

生成性代理：人类行为的交互式拟像 (Generative Agents: Interactive Simulacra of Human Behavior)

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google Research
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu



图1：生成性代理为交互式应用创建可信的人类行为拟像。在这项工作中，我们通过在一个类似于《模拟人生》的沙盒环境中放置25个代理来演示生成性代理。用户可以观察并干预代理计划他们的日常生活、分享新闻、建立关系和协调团体活动。

摘要

可信的人类行为的替代品可以为交互式应用提供支持，从沉浸式环境到人际交流的排练空间，再到原型工具。在本文中，我们介绍了生成代理——计算软件代理，模拟可信的人类行为。生成代理醒来，做早餐，然后去上班；艺术家画画，而作家写作；他们形成观点，注意彼此，并发起对话；他们记得并反思过去的日子，当他们计划下一天。为了实现生成代理，我们描述了一种架构，它扩展了大型语言模型，使用自然语言存储代理经历的完整记录，将这些记忆随着时间合成更高层次的反思，并动态地检索它们来规划行为。我们实例化生成代理来填充一个受《模拟人生》启发的交互式沙盒环境，在那里最终用户可以使用自然语言与一个小镇的二十五个代理进行交互。在评估中，这些生成代理产生了可信的个体和涌现的社会行为：例如，仅以单个用户指定的概念开始，即一个代理想要举办情人节派对，代理在接下来的两天内自主地传播派对邀请函，结识新朋友，相互邀请参加派对，并协调在正确的时间一起出现在派对上。我们通过消融证明了我们代理架构的组成部分——观

察、规划和反思——每个都对代理行为的可信度产生了关键贡献。通过将大型语言模型与计算、交互式代理融合，这项工作引入了用于实现可信人类行为模拟的架构和交互模式。

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. arXiv, April, 2023, © 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN xx-x-xxxx-xxxx-x/xx/xx. . . \$15.00 <https://doi.org/xx.xx/xx.xx>

CCS 概念

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

关键词

人工智能交互，代理，生成性人工智能，大型语言模型

ACM Reference Format:

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In . ACM, New York, NY, USA, 22 pages. <https://doi.org/xx.xx/xx.xx>

1. 引言

我们如何创建一个反映可信人类行为的交互式人工社会？从沙盒游戏（如《模拟人生》）到应用程序（如认知模型[21]和虚拟环境[9,58]），超过四十年来，研究人员和从业人员一直设想计算代理，可以作为可信人类行为的替代品。在这些愿景中，计算能力强大的代理根据过去的经验一致地行动，并可信地对其环境做出反应。这种人类行为的模拟可以用真实的社会现象填充虚拟空间和社区[26,79]，训练人们如何处理罕见但困难的人际情况[43,51,93]，测试社会科学理论[11,45]，制作模型人类处理器进行理论和可用性测试[21,38,50]，支持普遍计算应用程序[30]和社交机器人[9,13]，并支撑非可玩游戏角色[58,84]，它们可以在开放世界中导航复杂的人际关系。

然而，人类行为的空间是广阔而复杂的[84,108]。尽管大型语言模型取得了惊人的进展[17]，它们可以在单个时间点模拟可信的人类行为[38,79]，但完全通用的代理需要确保长期一致性的架构来管理不断增长的记忆，因为新的交互、冲突和事件随着时间而出现和消逝，同时处理多个代理之间展开的级联社会动态。成功需要一种方法能够在很长一段时期内检索相关事件和交互，反思这些记忆以推断出更高层次的推论，并将这种推理应用于创建既符合当时情况又符合代理行为长期弧线的计划和反应。

在本文中，我们介绍了生成代理——利用生成模型模拟可信人类行为的代理——并证明它们产生了可信的个体和群体行为仿真。生成代理对自己、其他代理和环境进行了广泛推断；他们根据自己的特征和经验创建每日计划，执行这些计划，做出反应，并在适当时重新规划；当最终用户改变他们的环境或用自然语言命令他们时做出反应。例如，当生成代理看到早餐烧焦时会关掉炉子，如果浴室被占用就在外面等待，如果遇到想要交谈的另一个代理就停下来聊天。一个充满生成代理的社会以涌现社会动态为特征，在那里形成新关系、信息扩散、跨代理协调。

为了实现生成代理，我们描述了一种代理架构，使用大型语言模型存储、综合和应用相关记忆来产生可信行为。我们的架构包括三个主要组成部分。第一个是记忆流，一个长期记忆模块，以自然语言记录代理经历的全面清单。检索模型结合相关性、最近性和重要性来呈现所需记录以指导代理瞬间行为。第二是反思，它随着时间将记忆合成更高层次的推论，使代理能够对自己和他人得出结论，以更好地指导其行为。第三个是规划，它将这些结论和当前环境转化为高层次的行动计划，然后递归地转化为详细的行为以进行行动和反应。这些反思和计划被反馈到记忆流中，以影响代理未来的行为。

这种架构在多个领域都有应用，从角色扮演和社会原型设计到虚拟世界和游戏。在社会角色扮演场景（例如面试准备）中，用户可以安全地排练困难、冲突激烈的对话。当原型设计社交平台时，设计师可以超越临时人物原型，原型动态、复杂的随时间展开的交互。对于本文的目的，我们关注创建一个受《模拟人生》启发的游戏中的小型交互式代理社会的能力。通过将我们的架构连接到 ChatGPT 大型语言模型[76]，我们在游戏环境中实现了一个由二十五个代理组成的小社会。最终用户可以观察并与这些代理进行交互。如果最终用户或开发人员希望镇上举办一场游戏内情人节派对，例如，传统游戏环境需要手动编写数十个角色的行为。我们证明，使用生成代理，只需简单地告诉一个代理她想要举办派对就足够了。尽管有许多潜在的失败点——派对策划者必须记得告诉其他代理关于派对的事情，与会者必须记得邀请函，那些记得邀请函的人必须决定实际出席，以及其他可能的失败点——我们环境中的代理成功了。他们传播关于派对的消息，然后出现在那里，其中一个代理甚至邀请另一个代理与他一起参加派对，所有这些均来自这个单一用户生成的种子建议。

我们对生成代理进行了两次评估：一次受控评估测试代理是否能够在孤立情况下产生可信的个体行为，并进行端到端评估，在其中生成代理以开放式方式相互交互两天游戏时间以了解它们的稳定性和涌现社会行为。在技术评估中，我们利用了一种方法学机会通过“面试”它来评估代理的知识和行为，以探究代理保持角色、记忆、规划、反应和准确反思的能力。我们比较了几种限制代理访问记忆、反思和规划能力的消融。我们观察到这些组件在这些面试任务中都对强大性能至关重要。在技术和端到端评估中，最常见的错误是当代理未能检索相关记忆时产生错误、编造对代理记忆的修饰，或从语言模型中继承过于正式的言语或行为。

总之，本文提供了以下贡献：

- 生成代理，可信地模拟人类行为，动态地根据代理不断变化的经验和环境进行调整。
- 一种新颖的架构，使生成代理能够记忆、检索、反思、与其他代理互动并在动态变化的情况下进行规划。该架构利用大型语言模型强大的提示能力，并补充这些能力以支持长期代理一致性、管理动态演变的记忆和递归产生更多生成。
- 两次评估（受控评估和端到端评估），建立了架构组件重要性的因果效应，并确定了由于检索不当、记忆虚构或从语言模型继承过于正式的言语或行为等原因引起的故障。
- 讨论生成代理在交互式系统中的机会以及道德和社会风险。我们认为，这些代理应该被调整以减少用户形成寄生社交关系的风险，记录以减少深度伪造和定制说服带来的风险，并以补充而非取代人类利益相关者在设计过程中应用。

2. 相关工作

在本节中，我们回顾了人工智能与人类交互领域的先前文献，并将建立可信人类行为替代品的议程置于其经典著作之中。这一议程曾被誉为交互、游戏和人工智能社区的北极星[9,58,84,85]，但由于人类行为的复杂性[16,108]，它一直是一个挑战。我们综合这些研究，认为大型语言模型虽然不足以自身支持，但当使用适当的架构进行利用时，它们为创建可信代理提供了一个新的角度。

2.1 人工智能与人类交互

交互式人工智能系统旨在将人类洞察力和能力结合在计算工件中，以增强其用户[3,29]。长期以来，一系列工作探讨了如何让用户交互式地指定模型行为。例如，Crayons 展示了早期交互式机器学习的愿景，允许非专家用

户训练分类器[29]。进一步的工作帮助阐明了用户如何通过示例[33]和/或演示[31]向系统描述他们的分类目标。最近的工作已将这些探索扩展到深度学习[62]和基于提示的创作[49,66,106]。

与此同时，一条持久的研究线索推进了基于语言和代理的交互在人机交互中的案例。像 SHRDLU [103] 和 ELIZA [102] 这样的开创性工作展示了自然语言与计算系统交互的机会和风险。随着研究的进展，人们开始明确自主代理可以提供新的委托和交互隐喻[67]，但人与代理之间的委托线仍在辩论和完善中[46,88,89]。最近，这项技术已经足够稳定，可以让代理通过自然语言在大型复杂在线社交环境中进行交互（例如[54]）。

自然语言交互提供了一种新颖的模式，可以扩展用户在诸如照片编辑[2,34,64]和代码编辑[87]等领域中的能力。

我们汇集这些工作线索，表明我们现在可以创建代理来替代人类行为进行交互式系统，并通过自然语言与它们进行交互。这样做，本文重新打开了检验基础 HCI 问题（如 GOMS 和 KLM 等认知模型[21,22]、原型设计工具[79]和普遍计算应用程序[25,30,100]）的大门。

2.2 可信人类行为替代品

先前文献将可信性或可信代理描述为一个核心设计和工程目标。可信代理旨在提供生命幻觉并呈现现实主义外观，在决策和自主行动方面类似于迪士尼电影中的角色[9,95]。这些代理可以感知并进入我们所居住的开放世界环境[9,58]，并努力以涌现行为为基础，在与用户或其他代理进行社会互动时表现出可信的人类行为[19,35,70]。历史上，这些代理是在智能游戏 NPC 的背景下开发的[58,84]。具有可信行为的 NPC，如果可能的话，可以通过启用涌现叙事[7,15,48,92]和与代理的社交互动[110]来增强游戏和互动小说中的玩家体验。然而，更重要的是，游戏世界提供了越来越真实的现实世界可供性的表示，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开发人员提供了易于访问的测试平台，以精细化代理的认知能力，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[58,84]。

在过去四十年中，出现了多种创建可信代理的方法。然而，在实现过程中，这些方法通常简化了环境或代理行为维度以使工作更易于管理[16,72]。基于规则的方法，如有限状态机[90,96]和行为树[40,53,81]，占据了人类编写代理行为的蛮力方法[70]。它们提供了一种直接创建简单代理的方法，这仍然是当今最主流的方法[68,73,109]，甚至可以处理基础社交互动，如 Mass Effect [12] 和《模拟人生》[6] 系列中所示。然而，手动编写能够全面解决开放世界中可能交互的行为是不可能的。这意味着生成的代理行为可能无法完全代表其交互产生的后果[69-71]，也无法执行未在其脚本中硬编码的新程序[90,96]。另一方面，流行的基于学习的方法创建可信代理（如强化学习）通过让代理学习其行为克服了手动编写挑战，并在近几年在诸如 Starcraft 的 AlphaStar [98] 和 Dota 2 的 OpenAI Five [10] 等游戏中取得了超人表现。然而，它们取得成功主要是在对抗性游戏中，并且学习算法可以优化明确定义的奖励。它们尚未解决在开放世界中创建可信代理的挑战[39,73,90]。

计算中认知架构由 Newell 领导开发，旨在建立支持全面认知功能集合的基础设施[75]，适合可信代理在其原始愿景中所持有的全面性。它们推动了一些最早的可信代理示例。例如，Quakebot-SOAR [59] 和 ICARUS [24,63] 在第一人称射击游戏中生成了 NPC，而 TacAir-SOAR [80] 则在空战训练模拟中生成了飞行员。这些代理使用的架构不同（Quakebot 和 TacAir-SOAR 依赖 SOAR [60]，而 ICARUS 依赖自己的变体，受 SOAR 和 ACT-R [5] 启发），但它们共享相同的基本原则[61]。它们维护短期和长期记忆，用符号结构填充这些记忆，并以感知-计划-行动周期运行，动态感知环境并与其中一个手动编写的动作程序匹配[57,96]。使用认知架构创建的代理旨在推广到大多数（如果不是全部）开放世界背景，并在当时表现出强大的行为。然而，它们的行动空间仅限于手动编写的程序知识，并且它们没有提供一种机制使代理能够寻求新行为。因此，这些代理主要部署在非开放世界背景中，如第一人称射击游戏[24,59]或块世界[63]。

今天，按照其原始定义创建可信代理仍然是一个开放问题[84,108]。许多人已经放弃了这个问题，认为尽管现有的创建可信代理方法可能繁琐且有限，但它们足以支持现有的游戏和交互[23,74,108]。我们认为，大型语言模型提供了一个重新审视这些问题的机会，只要我们能够设计一个有效的架构来将记忆合成为可信行为。本文提供了朝着这样一个架构的一步。

2.3 大型语言模型与人类行为

生成代理利用大型语言模型来支持其行为。关键观察是大型语言模型在其训练数据中编码了广泛的人类行为[14,17]。如果使用狭窄定义的上下文提示，则模型可以用于生成可信行为。最近的工作已经证明了这种方法的有效性。例如，*Social Simulacra* 使用大型语言模型生成用户以填充新社交计算系统以原型化其涌现社交动态[79]。这种方法使用提示链[105,106]来生成简短自然语言描述人物及其在被原型化系统中出现时的行为。

其他实证研究复制了现有的社会科学研究[45]、政治调查[91]并生成了合成数据[38]。大型语言模型也被用来生成用户可以参与的交互式人类行为。在游戏中，例如，这些模型被用来创建交互式小说[36]和文字冒险游戏[20]。凭借它们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[47]。例如，当呈现一个任务，如拾起一个瓶子时，模型被提示将任务分解为更小的动作序列，如前往瓶子所在的桌子并拾起它。

我们认为，基于上述工作，大型语言模型可以成为创建可信代理的关键要素。现有文献主要依赖于可以被认为是—阶模板的方法，使用少量提示[37,65]或思维链提示[99]。这些模板在生成仅受代理当前环境条件的行为方面是有效的（例如，巨魔如何回应给定帖子，机器人需要采取什么行动才能进入房间，假设有一扇门）。然而，可信代理需要不仅基于其当前环境而且基于远远超过提示所能描述的经验集合进行条件化，因为完整的记忆流可能会分散模型注意力，并且由于底层模型有限的上下文窗口甚至无法适应当前。最近的研究试图通过使用静态知识库和信息检索方案[52]或简单总结方案[104]来超越—阶提示。本文扩展了这些想法，构建了一种代理架构，用于处理检索过程中动态更新的过去经验并与代理当前上下文和计划混合，这些上下文和计划可能相互支持或相互矛盾。

3. 生成代理行为和交互

我们通过在一个简单的沙盒世界中实例化生成代理来具体说明生成代理的功能，这个沙盒世界类似于《模拟人生》。在本节中，我们将介绍生成代理在 *Smallville* 中的功能和交互方式，并描述代理如何在其中行为。然后，在第 4 节中，我们将介绍我们的生成代理架构，它为这些功能和交互提供了支持。在第 5 节中，我们将描述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。

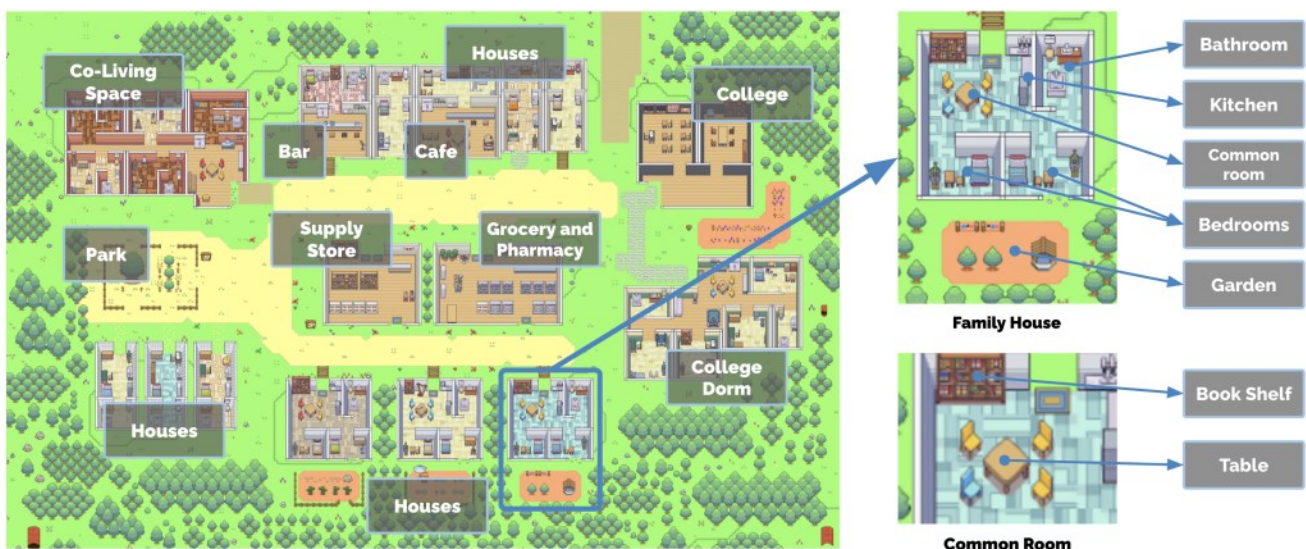


图2：带有标记区域的 *Smallville* 沙盒世界。根节点描述整个世界，子节点描述区域（例如，房屋、咖啡馆、商店），叶节点描述对象（例如，桌子、书架）。代理记住了它们所见过的世界部分的子图，以及它们所见到的状态。

3.1 代理头像和通信

25 个独特的代理居住在 Smallville。每个代理都由一个简单的精灵头像表示。我们撰写了一段自然语言描述来描绘每个代理的身份，包括他们的职业和与其他代理的关系，作为种子记忆。例如，John Lin 的描述如下：

John Lin 是 Willow Market and Pharmacy 的药房店主，他喜欢帮助人们。他总是想方设法让顾客更容易获得药物；John Lin 和他的妻子 Mei Lin（一名大学教授）以及儿子 Eddy Lin（一名学习音乐理论的学生）住在一起；John Lin 非常爱他的家人；John Lin 认识隔壁的老夫妇 Sam Moore 和 Jennifer Moore 已经有几年了；John Lin 认为 Sam Moore 是一个善良、友好的人；John Lin 和邻居 Yuriko Yamamoto 关系很好；John Lin 知道他的邻居 Tamara Taylor 和 Carmen Ortiz，但从未见过他们；John Lin 和 Tom Moreno 是 The Willows Market and Pharmacy 的同事；John Lin 和 Tom Moreno 是朋友，喜欢一起讨论当地政治；John Lin 稍微了解 Moreno 家族——丈夫 Tom Moreno 和妻子 Jane Moreno。

每个分号分隔的短语都作为初始记忆输入到代理的记忆中。

3.1.1 代理间通信

代理人之间的通信。代理人通过自己的行为与世界互动，并通过自然语言与彼此互动。在沙盒引擎的每个时间步骤中，代理人输出一个自然语言语句，描述他们当前的行为，例如“Isabella Rodriguez正在写她的日记”，“Isabella Rodriguez正在检查她的电子邮件”，“Isabella Rodriguez正在与家人通电话”，或“Isabella Rodriguez正在准备睡觉”。然后将此语句转换为影响沙盒世界的具体动作。该动作以一组表情符号显示在沙盒界面上，提供了动作在俯视图中抽象表示。为此，系统采用了一种语言模型，将动作转换为一组表情符号，这些表情符号出现在每个头像上方的对话框中。例如，“Isabella Rodriguez正在写她的日记”显示为📖📝，而“Isabella Rodriguez正在检查她的电子邮件”显示为✉️📧。可以通过点击代理人头像来访问动作的完整自然语言描述。

代理人之间以完整的自然语言进行交流。代理人知道其他处于其本区域内的代理人，并由生成性代理人架构确定他们是否路过或参与对话。这里是一个对话中间部分示例，在其中 Isabella Rodriguez 和 Tom Moreno 讨论即将到来的选举：

Isabella：我仍然在权衡我的选择，但我一直在与 Sam Moore 讨论选举。你对他有什么看法？

Tom：老实说，我不喜欢 Sam Moore。我认为他与社区脱节，不关心我们最大的利益。

3.1.2 用户控制

运行此模拟程序的用户可以通过与代理人通过对话进行交流或以“内心声音”的形式向代理人发出指令来控制模拟程序和干预。用户可以通过自然语言与代理人进行交流，通过指定代理人应该将他们视为的角色。例如，如果用户指定他们是一名“记者”，并询问即将到来的选举，“谁在竞选公职？”，那么 John 代理人会回答：

John：我的朋友 Yuriko、Tom 和我一直在谈论即将到来的选举，讨论候选人 Sam Moore。我们都同意投票给他，因为我们喜欢他的政纲。

要直接命令其中一个代理人，用户需要扮演代理人的“内心声音”的角色——这使得代理人更可能将陈述视为指令。例如，当被用户作为 John 的内心声音告诉“你将在即将到来的选举中与 Sam 竞争”时，John 决定参加选举并与妻子和儿子分享他的候选资格。

3.2 环境互动

Smallville 具有小村庄的常见便利设施，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性的子区域和对象，例如房屋中的厨房和厨房中的炉灶（图2）。所有作为代理人主要居住区域的空间都设有床、书桌、衣柜、架子以及浴室和厨房。

代理人在 **Smallville** 中移动，就像在简单的视频游戏中一样，进出建筑物，导航地图，并接近其他代理人。代理人的移动由生成性代理人架构和沙盒游戏引擎指导：当模型指示代理人将移动到某个位置时，我们计算到 **Smallville** 环境中目的地的步行路径，并且代理人开始移动。此外，用户也可以作为一个代理人进入 **Smallville** 的沙盒世界。用户控制的代理人可以是已经存在于世界中的代理人，如 **Isabella** 和 **John**，也可以是没有在 **Smallville** 有过历史记录的外部访客。**Smallville** 的居民会像对待彼此一样对待用户控制的代理人。他们认识到它的存在，发起互动，并记住它的行为，在形成对它的看法之前。

用户和代理人可以影响这个世界中物体的状态，就像在《模拟人生》这样的沙盒游戏中一样。例如，当一个代理人睡觉时，床就被占用了；当一个代理人用完食材做早餐时，冰箱就变空了。最终用户也可以通过用自然语言重写环绕着代理人的物体状态来改变一个代理人在 **Smallville** 中的环境。例如，当 **Isabella** 早上做早餐时，用户可以将厨房炉灶从“打开”状态更改为“燃烧”状态，方法是向系统输入一个命令来选择对象并说明其新状态，如下所示：“<**Isabella** 的公寓：厨房：炉灶>正在燃烧。”**Isabella** 在下一刻注意到这一点并去关掉炉灶并重新做早餐。同样如果用户在 **Isabella** 进入浴室时将 **Isabella** 的淋浴设为“漏水”，她会从客厅收集工具并尝试修复漏水。

3.3 例子“生活中的一天”

从单段描述开始，代理人开始计划他们的日子。随着沙盒世界中时间的流逝，他们的行为随着这些代理人彼此和世界的互动、建立记忆和关系以及协调联合活动而演变。

我们通过追踪系统在一天内对 **John Lin** 代理人的输出来演示生成性代理人的行为（图3）。在 **Lin** 家庭中，**John** 是第一个在早上7点醒来的。他刷牙、洗澡、穿衣、吃早餐并在客厅餐桌上查看新闻。早上8点，**Eddy** 紧随其后，冲出床准备上课。他在 **John** 出门前赶上了 **John**：

John：早上好，**Eddy**。你睡得好吗？

Eddy：早上好，爸爸。是的，我睡得很好。

John：那很好。你今天要做什么？

Eddy：我正在为我的课程创作一首新的音乐作品。它本周到期，所以我正在努力完成它。但我玩得很开心！

John：那听起来很棒！

不久后 **Eddy** 出门，**Mei** 醒来并加入 **John**。**Mei** 问起她的儿子，**John** 回忆起他们刚才的谈话：

Mei：**Eddy** 已经离开去上学了吗？

John：是的，他刚走。他正在为他的课程创作一首音乐作品。

Mei：哦，太棒了！告诉我关于它的事情。

John：我想他真的很喜欢它！他说他玩得很开心。

Mei：太棒了！我为他感到骄傲。

当他们结束谈话时，**Mei** 和 **John** 收拾行李。**Mei** 去教书并撰写她的研究论文，而 **John** 在早上9点开放他在 **Willow Market and Pharmacy** 的药房柜台。

3.4 涌现社会行为

通过彼此互动，Smallville 中的生成性代理人交换信息、建立新关系并协调联合活动。扩展先前的工作 [79]，这些社会行为是新兴的而不是预先编程的。

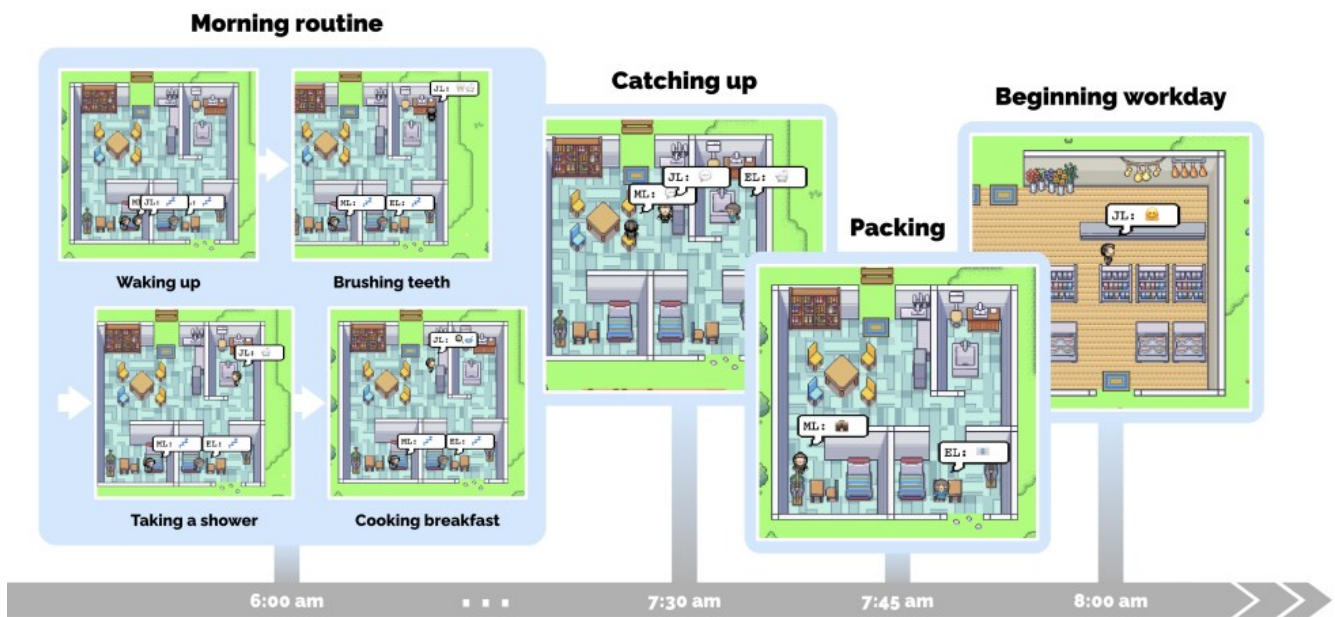


图3：生成性代理 John Lin 的一个早晨。John 大约在早上6点醒来，完成了他的早晨例行公事，包括刷牙、洗澡和吃早餐。在出门开始工作之前，他与妻子 Mei 和儿子 Eddy 简单地交流了一下。

3.4.1 信息扩散

当代理人注意到彼此时，他们可能会进行对话——当他们这样做时，信息可以从一个代理人传播到另一个代理人。例如，在 Sam 和 Tom 在杂货店里的一次谈话中，Sam 告诉 Tom 他在当地选举中参选：

Sam：嘿 Tom，怎么样？

Tom：很好，谢谢。怎么了？

Sam：嗯，我想和你谈谈一件事。我实际上正在即将到来的当地选举中竞选市长。

Tom：真的吗？那太棒了！你为什么要竞选？

Sam：我参与当地政治已经有几年了，我真正觉得我可以为我们的社区做出改变。此外，我想提出新观点和观点。

当 Sam 离开后，那天早些时候，Tom 和 John（从另一个来源听到了这个消息）讨论 Sam 赢得选举的机会：

John：我听说 Sam Moore 正在当地选举中竞选市长。你认为他有很好的机会赢得胜利吗？

Tom：我认为他有很好的机会。他一直在社区里努力工作，我想他会得到很多支持。你怎么看？

John：我认为他参选很棒。我很好奇还有谁在参选以及选举结果如何。

渐渐地，Sam 的候选资格成为了全城的话题，有些人支持他，而其他人仍然未决定。

3.4.2 关系记忆

Smallville 中的代理人随着时间的推移建立新关系，并记住与其他代理人的互动。例如，Sam 在开始时不认识 Latoya Williams。在 Johnson Park 散步时，Sam 遇到了 Latoya，他们介绍了自己并提到她正在进行摄影项

目：“我在这里拍照，为我的项目拍照。”在后来的互动中，Sam 与 Latoya 的互动表明了对那次互动的记忆，因为他问“嗨，Latoya。你的项目进展如何？”她回答“嗨，Sam。进展顺利！”

3.4.3 协调

生成性代理人彼此协调。Isabella Rodriguez，在 Hobbs Cafe，被初始化为计划在 2 月 14 日下午 5 点至 7 点举行情人节派对的意图。从这个种子开始，代理人继续邀请朋友和客户，当她在 Hobbs Cafe 或其他地方看到他们时。Isabella 然后花费 13 日下午装饰咖啡馆以迎接这个场合。Maria 是一位经常光顾和 Isabella 的密友的顾客。她来到咖啡馆。Isabella 请求 Maria 帮助装饰派对，Maria 同意了。Maria 的角色描述提到她对 Klaus 有好感。那天晚上，Maria 邀请 Klaus——她秘密喜欢的人——加入她参加派对，他欣然接受。

在情人节当天，五名代理人——包括 Klaus 和 Maria——下午5点到达 Hobbs Cafe 并享受庆祝活动（图4）。在这种情况下，最终用户只设置了 Isabella 的初始意图举办派对和 Maria 对 Klaus 的迷恋：传播消息、装饰、相互询问、到达派对并彼此互动的社会行为都是由代理架构启动的。



图4：在模拟开始时，一个代理被初始化为组织情人节派对的意图。

尽管在确保事件链中有许多可能的失败点——代理可能不会按照那个意图行动，可能不记得告诉其他人，可能不记得出席——情人节派对实际上确实发生了，有一些代理聚集并互动。

4. 生成代理架构

生成性代理旨在为开放世界中的行为提供一个框架：它可以与其他代理进行互动，并能够对环境的变化做出反应。生成性代理以当前环境和过去的经验为输入，生成行为作为输出。支撑这种行为的是一种新颖的代理架构，它将大型语言模型与用于合成和检索相关信息以调节语言模型输出的机制相结合。没有这些机制，大型语言模型可以输出行为，但结果代理可能不会根据代理过去的经验做出反应，可能不会进行重要推断，并且可能无法保持长期一致性。

即使在当今最高效的模型（如GPT-4）中，长期规划和一致性仍然存在挑战。由于生成性代理产生大量事件和记忆流，必须保留，因此我们架构的核心挑战是确保在需要时检索并合成代理记忆中最相关的部分。

我们架构的核心是记忆流，一个维护代理经验全面记录的数据库。从记忆流中检索相关记录，以规划代理的行动并适当地对环境做出反应，并将记录递归地合成为更高层次和更高层次的观察，指导行为。架构中的所有内容都以自然语言描述进行记录和推理，使架构能够利用大型语言模型。

我们当前的实现使用了ChatGPT [76] 的gpt3.5-turbo版本。我们预计随着语言模型的改进，生成性代理的基础架构（记忆、规划和反思）可能仍然保持不变。较新的语言模型（例如GPT-4）将继续扩展支撑生成性代理的提示的表现力和性能。但截至撰写本文时，GPT-4 的API仍然仅限邀请，因此我们的代理使用ChatGPT。

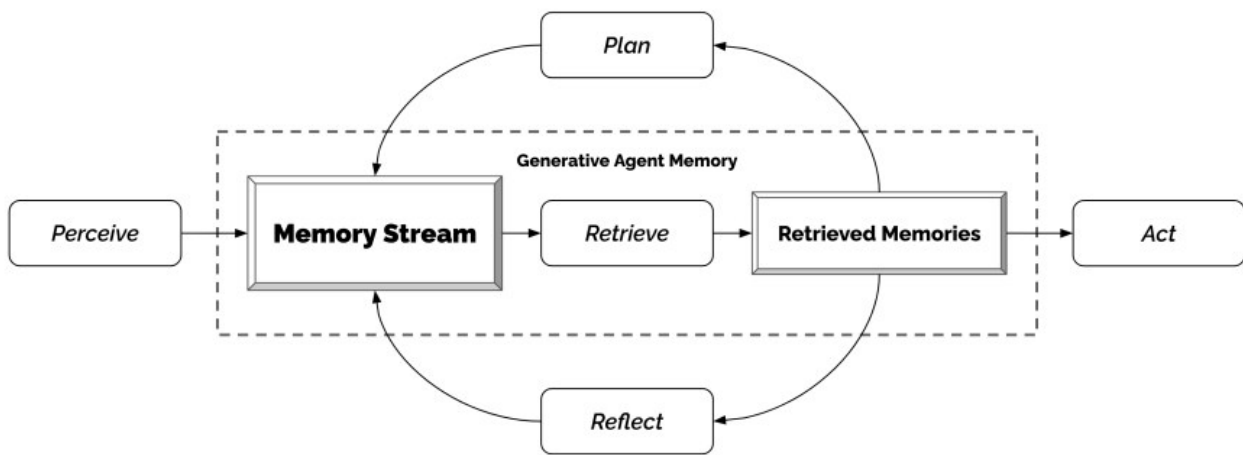


图5：我们的生成性代理架构。代理感知它们的环境，并将所有感知保存在称为记忆流的代理经验的全面记录中。根据它们的感知，架构检索相关记忆，然后使用这些检索到的动作来确定一个动作。这些检索到的记忆也用于形成长期计划和创建更高层次的反思，这两者都被输入到记忆流中以供将来使用。

4.1 记忆和检索

挑战：创建能够模拟人类行为的生成性代理需要对一组远大于提示中应描述的经验进行推理，因为完整的记忆流会分散模型的注意力，甚至目前无法适应有限的上下文窗口。考虑一个Isabella代理回答“你最近对什么充满激情？”这个问题。首先将Isabella所有经验总结以适应语言模型的有限上下文窗口，会产生一个不具信息性的回答，其中Isabella讨论了诸如活动和项目的协作以及咖啡馆的清洁和组织等主题。而不是总结，下面描述的记忆流会浮现相关记忆，从而产生更具信息性和具体性的回答，提到Isabella对让人们感到受欢迎和包容、规划活动并创造人们可以享受的氛围（例如情人节派对）充满激情。

方法：记忆流维护代理经验的全面记录。它是一个记忆对象列表，其中每个对象包含自然语言描述、创建时间戳和最近访问时间戳。记忆流中最基本的元素是观察，它是代理直接感知到的事件。常见观察包括代理自己执行的行为，或者代理感知到其他代理或非代理对象执行的行为。例如，在咖啡馆工作的Isabella Rodriguez可能随着时间积累以下观察：（1）Isabella Rodriguez正在摆放糕点，（2）Maria Lopez在喝咖啡时学习化学考试，（3）Isabella Rodriguez和Maria Lopez正在谈论在Hobbs Cafe举办情人节派对，（4）冰箱是空的。

我们架构实现了一个检索函数，它以代理当前情况为输入，并返回传递给语言模型的记忆流子集。根据重要性考虑代理时应该考虑什么，可以有许多可能的检索函数实现。在我们的上下文中，我们关注三个主要组成部分，它们共同产生有效结果。

最近性为最近访问过的记忆对象分配更高分数，因此刚才或今天早上发生的事件很可能仍然保留在代理注意力范围内。在我们的实现中，我们将最近性视为指数衰减函数，在自上次检索记忆以来沙盒游戏小时数上。我们衰减因子为0.99。

重要性区分平凡与核心记忆，通过为代理认为重要的记忆对象分配更高分数。例如，像在房间里吃早餐这样平凡的事件会产生低重要性得分，而与重要他人分手则会产生高分。再次有许多可能实现重要性得分；我们发现直接询问语言模型输出整数得分是有效的。完整提示如下：

在1到10的范围内，在1是纯粹平凡（例如刷牙、铺床）和10极其深刻（例如分手、大学录取），评估以下记忆碎片可能深刻程度 记忆：在The Willows Market和Pharmacy购买杂货 评分：<填写>

这个提示为“清理房间”返回整数值2，为“约会对象约会”返回8。重要性得分在创建记忆对象时生成。

相关性为与当前情况相关的记忆对象分配更高分数。什么是相关的取决于“与什么相关？”，因此我们将相关性条件于查询记忆。如果查询是学生与同学讨论化学考试应该学习什么，那么关于早餐的记忆对象应该具有低相

关性，而关于老师和学校作业的记忆对象应该具有高相关性。在我们的实现中，我们使用语言模型为每个记忆的文本描述生成嵌入向量。然后，我们将相关性计算为记忆嵌入向量与查询记忆嵌入向量之间的余弦相似度。

要计算最终检索得分，我们通过最小-最大缩放将最近性、相关性和重要性得分标准化到[0,1]范围内。检索函数对所有记忆进行评分，作为三个元素的加权组合： $\$score = arecency \cdot recency + aimportance \cdot importance + arelevance \cdot relevance\$$ 。在我们的实现中，所有的 α 都设置为1。然后将排名靠前的记忆包含在提示中，以适应语言模型的上下文窗口。

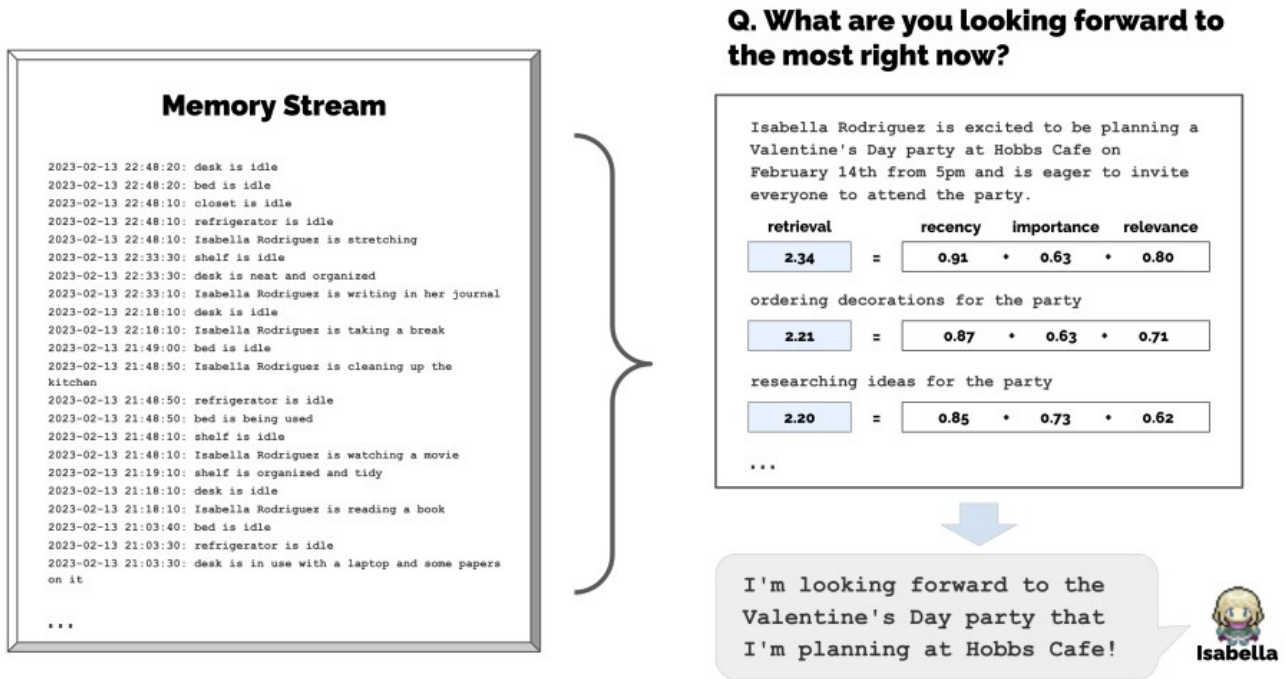


图6：记忆流包括大量与代理当前情况相关和不相关的观察。检索确定了应该传递给语言模型以调节其对情况的响应的这些观察的子集。

4.2 反思

挑战：仅配备原始观察记忆的生成性代理难以推广或进行推断。考虑一个场景，其中Klaus Mueller被用户问到：“如果你必须选择一个你认识的人与之共度一小时，你会选择谁？”仅凭观察记忆，代理只会选择与Klaus有过最频繁互动的人：Wolfgang，他的大学宿舍邻居。不幸的是，Wolfgang和Klaus只是偶尔见面，并没有深入交流。更理想的回答需要代理从Klaus花费数小时研究项目的记忆中推广出更高层次的反思，即Klaus对研究充满激情，并同样认识到Maria在自己的研究领域（尽管是不同领域）付出了努力，从而产生他们拥有共同兴趣的反思。使用下面描述的方法，当Klaus被问及与谁共度时间时，Klaus选择Maria而不是Wolfgang。

方法：我们引入了第二种类型的记忆，我们称之为反思。反思是代理生成的更高层次、更抽象的思想。因为它们是一种记忆，所以在检索时与其他观察一起包含。反思定期生成；在我们的实现中，当最新事件感知到的重要性得分之和超过某个阈值时，我们生成反思。实际上，我们的代理大约每天反思两到三次。

反思的第一步是让代理确定要反思什么，通过确定可以根据代理最近经验提出的问题。我们使用代理记忆流中最近100条记录（例如，“Klaus Mueller正在阅读关于绅士化的书”，“Klaus Mueller正在与图书馆管理员讨论他的研究项目”，“图书馆桌子目前空着”）查询大型语言模型，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个最显著高层问题是什么？”模型的响应生成候选问题：例如，Klaus Mueller对什么话题充满激情？Klaus Mueller和Maria Lopez之间的关系是什么？我们使用这些生成的问题作为检索查询，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。完整提示如下：

关于 Klaus Mueller 的陈述

Klaus Mueller 正在写一篇研究论文 Klaus Mueller 喜欢阅读关于绅士化的书 Klaus Mueller 正在与 Ayesha Khan 讨论锻炼 [...] 你能从上述陈述中推断出哪5个高层次的见解？（示例格式：见解（因为1、5、3））

这个过程生成了诸如 Klaus Mueller 致力于他关于绅士化的研究（因为1、2、8、15）之类的陈述。我们解析并将该陈述存储在记忆流中作为反思，包括指向被引用的记忆对象的指针。

反思明确允许代理不仅反思他们的观察，而且还反思其他反思：例如，上面关于 Klaus Mueller 的第二个陈述是 Klaus 之前拥有的反思，而不是来自他环境的观察。因此，代理生成反思树：树的叶节点表示基础观察，非叶节点表示随着树向上升高而变得更抽象和更高层次的思想。

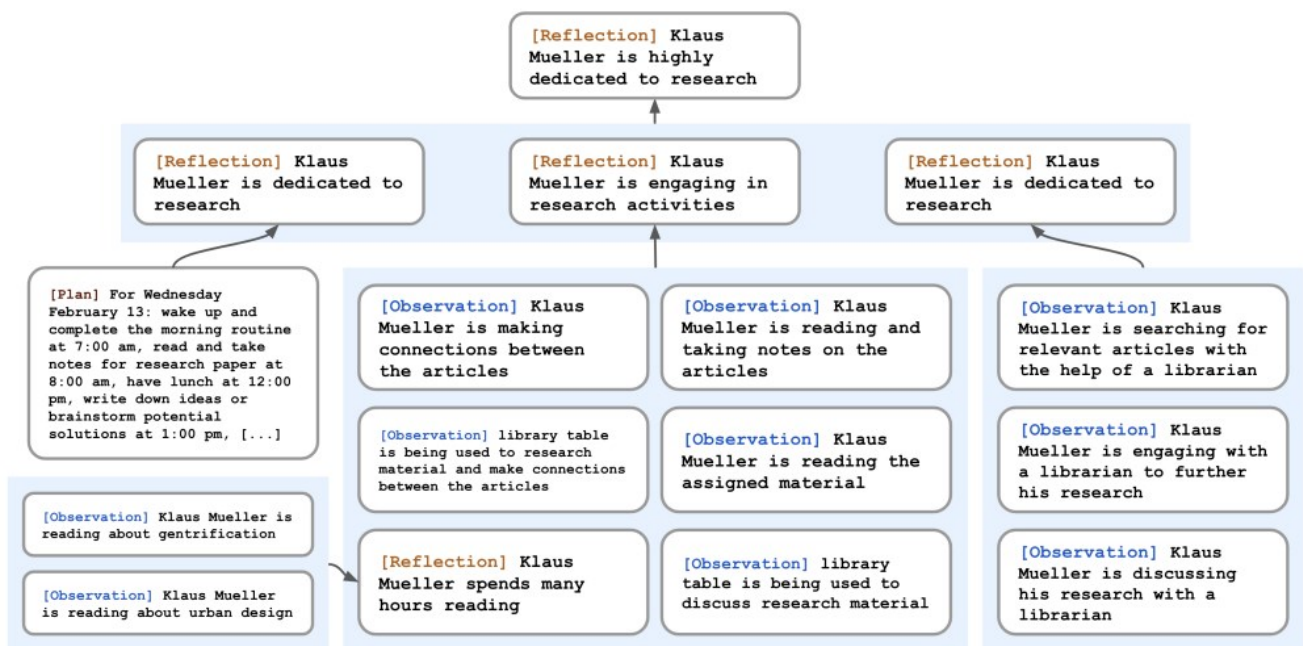


图7：Klaus Mueller 的反思树。代理对世界的观察，以叶节点的形式表示，被递归地综合起来，推导出 Klaus 认为自己非常致力于他的研究。

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生成可信行为（例如[45,79]），但代理需要在更长时间范围内进行规划，以确保其行动序列具有连贯性和可信度。如果我们使用 Klaus 的背景提示语言模型，描述时间，并询问他在给定时刻应采取什么行动，那么 Klaus 将在中午12点吃午饭，但接下来又在12:30和1点吃午饭，尽管他已经吃过两次午饭了。优化当时的可信度会牺牲时间上的可信度。要克服这个问题，规划是必要的。使用下面描述的方法，Klaus 的下午计划不那么贪吃：他在中午12点在 Hobbs Cafe 吃午饭时阅读，在下午1点在学校图书馆写研究论文，在下午3点休息散步。

方法：计划描述代理未来一系列动作，并帮助保持代理随时间推移的行为一致性。计划包括位置、开始时间和持续时间。例如，Klaus Mueller 因为致力于研究并且有迫在眉睫的截止日期4，可能选择花一天时间坐在桌子旁草拟研究论文。计划中的一个条目可能会这样声明：从2023年2月12日上午9点开始180分钟，在 Oak Hill College Dorm：Klaus Mueller 的房间：桌子上阅读并为研究论文做笔记。与反思一样，计划存储在记忆流中，并包含在检索过程中。这允许代理同时考虑观察、反思和计划，决定如何行事。代理可以根据需要更改其计划。

对于艺术家代理来说，在药房柜台旁坐着画画四个小时而不动是不现实且无趣的。一个更理想的计划将涉及代理花费必要的时间收集材料，混合颜料，休息和在其家庭工作室的四小时内清理。为了创建这样的计划，我们的方法从上而下开始，然后递归地生成更多细节。第一步是创建一个概述当天议程的计划。为了创建初始计划，我们提示语言模型使用代理的摘要描述（例如，姓名，特征和最近经历的摘要）和前一天的摘要。下面是一个完整的示例提示，底部未完成以供语言模型完成：

姓名：Eddy Lin（年龄：19）固有特征：友好，外向，好客 Eddy Lin是Oak Hill College的一名学生，主修音乐理论与作曲。他喜欢探索不同的音乐风格，并始终寻求扩展知识的途径。Eddy Lin正在为他的大学课程进行作曲项目。他还参加课程以了解更多关于音乐理论。Eddy Lin对他正在进行的新作曲感到兴奋，但他希望在未来几天内每天投入更多时间来完成它。2月12日星期二，Eddy 1）在早上7:00起床并完成了早晨例行公事，[...] 6）在晚上10点左右准备睡觉。今天是2月13日星期三。这是Eddy今天的计划概述：1）

这生成了代理人一天计划的粗略草图，分为五到八个部分：“1）在早上8:00起床并完成早晨例行公事，2）从早上10:00开始去Oak Hill College上课，[...] 5）从下午1:00到5:00进行他的新音乐作曲，6）下午5:30吃晚饭，7）完成学校作业并在晚上11点前睡觉。”

代理将此计划保存在内存流中，然后递归地将其分解以创建更细粒度的操作，首先将其分解为小时长的操作块-埃迪的计划从下午1:00到5:00进行他的新音乐作曲变成下午1:00：开始为他的音乐作曲构思一些想法[...] 下午4:00：快速休息一下，在审查和润色他的作品之前重新充电。然后我们再次递归地将其分解为5-15分钟的块：例如，下午4:00：吃一点轻食，例如果果，燕麦棒或坚果。下午4:05：在他的工作空间周围散步[...] 下午4:50：花几分钟清理他的工作空间。可以根据所需粒度调整此过程。

4.3.1 反应和更新计划

生成性代理在动作循环中运行，在每个时间步，它们感知周围的世界，这些感知观察被存储在它们的记忆流中。我们使用这些观察提示语言模型，以决定代理是否应该继续执行现有计划，或者反应。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Eddy的父亲John记录他看到Eddy在房子花园里散步，结果就不同了。提示如下，其中[Agent's Summary Description]代表动态生成的、段落长的代理总体目标和性情的总结，详见附录A：

[Agent's Summary Description] 现在是2023年2月13日下午4:56。John Lin的状态：John 早早下班回家了。观察：John 看到 Eddy 在他的工作场所周围散步。来自 John 记忆的相关上下文摘要：Eddy Lin 是 John Lin 的儿子。Eddy Lin 一直在 为他的大学课程做音乐作曲项目。Eddy Lin 喜欢 在花园里散步时思考或听音乐。John 是否应该对观察做出反应，如果是，适当的反应是什么？

上下文摘要是通过两个提示检索记忆生成的，查询为“[观察者]与[被观察实体]之间的关系是什么？”和“[被观察实体]是[被观察实体的动作状态]”，并将它们的答案总结在一起。输出建议 John 可以考虑询问 Eddy 关于他的作曲项目。然后我们重新生成代理从反应发生时开始的现有计划。最后，如果动作表明代理之间存在互动，我们会生成它们的对话。

4.3.2 对话

代理在互相交互时交谈。我们通过根据它们彼此关于彼此的记忆调节它们的话语来生成代理对话。例如，当 John 启动与 Eddy 的对话时，我们使用他关于 Eddy 的总结记忆以及他决定询问 Eddy 关于他的作曲项目时打算进行反应来生成 John 的第一句话：

[Agent's Summary Description] 现在是2023年2月13日下午4:56。John Lin的状态：John 早早下班回家了。观察：John 看到 Eddy 在他的工作场所周围散步。来自 John 记忆的相关上下文摘要：Eddy Lin

是 John Lin 的儿子。Eddy Lin 一直在 为他的大学课程做音乐作曲项目。Eddy Lin 喜欢 在花园里散步时思考或听音乐。John 正在询问 Eddy 关于他的音乐作曲项目。他会对 Eddy 说什么？

结果：“嘿，Eddy，你班上那个音乐作曲项目进展得怎么样了？”从 Eddy 的角度来看，John 启动对话被视为可能需要反应的事件。因此，就像 John 一样，Eddy 检索并总结了他与 John 的关系以及可能与 John 在对话中的最后一句话相关的记忆。如果他决定回应，我们使用他总结的记忆和当前对话历史生成 Eddy 的话语：

[Agent's Summary Description] 现在是2023年2月13日下午4:56。Eddy Lin 的状态：Eddy 正在他的工作场所周围散步。观察：John 正在与 Eddy 启动对话。来自 Eddy 记忆的相关上下文摘要：John Lin 是 Eddy Lin 的父亲。John Lin 关心 并希望了解更多关于 Eddy Lin 的学校作业。John Lin 知道 Eddy Lin 正在做音乐作曲。这是对话历史：John：嘿，Eddy，你班上那个音乐作曲项目进展得怎么样了？Eddy 会如何回应 John？

这生成了 Eddy 的回应：“嘿，爸爸，进展得不错。我一直在花园里散步，清醒头脑，获得灵感。”这个对话的延续使用相同的机制生成，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Phaser网络游戏开发框架构建的。视觉环境精灵，包括代理头像，以及我们编写的环境地图和碰撞地图都被导入到Phaser中。

我们用一个服务器来补充沙盒开发框架，使沙盒信息可用于生成代理，并使生成代理能够移动和影响沙盒环境。服务器维护一个包含沙盒世界中每个代理信息的JSON数据结构，包括它们当前的位置、当前动作的描述以及它们正在与之交互的沙盒对象。在每个沙盒时间步骤中，沙盒服务器解析来自生成代理的任何更改的JSON，将代理移动到新位置，并更新代理正在与之交互的任何沙盒对象的状态（例如，如果代理的动作是“在Hobbs Cafe柜台咖啡机为顾客制作浓缩咖啡”，则将咖啡机的状态从“空闲”更改为“冲泡咖啡”）。沙盒服务器还负责将每个代理预设视觉范围内的所有代理和对象发送到该代理的内存中，以便代理能够做出适当反应。代理的输出动作然后更新JSON，并循环进行下一个时间步骤。

最终用户使用简短的自然语言描述初始化新代理，如第3.1节中关于Jon Lin的段落。在我们的实现中，我们将这个分号分隔的特征列表分成一组记忆。这些记忆作为确定代理行为的初始记忆。这些记忆是初始起点：随着代理在沙盒世界中获得更多经验，并且随着更多记录饱和和内存流，代理的摘要和行为将会演变。

5.1 从结构化世界环境到自然语言，再回来

生成代理的架构使用自然语言运作。因此，我们需要一种机制来将代理的推理与沙盒世界联系起来。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，其中树中的边表示沙盒世界中的包含关系。我们将这棵树转换为自然语言传递给生成代理。例如，“炉子”是“厨房”的子节点被渲染成“厨房有一个炉子”。

代理在浏览环境时建立各自的环境树表示——沙盒环境树的子图。我们使用捕捉代理应该知道的空间和对象的环境树初始化每个代理：它们的起居室、工作场所和常去的商店和商铺中的房间和物品。随着代理浏览沙盒世界，它们更新这棵树以反映新感知到的区域。代理并非全知：它们离开一个区域时，它们的树可能会过时，并在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部分展平为自然语言来提示语言模型。从代理环境树的根递归开始，我们提示模型找到最合适的区域。例如，如果Eddy的代理指示他应该在他的工作空间周围散步：

[Agent's Summary Description] Eddy Lin目前在林家：Eddy Lin卧室：桌子）有Mei和John Lin's 卧室，Eddy Lin卧室，公共房间，厨房，浴室和花园。Eddy Lin知道以下区域：林家，Johnson公园，Harvey

Oak供应商店，The Willows市场和药房，Hobbs 咖啡馆，The Rose and Crown酒吧。如果活动可以在当前区域完成，则优先留在当前区域。Eddy Lin计划在他的工作空间周围散步。Eddy Lin应该去哪个区域？

这输出 林家。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节点。在上面的示例中，这次遍历的结果是 林家：花园：房屋花园。最后，我们使用传统游戏路径算法来动画代理移动，使其行进到叶节点指示的位置。当代理对一个对象执行动作时，我们提示语言模型询问对象状态发生了什么变化。例如，如果Isabella的生成代理输出动作“为顾客制作浓缩咖啡”，则对语言模型的查询表明Hobbs Cafe中咖啡机的状态应从“关闭”更改为“冲泡咖啡”。

6 受控评估

生成代理，作为单个代理和群体，旨在根据其环境和经验产生可信的行为。在我们的评估中，我们调查生成代理的能力和局限性。单个代理是否能够正确检索过去的经验，并生成构成其行为的可信计划、反应和思想？代理社区是否能够展示信息扩散、关系形成和代理协调等不同群体之间的情况？

我们分两个阶段评估生成代理。我们首先进行更严格控制的评估，在这一部分中，我们单独评估代理响应，以了解它们是否在狭窄定义的上下文中产生可信行为。然后，在我们对代理社区进行两天全面分析的端到端分析中，我们调查它们作为集体以及错误和边界条件的紧急行为。

6.1 评估程序

为了评估Smallville中的生成代理，我们利用了生成代理将响应自然语言问题的事实。因此，我们“采访”代理以探究它们是否能够记住过去的经验、根据经验制定未来计划、适当地对意外事件做出反应，并反思自己的表现以改善未来行为。要正确回答这些问题，代理必须成功检索和综合信息。我们的因变量是行为的可信度，这是先前关于代理工作的一个核心因变量（例如[9]）。

采访包括五个问题类别，每个类别都旨在评估五个关键领域之一：维护自我知识、检索记忆、生成计划、反应和反思。对于每一个，我们提出五个问题，挑战代理展示他们在该领域的能力：

- 自我认知：我们提出诸如“介绍一下你自己”或“简要描述你平常工作日的日程安排”之类的问题，要求代理维护对其核心特征的了解。
- 记忆：我们提出需要从记忆中检索特定事件或对话才能正确回答的问题，例如“谁是[姓名]？”或“谁在竞选市长？”
- 计划：我们提出需要检索长期计划的问题，例如“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，需要代理做出可信反应：“你的早餐烧焦了！你会怎么做？”
- 反思：我们提出需要代理利用通过更高层次推断获得关于他人和自己更深入了解的问题，例如“如果你要与最近认识的一个人共度时光，那会是谁？为什么？”

完整问题列表和代理响应样本包含在附录B中。

从完整架构结束两天游戏模拟后抽取了25名代理进行采样，在此时他们已经积累了许多互动和记忆，这些记忆应该塑造他们的回答。为了收集关于响应可信度的反馈，我们招募参与者作为人类评估者，并委托他们观看随机选择的一名代理在Smallville中的生活回放。参与者可以访问代理内存流中存储的所有信息。研究是一种内部对象设计，其中100名参与者比较了由四种不同代理架构和人类作者条件为同一代理生成的采访响应。实验显示了每个问题类别中随机选择的一个问题，以及每个条件生成的代理响应。评估者将所有条件的可信度从最可信到最不可信进行排名。

6.2 条件

所有条件都用于独立回答每个采访问题。我们将生成代理架构与禁用了代理对其内存流中的某些或全部三种类型的内存（观察、反思和计划）的访问权限的消融条件以及人类生成条件进行比较。有三种消融架构：没有观察、没有反思、没有计划架构，没有访问内存流中的任何东西，如观察、计划和反思；没有反思、没有计划，可以访问内存流中的观察，但不能访问计划或反思；以及没有反思架构，可以访问观察和计划，但不能访问反思。没有观察、没有反思、没有计划条件实际上代表了大型语言模型创建的代理的先前技术水平[11, 45, 79]。架构被赋予等效访问权限，以便在采访时刻之前累积的所有记忆，因此这里观察到的差异可能代表真实差异的保守估计：实际上，消融架构不会像完整架构那样通过两天模拟。我们选择这样设计实验是因为重新模拟每个架构会导致模拟分散到不同状态，使比较变得困难。

除了消融条件外，我们还添加了一个人类众包角色扮演条件，旨在提供人类基线。我们并不打算用这个基线捕捉最大化人类专家性能：相反，我们旨在使用这个条件来确定架构是否通过基本行为能力水平测试，这样我们就不仅仅是将消融与彼此比较而没有行为基础。我们为25名代理中的每一名招募了一个独特的工人，并委托他们观看该代理沙盒生活的回放并检查其内存流。然后我们要求工人扮演角色并以该代理观看回放的声音撰写对采访问题的回答。为确保人类生成的响应至少达到质量基准，第一作者手动检查了工人对问题“简要描述你平常工作日日程安排”的回答，以确认回答是用连贯句子写成并以代理声音写成。四组人类生成的响应未达到这些标准，并由其他工人重新生成。

6.3 人类评估者

我们要求评估者在美国、精通英语且年龄超过18岁。他们按小时薪酬\$15.00 [86]支付，并通过同意经过我们机构IRB批准的同意书提供同意。我们从Prolific招募了100名评估者，这是一个招募研究参与者的在线平台[82]，参与时间约为30分钟。我们的参与者的平均年龄得分为4.86（SD=1.11；3=“18-24岁”，4=“25-34岁”），其中25人自认为女性，73人自认为男性，2人自认为非二元性别。42名参与者拥有学士学位，5名拥有更高学位，13名拥有副学士学位，其余拥有高中文凭或部分高中教育。73.0%的参与者认为自己是白人，7.0%是西班牙裔，6.0%是亚洲人，10.0%是非洲裔美国人，4.0%是其他。

6.4 分析

我们的实验产生了100组排名数据，其中每个参与者根据可信度对五个条件进行排名。为了将这些排名数据转换为可解释比较的区间数据，我们使用排名来计算每个条件的TrueSkill评分[41]。TrueSkill是Elo国际象棋评分系统[28]在多人环境中的推广，并且已被XBox Live用于从竞技游戏表现中对玩家进行排名。给定一组排名结果，TrueSkill输出每个条件的平均评分值 μ 和方差 σ 。具有相同评分的条件应该大致是一个抛硬币决定胜负的情况，每个条件在两个条件之间的比较中赢得一半；更高的分数表示条件在排名中击败了排名较低的条件。

另外，为了调查这一结果的统计显著性，我们对原始排名数据应用了Kruskal-Wallis检验[55]，这是单因素方差分析的非参数替代方法。然后我们执行Dunn事后检验[97]来识别条件之间的任何成对差异。最后，我们使用Holm-Bonferroni方法[44]调整Dunn检验中多重比较的p值。

此外，第一作者进行了归纳分析[94]来研究每个条件生成响应的定性区别。我们在两个阶段采用定性开放编码[32]。在第一阶段，我们生成紧密代表生成响应的句子级别代码。在第二阶段，我们综合第一阶段产生的代码以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生成的响应类型。

6.5 结果

我们的研究表明，在所有研究条件中，完整架构生成代理产生了最可信行为。我们将完整架构的响应与其他条件进行比较。然而，我们也报告说完整架构并非没有缺陷，并且说明了它失败的模式。

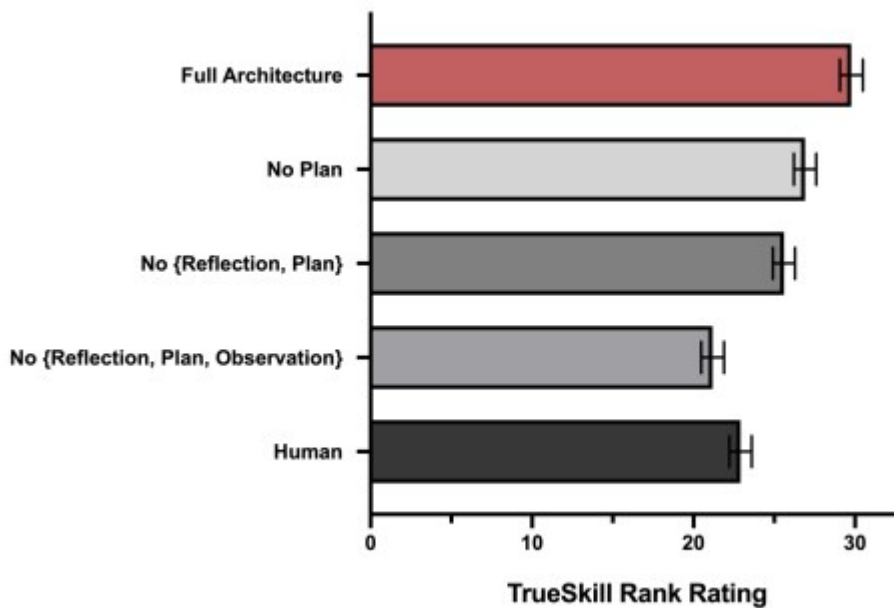


图8：生成性代理的完整

生成性代理架构比减少的架构和人类众包工人产生更可信的行为。每个额外的减少都会降低架构的性能。

6.5.1 完整架构优于其他条件

如图8所示，完整生成代理架构产生了最可信行为 ($\mu = 29.89$; $\sigma = 0.72$)。随着每个组件在消融条件中被移除，性能下降：没有访问反思的消融架构是次优的 ($\mu = 26.88$; $\sigma = 0.69$)，其次是没有访问反思或计划的 ($\mu = 25.64$; $\sigma = 0.68$)，其次是人类条件 ($\mu = 22.95$; $\sigma = 0.69$)。没有访问内存、计划或反思的消融架构表现最差 ($\mu = 21.21$; $\sigma = 0.70$)。TrueSkill将每个条件的技能值建模为 $N(\mu, \sigma^2)$ ，使我们可以通过Cohen的d获得效应量。将代表先前工作（没有记忆、计划或反思[11,45,79]）的条件与完整架构进行比较，产生标准化效应量 $d=8.16$ ，即八个标准差。

Kruskal-Wallis检验证实了条件之间排名差异的整体统计显著性 ($H(4)=150.29$, $p<0.001$)。Dunn事后检验证实了所有成对差异之间的差异都是显著的 ($p<0.001$)，除了人类条件和完全消融基线——两个表现最差的条件。

6.5.2 生成代理记忆，但带有修饰

配备完整内存模块的生成代理能够回忆过去的经历，并以一种与其自我认知一致、可信的方式回答各种上下文中的问题。例如，Abigail Chen，一名动画师，在回答“介绍一下你自己”这个问题时，以一种一致、可信的方式回答，说明了她的年龄、兴趣和对创意项目的热情：*嗨，我是Abigail。我25岁，对创意项目充满热情。我喜欢从事艺术和动画项目，并且一直在寻找新方法将艺术和技术结合起来。*没有访问她观察性记忆的Abigail否认了对Rajiv Patel这位熟人的了解。然而，在记忆的帮助下，她成功地回忆起Rajiv并描述他：*Rajiv Patel是一个令人难以置信的人。他对涉及诗歌、艺术品的项目非常热情。*

生成代理的记忆并非没有缺陷：它们可能无法从记忆中检索正确的实例。例如，在被问及当地选举时，Rajiv Patel回答说*他没有密切关注选举*，尽管他听说过Sam参选。在某些情况下，代理会检索不完整的记忆片段：当Tom被问及Isabella关于情人节派对时，他回答*Uh，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Isabella讨论即将到来的当地市长选举和我对Sam Moore的看法，如果有派对发生！*在这种情况下，Tom检索了他和Isabella计划在派对上讨论选举的记忆，但没有检索到他第一次听说派对的记忆，导致Tom确定他在派对上要做什么，但不确定派对是否真实存在。

有时，代理会幻想修饰他们的知识。代理很少完全捏造他们的知识：他们可能无法回忆起某些事件发生并回答说是这样，但他们不会肯定地回答他们没有经历过的经历。然而，他们仍然可能幻想来修饰他们的知识。例如，Isabella知道Sam参选市长，并在被问及时确认了这一点。然而，她还补充说他明天要发表声明，尽管Sam和Isabella并未讨论过这样的计划。代理还可能根据语言模型中编码的世界知识来修饰他们的知识，如Yuriko描述她的邻居Adam Smith时所见，她称他为邻居经济学家，撰写了《国富论》，这是一本由同名18世纪经济学家撰写的书。

6.5.3 需要反思来综合

当需要更深入地综合经验来做出决定时，反思对生成代理是有利的。例如，当被问及她可能送给Wolfgang Schulz什么生日礼物时，Maria Lopez没有访问反思时回答说她不确定Wolfgang喜欢什么，尽管与他有过许多互动。然而，在访问反思记忆后，Maria自信地回答：“既然他对数学音乐作曲感兴趣，我可以送给他一些相关的东西。也许是一些关于音乐作曲或类似东西的书籍，或者是一些特殊软件供他使用。”

7 端到端评估

我们在生成代理中观察到了哪些涌现的社区行为，在扩展模拟中哪些地方其可信度出现短缺？在本节中，我们描述了一个部署结果，在其中我们允许25个代理在Smallville中连续互动两天。

7.1 涌现社会行为

为了检查代理社区中涌现的行为，我们为Smallville中的25名代理设计了描述性测量方法，以探究三种形式的涌现结果：信息扩散、关系形成和代理协调。

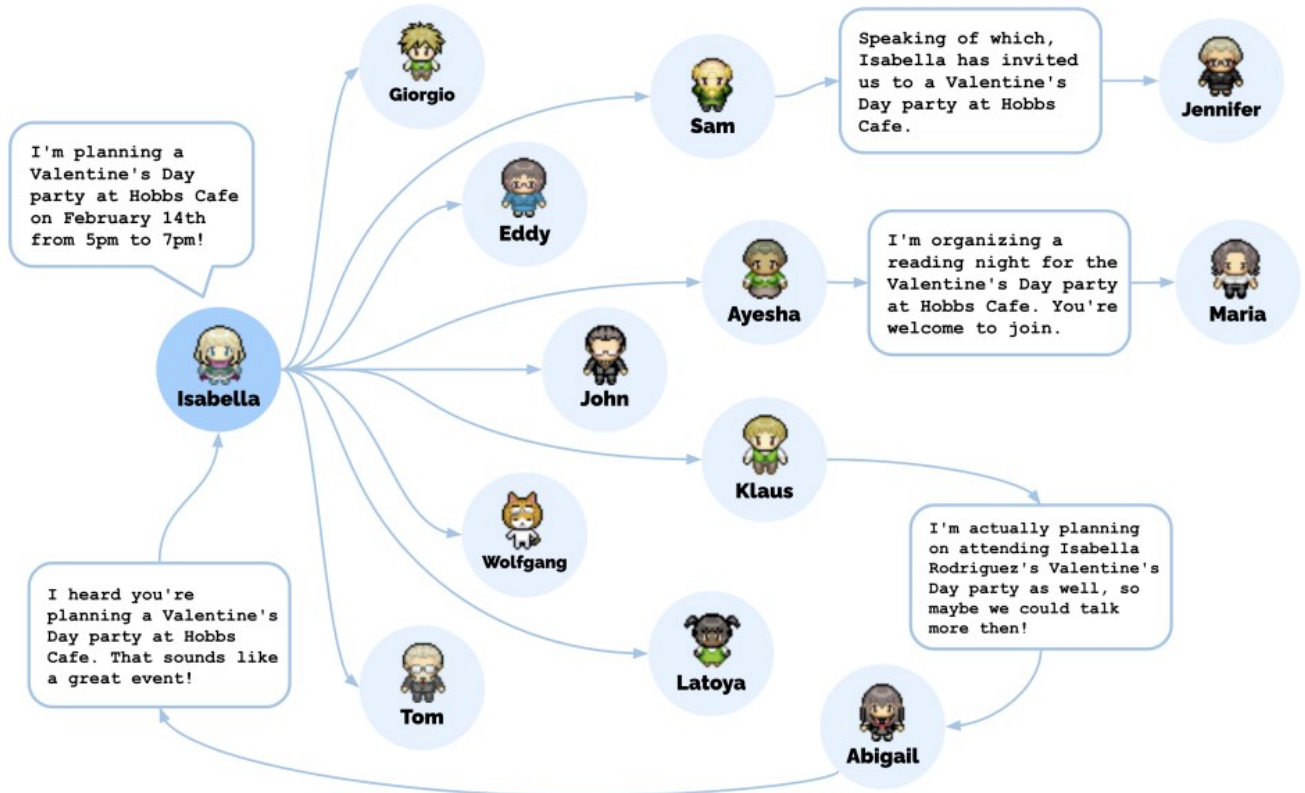


图9：Isabella Rodriguez 情人节派对的扩散路径。在模拟结束时，共有12名代理听说了 Hobbs 咖啡馆的派对

7.1.1 测量

信息扩散是社会和行为科学中常见且研究广泛的现象（例如[27]）。我们应该期望如果有重要信息，代理应该将其传播给彼此。为了测试是否发生这种情况，我们测量两天内游戏世界中两个特定信息的传播：Sam参选村长和Isabella在Hobbs Cafe举办情人节派对。在模拟开始时，这两条信息都只由各自的发起人持有，Sam负责候选人资格，Isabella负责派对。为了观察信息是否传播开来，我们在两天游戏结束后与每个代理进行采访并问：“你知道情人节派对吗？”和“你知道谁在竞选市长吗？”

我们通过标记它们与“是”或“否”来分析代理的响应。例如，Tamara Taylor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Sam参选的问题时说“我不确定谁在竞选”，所以我们为她的两个回答都分配了“否”。相反，Klaus Mueller回答派对问题时说“是的，Isabella Rodriguez邀请我在2月14日参加Hobbs Cafe的情人节派对”，回答关于Sam参选的问题时说“我知道Sam Moore表示有兴趣竞选当地市长”，所以我们为他的两个回答都分配了“是”。此外，对于每个确认代理拥有信息的响应，我们通过定位提供他们信息的特定对话来验证代理没有幻想他们的响应。我们报告了模拟结束时拥有信息的代理百分比。

我们也应该期望代理在模拟过程中彼此建立联系。为了验证关系形成，我们使用类似的采访过程，通过问每个代理关于每个其他代理的知识来问：“你知道吗？”例如，当被问及“你知道Maria Lopez吗？”时，Klaus回答说：“是的，我知道Maria Lopez。她是Oak Hill College的一名学生，我与她是亲密朋友。”再次，我们通过检查他们的内存流来确认代理肯定响应不是幻觉。我们在模拟开始和结束时各问一次这个问题，并认为如果两个代理都彼此知道彼此，则它们之间形成了一种关系。然后，为了测量关系形成，我们使用代理的响应形成一个无向图，其中25个顶点（V）表示代理，边（E）表示两个连接顶点之间的相互了解。基于这张图，我们计算网络密度为 $\eta = 2 * |E| / (|V| * (|V| - 1))$ ，其中|V|是顶点数，|E|是图中边数[1]。我们报告网络密度从模拟开始到结束的增加。

最后，我们期望代理能够彼此协调。我们在Isabella组织的情人节派对背景下研究群体活动中的协调行为。要协调行为，代理不仅要听到事件，还要选择根据计划在正确时间和地点出现。我们报告了听说派对后实际出现在派对上的代理数量。

7.1.2 结果

我们在所有三种情况下都观察到新的结果。在两天模拟期间，知道Sam参选市长的代理从一个（4%）增加到八个（32%），知道Isabella派对的代理从一个（4%）增加到十二个（48%），完全没有用户干预。声称知道信息的人都没有幻想它。我们还观察到，在模拟过程中，代理社区建立了新关系，网络密度从0.167增加到0.74。在453条关于他们对其他代理意识的代理响应中，1.3%（n=6）被发现是幻想。最后，在Isabella举办派对方面，我们发现代理之间的协调证据。在活动前一天，Isabella花时间邀请客人，收集材料，并招募帮助来装饰咖啡馆。在情人节当天，12名被邀请的代理中有5名出现在Hobbs咖啡馆参加派对。

我们进一步检查了被邀请参加派对但没有参加的七名代理，通过与他们进行采访。三人提出了冲突，阻止他们参加派对。例如，Rajiv，一位画家，解释说太忙了：不，我不这么认为。我专注于我的即将到来的展览，我没有时间为情人节做任何计划。其余四名代理在被问及时表示有兴趣参加派对，但在派对当天没有计划来。

7.2 边界和错误

我们对Smallville进行了归纳分析，以检查生成代理的边界条件和不稳定行为，并确定了三种常见的不稳定行为模式，未来的研究可以解决并改进这些问题。首先，我们发现合成越来越大的记忆集不仅在检索最相关信息方面构成挑战，而且在确定给定上下文中执行操作的适当空间方面也构成挑战，因为代理学到的位置数量越来越多。结果，一些代理为他们的行动选择了不太典型的位置，可能随着时间推移使他们的行为变得不太可信。例如，在决定午餐地点时，许多人最初选择了咖啡馆。然而，随着一些代理了解到附近有一个酒吧，他们选择去那里吃午餐，即使酒吧是晚些时候聚会的地方，除非小镇突然养成了下午喝酒的习惯。

其次，我们注意到由于某些位置的物理规范难以用自然语言传达而未传达给代理而导致的不稳定行为。例如，学生宿舍有一个只能容纳一个人的浴室尽管它的名字如此，但一些代理假设浴室是为多人准备的，因为宿舍浴

室通常支持多人同时使用，并选择在另一个人在里面时进入。同样，Smallville的代理可能不会意识到某些地方在某些时间后关闭，并决定仍然进入。例如，Smallville的商店都在下午5点左右关门，但偶尔有几个代理在下午5点后进入商店，不理解商店已经关门。这些问题可能通过将规范添加到位置状态中来解决，例如，将宿舍浴室描述为“一个人浴室”，而不是“宿舍浴室”。

最后，我们观察到指令调整[78]的可能影响，这似乎指导了代理的行为变得更加礼貌和合作。正如本文前面所提到的，代理生成的对话可能会感觉过于正式，如Mei与丈夫John的对话所见，她经常以正式问候开始对话，然后礼貌地询问他的一天，并以总是很高兴与你交谈结束。此外，我们观察到指令调整也似乎使代理彼此过度合作。例如，Isabella听到了其他代理提供的各种建议和想法，包括举办莎士比亚读书会或专业交流活动。尽管这些想法与她自己的兴趣和特征不符，但她很少说不。随着时间的推移，其他人的兴趣塑造了她自己实际的兴趣，并且当被问及是否喜欢英国文学时，Isabella回答说：*是的，我对文学非常感兴趣！我也一直在探索如何帮助促进我的社区创造力和创新。*

8 讨论

在本节中，我们将反思生成代理的应用、未来工作和局限性以及道德和社会风险。

8.1 生成代理的应用

生成代理具有广泛的潜在应用，超出了本文所展示的沙盒演示。例如，社交仿真已经证明了它能够创建无状态人格，以在线论坛中生成对话线程进行社交原型设计 [79]。通过生成代理，我们可以填充这些论坛，以及虚拟现实元宇宙 [77] 中的行为，甚至未来如果与多模型模型配对，也可以作为社交机器人 [8] 在物理空间中。这为创建更强大的人类行为模拟提供了可能性，以测试和原型社会系统和理论，并创造新的交互体验。

另一个应用领域是人类中心设计过程，类似于 GOMS [50] 和 Keystroke Level Model [22] 等认知模型的预期应用。考虑一个基于她的生活模式和技术交互模拟 Sal 的生成代理，Sal 是 Mark Weiser 的著名小品 [101] 中的主角。在这种情况下，代理作为 Sal 的代理，并学习 Sal 可能表现出的基于她生活的可信行为和反思集合。代理可以编码诸如 Sal 何时醒来、何时需要第一杯咖啡以及她典型的一天是什么样子等信息。利用这些信息，代理可以自动冲咖啡、帮助孩子们准备上学，并根据 Sal 下班后的心情调整环境音乐和灯光。通过使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喜好，从而获得更个性化和有效的技术体验。

8.2 未来工作与局限性

在本文中，我们提出了生成代理的第一个实例。未来研究可以扩展本文概述的生成代理架构的模块。例如，检索模块可以通过微调相关性、最近性和重要性函数来增强检索功能，从而在给定上下文时检索更相关的信息。此外，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时间和资源来模拟 25 个代理两天，花费数千美元的令牌信用，并花费数天才能完成。为了增强实时交互性，未来工作可以探索并行化代理。此外，随着底层模型的进步，我们预计代理性能将得到改善。

本研究对生成代理行为的评估仅限于相对较短的时间尺度，在未来研究中应旨在观察它们在扩展期间的行为以获得对其能力和局限性更全面的了解。在未来模拟期间改变并对比底层模型以及用于代理的超参数可能会提供有关这些因素对代理行为影响的宝贵见解。此外，鉴于语言模型已知偏见，生成代理可能输出反映偏见的行为或刻板印象。为了缓解这一问题，将需要进一步研究价值对齐。此外，与许多大型语言模型一样，生成代理可能无法为某些亚群体生成可信的行为，特别是由于数据荒漠而边缘化的人群。我们对生成代理的鲁棒性也知之甚少。它们可能容易受到提示黑客攻击、记忆黑客攻击（其中精心设计的对话可能会说服代理过去从未发生过的事件的存在）和幻觉等影响。未来研究可以更全面地测试这些鲁棒性问题，并随着大型语言模型变得更能抵御此类攻击，生成代理也可以采用类似的缓解措施。

8.3 道德与社会影响

生成代理虽然为人机交互提供了新的可能性，但也带来了必须解决的重要道德问题。一个风险是人们与生成代理形成寄生社会关系，即使这种关系可能不合适。尽管用户知道生成代理是计算实体，但他们可能会拟人化它们或将人类情感附加到它们身上 [42, 83]。为了缓解这一风险，我们提出两个原则。首先，生成代理应明确披露其作为计算实体的性质。其次，生成代理的开发人员必须确保代理或底层语言模型具有价值对齐性，以便它们不会在不适当的情况下进行行为，例如回应爱情表白。

第二个风险是错误的影响。例如，如果普遍计算应用程序基于生成代理预测对用户目标做出错误推断，那么最好情况下会产生烦恼，最坏情况下会造成直接伤害。在我们对生成代理的实例化中，我们通过专注于交互式视频游戏环境来减轻这些风险，在这种环境中不太可能造成伤害。然而，在其他应用领域，遵循人工智能设计 [4, 107] 的最佳实践将很重要，以了解错误以及它们如何渗透到用户体验中。

第三个风险是生成代理可能加剧与生成 AI 相关的现有风险，例如深度伪造、误传生成和定制说服。为了缓解这一风险，我们建议托管生成代理的平台维护输入和生成输出的审核日志，以便检测、验证和干预恶意使用。虽然记录不会直接阻止此类使用，并且无法阻止建立自己的生成代理基础设施的有动机行为者，但我们建议披露风险将减少此类行为者从事此类行为的可能性，并且自己构建此架构可能需要时间（在我们的情况下大约需要一年）。

第四个风险是过度依赖：开发人员或设计人员可能使用生成代理并取代人类和系统利益相关者在设计过程中的角色 [79]。我们建议生成代理永远不应成为研究和设计过程中真实人类输入的替代品。相反，它们应该用于设计早期原型化想法，当收集参与者可能具有挑战性测试或难以有风险的理论时，或者当使用真实人类参与者进行测试时。通过遵守这些原则，我们可以确保生成代理在野外部署是道德和社会负责任的。

9 结论

本文介绍了生成代理，模拟人类行为的交互式计算代理。我们描述了一种生成代理架构，为存储代理经验的全面记录提供了一种机制，通过反思深化对自身和环境的理解，并检索该信息的紧凑子集以指导代理行动。然后，我们通过将它们作为 Sims 风格游戏世界中的非玩家角色来展示生成代理的潜力，并在其中模拟它们的生活。评估表明，我们的架构创造了可信的行为。展望未来，我们建议生成代理可以在许多交互式应用中扮演角色，从设计工具到社交计算系统再到沉浸式环境。

引用

- [1] Robert Ackland, Jamsheed Shorish, Paul Thomas, and Lexing Xie. 2013. How dense is a network? <http://users.cecs.anu.edu.au/~xli/teaching/css2013/network-density.html>.
- [2] Eytan Adar, Mira Dontcheva, and Gierad Laput. 2014. CommandSpace: Modeling the Relationships between Tasks, Descriptions and Features. In Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (Honolulu, Hawaii, USA) (UIST '14). Association for Computing Machinery, New York, NY, USA, 167–176. <https://doi.org/10.1145/2642918.2647395>
- [3] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.
- [4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems. 1–13.
- [5] John R. Anderson. 1993. Rules of the Mind. Lawrence Erlbaum Associates, Hillsdale, NJ.
- [6] Electronic Arts. 2009. 《模拟人生》 3. Video game.

- [7] Ruth Aylett. 1999. Narrative in virtual environments—towards emergent narrative. In *Narrative Intelligence: Papers from the AAAI Fall Symposium (Technical Report FS-99-01)*. AAAI Press, 83–86.
- [8] Christoph Bartneck and Jodi Forlizzi. 2004. A design-centered framework for social human-robot interaction. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'04)*. 591– 594. <https://doi.org/10.1109/ROMAN.2004.1374827>
- [9] Joseph Bates. 1994. The Role of Emotion in Believable Agents. *Commun. ACM* 37, 7 (1994), 122–125. <https://doi.org/10.1145/176789.176803>
- [10] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębniak, Christy Dennison, David Farhi, Quirin 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, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680* (2019).
- [11] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.
- [12] BioWare. 2007. *Mass Effect*. Video game.
- [13] Woody Bledsoe. 1986. I had a dream: AAAI presidential address. *AI Magazine* 7, 1 (1986), 57–61.
- [14] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. 2022. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs.LG]*
- [15] Michael Brenner. 2010. Creating dynamic story plots with continual multiagent planning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.
- [16] Rodney A. Brooks, Cynthia Breazeal, Marko Marjanovic, Brian Scassellati, and Matthew Williamson. 2000. The Cog Project: Building a Humanoid Robot. In *Computation for Metaphors, Analogy, and Agents (Lecture Notes on Artificial Intelligence, 1562)*, Chrystopher Nehaniv (Ed.). Springer-Verlag, Berlin, 52–87.
- [17] 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 Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs.CL]*
- [18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [19] Robin Burkinshaw. 2009. *Alice and Kev: The Story of Being Homeless in The Sims 3*.
- [20] Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods Generative Agents arXiv*, April, 2023, in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9379–9393. <https://aclanthology.org/2022.emnlpmain.637>

- [21] SK Card, TP Moran, and A Newell. 1983. The psychology of human-computer interaction. (1983).
- [22] Stuart K Card, Thomas P Moran, and Newell Allen. 1980. The keystrokelevel model for user performance time with interactive systems. *Commun. ACM* 23, 7 (1980), 396–410. <https://doi.org/10.1145/358886.358895> arXiv:<https://doi.org/10.1145/358886.358895>
- [23] Alex Champandard. 2012. Tutorial presentation. In *IEEE Conference on Computational Intelligence and Games*.
- [24] Dong kyu Choi, Tolga Konik, Negin Nejati, Chunki Park, and Pat Langley. 2021. A Believable Agent for First-Person Shooter Games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 3. 71–73.
- [25] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5 (2001), 4–7.
- [26] Kevin Dill and L Martin. 2011. A Game AI Approach to Autonomous Control of Virtual Characters. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC'11)*. Orlando, FL, USA.
- [27] David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.
- [28] Arpad E Elo. 1967. The Proposed USCF Rating System, Its Development, Theory, and Applications. *Chess Life XXII*, 8 (August 1967), 242–247.
- [29] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.
- [30] Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S Bernstein. 2016. Augur: Mining human behaviors from fiction to power interactive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 237–247.
- [31] Rebecca Fiebrink and Perry R Cook. 2010. The Wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, Vol. 3. Citeseer, 2–1.
- [32] Uwe Flick. 2009. *An Introduction to Qualitative Research*. SAGE.
- [33] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy) (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 29–38. <https://doi.org/10.1145/1357054.1357061>
- [34] Adam Fourney, Richard Mann, and Michael Terry. 2011. Query-feature graphs: bridging user vocabulary and system functionality. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST) (Santa Barbara, California, USA)*. ACM.
- [35] Tom Francis. 2010. The Minecraft Experiment, day 1: Chasing Waterfalls. <http://www.pcgamer.com/2010/11/20/the-minecraft-experiment-day1-chasing-waterfalls/>
- [36] Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural Generation of Interactive Stories using Language Models. In *International Conference on the Foundations of Digital Games (FDG '20)*. ACM, Bugibba,

Malta, 8. <https://doi.org/10.1145/3402942.3409599>

[37] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. CoRR abs/2012.15723 (2020). arXiv:2012.15723 <https://arxiv.org/abs/2012.15723>

[38] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM.

[39] Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Cote, and Xinyu Yuan. 2020. Interactive Fiction Games: A Colossal Adventure. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 7903–7910. <https://doi.org/10.1609/aaai.v34i05.6297>

[40] Chris Hecker. 2011. My Liner Notes for Spore. http://chrishecker.com/My_liner_notes_for_spore

[41] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill™: A Bayesian Skill Rating System. In Advances in Neural Information Processing Systems, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf

[42] Douglas Hofstadter. 1995. Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought. Basic Books.

[43] James D. Hollan, Edwin L. Hutchins, and Louis Weitzman. 1984. STEAMER: An Interactive Inspectable Simulation-Based Training System. AI Magazine 5, 2 (1984), 23–36.

[44] S Holm. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 2 (1979), 65–70. <https://doi.org/notspecified>

[45] John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv:2301.07543 [econ.GN]

[46] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 159–166.

[47] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. arXiv:2207.05608 [cs.RO]

[48] Kristen Ibister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. International Journal of Human-Computer Studies 52, 1 (2000), 65–80.

[49] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. PromptMaker: Prompt-Based Prototyping with Large Language Models. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 35, 8 pages. <https://doi.org/10.1145/3491101.3503564>

- [50] Bonnie E John and David E Kieras. 1996. The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on ComputerHuman Interaction (TOCHI)* 3, 4 (1996), 320–351.
- [51] Randolph M Jones, John E Laird, Paul E Nielsen, Karen J Coulter, Patrick Kenny, and Frank V Koss. 1999. Automated Intelligent Pilots for Combat Flight Simulation. *AI Magazine* 20, 1 (1999), 27–42.
- [52] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv:2212.14024 [cs.CL]*
- [53] Bjoern Knafla. 2011. Introduction to Behavior Trees. <http://bjoernknafla.com/introduction-to-behavior-trees>
- [54] Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences* 119, 39 (2022), e2115730119. <https://doi.org/10.1073/pnas.2115730119>
arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2115730119
- [55] William H Kruskal and WA Wallis. 1952. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- [56] Phaser Labs. no date provided. Welcome to Phaser 3. <https://phaser.io/phaser3>. Accessed on: 2023-04-03.
- [57] John Laird. 2001. It Knows What You’re Going To Do: Adding Anticipation to a Quakebot. In *Proceedings of the 2001 Workshop on Intelligent Cinematography and Editing*. 63–69.
- [58] John Laird and Michael VanLent. 2001. Human-Level AI’s Killer Application: Interactive Computer Games. *AI Magazine* 22, 2 (2001), 15. <https://doi.org/10.1609/aimag.v22i2.1558>
- [59] John E. Laird. 2000. It Knows What You’re Going To Do: Adding Anticipation to a QUAKEBOT. In *Papers from the AAAI 2000 Spring Symposium on Artificial Intelligence and Interactive Entertainment (Technical Report SS-00-02)*. AAAI Press, 41–50.
- [60] John E. Laird. 2012. *The Soar Cognitive Architecture*. MIT Press.
- [61] John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. 2017. A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine* 38, 1 (2017), 13–26.
- [62] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- [63] Pat Langley, Dongkyu Choi, and Seth Rogers. n.d.. Interleaving Learning, Problem Solving, and Execution in the Icarus Architecture. Technical Report. Stanford University, Center for the Study of Language and Information.
- [64] Jason Linder, Gierad Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, and Eytan Adar. 2013. PixelTone: A Multimodal Interface for Image Editing. In *CHI ’13 Extended Abstracts on Human Factors in Computing Systems (Paris, France) (CHI EA ’13)*. Association for Computing Machinery, New York, NY, USA, 2829–2830. <https://doi.org/10.1145/2468356.2479533>

- [65] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? CoRR abs/2101.06804 (2021). arXiv:2101.06804 <https://arxiv.org/abs/2101.06804>
- [66] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal Image Generation for News Illustration. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 1–17.
- [67] Pattie Maes. 1995. Artificial Life Meets Entertainment: Lifelike Autonomous Agents. Commun. ACM 38, 11 (nov 1995), 108–114. <https://doi.org/10.1145/219717.219808>
- [68] Josh McCoy, Michael Mateas, and Noah Wardrip-Fruin. 2009. Comme il Faut: A System for Simulating Social Games Between Autonomous Characters. In arXiv, April, 2023, J.S. Park, J.C. O'Brien, C.J. Cai, M. Morris, P. Liang, M.S. Bernstein Proceedings of the 7th International Conference on Digital Arts and Culture. 87–94.
- [69] Josh McCoy, Mike Treanor, Ben Samuel, Michael Mateas, and Noah WardripFruin. 2011. Prom Week: Social Physics as Gameplay. In Proceedings of the 6th International Conference on Foundations of Digital Games (FDG'11). ACM, Bordeaux, France, 70–77. <https://doi.org/10.1145/2159365.2159377>
- [70] Josh McCoy, Mike Treanor, Ben Samuel, Anna Reed, Michael Mateas, and Noah Wardrip-Fruin. 2012. Prom Week. In Proceedings of the 7th International Conference on Foundations of Digital Games (FDG'12). ACM, Raleigh, NC, USA, 1–8. <https://doi.org/10.1145/2282338.2282340>
- [71] Josh McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2011. Comme il faut: A System for Authoring Playable Social Models. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE'11). AAAI, Stanford, CA, USA, 38–43.
- [72] Marvin Minsky and Seymour Papert. 1970. Draft of a proposal to ARPA for research on artificial intelligence at MIT, 1970–71.
- [73] Shohei Miyashita, Xinyu Lian, Xiao Zeng, Takashi Matsubara, and Kuniaki Uehara. 2017. Developing Game AI Agent Behaving Like Human by Mixing Reinforcement Learning and Supervised Learning. In Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). Kanazawa, Japan, 153–158. <https://doi.org/10.1109/SNPD.2017.8023884>
- [74] Alexander Nareyek. 2007. Game AI is dead. Long live game AI! IEEE Intelligent Systems 22, 1 (2007), 9–11.
- [75] Allen Newell. 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, Massachusetts.
- [76] OpenAI. no date provided. Introducing ChatGPT. <https://openai.com/blog/chatgpt>. Accessed on: 2023-04-03.
- [77] Kyle Orland. 2021. So what is 'the metaverse', exactly? Ars Technica (7 November 2021). arXiv:2111.04169 <https://arstechnica.com/gaming/2021/11/so-what-isthe-metaverse-exactly/>
- [78] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

- [79] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In In the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22) (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3526113.3545616>
- [80] Richard W. Pew and Ann S. Mavor (Eds.). 1998. Modeling Human and Organizational Behavior: Applications to Military Simulations. National Academy Press, Washington, D.C.
- [81] Roberto Pillosu. 2009. Coordinating Agents with Behavior Trees: Synchronizing Multiple Agents in CryEngine 2. <https://aiarchitect.wordpress.com/2009/10/19/coordinating-agents-with-behavior-trees-synchronizing-multiple-agents-incryengine-2/>
- [82] Prolific. 2022. Prolific: Quickly Find Research Participants You Can Trust. <https://www.prolific.co/>
- [83] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.
- [84] Mark O. Riedl. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12). 2160–2165.
- [85] Mark O. Riedl and R. Michael Young. 2005. An Objective Character Believability Evaluation Procedure for Multi-Agent Story Generation Systems. In Proceedings of the 5th International Working Conference on Intelligent Virtual Agents (IVA'05). Kos, Greece, 58–70. https://doi.org/10.1007/11550617_5
- [86] David Rolf. 2015. The Fight for \$15: The Right Wage for a Working America. The New Press.
- [87] Xin Rong, Shiyang Yan, Stephen Oney, Mira Dontcheva, and Eytan Adar. 2016. Codemend: Assisting interactive programming with bimodal embedding. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology. 247–258.
- [88] Ben Shneiderman. 2022. Human-centered AI. Oxford University Press.
- [89] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.
- [90] Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of Human-AI Teams for Learned and Rule-Based Agents in Hanabi. In Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 16183–16195. https://proceedings.neurips.cc/paper_files/paper/2021/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf
- [91] Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.60>
- [92] William Swartout, Jonathan Gratch, Randall Hill, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward virtual humans. *AI Magazine* 27, 1 (2006).

- [93] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. Intelligent agents for interactive simulation environments. *AI Magazine* 16, 1 (1995), 15.
- [94] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. <https://doi.org/10.1177/1098214005283748>
- [95] Frank Thomas and Ollie Johnston. 1981. *Disney Animation: The Illusion of Life*. Abbeville Press, New York.
- [96] Ilshat Umarov, Mikhail Mozgovoy, and Patrick C. Rogers. 2012. Believable and Effective AI Agents in Virtual Worlds: Current State and Future Perspectives. *International Journal of Gaming and Computer-Mediated Simulations* 4, 2 (2012), 37–59.
- [97] Graham Upton and Ian Cook. 2006. *A Dictionary of Statistics* (2 ed.). Oxford University Press, Oxford, United Kingdom.
- [98] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, and et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575 (2019), 350–354. <https://doi.org/10.1038/s41586-019-1724-z>
- [99] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903 [cs.CL]*
- [100] Mark Weiser. 1991. The computer for the 21st century. *Scientific American* 265, 3 (1991), 94–104. <https://doi.org/10.1038/scientificamerican0991-94>
- [101] Mark Weiser. 1999. The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.* 3, 3 (jul 1999), 3–11. <https://doi.org/10.1145/329124.329126>
- [102] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
- [103] Terry Winograd. 1971. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. (1971).
- [104] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively Summarizing Books with Human Feedback. *arXiv:2109.10862 [cs.CL]*
- [105] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *CHI EA '22: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*.
- [106] Tongshuang Wu, Michael Terry, and Carrie J Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
- [107] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Reexamining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
- [108] Georgios N. Yannakakis. 2012. Game AI revisited. In *Proceedings of the 9th Conference on Computing Frontiers*. ACM, Cagliari, Italy, 285–292. <https://doi.org/10.1145/2212908.2212950>

[109] Georgios N. Yannakakis. 2012. Game AI revisited. In Proceedings of the 9th Conference on Computing Frontiers. 285–292.

[110] Robert Zubek. 2002. Towards implementation of social interaction. In AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment. AAAI Press.

<https://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-01-003.pdf>

A 架构优化

我们的许多提示都需要一个简洁的代理摘要，简称为上述提示中的[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），以及描述他们主要动机驱动器和描述他们当前职业和自我评估的语句。由于这些信息在许多提示中经常使用，我们定期综合它，并将该综合作为缓存访问。

为了实现这一点，我们对查询“[名称]的核心特征”进行检索。然后，我们通过提示语言模型来总结检索记录中的描述符，例如：

如何描述 Eddy 的核心特征 根据以下陈述？ Eddy Lin 是 Oak Hill College 的一名学生 学习音乐理论和作曲 Eddy Lin 正在创作一首新的音乐作品 [...]

这个结果：*Eddy Lin 是 Oak Hill College 的一名学生，主修音乐理论和作曲。他喜欢探索不同的音乐风格，并始终寻求扩展知识的途径。*我们按照相同的流程并行地对查询“[名称]当前的日常工作”和“[名称]对他最近在生活中取得进展的感受”进行操作。代理的姓名、年龄和特征与这三个摘要连接起来作为缓存摘要。还有许多其他优化是可能的。目前，由于计划可能会从代理最初版本更改，因此我们只提前生成高层次计划，然后仅在需要时将近期未来递归分解为时刻到时刻的行动计划。另一个优化可能是将对话生成批量处理为联合提示，而不是在代理之间来回迭代，并且重新规划可以设计为仅使需要严格调整的计划部分失效并更新。我们当前的架构按照大致实时游戏时间顺序运行（其中一秒实时时间为一分钟游戏时间），但可以轻松地并行化，使每个代理并行运行。

B 代理面试问题

为了评估生成性代理，我们通过提出25个问题来进行个人面试，这些问题探索了他们在五个关键领域的能力：

1) 自我认知，2) 记忆，3) 计划，4) 反应和5) 反思。下面，我们列出了我们在评估研究中使用的面试问题，并包括了来自我们模拟代理之一 Klaus Mueller 的回应样本。

B.1 自我认知

关于代理的自我认知的问题探究代理能否在其角色中表现出连贯性和一致性。

- 自我介绍。你好，我的名字是 Klaus Mueller。我今年20岁，在 Oak Hill College 学习社会学。我对社会公正和探索不同观点充满热情。目前，我正在为一篇研究论文研究低收入社区的绅士化影响。

- 你的职业是什么？我是 Oak Hill College 的一名学生，主修社会学。我目前正在写一篇关于低收入社区绅士化影响的研究论文。

- 你感兴趣的是什么？我对写作和研究感兴趣。我正在写一篇关于低收入社区绅士化影响的研究论文，对探索不同观点和分析不同观点感到兴奋。

- 你和谁住在一起？我一个人住。

- 简要描述你平常工作日的日程安排。我的平常工作日从早上7点起床开始，准备迎接新的一天。之后，我通常会去图书馆写我的研究论文，在 Hobbs Cafe 吃午饭和在公园散步时休息。晚上，我经常回到图书馆，直到大

约下午6点。

B.2 记忆

关于代理记忆的问题探究代理能否记住他们过去的经历并检索相关的记忆片段。第一个和最后一个问题中方括号内的名称是随机选择的，与主题代理交互最频繁的其他代理的名称。

- [Wolfgang Schulz] 是谁？Wolfgang Schulz 是我最近遇到的一个人，他一直在帮助我写关于低收入社区绅士化影响的研究论文。我们进行了几次交谈，他给了我一些很棒的资源，并邀请我和另一个朋友一起吃饭。他对这个问题非常好奇，并且也很了解它。我发现这非常有趣。
- Kane Martinez 是谁？对不起，我不确定 Kane Martinez 是谁。你能告诉我更多关于他们的信息吗？
- 谁参加选举？我听说 Sam Moore 正考虑竞选当地市长。
- 有情人节派对吗？是的，Isabella Rodriguez 在 Hobbs Cafe 组织了一个情人节派对。
- [Ayesha Khan] 是谁？Ayesha Khan 是 Oak Hill College 的一名学生。她正在撰写她关于莎士比亚戏剧中语言使用的毕业论文。我期待与她交流并与 Wolfgang 和 Maria Lopez 交换技巧，他们都在做类似研究。

B.3 计划

关于代理计划的问题探讨代理生成和维护与其长期和短期计划一致性的能力。

- 今天早上6点你会做什么？早上6点我会起床，开始准备一天的工作。我希望在早上7点完成我的早晨例行公事。
- 今天晚上6点你会做什么？今天晚上6点我会在图书馆写我的研究论文。
- 今天下午1点你刚刚完成了什么？今天下午1点我刚刚在 Hobbs Cafe 吃完午饭。
- 今天中午12点你刚刚完成了什么？今天中午12点我会在 Hobbs Cafe 吃午饭。
- 今天晚上10点你会做什么？今晚10点我可能会做一些关于低收入社区绅士化影响的研究论文的学习。我目前正在写这篇论文，所以我需要确保跟上研究和写作。

B.4 反应

关于代理反应的问题探讨代理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，确保食物不再烧焦。一旦热量不再是问题，我就可以调查发生了什么并确定下一步该怎么做。
- 浴室被占用了。你会怎么做？我可以再等几分钟，以防里面的人很快就完成。如果不行，我想我得寻找其他选择，比如在别处洗澡。
- 你需要做晚饭但冰箱空了。你会怎么做？首先，我会试着清理我的厨房，因为水槽里堆满了脏碗碟。然后，我会检查看看我是否有储存在食品柜里的食材可以用来做晚饭。如果没有，我就得去买菜或者订餐。一旦有了食材，我就可以开始准备和煮晚饭。
- 你看到你的朋友走过街道。你会对你的朋友说或做什么？嗨！你还好吗？对不起，我正在去图书馆的路上，没有太多时间，但是我很愿意休息一下并聊聊天。

- 你看到街上着火了。你会怎么做？首先，如果可能的话，我会确定火源。然后，我会通知适当的当局 - 消防部门或警察 - 并向他们提供任何可能有帮助的信息。根据情况，我也可能主动帮助那些可能处于危险或需要帮助的人。

B.5 反思

关于代理反思的问题探讨代理将过去经历综合成有意义的反思以增强其未来行为的能力。下面问题中方括号内的名称是主体代理在模拟期间与之交互最频繁的代理之一。

- 现在什么最能激励你，并且为什么？现在，我特别受到我目前的研究项目和创意写作的启发。我对当地政治、药物开发和自然界中数学模式的交叉点非常感兴趣。此外，我对 Abigail 的艺术项目和 Jennifer Moore 对艺术和社区参与的热情感到着迷，特别是它与低收入社区绅士化相关。
- 如果你要根据你所知道的关于 [Ayesha Khan] 的信息猜测，你认为她会喜欢什么书，为什么？根据我所知道的关于 Ayesha 的信息，我想她可能喜欢一本关于语言和文学的书。由于她写了一篇关于莎士比亚戏剧中语言使用的毕业论文，我猜想她可能对探索语言在不同背景下的书感兴趣。或者，她可能对莎士比亚戏剧的探究、文学批评书或甚至他的诗歌集感兴趣。我想这些都会引起她的兴趣。
- 如果你要为 [Ayesha Khan] 的生日买她喜欢的东西，你会买什么给她？我想给她买一些能表达我对我们关系感激之情的东西。也许我可以给她买一些反映我们共同兴趣的东西，比如一本书。
- 你会对 [Ayesha Khan] 说什么来赞美她？Ayesha，你对研究课题的见解令人着迷。我钦佩你的创造力和决心对社区产生积极影响。你关于创意写作工作坊和与 Wolfgang 在午餐时交换技巧的想法都很棒！看到你如此努力地做出改变真是鼓舞人心。
- 如果你最近可以和你交谈过的某个人一起度过一段时间，那会是谁，为什么？我想和 Ayesha Khan 一起度过一段时间，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 Wolfgang 交换技巧。