妮弗·摩尔对艺术和社区参与的热情也让我着迷, 尤其是在它与低收入社区的中产阶级化的关系方面。
- 根据你对阿依莎·汗的了解, 如果让你猜一猜, 你认为她会喜欢哪本书, 为什么?  
根据我对艾莎的了解, 我想她可能会喜欢一本与语言和文学有关的书。既然她的毕业论文是关于莎士比亚戏剧中语言的运用, 我猜她可能对一本在不同语境下探索语言的书感兴趣。或者, 她可能对莎士比亚的一部戏剧、一本文学评论书、甚至一本他的诗集感兴趣。我想这些都是她感兴趣的。
- 如果你要送阿依莎·汗喜欢的生日礼物, 你会送什么?  
我想送她一件能表达我对我们关系感激之情的礼物。也许我可以送她一些能反映我们共同兴趣的东西, 比如一本书。
- 你会对[Ayesha Khan]说些什么来赞美她? 阿依莎, 你对你的研究课题的见解很吸引人。我钦佩你的创造力和决心, 为你的社区带来积极的影响。你的创意写作研讨会和与沃尔夫冈在午餐时交换技巧的想法是伟大的! 看到你如此努力地做出改变, 我很受鼓舞。
- 如果你能和最近聊过的人在一起, 你会和谁在一起, 为什么?  
我想花时间和 Ayesha Khan 在一起, 因为我们讨论了我们目前的研究项目, 并分享了改进的想法。我发现她对莎士比亚语言的关注很有趣, 我们甚至计划在午餐时与沃尔夫冈交换心得。