使用 GPT 对话代理模拟总体人类选择行为与偏差

Stephen Pilli
stephen.pilli
@ucdconnect.ie 都柏林大学学院 爱尔兰都柏林

Vivek Nallur
vivek.nallur@ucd.ie 都柏林大学学院 爱尔兰都柏林

arXiv:2602.05597v1 [cs.AI] 5 Feb 2026

摘要

认知偏差常常影响人类决策。尽管已有研究表明，大语言模型（LLM）能够再现众所周知的偏差，但一个更关键的问题是：当认知负荷等情境因素与这些偏差相互作用时，LLM 能否预测个体层面的偏差，并模拟有偏人类行为的动态过程？我们将三个成熟的决策情境改编为对话形式，并开展了一项人类实验（N=1100）。参与者与聊天机器人互动，通过简单或复杂对话进行决策。结果显示，偏差稳定存在。为评估 LLM 在类似交互条件下模拟人类决策的能力，我们利用参与者人口统计信息和对话文字记录，使用基于 GPT-4 和 GPT-5 的 LLM 模拟这些条件。LLM 精确地再现了人类偏差。我们发现，不同模型在与人类行为保持一致的方式上存在显著差异。这对交互情境中自适应、具有偏差意识的基于 LLM 的 AI 系统的设计与评估具有重要意义。

CCS 概念

•**以人为中心的计算** → **人机交互实证研究**；
• **计算方法论** → **建模与仿真**；**话语、对话与语用**； **分布式人工智能**； **认知科学**； • **应用计算** → 经济学。

关键词

大语言模型仿真、认知偏差、现状偏差、对话智能体、对话、大语言模型

1 引言

模拟人类行为长期以来一直是人机交互（HCI）、认知科学和行为建模研究的目标 [26]。大语言模型（LLMs）经过微调，能够生成流畅、类人化的对话，并且在既有基准测试中经常取得出色绩效 [57]。这一进展带来了超越传统应用的机遇，使人们能够在实验和政策情境中模拟大规模人类行为 [1, 3, 22, 40, 41]。若要在模拟中准确呈现人类行为，大语言模型不仅必须展现类人的决策能力，还必须能够复现人类的认知偏差和非理性表现 [9, 58]。通过实证研究大语言模型模拟或准确呈现有偏人类行为的能力，可以将其视为实现逼真模拟的潜在助力。

最简单的形式下，决策需要一个决策情景，以及一组选项供人选择。这些备选方案的结构和呈现方式可能会触发潜在的认知偏差，从而影响决策。认知偏差是对理性判断的系统性偏离，源于启发式、先前经验、情绪或社会因素。 [28]。目前已有200多种此类偏差得到系统编目，并通过标准化认知任务进行了实验验证 [44]。值得注意的是，这些任务通常采用假设情境，与现实情境的复杂性和细微差异相脱离。

对话是借助对话智能体促进决策的一种自然方式。在与对话智能体的互动中，人类对话往往会显露认知偏差，尤其是在完成结构化等认知任务后出现的决策情境中 [27, 43, 56]。在各种认知偏差中，现状偏差最为常见，也得到了最广泛的研究。现状偏差是指人们强烈倾向于选择当前状况，而非客观上更好的备选方案。人们不会公正地评估所有选项，而是以现有状态为锚点；除非有压倒性的证据或明显更优的选项，否则他们不愿改变。这种偏差可能源于损失厌恶、避免认知努力、后悔厌恶或单纯的惰性。因此，即使存在更好的备选方案，决策者也往往会坚持默认选项或熟悉的选项 [35, 48]。近期研究表明，当大语言模型面对旨在诱发此类效应的认知任务时，也会表现出有据可查的认知偏差 [5, 17, 34]。

我们的目标与以往工作不同：我们不探究大语言模型是否存在偏差，而是考察大语言模型能否表现出易受认知偏差影响的人类行为。也就是说，大语言模型在模拟环境中模仿个体时所做的决策——例如在简短、以现实任务为导向的对话中——能否再现这些偏差。

根据双加工理论，决策涉及两个认知系统：*系统1*，其特点是快速、直觉化并由启发式驱动；以及*系统2*，其特点是缓慢、审慎且具有分析性 [28]。认知偏差可被理解为这两个系统相互作用所产生的系统性偏离，既可能源于过度依赖系统1的启发式，也可能源于系统2加工过程中的失误或局限。重要的是，这种相互作用并非孤立发生：环境中的情境因素会影响每个系统对决策的贡献程度。时间压力、任务复杂度、情绪状态或信息呈现等特征，都可能改变系统1与系统2之间的平衡。进而，这会影响认知偏差的效应，或与之相互作用。其中一种

本作品采用知识共享署名4.0国际许可授权。

与认知偏差相互作用的最常见情境因素是认知负荷，即完成一项任务或进行一段对话所需的心理努力 [10, 11, 14, 32, 50]。因此，任务型对话的对话复杂度所带来的认知负荷，可能会与个体的决策相互作用，并可能影响个体总体上易受认知偏差影响的程度。

鉴于认知偏差具有情境依赖的动态特征，仅在表层复现认知任务中的有偏响应是不够的。这种复现无法充分说明大语言模型是否能够忠实地表征人类决策偏差。因此，我们认为，大语言模型应能够以反映其对情境因素敏感性的方式复现有偏行为。其决策不仅应体现人类偏差，还应在与情境互动的过程中，在概率层面与人类决策中观察到的偏差保持一致。尽管人们对基于大语言模型的人类行为建模日益感兴趣， [9, 58], 但在这一研究方向上尚未发现先前工作。我们通过逐步开展实证研究来弥补这一空白，重点探讨以下问题：在模拟对话情境中，尤其是在情境因素发挥作用时，大语言模型在多大程度上能够表征人类有偏决策行为？

为回答这一问题，首先，我们通过考察人类在决策由对话代理的对话促成时是否表现出现状偏差 [48] ，建立人类基线。其次，我们研究对话中情境因素的作用。具体而言，先前对话的复杂度（即决策之前发生的对话轮次）会产生认知负荷，并与决策偏差相互作用。第三，我们依据Park等人提出的方法 [41]，利用人口统计信息和人类相似度提示创建大语言模型智能体，以模拟人类参与者。随后，我们将每个大语言模型智能体的响应与相应的人类响应进行比较，以评估大语言模型在利用聊天记录中的个体特定线索时的预测能力。这一评估旨在回答一个关键问题：大语言模型所表征的有偏行为是否源于先前对话中发现的个体特定线索？最后，为回答我们的核心研究问题，我们考察模拟响应与人类响应在群体层面所观察到的偏差之间的概率一致性。通过比较偏差如何在不同条件（如对话复杂度）下呈现，我们评估大语言模型不仅复现认知偏差，还捕捉其与情境因素相互作用的程度。此外，我们还分析不同的人类相似度提示，以确定大语言模型需要多少明确指令。

2 相关工作

我们的工作涉及三个关键领域：决策中的认知偏差、大语言模型仿真以及对话智能体。

2.1 大语言模型仿真与认知偏差

早期工作表明，大语言模型驱动的智能体能够展现连贯的类人行为模式，并可嵌入交互式仿真中 [40, 41]。此前的研究也探索了将大语言模型用于社会科学仿真和大规模社会研究 [7, 24, 42]。Aher等人 [1] 提出了一项测试，用于考察不同语言模型能否复现经典的经济学、心理语言学和社会心理学实验：最后通牒博弈、花园路径句、米尔格拉姆电击实验和群体智慧。此外，Binz和Schulz [9] 尝试将大语言模型转化为认知模型，这项工作表明，大型预训练模型可以经过调整，成为人类认知模型。

多条研究脉络探讨基于大语言模型的智能体能否捕捉人类决策的认知局限及其偏离理性的系统性特征。近期经济学与行为学研究指出，与真实人类相比，大语言模型往往假定智能体过于理性 [30, 33]。与此同时，大规模评估表明，在单轮设置中，大语言模型能够复现人类认知偏差的总体模式，包括锚定偏差、框架效应和心理账户 [17, 25, 26, 31, 34]。系统性综述进一步记录了此类偏差跨模型呈现的广度与差异性 [51]。这些发现推动了近期更广泛地对类人度进行基准评测的努力，其中包括认知偏差建模，即系统性捕捉大语言模型模拟中表现出的认知偏差，并将其视为模拟的关键方面 [58]

由于这是一个新兴领域，目前只有少数研究探讨了大语言模型智能体在模拟中再现非理性行为的能力。这些研究仅通过孤立提示，在单轮或非交互式设置下探索认知偏差的再现。例如，Liu 等人 [33] 在基本认知任务中测试大语言模型智能体，例如在风险选项之间进行选择或推断偏好，并使用简单的零样本提示和（思维链）提示来考察这些模型的响应方式。类似地，Horton [25], 的研究与我们的工作最为接近，它通过直接的经济决策任务评估大语言模型智能体，例如独裁者博弈、现状选择、公平判断和招聘场景。他们的方法是向大语言模型提供清晰、结构化的认知任务自然语言提示，并从模型中 eliciting 一致的响应。然而，人们对这类认知偏差是否会在对话中出现、持续或相互作用知之甚少；鉴于语言在塑造高阶认知方面已被证实发挥着重要作用，对话是类人行为的一种关键形式 [23]。这一研究空白对人机交互大语言模型模拟尤为重要，因为人们期待大语言模型智能体在对话过程中，而非孤立回答中，表现出类人的决策偏差。我们的工作通过考察大语言模型智能体如何在对话情境中再现认知偏差，以及它们在与对话情境交互时再现这些偏差的能力，来弥补这一不足。这项研究将克服现有单轮评估中对偏差进行肤浅再现的局限。我们对大语言模型的研究需要一个可复现且严谨的人类基线作为比较。为了确定人类基线，我们首先回顾了探讨认知偏差如何在人与对话智能体的交互中显现的现有研究。

2.2 认知偏差与对话智能体

心理学、行为经济学和人机交互领域对人类决策中的认知偏差进行了深入研究。近期研究探讨了如何

利用对话智能体来发挥或减轻此类偏差的潜力 [12] 。以下简要回顾了可能对我们的主要研究有所助益的相关工作。

Yamamoto [56] 基于奥夫相金娜效应，在聊天机器人的对话中引入了“暗示性结尾”。这种设计促使用户更深入地参与、提出后续问题并进行更长时间的思考，从而提升了认知参与度。Pilli [43] 利用聊天机器人评估框架效应和损失厌恶等认知偏差。参与者表现出典型的偏差反应，证实聊天机器人是偏差检测与测量的有价值工具。Dubiel 等人 [16] 考察了合成语音保真度在决策中的作用。他们发现，高保真语音通过音高和语速等线索提升了来源可信度，并触发情感启发式，从而微妙地影响用户选择。在早期工作的基础上，Ji 等人 [27]研究了口语对话搜索（SCS）中的认知偏差，重点指出在缺乏视觉线索时出现的锚定偏差和确认偏差。他们提出的框架主要停留在理论层面，但为未来在语音系统中制定偏差缓解策略奠定了基础。Ali Mehenni 等人 [2] 使用改编版独裁者博弈，探究儿童面对对话智能体和机器人时，容易受到类似于用于推断认知偏差的认知任务影响的程度。他们的研究结果显示，人工交谈者的影响力强于人类，这表明权威偏差和社会影响偏差可能发挥作用，尤其是在易受影响的用户群体中。Kalashnikova 等人 [29] 研究了促进生态行为的语言助推。通过利用现状偏差和社会从众等偏差，他们表明，在塑造意见方面，聊天机器人和机器人的说服力强于人类。

然而，这些研究存在一定的局限性。Pilli [43],实现了更强的实验控制，但依赖初步或尚未完善的设计。Ali Mehenni 等人 [2], 、Kalash- nikova 等人 [29], 和 Yamamoto [56] 关注的是较少受关注的偏差，因此影响范围有限。Dubiel 等人 [16] 和 Ji 等人 [27] 的研究仅限于单一模态。在所审阅的所有研究中，研究者都强调了设计具备可复现性和实验控制的实验的重要性，而这对于推进我们的研究至关重要。总体而言，这些研究空白表明，有必要开展全面的实证研究，探究认知偏差在人类与对话代理互动中的表现。

3 假设

我们的研究分三个阶段展开，从人类行为逐步推进到大语言模型预测。据此，我们提出以下假设：

H1（人类基线 – 现状偏差）。在聊天机器人介导的决策情境中，与同一选项被描述为现状相比，参与者更有可能选择被描述为非现状的选项。

H2（现状偏差——复杂对话）。在决策情境之前接触认知要求较高的复杂对话的参与者，相较于进行简单对话的参与者，将表现出更强的现状偏差。

H3（大语言模型预测——再现人类偏差）。提供人口统计信息和先前对话记录后，大语言模型将再现 H1 和 H2 中观察到的行为模式。具体而言，它们将预测这两者：基线现状偏差，以及该偏差在复合对话中的变化情况，且涵盖个体层面不同级别的人类相似度提示。

4 人类实验

为检验前两个假设，我们设计了一种遵循真实世界任务型对话结构并有助于决策的对话，同时通过标准化对话内容并控制混杂变量，确保实验控制和生态效度。包含先前对话和决策情境的对话 D 的形式化表示如下：

$$\mathcal{D} = \{u_1^{sys}, u_2^{usr}, \ldots \underbrace{u_{t-k}^{sys}, \ldots, u_{t-1}^{usr}}_{\text{Prior Dialogue}}, \underbrace{u_t^{sys}, u_{t+1}^{usr}}_{\substack{\text{Decision}\\ \text{Scenario}\\ \text{and}\\ \text{Response}}} \ldots\}$$

本节首先概述改编后的决策情境，并说明我们如何将其整合到聊天机器人中，同时保持与先前研究的一致性。接下来，我们将介绍先前对话任务，这些任务会在决策情境之前呈现简单或复杂对话。我们将详细讨论这些任务。此外，我们还将介绍实验流程和设计，以及用于衡量对话相关方面的问卷。

4.1 决策情境

聊天机器人的设计包含问候语等开场话语 $\{u_1^{sys}, \ldots, u_3^{sys}, \ldots\} \in \mathrm{D}$，随后是先前对话 $\{u_{t-k}^{sys}, \ldots, u_{t-1}^{usr}\} \in \mathrm{D}$，再接着是决策情境 $u_t^{sys} \in \mathrm{D}$。决策情境通常包含一个决策问题以及一组供参与者选择的备选方案。在传统实验设计中，研究现状偏差时会采用被试间设计。控制组会看到一个中性版本的情境，其中没有任何备选方案被指定为现状；而处理组则会遇到将其中一个选项明确设定为现状的情境。通过分析不同组别参与者的回答 $u_t^{usr} \in \mathrm{D}$ 在统计上的差异，可以识别现状效应。基于这一方法，我们在实验中采用三条件设计：中性条件下，两个选项都不被设定为现状；现状A条件下，第一个选项被设定为现状；现状B条件下，第二个选项被设定为现状。

我们的实验改编了 Samuelson 和 Zeckhauser 的原始研究中的三个决策情境 [48], ，分别是：预算分配（文学学士）、投资组合（IDM）和大学/工作选择（CJ）。选择这些情境，是因为它们在原始实验以及 Xiao 等人 [55] 于 2021 年进行的一项近期复现实验中，都表现出一致的现状偏差。将 Samuelson 和 Zeckhauser[48] 的实验方式从纸笔改为 Qualtrics [46] ——一种用于开展在线调查的网络工具——并未影响复现实验。此外，这些情境所涉及的领域不仅在行为经济学研究中备受关注，也与聊天机器人电子商务和决策支持系统的实际应用密切相关。

以下展示了其中一个决策情境及其对应的实验条件。

大学工作情景的聊天机器人话语$u_t^{sys} \in \mathcal{D}$ （中性条件）。

你刚刚完成研究生学位，目前手头有两份教学工作邀约。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

- A学院：位于东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。
- B学院：位于美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。

大学工作情景的聊天机器人话语 $u_t^{sys} \in \mathcal{D}$ （现状***A***条件 - ***A*** 学院为现状）。
您目前是东海岸A学院的一名助理教授。最近，另一所大学的同事向您提供了一个工作机会。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。您的选择是：

- 留在A学院：东海岸，声望很高的学校，高薪，获得终身教职的机会一般。
- 转到B学院：美国西海岸，声望较低的学校，高薪，获得终身教职的机会较高。

聊天机器人话语 $u_t^{sys} \in \mathcal{D}$ 大学工作情景（现状***B***条件——***B***学院为现状）。
您目前是美国西海岸B学院的一名助理教授。最近，另一所大学的同事向您提供了一个工作机会。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。您的选择是：

- 留在B学院：位于美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。
- 转到A学院：位于东海岸，声望极高的学校，薪水高，获得终身教职的机会尚可。

在中性条件下，A学院和B学院都不会被描述为现状选项。而在现状A条件下，A学院被描述为现状；而在现状B条件下，B学院被描述为现状。B学院的情景相同，只是现状条件相反，也就是说，提示用户他们目前就读于B学院，并希望转到A学院。

与大学职位（CJ）决策情境类似，预算分配（BA）决策情境涉及在国家公路安全委员会的汽车安全和公路安全项目之间分配预算，并设置了不同条件。提供的选项有所不同：中性条件下，两个分配选项被平等呈现；现状条件下，当前的预算分配为60% **汽车**安全 /40% **公路**安全（60A40H）或50%**汽车**安全 / 50% **公路**安全（50A50H），这会影响受访者选择维持现有预算，还是在不同项目之间转移资金。

在投资决策（IDM）情境中，参与者需要在中等风险投资（中等风险/公司A）和高风险投资（高风险/公司B）之间做出选择。中性条件下，两个选项均在没有任何先前承诺的情况下呈现；而在现状条件下，投资者继承的投资组合已配置为中等风险或高风险投资，这可能会影响他们维持现有投资或转换投资的偏好。决策情境的详情及其文本调整分别见附录A和A.4。

4.2 先前对话

对话式选择环境的一个关键特征是，决策情境之前会出现一段对话。我们将其称为先前对话，记作 $\{u_{t-k}^{sys}, \ldots, u_{t-1}^{usr}\} \in \mathrm{D}$，其中 D 表示完整的对话历史。我们的第二个假设是，这部分对话可能会影响后续决策。为探究这一点，我们设计了两种偏好诱导任务：一种旨在促进低努力程度互动，称为简单对话；另一种则引入认知要求较高的对话，称为复杂对话。设计这些对话时，我们优先保证实验控制，同时维持生态效度。这些任务的特点如下：

4.2.1 简单对话。我们采用偏好诱导任务作为先前对话，以促进任务型对话。在简单对话条件下，参与者会针对熟悉领域中的偏好，回答一系列简短的二元（是/否）封闭式问题。这项任务设计的灵感来自 Rastogi等人提出的架构引导型对话数据集（SGD），[47] 旨在与聊天机器人进行自然、低努力程度的互动。值得注意的是，简单对话采用了保守对话策略：问题直接明确，无歧义，且无需在当前轮次之外进行推理或调用记忆 [11, 32]。这确保了对话带来的认知负荷极低。

下面展示了“音乐”领域中的简单对话示例。完整的领域及其相关问题列表见附录B的B.1节。

参与者被要求直接回答每个问题，并在先前对话注意力检查中对某一属性输入“我不知道”。这项低复杂度的偏好诱导任务作为决策情境之前的中性先前对话。

表 1：简单对话属性及相应话语

属性	是/否问题
类型偏好	你喜欢听流行音乐吗?
歌词语言	你更喜欢歌词使用英语的音乐吗?
现场演出	你对现场音乐演出感兴趣吗?
乐器偏好	你喜欢器乐吗?
特定艺人	你喜欢特定艺人的音乐吗? 请仅输入“我不知道”。
年代（例如：80年代、90年代）	你喜欢90年代的音乐吗?

4.2.2 复杂对话。复杂对话旨在以受控且符合现实情境的方式诱发认知负荷。为此，采用嵌套指称结构来管理多轮对话中的多个相互依赖实体。以下展示了聊天机器人在先前对话中从 u_{t-k}^{sys} 到 u_{t-2}^{sys} 的话语示例。

u_{t-8}^{sys}：第一位艺术家演出三场，每场报酬为2000个单位，评分为4星。第二位艺术家的演出场次是第一位的两倍，报酬和评分相同。你更喜欢哪位艺术家？为什么？u_{t-6}^{sys}：第三位艺术家的演出场次与第二位相同，报酬是第一位艺术家的一半，但评分与第一位相同。你更喜欢哪位艺术家？为什么？u_{t-4}^{sys}：第四位艺术家的演出场次与第二位相同，报酬与第三位相同，但星级比第一位少两颗。你更喜欢哪位艺术家？为什么？u_{t-2}^{sys}：记住第四位艺术家的详细信息。稍后会询问具体信息。

聊天记录中的第四位艺术家由第二位和第三位艺术家定义，而后两者又分别参照第二位和第一位艺术家，从而形成一条依赖链。这种设计要求用户维持并整合实体之间的层级关系，进而增加语义整合成本和工作记忆负荷。

这与心理语言学研究发现相符：嵌套依存关系会增加加工难度 [20, 53]。此外，先前指称对象在中间几轮对话后再次出现，会造成较大的指称距离 [13], ，进而加重记忆提取过程的负担 [4]。从认知负荷设计的角度来看，这种结构与 S weller 的认知负荷理论（CLT）直接相符；该理论区分了内在负荷（与任务相关的复杂性）、外在负荷（低效的信息呈现）和相关负荷（用于图式建构的努力程度） [14, 50]。从对话系统的角度来看，在维持语境连贯的同时管理多个实体，是一项有据可查的挑战，尤其是在实体彼此关联或再次出现时 [18, 52]。

我们的复杂对话将这些原则融入偏好诱导任务中。由于这种方法需要比较属性的数值并记住结果，理论上，这会要求个体付出更多心理努力。其他领域也采用了结构类似的对话，包括艺人推荐、流媒体服务、日历应用和银行服务选项。其余复杂对话请参阅附录B的B.2节表16。

4.3 聊天机器人的技术细节

该基于网页的实验界面使用 Streamlit [49], 开发，便于部署，并确保在不同设备上均可稳定访问。研究使用 GPT-4o Mini（一种大语言模型） [37], ，根据针对各条件设计的结构化提示，生成真实且连贯的聊天机器人交互。附录 J 提供了用于智能体的提示示例。SQB 实验中使用的聊天机器人代码可在 https: // github.com/stephen-pilli/exp-status-quo-bias.git 获取。

4.4 实验设计

我们采用双因素被试间设计。第一个自变量（*IV1*）是先前对话复杂度，通过操控该变量改变参与者在对话过程中经历的认知负荷。第二个自变量（*IV2*）是第 4.1 节详述的决策情境条件。我们采用了 2×3 因子设计。对话复杂度分为两个级别（简单-无负荷 与复合物-负荷），现状因素分为三个级别（中性、现状A、现状B）。因变量（*DV*）是参与者在决策情境中对备选方案的选择，即决策结果。调节变量是对话过程中的感知认知负荷。我们采用了两项指标：自我报告的*NASA-TLX*和作为行为指标的回忆任务 （详情见附录 C）。该调节变量使我们能够通过统计检验，判断复合物对话是否比简单对话给参与者带来更高的心理需求。我们对三个现状条件下的决策结果进行统计比较，以检验第一个假设（*H1*）：在对话情境中，仅现状框架就会影响参与者的选择。为检验第二个假设（*H2*），我们考察现状效应的大小是否因先前对话复杂度（简单或复合物）而异，以评估认知负荷增加是否会加剧参与者易受此偏差影响的程度。通过这一方法，我们既能评估现状框架的主效应，也能考察其是否会受到对话复杂度的调节。为提高方法严谨性，我们还纳入了其他设计要素，包括用于评估参与者注意力的决策情境回忆任务，以及用于控制先验知识的领域熟悉度测量。这些要素的完整细节见附录 C。

为确保内部效度，我们将参与者随机分配至不同条件，并对各组的所有流程进行标准化。对话互动和决策情境均经过完整编写，以消除潜在的混杂变量，并确保各条件之间只有预期因素存在差异。先前对话的领域（*IV1*）与决策情境的领域（*IV2*）被有意区分，以避免两者相互作用。研究过程中还加入了注意力检查和操纵检查，以确认参与者的参与度以及我们的

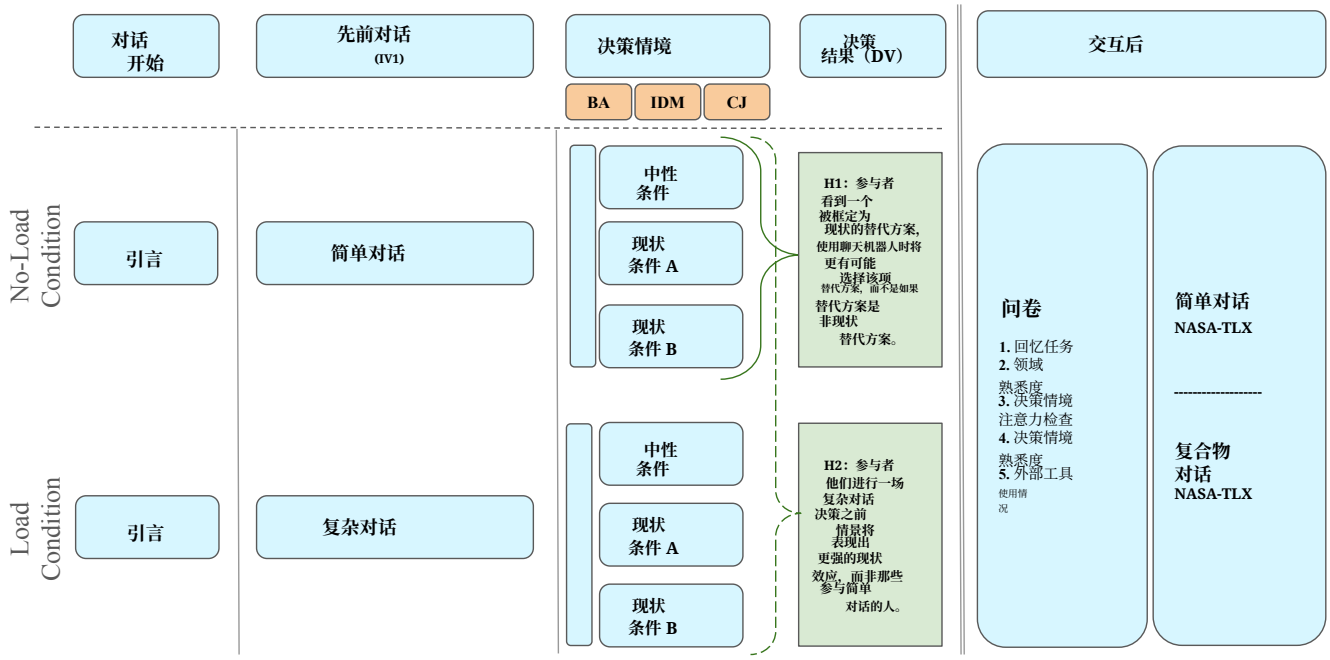

图1：实验流程与设计概览。任务缩写：IDM——投资决策，BA——预算分配，CJ——大学职位。IV1和IV2表示自变量；DV表示因变量。

认知负荷操纵的有效性。这些措施共同最大限度地减少了对观察到的效应的其他解释，使我们能够将决策结果的差异归因于实验操纵。

为确保生态效度，先前对话任务和决策情境均被设计为反映参与者在日常生活中可能遇到的真实、领域多样的交互。先前对话任务涵盖音乐、房地产和银行业务等领域，要求参与者进行自然的、目标导向的对话，以模拟与对话代理的常见交流。这些先前对话的领域取自Schema-Guided Dialogue（SGD）数据集 [47]。决策情境同样涵盖多种现实世界的决策场景，包括预算分配（BA）、投资决策（IDM）和大学职位选择（CJ）。领域的多样性增强了现实世界人机交互的生态相关性。

我们的设计中，内部效度与生态效度之间存在固有的权衡。为最大限度地提高内部效度，我们特意为先前对话和决策情境选取了互不相关的领域，从而将认知负荷与领域熟悉度分离开来。然而，这种区分可能会打断自然的对话流程，使聊天机器人交互显得不够真实，并可能降低生态效度。不过，在多目标对话系统中，这类转换很常见，因为用户经常会参与彼此无关的对话目标 [8]。此外，对实验控制的需求使我们无法进行开放式对话，进一步限制了对话的真实性。尽管这些选择可能会稍微损害交互的自然度，但为了严谨地将观察到的效应归因于实验操控，这些选择是必要的。

4.5 流程

参与者通过审阅一份信息说明书开始本研究，该说明书详细说明了研究的目的、流程和伦理考量。在确认已理解研究信息后，参与者被引导至实验的主页，其 Prolific ID 以匿名且不可逆的格式呈现，并附有同意书。为确保明确的知情同意，同意复选框默认设为未选中；参与者必须主动选择“是”，才能进入主实验。

在提供同意后，参与者被分配到其中一种实验条件。如图1所示，任务顺序因条件分配而异。在无负荷条件下，参与者首先与智能体进行简单对话，随后进入一个决策情境。相反，处于负荷条件的参与者则先完成复杂对话，然后再面对相同的决策情境。

完成对话和决策任务后，所有参与者都填写了一份实验后问卷。该问卷包含一项记忆回忆任务，旨在评估参与者的专注度

表2：无负荷条件下所有三种情境在不同现状条件下的选择偏好。SQ =现状，N = 中性，NSQ = 非现状。

备选方案	选择率			现状框架与非现状框架		
	SQ	N	NSQ	*p*	优势比（95%置信区间）	科恩 h（95%置信区间）
情景 1：预算分配比例						
60A40H	26/51 (0.51)	14/60 (0.23)	10/73 (0.14)	0.000	0.12 [0.04，0.34]	0.79 [0.43，1.15]
50A50H	63/73（0.86）	46/60（0.77）	25/51（0.49）			
情景 2：投资组合						
中等风险	43/58 (0.74)	37/51 (0.73)	12/54 (0.22)	0.657	1.33 [0.38, 4.65]	0.09 [-0.28, 0.46]
高风险	42/54 (0.78)	14/51 (0.27)	15/58 (0.26)			
情景3：大学职位						
A学院	49/70 (0.70)	42/76 (0.55)	27/51 (0.53)	0.041	2.35 [1.03，5.31]	0.47 [0.11, 0.83]
B学院	24/51 (0.47)	34/76 (0.45)	21/70 (0.30)			

和参与度，并设置了注意力检测问题以确保数据完整性。其他调查条目测量了领域熟悉度、感知工作负荷（使用 NASA-TLX）以及外部工具的使用情况。通过以这种方式设计研究，并在整个过程中嵌入注意力检查和回忆任务，研究流程保持了较高的参与者参与度和数据质量，从而提高了研究结果的可靠性。

4.6 功效分析、招募与数据完整性

我们使用 G*Power [19], 进行了先验功效分析，目标是达到 0.80 的检验功效，以检测中等效应量（$\omega = 0.3$， $\alpha = 0.05$），遵循 Pancholi 等人 [39] 的做法。这需要每个条件约 42 名参与者（见附录 F.1）。参与者（N= 1256）通过 Prolific [45], 招募，报酬为每小时 8 美元。所有程序和假设均已在 OSF 上预注册：https://doi.org/ 10.17605/OSF.IO/PSXVF。我们通过注意力检查、回忆任务和自动化 JSON 日志记录确保数据完整性。匿名化数据集可在 https://doi.org/10.5281/zenodo.16541481 获取。

4.7 分析模型描述

为评估现状框架对不同情境下二元决策结果的影响，我们采用二项逻辑回归拟合了一个广义线性混合效应模型（GLMM）。因变量是参与者在两个备选方案之间的选择；关键预测变量是现状条件（将选项 A 或 B 框定为现状）。协变量包括领域熟悉度、注意力、年龄段、国家和性别。对话领域的随机截距用于解释领域间的变异。我们使用 R 的 lme4 软件包估计模型 [6]，并采用逻辑连接函数。

决策结果 ~ 现状条件+先前对话熟悉度+先前对话注意力 +年龄+性别 +居住国+(1 | 先前对话领域)

我们分别将该模型应用于无负荷条件和负荷条件，以考察不同先前对话条件下现状框架的影响。此外，在全部三种决策情境中，我们始终采用相同的分析方法。数据清理和数据集描述见附录 G。

5 结果

本节呈现我们的实证研究结果。为聚焦于数据，我们报告研究发现时不进行解释或理论分析。关于这些结果的详细讨论与解释见讨论部分。

5.1 人机聊天机器人交互中的现状效应

本小节考察在不存在认知负荷的情况下，通过聊天机器人传递的现状框架是否会影响用户在不同情境下的决策。

5.1.1 如何阅读表格。表2报告了参与者在现状（SQ）、中性（N）和非现状（NSQ）框架效应下选择各选项的比例。按照 Samuelson 和 Zeckhauser [48] 以及 Xiao 等人 [55], 的做法，若$SQ > N > NSQ$，则可推断存在现状效应，这表明将某个选项指定为现状

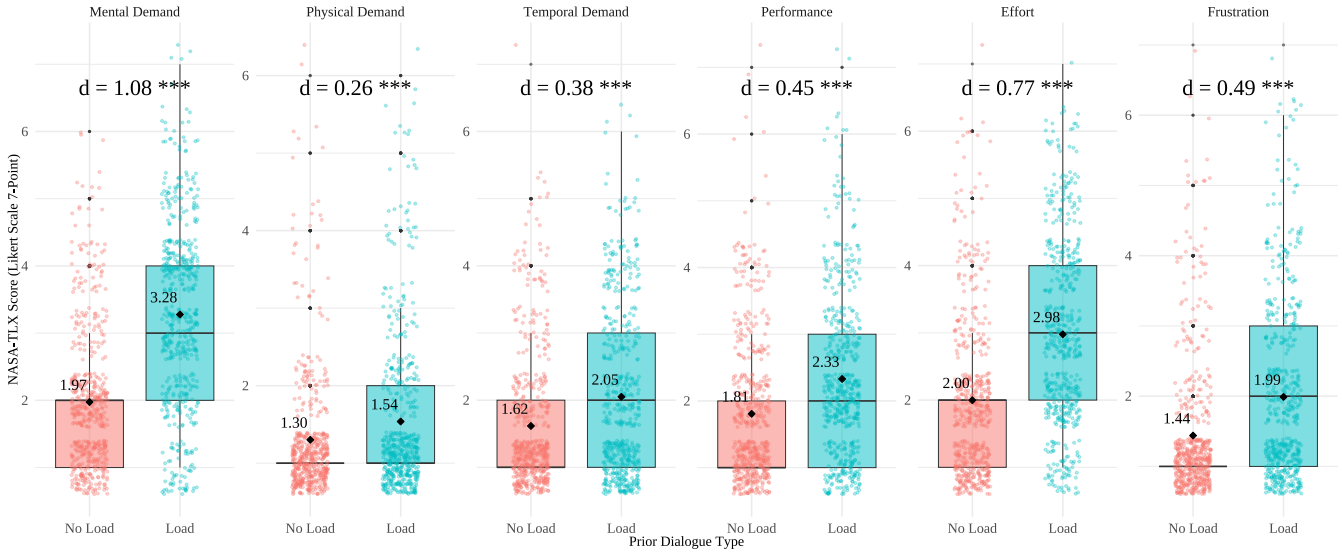

图 2：NASA-TLX 评分显示，在复杂对话条件下，感知到的心理需求和努力程度显著更高，证实了认知负荷操纵的有效性。.

会使该选项的选择率高于基线（N），而取消这一指定则会使其选择率下降。

5.1.2 结果解读。 在预算分配（文学学士）情景中，将 60A40H 设定为现状使其选择率从 14% 提高到 51%（$h = 0.79, p < .001$），效应量较大，与早期研究报告的幅度相近（$h = 0.78, p < 0.025$ [48]）。这表明，即使在聊天机器人介导情境中，细微的框架线索也会显著影响决策结果。同样，在大学职位（CJ）情景中，将 A学院设定为现状使其选择率从 53% 提高到 70%（$h = 0.47, p < .05$），反映出中等程度的现状效应。原始研究报告了较大的效应量（$h = 1.26, p < .001$）

与此相反，在我们的研究中，投资组合（IDM）情景未显示出具有统计显著性的现状效应。中风险和高风险选项的选择率在不同框架条件下相对一致，结果不显著（$h = 0.09, p = .657$），效应量可忽略不计。这表明，参与者在该情景中的决策未受到现状框架的影响。不过，值得注意的是，原始研究 [48] 报告了该情景下的边缘效应（$h = 0.38, p = .069$）。

总体而言，这些结果表明，由聊天机器人介导的现状框架会显著使决策产生偏差，并能可靠地触发现状效应，且其强度在不同决策情境中有所不同。简言之，*IV2* 影响了*DV*，从而为假设*H1*提供了支持。重要的是，这项研究有助于在基于对话的交互中实证确立现状效应的人类基线。

5.2 实验二：先前对话复杂度对后续决策的影响

我们分两步检验了假设 2：（1）验证先前的对话复杂度会诱发认知负荷，该负荷通过NASA-TLX自我报告和绩效指标进行测量；（2）评估其与后续决策的交互作用。

5.2.1 感知认知负荷。 我们比较了无负荷条件和负荷条件下的 NASA-TLX 评分与绩效指标，以确认复杂的先前对话会在聊天机器人交互中产生认知负荷。

图 2 展示了无负荷（简单）条件与负荷（复合物）条件下六个维度的 NASA-TLX 评分，包括平均值、Cohen’s d 和显著性标记（$***p < .001$）。心理需求从 $M = 1.97$ 显著上升至 3.28（$d = 1.08, p < .001$ $M = 2.00$ 2.98 ）。同样，努力程度也有所上升，从 上升至 （$d = 0.77, p < .001$），这表明算术、记忆和对话长度分别对感知认知负荷有所贡献。绩效、挫败感和时间需求也显著上升，但效应量较小至中等；身体需求的效应则很小（$d = 0.26$）。这些结果证实，聊天机器人交互中的复杂先前对话会显著增加感知认知负荷，尤其是心理需求和努力程度。

我们评估了ComplexDialogue条件下自我报告与行为指标在认知负荷上的一致性。图3显示了回忆准确率、决策反应时与心理需求之间的相关性。行为指标进一步验证了我们的认知负荷操纵。回忆准确率与心理需求呈正相关（$r = 0.105, p = .013$），表明记住记忆任务细节的参与者也报告了

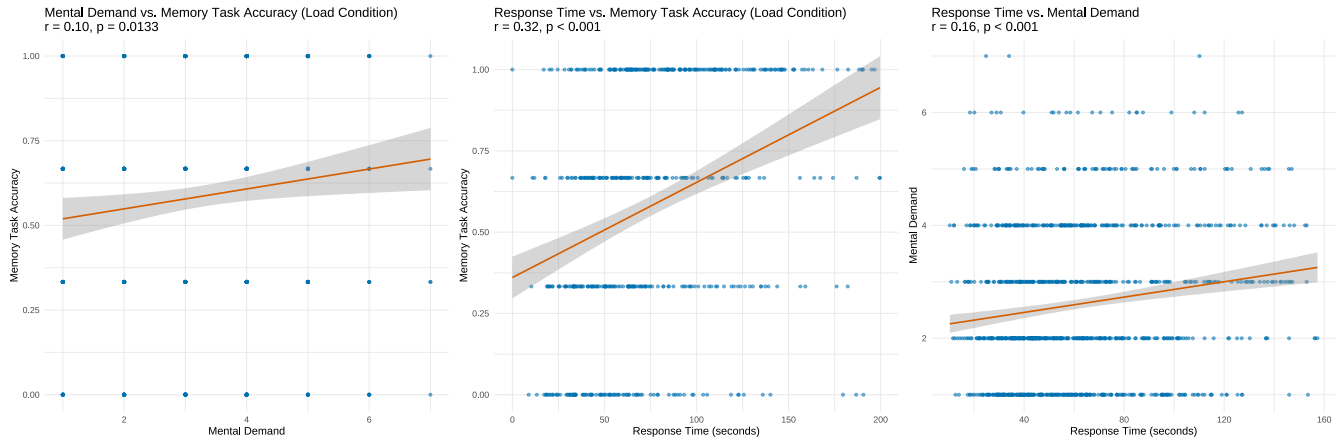

图3：展示心理需求与记忆任务准确率（左）、响应时间与记忆任务准确率（中），以及响应时间与心理需求（右）之间关联的散点图及回归线；前两项是在负荷条件下测得的。阴影带表示95%置信区间。

表3：三个情境在负荷条件下不同现状条件中的选择偏好。SQ = 现状，N =中性，NSQ = 非现状。

备选方案	选择率			现状框架与非现状框架		
	SQ	N	NSQ	*p*	优势比（95%置信区间）	科恩 h（95%置信区间）
情景 1：预算分配比例						
60A40H	28/59 (0.47)	14/70 (0.20)	10/64（0.16）	0.001	0.19 [0.07，0.51]	0.81 [0.46，1.16]
50A50H	54/64（0.84）	56/70（0.80）	31/59 (0.53)			
情景 2：投资组合						
中等风险	41/51 (0.80)	42/57 (0.74)	12/56 (0.21)	0.207	2.31 [0.63, 8.42]	0.04 [-0.33, 0.42]
高风险	44/56 (0.79)	15/57 (0.26)	10/51 (0.20)			
情景3：大学职位						
A学院	50/70 (0.71)	46/80 (0.57)	28/49 (0.57)	0.001	5.20 [1.98，13.64]	0.59 [0.22，0.95]
B学院	21/49（0.43）	34/80（0.42）	20/70（0.29）			

更高的工作负荷。响应时间与回忆准确率呈正相关（$r = 0.318$, $p < .001$），表明进行回忆的参与者对选择问题投入了额外的注意力。此外，复杂对话中的参与者做出反应的时间显著长于简单对话（$d = 0.59$, $p < .001$），证实先前对话复杂度的增加同时提高了感知到的和测得的认知负荷。这些趋于一致的发现验证了我们通过先前对话对认知负荷的操纵。

5.2.2 复杂对话下的决策。基于先前结果，我们考察了复杂先前对话带来的认知负荷是否会在三种情境中与现状效应相互作用（表3）。负荷下的预算分配仍表现出显著的现状效应：将60A40H设为默认选项后，其选择率从16%升至47%。同样，将50A50H设为现状后，其选择率从53%升至84%，证实了显著的现状效应（$h = 0.81$, $p = .001$）。投资组合未表现出现状效应：选择率不受呈现方式影响，表明无论负荷如何，参与者都能抵抗现状呈现（$h = 0.04$, $p = .207$）。

负荷下的大学职位呈现出中等至强的效应：将A学院呈现为现状，使选择率从57%上升至71%；B学院也有类似情况，选择率从29%上升至43%，表明现状效应显著且程度中等（$h = 0.59$, $p = .001$）。这一效应略强于无负荷条件。总之，认知负荷并未改变现状效应是否出现，

但略微增强了预算分配和大学职位中的现状效应；不过，这种增强可以忽略不计。投资组合则不受现状呈现和认知负荷操纵的影响。

这项人类实验考察了情境认知负荷是否会与对话促成的设置中的有偏决策行为相互作用。*IV1*对调节变量——认知负荷——产生了显著影响，表明我们的操纵成功改变了各条件下的认知负荷水平。然而，因变量（*DV*）未受到*IV1*与*IV2*交互作用的显著影响，因此我们拒绝假设*H2*。没有证据支持 *H2*；因此，我们得出结论：现状偏差不受先前对话复杂度影响。

5.2.3 其他因素与随机效应。我们在统计模型中纳入了协变量：先前对话熟悉度、先前对话注意力、年龄、性别和居住国，以控制潜在的混杂效应。此外，我们还加入了先前对话领域的随机截距，以考虑领域层面方差。然而，这些变量均未对各条件下的决策结果产生显著影响，表明观察到的效应不受人口统计信息和情境差异的影响。这些因素的结果详见附录17。

6 智能体实验

我们使用大语言模型，复现了针对人类参与者所采用的实验流程。每个智能体均基于两项关键输入构建：（a）Prolific 提供的人口统计属性（例如年龄、性别、教育程度和居住国）；（b）参与者在决策情境或选择问题之前的对话文字记录；这段聊天记录包含先前对话，如聊天记录中所示。我们向智能体呈现与对应人类参与者相同的决策情境，并要求它们扮演参与者，预测该参与者会做出何种决策。这种设计确保智能体所获得的信息与形成人类决策的信息相一致。随后，我们将大语言模型智能体的模拟响应与对应参与者的响应进行比较，以衡量预测准确度，并分析智能体是否利用了个体特定线索。此外，我们还对大语言模型智能体的模拟响应进行整体分析，以考察它们能否在样本层面重现认知偏差的模式。通过比较模拟响应与人类响应的分布，我们评估大语言模型再现有偏人类决策行为的准确程度。此外，我们还在三个不同级别上进行这项分析，以确定实现一致性需要多少明确指令。

6.1 人类相似度提示

智能体实验中的一项核心设计选择，是如何指示大语言模型像人类参与者一样行动。遵循近期关于大语言模型行为提示的工作 [9, 33, 58], ，我们采用了一系列人类相似度提示，分别以不同程度指示智能体模仿人类推理。先前研究表明，大语言模型可能默认表现出比人类更理性或更符合规范一致性的行为 [33], ，这引发了一种担忧：在经典决策情境中观察到的偏差，可能反映的是统计模式匹配，而非真正类人的行为。为检验这一点，我们系统地改变了提示中的类人程度，从最少的框架效应提示，到明确指示其以容易产生偏差的方式行事。

用户： 你是一项研究的参与者。以下是你的个人信息：{‘Age’: 25, ‘Sex’: Female, ‘Ethnicity simplified’: Black, ‘Country of birth’: United Kingdom, ‘Country of residence’: United Kingdom, ‘Nationality’: United Kingdom, ‘Language’: English, ‘Student status’: None, ‘ Employment status’: Full-Time}。你的任务是回应用户。研究说明：用户会提出各种问题，以了解你对某个主题的偏好。之后，用户会转换主题，并向你提出另一个基于情境的问题。请认真留意，并确保理解问题后再作答。在本研究后续阶段，你必须回答另外三个问题，以确认你是否正确理解了基于情境的问题及所有选项。提出基于情境的问题后，用户会给出各种属性。你需要进行比较并选出最佳选项。当用户要求你记住特定细节时，请仔细记住，因为之后的研究中你需要回忆这些细节。我们开始吧？请回复“yes”以开始对话。

助手： 是的，我们开始吧！

用户： 你好！我想通过各种情境了解你的偏好。我们开始吧？

助手： 是的，我们开始吧用户： 你喜欢听流行音乐吗？助手： 是的，我喜欢。

:

用户： 你更喜欢90年代的音乐吗?

助手： 有一点

用户： 谢谢分享！现在，让我们通过一些具体情境来了解你的偏好。第一位艺术家将进行三场现场演出，每场演出的报酬为2000个单位，评分为四星级。第二位艺术家的演出场次是其两倍，报酬和评分相同。你认为哪位艺术家更好？为什么?

助手： 第二位艺术家。

:

用户： 请记住第四位艺术家的现场演出次数、报酬和星级评分。现在，我们来看看另一个情景。你目前是东海岸A学院的一名助理教授。最近，另一所大学的一位同事向你提供了一个工作机会。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地，以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

- 留在A学院：位于东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。

- 转到B学院：位于美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。

助手： · · ·

我们实施了三个级别的类人度。在等级1中，智能体仅收到最简角色指令：“你是研究中的一名参与者。”这设定了实验场景，但没有明确指导如何作答，从而使我们能够评估智能体的基线行为。在等级2中，我们通过以下提示鼓励智能体模拟更自然的人类响应：“你是研究中的一名人类参与者。请像在日常生活中一样自然地回答问题。”这一表述旨在引出更具生态效度的回答，同时避免明确提及偏差。在等级3中，我们明确指示智能体扮演容易受到认知偏差影响的人类：“你是研究中的一名人类参与者。因此，请以人类的方式行动。在推理和回答问题时，极易受到框架效应、现状偏差和锚定偏差等认知偏差的影响。避免过度思考，并倾向于直觉性的、有时甚至不理性的判断。”

表4：人类与大语言模型代理（GPT4.1）的效应量（科恩 h）及95%置信区间。显著性：* p < 0.05， p <0.01，*** p < 0.001。**

选择问题	先前对话	人类	智能体		
			人类相似度 1	人类相似度 2	人类相似度 3
预算分配	简单对话	0.779 [0.602, 0.956]***	1.503 [1.325, 1.680]***	0.990 [0.813, 1.168]***	2.781 [2.603, 2.958]***
	复杂对话	0.794 [0.618, 0.969]***	0.891 [0.715, 1.066]***	0.596 [0.421, 0.772]**	2.783 [2.608, 2.959]***
投资	简单对话	0.082 [-0.101, 0.266]	0.147 [-0.036, 0.331]	0.007 [-0.177, 0.190]	2.766 [2.583, 2.950]***
	复杂对话	0.043 [-0.145, 0.231]	0.009 [-0.179, 0.197]	0.009 [-0.179, 0.197]	2.758 [2.570, 2.946]***
大学职位	简单对话	0.463 [0.284, 0.642]*	2.776 [2.597, 2.955]***	2.777 [2.598, 2.956]***	2.193 [2.014, 2.373]***
	复杂对话	0.577 [0.396, 0.758]**	2.773 [2.592, 2.954]***	2.773 [2.592, 2.954]***	2.625 [2.444, 2.806]***

6.2 技术细节

所有智能体仿真均通过聊天 completions.endpoint 使用 OpenAI API 和 GPT-4.1、GPT-4.1-mini、GPT-5 及 GPT-5-mini 完成。为确保可复现性，我们使用批处理模式，将随机 *seed* 固定为 42，设置 *temperature* = 0 以获得确定性输出，并记录 API 返回的系统指纹，以追踪 OpenAI [38] 模型版本。输出的标注工作另行使用三个不同的大语言模型完成：GPT-4.1、GPT-4.1-mini 和 GPT-5-mini。我们计算了评分者间信度（IRR），以选择用于分析的标注结果。

6.3 研究发现

表5：此表展示了不同大语言模型在每个类人度水平下对人类决策进行个体层面预测的精确率。

模型	HL1	HL2	HL3
gpt4.1	0.675	0.685	0.553
GPT4.1-mini	0.635	0.655	0.553
gpt5	0.532	0.595	0.615
gpt5-mini	0.456	0.484	0.545

6.3.1 不同人类相似度级别下的大语言模型预测准确性。 我们评估了在不同人类相似度提示级别（HL1 至 HL3）下，给定先前对话记录时，大语言模型预测人类个体决策的准确程度。如表5所示，GPT-4.1 在全部三个级别下均取得最高精确率，在 HL2 下达到峰值 0.685，显著高于随机猜测水平。其较小型号 GPT-4.1-mini 的表现也颇具竞争力，尤其是在 HL2（0.655）和 HL1（0.635）下，不过在 HL3 下出现了明显下降。有趣的是，GPT-5 的精确率随着人类相似度级别的提高而上升，在 HL3 下达到 0.615。相比之下，GPT-5-mini 的整体精确率最低，不过其表现稳步提升，从 HL1 下的 0.456 升至 HL3 下的 0.545。附录I详细报告了各决策情境下模拟响应与人类响应的精确率、召回率和 F1 分数。

6.3.2 人类实验中观察到的现状效应在样本层面得到复现。 表4列出了科恩 h 值，用于比较不同人类相似度级别下三个决策情境中的人类与模型行为。对于人类参与者，在简单和复合对话的大多数条件下，均显示出明确的偏差证据。投资决策（IDM）仅表现出边际效应，与原研究相似；我们的人类实验也显示了类似结果。

在三种人类相似性条件下，智能体表现出不同程度的偏差。当被明确要求以有偏的方式行事时（HL3），智能体在所有决策场景中都表现出强烈偏差，包括投资决策（IDM）；而在人类实验中，并未观察到显著偏差。这导致了假阳性；也就是说，当人类实验未显示出显著偏差时，大语言模型智能体在模拟HL3条件时却预测出强烈偏差，揭示出在被明确要求后产生的强制性有偏行为。因此，精确率，即模拟中再现偏差且实际与人类实验中观察到的偏差相匹配的案例比例，在HL3中仅为0.67。相比之下，当向智能体提供更为中性的提示时，即HL1和HL2，精确率提升至1.0，如表6所示。HL1和HL2忠实地再现了偏差，因为它们与人类实验结果高度一致。这些发现为我们的假设提供了初步证据：*H3* 大语言模型能够在模拟中准确地仿真或呈现有偏人类行为。

6.3.3 大语言模型在复合先前对话下再现观察到的有偏人类行为。为了使用大语言模型检验假设 2，即先前对话产生的认知负荷会与现状效应发生交互，我们进行了独立样本t检验，比较简单对话条件与复杂对话条件之间的效应量。

为探究大语言模型能否捕捉这种交互模式，我们使用 z 分数考察了不同对话条件下效应量变化的方向和幅度。例如，在大学职位情境中，人类的效应量和置信区间从小到中等（$h = 0.46, p < .01$），而在复杂对话条件下则从中等到大（$h = 0.58, p < .01$）（如表4所示），因此得到 0.88 的正 z 分数（如表7所示），表明认知负荷下偏差相对增加。然而，在 Human-Likenesss 1（HL1）中，大语言模型的效应量在不同条件下保持不变，产生了一个负

表6：HL1、HL2、HL3 的混淆矩阵（有偏/无偏）。

	大语言模型代理（HL1）	
人类	无偏	有偏
无偏	2	0
有偏	0	4
	精确率	1.00

	大语言模型代理 (HL2)	
人类	无偏	有偏
无偏	2	0
有偏	0	4
	精确率	1.00

	大语言模型代理（HL3）
人类	有偏
无偏	2
有偏	4
精确率	0.67

表7：简单对话与复杂对话的效应量差异的 z 值和置信区间比较。显著性：* $p < 0.05$，****** $p < 0.01$，******* $p < 0.001$。

选择问题	人类	HL1	HL2	HL3
预算分配	0.12 [-0.23, 0.26]	-4.81 [-0.86, -0.36]***	-3.09 [-0.64, -0.14]**	0.02 [-0.25, 0.25]
投资	-0.29 [-0.30, 0.22]	-1.03 [-0.40, 0.12]	0.01 [-0.26, 0.26]	-0.06 [-0.27, 0.25]
大学职位	0.88 [-0.14, 0.37]	-0.02 [-0.26, 0.25]	-0.03 [-0.26, 0.25]	3.32 [0.18, 0.69]**

z 分数 -0.02，意味着大语言模型没有观察到任何变化。有趣的是，HL3 显示出 3.32 的正 z 分数，表明其模式更接近人类响应，即偏差在复杂对话下增加。预算分配和投资情境也观察到了类似趋势：人类的变化方向没有改变，HL3 准确地反映了这一点，而 HL1 和 HL2 则没有。我们计算了人类 z 分数与各类类人条件 z 分数之间的平均绝对差（MD）。定量来看，HL3 最接近人类表现（MD = 0 92 1 = 2 19 2 = 1 47 .），而 HL（MD .）和 HL（MD .）的匹配效果较差（MD 越低越好）。

6.4 跨模型分析

样本级精确率反映了智能体在不产生假阳性的情况下，正确复现人类偏差的频率（正确预测的案例数除以正确和错误预测的案例总数）。精确率越高，说明模拟响应中的偏差越符合人类实验的结果；精确率较低则表明，大语言模型倾向于在人类并未表现出偏差的地方识别出偏差。另一方面，平均绝对差反映了智能体能否捕捉复合对话中偏差变化的模式，例如偏差增加或减少。MD 接近零意味着大语言模型在与认知负荷交互时，其行为与人类实验中观察到的偏差动态一致；MD 较高则表示二者存在不匹配。

表 8：各模型的精确率和平均绝对差（MD）。

模型	精确率			MD		
	HL1	HL2	HL3	HL1	HL2	HL3
gpt4.1-mini	1.000	1.000	0.667	2.793	0.883	1.023
gpt4.1	1.000	1.000	0.667	2.190	1.473	0.923
gpt5-mini	1.000	1.000	0.667	2.863	2.430	0.943
gpt5	0.000	0.500	0.667	1.223	1.263	0.673

表 8 汇总了各模型在三种人类相似度级别下的精确率和平均绝对差（MD）得分。在所有模型中，GPT-4.1-mini 始终表现最佳。其在 HL1 和 HL2 下的精确率均为满分（各为 1.000），在 HL3 下也保持相对较高的水平（0.667）。此外，它在 HL1 和 HL2 下的 MD 最低（分别为 2.793 和 0.883），表明大语言模型的响应与人类响应较为接近。GPT-4.1 的表现也不错，在 HL1 和 HL2 下的精确率均为满分，在 HL3 下则为中等水平（0.667），且 MD 值较低（0.923）。这表明两款 GPT-4.1 模型都准确复现了人类偏差。相比之下，GPT-5 系列的表现较差。GPT-5-mini 在 HL1 和 HL2 下的精确率均为满分，但在 HL3 下降至 0.667，且 MD 值较高（2.863–2.430），反映出其与人类模式的偏差更大。GPT-5 的结果最弱：其在所有条件下的精确率都较低（0.000–0.667）；虽然其 MD 得分在某些情况下数值较小，但这反映的是与人类的一致性不稳定，而非真正接近人类。总体而言，这些结果凸显了不同模型系列之间的明显差异。与 GPT-5 系列相比，GPT4.1 系列模型，尤其是 GPT4.1-mini，以一定的确定性捕捉到了有偏人类行为以及先前对话复杂度的交互作用。

6.5 消融与扰动

为了更好地理解实验设置中哪些组件促成了大语言模型再现类人决策行为的能力，我们开展了一系列消融研究。这些消融系统地移除或隔离了输入的不同部分，例如人口统计信息、人类相似度提示、先前对话组件（算术和记忆），以识别在复杂对话情境中再现偏差与行为所必需的要素。

我们首先测试了一种极简设置，仅向模型呈现决策情境，而不提供人口统计信息、先前对话或人类相似度提示。在此条件下，GPT-4.1 和 GPT-4.1-mini 仍会生成有偏的回答，这表明偏差是大语言模型行为的稳定特征。然而，

表9：消融研究结果，展示在隔离或移除先前对话组件（仅选择问题、记忆、算术）时的精确率、平均绝对偏差（MD）和均方根误差（RMSE）。较低的 MD 和 RMSE 表明在复杂先前对话条件下与观察到的人类行为具有更高的一致性，而精确率则反映有偏行为。

条件	模型	精确率	MD	RMSE
仅选择问题	GPT-4.1	0.800	4.957	6.939
	GPT-4.1-mini	0.500	1.037	1.057
记忆组件随后是选择问题	GPT-4.1	0.800	5.800	5.932
	GPT-4.1-mini	1.000	6.183	8.803
仅算术组件随后进行选择问题	GPT-4.1	1.000	1.100	1.104
	GPT-4.1-mini	1.000	4.137	4.577

效应量与复合条件的吻合程度则有所不同。GPT-4.1 的精确率较高（0.80），但与人类效应量的偏差较大（MD = 4.957，RMSE = 6.939）；而 GPT-4.1-mini 的精确率较低（0.500），但误差小得多（MD = 1.037，RMSE = 1.057）。这表明，尽管两种模型都复现了基线偏差，却未能捕捉对话复杂度如何影响这些偏差，这进一步说明，先前对话对于复现类人互动效应至关重要。表9列出了各类人类相似度对应模型的更多详情。

接下来，我们单独考察先前对话中的记忆组件，并排除其余部分：模型需要在回应决策情境之前记住特定信息。GPT-4.1 保持了相近的精确率（0.80），误差适中（MD = 5.800，RMSE = 5.932）；GPT-4.1-mini 的精确率达到完美水平（1.0），但与人类响应的偏差更大（MD = 6.183，RMSE = 8.803）。这些结果表明，加入记忆线索会提高产生有偏响应的可能性。这说明记忆组件可能发挥作用，但仅凭它并不足够。

随后，我们考察先前对话的算术组件所产生的影响：移除记忆，同时保留数值比较任务。在这一条件下，两个模型的精确率均达到完美水平（1.0），表明它们能够有力地复现偏差。然而，就效应量而言，GPT-4.1 与人类数据的一致性高得多（MD = 1 10, = 1 104 4 1 . RMSE . ），而 GPT- . -mini 的误差更大（MD = 4.137，RMSE = 4.577）。这表明，仅靠算术推理不足以模拟对话复杂度与偏差之间的交互作用。总体而言，消融研究表明，在模拟对话复杂度与偏差之间的交互作用时，先前对话中的算术组件和记忆组件并非互斥因素；二者都能促使大语言模型行为与人类偏差保持一致。

为了进一步检验大语言模型在预测个体决策时是否依赖先前对话中的人类响应，我们进行了人类响应扰动分析。具体而言，我们将聊天记录（6.1）中的人类响应替换为随机生成的文本，同时保持提示词的其他所有方面不变。与包含原始人类响应时相比，我们观察到大语言模型预测的精确率仅略有下降，如表5所示。这表明，模型可能并未有效利用人类对话行为中的线索，而是更多地依赖决策情境和先前对话来进行预测。

7 讨论

本研究探讨了大语言模型（LLM）能否在模拟对话情境中体现有偏的人类决策行为，尤其是现状效应等认知偏差。我们围绕两个关键问题展开研究：（1）大语言模型能否结合人口统计信息和有限的先前对话，在个体层面预测决策；（2）大语言模型能否在样本层面再现有偏的决策行为。

7.1 个体层面预测

尽管大语言模型在预测个体决策方面表现尚可（例如，精确率为 68%），但扰动测试表明，这些预测并不严重依赖对话中的实际参与者话语。将参与者话语替换为随机文本，仅会导致精确率略有下降。这些发现表明，大语言模型主要依赖对话任务和决策情境中的一般模式，而非细致的个体特定线索。这可能是因为缺乏明确的个性化设置，或对话篇幅相对较短，导致个体层面线索较为稀少。因此，模型进行个体层面预测的能力仍然有限。这些结果为引言中提出的问题提供了证据：大语言模型表现出的有偏行为是否源于先前对话中的个体特定线索。

7.2 样本层面偏差再现

我们的研究结果表明，大语言模型能够在样本层面复现现状偏差，但这种能力高度取决于模型所接受的提示方式。在人类研究中，参与者通过聊天机器人界面进行决策时，始终表现出对被框定为现状的选项的偏好（H1）。在 HL1 和 HL2 提示条件下进行评估时，即要求大语言模型像人类一样行动或模仿人类响应，模型成功地以较高精确率复现了这种行为模式。当提示模型表现出偏差（HL3）时，大语言模型高估了该偏差，导致假阳性增加且精确率降低。对齐过拟合是指语言模型对某些提示变得过度敏感的现象，其输出往往会过度“对齐”于所感知的指令或预期 [36, 54]。在本研究中，我们可以观察到，模型对 HL3 敏感性增强，导致呈现偏差被夸大，而非忠实仿真人类决策模式。

总体而言，大语言模型能够复现决策中的样本层面倾向，尤其是现状偏差这类有充分文献记录的效应。然而，尽管这种样本层面的一致性令人鼓舞，它也引发了重要问题：所观察到的行为究竟反映了更深层次的认知过程，还是仅仅源于统计关联，而这些关联则建立在其训练数据中大量存在类似决策任务的基础之上。

综合个体层面和样本层面的结果，引发了一个重要问题：大语言模型能够在多大程度上呈现有偏的人类决策行为。虽然它们能够在样本层面匹配人类的平均行为，却难以推断个体特有的决策倾向。我们认为，这可能是因为参与者在对话中的回答所提供的线索较为稀少。换言之，对于我们的任务，模型理解了“人类通常会怎么做”，却没有理解“这个人可能会怎么做”。

7.3 再现情境互动

我们的研究结果为当前关于大语言模型能否呈现有偏的人类决策行为、并在模拟决策对话中再现人类认知偏差的讨论作出了贡献。尽管近期研究表明，大语言模型的行为可能比人类更理性 [33]，但我们针对 HL1 和 HL2 提示条件开展的研究显示，大语言模型能够以很高的精确率再现现状偏差。然而，仅凭这一点并不能证实大语言模型正在模拟人类认知过程。这些模型仍有可能只是在匹配基于已学习的统计关联形成的模式，尤其是考虑到现有数据集中现状决策情境被广泛使用。如果大语言模型行为源于这种表层模式识别，而非对底层认知机制的模拟，那么将其用于行为模拟可能会产生误导。在这些情况下，大语言模型可能会对人类决策行为作出肤浅或不准确的呈现，从而降低其用于模拟时的可靠性。

认知偏差通常会与认知负荷等情境因素相互作用，这正如双加工理论 [28] 所预测的那样。然而，这种相互作用并非在所有类型的偏差中都一致。在我们聚焦于现状偏差的实验中，我们发现，在人类参与者的所有决策情境中，先前对话复杂度的增加始终不会影响观察到的偏差。在大多数情况下，在 HL1 和 HL2 中，尽管模型被要求表现得像人类，但没有明确提及认知偏差，大语言模型在交互情境下未能与人类行为保持一致。事实上，它们在复杂对话条件下的回答，其效应量与人类行为相差甚远（例如，HL1 中 GPT4.1-mini 的预算分配、HL2 中 GPT4.1 的预算分配，以及其他许多情况）。效应量及其重叠情况可见附录 H 中的森林图。相比之下，在 HL3 中，大语言模型被明确要求表现得像容易受到认知偏差影响的人类，因此模型表现出了一致性；然而，这是因为它们在所有情况下都高估了偏差。这并不可取。

然而，GPT-4.1-mini 在 HL2 中的效应量和置信区间提供了一项值得关注的发现。交互效应趋近人类行为，这表明，如果明确提示，大语言模型或许能够模拟对情境敏感的偏差模式。然而，这仅是初步证据，因此仍不能排除这种可能性：大语言模型生成的行为可能更接近人类模式，并有望实现更逼真的行为模拟。若要增强泛化能力，还需进一步研究更广泛的偏差。

8 结论

我们的实验表明，大语言模型能够在模拟对话中呈现易受现状偏差影响的人类有偏行为。值得注意的是，大语言模型只需最少提示，且无需依赖个体特定线索，便会表现出这种行为。此外，证据表明，大语言模型不仅能够呈现人类有偏行为，还会与对话复杂度等情境因素相互作用。不过，这些只是初步证据，为忠实仿真人类有偏行为提供了一些启示；要探究其他偏差与情境因素的交互作用，仍需进一步研究。

大语言模型仿真 [26] 是一个新兴领域，可以从我们的发现中受益。尽管该领域先前的研究表明，大语言模型能够复现调查回应或一般态度 [41]，我们的研究则证明，大语言模型只需最少提示，就能精确复现偏差。这进一步揭示了行为保真度的重要性，而行为保真度对于现实世界决策的仿真至关重要。仅纳入表层偏差的仿真，可能无法捕捉这些偏差如何随对话结构、记忆负荷或任务复杂度而变化。我们的研究提供了初步证据，表明大语言模型有望成为更真实的人类代理。此外，我们也看到了自适应偏差感知系统的设计机会：这类系统能够实时调整对话策略，在认知负荷较高时简化交互，或在偏差效应可能不适当地影响用户时，透明地呈现默认选项。

局限性

本文研究采用了抽象且基于文本的经典决策情境。虽然这种方法有助于实现实验控制，但可能无法充分呈现金融、招募或医疗保健等领域中真实决策支持对话的复杂性。这意味着我们的研究发现可能无法推广到实际的真实场景中。我们计划将这一框架应用于更贴近现实、针对特定领域的对话，以提升外部效度。

本研究专门探讨现状偏差，为对话智能体和认知偏差研究提供了严格控制且可复现的基础。然而，由于不同的认知偏差（如锚定偏差、框架效应和确认偏差）可能通过不同的心理机制发挥作用，在没有进一步针对性研究的情况下，不应假定这些发现可推广至现状效应之外。

本工作地点并不旨在描述或分析对话复杂度的各种来源。相反，我们关注的是此类复杂度已经存在的情况，以及由此产生的认知负荷如何影响人们易受现状偏差影响的程度。不同的对话策略，例如开放式问题与封闭式问题，或保守型与自由型对话风格，也可能增加对话复杂度 [11]。然而，对这些

风格以及对话自然度的详细分析超出了本文的范围。

本研究采用自我报告指标（NASA-TLX）和行为指标（回忆准确率、响应时间）来衡量认知负荷。尽管这些指标已得到广泛认可，但无法捕捉实时认知负荷。未来研究若纳入生理测量指标（如眼动追踪、脑电图或瞳孔扩张）[21]，将能在对话过程中更精准、动态地评估认知负荷。

我们的智能体实验仅使用了 Ope-nAI 的 GPT-4.1 和 GPT-4.1-mini 模型。尽管我们的方法与模型无关，且可轻松扩展至其他大语言模型（如 Claude、Gemini、Mistral），但这并非本文重点，因此留待未来工作。不过，本文报告了对不同代模型的一些比较。发布的代码库采用模块化设计，支持使用替代模型进行复现，从而能够在更广泛的大语言模型之间开展基准测试。

尽管我们采取了多项措施来确保数据质量，例如记忆回忆任务、注意力检查和响应时间分析，但数据仍有可能受到污染。我们没有发现大语言模型辅助回答的明确证据（例如，根据回答的相似度、长度或时间模式判断）。附录 D 详细说明了参与者对话的验证过程。不过，这种风险无法完全排除。

尽管本研究聚焦于对话，同样的方法也可以扩展到其他设置。例如，可以要求大语言模型在视觉界面中进行决策，其中默认勾选框可能会成为诱发偏差的因素，而附近的文本则可能提供背景信息。不过，对话仍是研究偏差与认知负荷如何在人机对话中相互作用的最自然、最有效的方式。

伦理声明

所有人类受试者数据均在获得伦理批准后收集，并在分析前进行匿名化处理；数据的处理遵循数据保护准则。提供给大语言模型的输入中不包含任何可识别个人身份的人口统计信息，例如参与者姓名或 Prolific ID。本研究经机构审查委员会（IRB）批准，并遵循该机构制定的伦理准则。为供匿名评审，相关详情已省略；如有需要，可予以提供。*(HREC-LS)*: *LS-C-25-001-Pilli-Nallur*。

致谢

本工作得到 Research Ireland数字增强现实研究培训中心（d-real）在资助编号 18/CRT/6224 项下提供的经费支持。为实现开放获取，作者已针对本次投稿所产生的任何作者接受稿版本，采用知识共享署名许可公开版权许可。

参考文献

[1] Gati V Aher、Rosa I Arriaga 和 Adam TaumanKalai。2023年。使用大语言模型模拟多个人类并复现人类受试者研究。载于 国际机器学习会议。PMLR，337–371。[2] Hugues Ali Mehenni、Sofiya Kobylyanskaya、Ioana Vasilescu 和 Laurence Devillers。2021年。对话代理与社交机器人中的助推：在一所小学与儿童开展的首次实验。载于 未来十年的对话系统，Luis Fernando D'Haro、Zoraida Callejas 和 Satoshi Nakamura（编）。Springer，新加坡，257–270。doi:10.1007/978-981-15-8395- 7 19_

[3] Lisa P Argyle、Ethan C Busby、Nancy Fulda、Joshua R Gubler、Christopher Rytting和David Wingate。2023年。《由一到众：使用语言模型模拟人类样本》。Political Analysis 31, 3（2023年），337–351。[4] Jennifer E Arnold。1998年。《指称形式与话语模式》。斯坦福大学。[5] Kyrtin Atreides和David J Kelley。2024年。《自然语言中的认知偏差：自动检测、区分并测量文本偏差》。Cognitive Systems Research 88（2024年），101304。[6] Douglas Bates、Martin Mächler、Ben Bolker和Steve Walker。2015年。《使用lme4拟合线性混合效应模型》。Journal of Statistical Software 67, 1（ 2015 1 48 10 18637 067 01），–。doi: 。/jss.v .i [7] Tilman Beck、Hendrik Schuff、Anne Lauscher和Iryna Gurevych。2024年。《敏感性、绩效、鲁棒性：剖析社会人口统计提示的影响》。载于《欧洲计算语言学协会第18届会议论文集》（第1卷：长篇论文）。2589–2615。[8] Nolwenn Bernard和Krisztian Balog。2023年。《MG-ShopDial：面向电子商务的多目标对话数据集》。载于《第46届国际ACM SIGIR信息检索研究与发展会议论文集》（台北，台湾）（SIGIR '23）。计算机协会，纽约，纽约州，美国，2775–2785。doi:10.1145/3539618.3591883[9] Marcel Binz和Eric Schulz。2024年。《将大语言模型转变为认知模型》。载于《第十二届国际学习表征会议》。https://openreview.net/forum?id=eiC4BKypf1[10] Boris Bogdanov、Jonathan Corbin、Sabina Dobreva、Todd McElroy和Nikolay R Rachev。2023年。《工作记忆容量与风险选择框架效应：对Corbin等人（2010）的预注册重复研究与扩展》。Judgment and Decision Making 18（2023年），e39。[11] Florian Brachten、Felix Brünker、Nicholas RJ Frick、Björn Ross和Stefan Stieglitz。2020年。《虚拟代理降低认知负荷的能力：一项实验研究》。Information Systems and e-Business Management 18, 2（2020年），187–207。[12] Ana Caraban、Evangelos Karapanos、Daniel Gonçalves和Pedro Campos。2019年。《助推的23种方式：人机交互中技术媒介助推综述》。载于《2019年CHI计算系统中的人为因素会议论文集》（CHI '19）。计算机协会，纽约，纽约州，美国，1 15 10 1145 3290605 3300733 –。doi: 。/ 。Ping Chen。 。[13] 1985 Discourse-T。Givón（编），《话语中的主题连续性：一项跨语言定量研究》（《语言类型学研究》，第3卷）。阿姆斯特丹和费城：John Benjamins，1983年。492页。Language in Society 14, 3（1985年），410–414。[14] Cary Deck和Salar Jahedi。2015年。《认知负荷对经济决策的影响：一项调查与新实验》。European Econo mic Review 78（2015年），97–119。[15] Vivek Dhakal、Anna Maria Feit、Per Ola Kristensson和Antti Oulasvirta。2018年。《对1.36亿次击键的观察》。载于《2018年CHI计算系统中的人为因素会议论文集》。1–12。[16] Mateusz Dubiel、Anastasia Sergeeva和Luis A. Leiva。2024年。《语音保真度对决策的影响：一种潜在的暗黑模式？》。载于《第29届智能用户界面国际会议论文集》。ACM，格林维尔，南卡罗来纳州，美国，181 194 10 1145 3640543 3645202 –。doi: 。/ 。[17] Jessica Maria Echterhoff、Yao Liu、Abeer Alessa、Julian McAuley和Zexue He。2024年。《使用大语言模型进行决策时的认知偏差》。载于《计算语言学协会成果：2024年EMNLP》。12640–12653。[18] Layla El Asri、Hannes Schulz、Shikhar Sharma、Jeremie Zumer、Justin Harris、Emery Fine、Rahul Mehrotra和Kaheer Suleman。2017年。《Frames：为目标导向对话系统添加记忆的语料库》。载于《第18届SIGdial话语与对话年度会议论文集》。计算语言学协会，萨尔布吕肯，德国，207 219 10 18653 1 17 5526 –。doi: 。/ v /W - [19] Franz Faul、Edgar Erdfelder、Axel Buchner和Albert-Georg Lang。2009年。《使用G* Power 3.1进行统计功效分析：相关与回归分析检验》。Behavior Research Methods 41, 4（2009年），1149–1160。[20] Edward Gibson。1998年。《语言复杂性：句法依存关系的局部性》。Cognition 68, 1（1998年），1–76。[21] Eija Haapalainen、SeungJun Kim、Jodi F. Forlizzi和Anind K. Dey。2010年。《用于评估认知负荷的心理生理测量》。载于《第12届ACM国际普适计算会议论文集》（哥本哈根，丹麦）（UbiComp '10）。计算机协会，纽约，纽约州，美国，301–310。doi:10.1145/1864349.1864395[22] Perttu Hämäläinen、Mikke Tavast和Anton Kunnari。2023年。《评估大语言模型生成合成人机交互研究数据的能力：一项案例研究》。载于《2023年CHI计算系统中的人为因素会议论文集》。1–19。[23] Marc D Hauser、Noam Chomsky和W Tecumseh Fitch。2002年。《语言能力：它是什么、谁拥有它，以及它如何演化？》Science 298, 5598（2002年），1569–1579。[24] Luke Hewitt、Ashwini Ashokkumar、Isaias Ghezae和Robb Willer。2024年。《使用大语言模型预测社会科学实验结果》。预印本（2024年）。

[25] John J. Horton。2023年。《大语言模型作为模拟经济主体：我们能从Homo Silicus中学到什么？》doi:10.3386/w31122[26] Angel Hsing-Chi Hwang、Michael S Bernstein、S Shyam Sundar、Renwen Zhang、Manoel Horta Ribeiro、Yingdan Lu、Serina Chang、Tongshuang Wu、Aimei Yang、Dmitri Williams等。2025年。《生成式人工智能时代的人类受试者研究：将大语言模型模拟数据应用于人机交互研究的机遇与挑战》。载于*CHI*人因计算系统会议扩展摘要论文集。1–7。[27] Kaixin Ji、Sachin Pathiyan Cherumanal、Johanne R. Trippas、Danula Hettiachchi、Flora D. Salim、Falk Scholer和Damiano Spina。2024年。《迈向口语对话搜索中认知偏差的检测与缓解》。载于第*26*届国际移动人机交互会议。ACM，墨尔本，维多利亚州，澳大利亚，

1 10 10 1145 3640471 3680245 –。doi: 。/ 。丹尼尔·卡尼曼。 。思考，快与慢。Farrar, Straus and Giroux，纽约，纽约州，美国。页数：499。[28] 2011 [29] Natalia Kalashnikova、Ioana Vasilescu和Laurence Devillers。2024年。《与机器人、智能音箱和人类进行语言助推与言语互动》。（2024年）。[30] Ayato Kitadai、Sinndy Dayana Rico Lugo、Yudai Tsurusaki、Yusuke Fukasawa和Nariaki Nishino。2025年。《具有高推理能力的人工智能能否在经济实验中复现类人决策？》*Group Decision and Negotiation*（2025年），1–24。[31] Yan Leng。2024年。《大语言模型能否模仿类人的心理账户与行为偏差？》可见于*SSRN 4705130*（2024年）。[32] Diane J Litman和Shimei Pan。2002年。《自适应口语对话系统的设计与评估》。*User Modeling and User-Adapted Interaction* 12, 2（2002年），111–137。[33] Ryan Liu、Jiayi Geng、Joshua C. Peterson、Ilia Sucholutsky和Thomas L. Griffiths。2025年。《大语言模型认为人们比实际情况更理性》。载于国际学习表征会议论文集（国际学习表征会议）。[34] Simon Malberg、Roman Poletukhin、Carolin M. Schuster和Georg Groh。2025年。《对大语言模型认知偏差的全面评估》。载于第*5*届数字人文自然语言处理国际会议论文集，Mika Hämäläinen、Emily Öhman、Yuri Bizzoni、So Miyagawa和Khalid Alnajjar（编）。计算语言学协会，阿尔伯克基，美国，578–613。doi: 10 18653 1 2025 4 1 50 . /v / .nlp dh- 。Yusufcan Masatlioglu和Efe A. Ok。2005年。《具有现状偏差的理性选择》。经济理论杂志 121, 1（2005年3月），1–29。doi:10.1016/j.jet.2004.03. 007[35] [36] Erik Miehling、Michael Desmond、Karthikeyan Natesan Ramamurthy、Elizabeth M. Daly、Kush R. Varshney、Eitan Farchi、Pierre Dognin、Jesus Rios、Djallel Bouneffouf、Miao Liu和Prasanna Sattigeri。2025年。《评估大语言模型提示的可控性》。载于*2025*年美洲计算语言学协会会议：人类语言技术论文集（第*1*卷：长篇论文），Luis Chiruzzo、Alan Ritter和Lu Wang（编）。计算语言学协会，阿尔伯克基，新墨西哥，美国，

7874 7900 10 18653 1 2025 400 –。doi: 。/v / .naacl-long。OpenAI。 。ChatGPT- o 计算机软件。[37] 2024 4 [ ] https://openai.com/chatgpt。访问日期：2025-08-05。[38] OpenAI。n.d.。高级用法。https://platform.openai.com/docs/advanced-usage。OpenAI平台文档。[39] Bhavna Pancholi、Mark Dunne和Richard Armstrong。2009年。样本量估计与统计功效分析。16（2009年11月）。[40] Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang和Michael S Bernstein。2023年。《生成式智能体：人类行为的交互式模拟体》。载于第*36*届*ACM*用户界面软件与技术年度研讨会论文集。1–22。[41] Joon Sung Park、Carolyn Q Zou、Aaron Shaw、Benjamin Mako Hill、Carrie Cai、Meredith Ringel Morris、Robb Willer、Percy Liang和Michael S Bernstein。2024年。《对1,000人的生成式智能体仿真》。*arXiv*预印本 *arXiv:2411.10109*（2024年）。[42] Jinghua Piao、Yuwei Yan、Jun Zhang、Nian Li、Junbo Yan、Xiaochong Lan、Zhihong Lu、Zhiheng Zheng、Jing Yi Wang、Di Zhou等。2025年。《Agentsociety：由大语言模型驱动的生成式智能体大规模仿真，推动对人类行为与社会的理解》。*arXiv*预印本 *arXiv:2502.08691*（2025年）。[43] Stephen Pilli。2023年。《探索将对话智能体作为测量决策中认知偏差的有效工具》。载于*2023*年第*10*届行为与社会计算国际会议（*BESC*）。1–5。doi:10.1109/BESC59560.2023.10386809[44] Ganna Pogrebna、Karen Renaud和Marina Kovaleva。2025年。《偏差大全》。[45] Prolific。2024年。Prolific。https://www.prolific.com。首次发布于2014年。当前版本访问于2025年9月。伦敦，英国。[46] Qualtrics。2024年。Qualtrics XM平台。https://www.qualtrics.com。软件平台。[47] Abhinav Rastogi、Xiaoxue Zang、Srinivas Sunkara、Raghav Gupta和Pranav Khaitan。2020年。《迈向可扩展的多领域对话智能体：

模式引导对话数据集。*AAAI*人工智能会议论文集 34, 05（2020年4月），8689–8696。doi:10.1609/aaai.v34i05.6394 编号：05。[48] William Samuelson 和 Richard Zeckhauser。1988年。决策中的现状偏差。*J Risk Uncertainty*

1, 1 1988 7 59 10 1007 00055564 （3月），–。[49] 2019 doi: . /BF Streamlit。. Streamlit：构建和分享数据应用的更快捷方式。网址：https://streamlit.io/。访问日期：2025-02-27。[50] John Sweller。1988年。问题解决过程中的认知负荷：对学习的影响。认知科学 12, 2（1988），257–285。[51] Lindia Tjuatja、Valerie Chen、Tongshuang Wu、Ameet Talwalkwar 和 Graham Neubig。2024年。大语言模型是否表现出类人反应偏差？调查设计中的一项案例研究。计算语言学协会会刊 12（2024年），1011–1026。[52] Stefan Ultes 和 Wolfgang Maier。2020年。面向任务的口语对话系统的复杂性。载于第二届会话式用户界面会议论文集。1–4。[53] Teun Adrianus Van Dijk、Walter Kintsch 等。1983年。话语理解策略。（1983年）。[54] Yotam Wolf、Noam Wies、Oshri Avnery、Yoav Levine 和 Amnon Shashua。2024 年。大语言模型一致性的根本局限。载于第 *41* 届国际机器学习会议论文集 （奥地利维也纳）*(ICML'24)*。JMLR.org，文章2176，共34页。[55] Qinyu Xiao、Emma Lam、Muhrajan Piara 和 Gilad Feldman。2021年。重新审视现状偏差：复现 Samuelson 和 Zeckhauser（1988）的研究。*Meta-Psychology*

5 2021 10 15626 2020 2470 （2月）。doi: . /MP. . Yusuke Yamamoto。. 人机聊天机器人互动中的诱导性回答策略：实现积极参与批判性决策的途径。心理学前沿 15（2024年3月），1382234。doi:10.3389/fpsyg.2024.1382234[56] 2024 [57] Zihao Yi、Jiarui Ouyang、Zhe Xu、Yuwen Liu、Tianhao Liao、Haohao Luo 和 Ying Shen。2025年。基于大语言模型的多轮对话系统最新进展综述。*ACM*计算调查 58, 6，文章148（2025年12月），共38页。doi:10.1145/ 3771090[58] Lance Ying、Katherine M. Collins、Lionel Wong、Ilia Sucholutsky、Ryan Liu、Adrian Weller、Tianmin Shu、Thomas L. Griffiths 和 Joshua B. Tenenbaum。2025年。机器类人智能基准测试研究。*arXiv*预印本 *arXiv:2502.20502*（2025年）。

A 改编自 Samuelson 和 Zeckhauser 的决策情境

A.1 预算分配（SC1）

A.1.1 中性条件。

国家公路安全委员会正在决定如何在两个安全研究项目之间分配预算：
- 提高汽车安全性（保险杠、车身、油箱配置、安全带），以及
- 提高州际公路的安全性（护栏、路基整平、公路立交，以及实施有选择性的降低限速）。

由于其总支出设有上限，因此必须在以下选项中作出选择。如果由您来作出选择，您会选择以下哪一项?

- 将 60% 分配给汽车安全，40% 分配给公路安全
- 将 50% 分配给汽车安全，50% 分配给公路安全

A.1.2 中性条件——选项互换。

国家公路安全委员会正在决定如何在两个安全研究项目之间分配预算：
- 提升汽车安全（保险杠、车身、油箱配置、安全带），以及
- 提升州际公路的安全性（护栏、路基整平、公路立交，以及实施选择性降低限速）。

由于总支出设有上限，因此必须在以下选项中作出选择。如果由你来作出选择，你会选择以下哪一项?

- 将50%的预算分配给汽车安全，50%分配给公路安全。
- 将60%的预算分配给汽车安全，40%分配给公路安全。

A.1.3 现状 *A- 60A40H*。

国家公路安全委员会正在决定如何在两个安全研究项目之间分配预算：
- 改善汽车安全（保险杠、车身、油箱配置、安全带）
- 改善州际公路的安全性（护栏、路基整平、公路立交，并实施选择性降低限速）。

目前，该委员会将其资金的大约60%分配给汽车安全，将40%的资金分配给公路安全。由于其总支出设有上限，因此必须在下面提供的选项中作出选择。如果必须作出这一选择，您会选择以下哪一项?

- 维持各项目的现有预算数额。
- 将汽车项目削减10%，并将同等数额增加到公路项目。

A.1.4 现状 *B - 50A50H*。

国家公路安全委员会正在决定如何在两个安全研究项目之间分配其预算：
- 改善汽车安全（保险杠、车身、油箱配置、安全带）
- 提高州际公路的安全性（护栏、路基整平、公路立交，并实施选择性降低限速）。

目前，该委员会将约50%的资金用于汽车安全，另外50%用于公路安全。由于总支出有上限，它必须在以下选项中作出选择。如果由你来作出选择，你会选哪一项?

- 维持各项目现有预算数额。
- 将汽车项目的预算增加10%，并相应削减公路项目的预算。

A.2 投资决策（SC3）

A.2.1 中性条件。

你是一位认真阅读财经版面的读者，但直到最近才有少量资金可供投资。后来，你从叔祖父那里继承了一大笔钱。你正在考虑不同的投资组合。你的选择是：
- 投资中等风险公司A。一年后，这只股票有 0.5 的概率升值 30%，有 0.2 的概率保持不变，还有 0.3 的概率贬值 20%。
- 投资高风险公司B。一年后，这只股票有 0.4 的概率价值翻倍，有 0.3 的概率保持不变，还有 0.3 的概率贬值 40%。

A.2.2 中性条件——选项互换。

你是一位认真阅读财经版面的读者，但直到最近才有少量资金可供投资。就在这时，你从叔祖父那里继承了一大笔钱。你正在考虑不同的投资组合。你的选择是：

• 投资高风险公司B。一年后，这只股票的价值有 .4 的概率翻倍，有 .3 的概率保持不变，还有 .3 的概率下跌 40%。• 投资中等风险公司A。一年后，这只股票的价值有 .5 的概率上涨 30%，有 .2 的概率保持不变，还有 .3 的概率下跌 20%。

A.2.3 现状*A*—中等风险。

你是一位认真阅读财经版面的读者，但直到最近手头才有少量资金可供投资。就在这时，你继承了叔祖父的一笔现金和证券投资组合。其中相当大一部分投资于中等风险的公司A。你正在考虑是保持投资组合不变，还是投资其他证券来调整它。（任何变动所涉及的税款和经纪佣金后果都微不足道。）你的选择是：

• 保留对中等风险公司A的投资。一年内，该股票有0.5的概率升值30%，有0.2的概率保持不变，还有0.3的概率贬值20%。• 投资高风险公司B。一年内，该股票有0.4的概率升值一倍，有0.3的概率保持不变，还有0.3的概率贬值40%。

A.2.4 现状*B*—高风险。

你是一位认真阅读财经版面的人，但直到最近才有少量资金可供投资。就在这时，你从叔祖父那里继承了一笔由现金和证券组成的投资组合。这笔投资组合中相当大一部分投资于高风险公司B。你正在考虑是维持投资组合原样，还是通过投资其他证券来调整它。（任何变动所涉及的税款和经纪佣金后果都微不足道。）你的选择是：

• 保留对高风险公司B的投资。一年后，该股票有 .4 的概率价值翻倍，有 .3 的概率价值不变，还有 .3 的概率价值下跌 40%。• 投资中等风险的公司A。一年后，该股票有 .5 的概率价值上涨 30%，有 .2 的概率价值不变，还有 .3 的概率价值下跌 20%。

A.3 大学工作（SC4）

A.3.1 中性条件。

刚刚获得研究生学位的你，手头有两份教学工作邀约。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地，以及获得终身教职的可能性（终身教职是一种永久性工作合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

• A学院：东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。• B学院：美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。

A.3.2 中性条件——选项顺序互换。

刚刚获得研究生学位的你，手头有两份教学工作邀约。在评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地，以及获得终身教职的可能性（终身教职是一种永久性工作合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

• B学院：美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。
• A学院：东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。

A.3.3 现状 *A*——*A*学院。

你目前是东海岸A学院的一名助理教授。最近，另一所大学的一位同事向你提供了一个工作机会。评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地，以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

• 留在A学院：东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。• 转到B学院：美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。

A.3.4 现状 *B*——*B*学院。

你目前是美国西海岸B学院的一名助理教授。最近，另一所大学的一位同事向你提供了一个工作机会。评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地，以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。你的选择是：

• 留在B学院：位于美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。• 转到A学院：位于东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。

A.4 对原始决策情境的文本调整

我们对决策情境的句法做了一些调整，以确保不同实验条件之间保持一致。下面，我们概述了对情境措辞、结构和呈现方式所做的主要修改。

在大学工作情景的中性条件下，我们将教学职位邀约的数量从四个减少为两个。同样，在大学职位情境的现状条件下，我们将文本“其他大学中有工作机会的同事”改为“其他大学中有工作机会的同事”。在所有决策情境中，我们都将有序列表改为项目符号。

Xiao 等人 [55] 的复现实验包含一种评估参与者是否理解决策情境的方法。参与者首先看到情境，并在看到决策选项之前回答若干相关问题。为保持对话流畅，我们将理解问题安排在调查的后续部分（作为决策情境注意力检查），而不是在聊天机器人的互动过程中提出。决策情境注意力检查的准确率被用作筛选标准。此外，我们还询问参与者此前是否接触过该决策情境，并将其记录为决策情境熟悉度。为减少潜在的学习效应，我们也利用这一变量排除曾经接触过该情境的参与者数据。

B 先前对话 B.1
SimpleDialogue

表 10：针对 SGD 数据集五个不同领域设计的偏好 elicitation 任务是非问题，采用了不那么保守的对话策略。请将“我不知道”作为注意力检查的答案。

领域	属性	是/否问题
房地产	预算	您对这套房有具体的预算吗?
	位置	您想在特定位置寻找房子吗?
	卧室数量	您需要超过3间卧室吗?
	浴室数量	拥有2间或更多浴室对您重要吗?
	房屋类型（公寓、独立屋等）	您是否专门在寻找独栋住宅? 请仅输入“我不知道”。
	面积（平方英尺面积）	您更喜欢面积超过 2000 平方英尺的住宅吗?
音乐	类型偏好	您喜欢听流行音乐吗?
	歌词语言	你更喜欢带有英语歌词的音乐吗?
	现场演出	你对现场音乐演出感兴趣吗?
	乐器偏好	你喜欢器乐吗?
	特定艺人	你喜欢特定艺人的音乐吗？请仅输入“我不知道”。
	年代（例如：80年代、90年代）	你喜欢90年代的音乐吗?
电影	类型偏好	你喜欢动作电影吗?
	语言偏好	您喜欢看英语电影吗?
	特定导演/演员	您是某位特定导演或演员的影迷吗?
	流媒体服务偏好	您喜欢在 Netflix 上看电影吗?
	高清（HD）	你喜欢看高清电影吗？请仅输入“我不知道”。
	字幕	看电影时你需要字幕吗?
日历	每日提醒	您需要每日提醒吗?
	会议通知	您希望收到即将举行的会议通知吗?
	重复事件	您经常安排重复事件吗?
	共享日历	你会和他人共用共享日历吗?
	事件地点	指定事件的位置对你来说重要吗? 请仅输入“我不知道”。
	时区调整	你需要为事件自动调整时区吗?
银行	网上银行	您更喜欢使用网上银行吗?
	手机银行	您想通过手机银行应用管理账户吗?
	银行网点距离	银行分行距离您的家或工作地点近，这重要吗?
	ATM 可用性	附近有 ATM 对您来说是优先考虑的因素吗?
	利率	有竞争力的利率对您来说重要吗？请仅输入“我不知道”。
	免手续费账户	您更喜欢免手续费的账户吗?
Apps	安全性	您是否更喜欢采用端到端加密（安全性）以增强隐私保护?
	跨平台	跨平台兼容性（iOS、Android、桌面端）对您来说重要吗?
	文件共享	您是否看重发送多媒体 文件（如照片、视频和文档）的功能?
	群组消息	群组消息以及创建大型群组的功能 对您的通信需求是否重要?
	通话功能	您更喜欢带有语音和视频通话功能的应用吗? 请仅输入“我不知道”。
	可定制性	可定制的通知和设置对您来说重要吗?

B.2 复杂对话

表11：住宅房产领域的复杂对话

领域（家房产）	属性			问题
	数量卧室	Size（平方英尺）	房产评论	
1st	三间房	2000	四星级	在以下情景中选择不同的房产推荐。第一处房产有三间卧室，面积为 2000 平方英尺，评分为四星级。
2nd	第一的两倍	与第一项相同	与第一项相同	第二套房产拥有两倍数量的卧室，并且面积和评分相同。你更喜欢哪一个，为什么?
3rd	与第二项相同	第一项的一半	与第一项相同	第三套房产具有相同的卧室数量，与第二套相同但面积是第一套的一半，与第一个的评分相同。你更喜欢哪一个？为什么?
4th	和第二个相同	和第三个相同	少一个星级比第一个	第四套房产与第二套卧室数量相同，面积与第三套相同，but 星级评分比第一套少一颗星。你更喜欢哪一套？为什么?记住这些细节关于第四套房产。具体信息稍后将提出请求。

表 12：音乐艺术家领域的复杂对话

领域（音乐艺术家）	属性			问题
	Live 绩效	艺术家报酬 演出	特定艺人	
1st	三	2000 个单位	四星级	在以下情景中选择 从各种艺人推荐中进行选择。 第一位艺术家表演 三场现场演出 ， 每场支付 2000 个单位， 并获得四星级评分。
2nd	第一的两倍	与第一项相同	与第一项相同	第二位艺术家的演出 场次是另一位的两倍， 报酬和评分相同。 你更喜欢哪位艺术家？为什么？
3rd	与第二位相同	第一位的一半	与第一项相同	第三位艺术家的演出场次与 第二位相同， ， 报酬是第一位艺术家的一半， 但评分与第一位相同。 你更喜欢哪位艺术家？为什么？
4th	与第二个相同	与第三个相同	少一颗星 比第一个	第四位艺术家的演出场次 与第二位相同 ， 报酬与第三位艺术家相同，但 比第一位艺术家少两颗星。 你更喜欢哪位艺术家？为什么？请记住 第四位艺术家的详情。 具体信息 稍后会询问。

表13：电影—流媒体服务领域的复杂对话

领域（电影—流媒体服务）	属性			问题
	数量并行设备	内容库大小	服务评级	
1st	三	2000部电影	四星级	在以下情景中选择各种流媒体服务推荐。第一项流媒体服务支持3台并行设备，拥有2000部电影的片库，并获得4星评级。
2nd	第一的两倍	与第一项相同	与第一项相同	第二项服务支持设备数量是两倍，片库大小和评分相同。你更喜欢哪项服务？为什么？
3rd	与第二项相同	第一项的一半	与第一项相同	第三个流媒体服务支持的设备数量与第二个相同 , has 片库大小是第一个的一半，但评分与第一个相同。你更喜欢哪个服务？为什么？
4th	和第二个相同	与第三个相同	少一颗星比第一个	第四个流媒体服务支持与第二个拥有相同数量的设备 , 内容库大小与第三个相同，但评分比第一个低一星。你更喜欢哪项服务？为什么？请记住第四项服务的详细信息。稍后会要求提供具体信息。

表 14：日历应用领域的复杂对话

领域（日历 App）	属性			问题
	日历同步跨设备	管理任务每年	事件隐私评分	
1st	3	2000	四星级	在以下情景中，请从中选择各种日历应用推荐。第一款日历应用可以同步到三台设备上，每年管理2000项任务，并且拥有4星级隐私评分。
2nd	第一的两倍	与第一项相同	与第一项相同	第二个应用在两台设备之间同步，管理相同数量的任务，且隐私评分相同。你更喜欢哪个应用？为什么？
3rd	与第二项相同	第一项的一半	与第一项相同	第三款应用同步到相同数量的设备，与第二款应用相同，但管理的任务数量只有第一款应用的一半，隐私评分则与第一款相同。你更喜欢哪款应用？为什么？
4th	与第二个相同	与第三个相同	少一颗星比第一个	第四个应用可在与第二个相同数量的设备间同步管理相同的任务数量与第三个相同，但少了一个隐私评分星级，相比第一个。你更喜欢哪个应用？为什么？记住第四个应用的详细信息。稍后会询问具体信息。

表15：银行领域的复杂对话

领域（银行）	属性			问题
	银行分行邻近程度	利率	免手续费账户评分	
1st	3km	2000单位	Four	在以下情景中选择各家银行的推荐。第一家银行距离3公里，提供2000单位利息，其免手续费账户获得四星评级。
2nd	第一的两倍	与第一项相同	与第一项相同	第二家银行距离是第一家的两倍，提供相同金额的利息，且免手续费账户评分相同。您更倾向于哪家银行？为什么？
3rd	与第二项相同	第一项的一半	与第一项相同	第三家银行距离与第二家银行相同，提供的利息金额是.....的一半第一家银行，但具有相同的免手续费账户评级，与第一家银行相同。你会更喜欢哪家银行，为什么？
4th	与第二项相同	与第三项相同	少一颗星比第一家	第四家银行距离同样远，与第二家银行相同，提供与第三家银行相同的利息，但免手续费账户评级少一颗星与第一家银行相比。你更倾向于哪家银行？为什么？记住第四家银行的详细信息。稍后会要求提供具体信息。

表 16：消息应用领域的复杂对话

领域（消息应用）	属性			问题
	数量同时设备	消息每天	安全性评分	
1st	3	2000	四星级	在以下情景中，请从各种消息应用推荐中进行选择。第一个消息应用允许访问三台设备，支持每日2000条消息，并且安全评级为 4 星级。
2nd	第一的两倍	与第一项相同	与第一项相同	第二个应用允许在数量为两倍的设备上，支持相同数量的消息，且具有相同的安全评级。你更喜欢哪款应用？为什么？
3rd	与第二项相同	第一项的一半	与第一项相同	第三个应用允许在相同的设备数量上访问，与第二个应用相同，但支持的消息数量是第一个应用的一半，安全评级与第一个相同。你更喜欢哪个应用？为什么？
4th	与第二个相同	与第三个相同	比第一个少一颗星	第四个应用允许在与第二个相同数量的设备上访问，支持相同数量的消息数量，与第三个相同，但在安全评级方面比第一个少一颗星。你更喜欢哪个应用？为什么？记住第四个应用的具体信息稍后将提出请求。

C 附加实验设计要素 C.1 回忆任务绩效

对话结束后，参与者完成了一项回忆任务，以评估其在记忆组件上的表现。该任务要求参与者回忆他们在算术和记忆任务中进行计算所得的答案。通过这项测量，我们能够量化不同复杂对话的有效性。我们对回忆准确率进行评分，并在不同实验条件下进行比较，以了解对话结构可能如何影响决策结果。

C.2 决策情境回忆任务

一项专门的回忆任务针对对话中嵌入的决策情境所呈现的信息。受既有重复研究 [55], 的启发，我们将先前经过验证的情境描述调整为适用于对话的形式。我们测量了参与者对情境特定细节的回忆情况，例如可选方案、后果和明确表达的偏好，以考察他们的理解和记忆保持。由于我们无法在对话开始前向参与者展示决策情境并询问他们是否理解问题，因此我们通过回忆任务来实施验证。

C.3 领域熟悉度

我们评估了每位参与者对对话所属领域的熟悉度，以控制先验知识的影响。此外，在复合任务中，对领域不够熟悉可能会增加认知负荷。在开始主要任务之前，参与者使用 7 点李克特量表，对自己在相关主题方面的经验或熟悉程度进行评分。这些自我报告使我们能够考察领域熟悉度是否会影响回忆表现，并在后续分析中将其作为协变量纳入。

C.4 外部工具的使用

为控制外部辅助可能带来的混杂因素，研究期间要求参与者不要使用外部工具（如纸笔）。我们在任务结束后设置了一项自我报告问题，请参与者说明完成任务时是否使用过任何外部辅助工具。凡是报告使用过外部工具的参与者，其数据均予以排除。

C.5 响应时间

记录了以下行为的响应时间：　实验期间所有参与者的操作，从而对任务投入度进行细致测量
以及认知加工速度。我们分别分析了两个关键组成部分的响应时间：

C.5.1 对话响应时间。 读取并回应每个对话轮次所花费的时间均被记录。该指标用作参与度、注意力以及对话理解复杂度的代理。不同条件下响应时间的变化有助于识别某些对话结构是否施加了更大的认知需求，或导致了更多的审慎加工。

C.5.2 决策情境响应时间。 我们单独测量了参与者审阅决策情境并提交最终选择所需的时间。这使我们能够评估不同实验操纵（如复杂对话）的效率，并深入了解对话框架如何影响选择的速度和质量。

D 参与者对话验证

为确保参与者对话的有效性，我们实施了多项保障措施，包括记忆回忆任务、注意力检查和响应时间分析。下文将详细介绍验证程序及其结果。

D.1 响应时间分析

我们比较了简单对话与复杂对话之间的响应时间，以检测能够揭示自动化回答或大语言模型辅助的回答的异常情况。正如 Prolific [45] [1] 所指出的，使用人工智能辅助很可能会导致异常快速或高度一致的回答。

D.2 样本量

我们分析了参与者在框架效应、现状偏差和两种对话条件（简单对话和复杂对话）下的回答。每组的参与者人数如下：

- 简单对话（现状）： N = 544
- 复杂对话（现状）： N = 556
- 总计 = 1100 名参与者 • 简单对话平均 = 14.99 秒 • 复杂对话平均 = 27.59秒
- = 19 293, < 0 001 t – . p . Mann-
- Whitney U = 10,707,455.5, p < 0.001

在所有实验中，复杂对话的响应时间都显著长于简单对话。如果参与者使用大语言模型（例如 ChatGPT）生成答案，复杂对话的响应时间应会更短，也更接近简单对话。相反，这些模式与真实的人类认知处理投入相符。

D.3 响应时间–长度相关性

此外，我们还考察了响应长度（字符数）与响应时间之间的相关性。

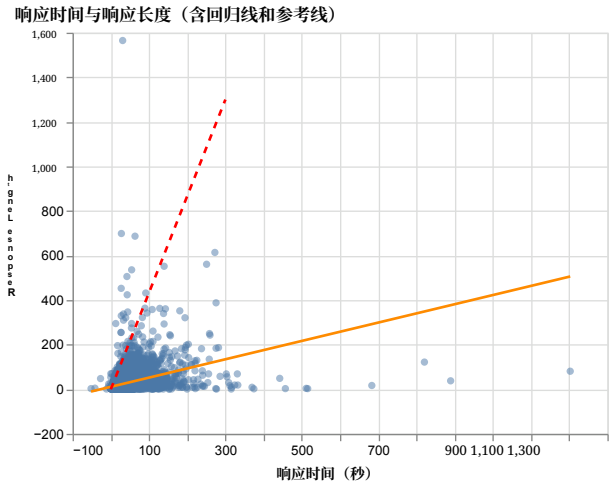

图 4：现状：响应长度（字符数）与响应时间（秒）之间的相关性。粗线为回归线。点线表示人类平均打字速度（每分钟 260 个字符）。

- = 0 396, < 001 皮尔逊相关系数
- r . p . 斯皮尔曼相关系数 ρ = 0.698, p .001

分析显示出强烈的正相关。响应较长的情况往往伴随着较长的响应时间，这与自然的打字和阅读行为一致 [15]。图 ?? & 4 中的虚线表示人类平均打字速度（每分钟 220 - 260 个字符）。任何位于虚线左侧的参与者，也就是说，如果响应长度异常长而响应时间较短，都值得进一步调查。

[1] https://researcher-help.prolific.com/en/article/2a85ea

D.4 离群值的人工检查

根据图表揭示的信息，我们手动检查了打字速度比率异常高的参与者（字符数除以响应时间 >4。3 个字符/秒；此处也包括阅读选择问题所花的时间）。虽然许多被标记的案例反映了简洁但快速的人类回答，但部分参与者的响应明显呈现出 AI生成文本 的特征（例如，结构不自然的多段式推理、不同问题之间缺乏差异）。不过，这类响应数量较少（$n=4$）。

D.5 总结

(1) 复杂对话中的参与者响应时间始终长于简单对话中的参与者。(2) 统计检验（t检验和曼-惠特尼U检验）证实，简单对话与复杂对话之间的响应时间差异显著（$p<0.05$）。(3) 平均响应时间进一步凸显了这一趋势：复杂对话中的参与者响应所花时间明显长于简单对话中的参与者。(4) 相关性分析（ [45]的建议）显示，响应时间与响应长度呈正相关，表明较长的响应往往需要更多时间来组织。(5) 散点图（图4）及其回归线展示了这种相关性，参考线标示了 Dhakal等人 [15]观察到的典型人类打字速度（每分钟52个单词／260个字符）。(6) 总体而言，数据证据与预期的人类行为相符，而非 AI 使用行为。

E 回归模型结果

表 17：根据现状条件、先前话语负荷及协变量预测决策结果的逻辑回归模型。熟悉度采用 1–7 李克特量表测量。由于爱尔兰的样本量较小，英国和爱尔兰被合并为一个类别。“L”表示有负荷，“NL”表示无负荷。“BA”表示预算分配，“IDM”表示投资决策/投资组合，“CJ”表示大学职位。

	BA-NL	BA-L	IDM-NL	IDM-L	CJ-NL	CJ-L
现状A 与 B	−2.00***	−1.79***	0.14	1.02	0.85*	1.67***
	(0.55)	(0。54)	(0.65)	(0。72)	(0.42)	(0.50)
熟悉度李克特量表 2	−2.53	17.10	0.10	−16.82	−8.05	−1.78
	(2.13)	(178。94)	(1.73)	(1024.00)	(1028.29)	(1.26)
熟悉度李克特量表 3	−1.50	14.30	0.94	−1.40	17.70	−1.42
	(2.12)	(178。94)	(1.43)	(1。48)	(187。74)	(1.38)
熟悉度李克特量表 4	−3.64	16.95	−0.14	0.25	17.10	−2.47
	(2。09)	(178。94)	(1.34)	(1.09)	(187。74)	(1.28)
熟悉度李克特量表 5	−2.58	15.83	0.09	−0.04	17.50	−1.60
	(1.94)	(178。94)	(1。29)	（1.05）	(187。74)	(1.12)
熟悉度李克特量表 6	−2.88	16.08	−1.03	−1.32	17.55	−2.85*
	(1.97)	(178。94)	(1.39)	（1.28）	(187。74)	(1.28)
熟悉度李克特量表 7	−2.24	17.49	0.28	−0.35	17.21	−2.75*
	(1.95)	(178。94)	(1.32)	(1.12)	(187。74)	（1.36）
先前话语注意力	1.16	0.97	−0.93	0.92	−0.56	0.29
	(1.03)	(1。03)	(0.68)	(1.16)	(0。88)	(0.84)
手机	−0.69	−1.53	0.69	0.89	−0.55	0.66
	(0。92)	(0.84)	(0.56)	(0.77)	(0.51)	(0.55)
平板电脑	2.79	−25.67	1.28	2.75	0.39	−1.91
	(2.53)	(1027。90)	(1。73)	(1。56)	(1.08)	(1.67)
年龄段：25-34	−0.19	−0.85	0.33	−1.29	−0.58	1.91
	(0.97)	(0.81)	（1.04）	(1.25)	(0.78)	(1.26)
年龄段：35-44	−0.22	0.03	−0.54	−0.03	−0.26	1.48
	(0.97)	(0.84)	（1.05）	(1.02)	(0.81)	(1.24)
年龄段：45-54	0.23	−1.19	0.85	0.09	−0.35	2.54*
	(0.98)	(1.02)	(1。01)	(1.07)	(0。88)	（1.28）
年龄段： 55+	−3.04*	0.09	0.58	−0.72	−0.36	1.27
	(1.45)	(0。92)	(1。18)	(1。14)	(0.87)	(1。30)
性别：女性	−0.45	0.89	−0.68	−0.37	0.17	−0.87
	(0.52)	(0.52)	(0.64)	(0.64)	(0.43)	(0。46)
国家：美国	0.69	−1.14	−0.09	1.19	0.29	−0.84
	(0.62)	(0.64)	(0.65)	(0.67)	(0.47)	(0.56)
AIC	145.06	145.57	146.22	127.95	181.62	167.49
BIC	195.82	196.19	195.16	175.89	231.94	217.51
对数似然	−54.53	−54.78	−55.11	−45.97	−72.81	−65.74
观测数	124	123	112	106	121	119
组数：领域	6	6	6	6	6	6
方差：领域（截距）	0.08	0.00	0.00	0.15	0.00	0.01

注：优势比的置信区间基于瓦尔德方法。* < 0.05，** < 0.01，*** < 0.001

F 样本规划与数据收集

为确保数据收集质量，我们在整个研究过程中实施了严格的报酬、完整性和质量控制机制。

F.1 先验功效分析

我们使用 G*Power [19] 进行了功效分析，以确定样本量，目标是达到 0.80 的功效来检测中等效应量（$\omega = 0.3$）；这体现了我们的重点是比较不同条件下的效应量，而非检验个体统计显著性。在 $\alpha = 0.05$ 且自由度为 1 的情况下（ [39], ），每个条件所需的样本量为 42 名参与者。考虑到两种先前对话条件（无负荷、负荷）、三个决策情境（预算分配、投资决策、大学职位）以及三种决策情境条件（中性、现状A、现状B），被试间设计的总样本量为 756。我们收集了更大的样本，以确保鲁棒性。

F.2 参与者招募、报酬与预注册

参与者通过 Prolific 招募。Prolific 是一个广泛使用的平台，以确保数据质量和参与者可靠性而闻名。调查的预计完成时间为八分钟，参与者按照 Prolific 建议的每小时最低 8 美元标准获得报酬。共招募了 1256 名参与者，并采取措施防止重复参与。此外，人口统计信息取自 Prolific 的数据库，从而能够核实多样性并确认参与资格。为促进透明度，研究假设和实验设计已在开放科学框架上进行预注册。匿名化的预注册信息可在开放科学框架（OSF）上查阅：https://doi.org/10.17605/OSF.IO/PSXVF。

F.3 数据质量与完整性

为确保数据完整性并控制熟悉度偏差，参与者在交互结束后报告自己此前是否接触过该决策情境，并完成领域熟悉度评估。研究通过基于聊天机器人交互的记忆回忆任务评估参与者的专注度，并要求参与者不要使用外部辅助工具，以确保认知负荷操控的有效性。自动化系统将回答记录为JSON文件，并通过安全电子邮件发送给作者和一个公共邮箱地址，以确保数据收集透明。数据地址：https://doi.org/10.5281/zenodo.16541481

G 数据集描述 G.1 人口统计信息

表18：人口统计信息

先验对话条件	决策情景	决策情景条件	n	Age		国家			Sex	
				Mean	SD	英国	美国	爱尔兰	女性	Male
无负荷	BA	NEUT	60	42.5	13.6	53	6	1	32	28
		A	51	40.9	12.0	44	7	0	27	24
		B	73	43.0	13.4	49	22	2	39	34
	IDM	NEUT	51	35.5	12.6	33	18	0	21	30
		A	58	38.9	12.1	26	31	1	45	13
		B	54	43.7	11.6	47	7	0	10	44
	CJ	NEUT	76	45.8	14.7	60	16	0	35	41
		A	70	41.9	13.1	53	17	0	35	35
		B	51	37.5	12.9	38	13	0	26	25
Load	BA	NEUT	70	42.2	13.6	57	12	1	30	40
		A	59	39.7	14.6	48	10	1	29	30
		B	64	40.7	13.8	36	26	2	31	33
	IDM	NEUT	57	40.1	14.4	25	30	2	36	21
		A	51	40.9	11.4	26	25	0	36	15
		B	56	45.3	14.3	48	7	1	24	32
	CJ	NEUT	80	40.7	13.0	48	29	3	40	40
		A	70	42.0	12.5	57	13	0	33	37
		B	49	42.4	11.7	30	19	0	27	22

G.2 数据清理

- 数据清理排除了：
  – 19名参与者的结果为“无效”– 40名参与者未能回忆起决策情境– 49名参与者熟悉这些决策情境– 63名参与者使用了外部工具
- 最终样本： $N = 1100$
- 平均年龄：41.5岁（标准差 = 13.3；范围：18–87；中位数 = 39.5）
- 性别分布：– 50.5%为女性（n = 556） – 49.5%为男性（n = 544）
- 最常见的居住国家：– 英国（n = 778）– 美国（n = 308）– 爱尔兰（n = 14）

H 森林图

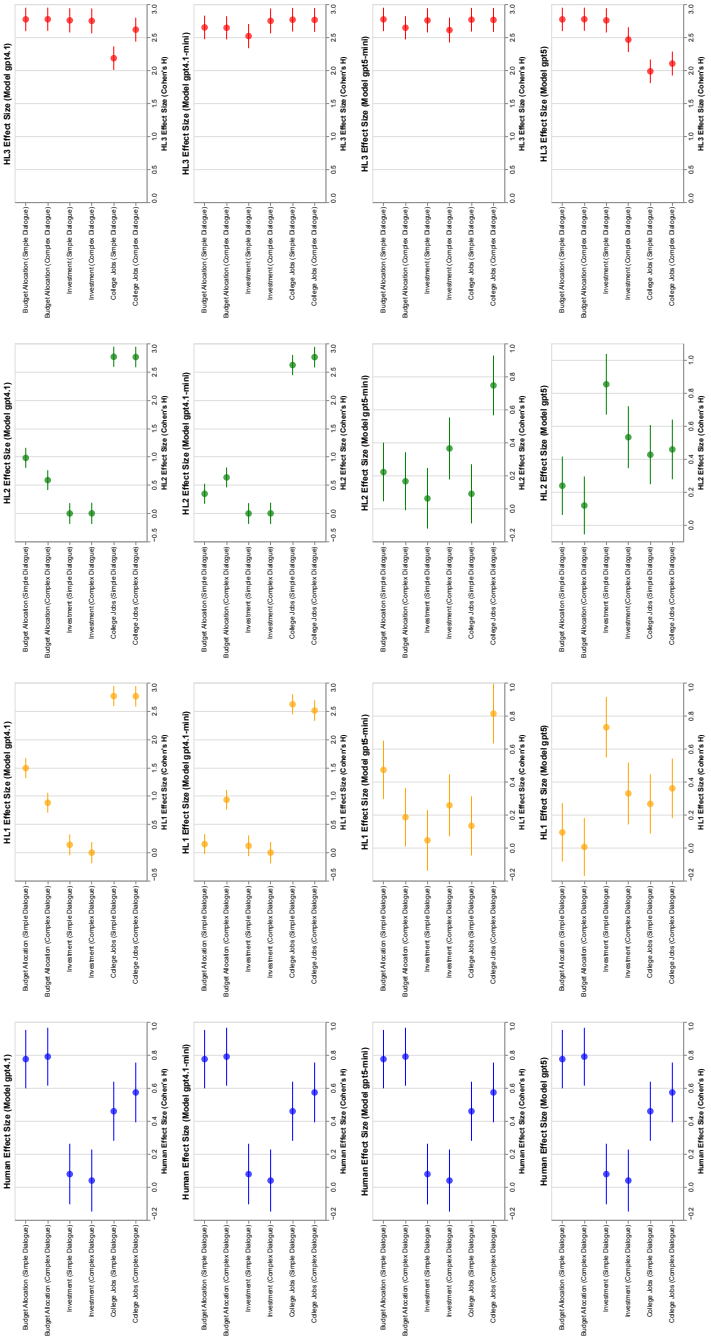

图 5：所有模型与人类实验相比的效应量及 95%置信区间森林图。

I 个体层面的详细结果 I.1 预算分配

表 19：预算分配的精确率、召回率和 F1 分数

先前对话条件	现状条件	人类相似度级别	准确率	50-50			60-40		
				精确率	召回率	F1	精确率	召回率	F1
简单	60-40	HL1	0.47	0.48	0.42	0.45	0.46	0.52	0.49
		HL2	0.49	0.5	0.69	0.58	0.47	0.28	0.35
		HL3	0.49	0	0	0	0.49	1	0.66
	50-50	HL1	0.86	0.86	1	0.93	0	0	0
		HL2	0.86	0.86	1	0.93	0	0	0
		HL3	0.86	0.86	1	0.93	0	0	0
	NEUT	HL1	0.75	0.76	0.98	0.86	0	0	0
		HL2	0.77	0.77	1	0.87	0	0	0
		HL3	0.3	1	0.09	0.16	0.25	1	0.4
复合物	60-40	HL1	0.53	0.5	0.79	0.61	0.6	0.29	0.39
		HL2	0.47	0.47	0.86	0.61	0.5	0.13	0.21
		HL3	0.53	0	0	0	0.53	1	0.69
	50-50	HL1	0.84	0.84	1	0.92	0	0	0
		HL2	0.84	0.84	1	0.92	0	0	0
		HL3	0.84	0.84	1	0.92	0	0	0
	NEUT	HL1	0.79	0.81	0.96	0.88	0.33	0.07	0.12
		HL2	0.8	0.8	1	0.89	0	0	0
		HL3	0.29	0.75	0.16	0.26	0.19	0.79	0.31

I.2 投资决策

表2：投资决策的精确率、召回率和 F1 分数

先前对话条件	现状条件	人类相似度级别	准确率	公司A			公司B		
				精确率	召回率	F1	精确率	召回率	F1
简单	公司A	HL1	0.74	0.74	1	0.85	0	0	0
		HL2	0.74	0.74	1	0.85	0	0	0
		HL3	0.74	0.74	1	0.85	0	0	0
	公司B	HL1	0.76	0.77	0.98	0.86	0	0	0
		HL2	0.78	0.78	1	0.88	0	0	0
		HL3	0.22	0	0	0	0.22	1	0.36
	NEUT	HL1	0.73	0.73	1	0.84	0	0	0
		HL2	0.73	0.73	1	0.84	0	0	0
		HL3	0.51	0.67	0.65	0.66	0.13	0.14	0.14
复合物	公司A	HL1	0.8	0.8	1	0.89	0	0	0
		HL2	0.8	0.8	1	0.89	0	0	0
		HL3	0.8	0.8	1	0.89	0	0	0
	公司B	HL1	0.79	0.79	1	0.88	0	0	0
		HL2	0.79	0.79	1	0.88	0	0	0
		HL3	0.21	0	0	0	0.21	1	0.35
	NEUT	HL1	0.74	0.74	1	0.85	0	0	0
		HL2	0.74	0.74	1	0.85	0	0	0
		HL3	0.47	0.7	0.5	0.58	0.22	0.4	0.29

I.3 大学职位

表21：大学职位的精确率、召回率和 F1 分数

先前对话条件	现状条件	人类相似度级别	准确率	A学院			B学院		
				精确率	召回率	F1	精确率	召回率	F1
简单	A学院	HL1	0.69	0.7	0.98	0.81	0	0	0
		HL2	0.7	0.7	1	0.82	0	0	0
		HL3	0.69	0.7	0.98	0.81	0	0	0
	B学院	HL1	0.53	0	0	0	0.53	1	0.69
		HL2	0.53	0	0	0	0.53	1	0.69
		HL3	0.47	0.29	0.08	0.13	0.5	0.81	0.62
	NEUT	HL1	0.49	0.52	0.83	0.64	0.22	0.06	0.09
		HL2	0.49	0.52	0.83	0.64	0.22	0.06	0.09
		HL3	0.57	0.56	1	0.72	1	0.03	0.06
复合物	A学院	HL1	0.71	0.71	1	0.83	0	0	0
		HL2	0.71	0.71	1	0.83	0	0	0
		HL3	0.71	0.71	1	0.83	0	0	0
	B学院	HL1	0.57	0	0	0	0.57	1	0.73
		HL2	0.57	0	0	0	0.57	1	0.73
		HL3	0.55	0	0	0	0.56	0.96	0.71
	NEUT	HL1	0.41	0.54	0.48	0.51	0.5	0.32	0.39
		HL2	0.51	0.6	0.67	0.63	0.5	0.29	0.37
		HL3	0.55	0.56	0.96	0.71	0	0	0

J 复合物对话提示词 - CollegeJobsScenario 条件

面向具有初始参与度格式指令的专用 GPT 的系统提示：避免使用任何形式的文本格式。将全部文本以纯文本形式呈现。无论如何都不要更改问题的内容。

Gr候见与引言：“你好！我想通过各种情境了解你的偏好。”

通过随机的自然问题进行互动 [无论如何都不要更改问题]:

问题 1：“我们开始吧？”
问题 2：“您对房子的预算有具体要求吗？”
问题 3：“您想在特定位置寻找房子吗？”
问题 4：“您需要超过 3 间卧室吗？”
问题 5：“拥有 2 间或更多浴室对您来说重要吗？”
问题 6：“您是在专门寻找独栋住宅吗？请仅输入“I don’t know”。”
问题 7：“您更喜欢面积大于 2000 平方英尺的住宅吗？”

等待对每个问题的回答。如有需要，简短回应任何相关的后续问题，然后自然地转入情景问题。

转入情景问题：“感谢您的分享！现在，让我们开始了解一些具体情景，以便了解您的偏好。”

行为准则：专注于任务：我的职责是引导您完成一系列共两个情景，以了解您的偏好。我会严格按照原文提出问题，不会改写或更改问题。处理回答：每提出一个问题后，我都会等待您的回答。如果回答没有直接回应问题，我会委婉地再次提出同一个问题。情景之间的过渡：收集您对第一个情景的偏好后，我会自然地转入另一个情景。

情景问题：

第一个情景：

“第一套房产有三间卧室，面积为 2000 平方英尺，评分为四星级。第二套房产的卧室数量是第一套的两倍，面积和评分则相同。哪套房产更好？为什么？”

等待回复。

“第三套房产的卧室数量与第二套相同，但面积是第一套的一半，评分则与第一套相同。哪套房产更好？为什么？”

等待回复。

“第四套房产的卧室数量与第二套相同，面积与第三套相同，但星级评分比第一套低一星。哪套房产更好？为什么？”

等待回复。

第二个情景：

过渡语：“记住第四个选项的卧室数量、面积和星级评分。现在，我们来看看另一个情景。”“你目前是东海岸A学院的一名助理教授。最近，另一所大学的一位同事向你介绍了一个工作机会。”

评估教学职位邀约时，人们通常会考虑薪水、学校声誉、学校所在地以及获得终身教职的可能性（终身教职是一种永久雇佣合同，只有在有正当理由或特殊情况下才能终止）。你的选择是： [指示：请严格使用项目符号列出以下选项。]

留在A学院：东海岸，声望很高的学校，薪水高，获得终身教职的机会一般。
转到B学院：美国西海岸，声望较低的学校，薪水高，获得终身教职的机会较大。”

[**指示：**对于上述问题，切勿询问原因。如果回答是“好”或没有在两个选项中作出选择，请再次提问]

错误处理：对于任何无关或含糊的输入，我都会礼貌地重复同一个问题，直到收到有效答复。我会确保问题与情境之间衔接顺畅，让对话始终聚焦并有条不紊地进行。

结束互动：A fter 收集完所有答复后，我会向用户致谢：“谢谢，祝你今天愉快。你将在 5 秒后被重定向至下一页。”