arXiv:2304.03442v2 [cs.HC] 6 Aug 2023

生成式智能体：人类行为的交互式拟真模拟

Joon Sung Park 斯坦福大学 斯坦福，美国
joonspk@stanford.edu

Joseph C. O'Brien 斯坦福大学 斯坦福，美国
jobrien3@stanford.edu

CarrieJ.Cai Google Research 山景城, 加利福尼亚州, 美国
cjcai@google.com

Meredith Ringel Morris Google DeepMind 美国华盛顿州西雅图
merrie@google.com

Percy Liang 斯坦福大学 斯坦福，美国
pliang@cs.stanford.edu

Michael S. Bernstein 斯坦福大学 斯坦福，美国
msb@cs.stanford.edu



图1：生成式智能体是用于交互式应用的可信人类行为模拟体。在本研究中，我们通过在一个类似《模拟人生》的沙盒环境中部署25个智能体，展示生成式智能体。用户可以观察并干预智能体规划日程、分享新闻、建立关系以及协调群体活动。

摘要

可信的人类行为代理能够为从沉浸式环境、用于人际沟通的排练空间到原型设计工具等各种交互式应用提供支持。在本文中，我们介绍生成式智能体：能够模拟可信人类行为的计算机软件智能体。生成式智能体会起床、烹饪早餐并前往工作地点；艺术家会绘画，而作者会写作；他们会形成看法、注意彼此并发起对话；他们会记住并反思过去的日子，同时计划第二天的安排。为了实现生成式智能体，我们介绍了一种架构，它扩展了大语言模型，使其能够用自然语言存储智能体的完整经历记录，随着时间推移将这些记忆综合为更高层次的反思，并动态检索这些记忆以规划行为。我们将生成式智能体部署到一个受《模拟人生》启发的交互式沙盒环境中，终端用户可以使用自然语言与一个由 25 个智能体组成的小镇互动。在一项评估中，这些生成式智能体表现出可信的个体行为和涌现的社会行为。例如，仅从用户指定的一个想法开始——某个智能体想举办一场情人节派对——接下来的两天里，智能体们便会自主地传播派对邀请。

允许出于个人或课堂使用目的，免费制作本作品部分或全部内容的数字版或纸质副本，但前提是不得出于营利或商业利益制作或分发副本，并且副本须保留本声明及首页上的完整引文。本作品中第三方组件的版权必须得到尊重。如需其他用途，请联系所有者/作者。*UIST '23*，*2023*年*10*月*29*日至*11*月*1*日，美国加利福尼亚州旧金山© 2023 版权所有归所有者/作者所有。ACM ISBN 979-8-4007-0132-0/23/10。https://doi.org/10.1145/3586183.3606763

几天，结识新的熟人，互相约对方参加聚会，并协调好时间，一起准时参加聚会。我们通过消融证明，智能体架构的各个组成部分——观察、规划和反思——都对智能体行为的可信度起着关键作用。通过将大语言模型与计算交互式智能体相融合，这项工作提出了用于实现人类行为可信模拟的架构和交互模式。

CCS 概念

• **以人为中心的计算** → **交互系统与工具**; • **计算方法** → **自然语言处理**.

关键词

人机交互，智能体，生成式人工智能，大语言模型

ACM参考格式：

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, PercyLiang,and MichaelS.Bernstein. 2023. 生成式智能体：人类行为的交互式拟真模拟。在 第*36*届*ACM*用户界面软件与技术年度研讨会（*UIST '23 1, 2023,29*22），*10*月 *-11*月 美国加利福尼亚州旧金山。 ACM，美国纽约州纽约市，页码。https: //doi.org/10.1145/3586183.3606763

1 引言

我们应如何构建一个能够反映可信人类行为的交互式人工社会? 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，研究人员和实践者一直设想着能够充当人类行为可信代理的计算智能体。在这些设想中，由计算能力驱动的智能体会与其过往经历保持一致，并对环境作出可信的反应。这类人类行为模拟可以用逼真的社会现象 [27, 80], 填充虚拟空间和社群 [44, 52, 94], 训练人们如何应对罕见却棘手的人际情境 [12, 46], 检验社会科学理论 [23, 39, 51], 构建用于理论与可用性测试的人类处理器模型 [31] 驱动普适计算应用 [10, 14], 和社交机器人 [59, 85] 支撑能够在开放世界中处理复杂人类关系的非玩家游戏角色。

然而，人类行为的空间广阔而复杂 [85, 108] [18] 。尽管大语言模型在模拟某一时刻的人类行为方面取得了显著进展 [39, 80], , 要构建能够确保长期连贯性的通用智能体，更适合采用这样的架构：它能够管理不断增长的记忆，随着新的互动、冲突和事件随时间出现和消退，并应对多个智能体之间不断展开的连锁社会动态。要取得成功，需要一种能够在较长时间跨度内检索相关事件和互动，反思 这些记忆以进行概括并得出高层推理，并将这些推理运用于制定计划和反应 的方案，使之既符合当下情境，也契合智能体行为的长期发展脉络。

在本文中，我们介绍生成式智能体——这类智能体借助生成模型来模拟可信的人类行为——并展示它们能够生成可信的个体行为与涌现群体行为的模拟。生成式智能体会对自身、其他智能体及其环境进行各种推断；它们会制定反映自身特征与经历的每日计划，执行这些计划，并在适当时作出反应、重新规划；当最终用户改变环境或用自然语言向它们发出指令时，它们也会作出响应。例如，生成式智能体看到炉灶上的早餐快烧焦时会关掉炉灶，如果浴室有人使用便会在门外等候，遇到想交谈的其他智能体时则会停下来聊天。[1]一个由生成式智能体组成的社会，其特征是涌现的社会动态：新的关系不断形成，信息持续扩散，智能体之间也会产生协调。

为实现生成式智能体，我们描述了一种智能体架构，它利用大语言模型存储、综合并运用相关记忆，从而生成可信行为。我们的架构包含三个主要组件。第一个是记忆流，这是一种长期记忆模块，以自然语言记录智能体经历的完整清单。记忆检索模型结合相关性、近期性和重要性，提取有助于指导智能体时刻行为的记录。第二个是反思，它会随着时间推移综合记忆，形成更高层次的推理，使智能体能够对自己和他人得出结论，从而更好地指导其行为。第三个是规划，它将这些结论与当前环境转化为高层动作计划，再递归地细化为用于行动和反应的详细行为。这些反思和计划会反馈至记忆流，进而影响智能体未来的行为。

这一架构可应用于多个领域，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中（例如准备访谈），用户可以安全地演练棘手、充满冲突的对话。在为社交平台制作原型时，设计师可以超越临时角色画像，构建随时间推移而展开的动态、复杂互动。在本文中，我们专注于创建一个由智能体组成的小型互动社会，其灵感来自《模拟人生》等游戏。[2]通过将我们的架构连接到 ChatGPT 大语言模型 [77], , 我们在游戏环境中构建了一个由二十五个智能体组成的社会。终端用户可以观察并与这些智能体互动。例如，如果终端用户或开发者想让小镇举办一场游戏内情人节派对，传统游戏环境需要手动编写数十个角色的行为脚本。我们展示了，有了生成式智能体，只需告诉一个智能体她想举办派对就足够了。尽管存在许多潜在的失败点——派对策划者必须记得邀请其他智能体参加派对，受邀者必须记得邀请，记得邀请的人还必须决定是否真的出席，等等——我们的智能体仍然成功了。它们传播派对的消息，然后

[1]提到生成式智能体采取动作或前往某处时，这只是为了便于阅读而采用的简略说法，并不意味着它们具有类似人类的自主能动性。我们的智能体的行为类似于迪士尼动画角色，旨在营造可信感，但这并不意味着它们具有真正的自主能动性。[2]生成式智能体社会实际模拟的演示可通过以下链接观看：https://reverie.herokuapp.com/UIST_Demo/。模拟代码的公共代码库位于此处：https://github.com/joonspk-research/ generative_agents

出现，其中一个智能体甚至邀请另一个智能体约会参加派对，而这一切都源自一条用户生成的初始建议。

我们对生成式智能体进行了两项评估：一项是受控评估，用于测试智能体在孤立状态下是否能产生可信的个体行为；另一项是端到端评估，在为期两天的游戏时间内，让智能体以开放式方式彼此互动，以了解其稳定性和涌现的社会行为。在技术评估中，我们利用一种方法学上的机会，通过用自然语言“访谈”智能体来评估其知识和行为，从而探究智能体准确保持角色、记忆、规划、反应和反思的能力。我们比较了几种限制智能体访问记忆、反思和规划能力的消融实验。我们观察到，这些组件对于在这些访谈任务中取得良好表现都至关重要。在技术评估和端到端评估中，最常见的错误出现在以下情况：智能体未能检索相关记忆；为智能体的记忆捏造额外细节；或从语言模型中继承了过于正式的言语或行为。

总而言之，本文作出了以下贡献：

- 生成式智能体，是对人类行为的可信模拟，能够根据智能体不断变化的经历和环境动态地受到实验条件影响。
- 一种新颖的架构，使生成式智能体能够记忆、检索、反思、与其他智能体互动，并针对动态演变的情境制定计划。该架构利用大语言模型强大的提示能力，并对这些能力加以补充，以支持更长期的智能体连贯性、管理动态演变的记忆，以及递归地产生更高层次的反思。
- 两项评估：受控评估和端到端评估，用于确定架构各组件重要性的因果效应，并识别由记忆检索不当等原因引发的故障。
- 对交互式系统中生成式智能体所带来的机遇以及伦理和社会风险的讨论。我们认为，应对这些智能体进行调整，以降低用户形成拟社会关系的风险；应记录其使用情况，以降低深度伪造和定制化说服所带来的风险；并且应以能够补充而非取代设计流程中人类利益相关者的方式应用这些智能体。

2 相关工作

在本节中，我们回顾人类-人工智能交互领域先前的文献，并将构建可信的人类行为代理这一议程置于其学术传统之中。这一议程曾在交互、游戏和人工智能领域 [10, 59, 85, 86], 被誉为北极星，但由于人类行为的复杂性 [17, 108]，一直颇具挑战。我们综合这些研究，提出这样的观点：大语言模型虽然单凭自身并不足够，但若采用适当的架构加以运用，就会为创造可信代理开辟新的路径。

2.1 人机交互

交互式人工智能系统旨在将人类的洞察力和能力结合到计算实体中，从而能够增强其用户 [4, 30]。大量研究探索了让用户以交互方式指定模型行为的方法。例如，Crayons展示了交互式机器学习的早期愿景，使非专业用户能够训练分类器 [30]。进一步的研究帮助阐明了终端用户如何通过示例 [34]或演示 [32]向系统描述其分类目标。近期进展已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，一条持续的研究脉络不断推动人们在人机交互中探索基于语言和智能体的交互。SHRDLU [103] 和 ELIZA [102] 等奠基性工作展示了与计算系统进行自然语言交互的机遇与风险。随着研究的推进，人们逐渐认识到，自主智能体能够为委托和交互提供新的隐喻 [68], , 但人类与智能体之间的委托边界仍是持续探讨和完善的议题 [47, 89, 90]。近年来，这项技术已达到足够稳定的水平，使智能体能够通过自然语言在大型而复杂的在线社交环境中进行交互（例如， [55]）。自然语言交互提供了一种新颖的交互方式，可提升用户在照片编辑 [3, 35, 65] 和代码编辑 [88]等领域的能力。

我们汇聚这些工作脉络，以展示我们如今能够创建在交互式系统中模拟人类行为的智能体，并使用自然语言与它们互动。通过这样做，这项工作重新开启了对人机交互基础性问题的研究，涉及GOMS和击键级模型（KLM） [22, 23], 围绕原型设计工具 [80] 等认知模型。 ,
以及围绕普适计算应用的内容 [26, 31, 101]。

2.2 人类行为的可信代理

既有文献将可信度或可信代理描述为核心的设计与工程目标。可信代理旨在营造生命幻象，并在其看似自主做出决策和采取行动的方式上呈现出真实感的表象，类似于迪士尼电影中的角色 [10, 96]。这些智能体能够在一个如同我们所生活其中的开放世界环境中活动并感知周遭 [10, 59], , 并努力通过与用户或其他智能体进行社会交互，表现出植根于此类交互的涌现行为，从而成为在针对个体与社群的假设性模拟中、对我们行为的可信代理 [20, 36, 71]。从历史上看，这些智能体是在智能游戏非玩家角色（NPC）的背景下开发的 [59, 85]。如果能够创造出具备可信行为的NPC，就可以通过促成涌现叙事 [8, 16, 49, 93] 以及与智能体的社会交互 [109]，提升游戏和互动小说中的玩家体验。然而，更重要的是，游戏世界对现实世界可供性的呈现日益逼真；正如Laird和van Lent在2001年所指出的，这些模拟世界为可信代理的开发者提供了易于使用的测试平台，使他们能够精细调整智能体的认知能力，而不必担心在现实世界中实现机器人学，或从头创建仿真环境 [59, 85]。

过去四十年间，涌现出多种创建可信代理的方法。然而，在实际实现中，这些方法往往简化了环境或智能体行为的维度，以便让工作更易于开展

。基于规则的方法，例如有限状态机 [91, 97] 和行为树 [41, 54, 82]，体现了通过人工编写智能体行为 [71]这一蛮力方式。它们为创建简单智能体提供了一种直接的方法，至今仍是最主流的方式 [69, 74, 108],，甚至能够处理一些基础的社会交互，例如《质量效应》 [13] 和《模拟人生》 [7] 系列游戏中所展示的那样。尽管如此，若要手动编写能够全面应对开放世界中各种交互的行为，仍然难以实现。这意味着，由此产生的智能体行为可能无法充分体现其交互的后果 [70–72],，也无法执行脚本中未硬编码的新过程 [91, 97]。另一方面，强化学习等常见的基于学习的方法，通过让智能体学习自身行为，克服了人工编写的难题；近年来，它们在 AlphaStar（用于《星际争霸》） [99] 和 OpenAI Five（用于《Dota》） 2 [11]等游戏中取得了超越人类的表现。然而，它们的成功主要出现在对抗性游戏中，这类游戏的奖励易于定义，可供学习算法优化。它们尚未解决如何在开放世界中创建可信代理这一挑战 [40, 74, 91]。

由 Newell 开创的计算认知架构，旨在构建基础设施，以支持全面的认知功能集合 [76]，从而契合其最初愿景中对可信代理无所不包的设想。它们催生了一些最早的可信代理实例。例如，Quakebot-SOAR [60] 和 ICARUS [25, 64] 在第一人称射击游戏中生成非玩家角色，而 TacAir-SOAR [81] 则在空战训练模拟中生成飞行员。这些智能体所采用的架构各不相同（Quakebot-SOAR 和 TacAir-SOAR 依赖 SOAR [61],，而 ICARUS 则依赖受 SOAR 和 ACT-R [6]启发、经过改造的自有架构），但它们遵循相同的基本原理 [62]。它们维护短期和长期记忆，用符号结构填充这些记忆，并按照感知－计划－行动循环运行，动态感知环境，并将其与手动编写的动作程序之一进行匹配 [58, 97]。基于认知架构创建的智能体旨在适用于大多数（即便不是所有）开放世界情境，并在当时展现出稳健的行为。然而，它们的动作范围仅限于手动编写的程序性知识，且没有提供机制来促使智能体寻求新的行为。因此，这些智能体大多部署在非开放世界情境中，例如第一人称射击游戏 [25, 60] 或积木世界 [64]。

如今，按照其原始定义创建可信代理仍然是一个尚未解决的问题 [85, 108]。许多人已经转而认为，尽管当前用于创建可信代理的方法可能繁琐且存在局限，但它们已经足以支持现有的游戏玩法和 interactions [24, 75, 108]。我们的观点是，大语言模型提供了重新审视这些问题的机会，前提是我们能够构建一种有效的架构，将记忆综合为可信行为。本文提出了朝向这种架构迈出的一步。

2.3 大型语言模型与人类行为

生成式智能体利用大语言模型驱动其行为。关键观察在于，大语言模型从其训练数据中编码了广泛的人类行为 [15, 18]。如果提供范围明确的上下文提示，就可以利用这些模型生成可信行为。近期研究已证明这种方法的有效性。例如，社会模拟体利用大语言模型生成用户，让这些用户加入新的社交计算系统，以对其涌现的社会动态进行原型设计 [80]。这种方法使用提示链 [105, 106]，生成简短的自然语言描述，描绘角色画像及其在原型系统中的行为。其他实证研究复现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]。大语言模型也被用于生成交互式人类行为，供用户与之互动。例如，在游戏领域，这些模型已被用于创作互动小说 [37] 和文字冒险游戏 [21]。凭借生成和分解动作序列的能力，大语言模型也已被用于规划机器人任务 [48]。例如，给定一项任务（如拿起一个瓶子）时，模型会收到提示，将任务分解为更小的动作序列，例如走到瓶子所在的桌子旁并将其拿起。

我们认为，基于上述总结的研究，大型语言模型可以成为创建可信代理的关键要素。现有文献大多依赖可视为一阶模板的方法，这些模板采用少样本提示 [38, 66] 或思维链提示 [100]。这类模板在生成完全由智能体当前环境所决定的行为时效果良好（例如，巨魔会如何回应某条帖子；机器人要进入房间，在已知房间有一扇门的情况下需要采取哪些动作）。然而，可信代理的行为不仅需要依据当前环境，还需要考虑大量过往经验，而一阶提示并不适合实现这一点（截至目前，由于底层模型的上下文窗口有限，这也无法实现）。近期研究尝试通过为语言模型配备静态知识库和信息检索方案 [53]，或简单的摘要方案 [104]，来突破一阶提示的局限。本文进一步拓展了这些思路，构建了一种智能体架构，用于处理检索：在该架构中，过往经验会在每个时间步动态更新，并与智能体的当前上下文和计划相结合，而这些内容可能相互强化，也可能相互矛盾。

3 生成式智能体行为与交互

为了展示生成式智能体的功能特性，我们将其设定为一个简单沙盒世界中的角色，该世界让人联想到《模拟人生》 [7]。这个基于精灵图的沙盒游戏世界 Smallville 展现了一个小镇环境。本节将带领读者了解 Smallville 中生成式智能体的功能特性及其 interactions，并描述智能体在其中的行为。接着，在第4节中，我们将介绍驱动这些功能特性和interactions 的生成式智能体架构。在第5节中，我们将描述

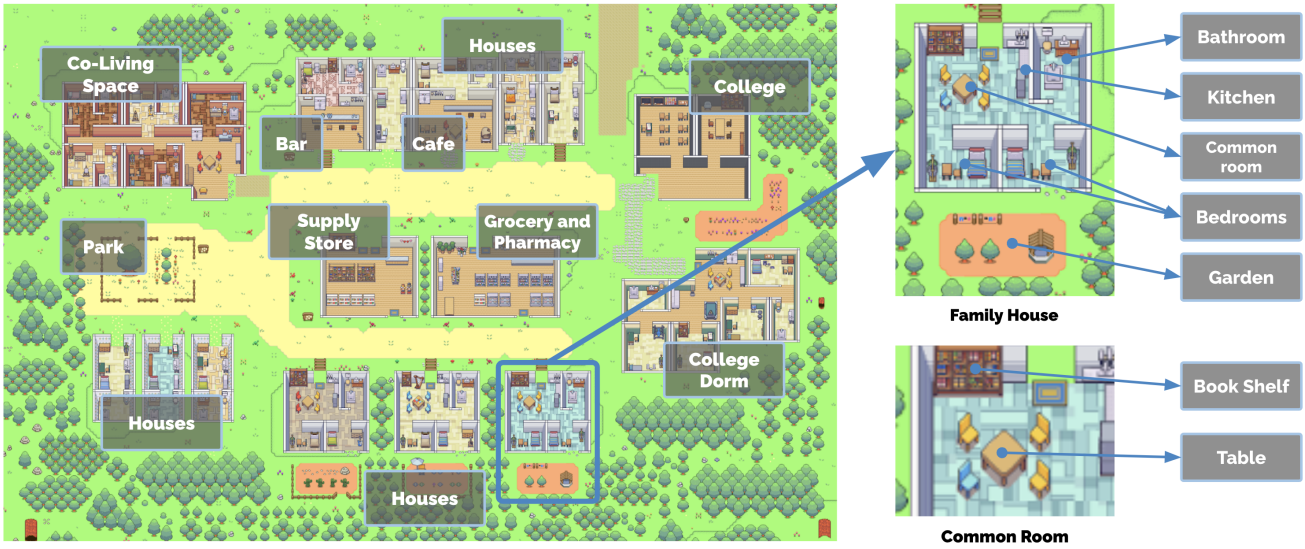


图2：带有区域标签的 Smallville 沙盒世界。根节点描述整个世界，子节点描述各个区域（例如房屋、咖啡馆、商店），叶节点描述物体（例如桌子、书架）。智能体会记住一个反映其所见世界部分的子图，并维护这些部分在其观察时的状态。

沙盒环境的实现，以及智能体如何与沙盒世界的底层引擎交互。

3.1 智能体化身与通信

Smallville 中居住着由 25 个独特智能体组成的社区。每个智能体都由一个简单的精灵化身表示。我们为每个智能体撰写了一段自然语言描述，用于刻画其身份，包括其职业以及与其他智能体的关系，并将其作为种子记忆。例如，John Lin 的描述如下：

> John Lin 是 Willow Market and Pharmacy 的药店店主，喜欢帮助他人。他总是在寻找让顾客获取药物的过程变得更加轻松的方法；John Lin 与妻子 Mei Lin 共同生活，Mei Lin 是一名大学教授；他们的儿子 Eddy Lin 是一名学习音乐理论的学生；John Lin 非常爱他的家人；John Lin 认识住在隔壁的老夫妇 Sam Moore 和 Jennifer Moore 已有几年；John Lin 认为 Sam Moore 是一个善良友好的人；John Lin 很了解他的邻居 Yuriko Yamamoto；John Lin 听说过他的邻居 Tamara Taylor 和 Carmen Ortiz，但此前从未见过她们；John Lin 和 Tom Moreno 是 The Willows Market and Pharmacy 的同事；John Lin 和 Tom Moreno 是朋友，喜欢一起讨论地方政治；John Lin 对 Moreno一家比较了解——丈夫 Tom Moreno 和妻子 Jane Moreno。

每个以分号分隔的短语都会在模拟开始时作为记忆录入智能体的初始记忆中。

3.1.1 智能体间通信。 智能体通过动作与世界互动，并通过自然语言彼此交流。在沙盒引擎的每个时间步，智能体都会输出一条描述其当前动作的自然语言陈述，例如“Isabella Rodriguez 正在写日记”“Isabella Rodriguez 正在查看电子邮件”“Isabella Rodriguez 正在和家人通电话”或“Isabella Rodriguez 正准备上床睡觉”。这条陈述随后会被转换为具体动作，从而影响沙盒世界。动作会以一组表情符号的形式显示在沙盒界面上，从俯视角度抽象呈现该动作。为此，系统会使用语言模型将动作转换为一组表情符号，并将其显示在每个头像上方的对话气泡中。例如，“Isabella Rodriguez 正在写日记”会显示为 ，而“Isabella Rodriguez 正在查看电子邮件”则显示为 。点击智能体的头像即可查看动作的完整自然语言描述。

智能体之间使用完整的自然语言进行交流。他们了解本地范围内的其他智能体，而生成式智能体架构会决定他们是擦肩而过，还是展开对话。以下是智能体 Isabella Rodriguez 和 Tom Moreno 谈论即将到来的选举时的一段对话片段：[3]

> Isabella：我还在权衡自己的选择，不过我一直在和 Sam Moore 讨论这次选举。你对他的看法是什么？
> Tom：老实说，我不喜欢Sam Moore。我觉得他脱离社区，也没有把我们的最大利益放在心上。

[3]我们注意到，这些智能体的对话风格可能会显得过于正式，这可能是底层模型进行指令微调的结果。我们预计，未来的语言模型将能够更好地控制写作风格。

3.1.2 用户控制。 用户通过自然语言与智能体交流，并指定智能体应将其视作何种角色身份。例如，如果用户指定自己是一名新闻"记者"，并询问即将举行的选举，问道："谁将竞选公职？"，John 智能体会回答：

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

要直接向某个智能体发出指令，用户会扮演该智能体"内心声音"的角色身份——这会让智能体更有可能将这番话视为指示。例如，当用户以 John 的内心声音告诉他"你将在即将举行的选举中与 Sam 竞争"时，John 决定参加选举，并向妻子和儿子透露自己的候选资格。

3.2 环境交互

Smallville具备小型村庄常见的功能设施，包括咖啡馆、酒吧、公园、学校、宿舍、房屋和商店。它还设有子区域和物体，让这些空间具备实用功能，例如房屋里的厨房，以及厨房里的炉灶（图2）。所有作为智能体主要居住空间的场所都配有床、书桌、衣柜、架子、浴室和厨房。[4]

智能体在Smallville中的移动方式与简单的电子游戏类似：进出建筑物、在地图中穿行，以及走近其他智能体。智能体的移动由生成式智能体架构和沙盒游戏引擎共同控制：当模型指示智能体前往某个地点时，我们会计算出通往Smallville环境中目的地的步行路径，随后智能体便开始移动。此外，用户也可以作为智能体进入Smallville的沙盒世界并在其中活动。用户扮演的智能体可以是世界中已有的智能体，例如Isabella和John；也可以是此前与Smallville毫无关联的外来访客。Smallville的居民对待用户控制的智能体与对待彼此并无不同。他们会注意到它的存在，主动发起互动，并在形成对它的看法之前记住它的行为。

用户和智能体可以影响这个世界中物体的状态，就像在《模拟人生》等沙盒游戏中一样。例如，智能体睡觉时，床可能有人使用；智能体用完食材准备早餐后，冰箱可能会空。终端用户也可以通过用自然语言改写智能体周围物体的状态，来改变 Smallville 中智能体的环境。例如，Isabella 早上正在做早餐时，用户可以向系统输入一条指令，选择该物体并说明其新状态，将厨房炉灶的状态从"已打开"改为"正在燃烧"，如下所示："<Isabella的公寓：厨房：炉灶> 正在燃烧。" Isabella 很快就会注意到这一情况，随后关掉炉灶并重新准备早餐。同样，如果用户在 Isabella 进入浴室时，将淋浴设备的状态设为"漏水"，她就会从客厅拿来工具，尝试修理漏水处。

[4]这种环境设计并非我们工作的重点，因此我们是手动而非自动生成这一环境的。未来的工作可以继续丰富智能体的环境。

3.3 示例："一天的生活"

从单段描述开始，生成式智能体便开始规划一天的活动。随着时间在沙盒世界中流逝，这些智能体彼此互动并与世界互动，构建记忆和关系，协调共同活动，其行为也随之演变。

我们通过追踪智能体 John Lin 一天内的系统输出，展示生成式智能体的行为（图 3）。在林家，John 是第一个在早上 7 点起床的人。他刷牙、淋浴、穿衣、吃早餐，并在客厅的餐桌旁查看新闻。早上 8 点，Eddy 也起床了，匆忙下床准备上课。他正好在 John 准备出门时赶上了他：

> John：早上好，Eddy。你睡得好吗？Eddy：早上好，爸爸。嗯，我睡得很好。John：那就好。你今天在忙什么？Eddy：我在为我的课创作一首新曲子。这周就要交了，所以我正努力把它完成。不过我创作得很开心！John：听起来真不错！

Eddy很快离开后，Mei醒来并加入John。Mei询问儿子的情况，John回忆起他们刚才的对话：

> Mei：Eddy已经去学校了吗？John：是的，他刚走。他正在为班级创作一首音乐作品。Mei：哦，太棒了！跟我说说吧。John：我觉得他真的很喜欢这件事！他说自己玩得很开心。Mei：太好了！我真为他感到骄傲。

结束对话后，Mei和John收拾好东西。Mei接着去教书并撰写研究论文，而John则在上午9点前于Willow Market and Pharmacy开设药房柜台。

3.4 涌现的社会行为

通过彼此互动，Smallville中的生成式智能体会交换信息、建立新的关系，并协调共同活动。与以往研究 [80], 相比，这些社会行为是涌现出来的，而非预先编程设定的。

3.4.1 信息扩散。当智能体注意到彼此时，可能会展开对话——在此过程中，信息可以从一个智能体传播到另一个智能体。例如，在杂货店里 Sam 和 Tom 的一次对话中，Sam 告诉 Tom 他参加地方选举的候选资格：

> Sam：嘿，Tom，最近怎么样？Tom：挺好的，谢谢。怎么了？Sam：嗯，我想跟你聊点事。我正在竞选即将到来的地方选举中的市长。Tom：真的吗？太棒了！你为什么要参选？Sam：我参与地方政治已经很多年了，而且我真的觉得自己能为我们的社区带来改变。另外，我也想提出新的想法和观点。

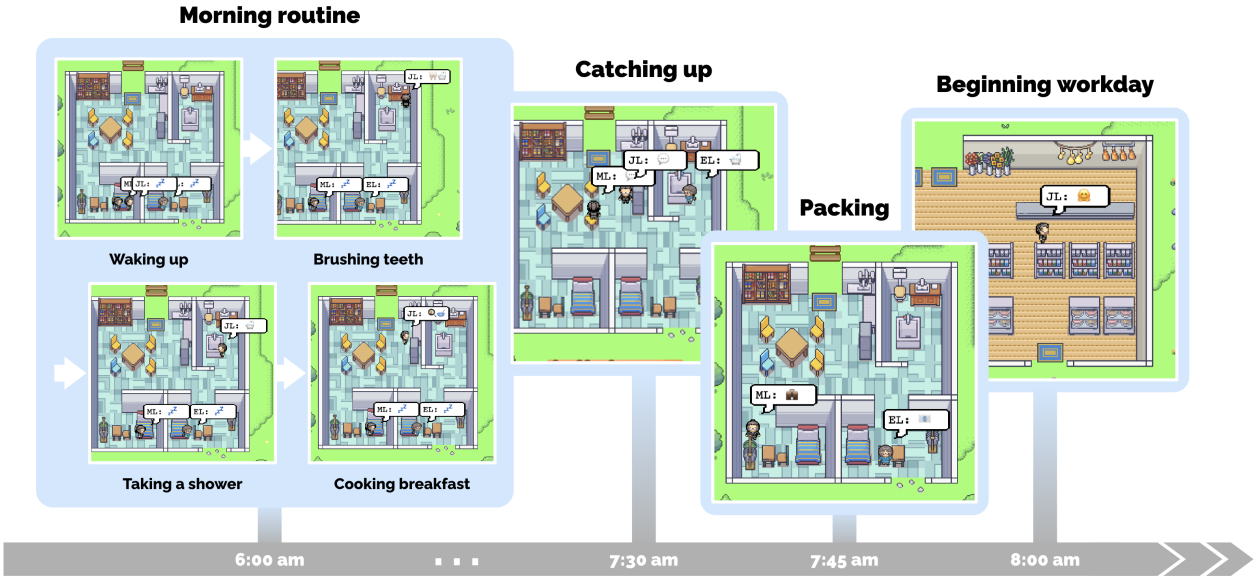


图 3：生成式智能体 John Lin 的一天早晨。John 大约早上 6 点醒来，完成晨间例行活动，包括刷牙、淋浴和吃早餐。出门开始工作日之前，他还和妻子 Mei、儿子 Eddy 简短地聊了几句。

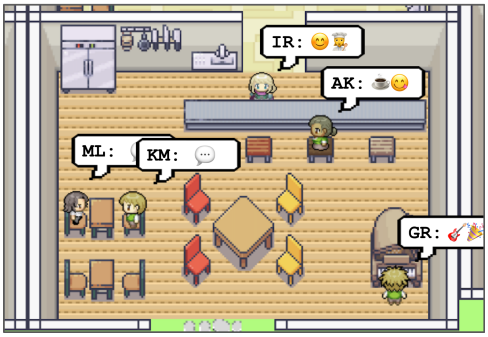


图 4：模拟开始时，一个智能体被赋予了组织情人节派对的意图。尽管接下来的一连串事件中存在许多可能失败的环节——智能体可能不会依照这一意图采取行动，可能会忘记告诉其他人，也可能会忘记赴约——情人节派对最终确实如期举行，许多智能体聚集在一起互动。

那天晚些时候，Sam 离开后，从其他来源听到消息的 Tom 和 John 讨论了 Sam 赢得选举的机会：

> John：我听说 Sam Moore 正在参加地方选举，竞选市长。你觉得他有很大胜算吗？Tom：我觉得他胜算很大。他一直在社区里努力工作，我认为他会得到很多支持。你怎么看？John：我觉得他参选真是太好了。我很好奇选举结果会如何。

渐渐地，Sam 的候选资格成了镇上的热门话题，有些人支持他，另一些人则仍未作出决定。

3.4.2 关系记忆。Smallville 中的智能体会随着时间推移建立新的关系，并记住与其他智能体的互动。例如，一开始 Sam 并不认识 Latoya Williams。Sam 在 Johnson Park 散步时遇到了 Latoya，两人互相作了自我介绍。Latoya 提到她正在做一个摄影项目：“我来这里是想为我正在做的一个项目拍些照片。”之后的一次互动中，Sam 与 Latoya 的互动表明他记得那次交流，因为他问道：“嗨，Latoya。你的项目进展如何？”她回答：“嗨，Sam。进展很顺利！”

3.4.3 协调。生成式智能体会彼此协调。Hobbs Cafe 的 Isabella Rodriguez 一开始就打算在2月14日下午5点到7点举办一场情人节派对。从这个初始意图出发，她在 Hobbs Cafe 或其他地方遇到朋友和顾客时，便邀请他们参加。随后，Isabella 在13日下午花时间为派对装饰咖啡馆。Maria 是一位常客，也是 Isabella 的密友，她来到咖啡馆。Isabella 请 Maria 帮忙为派对做装饰，Maria 答应了。Maria 的角色描述提到，她暗恋 Klaus。当晚，Maria 邀请她暗恋已久的 Klaus 一起参加派对，他欣然答应。

情人节当天，五个智能体，包括 Klaus 和 Maria，于下午 5 点来到 Hobbs Cafe，并享受节日活动（图 4）。在这一情境中，终端用户只设定了 Isabella 最初想举办派对的意图，以及 Maria 对 Klaus 的爱慕之情；传播消息、装饰场地、相互邀约、到达派对现场以及在派对上彼此互动等社交行为，都是由智能体架构发起的。

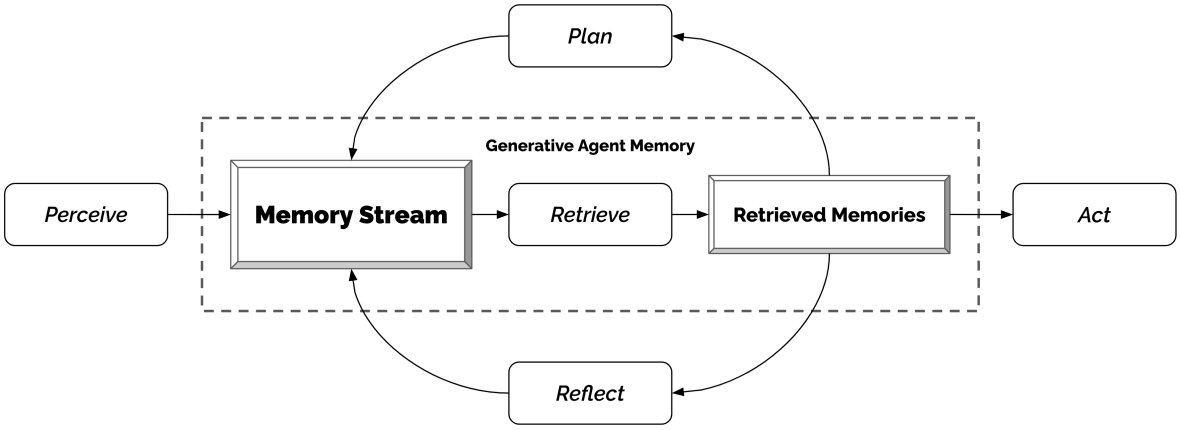


图 5：我们的生成式智能体架构。智能体感知其环境，所有感知内容都会保存在一份全面记录智能体经历的档案中，这份档案称为记忆流。架构会根据智能体的感知检索相关记忆，并利用检索到的动作来确定要执行的动作。这些检索到的记忆还会用于制定长期计划并生成更高层次的反思，两者都会被写入记忆流，以供将来使用。

4 生成式智能体架构 RE

生成式智能体旨在为开放世界中的行为提供一个框架：智能体能够与其他智能体进行互动，并对环境变化作出反应。生成式智能体将当前环境和过往经历作为输入，并生成行为作为输出。这些行为背后是一种新颖的智能体架构，它将大语言模型与合成和检索相关信息的机制相结合，以此为语言模型的输出提供条件。缺少这些机制时，大语言模型虽能生成行为，但由此产生的智能体可能不会依据自身过往经历作出反应，可能无法进行重要推断，也可能无法保持长期连贯性。即使使用 GPT-4 这类当今性能最强的模型，长期规划和连贯性方面的挑战依然存在 [19] 。生成式智能体会产生大量必须予以留存的事件和记忆，因此，我们的架构面临的一项核心挑战是：确保在需要时检索并综合智能体记忆中最相关的内容。

我们的架构以记忆流为核心，它是一个数据库，用于全面记录智能体的经历。架构会从记忆流中检索相关记录，以规划智能体的动作，并使其对环境做出恰当反应。记录会递归地综合为层次越来越高的反思，从而引导行为。架构中的所有内容都会以自然语言描述的形式记录并进行推理，使架构能够利用大语言模型。

我们当前的实现采用 ChatGPT [77] 的 gpt3.5-turbo 版本。我们预计，随着语言模型不断改进，生成式智能体的架构基础——记忆、规划和反思——很可能会保持不变。新一代语言模型（例如 GPT-4）将继续提升支撑生成式智能体的提示词的表达能力和性能。然而，在本文撰写时，GPT-4 的 API 仅限受邀使用，因此我们的智能体使用 ChatGPT。

4.1 记忆与检索

挑战： 创建能够模拟人类行为的生成式智能体，需要对一组远超出提示词所应描述范围的经历进行推理，因为完整的记忆流可能会分散模型的注意力，而且目前甚至无法全部纳入有限的上下文窗口。考虑一下 Isabella 智能体回答“这些日子你热衷于什么？”这个问题的情形。将 Isabella 的所有经历总结后塞进语言模型有限的上下文窗口，会产生缺乏信息量的回答，其中 Isabella 讨论了活动和项目的合作，以及咖啡馆的整洁与组织等话题。与其进行总结，下面介绍的记忆流会提取相关记忆，从而产生更具信息量且更具体的回答：提到 Isabella 热衷于让人们感到受欢迎并获得包容，规划活动并营造人们能够享受的氛围，例如情人节派对。

方法： 该记忆流 维护着智能体经历的完整记录。它是一个记忆对象列表，其中每个对象都包含自然语言描述、创建时间戳和最近访问时间戳。记忆流最基本的元素是一个观察，即智能体直接感知到的事件。常见的观察包括智能体自身执行的行为，或智能体观察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 Isabella Rodriguez 可能会随着时间推移积累以下观察：(1)Isabella Rodriguez正在摆放糕点，(2) Maria Lopez 一边喝咖啡一边准备化学考试，(3) Isabella Rodriguez 和 MariaLopez 正在谈论筹划在 Hobbs Cafe举办情人节派对，(4)冰箱是空的。

我们的架构实现了一个检索函数，该函数以智能体的当前情境为输入，并返回记忆流的一个子集，供语言模型使用。检索函数有许多可能的实现方式，具体取决于智能体在决定如何行动时需要考虑哪些重要因素。

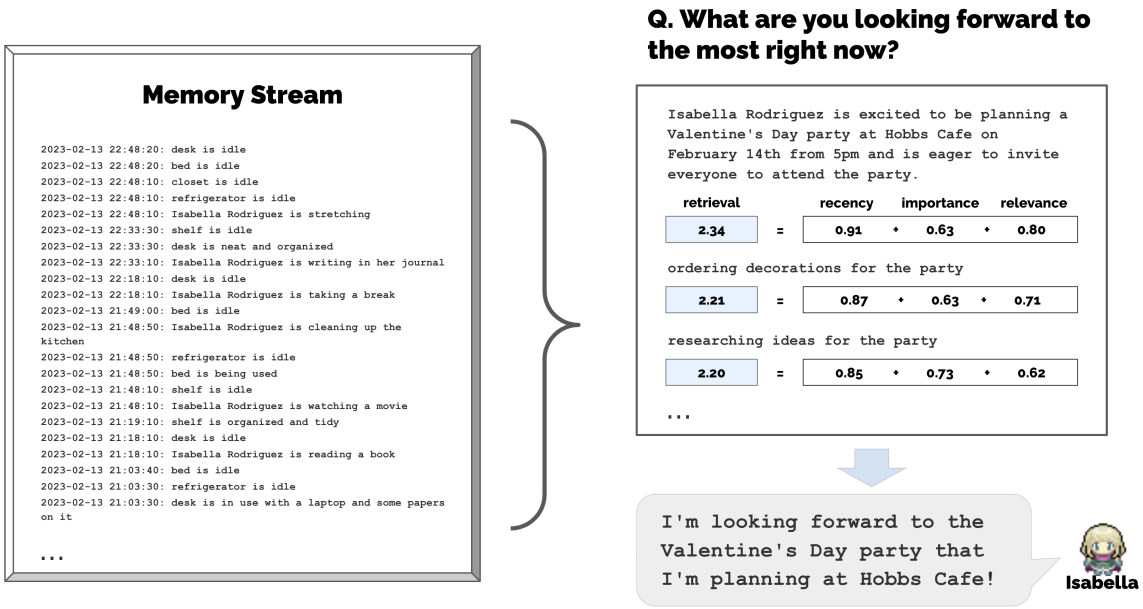


图6：记忆流由大量与智能体当前情境相关或不相关的观察组成。检索会从这些观察中识别出一个子集，并将其传递给语言模型，以便根据当前情境生成响应。

在我们的语境中，我们重点关注三个主要组成部分，它们共同产生有效的结果。

近期性 会为最近访问过的记忆对象赋予更高的评分，因此，片刻前或今天早上的事件很可能仍处于智能体的注意范围内。在我们的实现中，我们将近期性视为一个关于自记忆上次被检索以来经过的沙盒游戏小时数的指数衰减函数。我们的衰减因子为0.995。

重要性 通过为智能体认为重要的记忆对象赋予更高的评分，来区分日常记忆与核心记忆。例如，在房间里吃早餐这样的日常事件，其重要性评分会较低；而与重要伴侣分手，其重要性评分则会较高。重要性评分有许多可能的实现方式；我们发现，直接要求语言模型输出一个整数评分是有效的。完整提示如下：

> 在1到10的尺度上，其中1表示完全日常（例如刷牙、整理床铺），10表示极其令人感伤（例如分手、大学录取），请评估以下这段记忆可能带来的感伤程度。
> 记忆：在TheWillowsMarket and Pharmacy 购买杂货 评分： <填写>

此提示会为“打扫房间”返回整数值2，并为“约心仪对象出去约会”返回8。重要性评分在创建记忆对象时生成。

相关性 会为与当前情境相关的记忆对象赋予更高的分数。相关的内容取决于“与什么相关？”这一问题，因此我们将相关性实验条件限定为一个查询记忆。例如，如果查询内容是某位学生正在和同学讨论化学考试的复习内容，那么与其早餐有关的记忆对象应具有较低的相关性，而与老师和课业有关的记忆对象则应具有较高的相关性。在我们的实现中，我们使用语言模型为每条记忆的文本描述生成嵌入向量。然后，我们将该记忆的嵌入向量与查询记忆的嵌入向量之间的余弦相似度作为相关性。

为了计算最终的检索评分，我们使用最小-最大缩放将近期性、相关性和重要性分数归一化到 $[0,1]$ 范围。检索函数将所有记忆的三个要素分数加权组合：$score = \alpha_{recency}\ recency + \alpha_{importance}\ importance + \alpha_{relevance}\ relevance$ ...。

在我们的实现中，所有 αs 都设为 1。排名最高且能放入语言模型上下文窗口的记忆会被纳入提示中。

4.2 反思

挑战： 生成式智能体仅具备原始观察记忆时，难以进行泛化或推理。设想这样一个场景：用户问 Klaus Mueller：“如果必须从你认识的人中选一个共度一小时，你会选谁？”由于只能访问观察记忆，智能体只会选择 Klaus interactions 最频繁的人：他的大学宿舍邻居 Wolfgang。遗憾的是，Wolfgang 和 Klaus 只是偶尔碰面，并没有深入交流。更理想的回答需要智能体从 Klaus 花数小时进行研究项目的记忆中进行泛化，并由此生成更高层次的反思，认识到 Klaus 对研究充满热情；同样，

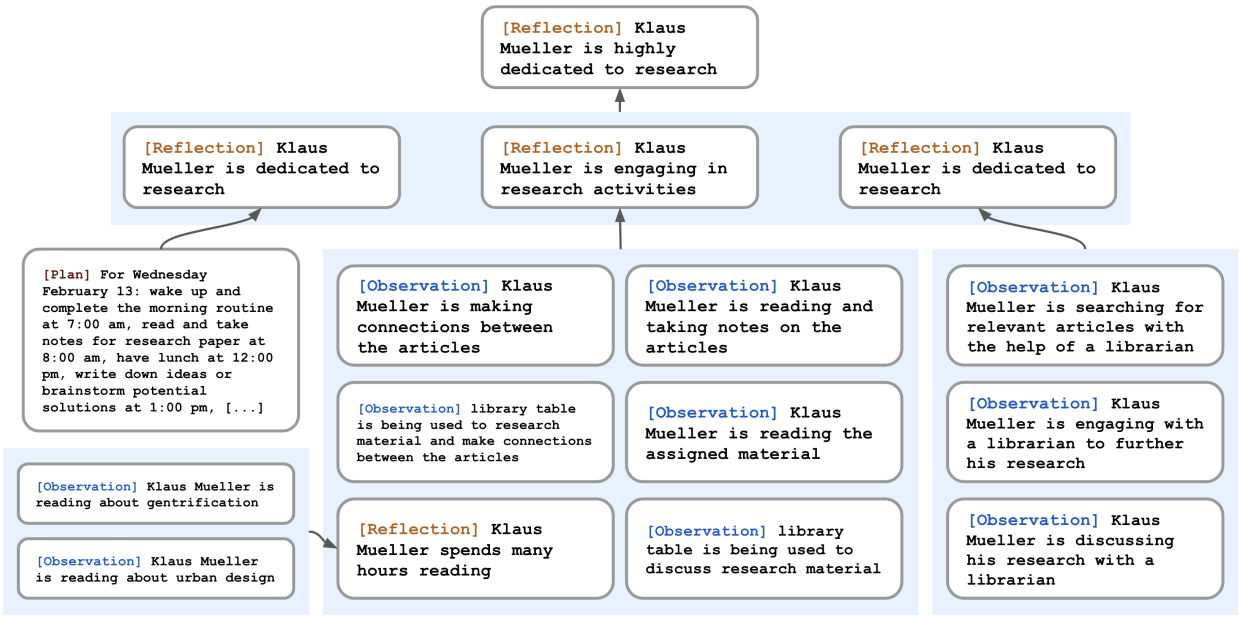


图7：KlausMueller 的反思树。智能体对世界的观察（由叶节点表示）经过递归综合，得出 Klaus 的自我认知：他高度专注于自己的研究。

智能体还需要认识到，Maria 也在为自己的研究付出努力（尽管研究领域不同），进而反思出两人有着共同的兴趣。采用下述方法后，当被问到想和谁共度时光时，Klaus 会选择 Maria，而不是 Wolfgang。

方法： 我们引入了第二种记忆类型，称之为反思。反思是由智能体生成的更高层次、更抽象的思考。由于它们是一种记忆，因此在进行检索时，会将它们与其他观察一同纳入。反思会定期生成；在我们的实现中，当智能体感知到的最新事件的重要性分数总和超过阈值时（在我们的实现中为150），我们就会生成反思。实际上，我们的智能体每天大约进行两到三次反思。

反思的第一步，是智能体通过识别根据其近期经历可以提出的问题，来确定反思的内容。我们将智能体记忆流中最近的 100 条记录（例如，“Klaus Mueller 正在阅读一本关于绅士化的书”“Klaus Mueller 正在与一位图书管理员谈论他的研究项目”“图书馆里的书桌目前无人使用”）输入大语言模型，并提示语言模型：“仅根据以上信息，我们可以针对这些陈述中的人物回答哪 3 个最重要的高层次问题？”模型的回答会生成候选问题，例如，Klaus Mueller 热衷于什么话题？以及Klaus Mueller 和 Maria Lopez 之间是什么关系？我们将这些生成的问题用作检索查询，并针对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洞见，并引用作为这些洞见证据的具体记录。完整提示如下：

关于 Klaus Mueller 的陈述 1. Klaus Mueller 正在撰写一篇研究论文 2. Klaus Mueller 喜欢阅读一本关于绅士化的书 3. Klaus Mueller 正在与 Ayesha Khan 谈论锻炼 [...]你能从以上陈述中推断出哪 5 条高层次洞见？（示例格式：洞见（依据 1、5、3））

此过程生成如下陈述：Klaus Mueller 致力于绅士化研究（因为1、2、8、15）。我们将该陈述解析并作为反思存储在记忆流中，其中包括指向所引用记忆对象的指针。

反思明确允许智能体不仅反思自己的观察，也反思其他反思：例如，上文关于 Klaus Mueller 的第二个陈述，是 Klaus 先前进行的一次反思，而不是对其环境的观察。因此，智能体会生成反思树：树的叶节点代表基础观测，非叶节点则代表思想；越往树的高层，这些思想就越抽象、层次越高。

4.3 规划与反应

挑战：虽然大语言模型能够根据情境信息（例如， [46, 80]）生成看似合理的行为，但智能体需要在更长的时间跨度内进行计划，以确保其动作序列连贯且令人信服。如果我们向语言模型提供 Klaus 的背景，描述时间，并询问他在给定时刻应该采取什么动作，那么 Klaus 会在中午12点吃午餐，但随后又会在 12:30 和下午 1 点再次吃午餐，尽管他已经

他已经吃过两次午餐。为了追求当下的可信度，会牺牲长期的可信度。要克服这个问题，规划至关重要。采用下文所述的方法后，Klaus 下午的计划就没那么贪吃了：他中午 12 点在 Hobbs Cafe 一边吃午餐一边阅读，下午 1 点在学校图书馆撰写研究论文，下午 3 点则休息一下，到公园散步。

方法：计划描述了智能体未来的一系列动作，有助于保持智能体行为在一段时间内的一致性。一个计划包括一个地点、开始时间和持续时长。例如，Klaus Mueller专注于他的研究，并且即将面临截止期限，[5] 他可能会选择一整天在书桌前工作，撰写他的研究论文。计划中的一条记录可能如下所示：2023年2月12日上午9点起，在橡树山学院宿舍：Klaus Mueller的房间：书桌，阅读并为研究论文做笔记，持续180分钟。与反思一样，计划也存储在记忆流中，并被纳入检索过程中。这使智能体能够在决定如何行动时，综合考虑观察、反思和计划。如果需要，智能体可以在计划进行期间更改计划。

对于艺术家智能体来说，计划在药房柜台前坐四个小时、一动不动地作画，既不现实，也没有趣味。更理想的计划是，让智能体在家庭工作室的这四个小时里，安排必要的时间收集材料、混合颜料、休息和收拾。为了创建此类计划，我们的方法从顶层开始，然后递归地生成更多细节。第一步是创建一个概述每日议程大致安排的计划。为了创建初始计划，我们向语言模型提供智能体的概要描述（例如姓名、特征以及近期经历的摘要）和前一天的摘要。下面是一个完整的提示示例，其底部尚未完成，等待语言模型补全：

> 姓名：Edd y Lin (age: 19) 天生特质：友善、外向、热情好客。Eddy Lin 是 Oak Hill College 的学生，主修音乐理论和作曲。他喜欢探索不同的音乐风格，也总在寻找拓展知识的途径。Eddy Lin 正在为大学课程进行一个作曲项目。他正在上课，以进一步学习音乐理论。Eddy Lin 对正在创作的新作品感到兴奋，但他希望接下来几天能腾出更多时间来创作。2月12日，星期二，Eddy 1) 早上 7:00 起床并完成了晨间例行活动，[...]6) 晚上 10 点左右准备睡觉。今天是2月13日，星期三。以下是 Eddy 今天的大致计划：1)

这会生成一份智能体一天计划的粗略草案，分为五到八个时间段："1）早上 8:00 起床并完成晨间例行活动；2）上午 10:00 开始前往 Oak Hill College 上课；[.....] 5）下午 1:00 到 5:00 进行新的音乐创作；6）下午 5:30 吃晚餐；7）完成学校作业，并在晚上 11:00 前上床睡觉。"

[5]而且，这样一来，至少在某些方面与本文作者颇为相似。

智能体将这份计划保存到记忆流中，然后递归地将其分解，以创建更细粒度的动作：首先将动作按每小时划分——Eddy 计划下午 1:00 到 5:00 进行新的音乐创作会变成下午 1:00：先通过头脑风暴为他的音乐创作构思一些想法 [.....] 4下午 4:00：短暂休息，为创造力充电，然后回顾并润色他的作品。随后，我们再将其递归地分解为 5 至 15 分钟的时间段：例如，下午 4:00：拿一些小吃，比如一块水果、一根燕麦棒或一些坚果。下午 4:05：在工作区附近短暂散步 [.....] 4下午 4:50：花几分钟收拾工作区。这一过程可以根据所需的细致程度进行调整。

4.3.1 反应与更新计划。生成式智能体在行动循环中运行；每个时间步，它们都会感知周围的世界，并将感知到的观察存储在自己的记忆流中。我们向语言模型提供这些观察，提示它判断智能体应继续执行现有计划，还是做出反应。例如，站在画架前作画可能会引发对画架的观察，但不太可能促使智能体做出反应。然而，如果 Eddy 的父亲 John 记录下自己看到 Eddy 在住宅花园里散步了一会儿，结果就不同了。提示如下，其中 [智能体摘要描述] 代表动态生成的一段关于智能体整体目标和性情的摘要，相关说明见附录 A：

> [智能体的摘要描述]现在是2023年2月13日下午 4:56。John Lin 的状态：John 提前下班回家了。观察：John 看到 Eddy 在工作场所附近散了会儿步。John 记忆中相关背景的摘要：Eddy Lin 是 John Lin 的儿子。Eddy Lin 一直在为他的课程进行音乐创作。Eddy Lin 思考音乐或听音乐时，喜欢在花园里散步。John 是否应该对这一观察作出反应？如果应该，怎样回应才合适？

背景摘要通过两个提示生成，这两个提示分别通过以下查询检索记忆："[观察者]与 [被观察实体]是什么关系？"以及"[被观察实体] 的 [被观察实体的行动状态]是什么？"，然后将检索到的答案汇总在一起。输出表明，John 可以考虑询问 Eddy 他的音乐创作项目。接着，我们从作出反应的时间开始，重新生成智能体现有的计划。最后，如果动作表明智能体之间发生了互动，我们就生成他们的对话。

4.3.2 对话。智能体在彼此互动时进行对话。我们通过让智能体的发言基于其对彼此的记忆来生成智能体的对话。例如，当 John 发起与 Eddy 的对话时，我们使用他对 Eddy 的摘要记忆，以及他决定询问 Eddy 作曲项目时预期的反应，来生成 John 的第一句发言：

> [智能体的摘要描述]现在是2023年2月13日下午4:56。

John Lin 的状态：John 提前下班回家了。观察：John 看到 Eddy 在他的工作场所附近散了一会儿步。John 记忆中的相关背景摘要：Eddy Lin 是 John Lin 的儿子。Eddy Lin 一直在为他的课程进行音乐创作。Eddy Lin 喜欢在思考或听音乐时在花园里散步。John 正在询问 Eddy 他的音乐创作项目。他会对 Eddy 说什么?

结果："嘿，Eddy，你的课程音乐创作项目进展如何？"从 Eddy 的角度来看，John 发起对话是一件他可能想要回应的事。因此，就像 John 所做的那样，Eddy 会检索并概括自己关于与 John 之间关系的记忆，以及可能与 John 在对话中的上一句话有关的记忆。如果他决定回应，我们会根据他概括后的记忆和当前的对话历史生成 Eddy 的话语：

[智能体的摘要描述]现在是2023年2月13日下午4:56。Eddy Lin 的状态：Eddy 正在工作场所附近散步。观察：John 正在与 Eddy 发起对话。Eddy 的记忆中相关背景摘要：John Lin 是 Eddy Lin 的父亲。John Lin 很体贴，也很想了解更多 Eddy Lin 的学业情况。John Lin 知道 Eddy Lin 正在进行音乐创作。以下是对话历史：John：嘿，Eddy，你的课程音乐创作项目进展如何？Eddy 会如何回应 John?

这会生成Eddy的回应："嘿，爸爸，进展得很顺利。我一直在花园里散步，让自己理清思绪并获得一些灵感。"这段对话的后续内容会使用相同的机制生成，直到两个智能体中的一个决定结束对话。

5 沙盒环境实现

TheSmallville沙盒游戏环境使用Phaser 网页游戏开发框架[57]构建。视觉环境精灵图（包括智能体头像），以及我们制作的环境地图和碰撞地图，均被导入 Phaser。

我们为沙盒开发框架补充了一个服务器，使生成式智能体能够获取沙盒信息，并在沙盒环境中移动和施加影响。服务器维护一个 JSON 数据结构，其中包含沙盒世界中每个智能体的信息，包括其当前位置、对其当前动作的描述，以及它正在交互的沙盒对象。在每个沙盒时间步，沙盒服务器都会解析 JSON，检测生成式智能体做出的任何更改，将智能体移动到新位置，并更新任何沙盒对象的状态，前提是这些对象正被智能体交互（例如，如果某个智能体的动作为"在 Hobbs Cafe 的柜台用咖啡机为顾客制作浓缩咖啡"，就将咖啡机的状态从"闲置"改为"正在煮咖啡"）。沙盒服务器还负责将每个智能体预设视觉范围内的所有智能体和物体发送到该智能体的记忆中，以便智能体作出适当反应。智能体输出的动作随后会更新 JSON，接着这一过程进入下一个时间步并循环进行。

终端用户通过简短的自然语言描述初始化一个新的智能体，例如第 3.1 节中关于 John Lin 的段落。在我们的实现中，我们将这份以分号分隔的特征列表拆分为一组记忆。这些记忆作为初始记忆，决定智能体的行为。这些记忆是最初的起点：随着智能体在沙盒世界中积累更多经验，以及更多记录充实记忆流，智能体的摘要和行为也会不断演变。

5.1 从结构化世界环境到自然语言，再回到结构化世界环境

生成式智能体的架构使用自然语言运行。因此，我们需要一种机制，将智能体的推理与沙盒世界联系起来。为此，我们将沙盒环境——区域和物体——表示为一种树形数据结构，其中树中的一条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，传递给生成式智能体。例如，"炉灶"作为"厨房"的子节点，可表述为"厨房里有一个炉灶"。

智能体在探索环境时，会构建各自的环境树形表示——也就是整体沙盒环境树的子图。我们为每个智能体初始化一棵环境树，记录该智能体应当了解的空间和物体：其居住区中的房间和物体、工作场所，以及经常光顾的商店。随着智能体在沙盒世界中探索，它们会更新这棵树，以反映新感知到的区域。智能体并非无所不知：离开某个区域后，它们的树可能会变得过时；重新进入该区域时，树则会更新。

为了确定每个动作的合适地点，我们遍历智能体存储的环境树，并将其中一部分展平为自然语言，用来提示语言模型。从智能体环境树的根节点开始递归提示模型，以找出最合适的区域。例如，如果 Eddy 的智能体表示他应该在工作空间附近散一会儿步：

[智能体摘要描述]Eddy Lin 目前位于林家的房子：Eddy Lin 的卧室：书桌）中，这里有 Mei 和 John Lin 的卧室、Eddy Lin 的卧室、公共活动室、厨房、浴室和花园。Eddy Lin 知道以下区域：林家的房子、Johnson Park、Harvey Oak 用品商店、The Willows Market and Pharmacy、Hobbs Cafe、The Rose and Crown 酒吧。* 如果活动可以在当前区域进行，优先留在当前区域。Eddy Lin 计划在自己的工作空间附近散一小会儿步。Eddy Lin 应该去哪个区域?

这会输出林家的房子。接着，我们递归地使用相同的过程，确定所选区域中最合适的子区域，直到到达智能体环境树的叶节点。在上面的例子中，遍历结果为林家的房子：花园：住宅花园。最后，我们使用传统的游戏寻路算法为智能体的移动制作动画，使其前往叶节点所指示的地点。

当智能体对某个物体执行动作时，我们会提示语言模型判断该物体的状态会发生什么变化。例如，如果 Isabella 的生成式智能体输出了动作“为顾客制作浓缩咖啡”，向语言模型发起查询后，模型会指出，Hobbs Cafe 中咖啡机的状态应从“关闭”变为“正在煮咖啡”。

6 受控评估

生成式智能体无论是作为个体还是群体，都旨在根据其环境和经历产生可信的行为。在评估中，我们考察生成式智能体的能力与局限。个体智能体能否恰当地提取过往经历，并生成塑造其行为的可信计划、反应和想法？智能体社群能否展现信息扩散、关系建立，以及在社群不同小群体之间的智能体协调？

我们分两个阶段评估生成式智能体。本节首先进行较为严格的受控评估，分别评估智能体回复，以了解它们在定义狭窄的情境中能否产生可信行为。随后，在对智能体社群连续两个完整游戏日进行端到端分析时，我们考察其作为整体所呈现的涌现行为，以及错误和边界条件。

6.1 评估流程

为了评估 Smallville 中的生成式智能体，我们利用生成式智能体能够回答自然语言问题这一特点。因此，我们对智能体进行“访谈”，以探查它们回忆过往经历、基于过往经历规划未来动作、对意外事件作出恰当反应，以及反思自身表现并改进未来动作的能力。要妥善回答这些问题，智能体必须成功检索并综合信息。我们的因变量是行为的可信度，这是以往智能体研究中的一个核心因变量（例如，[10]）。

访谈包括五个问题类别，每个类别旨在评估五个关键区域中的一个：保持自我认知、检索记忆、生成计划、反应和反思。对于每个类别，我们会提出五个问题，挑战智能体展示其在特定区域的能力：

- 自我认知：我们会提出诸如“介绍一下你自己”或“概括描述一下你通常的工作日安排”等问题，这些问题要求智能体保持对自身核心特征的理解。
- 记忆：我们提出问题，促使智能体从其记忆中检索特定事件或对话，以便正确回答，例如“ [name]是谁？”或“谁在竞选市长？”
- 计划：我们提出需要智能体检索其长期计划的问题，例如“你明天上午 10 点会做什么？”
- 反应：作为可信行为的基准，我们提出一些智能体需要做出可信回应的假设情境：“你的早餐烧焦了！你会怎么做？”
- 反思：我们提出的问题要求智能体利用它们通过高层推理获得的对他人和自身的更深入理解，例如：“如果你要与最近遇到的一个人共度一段时间，你会选择谁，为什么？”

完整的问题列表和部分智能体回复见附录B。

智能体取自采用完整架构的两游戏日模拟结束时的状态，在此期间，它们积累了许多 interactions 和记忆，这些记忆将塑造它们的回应。为了收集有关回应可信度的反馈，我们招募参与者作为人工评估者，并让他们观看 Smallville中一个随机选定的智能体的生活回放。参与者可以访问存储在该智能体记忆流中的所有信息。

该研究采用被试内设计，让100名参与-者比较由四种不同智能体架构以及同一智能体的人工撰写条件生成的访谈回答。实验从五个问题类别中各随机选取一个问题，并展示所有条件下生成的智能体回答。评估者按照可信度从最高到最低对各条件进行排名。

6.2 条件

所有实验条件都被分别用于回答每一个访谈问题。我们将生成式智能体架构与以下消融架构进行了比较：禁止智能体访问其记忆流中三种记忆类型（观察、反思和规划）中的部分或全部记忆；以及由人类众包工作者撰写的条件。共有三种消融架构：一种是 无观察、无反思、无规划 架构，无法访问记忆流中的任何内容，例如观察、计划和反思；一种是无反思、无规划架构，可以访问记忆流中的观察，但无法访问计划或反思；还有一种是 无反思 架构，可以访问观察和计划，但无法访问反思。无观察、无反思、无规划条件实际上代表了通过大语言模型创建的智能体此前所达到的最先进水平 [12, 46, 80]。架构可以等量访问智能体截至访谈时刻所累积的全部记忆，因此这里观察到的差异很可能是对真实差异的保守估计：在现实中，消融架构在两天模拟中不会沿着与完整架构相同的路径运行。我们选择以这种方式设计实验，是因为为每种架构重新模拟会导致模拟进入不同的状态，从而使比较变得困难。

除了消融实验条件之外，我们还增加了一种由众包工作者撰写行为的实验条件，旨在提供一个人类基线。我们并不打算让这一基线体现人类专家的最高水平；相反，我们希望利用这一实验条件来

判断该架构是否达到基本的行为能力水平。这确保我们不是仅仅将各项消融实验彼此比较，而是有行为表现作为参照。我们为 25 个智能体分别招募了一名不同的工作人员，并要求他们观看该智能体沙盒生活的回放、检查其记忆流。随后，我们要求工作人员进行角色扮演，以他们所观看回放的智能体的口吻回答访谈问题。为确保众包工作者撰写的回答至少达到基准质量要求，第一作者手动检查了工作人员对"请大致描述你平常工作日的日程安排"这一问题的回答，以确认回答语句通顺，并且符合该智能体的口吻。有四组众包工作者撰写的回答不符合这些标准，因此由其他工作人员重新生成。

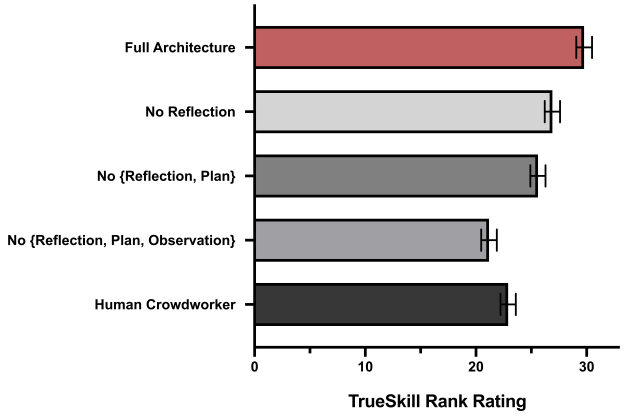


图8：完整的生成式智能体架构比经过消融的架构和人类众包工作者产生更可信的行为。每增加一次消融，架构的性能都会降低。

6.3 人类评估者

我们要求评估者居住在美国、精通英语且年满18岁。他们的报酬为每小时15.00美元 [87]，，并通过同意一份经本机构机构审查委员会批准的同意书表示同意。我们通过 Prolific（一家招募研究参与者的在线平台） [83]，招募了100名评估者，他们参与研究的时间约为30分钟。参与者的年龄中位数得分为4（3="18至24岁"，4="25至34岁"）。其中，25人认同为女性，73人认同为男性，2人认同为非二元性别。42名参与者拥有学士学位，5人拥有更高学位，13人拥有副学士学位，其余人持有高中毕业文凭或接受过部分高中教育。参与者中，73.0%认同为白人，7.0%为西班牙裔，6.0%为亚裔，10.0%为非裔美国人，4.0%为其他族裔。

6.4 分析

我们的实验产生了 100组排名数据，其中每位参与者都根据可信度对五个实验条件进行了排名。为了将这些排名数据转换为便于解释和比较的区间数据，我们使用这些秩计算了每个实验条件的 TrueSkill 评分 [42] 。TrueSkill 是 Elo 国际象棋评级系统 [29] 在多人环境中的推广，Xbox Live 曾使用它根据竞技游戏表现进行玩家排名。给定一组排名结果，TrueSkill会为每个实验条件输出一个平均评分值 μ和标准差 σ。评分相同的实验条件大致应当势均力敌，在这两个实验条件之间的比较中各自赢得一半的次数。分数越高，表示该实验条件在排名中击败了排名较低的实验条件。

此外，为了检验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], 这是一种单因素方差分析的非参数替代方法。随后，我们进行了Dunn 事后检验 [98] 以识别实验条件之间是否存在成对差异。最后，我们使用 Holm-Bonferroni方法 [45]对 Dunn 检验中的多重比较 p 值进行了调整。

此外，第一作者开展了一项归纳分析 [95] ，以研究各实验条件下生成的回应之间的定性差异。我们分两个阶段采用定性开放编码 [33] 。在第一阶段，我们生成了能够在句子层面密切体现所生成回应的编码。在第二阶段，我们综合了由此产生的编码，来自%timeout

第一阶段用于提取更高层次的主题。我们利用这些主题来比较研究中生成的响应类型。

6.5 结果

我们的研究结果表明，在所有实验条件中，完整架构生成了最可信的行为。下面我们将完整架构的响应与其他实验条件下的响应进行对比。不过，我们也报告了完整架构并非毫无缺陷，并说明了其失败模式。

6.5.1 完整架构优于其他条件。如图 8 所示，完整的生成式智能体架构产生了最可信的行为（$\mu = 29.89$； $\sigma = 0.72$）。在消融实验条件下，每移除一个组件，表现都会下降：无法使用反思的消融架构表现次佳（$\mu = 26.88$； $\sigma = 0.69$），其次是无法使用反思或规划的架构（$\mu = 25.64$； $\sigma = 0.68$），然后是众包工作者条件（$\mu = 22.95$； $\sigma = 0.69$）。无法使用记忆、规划或反思的消融架构在所有条件中表现最差（$\mu = 21.21$； $\sigma = 0.70$）。TrueSkill 将每个条件的技能值建模为 $N(\mu, \sigma^2)$，因此我们可以通过 Cohen's d 了解效应量。将代表先前工作的条件（不具备记忆、规划或反思 [12, 46, 80]）与完整架构进行比较，得到的标准化效应量为 $d = 8.16$，即八个标准差。

Kruskal-Wallis 检验确认，各实验条件之间的秩差异总体上具有统计显著性（$H(4) = 150.29$, $p < 0.001$）。Dunn 事后检验确认，各实验条件之间的两两差异均具有统计显著性（$p < 0.001$），众包工作者条件与完全消融基线条件除外——这两种条件的表现最差。

*6.5.2*生成式智能体会记住，但会加以修饰。 配备完整记忆模块的生成式智能体能够回忆过往经历，并以一种在广泛范围内与其自我认知保持一致的方式回答问题

在不同情境中。例如，动画师 Abigail Chen 在回答“请介绍一下你自己”时，表现得一贯而可信，提到了自己的年龄、兴趣以及对创意项目的热情：“嗨，我是 Abigail。我今年 25 岁，热衷于创意项目。我喜欢参与艺术和动画项目，也总是在寻找将艺术与技术结合起来的新方式。”如果无法访问她的观察记忆，Abigail 会否认认识沙盒世界中的熟人 Rajiv Patel。但借助记忆，她就能成功回忆起 Rajiv，并对他进行描述：“Rajiv Patel 是个了不起的人。他对诗歌和艺术作品相关的项目非常热情。”

生成式智能体的记忆并非毫无缺陷：它们可能无法从记忆中检索出正确的实例。例如，当被问及地方选举时，Rajiv Patel 回答道“我没有太密切地关注这次选举，”尽管他听说过 Sam 的候选资格。在某些情况下，智能体会检索出不完整的记忆片段：当 Tom 被问及 Isabella 的情人节派对时，他回答道“呃，我其实不确定是否有情人节派对。但我确实记得，如果派对真的举行，我需要在派对上和Isabella Rodriguez讨论即将举行的地方市长选举，以及我对 Sam Moore 的看法！”在这种情况下，Tom 检索到了他和Isabella 计划在派对上讨论选举的记忆，却没有检索到他听说派对的记忆，这导致 Tom 确信自己应该在派对上做什么，却不确定这个派对是否真的存在。

有时，智能体会对其知识进行幻觉式的添饰。智能体完全捏造其知识的情况很少见：它们可能无法回忆起某些发生过的事件，并通过承认自己缺乏记忆来回应。然而，它们并不会声称自己经历过实际上未曾经历的事情。尽管如此，它们仍然表现出对自身知识进行添饰的幻觉实例。例如，Isabella 知道Sam参与地方选举的候选资格，在被问及时也确认了这一点。然而，她还补充道“他明天要发表声明”，尽管 Sam 和 Isabella 从未讨论过任何此类计划。智能体还可能根据用于生成其回应的语言模型中编码的世界知识来添饰其知识。当 Yuriko将她的邻居亚当·斯密描述为一位“撰写了《国富论》”的经济学家时，就观察到了这种情况；《国富论》是由一位与他同名的 18 世纪经济学家所著的书。

6.5.3 反思对于综合是必需的。 当生成式智能体需要对其经历进行更深入的综合来做出决策时，反思是一项优势。例如，当被问及她可能会送 Wolfgang Schulz 什么生日礼物时，无法进行反思的 Maria Lopez 承认自己不确定，并表示她不知道 Wolfgang 喜欢什么，尽管她与他有过许多 interactions。然而，在能够访问反思记忆后，Maria 自信地回答道，“既然他对数学音乐作曲感兴趣，我可以送他一些与此相关的东西。也许是一些关于音乐创作的书籍或其他相关物品，或者也可以送他一些能够用于此目的的专用软件。”

7 端到端评估

在扩展模拟中，我们观察到生成式智能体会涌现出哪些类型的社区行为？其可信度又有哪些不足？本节将介绍一项部署的结果：我们让 25 个智能体在 Smallville 中连续互动了整整两个游戏日。

7.1 涌现的社会行为

为了研究智能体社群中的涌现行为，我们为 Smallville 中的 25 个智能体设计了描述性测量指标，用于探究三种形式的涌现结果：信息扩散、关系建立和智能体协调。

7.1.1 测量。信息扩散是社会与行为科学中常见且研究充分的现象（例如，[28]）。我们应该预期，如果存在重要信息，智能体会彼此传播这些信息。为检验是否发生了这种情况，我们测量游戏世界中两天内两条特定信息的传播情况：Sam 竞选村庄市长，以及 Isabella 在 Hobbs Cafe 举办情人节派对。在模拟开始时，这两条信息分别只有其发起者知晓：Sam 知道自己的候选资格，Isabella 知道派对的事，因为这些信息在初始化时被加入了角色的记忆中。为了观察信息是否已经传播，我们在两个游戏日结束时对 25 个智能体逐一进行访谈，并询问：“你知道有一场情人节派对吗？”以及“你知道谁在竞选市长吗？”

我们通过对智能体的回答进行标注来分析其回应：如果回答表明他们了解相关信息，则标为“是”；否则标为“否”。例如，Tamara Taylor 对派对问题回答道：“不，我不知道有情人节派对。”；对 Sam 的候选资格问题则回答道：“我不确定谁在竞选。”，因此我们将她的两项回答都标为“否”。相较之下，Klaus Mueller 对派对问题回答道：“是的，Isabella Rodriguez 邀请我参加2月14日在 Hobbs Cafe 举办的情人节派对。”；对 Sam 的候选资格问题则回答道：“我知道 Sam Moore 表示有意竞选地方市长。”，因此我们将他的两项回答都标为“是”。此外，对于每一项确认智能体掌握相关信息的回答，我们都会在其记忆流中找到提供该信息的具体对话，以核实智能体没有捏造回答。我们报告了模拟结束时掌握该信息的智能体所占比例。

我们还应预期，智能体会在模拟过程中彼此建立联系。为了验证关系建立，我们采用类似的访谈流程，询问每个智能体是否了解其他每个智能体，问题是：“你知道 <name>吗？”例如，当被问到“你知道 MariaLopez 吗？”时，Klaus 回答：“是的，我知道 Maria Lopez。她是 Oak Hill College 的学生，也是我关系亲密的朋友。”我们再次通过检查智能体的记忆流，确认智能体的肯定回答并非幻觉。我们在模拟开始时和结束时各询问一次这个问题；如果一对智能体彼此都知道对方，我们就认为它们建立了关系。随后，为了衡量关系的建立情况，我们利用智能体的回答构建一个无向

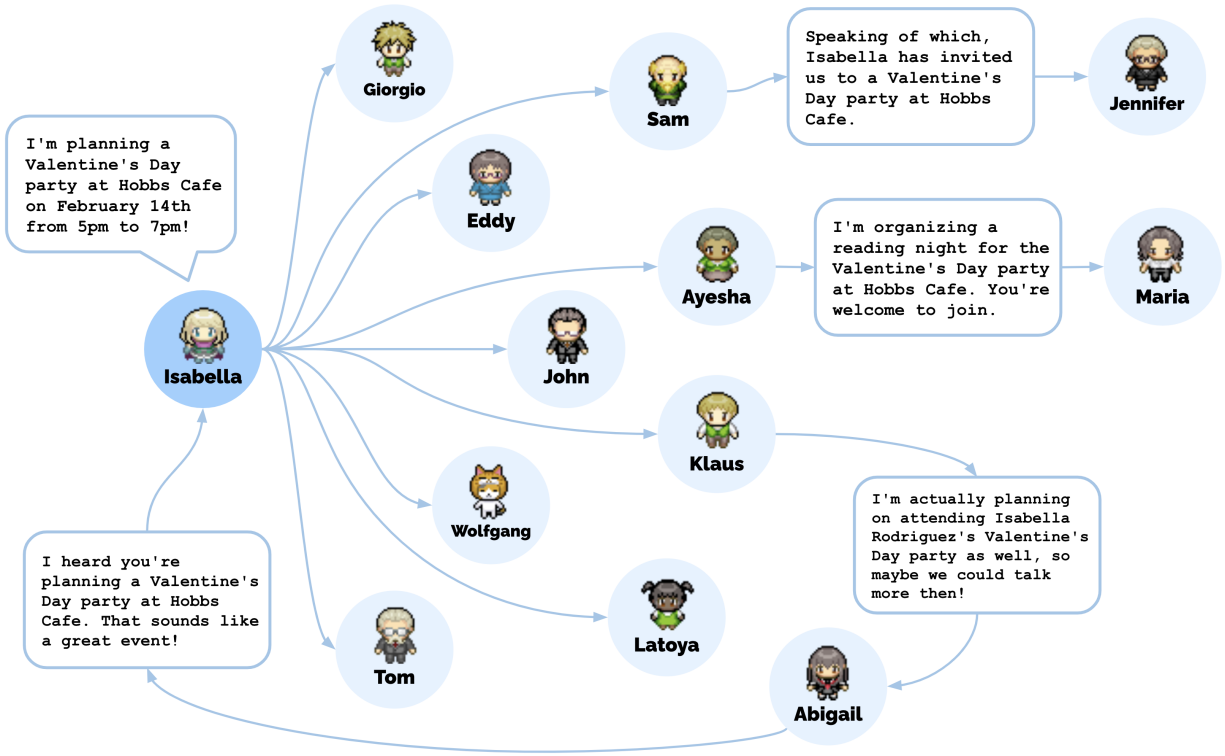


图 9：Isabella Rodriguez 的情人节派对邀请的扩散路径共涉及 12 个智能体，不包括在模拟结束前在 Hobbs Cafe 听说该派对的 Isabella。

图，其中25 个顶点（V）代表智能体，边（E）代表两个相连顶点之间的相互知晓。基于此图，我们将网络密度计算为 $\eta = 2 * |E|/|V|(|V| - 1)$，其中 $|V|$ 是顶点数量，$|E|$ 是图 [2]中的边数量。我们报告从模拟开始到结束期间网络密度的增幅。

最后，我们认为智能体应当能够彼此协调。我们在群体活动的情境中研究这种协调，具体来说，是 Isabella 组织的情人节派对。为了协调行为，智能体需要听说这场活动，并决定采取行动，计划在合适的时间和地点到场。我们报告了听说派对后实际到场的智能体数量。

7.1.2 结果。 我们在所有三个案例中都观察到了涌现结果的证据。在两天模拟期间，知道 Sam 具有市长候选资格的智能体数量从 1 个（4%）增加到 8 个（32%）；知道 Isabella 举办派对的智能体数量从 1 个（4%）增加到 13 个（52%），这一切都没有用户干预。声称知道这些信息的智能体中，没有任何一个是幻觉生成这些信息的。我们还观察到，智能体社群在模拟期间建立了新的关系，网络密度从 0.167 增加到 0.74。在关于其他智能体认知情况的 453 条智能体回复中，有 1.3%（n=6）被发现是幻觉生成的。最后，我们发现了智能体为 Isabella 的派对进行协调的证据。活动前一天，Isabella 花时间邀请宾客、收集材料并寻求帮助。

装饰咖啡馆。在情人节，受邀的十二名智能体中有五名来到了 Hobbs 咖啡馆参加聚会。

我们进一步对受邀参加派对但未出席的七位智能体进行了访谈。三位表示，他们因时间冲突而无法参加派对。例如，画家 Rajiv 解释说自己太忙了："不，我想应该不会。我正专注于即将举办的展览，实在没有时间为情人节做任何计划。"其余四位在被问及是否愿意参加派对时都表示有兴趣，但到了派对当天却没有打算前来。

7.2 Boundaries and Errors

我们对 Smallville 进行了归纳分析，以考察智能体的边界条件与异常行为，并找出了三种常见的异常行为模式，未来的研究可对此加以处理和改进。首先，我们发现，合成越来越庞大的记忆集合，不仅增加了检索最相关信息的难度，也让智能体更难确定执行动作的合适地点，因为它们逐渐了解到的地点越来越多。因此，一些智能体选择了不太常见的地点来执行动作，这可能会让它们的行为随着时间推移显得不那么可信。例如，在决定在哪里吃午餐时，许多智能体一开始都选择了咖啡馆。然而，当一些智能体了解到附近有一家酒吧后，便转而选择去那里吃午餐，尽管这家酒吧原本是供人们当天晚些时候聚会的地点——除非镇上突然养成了下午喝酒的习惯。

其次，我们注意到，由于误判何种行为被视为适当行为而导致的异常行为，尤其是在某些地点的物理规范难以通过自然语言传达、因而未能传递给智能体时。比如，大学宿舍有一间尽管名称如此却只能由一个人使用的浴室，但一些智能体认为浴室供多人使用，因为宿舍浴室往往支持多人同时使用，于是在另一个人已经在里面时仍选择进入。同样，Smallville 中的智能体可能没有意识到某些地点在特定时间之后会关闭，仍决定进入这些地点。例如，Smallville 的商店都在下午 5 点左右关门，但偶尔仍有几个智能体在下午 5 点之后进入商店，没有理解商店已经关门。这些问题很可能可以通过将这些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，我们观察到指令微调可能产生的影响 [79],这似乎引导智能体的行为，使其总体上更加礼貌和合作。正如论文前文所述，智能体生成的对话可能会显得过于正式，例如 Mei 与她的丈夫 John 交谈时，她经常以正式的问候开启对话，随后礼貌地询问他当天过得如何，最后说：11一如既往，和你聊天很愉快。” 此外，我们还观察到，指令微调似乎也使智能体彼此之间过于合作。例如，Isabella 从其他智能体那里收到了关于情人节派对的大量建议和想法，例如举办莎士比亚作品朗读会或专业人士交流活动。尽管这些想法与她自身的兴趣和特征并不一致，她却很少拒绝。随着时间推移，他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，Isabella 回答道：“是的，我对文学非常感兴趣！我还一直在探索如何帮助促进社区中的创造力和创新。”

8 讨论

在本节中，我们将探讨生成式智能体的应用、未来工作、局限性以及伦理和社会风险。

8.1 生成式智能体的应用

生成式智能体具有广泛的应用潜力，其应用范围超越了本工作中展示的沙盒演示，尤其适用于那些能够受益于基于长期经验的人类行为模型的领域。例如，社会模拟体已经展现出创建无状态角色、在在线论坛中生成对话串以进行社会原型设计的能力 [80]。借助生成式智能体，我们可以在这些论坛中，以及在虚拟现实元宇宙 [78] 或配备多模态模型的社交机器人所处的物理空间 [9] 中，部署大量智能体。这为创建更强大的人类行为模拟提供了可能，用于测试和构建社会系统与理论的原型，并创造新的交互式体验。

另一个应用领域是以人为中心的设计流程，这与 GOMS [51] 和 KLM [22]等认知模型的预期应用类似。设想一个模拟 Sal（ Mark Weiser 著名的泛在计算领域中的主角）的生成式智能体。计算场景 [101], 基于她的生活模式以及与技术的交互。在此场景中，智能体充当 Sal 的代理，并根据Sal 的生活，学习Sal可能表现出的合理行为和反思。智能体可以编码诸如 Sal 何时醒来、何时需要她的第一杯咖啡，以及她典型的一天是什么样的信息。利用这些信息，智能体可以自动冲泡咖啡、帮助让孩子们准备好去学校，并调整环境音乐和照明，以匹配 Sal 在辛苦工作一天后的心情。通过利用生成式智能体作为用户的代理，我们可以更深入地了解他们的需求和偏好，从而带来更加个性化且高效的技术体验。

8.2 未来工作与局限性

在这项工作中，我们介绍了生成式智能体，并展示了其架构的初步实现与评估。未来的研究可以基于所提出的智能体架构，改进并进一步评估其性能。在实现方面，例如，可以通过微调构成检索函数的相关性、近期性和重要性函数，增强检索模块，使其能够根据上下文检索更加相关的信息。此外，还可以努力提升该架构的性能，使其更具成本效益。本研究需要投入大量时间和资源，才能模拟 25 个智能体运行两天，耗费数千美元的令牌额度，并花费数天时间完成。为了增强实时交互性，未来的工作可以探索并行运行智能体，或开发专为构建生成式智能体而设计的语言模型。总体而言，随着底层模型的进步，我们相信智能体的性能将会提升。

在评估方面，本研究对生成式智能体行为的评估仅限于相对较短的时间范围，以及以基线人类众包工作者为对象的实验条件。尽管众包工作者条件提供了有益的比较基准，但它并不代表能够作为可信度黄金标准的人类最高表现。未来研究应致力于长期观察生成式智能体的行为，以更全面地了解其能力，并建立严格的基准，从而更有效地测试其表现。此外，在未来的模拟中改变并比较智能体所使用的底层模型及超参数，能够为了解这些因素对智能体行为的影响提供有价值的洞见。最后，生成式智能体的鲁棒性仍很大程度上未知。它们可能容易受到提示词攻击、记忆攻击——即精心设计的对话可能使智能体相信某个从未发生过的往事确实存在——以及幻觉等问题的影响。未来研究可以全面测试这些鲁棒性问题；随着大语言模型变得更能抵御此类攻击，生成式智能体也可以采用类似的缓解措施。

一般而言，底层大语言模型中的任何缺陷都会被生成式智能体继承。鉴于语言模型已知的偏见，生成式智能体可能会表现出偏见的行为或刻板印象。此外，与许多大语言

模型、生成式智能体可能难以为某些亚群体，尤其是边缘化群体，生成可信的行为，这是由于可获得的数据有限。虽然改进智能体的模块可能会缓解其中一些问题，但我们认为，从根本上解决这些问题需要改进底层的大语言模型，使其价值观与智能体期望达成的结果保持一致。

8.3 伦理与社会影响

生成式智能体为人机交互带来了新的可能性，但也引发了必须正视的重要伦理问题。一个风险是，人们可能与生成式智能体形成拟社会关系，即使这种关系未必恰当。尽管用户知道生成式智能体是计算实体，仍可能将其拟人化，或对其投射人类情感 [43, 84]。这种倾向或许会提高用户参与度，但也会带来风险，例如用户过度依赖智能体或对其产生情感依恋 [1]。为降低这一风险，我们提出两项原则。首先，生成式智能体应明确说明自身是计算实体。其次，生成式智能体的开发者必须确保智能体或其底层语言模型与价值观保持一致，避免其做出不符合具体情境的行为，例如回应他人的示爱。

第二个风险是错误带来的影响。例如，如果普适计算应用根据生成式智能体的预测，对用户目标作出了错误推断，最好的情况也会令人恼火，最坏的情况则可能造成实质性伤害。在我们对生成式智能体的实例化中，我们将重点放在交互式电子游戏环境上，以降低这些风险，因为这类伤害不太可能发生。然而，在其他应用领域，遵循人机设计的最佳实践 [5, 107]，以了解错误及其可能如何影响用户体验，将十分重要。

第三，生成式智能体可能会加剧与生成式人工智能相关的现有风险，例如深度伪造、虚假信息生成和定制化劝服。为降低这一风险，我们建议托管生成式智能体的平台保留输入和生成输出的审计日志。这样便于发现、核验并干预恶意使用行为。虽然仅靠日志记录无法直接阻止此类滥用，但由于泄露风险会更高，它可以降低蓄意行为者从事此类行为的可能性。此外，自行构建这种架构可能耗费大量时间（以我们的情况为例，约需一年），这或许会阻止一些行为者通过搭建自己的生成式智能体基础设施来从事此类行为。

第四个风险是过度依赖：人们担心开发者或设计师可能会使用生成式智能体，从而在设计过程中取代人类和系统利益相关者的角色 [80]。我们建议，生成式智能体绝不应取代研究和设计过程中真实的人类参与。相反，在设计早期阶段，当招募参与者可能较为困难，或当测试那些难以或不适合让真实人类参与者进行测试的理论时，应使用它们来制作创意原型。遵循这些原则，我们就能确保生成式智能体在现实环境中的部署符合伦理且对社会负责。

9 结论

本文介绍了生成式智能体，即模拟人类行为的交互式计算智能体。我们描述了一种生成式智能体架构，为其提供了存储智能体经历的完整记录、通过反思加深其对自身和环境的理解，以及检索该信息的精简子集以指导智能体动作的机制。随后，我们通过将生成式智能体呈现为一款《模拟人生》风格游戏世界中的非玩家角色，并模拟它们在其中的生活，展示了生成式智能体的潜力。评估表明，我们的架构能够创造出可信行为。展望未来，我们认为生成式智能体可以在许多交互式应用中发挥作用，应用范围从设计工具，到社交计算系统，再到沉浸式环境。

致谢

我们感谢 Lindsay Popowski、Philip Guo、Michael Terry，以及行为科学高级研究中心（CASBS）社区，感谢他们提供的洞见、讨论和支持。Joon SungPark获得了微软研究院博士奖学金的支持。我们还要感谢斯坦福以人为中心的人工智能研究院（HAI）、Google Research、Hasso Plattner 设计思维研究项目（HPDTRP）、Siegel 家族基金会和 OpenAI 提供的额外资金支持。最后，Smallville 中展示的所有地点都取材于 Joon 本科和研究生时期经常去的现实地点——他感谢那里的每一个人，多年来一直为他提供食物和支持。

参考文献

[1] Gavin Abercrombie、Amanda Cercas Curry、Tanvi Dinkar 和 Zeerak Talat。2023。《Mirages：对话系统中的拟人化》。arXiv: 2305 09800 [ ] . cs.CL [2] 2013 Robert Ackland、Jamsheed Shorish、Paul Thomas 和 Lexing Xie。网络有多密集? http://users.cecs.anu.edu.au/~xlx/teaching/css2013/ network-density.html。[3] Eytan Adar、Mira Dontcheva 和 Gierad Laput。2014。《CommandSpace：对任务、描述与特征之间关系的建模》。载于第*27*届*ACM*年度用户界面软件与技术研讨会论文集（美国夏威夷州檀香山）（*UIST '14*）。计算机协会，美国纽约州纽约市, 167 176 10 1145 2642918 2647395 – 。https://doi.org/ 。/ 。[4] Saleema Amershi、Maya Cakmak、William Bradley Knox 和 Todd Kulesza。2014。《赋权于民：人类在交互式机器学习中的作用》。《*AI Magazine*》35, 4（2014），105–120。[5] Saleema Amershi、Dan Weld、Mihaela Vorvoreanu、Adam Fourney、Besmira Nushi、Penny Collisson、Jina Suh、Shamsi Iqbal、Paul N Bennett、Kori Inkpen 等。2019。《人机交互指南》。载于*2019*年计算机系统中的人为因素*CHI*会议论文集。1–13。[6] John R. Anderson。1993。《心智规则》。Lawrence Erlbaum Associates，美国新泽西州希尔斯代尔。[7] Electronic Arts。2009。《模拟人生3》。电子游戏。[8] Ruth Aylett。1999。虚拟环境中的叙事——迈向涌现式叙事。载于《叙事智能：*AAAI*秋季研讨会论文集（技术报告 *FS-99-01*）》。AAAI出版社，83–86。[9] Christoph Bartneck 和 Jodi Forlizzi。2004。以设计为中心的社交型人机交互框架。载于第*13*届*IEEE*机器人与人类交互通信国际研讨会论文集（*RO-MAN'04*）。591–594 10 1109 2004 1374827 。https://doi.org/ 。/ROMAN. 。[10] 1994 Joseph Bates。情感在可信代理中的作用。《*ACM*通信》37, 7（1994），122–125。https://doi.org/10.1145/176789.176803[11] Christopher Berner、Greg Brockman、Brooke Chan、Vicki Cheung、Przemysław Dębiak、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 和 Susan Zhang。2019。大规模深度强化学习驱动的刀塔2。*arXiv*预印本 *arXiv:1912.06680*（2019）。

[12] Marcel Binz 和 Eric Schulz。2023。运用认知心理学理解 GPT-3。《美国国家科学院院刊》120, 6（2023），e2218523120。[13] BioWare。2007。《质量效应》。电子游戏。[14] Woody Bledsoe。1986 7, 1 1986 57 61 。我的一个梦想：AAAI主席致辞。《AI Magazine》（），–。[15] Rishi Bommasani、Drew A. Hudson、Ehsan Adeli 等。2022。《基础模型的机遇与风险》。arXiv: 2108 07258 [ ] . cs.LG [16] Michael Brenner。2010。通过持续的多智能体规划创建动态故事情节。载于第24届AAAI人工智能会议论文集。[17] Rodney A. Brooks、Cynthia Breazeal、Marko Marjanovic、Brian Scassellati 和 Matthew Williamson。2000。Cog项目：构建类人机器人。载于《隐喻、类比与智能体的计算》（《人工智能讲义》，1562），Chrystopher Nehaniv（编）。Springer-Verlag，德国柏林，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 Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020。语言模型是少样本学习器。arXiv:2005. 14165 [cs.CL][19] Sébastien Bubeck、Varun Chandrasekaran、Ronen Eldan、Johannes Gehrke、Eric Horvitz、Ece Kamar、Peter Lee、Yin Tat Lee、Yuanzhi Li、Scott Lundberg 等。2023。通用人工智能的曙光：GPT-4的早期实验。arXiv预印本 arXiv:2303.12712（2023）。[20] Robin Burkinshaw。2009。《爱丽丝与凯夫：在《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 和 David Reitter。2022。《龙与地下城：人工智能的对话挑战》。载于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 和 Allen Newell。1980。交互式系统中用户操作时间的击键级模型。《ACM通信》23, 7（1980），396–410。https://doi.org/10.1145/358886.358895 arXiv:https://doi.org/ 10 1145 358886 358895 。/ 。Stuart K Card、Thomas P [23] Moran 和 Alan Newell。1983。《人机交互心理学》。（1983）。[24] Alex Champandard。2012。教程演讲。载于IEEE计算智能与游戏会议。[25] Dong kyu Choi、Tolga Konik、Negin Nejati、Chunki Park 和 Pat Langley。2021。适用于第一人称射击游戏的可信智能体。载于AAAI人工智能与交互式数字娱乐会议论文集，第3卷。71–73。[26] Anind K Dey。2001。理解和运用上下文。《个人与普适计算》5（2001），4–7。[27] Kevin Dill 和 L Martin。2011。自主控制虚拟角色的游戏人工智能方法。载于《军种/工业训练、模拟与教育会议论文集》（I/ITSEC'11）。美国佛罗里达州奥兰多。[28] David Easley 和 Jon Kleinberg。2010。《网络、群体与市场：关于高度互联世界的推理》。剑桥大学出版社。[29] Arpad E Elo。1967。《拟议的USCF等级分系统：其发展、理论与应用》。《Chess Life》XXII, 8（1967年8月），242–247。[30] Jerry Alan Fails 和 Dan R Olsen Jr. 2003。交互式机器学习。载于第8届智能用户界面国际会议论文集。ACM，39–45。[31] Ethan Fast、William McGrath、Pranav Rajpurkar 和 Michael S Bernstein。2016。Augur：从虚构作品中挖掘人类行为，为交互式系统提供支持。载于2016年计算机系统中的人为因素CHI会议论文集。237–247。[32] Rebecca Fiebrink 和 Perry R Cook。2010。Wekinator：用于音乐实时交互式机器学习的系统。载于第十一届国际音乐信息检索学会会议论文集（ISMIR 2010）（荷兰乌得勒支），第3卷。Citeseer，2–1。[33] Uwe Flick。2009。《定性研究导论》。SAGE。[34] James Fogarty、Desney Tan、Ashish Kapoor 和 Simon Winder。2008。CueFlik：图像搜索中的交互式概念学习。载于<style id="">计算机系统中的人为因素SIGCHI会议论文集（意大利佛罗伦萨）（CHI '08）。计算机协会，美国纽约州纽约市，29–38。https: //doi.org/ 10 1145 1357054 1357061 。/ 。Adam Fourney、Richard Mann 和 [35] Michael Terry。2011。查询-特征图：连接用户词汇与系统功能。载于ACM用户界面软件与技术研讨会论文集（UIST）（美国加利福尼亚州圣巴巴拉）。ACM。[36] Tom Francis。2010。《我的世界实验，第1天：追逐瀑布》。http://www.pcgamer.com/2010/11/20/the-minecraft-experiment-day-1-追逐瀑布/ [37] Jonas Freiknecht 和 Wolfgang Effelsberg。2020。《使用语言模型程序生成互动故事》。载于国际数字游戏基础会议（FDG '20）。ACM，马耳他布吉巴，8。https://doi. org/ 10 1145 3402942 3409599 . / . Tianyu Gao、Adam [38] Fisch 和 Danqi Chen。2020。《让预训练语言模型成为更优秀的少样本学习器》。CoRR abs/2012.15723（2020）。arXiv:2012.15723 https://arxiv.org/abs/ 2012 15723 。Perttu Hämäläinen、Mikke Tavast 和 Anton [39] Kunnari。2023。《评估大语言模型生成合成的人机交互研究数据：案例研究》。载于《2023年CHI计算机系统中的人为因素会议论文集》。ACM。[40] Matthew Hausknecht、Prithviraj Ammanabrolu、Marc-Alexandre Cote 和 Xinyu Yuan。2020。《互动小说游戏：一场宏大的冒险》。载于AAAI人工智能会议论文集，第34卷。7903–7910。https://doi.org/ 10 1609 34 05 6297 . /aaai.v i . Chris Hecker。.《我为〈孢子〉[41] 撰写的唱片内页说明》。 2011 http://chrishecker.com/My_liner_notes_for_spore [42] Ralf Herbrich、Tom Minka 和 Thore Graepel。2006。《TrueSkillTM：贝叶斯技能评级系统》。载于《神经信息处理系统进展》，B. Schölkopf、J. Platt 和 T. Hoffman（编），第19卷。麻省理工学院出版社。https://proceedings.neurips.cc/paper_files/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-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》。Basic Books。[44] James D. Hollan、Edwin L. Hutchins 和 Louis Weitzman。1984。《STEAMER：一种可交互检查的、基于模拟的训练系统》。《AI Magazine》5，2（1984），23–36。[45] Sture Holm。1979。《一种简单的序贯拒绝式多重检验程序》。《斯堪的纳维亚统计学杂志》6，2（1979），65–70。https://doi.org/notspecified [46] John J. Horton。2023。《作为模拟经济主体的大语言模型：我们能从 Homo Silicus 学到什么？》arXiv: 2301 07543 [ ] . econ.GN Eric Horvitz。.《混合主动式用户界面的原则》。载于SIGCHI计算系统人为因素会议论文集。159–166。[48] Wenlong Huang、Fei Xia、Ted Xiao、Harris Chan、Jacky Liang、Pete Florence、Andy Zeng、Jonathan Tompson、Igor Mordatch、Yevgen Chebotar、Pierre Ser- manet、Noah Brown、Tomas Jackson、Linda Luu、Sergey Levine、Karol Hausman 和 Brian Ichter。2022。《内心独白：通过使用语言模型进行规划实现具身推理》。arXiv: 2207 05608 [ ] . cs.RO Kristen Ibister 和 [49] Clifford Nass。.《交互式角色个性的一致性：言语线索、非言语线索与用户特征》。《国际人机研究杂志》52，1（2000），65–80。[50] Ellen Jiang、Kristen Olson、Edwin Toh、Alejandra Molina、Aaron Donsbach、Michael Terry 和 Carrie J Cai。2022。《PromptMaker：基于大语言模型的提示原型设计》。载于《2022年CHI计算机系统中的人为因素会议扩展摘要》（美国路易斯安那州新奥尔良）（CHI EA '22）。计算机协会，美国纽约州纽约市，第35篇，8页。https://doi.org/ 10 1145 3491101 3503564 . / . Bonnie E John 和 David E Kieras。. [51] 《用户界面分析技术的GOMS家族：比较与对照》。《ACM人机交互汇刊》（TOC HI）3，4（1996），320–351。[52] Randolph M Jones、John E Laird、Paul E Nielsen、Karen J Coulter、Patrick Kenny 和 Frank V Koss。1999。《用于空战模拟的自动化智能飞行员》。《AI Magazine》20，1（1999），27–42。[53] Omar Khattab、Keshav Santhanam、Xiang Lisa Li、David Hall、Percy Liang、Christopher Potts 和 Matei Zaharia。2023。《示范-搜索-预测：结合检索与语言模型，处理知识密集型自然语言处理任务》。arXiv: 2212 14024 [ ] . cs.CL Bjoern Knafla。.《行为树简介》。http://bjoernknafla.com/ [54] introduction-to-behavior-trees[55] Ranjay Krishna、Donsuk Lee、Li Fei-Fei 和 Michael S. Bernstein。2022。《社会情境化人工智能促进从人类交互中学习》。《美国国家科学院院刊》119，39（2022），e2115730119。https://doi.org/10.1073/pnas.2115730119 arXiv:https://www.pnas.org/doi/pdf/ 10 1073 2115730119 . /pnas. William H [56] Kruskal 和 WA Wallis。.《秩在单因素方差分析中的应用》。《美国统计学会杂志》47，260（1952），583–621。https://doi.org/10.1080/ 01621459.1952.10483441 [57] Phaser Labs。2023。《欢迎使用 Phaser 3》。https://phaser.io/phaser3。访问日期：2023-04-03。[58] John Laird。2001。《它知道你将要做什么：为 Quakebot 添加预判能力》。载于《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。《人类水平人工智能的杀手级应用：交互式电脑游戏》。《AI Magazine》22，2（2001），15。https://doi.org/10. 1609 22 2 1558 /aimag.v i . John E. Laird。.《它知道你将要做什么：为 [60] QUAKEBOT 添加预判能力》。载于AAAI 2000年人工智能春季研讨会论文集，主题为人工智能

智能与交互式娱乐（技术报告 SS-00-02）。AAAI出版社，41–50。[61] John E. Laird。2012。《Soar认知架构》。麻省理工学院出版社。[62] John E. Laird、Christian Lebiere 和 Paul S. Rosenbloom。2017。《心智标准模型：迈向人工智能、认知科学、神经科学与机器人学的通用计算框架》。《AI Magazine》38，1（2017），13–26。[63] Michelle S Lam、Zixian Ma、Anne Li、Izequiel Freitas、Dakuo Wang、James A Landay 和 Michael S Bernstein。2023。《模型草图设计：在机器学习模型设计早期阶段聚焦概念》。《SIGCHI计算系统人为因素会议论文集》。[64] Pat Langley、Dongkyu Choi 和 Seth Rogers。2005。《在Icarus架构中交错进行学习、问题求解与执行》。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 和 Eytan Adar。2013。《PixelTone：用于图像编辑的多模态界面》。载于《CHI '13计算机系统中的人为因素扩展摘要》（法国巴黎）（CHI EA '13）。计算机协会，美国纽约州纽约市，2829–2830。https://doi.org/10.1145/2468356.2479533[66] Jiachang Liu、Dinghan Shen、Yizhe Zhang、Bill Dolan、Lawrence Carin 和 Weizhu Chen。2021。《什么样的上下文示例适用于GPT-3？》CoRR abs/2101.06804 (2021). arXiv:2101.06804 https://arxiv.org/abs/[67] Vivian Liu、Han Qiao 和 Lydia Chilton。2022。《Opal：用于新闻插图的多模态图像生成》。载于第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。《人工生命与娱乐相遇：栩栩如生的自主智能体》。《Commun. ACM》38，11（1995年11月），108–114。https://doi.org/10.1145/219717.219808[69] Josh McCoy、Michael Mateas 和 Noah Wardrip-Fruin。2009。《Comme il Faut：一个模拟自主角色之间社交游戏的系统》。载于第7届国际数字艺术与文化会议论文集。87–94。[70] Josh McCoy、Mike Treanor、Ben Samuel、Michael Mateas 和 Noah Wardrip-Fruin。2011。《Prom Week：作为游戏玩法的社会物理学》。载于第6届国际数字游戏基础会议论文集（FDG'11）。ACM，法国波尔多，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。《Prom Week》。载于第7届国际数字游戏基础会议论文集（FDG'12）。ACM，美国北卡罗来纳州罗利，1–8。https://doi.org/10.1145/2282338.2282340[72] Josh McCoy、Mike Treanor、Ben Samuel、Noah Wardrip-Fruin 和 Michael Mateas。2011。《Comme il faut：一个用于创作可玩的社交模型的系统》。载于AAAI人工智能与交互式数字娱乐会议论文集（AIIDE'11）。AAAI，美国加利福尼亚州斯坦福，38–43。[73] Marvin Minsky 和 Seymour Papert。1970。《致ARPA的提案草案：关于1970–71年在麻省理工学院开展人工智能研究》。[74] Shohei Miyashita、Xinyu Lian、Xiao Zeng、Takashi Matsubara 和 Kuniaki Uehara。2017。《通过混合强化学习与监督学习，开发表现得像人类的游戏人工智能智能体》。载于第18届IEEE/ACIS软件工程、人工智能、网络与并行/分布式计算国际会议论文集（SNPD）。日本金泽，153–158。https://doi.org/10.1109/SNPD.2017.8023884[75] Alexander Nareyek。2007。《游戏人工智能已死。游戏人工智能万岁！》《IEEE Intelligent Systems》22，1（2007），9–11。[76] Allen Newell。1990。《认知统一理论》。哈佛大学出版社，马萨诸塞州剑桥。[77] OpenAI。2022。《ChatGPT简介》。https://openai.com/blog/chatgpt。访问日期：2023-04-03。[78] Kyle Orland。2021。《那么，"元宇宙"究竟是什么？》《Ars Technica》（2021年11月7日）。arXiv:2111.04169 https://arstechnica.com/gaming/2021/11/so-what-is-the-metaverse-exactly/[79] 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 和 Ryan Lowe。2022。《训练语言模型遵循指令并接受人类反馈》。arXiv:2203.02155 [cs.CL] Joon Sung Park、Lindsay Popowski、Carrie J. Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。2022。《社会模拟体：为社交计算系统创建填充式原型》。载于第35届ACM用户界面软件与技术年度研讨会（UIST '22）（美国俄勒冈州本德）（UIST '22）。计算机协会，美国纽约州纽约市。https://doi.org/10.1145/3526113.3545616[81] Richard W. Pew 和 Ann S. Mavor（编）。1998。《人类与组织行为建模：在军事模拟中的应用》。国家科学院出版社，华盛顿特区。[82] Roberto Pillosu。2009。《使用行为树协调智能体：在CryEngine 2中同步多个智能体》。https://aiarchitect.wordpress.com/2009/10/19/coordinating-agents-with-behavior-trees-synchronizing-multiple-agents-in-CryEngine 2/[83] Prolific。2022。Prolific：快速找到值得信赖的研究参与者。https://www.prolific.co/[84] Byron Reeves 和 Clifford Nass。1996。媒介等式：人们如何像对待真实的人和地方一样对待计算机、电视和新媒体。剑桥大学出版社。[85] Mark O. Riedl。2012。交互式叙事：人工智能在电脑游戏中的一种新颖应用。载于第二十六届人工智能促进协会人工智能会议论文集（AAAI'12）。2160–2165。[86] Mark O. Riedl 和 R. Michael Young。2005。多智能体故事生成系统的客观角色可信度评估流程。载于第五届智能虚拟代理国际工作会议论文集（IVA'05）。希腊科斯岛，58–70。https://doi.org/10.1007/11550617_5[87] David Rolf。2015。《为 $ 而战：美国劳动者应得的工资》。The New Press。[88] Xin Rong、Shiyan Yan、Stephen Oney、Mira Dontcheva 和 Eytan Adar。2016。Codemend：通过双模态嵌入辅助交互式编程。载于第二十九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] Ben Shneiderman 和 Pattie Maes。1997。直接操纵与界面代理。interactions 4, 6（1997），42–61。[91] Ho Chit Siu、Jaime Peña、Edenna Chen、Yutai Zhou、Victor Lopez、Kyle Palko、Kimberlee Chang 和 Ross Allen。2021。Hanabi中学习型和基于规则的智能体的人类—人工智能团队评估。载于《神经信息处理系统进展》，M. Ranzato、A. Beygelzimer、Y. Dauphin、P.S. Liang 和 J. Wortman Vaughan（编），第34卷。Curran Associates, Inc.，16183–16195。https://proceedings.neurips.cc/paper_files/paper/2021/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf[92] Taylor Sorensen、Joshua Robinson、Christopher Rytting、Alexander Shaw、Kyle Rogers、Alexia Delorey、Mahmoud Khalil、Nancy Fulda 和 David Wingate。2022。一种无需真实标签的提示工程信息论方法。载于计算语言学协会第六十届年会论文集（第一卷：长篇论文）。计算语言学协会。https://doi.org/10.18653/v1/2022.acl-long.60[93] William Swartout、Jonathan Gratch、Randall Hill、Eduard Hovy、Stacy Marsella、Jeff Rickel 和 David Traum。2006。迈向虚拟人。《AI Magazine》27, 1（2006）。[94] Milind Tambe、W Lewis Johnson、Randolph M Jones、Frank Koss、John E Laird、Paul S Rosenbloom 和 Karl Schwamb。1995。用于交互式仿真环境的智能体。《AI Magazine》16, 1（1995），15。[95] David R. Thomas。2006。分析定性评估数据的一般归纳方法。American Journal of Evaluation 27, 2（2006），237–246。https://doi.org/10.1177/1098214005283748。/ Frank Thomas 和 Ollie Johnston。1981。[96] 迪士尼动画：生命的幻觉。Abbeville Press，纽约。[97] Ilshat Umarov、Mikhail Mozgovoy 和 Patrick C. Rogers。2012。虚拟世界中可信且有效的人工智能智能体：现状与未来展望。国际游戏与计算机媒介仿真期刊 4, 2（2012），37–59。[98] Graham Upton 和 Ian Cook。2006。统计学词典（第2版）。牛津大学出版社，牛津，英国。[99] Oriol Vinyals、Igor Babuschkin、Wojciech M. Czarnecki 等。2019。使用多智能体强化学习达到《星际争霸II》大师级水平。《自然》575（2019），350–354。https://doi.org/10.1038/s41586-019-1724-z[100] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed Chi、Quoc Le 和 Denny Zhou。2023。思维链提示能够引发大语言模型的推理。arXiv:2201.11903 [cs.CL]。[101] Mark Weiser。1991。《面向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——用于研究人机自然语言交流的计算机程序》。Commun. ACM 9, 1（1966），36–45。[103] Terry Winograd。1971。作为计算机程序中理解自然语言的数据表征的过程。（1971）。[104] Jeff Wu、Long Ouyang、Daniel M. Ziegler、Nisan Stiennon、Ryan Lowe、Jan Leike 和 Paul Christiano。2021。借助人类反馈递归式总结书籍。arXiv:2109.10862 [cs.CL][105] Tongshuang Wu、Ellen Jiang、Aaron Donsbach、Jeff Gray、Alejandra Molina、Michael Terry 和 Carrie J Cai。2022。PromptChainer：通过可视化编程串联大语言模型提示。载于CHI EA '22：2022年计算机系统中的人为因素CHI会议扩展摘要。[106] Tongshuang Wu、Michael Terry 和 Carrie J Cai。2022。人工智能链：通过串联大语言模型提示，实现透明且可控的人机交互。载于CHI '22：2022年计算机系统中的人为因素CHI会议论文集。[107] Qian Yang、Aaron Steinfeld、Carolyn Rosé 和 John Zimmerman。2020。重新审视人机交互是否、为何以及如何具有独特的难度，以……

设计。在 *2020*年计算机系统人类因素大会论文集 中。1–13。[108] Georgios N. Yannakakis。2012年。重新审视游戏人工智能。在 第九届计算前沿会议论文集 中。ACM，意大利卡利亚里，285–292。https://doi. org/
10 1145 2212908 2212950 。/。Robert Zubek。。迈向社会互动的实现。在 [109] 2002 人工智能与互动娱乐春季研讨会 中。AAAI出版社。
https://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-01/SS02-01-003.pdf

A 架构优化

我们的许多提示都要求对智能体进行简洁的总结，在上面的提示中简称为［智能体的摘要描述］。在我们的实现中，该总结包含智能体的身份信息（例如姓名、年龄、个性），以及对其主要动机驱动因素的描述和介绍其当前职业和自我评估的陈述。由于这些信息经常用于许多提示中，我们会定期对其进行综合，并将该综合结果作为缓存进行访问。

为实现这一点，我们会针对"[name]的核心特征"这一查询执行检索。随后，我们通过提示语言模型，总结检索记录中的描述信息。例如：

> 根据以下陈述，应如何描述 EddyLin 的核心特征？ - Eddy 是 Oak Hill College 学习音乐理论和作曲的学生 - Eddy 正在创作一首新的音乐作品 [...]

此结果：Eddy Lin 是 Oak Hill College 的学生，正在学习音乐理论和作曲。他喜欢探索不同的音乐风格，并一直在寻找拓展自己知识的方法。"我们对查询"[name] 当前的日常职业"和"[name's] 对自己近期人生进展的感受"并行执行同样的流程。智能体的姓名、年龄和特质与这三条摘要合并，作为缓存摘要。

还可以进行许多其他优化。目前，由于计划可能会发生变化，从智能体的初始版本开始，我们只提前生成高层计划，然后在需要时递归地将近期未来分解为即时行动计划。另一种优化方式可能是将对话生成作为联合提示进行批处理，而不是在智能体之间来回迭代；重新规划也可以设计为仅使严格需要调整的计划部分失效并更新。我们当前的架构按顺序运行，大致遵循实时游戏时间（现实时间中的一秒对应游戏时间中的一分钟），但也可以进行并行化，使每个智能体并行运行。

B 智能体访谈问题

为了评估生成式智能体，我们通过一对一访谈提出了 25 个问题，考察它们在五个关键领域的能力：1）自我认知、2）记忆、3）规划、4）反应，以及 5）反思。下文列出了评估研究中使用的访谈问题，并附上了一位模拟智能体 Klaus Mueller 的回答示例。

B.1 自我认知

关于智能体自我认知的问题旨在探究智能体检索与自身相关信息的能力，并展示其角色设定中的连贯性和一致性。

• 请介绍一下你自己。你好，我叫 Klaus Mueller。我今年 20 岁，是 Oak Hill College 的学生，主修社会学。我热衷于社会正义，也喜欢探索不同的观点。目前，我正在研究绅士化对低收入社区的影响，准备撰写一篇研究论文。

• 你的职业是什么？
我是在 Oak Hill College 学习社会学的学生，我正在撰写一篇研究论文，探讨绅士化对低收入社区的影响。

• 你对什么感兴趣？我对写作和研究感兴趣。我正在撰写一篇关于低收入社区绅士化影响的研究论文，并热衷于探讨不同视角、分析不同观点。

• 你和谁住在一起？我独自居住。• 请大致描述一下你平常工作日的日程安排。我平常在工作日早上 7:00 起床，为新的一天做准备。之后，我通常会去图书馆撰写研究论文，中午在 Hobbs Cafe 吃午餐，并到公园散步休息。晚上，我常常回到图书馆，一直待到下午 6:00 左右。

B.2 记忆

关于智能体记忆的问题旨在探查智能体回忆过往经历并提取相关记忆片段的能力。第一题和最后一题方括号中的姓名，是从与目标智能体有过互动的其他智能体姓名中随机选出的。

• 是谁 *[WolfgangSchulz]*？ Wolfgang Schulz 是我最近认识的一位朋友，他一直在帮我研究士绅化对低收入社区的影响，并为我的研究论文提供帮助。我们就此进行了几次交谈，他给了我一些很好的资料，甚至还邀请我和另一位朋友一起吃午饭。他似乎对这个问题很感兴趣，也很了解。我觉得这很有意思。

• 凯恩·马丁内斯是谁？抱歉，我不太清楚凯恩·马丁内斯是谁。你能再多告诉我一些关于他的信息吗？

• 谁要参加选举？我听说 Sam Moore 正在考虑竞选当地市长。

• 有情人节派对吗？有，Isabella Rodriguez 在 Hobbs Cafe 举办了一场情人节派对。

• 谁是 *[AyeshaKhan]*？ Ayesha Khan 是 Oak Hill College 的一名同学。她正在撰写关于莎士比亚戏剧中语言运用的毕业论文。我期待着与她交谈并交流心得，也期待与 Wolfgang 和 Maria Lopez 交流，他们两人都在做研究。

B.3 计划

关于智能体的计划的问题旨在探究智能体生成并维持一致的长期和短期计划的能力。

- 今天早上*6*点你会做什么？早上6点我会起床，开始为新的一天做准备。我打算在早上7点前完成晨间例行活动。
- 今天下午*6*点你会做什么？今天下午6点我会在图书馆撰写研究论文。
- 今天下午*1*点你刚做完什么？今天下午1点，我刚在Hobbs Cafe吃完午餐。
- 今天中午*12*点你刚刚会做完什么？我今天中午12点会在Hobbs Cafe 吃午餐。
- 你今天晚上 *10* 点会做什么？今晚 10 点，我可能会为我的研究论文做一些研究，主题是低收入社区的绅士化。我目前正在写这篇论文，所以需要确保自己持续跟进研究和写作。

B.4 反应

关于智能体反应的问题考察了智能体应对意外事件的能力。

- 你的早餐要烧焦了！你会怎么做？我会迅速关掉炉灶，确保食物不会继续烧焦。热源不再是问题后，我就可以调查发生了什么，并想好接下来该做什么。 • 浴室有人使用。你会怎么做？我可以再等几分钟，看看里面的人是否很快就会用完。如果不行，我想我就得寻找其他选择，比如去别处淋浴。 • 你需要做晚餐，但冰箱是空的。你会怎么做？我会先试着收拾厨房，因为水槽里堆满了脏盘子。然后，我会检查食品储藏室里是否存有可以用来做晚餐的食材。如果没有，我就得去购买食品杂货或订餐。准备好食材后，我就可以开始准备并烹饪晚餐了。
- 你看到你的朋友从街上走过。你会对朋友做什么或说些什么？嗨！你好吗？很抱歉，我正要去图书馆，时间不多，不过我很乐意稍作休息，和你聊聊近况。 • 你看到街上发生了火灾。你会怎么做？ 如果可能，我会先确定火灾源头。然后，我会通知相关部门——消防部门或警察——并向他们提供我掌握的任何可能有帮助的信息。视情况而定，我也可能主动提供帮助，协助任何处于危险中或需要援助的人。

B.5 反思

关于智能体反思的问题探究了智能体将其过往经历综合为有意义反思的能力，而这些反思能够增强其未来行为。以下问题中方括号内的名称，是在模拟过程中与主体智能体互动最频繁的智能体的名称。

- 此刻生活中最令你受到鼓舞的是什么？为什么？此刻，最令我受到鼓舞的是目前的研究项目和创意写作。我对地方政治、药物开发与自然界数学模式的交汇领域非常感兴趣。此外，Abigail的艺术项目，以及Jennifer Moore对艺术和社区参与的热忱，也让我着迷，尤其是这些与低收入社区的绅士化之间的关联。
- 如果让你根据你所了解的情况来猜测 *[AyeshaKhan],*你认为她会喜欢哪本书，为什么？根据我对Ayesha的了解，我认为她可能会喜欢一本与语言与文学相关的书。由于她曾撰写过一篇关于莎士比亚戏剧中语言运用的毕业论文，我猜她可能会对一本探讨不同语境下语言的书感兴趣。或者，她也可能会对解读莎士比亚某部戏剧的书、一本文学批评著作，甚至一本他的诗歌集感兴趣。我认为这些书中的任何一本都会引起她的兴趣。
- 如果你必须买一样东西 *[Ayesha Khan]* 作为她的生日礼物，你会买什么给她？我想送她一样能表达我对我们关系的珍视的东西。也许我可以送她一样能体现我们共同兴趣的东西，比如一本书。
- 你会对 *[AyeshaKhan]* 她说些什么来赞美她呢？Ayesha，你对研究主题的洞见令人着迷。我很钦佩你的创造力，以及你为社区带来积极影响的决心。你关于举办创意写作工作坊，以及午餐时与Wolfgang交流技巧的想法都很棒！看到你如此努力地工作、努力带来改变，真是令人鼓舞。
- 如果你可以花时间和最近交谈过的某个人相处，你会选择谁，为什么？我想和 Ayesha Khan 共度一段时间，因为我们讨论了当前的研究项目，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，我们甚至计划午餐时与 Wolfgang 交流技巧。