心理学报 2026, Vol. 58, No. 3, 416–436 https://doi.org/10.3724/SP.J.1041.2026.0416
Acta Psychologica Sinica https://cstr.cn/32110.14.2026.0416

能辨“单次–多次博弈”的大语言模型: 理解与干预风险决策*

周 蕾[1] 李立统[1] 王 旭[1] 区桦烽[1]
胡倩瑜[1] 李爱梅[2] 古晨妍[1]

([1]广东工业大学管理学院, 广州 510520) ([2]暨南大学管理学院, 广州 510632)

摘 要 风险决策的理论研究主要依赖行为结果的逆向推理和自我报告数据, 缺乏对决策过程的直接观测, 制约了其内在机制解释及有效行为干预方案开发。人工智能大语言模型(LLMs)的运用为克服以上局限提供了途径。本文通过三项研究系统考察了 LLMs 在风险决策中的模拟潜力, 基于 DeepSeek-R1 进行单次和多次博弈并生成决策依据, 并运用 GPT-4o 对其进行归纳性主题分析(ITA), 构建了 LLMs 生成决策策略文本的技术路径, 并将其用于决策干预。发现: (1) ChatGPT-3.5/4 能复现人类单次(更风险规避)与多次(更风险寻求)博弈的典型选择模式; (2) LLMs 能分清单次/多次博弈逻辑, 并正确分别运用规范性和描述性理论生成相应策略, 其策略被认可度高; (3) LLMs 基于不同策略生成的干预文本能有效影响人们在医疗、金融、内容创作和电商营销情境中固有的风险决策偏好。研究系统验证 LLMs 对行为偏好的模拟能力, 对决策的理解力, 并构建了基于生成式 AI 的决策干预新范式, 为人工智能辅助高风险决策提供了理论和实践基础。

关键词 风险决策, 单次/多次博弈, 大语言模型, 决策策略, 干预

分类号 B849: C91

1 引言

风险决策是行为研究的核心议题, 对人类生存发展至关重要。其理论发展主要分为规范性理论(normative theory)和描述性理论(descriptive theory)两大路径(Benartzi & Thaler, 1999), 单次与多次博弈范式为区分这两类理论提供了有效方法。然而, 现有理论研究多依赖行为结果的逆向推理和自我报告数据, 缺乏对决策过程的直接观测, 导致对内在机制的解释力不足, 也制约了行为干预的效果。近年来, 人工智能(Artificial Intelligence, AI)的发展为风险决策研究提供了新的视角, 特别是通过 AI 模拟人类决策的选择偏好并将抽象的决策策略外化。但随之产生的问题是: AI 生成的决策策略是否被人类认可? 这些策略能否干预人类决策? 通过回答上述问题, 不仅能揭示 AI 对人类决策偏好的解释力, 为决策模型构建提供新证据, 还能为基于 AI 的决策助推提供理论依据。本文基于人工智能大语言模型(large language models, LLMs)开展了三项研究, 系统考察 LLMs 对风险决策选择偏好及过程机制的模拟能力, 探究其对人类决策策略的理解, 构建运用 LLMs 生成针对单次与多次风险决策偏好的干预文本的方法, 并通过实验验证其干预效果。

1.1 单次与多次博弈的风险决策行为差异

风险决策指人们面临的选项具有多个结果, 且各个结果的发生概率已知的决策(Kahneman & Tversky, 1979)。其理论发展可分为规范性(normative)和描述性(descriptive)理论两大路径: 规范性理论

收稿日期: 2025-05-12

* 国家自然科学基金面上项目(72271066, 72571114), 国家自然科学基金青年科学基金项目(72102085), 广州市科技计划基础与应用基础研究专题项目(2025A04J5384)资助。

通信作者: 李爱梅, E-mail: tliaim@jnu.edu.cn; 古晨妍, E-mail: g_cy1989163@163.com

模型大多基于数学期望(mathematical expectation)假设, 遵循“加权求和”的期望法则(Von Neumann & Morgenstern, 1947)。描述性理论模型更多从心理学视角出发, 关注实际决策行为中的偏差和启发式(Hertwig & Erev, 2009)。

现实中的风险决策可分为单次或多次博弈场景。例如, 企业家可能在一定时期内遇到一个或多个相似的投资机会, 医生可能治疗一个或多个具有相似问题的患者(Redelmeier & Tversky, 1992)。研究发现, 人们在单次和多次博弈情境中常出现行为偏好反转(Anderson et al., 2023)。例如, Sun 等人(2014)设计了一项金融应急管理任务, 要求被试在两种应急方案中进行选择: 方案 A. 确定获得¥5000; 方案 B. 0.1%概率获得¥6, 000, 000, 99.9%概率获得¥0。被试需要分别在单次(方案实施 1 次)和多次(方案实施 100 次)情境中做出选择。结果发现, 人们在单次博弈中更偏好确定方案 A, 而在多次博弈中则更偏好风险方案 B。这种选择偏好差异在多项研究被反复验证(Barberis & Huang, 2009; Kumar & Lim, 2008; Lopes, 1996)。

单次和多次博弈的选择偏好可分别由描述性和规范性理论进行解释。根据经典规范性理论——期望效用理论(Expected Utility Theory)假设, 在多次博弈中, 个体倾向于通过各选项发生概率对结果加权求和得出期望值, 并选择最大期望值的选项(Von Neumann & Morgenstern, 1947)[1]。而在单次博弈中, 人们的选择常违背期望效用理论假设, 更符合描述性理论假设。如短视损失厌恶理论(Myopic Loss Aversion)提出, 在单次博弈中, 个体对损失的过度敏感使得他们倾向规避风险, 选择确定选项(Tversky & Bar-Hillel, 1983)。又如, 齐当别理论(equate-to-differentiate theory, Li, 2004)指出, 人们在决策时会把差异小的维度(例如, 概率维度)齐同处理, 而只依据差异大的维度(例如, 结果维度)进行。

传统风险决策理论研究存在以下局限: 首先, 主要基于对行为结果的逆向推理, 缺少对决策过程的系统性直接观测; 其次, 对决策策略的探讨依赖自我报告数据和选择结果, 导致对决策内在机制解释力不足。这不仅限制了理论对现实复杂决策情境的解释深度(Kelton et al., 2010; Morreale et al., 2018), 也制约了有效行为干预方案的开发。

1.2　人工智能助力决策理论研究

AI 的发展和运用, 为心理学研究提供了新的视角与途径。作为最为前沿的 AI 工具, LLMs 在多个领域展现出“类人”特征(焦丽颖 等, 2025; Salles et al., 2020)。在行为层面, LLMs 常直接被作为实验研究被试(Grossmann et al., 2023)。例如, Qin 等人(2024)利用 ChatGPT 以 93.2%的准确率成功复制了包括领导行为与职场互动在内的 22 项社会心理学实验结果; Argyle 等人(2023)发现 GPT-3 在社会科学研究中具有模拟不同人类子群体的能力, 能够生成在形式、语义与结构上高度类人的回答; Strachan 等人(2024)指出 LLMs 的心理理论(theory of mind)的能力可媲美人类; 在过程机制层面, 多个研究发现 LLMs 在特征提取路径和层级处理机制方面和大脑的神经反应模式趋同(Mischler et al., 2024), 体现了 LLMs 的语言加工机制与人脑的相似性。

在决策研究领域, LLMs 能模拟人类的决策偏好。例如, Zhao 等人(2024)发现, ChatGPT-4 在情绪激发下能够调整其风险决策行为, 尤其是在恐惧情绪激发时, 表现出较低的风险偏好; Aher 等人(2023)发现 ChatGPT 在囚徒困境、最后通牒博弈和信任博弈中表现出类人的合作行为, 此外, 也有研究发现 LLMs 在投票行为、经济博弈以及经典社会实验中也表现出高度拟人化(Dillion et al., 2023); Jones 和 Steinhardt (2022)揭示了 LLMs 在生成开放式文本时容易受到无关信息和初始输入的影响, 导致其输出偏离预期, 表现出类似于人类的框架效应和锚定效应; Park (2024)指出 LLMs 在“少即是多效应”(Less-is-better effect)实验中与人类相似, 更倾向选择低价商品。

此外, LLMs 在决策过程层面也展现出“类人性”。例如, 研究发现, LLMs 生成单词时通过重读上下文进行链式推理, 类似人类在解决数学问题或写文章时使用便签辅助思考或迭代完善论点的做法(Hagendorff et al., 2023)。此外, 随着模型复杂度的提升, LLMs 能够减少推理偏差, 表现出类似人类的系统 2(system 2)思维(Wei et al., 2022)。进一步地, Binz 和 Schulz (2023)通过两步任务(two-step task)实验揭示, GPT-3 表现出与人类相似的基于模型的强化学习(model-based reinforcement learning)特征, 即根据环境状态和奖励信息, 调整自己的行为策略以实现奖赏最大化。类似地, Webb 等人(2023)通过四种类比推理任务, 发现 GPT-3 表现出了与人

[1] 同理, 在 Sun 等人(2014)文中的应急方案：选项 A 的期望值为¥5000, 小于选项 B 的期望值¥6000, 因此在多次博弈中人们更偏好期望值更大的选项 B。

类相当甚至更强的推理能力, 且在类比推理和抽象模式识别上存在一定的相似性。而 Goli 和 Singh (2024)发现 GPT 在进行跨期决策时, 会和人一样考虑多个因素, 特别是关于风险与不确定性、机会成本与投资选择、紧迫性与即时需求这三个维度。

总结而言, LLMs 可模拟人类决策的选择偏好及过程机制, 因而有助于揭示认知“黑箱”中的关键步骤。此外, LLMs 这种能力也有助于将抽象理论转化为具体策略, 便于将决策逻辑具象化。以上优势为突破传统决策理论的研究局限提供了新的路径。

1.3 人工智能干预决策行为

LLMs 凭借其强大的内容生成能力和多模态交互能力, 能够有效地干预人们的行为并提升决策质量(章彦博 等, 2025; Gupta et al., 2024)。以医疗决策为例, ChatGPT 等 AI 聊天助手被验证对提升疫苗健康素养和接种信心具有显著效果, 且能有效纠正公众对疫苗的认知偏差, 并生成高质量的公共健康传播或疫苗倡导材料(Altay et al., 2023; Deiana et al., 2023; Karinshak et al., 2023; Xia et al., 2025)。此外, LLMs 干预不仅可作用于行为, 还能重塑个体的认知加工路径: 例如, Ferguson 等人(2023)发现 GPT-3 生成的微妙性别歧视解释文本, 能显著调整人们的理解和表达, 包括语言风格与论证角度。

与传统干预方式相比, LLMs 的干预具有以下优势: (1) 个性化程度高: 传统干预方法通常采用静态的“一刀切”策略, 易受到个体差异和情境等因素干扰(Jin & Han, 2014; Pavey & Churchill, 2014; Peters & Levin, 2008); LLMs 则可根据个体的行为和心理特征动态调整策略, 实现“千人千面”的个性化精准干预(Xia et al., 2025)。(2) 智能化程度高: 传统干预主要依赖人为设计与预设信息, 缺乏对复杂情境的信息整合与实时响应能力; LLMs 能实时整合多源数据, 自动化生成更具针对性和说服力的文本(Choi et al., 2025)。(3) 效果更稳定: 传统干预的适用性和稳定性存在争议。例如, Peng 等人(2021)指出消极框架更能提升疫苗接种意愿, 但 Liu 等人(2019)则认为积极框架更具说服力; LLMs 通过大规模语料训练获得强推理能力(de Varda et al., 2025; Strachan et al., 2024), 生成的干预文本可读性强、逻辑连贯, 具有稳定的跨情境适用性。

注意, 尽管 LLMs 可以根据指令生成干预性文本, 但由于其工作原理是基于信息输入去概率性地输出词语, 其效果高度依赖输入信息的质量(de Kok, 2025; Khalid & Witmer, 2025; Lin, 2024, 2025)。此外, 当缺乏足够的上下文或背景信息时, LLMs 容易生成不正确或虚构的信息, 因而需要通过示例、指令或微调来补充专业知识(Lin, 2023)。

为了克服以上局限, 可运用 LLMs 进行归纳性主题分析(Inductive Thematic Analysis, ITA)方法提高干预文本质量(Zhang et al., 2024)。相较于传统的 ITA 方法(如人工提取、LDA 及 BERTopic), 基于 LLMs 的 ITA 成本更低、对上下文理解力更强, 且能有效处理非结构化数据等缺陷(Prescott et al., 2024)。目前, 在医疗健康(Deiner et al., 2024)、教育(Katz et al., 2024)、交通(Carvalho et al., 2024)和顾客满意度研究(Dai et al., 2023)等领域展现出良好的应用前景。

可见, 人工智能, 特别是 LLMs 是决策支持的重要工具, 已在多个领域展现出显著的决策优化效果。

1.4 研究目的

综上, 传统的风险决策研究缺少对决策过程的系统性观测, 限制了理论发展和行为干预的效果。运用人工智能有助于厘清决策背后的策略和逻辑, 并基于此生成干预方案来辅助决策。本研究基于经典的单次/多次博弈的风险决策范式, 首先揭示 LLMs 对风险决策偏好和过程的模拟和理解能力, 然后探索如何应用 LLMs 对决策策略的理解, 生成决策干预文本, 并验证其对人类决策的干预效果。

本文包含三个研究: 研究 1 通过经典的单次与多次赌博游戏任务, 考察 LLMs (ChatGPT-3.5/4)模拟人类决策的表现。研究 2 首先构建 LLMs 决策策略文本的生成方法: 利用 DeepSeek-R1 生成决策依据, 再通过 GPT-4o 进行归纳性主题分析(ITA)并生成决策策略文本。其次, 运用 Sun 等人(2014)研究中的医疗和金融情境, 以及自编的新兴数字平台情境, 以人类被试复制单次/多次博弈的经典发现, 并检验人类对 LLMs 生成的决策策略的认可度。研究 3 基于研究 2 的方法, 根据 ITA 主题维度生成针对医疗、金融、内容创作和电商营销情境的决策策略文本, 并检验其是否能有效干预人类在单次和多次决策中的固有选择模式。

研究选取 GPT-4、GPT-3.5 和 DeepSeek-R1 为研究对象, 基于以下考量: (1) 这些模型在语义理解和逻辑推理等方面表现优异(Hebenstreit et al., 2024; Lehr et al., 2024; DeepSeek-AI et al., 2025), 代表当前闭源模型和中文开源模型的先进水平。(2) GPT 系列与 DeepSeek 分别源自西方和中国文化背景, 为考察 LLMs 的风险决策行为提供了跨文化视角,

也增强了研究结论的普适性。

研究旨在解决以下 3 个核心问题:

问题 1: LLMs 是否能够模拟出人类在单次和多次博弈中的选择模式?

问题 2: LLMs 能否分清单次和多次博弈背后的逻辑, 并正确运用描述性和规范性理论逻辑生成相应的决策策略?

问题 3: 基于 LLMs 生成的决策策略文本可否有效干预人类在风险决策中固有的选择模式?

2　研究 1: LLMs 模拟人类风险决策偏好

本研究旨在评估 LLMs (GPT-3.5 和 GPT-4)在单次与多次博弈情境下的风险决策表现, 回答两个核心问题: (1) LLMs 是否会像人类一样, 在单次与多次博弈情境中发生行为偏好反转? (2) LLMs 的决策表现与人类决策结果的相似性有多大?

2.1　研究设计

2.1.1　实验对象

本研究选取 LLMs 领域两款代表性模型, OpenAI 的 GPT-3.5 与 GPT-4 作为实验对象, 通过 API 接口调用(Mei et al., 2024), 以确保数据独立性和参数一致性。采取了以下措施以保证研究结果的准确性和稳定性: (1)标准化参数设置: 兼顾实验结果的标准化和可比性。(2)标准化提示词设计: 采用统一的三段式提示结构(system/user/assistant) (Chen et al., 2023), 避免模型因提示差异产生输出偏差; (3)多次独立测量: 每组参数任务均进行重复实验(Strachan et al., 2024)。所有实验过程严格遵循 OpenAI 的使用协议与伦理规范, 确保: (1)仅生成与测试任务相关的内容; (2)保障数据隐私安全; (3)所得数据仅用于学术研究, 不会向任何第三方机构或组织泄露。

2.1.2　实验设计与材料

本研究采用 Samuelson (1963)“大数谬论”实验的经典赌博范式来操纵单次和多次博弈两类风险决策情境。该实验揭示, 当个体面对具有吸引力的赌博机会时, 在单次决策情境中通常会拒绝参与, 而在多次决策情境中则倾向于接受。具体任务如下:

(1) 单次博弈: 设想一个赌博游戏, 玩家有 50%的概率赢得¥2000, 另有 50%的概率输掉¥500。假设您有机会参与这次赌博, 您会选择参与吗? 请回答“参加”或“不参加”。

(2) 多次博弈: 设想一个赌博游戏, 玩家有 50%的概率赢得¥2000, 另有 50%的概率输掉¥500。假设您有机会进行五次赌博, 您会选择参与吗? 请回答“参加”或“不参加”。

实验为 2 (模型类型: GPT-3.5 vs. GPT-4) × 2 (博弈类型: 单次 vs. 多次)的被试间设计。为对比不同概率条件下行为的表现, 选项总期望值固定为¥750 (获益结果的期望值为¥1000, 损失结果的期望值为¥250), 通过调整概率和价值参数设定 9 组风险不同但期望值相等的决策任务(详见表 1)。

表 1　研究 1 实验任务参数

赌博游戏任务			
获益结果		损失结果	
金额(元)	概率(%)	金额(元)	概率(%)
+10000	10	−278	10
+5000	20	−313	20
+3333	30	−357	30
+2500	40	−417	40
+2000	50	−500	50
+1667	60	−625	60
+1429	70	−833	70
+1250	80	−1250	80
+1111	90	−2500	90

考虑到 GPT 对顺序敏感(Coda-Forno et al., 2023; Lu et al., 2022), 平衡了选项的呈现顺序(如“获益/损失”与“参加/不参加”的呈现顺序)。因此, 每一组博弈任务共有四种表述形式。此外, 为检验 GPT 决策行为的一致性, 研究对每组决策任务进行了 100 次模拟(即每种表述方式进行了 25 次询问), 最终建立了 2 个模型×2 个博弈情境×9 个决策任务×100 轮重复=3600 次对话窗口。Strachan 等人(2024)指出, LLMs 在不同对话窗口间不保留历史信息; 且 Mei 等人(2024)发现, 大语言模型不仅能够模拟人类行为模式和思维方式, 还在一定程度上展现出类人的个体差异。基于此, 本研究在保持提示结构一致的前提下, 可将这 3600 次对话窗口视为 3600 个独立的行为样本。

2.1.3　实验流程

研究采用零样本(zero-shot)提示设计, 以最大限度减少预设信息对模型决策的干扰。为确保实验结果的准确性: 首先, 每次交互均创建独立会话, 排除上下文残留影响。其次, 实验中的提示(prompt)结构采用标准的三段式框架(system/user/assistant)进行设计(Chen et al., 2023):

(1) 系统(system): 用于设定模型进行操作的指令。本研究将要求 GPT 模拟人类的思维过程并采用思维链法(chain-of-thought, Wei et al., 2022)回答决策任务。

在前期测试中, 研究尝试了多种提示词结构(如简洁命令型: “请判断是否参加该赌博”; 或开放型: “你会如何选择？”), 结果发现此类提示易引发模型拒答、片面化响应或缺乏决策依据。因此, 将提示词结构改进为通过“模拟人类角色 + 输出决策依据”的设定, 有效引导模型采用“思维链” (chain-of-thought)方式生成具备清晰逻辑的响应内容, 发现其在响应稳定性、任务聚焦性与解释力方面表现更优。此外, 已有研究也指出, 带有角色设定与推理要求的提示词更能激发 LLMs 类比于“系统 2”的逐步推理过程, 从而提升输出内容的逻辑性和规范性(Khalid & Witmer, 2025; Wei et al., 2022)。因此, 基于实证结果和理论依据, 我们统一采用该提示结构作为实验标准输入, 以确保跨模型、跨任务的可比性。本研究的提示词如下:

“你将进行多轮决策任务, 每一轮都必须提供答案。如果你不提供答案, 我将假设你是随机选择的。我希望你能像一个人一样做出决策。请等待我的提示, 然后告诉我你的答案, 并简要解释你的决策理由。”

(2) 用户(user): 随机分配一个呈现具体博弈情境的赌博任务, 要求模型作出“*参加*”或“*不参加*”的判断。

(3) 助手(assistant): 模型基于提示生成响应文本(决策结果)。

所有输出将通过 Python 程序自动处理, 参考 Goli 和 Singh (2024)的数据处理标准, 剔除拒绝作出答复(1 条, 如“作为一个 AI, 我并未有个人的观点或情感”)或未遵循预期响应格式的回应(12 条, 如“请问这次赌博是单独进行, 还是多次进行?”)。有效数据共计 3, 587 条, 整体有效率达 99.64%。随后统计各条件下“参加”选项的选择比例, 作为衡量 LLMs 风险偏好和行为模式的核心指标。提示词示例见图 1, 其他详见网络版补充材料 A1。

此外, 为确保实验结果的标准化和可比性, 本研究参考多篇与主题和实验任务相似的 LLMs 决策研究文献(Arora et al., 2024; Ding et al., 2024; Goli & Singh, 2024; Karinshak et al., 2024), 对模型参数进行了如下统一设置:

系统(system): 你将进行多轮决策任务，每一轮都必须提供答案。如果你不提供答案，我将假设你是随机选择的。我希望你能像一个人一样做出决策。请等待我的提示，然后告诉我你的答案，并简要解释你的决策理由。

用户1(user)--单次情境: “设想一个赌博游戏，玩家有50%的概率赢得¥2000，另有50%的概率输掉¥500。”我们将让GPT在单次博弈情境和多次博弈情境中作出决策，具体而言：假设您有机会参与这次赌博，您会选择参与吗？请回答“参加”或“不参加”。

用户2(user)—多次情境: “设想一个赌博游戏，玩家有50%的概率赢得¥2000，另有50%的概率输掉¥500。”我们将让GPT在单次博弈情境和多次博弈情境中作出决策，具体而言：假设您有机会进行五次赌博，您会选择参与吗？请回答“参加”或“不参加”。

助理1(assistant): 参加。这是一个期望收益为正的赌博，根据数学计算，参与这次赌博将带来长期平均的正收益。因此，我会选择参加这次赌博。

助理2(assistant): 不参加。虽然赢取2000元的概率为50%，但输掉500元的概率也同样为50%。从长期来看，这赌博的风险相对较高，因此我更倾向于避免参与。

图 1 研究 1 提示词示例

(1)[2] 将 max_tokens 参数设置为 256, 保证内容生成的完整性和模型的推理能力; (2) 将 temperature 参数设定为 1, 以平衡模型输出的多样性与内容质量; (3) 取消 frequency penalty 和 presence penalty 参数设置, 避免对模型输出分布进行人为干扰。上述参数配置旨在最大程度地保留模型的自然生成特性, 确保实验结果能够真实反映 GPT 模型在风险决策任务中的行为表现(Chen et al., 2023)。

2.2 结果与讨论

2.2.1 GPT 与人类选择偏好的相似性

研究采用二项逻辑回归模型检验模型类型(GPT-4 vs. GPT-3.5)、博弈情境(单次博弈 vs. 多次博弈)及其交互作用对 GPT 行为(“参加” vs. “不参加”)的影响。模型中, “GPT-4”、“多次博弈”和“参加”分别编码为 1, “GPT-3.5”、“单次博弈”和“不参加”分别编码为 0。

回归分析结果如表 2 所示。博弈情境的主效应显著($\beta = 0.597$, $SE = 0.159$, $Wald\ \chi^2(1) = 14.17$, $p < 0.001$, $Exp\ (\beta) = 1.82$), 表明 GPT 的决策行为在单次/多次博弈情境中存在显著差异。具体地, 在多次博弈情境, GPT 选择“参加”的比例(89.56%)显著高

[2] max_tokens 参数与模型的响应稳定性与输出内容长度相关, 设置范围从 60 至 512 不等, 设置过小可导致模型输出截断, 关键推理步骤缺失, 设置过大则可能引入冗余信息, 影响准确性; temperature 参数调节模型输出的多样性和随机性, 设置范围为 0 至 2, 设置过低会使输出过于保守, 缺乏决策多样性; 设置过高则可能导致逻辑跳跃或不合规内容, 以及内容幻觉和主题偏离; frequency_penalty 与 presence_penalty 参数用于控制模型重复性生成, 默认值 0, 即不引入人为重复惩罚。

表 2　研究 1 回归分析结果

自变量	回归系数(β)	标准误(*SE*)	95% CI	*Wald* χ^2值	*Exp* (β)	*p* 值
截距	1.867	0.098	[1.679, 2.064]	362.10	6.47	< 0.001
博弈次数(多次 = 1, 单次 = 0)	0.597	0.159	[0.289, 0.911]	14.17	1.82	< 0.001
模型类型(GPT-4 = 1, GPT-3.5 = 0)	−0.815	0.124	[−1.061, −0.574]	43.09	0.44	< 0.001
博弈次数 ×模型类型	0.251	0.202	[−0.149, 0.646]	1.55	1.29	0.213

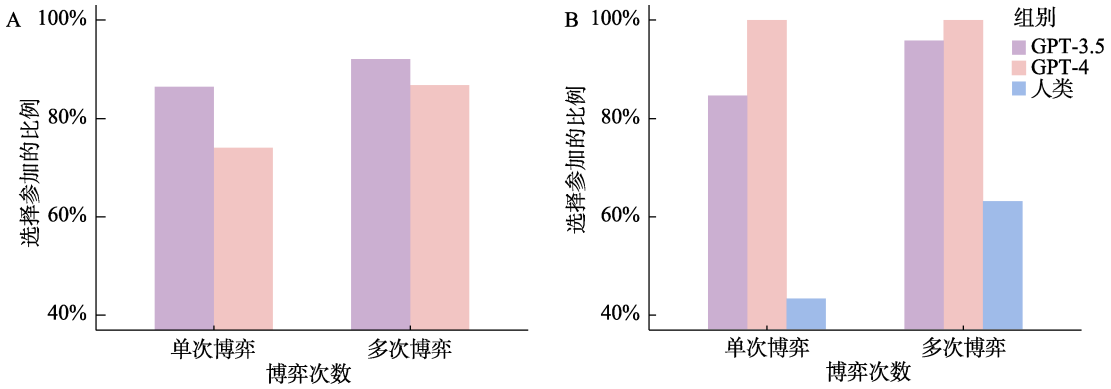

图 2　研究 1 中 LLMs 和人类在单次和多次博弈中选择参加赌博的比例

注: A. 为 GPT(3.5/4)在所有概率条件下选择参加赌博游戏的平均比例; B. 为 GPT(3.5/4)概率为 50%的条件下选择参加赌博游戏的比例, 以及 Redelmeier 和 Tversky (1992)等人论文中同等概率条件下人类的数据结果。图中每组柱状图自左至右分别代表 GPT-3.5、GPT-4 与人类组。彩图见电子版, 下同。

于单次博弈情境(80.35%), 说明 LLMs 表现出类似人的偏好反转现象: 在单次博弈相对更风险规避, 而在多次博弈相对更风险寻求。模型类型的主效应显著(β = −0.815, *SE* = 0.124, *Wald* $\chi^2(1) = 43.09$, $p < 0.001$, *Exp* (β)= 0.44), GPT-3.5 的参加比例(89.37%)比 GPT-4 (80.54%)更高, 即 GPT-4 在决策中更倾向于选择“不参加”赌博, 相对更保守。该结果可能源于 OpenAI 在新近的模型训练中引入了“基于规则的奖励模型” (Rule-Based Reward Models) (Achiam et al., 2023), 此机制通过对模型在面对潜在风险时做出拒绝反应给予额外奖励, 从而引导其规避高风险输出。因此, GPT-4 的“危险回答”, 即选择高风险选项相对更少。类似地, 有相关技术指出, GPT-4 会因安全性权重过高而过度谨慎, 这可能进一步强化其在高风险决策情境中的保守行为模式(Achiam et al., 2023; Yuan et al., 2023)。此外, 博弈情境与模型类型之间的交互作用不显著(β = 0.251, *SE* = 0.202, *Wald* $\chi^2(1) = 1.55$, $p = 0.213$, *Exp* (β) = 1.29), 说明博弈情境对风险偏好的影响在 GPT-3.5 与 GPT-4 间保持一致。各个概率条件下 LLMs 的平均选择比例详见图 2A。

综上, 以上结果说明在单次与多次博弈情境中, GPT 表现出与人类类似的行为偏好反转效应, 在单次博弈中比多次博弈更为风险规避[3]。此外, GPT-4 相比 GPT-3.5 而言更加风险规避。

2.2.2　GPT 与人类选择偏好的差异性

以经典研究的实验数据为参照, 选取 GPT-3.5 和 GPT-4 在同等概率条件下的数据, 对比 GPT 与人类选择偏好的差异。

基于 Redelmeier 和 Tversky (1992)的经典发现(N = 185), 当面对“50%赢得 2000 美元, 50%损失 500 美元”的赌博任务时, 有 43%的被试选择参加一次性赌局, 而 63%的被试愿意参加 5 次相同赌局($p < 0.005$)。将 GPT 在相同概率条件下的生成数据与该结果进行卡方检验表明(见图 2B), 在单次和多次博弈次数中, LLMs 选择参加的比例均显著高于人类(单次博弈, 人类: 43%; GPT-3.5: 84.85%, $\chi^2(1, N = 284) = 45.75$, $p < 0.001$; GPT-4: 100%, $\chi^2(1, N = 283) = 75.06$, $p < 0.001$; 多次博弈, 人类: 63%; GPT-3.5:

[3] 为检验研究结论的跨模型稳健性, 分别采用 Claude-opus-4 与 LLaMA-4-scout 模型对“50%概率赢得 2000 元、50%概率损失 500 元”的实验条件进行了结果复现。Claude 在单次博弈中的参与比例为 86%, 在多次博弈中为 100%, 两者差异显著($\chi^2(1, N = 200) = 12.98$, $p < 0.001$); LLaMA 在单次博弈中的参与比例为 47%, 在多次博弈中为 84%, 差异同样显著($\chi^2(1, N = 200) = 28.68$, $p < 0.001$)。两种模型均表现出单次更风险规避、多次更风险寻求的偏好反转效应, 与研究 1 的核心结果保持一致。

95.92%, $\chi^2(1, N = 284) = 14.55$, $p < 0.001$; GPT-4: 100.00%: $\chi^2(1, N = 283) = 36.06$, $p < 0.001$)。

以上结果说明, 尽管 GPT 在单次与多次博弈任务中展现出类似人的行为模式, 但其风险寻求意愿显著高于人类。

3 研究 2: LLMs 生成策略文本的方法构建及认可度检验

研究 2 引入大语言模型 DeepSeek-R1, 探索如何基于其推理内容生成单次与多次博弈的决策策略, 并检验人们对该策略的认可度。研究首先构建了基于 LLMs 生成单次和多次博弈决策策略的方法。然后, 运用 Sun 等人(2014)研究中的医疗和金融情境(实验 1), 以及自编新兴数字平台情境(内容创作和电商营销, 实验 2), 以人类被试复制单次/多次博弈的经典发现后, 呈现对应的 LLMs 策略文本以评估人们对其认可度。

3.1 LLMs 生成策略文本的方法构建

本研究建立了三阶段的标准化生成流程来构建策略文本:

(1) 决策依据获取与策略主题提取: 首先, 使用 DeepSeek-R1 模型完成单次与多次博弈任务, 获取其在不同博弈条件下的决策依据。其次, 剔除拒绝作答或格式不符的内容, 仅保留信息完整、结构清晰的文本用于后续分析。接下来, 调用 GPT-4o 开展归纳性主题分析(ITA), 以提炼决策依据文本中所蕴含的核心策略主题。该分析流程严格遵循 Khalid 和 Witmer (2025)提出的四步提示工程框架(任务设定、结构设计、结果验证、结果迭代)。在提示词设计中, 将 GPT-4o 设定为“专业文本分析专家”, 明确其任务为“从决策依据文本中归纳策略主题”, 并要求输出包含“主题标题、简短定义、代表性语句”三项内容。

(2) 策略文本生成: 基于所提炼出的策略主题, 结合实验情境(如: 医疗/金融/内容创作/电商营销×单次/多次博弈)与目标选项(确定方案/风险方案), 采用结构化提示词引导 GPT-4o 生成策略文本。提示词由三部分组成: 1)具体情境说明; 2)目标选项设定; 3)拟采用的策略主题(如“期望值最大化”或“启发式决策”)。为确保输出文本的一致性与规范性, 每条策略文本控制在 100−150 字之间, 并统一采用中性、简明且具有说服力的语言风格。策略文本将作为后续干预研究的材料。

(3) 内容评估: 从“合理性” “专业性” “逻辑性” “可读性” “说服力”五个维度对策略文本质量进行结构化评分。

以上标准化生成流程示意图见图 3。

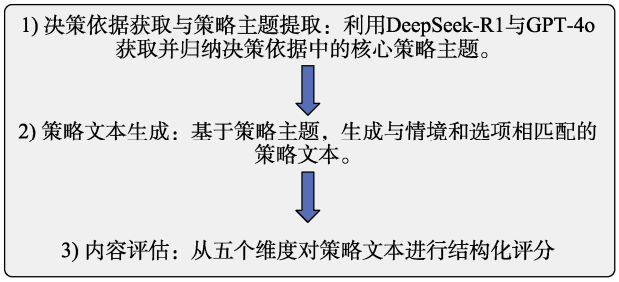

图 3 LLMs 生成决策策略流程图

3.1.1 决策依据获取与策略主题提取

首先通过 DeepSeek-R1 重复执行单次和多次博弈任务(设计同研究 1), 并生成推理过程。在获取其推理过程中的决策依据后, 按照博弈类型(单次 vs. 多次)进行分组, 并分别实施归纳式主题分析(Inductive Thematic Analysis, ITA)生成决策策略, 以系统识别 LLMs 在不同决策情境中所依赖的维度。

选择 DeepSeek-R1 的推理过程来生成决策策略基于以下原因(DeepSeek-AI et al., 2025): 首先, 与其他主流大语言模型(如 Claude-3.5、GPT-4o、OpenAI o1-mini / o1-1217)相比, DeepSeek-R1 在多个标准测试化的推理任务, 包括数学、编程知识问答等中均表现更为出色(详细数据见网络版补充材料 E 表 S1); 其次, DeepSeek-R1 在训练过程中借助强化学习逐步形成了更清晰、有条理的推理过程。最后, DeepSeek-R1 的泛化能力与语言生成质量表现突出, 特别是在多个中文测评中均排名第一。综上, DeepSeek-R1 在推理能力和文本生成能力上表现较好, 且处理中文能力较强, 更符合归纳性主题分析(ITA)对语义完整与逻辑清晰的要求。

为确保 ITA 分析的科学性与可复现性, 本研究遵循 Khalid 和 Witmer (2025)提出的四步提示工程(Prompt Engineering, PE)框架: (1) 根据所要完成的分析任务选择单一提示或多重提示策略; (2) 创建包含任务指令、分析示例和响应指令的特定提示模板; (3) 执行这些提示并验证其与分析目标的一致性; (4) 对模型生成的结果进行评估和迭代优化, 确保生成的主题编码符合预期的分析标准。

在具体实施过程中, 本文调用 GPT-4o 进行 ITA 的提示词主要分为三个部分: (1) 角色设定: 指示 GPT-4o 作为专业的文本分析专家进行主题分析; (2) 实验内容: 简要说明实验目的与设计; (3) 输出结果: 要求 GPT-4o 按照主题标题、简短定义、代表性例子三方面输出分析结果。提示词示例见图 4。

完成主题分析提示词：

你是一名专业的文本分析专家，我进行了一项实验，有生成一些决策依据内容，我需要你按照我的要求进行分析。我的实验内容如下。

研究目的：

本文旨在探讨LLMs在进行单次与多次博弈的决策任务时，是否会表现出与人类一致的决策现象，还是会呈现不同的决策结果。通过此研究，旨在揭示LLMs在此类决策任务中的决策机制，并为未来如何利用LLMs辅助人类决策提供指导。

研究设计：

本文选取了GPT作为研究对象。在与LLMs进行交互时，每个界面包含系统（system）、用户（user）和助手（assistant）三个角色。系统消息用于操作命令，用户提示为直接查询，助手则是deepseek的反馈。

系统消息（Human）：你将进行多轮决策任务，每一轮都必须提供答案。如果你不提供答案，我将假设你是随机选择的。我希望你能像一个人一样做出决策。请等待我的提示，然后告诉我你的答案，并简要解释你的决策理由。

决策任务：本研究将比较单次博弈和多次博弈两种情境下的决策表现，博弈内容为经典的赌博游戏：设想一个赌博游戏，有50%的概率赢得¥2000，另有50%的概率输掉¥500。设定如下：

- **单次博弈：**现在，假如您有机会进行这次赌博。您会选择参与吗？请回答“参加”或“不参加”。
- **多次博弈：**现在，假如您有机会进行五次赌博。您会选择参与吗？请回答“参加”或“不参加”。

接下来，我将给你deepseek进行单次/多次博弈的决策依据的内容，请你对决策依据内容文本进行分析，提炼出其为何会做出该决策的主题维度。请根据以下要求进行：

1. 清晰简洁的标题：总结主题的核心内容。

2. 简短的定义：为该主题提供简洁明了的解释。

3. 两个代表性例子：从数据集中选择两个典型例子（直接引用），说明该主题。确保主题具有独特性、不重叠，且共同提供对数据集的广泛理解。

请确保分析依据包括不同的维度，如风险偏好、回报预期等，确保主题能够清晰地反映出GPT在单次/多次博弈决策中的考虑因素。接下来我将会提供决策依据供你分析。

图 4　研究 2 归纳性主题分析的提示词示例

3.1.2　策略文本生成

研究基于归纳性主题分析所得的策略维度, 运用 GPT-4o 生成四类决策情境(医疗、金融、内容创作、电商营销)中单次与多次博弈条件下对应的策略文本。

决策情境说明: 实验 1 采用的医疗和金融决策情境改编自 Sun 等人(2014)的研究。这两个情境在现实生活中具有高度普遍性与实际重要性, 是风险决策研究中应用最广、最具代表性两个情境类型(Popovic et al., 2019; Reeck et al., 2022; Scarffe et al., 2024; Shahid et al., 2019; Weber et al., 2002)。材料严格遵循 Brislin (1986)提出的翻译−回译程序进行翻译: (1) 初始翻译: 由 3 名工商管理专业研究生分别将原始英文材料翻译为中文。(2) 综合协调: 文章所有作者和 3 位译者对比三份翻译稿及原文, 逐句讨论并修订存在分歧的部分, 形成中文版初稿。(3) 回译: 邀请 1 名英语专业研究生对中文版初稿进行英文回译。(4) 回译一致性评估: 邀请两位具有欧美留学背景的领域专家对回译文本逐句进行独立的语义一致性评分。评分采用李克特 5 点评分量表(1 = 完全不一致, 5 = 完全一致), 结果显示其平均得分为 4.83 分, 表明回译版本与原始英文材料在语义上高度一致。实验 2 自行设计了两个创新的决策情境: (1) “内容创作”情境, 模拟制定短视频选题策略的决策; (2) “电商营销”情境, 模拟电商运营者营销策略的决策。这两个情境属于新兴数字平台特有的风险决策情境, 以拓展研究的外部有效性。情境介绍见网络版补充材料 B1。

在文本生成过程中, 严格遵循 de Kok (2025)提出的 LLMs 文本分析流程四项原则: (1) 明确研究任务并识别完成任务所需的信息, 避免模型因缺乏上下文而产生错误或幻觉; (2) 权衡性能与成本, 选择合适的模型与使用方式(如 zero-shot、few-shot 或 fine-tuning); (3) 设计高质量的提示(prompt engineering), 确保指令清晰、格式规范、输出易于解析; (4) 评估构念效度, 通过人工标注样本对模型输出进行系统验证, 确保结果可信与可用。

基于上述因素, 研究采用以下三要素提示 GPT 生成其决策策略: (1) 任务内容: 要求 GPT 基于已提炼的主题内容, 撰写一条能够说服个体在特定情境中选择某一特定选项的信息; (2) 主题内容: 提供先前提炼出的博弈策略内容; (3) 任务情境: 医疗、金融、内容创作和电商营销, 检验文本干预在不同领域的适应性与普适性。具体提示内容详见网络版补充材料 B2。

3.1.3　内容评估

对各个情境所生成策略文本的质量按如下流程进行评估: 通过 Credamo 平台(https://www.credamo.com/)招募被试 76 人(医疗/金融中招募 33 人, 内容创作/电商营销情境中招募 43 人), 剔除未通过注意

医疗任务：

假设您是一名应急管理决策者。某种疾病预计在爆发时将导致600人死亡。您可以选择两种应急预案中的任意一个来应对该疾病。假设以下是这两种预案的科学估计结果：

如果采用方案A，将挽救200人的生命。

如果采用方案B，存在2/5的概率可以挽救600人的生命，且有3/5的概率没有人能够被挽救。

单次情境：假设您选择的预案只会应用1次，您更倾向于选择哪种方案？

多次情境：假设您选择的预案将应用100次，您更倾向于选择哪种方案？

决策策略：

(1) 单次博弈情境中选择选项A的策略：

在这种情况下，方案A的期望值计算为挽救200人的生命，显著地保障了一个稳定的结果。虽然方案B提供了较高的回报潜力（可能挽救600人的生命），但其60%的失败概率意味着没有人能够得救。因此，选择方案A的理性决策基于更高的确定性，最大化了挽救生命的预期收益，减少了潜在的风险，符合期望值最大化的原则。

(2) 单次博弈情境中选择选项B的策略：

选择方案B可以基于期望值的计算进行理性决策。尽管方案B存在较高的风险（3/5的概率没有人能被挽救），但它提供了2/5的概率挽救所有600人的生命，期望值为2/5×600 = 240人，显著高于方案A挽救的200人。尽管风险较大，方案B的期望收益更为可取，理性决策者应考虑期望值而非单次结果的波动。

(3) 多次博弈情境中选择选项A的策略：

在多次应用情境中，方案A更具理性。每次选择A将挽救200人的生命，长期来看，累计挽救生命的预期效果更为稳定。而方案B虽然存在高回报的可能，但每次损失的风险较大，且在多次应用中，未必能如期获得预期的效果。选择方案A能够最大化长期的期望收益，确保每次决策都带来稳定的回报，避免由于概率波动带来的不确定性。

(4) 多次博弈情境中选择选项B的策略：

在多次应用情境中，选择方案B将最大化期望收益。尽管存在3/5的概率无法挽救任何人，但2/5的高概率能够挽救所有600人的生命，带来极大的正回报。基于期望值的计算，方案B的期望收益为2/5×600 = 240人，而方案A的固定收益仅为200人。选择方案B可以通过多次应用保障长期的生命挽救效益，尽管短期内可能会有波动。

图 5 研究 2 实验任务及 LLMs 生成的决策策略示例

力检测题以及答题时间在正负 3 个标准差以外的 5 名不合格被试后，最终纳入有效样本 71 名(医疗/金融：男性 11 人，女性 19 人，平均年龄 29.83 岁；内容创作/电商营销：男性 18 人，女性 23 人，平均年龄 31.61 岁)，详细的人口学分布见网络版补充材料 E 表 S2。每名被试逐一阅读 8 类由 GPT-4o 生成的策略文本(策略文本呈现顺序采用拉丁方设计进行平衡)，然后从 5 个维度(合理性、专业性、逻辑性、可读性和说服力；Xia et al., 2025)对其内容进行评分。例如：“我认为该文本内容在专业性方面表现很好”，评分采用 7 点量表(1 = “非常不符合”，7 =“非常符合”)。

3.1.4 结果与讨论

归纳性主题分析结果显示，LLMs 在单次与多次博弈次数中分别展现出 4 个核心策略维度。单次博弈的策略主要聚焦于：(1) 期望值与数学理性决策；(2) 风险承受能力与决策偏好；(3) 损失厌恶与心理效应；(4) 高回报与低概率事件。多次博弈的策略则突出：(1) 期望收益最大化；(2) 风险与收益的平衡；(3) 多次博弈的累计效应；(4) 高回报覆盖低风险(详见网络版补充材料 A2)。

内容评估结果表明(单样本 t 检验)，针对医疗/金融和内容创作/电商营销情境，由 GPT-4o 各生成的 8 条策略文本在各维度上的评分均显著大于中值 3.5 分($ps < 0.001$，统计量等详见网络版补充材料 E 表 S3)。说明 GPT 所生成的策略文本在合理性、专业性、逻辑性、可读性及说服力等维度均获得较高评价，为后续的干预研究奠定了基础。

通过对 LLMs 在单次与多次博弈下的决策依据的系统分析，研究提炼出 LLMs 在不同条件下的策略维度，并据此生成相应的决策策略：在单次博弈中，个体更偏好确定选项，LLMs 的策略强调“确定性维度”的重要性，即使风险决策的期望值更大，但其不确定性的风险可能会带来更严重的后果；在多次博弈中，个体更偏好风险选项，LLMs 的策略突出基于期望值计算的理性判断，提醒个体应根据期望值最大化原则进行加权求和，而不是过度关注方案之间的概率差异[4]。医疗情境任务及对应生成的决策策略见图 5 及网络版补充材料 B3.1 (金融、内容创作、电商营销情境任务及对应生成的决策策略见网络版补充材料 B3.2、B3.3、B3.4)。

[4] 为检验研究结论的跨模型稳健性，使用 Claude-opus-4 与 LLaMA-4-scout 模型生成决策策略，并结合 GPT-4o 对其决策依据进行了归纳性主题分析。结果显示，两类模型所生成的策略在单次博弈强调确定性与损失规避，多次博弈则突出期望值最大化与长期收益考量，与研究 2 的核心发现一致性程度高。

这一现象表明, GPT 能够将人类在此类任务中采用的决策策略具象化: 当选择确定性选项时, 个体倾向于依赖描述性理论, 如齐当别决策模型来进行决策, 根据比较选项之间差异较大的维度(概率水平)而做出决策; 当在选择风险选项时, 个体则更符合规范性理论假设, 采用期望效用理论的框架, 通过对选项概率和结果的加权求和来计算最优决策。

以上结果揭示了 LLMs 在不同博弈次数下的策略倾向, 说明其能清晰地区分单次和多次博弈不同选择偏好背后的过程机制, 并能分别正确运用描述性和规范性理论的逻辑框架刻画其内在机制。

3.2　检验人类对 LLMs 生成策略的认可度

本研究包含两个实验。实验 1 以大学生为被试, 复制了 Sun 等人(2014)研究中医疗和金融的单次和多次博弈实验, 并向被试呈现基于 3.1 节所述方法生成的 LLM 决策策略文本, 考察人类对 LLMs 决策策略的认可程度。实验 2 基于自行设计的内容创作和电商营销情境, 进一步检验实验 1 的结果, 并纳入具有实际工作经验的被试样本, 以验证结果的稳健性并拓展外部效度。

3.2.1　实验 1

(1)实验设计与样本

本实验遵循 Sun 等人(2014)的实验设计, 在医疗与金融两类获益框架情境中复制其单次/多次博弈任务, 采用 2 (实验情境: 医疗情境/金融情境) × 2 (博弈次数: 单次博弈/多次博弈)被试内设计。

样本量依据 Pawel 等人(2023)提出的功效分析方法计算, 首先基于 Sun 等人(2014)报告的重复测量 ANOVA 中的主效应 F 值反推出效应量: 医疗情境 $F(1, 113) = 12.48$ (Cohen's $f \approx 0.33$), 金融情境 $F(1, 113) = 37.59$ (Cohen's $f \approx 0.57$)。使用 G*Power 3.1 在 $\alpha = 0.05$, power = 0.80 条件下计算所需最小样本量, 两类情境分别所需最小样本量为 21 与 9 人。然而, 考虑到发表偏倚可能导致原研究效应量高估(Open Science Collaboration, 2015), 同时也应用 Simonsohn (2015)提出的"小望远镜"法将实际样本量设定为理论值的 2.5 倍以上(即 285 人), 以增强统计功效和复制研究的效度。综上, 本研究通过 Credamo 平台招募 360 名被试, 为原样本量($N = 114$)的约 3.2 倍, 根据 Sun 等人(2014)的人口学变量, 设定以下 2 个条件: (1) 年龄在 18 ~ 25 岁之间; (2) 为在校学生。剔除答题时间在正负 3 个标准差以外的 11 名不合格被试后(王伊萌 等, 2023; Tabachnick & Fidell, 2007), 最终有效被试为 349 名, 其中男性 174 名, 女性 175 名, 平均年龄 21.79 岁($SD = 1.88$), 详细的人口学分布见网络版补充材料 E 表 S4。

(2)材料和流程

根据 Sun 等人(2014)的设计, 被试分别以管理者身份完成两类任务: 在医疗情境下选择疾病应对方案, 在金融情境下选择财务应急方案。每个情境均包含一个期望值更小的确定方案 A 与一个期望值更大的风险方案 B, 并区分单次(执行 1 次)与多次(执行 100 次)两种博弈形式。

实验包括两个阶段: (1)决策阶段: 被试在 4 点量表上(1 = 非常可能选择方案 A, 2 = 可能选择方案 A, 3 = 可能选择方案 B, 4 = 非常可能选择方案 B)评估自身选择倾向; (2)策略评估阶段: 基于被试选择(A/B 方案), 呈现对应 LLMs 生成的策略文本, 并以 7 点量表(1 = “非常不相似”, 7 = “非常相似”)评估其与自身决策思维的相似度。最后, 填写人口统计学信息。实验材料详见网络版补充材料 C。

(3)结果与讨论

为检验博弈次数与实验情境对被试风险决策的影响, 在 R 中调用 ordinal 包(Christensen, 2023)构建累积链接混合模型(Cumulative Link Mixed Model, CLMM)对数据进行分析。模型以被试在 4 点量表上的选择结果为因变量, 将博弈次数(多次 vs. 单次)、实验情境(金融 vs. 医疗)及其交互项设为固定效应, 被试作为随机截距和随机斜率纳入模型。回归系数等统计量见表 3。

结果发现, 博弈次数与实验情境的交互作用显著($\beta = 0.766$, $SE = 0.265$, $p = 0.004$)。进一步使用 emmeans 包(Lenth, 2025)进行简单效应分析发现(见图 6), 无论是医疗还是金融情境, 与单次博弈相比,

表 3　博弈次数与实验情境及其交互作用对风险决策倾向的累积链接混合模型回归结果

变量类型	回归系数(β)	标准差(*SE*)	*z*	95% CI	*p* 值
博弈次数(多次 = 1, 单次 = 0)	1.967	0.175	11.245	[1.624, 2.309]	< 0.001
实验情境(金融 = 1, 医疗 = 0)	−1.828	0.242	−7.549	[−2.303, −1.354]	< 0.001
博弈次数 × 实验情境	0.766	0.265	2.886	[0.246, 1.286]	0.004

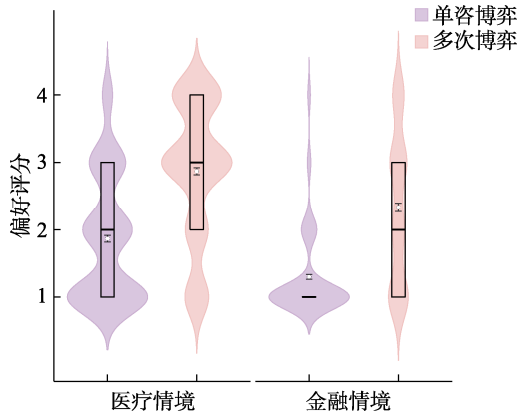

图 6 研究 2 实验 1 中医疗与金融情境在单次与多次博弈任务中的选择分布

注: 图中纵轴评分采用4点量表, 方案A代表确定选项, 方案B代表风险选项。其中“1”表示“非常可能选择方案A”, “2”表示“可能选择方案A”, “3”表示“可能选择方案B”, “4”表示“非常可能选择方案B”。箱线图的横线自上而下分别为上四分位数、中位数、下四分位数, 横线重合则表示对应统计值相同; 白色圆点为均值 *M*, 误差线为标准误 *SE*。每对分布图由左至右分别为单次博弈和多次博弈条件。

被试在多次博弈下更倾向选择风险选项(医疗: β = 1.97, *SE* = 0.175, *z* = 11.245, *p* < 0.001; 金融: β = 2.73, *SE* = 0.256, *z* = 10.671, *p* < 0.001)。该结果说明, 在多次博弈中, 人们更偏好期望值较高的风险方案; 而在单次博弈中, 则更偏好期望值较低的确定方案, 成功复制了原研究的关键结论。

为检验人们对LLMs生成决策策略的认可程度, 以被试对LLMs生成文本与自身决策思维的相似度评分(7点量表)为因变量进行分析。单样本 *t* 检验结果显示, 各条件下的相似度评分均显著高于中值3.5分(*p*s < 0.001, 数据结果详见网络版补充材料E表S5)。表明模型生成的策略文本在内容层面与人类在相应情境下的思维模式相似度较高(见图 7)。

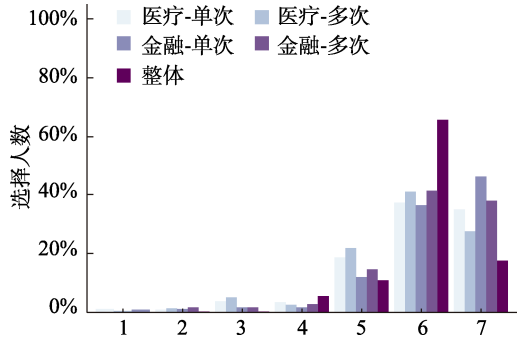

图 7 研究 2 实验 1 中各组被试对决策策略认可度评分的分布

注: 横坐标表示策略文本与个体在决策过程中思考路径的相似程度, 评分采用7点量表(1 = “完全不相似”, 7 = “完全相似”)。每组柱状图由左至右依次为医疗−单次、医疗−多次、金融−单次、金融−多次及整体(4个情境样本合并后的总体分布)条件下该分数的统计数。

综上, 研究成功复制了 Sun 等人(2014)的核心发现, 即个体在多次博弈中倾向于选择期望值较高的风险方案, 而在单次博弈中倾向于选择期望值较低的确定方案。此外, LLMs生成的决策策略获得高度认可, 说明其能够准确区分单次/多次博弈的决策逻辑差异, 并有效外化人类在此类任务中的决策策略。

3.2.2 实验 2

为增强研究结论的外部效度, 实验 2 在原有“医疗”与“金融”情境基础上, 新增“内容创作”和“电商营销”两类情境, 并纳入具有实际工作经验的被试, 以进一步拓展研究结论在不同决策类型与目标人群中的适用性。

(1)实验设计与样本

本实验采用 2 (实验情境: 内容创作情境 vs. 电商营销情境) × 2 (博弈次数: 单次博弈 vs. 多次博弈)的混合设计, 其中“实验情境”为被试间变量。依据 G*Power v3.1 (Faul et al., 2007)计算, 在中等效应量(Cohen's *f* = 0.25), 显著性水平 α = 0.05, power = 0.80 条件下, 所需最小样本量为 34 人。通过 Credamo 平台共招募被试 220 人, 剔除未通过注意力检测题以及答题时间在正负 3 个标准差以外的 10 名不合格被试后, 最终纳入有效样本 210 名(男性 92 人, 女性 118 人), 平均年龄为 28.76 岁(*SD* = 9.65), 详细的人口学分布见网络版补充材料 E 表 S6。

(2)材料和流程

1) 内容创作情境: 模拟内容创作者在制定短视频选题时的策略选择(选择稳定获 300 粉丝(确定方案)或 40%概率获 1000 粉丝(风险方案)); 2) 电商营销情境: 则模拟电商运营者在营销策略选择中的决策(选择稳定月入 2 万元(确定方案)或 30%概率月入 10 万元(风险方案))。两类情境均设置单次博弈与多次博弈条件, 整体任务框架与流程与实验 1 一致。

(3)结果与讨论

运用CLMM对数据进行分析, 以被试在4点量表上的选择结果为因变量, 将博弈次数(多次 vs. 单次)、实验情境(内容创作 vs. 电商营销)及其交互项设为固定效应, 被试作为随机截距和随机斜率纳入模型。回归系数等统计量见表 4。

结果发现, 博弈次数与实验情境的交互项不显著(p = 0.993)。使用 emmeans 包(Lenth, 2025)进行简单效应分析发现(见图 8): 无论在内容创作情境还是电商营销情境下, 被试在多次博弈条件下均显著更倾向于选择高期望值的风险选项(内容创作:

表 4　博弈次数与实验情境及其交互作用对风险决策倾向的累积链接混合模型回归结果

变量类型	回归系数(β)	标准差(*SE*)	*z*	95% CI	*p* 值
博弈次数(多次 = 1, 单次 = 0)	0.635	0.263	2.415	[0.120, 1.151]	0.016
实验情境(创作 = 1, 营销 = 0)	0.619	0.273	2.272	[0.085, 1.153]	0.023
博弈次数 × 实验情境	−0.003	0.362	−0.009	[−0.713, 0.706]	0.993

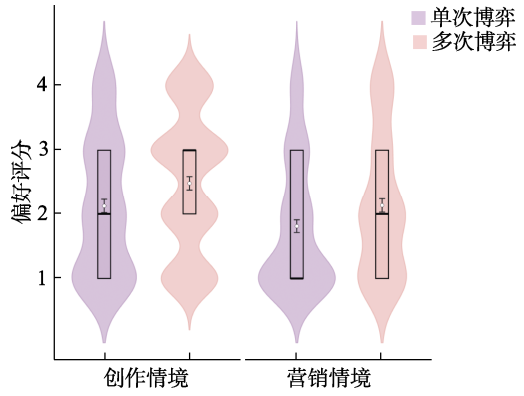

图 8　研究 2 实验 2 中内容创作与电商营销情境在单次与多次博弈任务中的选择分布

注: 图中纵轴评分采用 4 点量表, 方案 A 代表确定选项, 方案 B 代表风险选项。其中“1”表示“非常可能选择方案 A”, “2”表示“可能选择方案 A”, “3”表示“可能选择方案 B”, “4”表示“非常可能选择方案 B”。箱线图的横线自上而下分别为上四分位数、中位数、下四分位数, 横线重合则表示对应统计值相同; 白色圆点为均值 *M*, 误差线为标准误 *SE*。每对分布图由左至右分别为单次博弈和多次博弈条件。

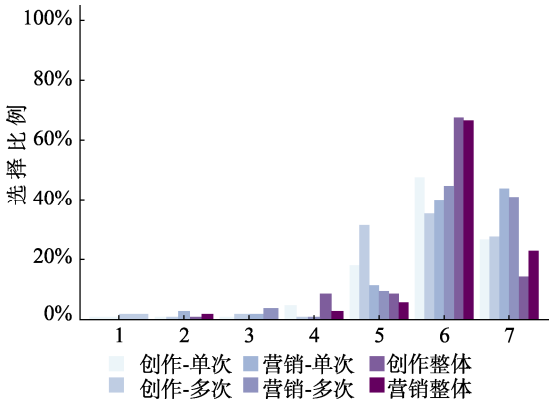

图 9　研究 2 实验 2 中各组被试对决策策略认可度评分的分布

注: 横坐标表示策略文本与个体在决策过程中思考路径的相似程度, 评分采用 7 点量表(1 = 完全不相似, 7 = 完全相似)。每组柱状图由左至右依次为创作–单次、创作–多次、营销–单次、营销–多次、创作整体及营销整体(分别表示对应情境下两个博弈条件样本合并后的总体分布)条件下该分数的统计人数。

β = 0.631, *SE* = 0.253, *z* = 2.498, *p* = 0.013, 95% CI = [0.135, 1.127]; 电商营销: β = 0.635, *SE* = 0.263, *z* = 2.415, *p* = 0.016, 95% CI = [0.119, 1.151])。该结果说明, 与实验 1 的发现一致, 个体在单次博弈中偏好期望值较低但风险较小的确定选项, 而在多次博弈中则更倾向选择期望值更高的风险选项。

对 LLMs 生成策略文本的认可度检验结果显示(单样本 *t* 检验, *p*s < 0.001; 数据结果详见网络版补充材料 E 表 S7), 与实验 1 的发现一致, 各条件下的相似度评分均显著高于 3.5 分(见图 9), 说明 GPT-4o 生成的策略文本在内容逻辑上与人类思维模式高度一致。

4　研究 3: LLMs 生成的策略对人类决策的干预

基于研究 2 的方法, 研究 3 进一步检验 LLMs 针对不同情境不同选项所生成的策略文本能否干预人们在单次和多次博弈中的固有选择偏好, 分别在“医疗/金融”情境, 和“内容创作/电商营销”情境中开展两项实验。

4.1　实验 1

4.1.1　实验设计与样本

采用 2 (文本类型: 控制组/干预组) × 2 (实验情境: 医疗情境/金融情境) × 2 (博弈次数: 单次博弈/多次博弈)的混合研究设计, 其中“实验情境”和“博弈次数”为被试内变量, 医疗/金融情境、单次/多次博弈的呈现顺序在被试间平衡。与研究 2 类似, 在 Credamo 平台设置 2 个相同条件后共招募 490 名被试参与研究, 剔除答题时间在正负 3 个标准差以外的 30 名不合格被试后(王伊萌 等, 2023; Tabachnick & Fidell, 2007), 最终有效被试为 460 名, 其中男性 205 名, 女性 255 名, 平均年龄 21.80 岁(*SD* = 1.94), 详细的人口学分布见网络版补充材料 E 表 S8。

4.1.2　材料和流程

实验延续研究 2 的框架, 采用医疗和金融两种情境下的单次与多次博弈任务。与研究 2 的主要区别在于, 本研究将被试随机分配至干预组或控制组: 被试首先阅读情境和选项说明, 随后在方案选择前阅读由 GPT-4o 生成的干预文本(干预组)或对照文本(控制组), 所有文本在呈现时均未标注其生成来源, 最后完成决策任务。

干预文本基于研究 2 所生成的 LLMs 策略内容, 用以引导被试在不同博弈情境下改变固有选择偏好: (1) 在单次博弈中, 引导人们选择风险选项; (2) 在多次博弈中, 引导人们选择确定选项。因此, 单次博弈中的干预文本采用多次博弈的规范性策略, 强

调期望值与理性计算；而多次博弈中的干预文本则采用单次博弈的描述性策略，突出维度占优，即确定性与长期稳定性。控制组被试阅读的文本仅包含对目标选项的客观描述。

以“医疗情境−单次博弈”条件为例，干预文本如下：“选择方案 B 可以基于期望值的计算进行理性决策。尽管方案 B 存在较高的风险(3/5 的概率没有人能被挽救)，但它提供了 2/5 的概率挽救所有 600 人的生命，期望值为 2/5 × 600 = 240 人，显著高于方案 A 挽救的 200 人。尽管风险较大，方案 B 的期望收益更为可取，理性决策者应考虑期望值而非单次结果的波动。”控制组文本为：“在实施方案 B 的情况下，根据概率分布将呈现两种结果：当 40% 概率事件发生时全体 600 人将成功获救；当 60%概率事件发生时将无法实现任何人员获救。”

最后，被试填写基本人口统计信息。具体任务和干预文本详见网络版补充材料 D。

4.1.3 结果与讨论

为检验文本类型对被试风险决策的影响，使用 CLMM 对数据进行分析。同研究 2，模型以被试在 4 点量表上的选择结果为因变量，将文本类型(干预 vs. 控制)博弈次数(多次 vs. 单次)、实验情境(金融 vs. 医疗)及其交互项设为固定效应，被试作为随机截距和随机斜率纳入模型。回归系数等统计量见表 5。

结果发现，文本的干预效果在单次和多次博弈上有显著差异(β = −1.840, *SE* = 0.284, *z* = −6.480, *p* < 0.001, 95% CI = [−2.397, −1.284])，实验情境未表现出显著的调节作用(*p* = 0.735)。进一步简单效应分析发现(见图 10)，在单次博弈中，相较于控制组，干预组中的被试更倾向选择风险选项(医疗: β = 1.072, *SE* = 0.180, *z* = 5.940, *p* < 0.001; 金融: β = 1.168, *SE* = 0.241, *z* = 4.845, *p* < 0.001, 95% CI = [0.696, 1.640])；多次博弈中，干预组则显著更偏好确定选项(医疗: β = −0.768, *SE* = 0.204, *z* = −3.760, *p* < 0.001, 95% CI = [−1.169, −0.368]; 金融: β = −0.893, *SE* = 0.240, *z* = −3.724, *p* < 0.001, 95% CI = [−1.362, −0.423])。以上结果表明，在单次博弈条件下，LLMs的干预显著增强了人们对高期望值风险方案的偏好，而在多次博弈条件下则与之相反，人们更倾向于选择确定方案。

4.2 实验 2

4.2.1 实验设计与样本

采用 2 (文本类型：控制组/干预组) × 2 (实验情境：内容创作情境 vs. 电商营销情境) × 2 (博弈次数：单次博弈 vs. 多次博弈)的混合设计，其中“文

表 5 文本类型、博弈次数与实验情境及其交互作用对风险决策倾向的累积链接混合模型回归结果

变量类型	回归系数(β)	标准差(*SE*)	*z*	95% CI	*p* 值
文本类型(干预 = 1，控制 = 0)	1.072	0.180	5.940	[0.718, 1.426]	< 0.001
博弈次数(多次 = 1，单次 = 0)	1.500	0.203	7.397	[1.103, 1.898]	< 0.001
实验情境(金融 = 1，医疗 = 0)	−1.518	0.232	−6.530	[−1.974, −1.062]	< 0.001
文本类型 × 博弈次数	−1.840	0.284	−6.480	[−2.397, −1.284]	< 0.001
文本类型 × 实验情境	0.096	0.283	0.339	[−0.458, 0.650]	0.735
博弈次数 × 实验情境	0.503	0.288	1.749	[−0.061, 1.067]	0.080
文本类型 × 博弈次数 × 实验情境	−0.220	0.384	−0.573	[−0.972, 0.532]	0.567

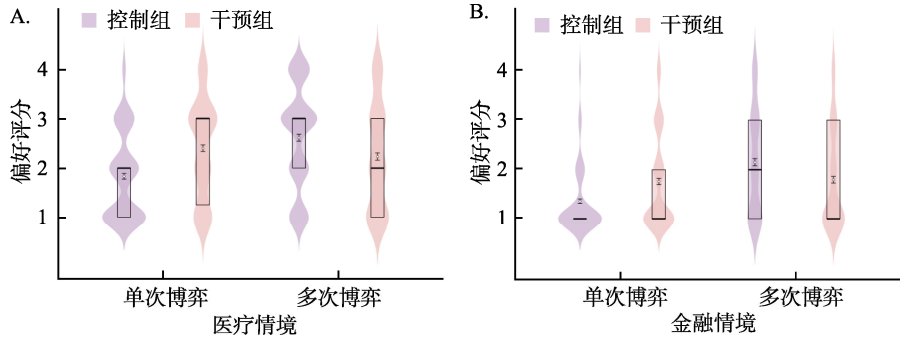

图 10 研究 3 实验 1 中医疗与金融情境在单次与多次博弈任务中的选择分布

注：在图 A 与图 B 中，纵轴评分采用 4 点量表，方案 A 代表确定选项，方案 B 代表风险选项。其中“1”表示“非常可能选择方案 A”，“2”表示“可能选择方案 A”，“3”表示“可能选择方案 B”，“4”表示“非常可能选择方案 B”。箱线图的横线自上而下分别为上四分位数、中位数、下四分位数，横线重合则表示对应统计值相同；白色圆点为均值 *M*，误差线为标准误 *SE*。每对分布图由左至右分别为控制组和干预组。

本类型”和“实验情境”为被试间变量, “博弈次数”为被试内变量。依据 G*Power v3.1 (Faul et al., 2007) 计算, 在中等效应量(Cohen’s f = 0.25), 显著性水平 α = 0.05, power = 0.80 条件下, 所需最小样本量为 48 人。通过 Credamo 平台共招募被试 245 人, 剔除未通过注意力检测题以及答题时间在正负 3 个标准差以外的 5 名不合格被试后, 最终纳入有效样本 240 名(男性 106 人, 女性 134 人), 平均年龄为 29.12 岁(SD = 9.19), 详细的人口学分布见网络版补充材料 E 表 S9。

4.2.2　材料与流程

实验流程与实验 1 保持一致。干预文本均由 GPT-4o 生成, 相关材料与问卷内容详见网络版补充材料 B 与补充材料 D。

4.2.3　结果与讨论

采用 CLMM 对数据进行分析, 以被试在 4 点量表上的选择结果为因变量, 将文本类型(干预 VS. 控制)博弈次数(多次 vs. 单次)、实验情境(内容创作 vs. 电商营销)及其交互项设为固定效应, 被试作为随机截距和随机斜率纳入模型。回归系数等统计量见表 6。

结果显示, 文本类型 × 博弈次数 × 实验情境三重交互项不显著(p = 0.916); 文本类型 × 博弈次数的二重交互项显著(β = −2.101, SE = 0.487, z = −4.314, p < 0.001, 95% CI = [−3.056, −1.147])。进一步简单效应分析发现(见图 11), 单次博弈中, 干预组相较于控制组更偏好风险选项(内容创作: β = 0.696, SE = 0.325, z = 2.143, p = 0.032, 95% CI = [0.059, 1.332]; 电商营销: β = 0.961, SE = 0.335, z = 2.867, p = 0.004, 95% CI = [0.304, 1.619]; 多次博弈中, 干预组则显著更偏好确定选项(内容创作: β = −1.335, SE = 0.333, z = −4.008, p < 0.001, 95% CI = [−1.988, −0.682]; 电商营销: β = −1.140, SE = 0.348, z = −3.273, p = 0.001, 95% CI = [−1.822, −0.457])。以上结果表明, 在单次博弈条件下, LLMs 的干预显著增强了人们对高期望值风险方案的偏好, 而在多次博弈条件下则促使人们更倾向于选择“确定”方案。

综上所述, 在医疗、金融、内容创作和电商营销情境中, 通过 LLMs 生成的干预文本均可干预人们在单次和多次博弈中固有的选择偏好。

表 6　文本类型、博弈次数与实验情境及其交互作用对风险决策倾向的累积链接混合模型回归结果

变量类型	回归系数(β)	标准差(SE)	z	95% CI	p 值
文本类型(干预 = 1, 控制 = 0)	0.961	0.335	2.867	[0.304, 1.619]	0.004
博弈次数(多次 = 1, 单次 = 0)	0.819	0.350	2.341	[0.133, 1.505]	0.019
实验情境(创作 = 1, 营销 = 0)	0.917	0.335	2.735	[0.260, 1.574]	0.006
文本类型 × 博弈次数	−2.101	0.487	−4.314	[−3.056, −1.147]	< 0.001
文本类型 × 实验情境	−0.266	0.464	−0.572	[−1.176, 0.645]	0.567
博弈次数 × 实验情境	−0.534	0.471	−1.132	[−1.457, 0.390]	0.257
文本类型 × 博弈次数 × 实验情境	0.070	0.666	0.105	[−1.235, 1.375]	0.916

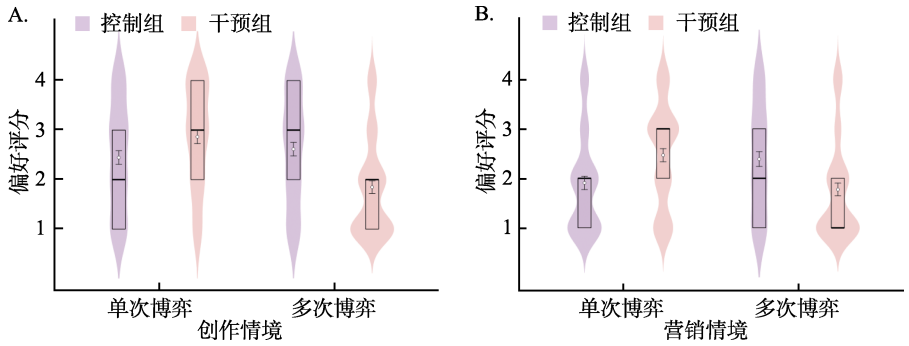

图 11　研究 3 实验 2 中内容创作与电商营销情境在单次与多次博弈任务中的选择分布

注: 在图 A 与图 B 中, 纵轴评分采用 4 点量表, 方案 A 代表确定选项, 方案 B 代表风险选项。其中“1”表示“非常可能选择方案 A”, “2”表示“可能选择方案 A”, “3”表示“可能选择方案 B”, “4”表示“非常可能选择方案 B”。箱线图的横线自上而下分别为上四分位数、中位数、下四分位数, 横线重合则表示对应统计值相同; 白色圆点为均值 M, 误差线为标准误 SE。每对分布图由左至右分别为控制组和干预组。

5 讨论

本文基于单次/多次博弈范式, 系统考察了 LLMs 对人类风险决策选择偏好和过程机制的模拟效度, 构建了 LLMs 生成策略文本的方法, 并检验人们对该文本的认可度及其对人类决策行为干预的有效性。发现: LLMs (GPT-3.5、4)能复现人类在单次(更风险规避)与多次(更风险寻求)博弈中的典型选择模式; LLMs (GPT-4o、DeepSeek-R1)能区分人们在单次/多次博弈中决策偏好差异的内部逻辑, 准确运用描述性和规范性理论分别生成单次和多次博弈差异性策略, 并获得人类认可; 在医疗、金融、内容创作和电商营销情境中, LLMs 生成的策略文本能够干预人们在单次与多次博弈中固有的选择偏好。

5.1 LLMs 对风险决策选择偏好的模拟

本研究基于单次/多次博弈范式发现, LLMs 能复现人类在单次和多次博弈下的选择模式, 但存在细节差异。具体地, 当面对期望值相同的赌博时, LLMs 在单次博弈中的参与率显著低于多次博弈(Benartzi & Thaler, 1999; Kumar & Lim, 2008)。然而, 与 Chen 等人(2023)的发现类似, 无论是单次还是多次博弈, LLMs 的参与率(均值约 95%)均比人类更高(均值约 53%; Redelmeier & Tversky, 1992), 即更加风险寻求。这一结果可能源于模型的训练语料。研究所选 GPT 系列模型的部分训练语料来源于 Common Crawl、Wikipedia 及投资论坛等文本(Brown et al., 2020), 其中不仅包括人类面对风险与收益的主观认知, 也包含了风险与收益的客观规律。当模型学习前者时, 可能习得与人相似的风险偏好, 但由于其无法真正体验损失带来的负面情绪, 当从客观规律中学习时, 可能形成比人类更强的风险寻求偏好。

通过揭示 LLMs 和人类在风险决策中的相似和差异性, 本研究在理论层面有利于区分数据驱动与生物认知的风险决策机制, 为建立人工智能决策的评价框架提供证据。在实践层面, 有助于优化人机协作模式, 实现优势互补。同时, 本研究提示, 在金融、医疗等高风险领域直接采用 LLMs 建议时需建立风险校准机制(如阈值控制或人工复核)。

5.2 LLMs 对人类风险决策策略的理解

本研究在医疗、金融、内容创作与电商营销四类情境中, 采用单次/多次博弈范式系统检验了 LLMs 对人类风险决策策略的理解与生成能力。其中, “理解”是指 LLMs 具备区分单次/多次博弈特征, 还能依据相应理论框架(描述性 vs. 规范性)生成匹配的行为偏好与策略文本的能力。这反映了 LLMs 对人类信息处理内在机制(包括逻辑推理、常识运用及因果推断等认知过程)的推理能力(Mischler et al., 2024; Wei et al., 2022), 而非仅停留在行为层面对人类决策模式的表面“模拟” (Argyle et al., 2023; Binz & Schulz, 2023)。

在多次博弈中, LLMs 捕捉到人类对高期望值风险选项的偏好, 生成符合规范性理论框架, 强调通过概率加权求和的理性决策策略。在单次博弈中, LLMs 识别出人类对确定选项的偏好, 其生成聚焦于“确定性”和“稳健性”维度的策略, 符合描述性理论的启发式决策(Brandstätter et al., 2006)和齐当别决策模型(Li, 2004)假设: 忽略/“齐同”掉差异较小的维度, 基于关键维度的比较做出决策。该发现不仅验证了 Sun 等人(2014)关于人类在单次和多次博弈中采用不同决策策略的理论假设, 也揭示了 LLMs 能够分别解释单次和多次博弈中的行为的不同逻辑。

通过 LLMs 的策略外化, 本研究不仅为风险决策的认知过程提供了直接的计算建模证据, 帮助研究者理清风险决策的内部逻辑, 有助于揭示传统研究中难以观测的决策黑箱, 也为运用 AI 干预风险决策奠定了理论和方法基础。

5.3 LLMs 对风险决策的干预

本研究创新性地利用 LLMs 的内容生成能力, 构建了基于策略文本的风险决策干预方法。通过医疗、金融、内容创作和电商营销四类情境的单次和多次博弈实验, 系统验证了该方法的有效性及其作用机制。

研究建立了三阶段的策略或干预文本生成流程, 并运用其来进行决策干预: (1) 利用 DeepSeek-R1 和 GPT-4o 获取决策依据并运用 ITA 提炼核心策略主题; (2) 基于策略主题, 生成策略文本; (3) 对文本进行内容评估。研究的干预逻辑基于“偏好反转”原理: 即在单次博弈中引导选择高期望值的风险选项, 在多次博弈中引导选择低期望值的确定选项。结果显示, 与对照组相比, LLMs 生成的文本在四类情境下均能显著改变人们固有的决策偏好: 在单次博弈中, 干预文本通过强调“期望收益最大化”, 符合期望价值最大化理论(expected value theory, Pascal, 1670)的核心逻辑, 引导决策者理性地基于期望价值计算来进行决策。相应地, 在多次博弈中,

与启发式决策(Brandstätter et al., 2006)和齐当别决策模型(Li, 2004)的逻辑一致, 干预文本则通过凸显风险选项的潜在损失(“每次损失的风险较大”)和确定选项的稳定性(“确保每次决策都带来稳定的回报”), 引导被试基于关键维度(如概率)差异来简化决策。

值得注意的是, LLMs 的干预效果在不同博弈情境间存在差异。例如, 在单次与多次博弈中, 干预组与对照组的个体对目标选项的选择比例变化量存在差异。这表明, LLMs 生成的干预文本对决策行为的影响程度, 可能受个体特征或对具体决策情境卷入度等因素所调节。

综上, 传统的风险决策多基于人们在多次博弈决策中的思路(符合规范性理论的“加权求和”策略)来指导人们单次博弈决策中的行为, 以致(1) 规范性理论与描述性理论分离; (2) 未能真正有效干预人们在真实生活中的风险决策。本文发现, LLMs 能够深入理解单次与多次博弈的决策逻辑, 并生成与人类思维高度一致、认可度较高的策略文本。同时该策略文本在行为干预中表现出出色的效果。

5.4　研究贡献

本文首次在单次/多次博弈范式中系统验证 LLMs 对行为偏好的模拟能力、对决策的理解力, 以及 LLMs 生成文本对人类行为的干预作用, 具有以下三方面理论贡献:

首先, 通过对比分析 LLMs 与人类在单次和多次风险决策任务中的选择偏好, 系统检验了大语言模型对人类风险决策行为的模拟效度。其重要性在于: (1) 揭示了 LLMs 能够同时捕捉规范性理论的理性计算原则和描述性理论的经验决策模式, 从而为检验决策理论分歧点提供了独特的实验技术; (2) LLMs 即便在不具备人类认知和情感的前提下, 仍能复现人类在单次与多次博弈中的行为偏好反转现象, 在认知层面揭示人类的单次和多次风险决策加工过程可能受到语言系统固有模式(linguistic patterning)与语义联结关系(semantic association)的驱动(Suri et al., 2024)。

其次, 本文深入探讨了 LLMs 在单次和多次风险决策任务中的推理机制: LLMs 在风险决策中展现出类人的双模式推理机制, 即在多次博弈条件下更符合规范性理论的理性计算, 而在单次博弈中则转向描述性理论的启发式策略, 这种动态切换能力使得 AI 能够重现人类决策的理性与直觉的辩证统一。当前, 随着人工智能模型能力的持续提升, 单纯依赖任务表现的评估方式已难以满足理解其内部机制的需求, 解析 LLMs 的推理路径对于揭示其“如何”做出类似人类决策至关重要, 同时也为风险决策领域的理论建模提供了新的证据。

最后, 本文创新性地构建了基于“LLMs 模拟决策——生成决策依据——归纳性主题分析——生成分情境的干预文本”这一闭环干预方式, 并验证了其对改变人类风险偏好的有效性。该发现(1)拓展了行为决策理论中人与 AI 交互的研究维度, 揭示出在单次和多次博弈的风险决策中, 即使脱离传统认知与情感过程, 语言中的模式和联结仍可以引发人类相应的决策反应, 进而影响决策的过程和结果。(2)为建立合理的“人-LLMs”互动规则提供了新的方向。通过对 LLMs 生成的决策依据进行文本分析, 解析出影响人类单次和多次博弈决策偏好的关键语言特征, 有助于提示词设计者(Prompt Engineering, Khalid & Witmer, 2025)优化交互策略, 从而引导 LLMs 生成更具实用可靠的输出内容。

已有关于 LLMs 的决策研究多基于经典的决策范式, 侧重于从行为层面检验 LLMs 与人类的相似性。与之相比, 本研究基于风险决策中经典的“单次/多次博弈”范式, 不仅考察了 LLMs 决策表现和机制的类人性, 以较为完整和系统的研究路径, 探索了 LLMs 对人类风险决策行为机制的理解能力和干预效果。

在实践层面, 本研究首先揭示了 LLMs 能清晰、一致地呈现复杂策略, 此优势使其可突破传统访谈的局限, 作为一种“理论建构工具”, 未来或可通过与之进行“结构化访谈”来高效探索决策机制, 推动理论发展。其次, 也为基于生成式人工智能的行为助推技术开发提供了重要参考, 不仅为矫正非理性决策行为提供了新思路, 也对金融投资、健康医疗、电商营销等领域的智能辅助决策系统设计具有指导价值。

5.5　研究不足与展望

本研究存在以下不足。

首先, 本研究仅基于 ChatGPT、DeepSeek 两种模型, 在中文语境开展研究, 没有系统比较不同 LLMs 在模拟和生成决策策略等层面的差异, 并基于不同语言探索研究结果中可能存在的文化差异。由于各模型在训练语料(Demszky et al., 2023; Thapa & Adhikari, 2023)、对齐策略(Zhao et al., 2023)等方面存在差异, 导致其在同一任务中的表

现会有所差异。研究使用 Claude 和 LLaMA 模型复现了部分实验结果, 以进行跨模型稳健性检验, 但结论仍可能存在模型特异性偏差。未来可引入 Claude、豆包、通义千问等主流和新兴模型, 在不同语言中进一步探索其风险决策的差异, 并检验研究结论的普适性。

其次, LLMs 的干预效果可能受到情境、个体差异(如人格特质、决策风格等)、文化背景等因素的影响。例如, 大五人格中的神经质特质、以及集体主义/个人主义的背景均可能调节风险倾向(Liu et al., 2021; Xia et al., 2019), 从而影响风险决策的干预效果。此外, 尽管 LLMs 相较传统干预具有个性化与智能化优势, 本文尚未直接比较两者效果。未来研究一方面可以进一步优化 LLMs 的文本生成算法, 控制或整合更多个体特征信息, 以提升干预的精度和生态效度, 推动个性化决策支持系统的发展。此外, 也有必要在相同任务条件下, 系统比较 LLMs 与传统信息框架干预的作用效果, 以界定其独特价值与优势。

再者, LLMs 的干预效果可能源于文本中嵌入的策略机制(如期望效用、齐当别等理论机制等), 也可能受到文本效应(如语言表达的专业性、简洁性、情绪色彩等)影响。本研究虽在一定程度上为策略机制的干预效应提供了证据支持, 但未能排除文本效应的影响。未来研究可进一步操纵语言风格或情绪色彩等变量, 以更严谨的实验设计来揭示 LLMs 文本干预作用的内在机制。

6 结论

本研究首先探讨了人工智能大语言模型对风险决策选择偏好和过程机制的模拟, 以及对人类决策策略的理解, 然后探索了如何运用大语言模型生成影响个体单次和多次风险决策选择偏好的干预文本, 并验证了其干预效果。发现:

(1) LLMs 能模拟人类在单次和多次博弈中选择偏好的差异模式, 但整体表现出比人类更高的风险偏好水平;

(2) LLMs 能准确区分单次和多次博弈背后的决策逻辑差异, 并分别运用描述性理论和规范性理论生成相应的策略, 且这些策略获得较高的人类认可度;

(3) 通过对 LLMs 决策过程实施归纳性主题分析所提炼出的策略文本, 能够有效干预人们在单次和多次博弈中固有的选择偏好模式。

致谢: 感谢卡耐基梅隆大学蔡晓红博士后、中国科学院心理研究所李纾研究员对本文的研究设计、数据分析以及论文撰写的指导帮助。

参 考 文 献

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... McGrew, B. (2023). GPT-4 technical report. *arXiv preprint*. https://doi.org/10.48550/arXiv.2303.08774

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 337−371). PMLR. https://proceedings.mlr.press/v202/aher23a.html

Altay, S., Hacquin, A. S., Chevallier, C., & Mercier, H. (2023). Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, *29*(1), 52−62. https://doi.org/10.1037/xap0000400

Anderson, M. A. B., Cox, D. J., & Dallery, J. (2023). Effects of economic context and reward amount on delay and probability discounting. *Journal of the Experimental Analysis of Behavior*, *120*(2), 204−213. https://doi.org/10.1002/jeab.868

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337−351. https://doi.org/10.1017/pan.2023.2

Arora, C., Sayeed, A. I., Licorish, S., Wang, F., & Treude, C. (2024). Optimizing large language model hyperparameters for code generation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2408.10577

Barberis, N., & Huang, M. (2009). Preferences with frames: A new utility specification that allows for the framing of risks. *Journal of Economic Dynamics and Control*, *33*(8), 1555−1576. https://doi.org/10.1016/j.jedc.2009.01.009

Benartzi, S., & Thaler, R. H. (1999). Risk aversion or myopia? Choices in repeated gambles and retirement investments. *Management Science, 45*(3), 364−381. https://doi.org/10.1287/mnsc.45.3.364

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409−432. https://doi.org/10.1037/0033-295X.113.2.409

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137−164). Sage Publications.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877−1901. https://doi.org/10.48550/arXiv.2005.14165

Carvalho, T., Negm, H., & El-Geneidy, A. (2024). A comparison of the results from artificial intelligence-based and human-based transport-related thematic analysis. *Findings*. https://doi.org/10.32866/001c.94401

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The

emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, *120*(51), e2316205120. https://doi.org/10.1073/pnas.2316205120

Choi, S., Kang, H., Kim, N., & Kim, J. (2025). How does artificial intelligence improve human decision-making? Evidence from the AI-powered Go program. *Strategic Management Journal, 46*(6), 1523−1554. https://doi.org/10.1002/smj.3694

Christensen, R. H. B. (2023). *ordinal: Regression models for ordinal data (R package version 2023.12-4.1)* [Computer software]. https://CRAN.R-project.org/package=ordinal

Coda-Forno, J., Witte, K., Jagadish, A. K., Binz, M., Akata, Z., & Schulz, E. (2023). Inducing anxiety in large language models can induce bias. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.11111

Dai, S. C., Xiong, A., & Ku, L. W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint.* https://doi.org/10.48550/arXiv.2310.15100

de Kok, T. (2025). ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Management Science, 71*(9), 7888−7906. https://doi.org/10.1287/mnsc.2023.03253

de Varda, A. G., Saponaro, C., & Marelli, M. (2025). High variability in LLMs' analogical reasoning. *Nature Human Behaviour*, *9*(7), 1339−1341. https://doi.org/10.1038/s41562-025-02224-3

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., … Zhang, Z. (2025). Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2501.12948

Deiana, G., Dettori, M., Arghittu, A., Azara, A., Gabutti, G., & Castiglia, P. (2023). Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*, *11*(7), 1217. https://doi.org/10.3390/vaccines11071217

Deiner, M. S., Honcharov, V., Li, J., Mackey, T. K., Porco, T. C., & Sarkar, U. (2024). Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study. *JMIR Infodemiology*, *4*(1), e59641. https://doi.org/10.2196/59641

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688−701. https://doi.org/10.1038/s44159-023-00241-5

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597−600. https://doi.org/10.1016/j.tics.2023.04.008

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., & Yang, M. (2024). Longrope: Extending LLM context window beyond 2 million tokens. *arXiv preprint*. https://doi.org/10.48550/arXiv.2402.13753

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175−191. https://doi.org/10.3758/BF03193146

Ferguson, S. A., Aoyagui, P. A., & Kuzminykh, A. (2023). Something borrowed: Exploring the influence of AI-generated explanation text on the composition of human explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1−7). ACM. https://doi.org/10.1145/3544549.3585727

Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, *43*(4), 709−722. https://doi.org/10.1287/mksc.2023.0306

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108−1109. https://doi.org/10.1126/science.adi1778

Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. *International Journal of Information Management Data Insights*, *4*(1), 100232. https://doi.org/10.1016/j.jjimei.2024.100232

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*(10), 833−838. https://doi.org/10.1038/s43588-023-00527-x

Hebenstreit, K., Praas, R., Kiesewetter, L. P., & Samwald, M. (2024). A comparison of chain-of-thought reasoning strategies across datasets and models. *PeerJ Computer Science*, *10*, e1999. https://doi.org/10.7717/peerj-cs.1999

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517−523. https://doi.org/10.1016/j.tics.2009.09.004

Jiao, L., Li, C., Chen, Z., Xu, H., & Xu, Y. (2025). When AI “possesses” personality: Roles of good and evil personalities influence moral judgment in large language models. *Acta Psychologica Sinica*, *57*(6), 929−946. https://doi.org/10.3724/SP.J.1041.2025.0929

[焦丽颖, 李昌锦, 陈圳, 许恒彬, 许燕. (2025). 当 AI“具有”人格: 善恶人格角色对大语言模型道德判断的影响. *心理学报*, *57*(6), 929−946.]

Jin, H. J., & Han, D. H. (2014). Interaction between message framing and consumers' prior subjective knowledge regarding food safety issues. *Food Policy*, *44*, 95−102. https://doi.org/10.1016/j.foodpol.2013.10.007

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, *35*, 11785−11799. https://doi.org/10.48550/arxiv.2202.12299

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263−292. https://doi.org/10.2307/1914185

Karinshak, E., Hu, A., Kong, K., Rao, V., Wang, J., Wang, J., & Zeng, Y. (2024). LLM-globe: A benchmark evaluating the cultural values embedded in LLM output. *arXiv preprint*. https://doi.org/10.48550/arXiv.2411.06032

Karinshak, E., Liu, S. X., Park, J. S., & Hancock, J. T. (2023). Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1−29. https://doi.org/10.1145/3579592

Katz, A., Fleming, G. C., & Main, J. (2024). Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. *arXiv preprint*. https://doi.org/10.48550/arXiv. 2410.03721

Kelton, A. S., Pennington, R. R., & Tuttle, B. M. (2010). The effects of information presentation format on judgment and decision making: A review of the information systems research. *Journal of Information Systems*, *24*(2), 79−105. https://doi.org/10.2308/jis.2010.24.2.79

Khalid, M. T., & Witmer, A. P. (2025). Prompt engineering for large language model-assisted inductive thematic analysis. *arXiv preprint*. https://doi.org/10.48550/arXiv.2503.22978

Kumar, A., & Lim, S. S. (2008). How do decision frames

influence the stock investment choices of individual investors? *Management Science*, *54*(6), 1052−1064. https://doi.org/10.1287/mnsc.1070.0845

Lehr, S. A., Caliskan, A., Liyanage, S., & Banaji, M. R. (2024). ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences, 121*(35), e2404328121. https://doi.org/10.1073/pnas.2404328121

Lenth, R. V. (2025). *Emmeans: Estimated marginal means, aka least-squares means* (R package version 1.11.0) [Computer software]. https://doi.org/10.32614/CRAN.package. emmeans

Li, S. (2004). A behavioral choice model when computational ability matters. *Applied Intelligence*, *20*(2), 147−163. https://doi.org/10.1023/B:APIN.0000013337.01711.c7

Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science*, *10*(8), 230658. https://doi.org/10.1098/rsos.230658

Lin, Z. (2024). How to write effective prompts for large language models. *Nature Human Behaviour*, *8*(4), 611−615. https://doi.org/10.1038/s41562-024-01847-2

Lin, Z. (2025). Techniques for supercharging academic writing with generative AI. *Nature Biomedical Engineering*, *9*(4), 426−431. https://doi.org/10.1038/s41551-024-01185-8

Liu, N., Zhou, L., Li, A. M., Hui, Q. S., Zhou, Y. R., & Zhang, Y. Y. (2021). Neuroticism and risk-taking: the role of competition with a former winner or loser. *Personality and Individual Differences, 179*, 110917. https://doi.org/10.1016/j.paid.2021.110917

Liu, S. X., Yang, J. Z., & Chu, H. R. (2019). Now or future? Analyzing the effects of message frame and format in motivating Chinese females to get HPV vaccines for their children. *Patient Education and Counseling, 102*(1), 61−67. https://doi.org/10.1016/j.pec.2018.09.005

Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes*, *65*(3), 179−189. https://doi.org/10.1006/obhd.1996.0017

Lu, J., Chen, Y., & Fang, Q. (2022). Promoting decision satisfaction: The effect of the decision target and strategy on process satisfaction. *Journal of Business Research*, *139*, 1231−1239. https://doi.org/10.1016/j.jbusres.2021.10.056

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence, 6*(10), 1467−1477. https://doi.org/10.1038/s42256-024-00925-4

Morreale, A., Stoklasa, J., Collan, M., & Lo Nigro, G. (2018). Uncertain outcome presentations bias decisions: Experimental evidence from Finland and Italy. *Annals of Operations Research*, *268*(1−2), 259−272. https://doi.org/10.1007/s10479-016-2349-3

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Park, P. S. (2024). Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*, *56*(6), 5754−5770. https://doi.org/10.3758/s13428-023-02307-x

Pascal, B. (1670). *Pensées* (W. F. Trotter, Trans.). Retrieved Nov. 22, 2018, from https://sourcebooks.fordham.edu/mod/1660pascal-pensees.asp

Pavey, L., & Churchill, S. (2014). Promoting the avoidance of high-calorie snacks: Priming autonomy moderates message framing effects. *PLoS One*, *9*(7), e103892. https://doi.org/10.1371/journal.pone.0103892

Pawel, S., Consonni, G., & Held, L. (2023). Bayesian approaches to designing replication studies. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000604

Peng, L., Guo, Y., & Hu, D. (2021). Information framing effect on public's intention to receive the COVID-19 vaccination in China. *Vaccines*, *9*(9), 995. https://doi.org/10.3390/vaccines9090995

Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, *3*(6), 435−448. https://doi.org/10.1017/s1930297500000012

Popovic, N. F., Pachur, T., & Gaissmaier, W. (2019). The gap between medical and monetary choices under risk persists in decisions for others. *Journal of Behavioral Decision Making*, *32*(4), 388−402. https://doi.org/10.1002/bdm.2121

Prescott, M. R., Yeager, S., Ham, L., Saldana, C. D. R., Serrano, V., Narez, J., … Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*(1), e54482. https://doi.org/10.2196/54482

Qin, X., Huang, M., & Ding, J. (2024). AITurk: Using ChatGPT for social science research. *PsyArXiv*. https://doi.org/10.31234/osf.io/xkd23

Redelmeier, D. A., & Tversky, A. (1992). On the framing of multiple prospects. *Psychological Science*, *3*(3), 191−193. https://doi.org/10.1111/j.1467-9280.1992.tb00025.x

Reeck, C., Mullette-Gillman, O. A., McLaurin, R. E., & Huettel, S. A. (2022). Beyond money: Risk preferences across both economic and non-economic contexts predict financial decisions. *PLoS One*, *17*(12), e0279125. https://doi.org/10.1371/journal.pone.0279125

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88−95. https://doi.org/10.1080/21507740.2020.1740350

Samuelson, P. A. (1963). Risk and uncertainty: A fallacy of large numbers. *Scientia*, *98*, 108−113.

Scarffe, A., Coates, A., Brand, K., & Michalowski, W. (2024). Decision threshold models in medical decision making: A scoping literature review. *BMC Medical Informatics and Decision Making*, *24*(1), 273. https://doi.org/10.1186/s12911-024-02681-2

Shahid, N., Rappon, T., & Berta, W. (2019). Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS One*, *14*(2), e0212356. https://doi.org/10.1371/journal.pone.0212356

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559−569. https://doi.org/10.1177/0956797614567341

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., … Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285−1295. https://doi.org/10.1038/s41562-024-01882-z

Sun, H. Y., Rao, L. L., Zhou, K., & Li, S. (2014). Formulating an emergency plan based on expectation-maximization is one thing, but applying it to a single case is another. *Journal of Risk Research*, *17*(7), 785−814. https://doi.org/10.1080/13669877.2013.816333

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do

large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General, 153*(4), 1066−1075. https://doi.org/10.1037/xge0001547

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon.

Thapa, S., & Adhikari, S. (2023). ChatGPT, Bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering, 51*(12), 2647−2651. https://doi.org/10.1007/s10439-023-03284-0

Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 713−717. https://doi.org/10.1037/0278-7393.9.4.713

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (2nd rev. ed.). Princeton University Press.

Wang, Y., Zhang, J., Wang, F., Xu, W., & Liu, W. (2023). Do not think any virtue trivial, and thus neglect it: Serial mediating role of social mindfulness and perspective taking. *Acta Psychologica Sinica, 55*(4), 626−641. https://doi.org/10.3724/SP.J.1041.2023.00626

[王伊萌, 张敬敏, 汪凤炎, 许文涛, 刘维婷. (2023). 勿以善小而不为: 正念与智慧——社会善念与观点采择的链式中介. *心理学报*, *55*(4), 626−641. https://doi.org/10.3724/SP.J.1041.2023.00626]

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526−1541. https://doi.org/10.48550/arXiv.2212.09196

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263−290. https://doi.org/10.1002/bdm.414

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824−24837. https://doi.org/10.48550/arxiv.2201.11903

Xia, D., Li, Y., He, Y., Zhang, T., Wang, Y., & Gu, J. (2019). Exploring the role of cultural individualism and collectivism on public acceptance of nuclear energy. *Energy Policy*, *132*, 208−215. https://doi.org/10.1016/j.enpol.2019.05.014

Xia, D., Song, M., & Zhu, T. (2025). A comparison of the persuasiveness of human and ChatGPT generated pro-vaccine messages for HPV. *Frontiers in Public Health*, *12*, 1515871. https://doi.org/10.3389/fpubh.2024.1515871

Yuan, Y., Jiao, W., Wang, W., Huang, J. T., He, P., Shi, S., & Tu, Z. (2023). Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint*. https://doi.org/10.48550/arXiv.2308.06463

Zhang, J., Li, H. A., & Allenby, G. M. (2024). Using text analysis in parallel mediation analysis. *Marketing Science,* 43(5), 953−970. https://doi.org/10.1287/mksc.2023.0045

Zhang, Y., Huang, F., Mo, L., Liu, X., & Zhu, T. (2025). Suicidal ideation data augmentation and recognition technology based on large language models. *Acta Psychologica Sinica*, *57*(6), 987−1000. https://doi.org/10.3724/SP.J.1041.2025.0987

[章彦博, 黄峰, 莫柳铃, 刘晓倩, 朱廷劭. (2025). 基于大语言模型的自杀意念文本数据增强与识别技术. *心理学报*, *57*(6), 987−1000.]

Zhao, F., Yu, F., & Shang, Y. (2024). A new method supporting qualitative data analysis through prompt generation for inductive coding. *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 164−169. https://doi.org/10.1109/IRI62200.2024.00043

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... Wen, J. R. (2023). A survey of large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2303.18223

Large language models capable of distinguishing between single and repeated gambles: Understanding and intervening in risky choice

ZHOU Lei[1], LI Litong[1], WANG Xu[1], OU Huafeng[1], HU Qianyu[1], LI Aimei[2], GU Chenyan[1]

([1] *School of Management, Guangdong University of Technology, Guangzhou* 510520, *China*)

([2] *School of Management, Jinan University, Guangzhou* 510632, *China*)

Abstract

Risky choice (RC) is a common and important form of decision making in daily life. Its theoretical development primarily follows two major theories: normative theory and descriptive theory. The paradigms of single- and repeated-play gambles can provide an effective framework for distinguishing between the theories. However, prior research lacks direct observations of the decision-making process, which can limit the deep understanding of individual behaviour and hinder the development of effective behavioural interventions. In recent years, large language models (LLMs) have demonstrated highly human-like characteristics by not only simulating human preferences in behavioural performance but also exhibiting similar reasoning pathways. This offers a promising solution to the aforementioned limitations. This study, which is grounded in the classic RC paradigms of single versus repeated gambles, investigates the capability of LLMs to simulate and understand risk preferences and decision-making processes. Specifically, this study explores the potential of LLMs' understanding of decision strategies to generate intervention texts and evaluates their effectiveness in influencing human decisions.

This work comprises three studies. In Study 1, GPT-3.5 and GPT-4 were employed to simulate human responses to gambling decisions under nine probability conditions (with constant expected value), which generated a total of 3, 600 responses across single and repeated gamble scenarios. In Study 2, LLM-generated strategies were constructed through a three-stage process (decision rationale extraction, strategy generation and quality evaluation), then the human participants were required to complete decision-making tasks in two experiments: Experiment 1 replicated the medical/financial scenarios (N = 349, N_{male} = 174, M_{age} = 21.79) of Sun et al. (2014) in a 2 (context: medical vs. financial) × 2 (application frequency: single vs. repeated) within-subjects design, and Experiment 2 examined digital contexts with a 2 (context: content creation vs. e-commerce marketing) × 2 (frequency: single vs. repeated) mixed design (context as between subjects). Subsequently, DeepSeek-R1 was used to perform the same tasks and generate strategy texts through the three-stage process. Finally, the participants were instructed to evaluate their acceptance of the LLM-generated strategies. Study 3 extended the Study 2 methodology to determine whether the LLM-generated intervention texts could reverse the participants' classic choice preference across the single versus repeated gamble scenarios. The Study 2 experimental contexts (Experiment 1: medical vs. financial, N = 460, N_{male} = 205, M_{age} = 21.80; Experiment 2: content creation vs. e-commerce marketing, N = 240, N_{male} = 106, M_{age} = 29.12) were mirrored in Study 3, in which strategically designed intervention texts were presented during the decision-making tasks to test their capacity to modify the participants' inherent risk preference between the single and repeated gamble conditions and evaluate the persuasive efficacy of LLM-generated strategies on human decision biases.

Study 1 shows that the LLMs (GPT-3.5 and GPT-4) can successfully replicate the typical human pattern of risk aversion in single-play scenarios and risk seeking in repeated-play scenarios, though both models demonstrated an overall stronger tendency toward risk seeking compared with the human participants. Study 2 demonstrates that the human participants preferred low-EV certain options in single-play contexts and high-EV risky options in repeated-play contexts in both experiments. The participants also showed high agreement with the strategies generated by the LLMs in different scenarios. Study 3 confirms that the LLM-generated intervention texts can significantly influence the participants' choice tendency in all four scenarios, with strong intervention effects observed in the single-play contexts. The LLM intervention strategies are characterised by reliance on expected value computations (normative) when promoting RCs and emphasis on certainty and robustness (descriptive) when promoting safe choices.

In summary, this study demonstrates that (1) LLMs can effectively simulate context-dependent human preferences in RC, particularly the shift from risk aversion in single plays to risk seeking in repeated plays; (2) LLMs can distinguish between the logic underlying single and repeated gambles and apply normative and descriptive reasoning accordingly to externalise decision strategies; and (3) the decision strategies extracted from LLM-generated reasoning can be used to construct effective intervention texts that can alter human preferences in classic risk decision tasks, thereby validating the feasibility and effectiveness of an LLM-based cognitive intervention pathway. This study offers a new technological paradigm for AI-assisted decision intervention and expands the application boundary of LLMs to human cognitive process modelling and regulation.

Keywords risk decision-making, single- vs. repeated-play gambles, large language models, decision strategy, intervention

补充材料 A

A1　归纳性主题分析的提示词设计

你是一名专业的文本分析专家, 我进行了一项实验, 有生成一些决策依据内容, 我需要你按照我的要求进行分析。我的实验内容如下。

研究目的:

本文旨在探讨 LLMs 在进行单次与多次博弈的决策任务时, 是否会表现出与人类一致的决策现象, 还是会呈现不同的决策结果。通过此研究, 旨在揭示 LLMs 在此类决策任务中的决策机制, 并为未来如何利用 LLMs 辅助人类决策提供指导。

研究设计:

本文选取了 deepseek 作为研究对象。在与 LLMs 进行交互时, 每个界面包含系统(system)、用户(user)和助手(assistant)三个角色。系统消息用于操作命令, 用户提示为直接查询, 助手则是 GPT 的反馈。

系统消息(Human): 你将进行多轮决策任务, 每一轮都必须提供答案。如果你不提供答案, 我将假设你是随机选择的。我希望你能像一个人一样做出决策。请等待我的提示, 然后告诉我你的答案, 并简要解释你的决策理由。

决策任务: 本研究将比较单次博弈和多次博弈两种情境下的决策表现, 博弈内容为经典的赌博游戏: 设想一个赌博游戏, 有 50%的概率赢得¥2000, 另有 50%的概率输掉¥500。设定如下:

- 单次博弈: 现在, 假如您有机会进行这次赌博。您会选择参与吗? 请回答“参加”或“不参加”。
- 多次博弈: 现在, 假如您有机会进行五次赌博。您会选择参与吗? 请回答“参加”或“不参加”。

接下来, 我将给你 deepseek 进行单次博弈/多次博弈[5]的决策依据的内容, 请你对决策依据内容文本进行分析, 提炼出其为何会做出该决策的主题维度。请根据以下要求进行:

1. 清晰简洁的标题: 总结主题的核心内容。

2. 简短的定义: 为该主题提供简洁明了的解释。

3. 两个代表性例子: 从数据集中选择两个典型例子(直接引用), 说明该主题。确保主题具有独特性、不重叠, 且共同提供对数据集的广泛理解。

请确保分析依据包括不同的维度, 如风险偏好、回报预期等, 确保主题能够清晰地反映出 deepseek 在单次博弈/多次博弈决策中的考虑因素。接下来我将会提供决策依据供你分析。

[5] 在提取单次博弈的决策依据时, 提示词限定到单次博弈; 在提取多次博弈的决策依据时, 提示词限定到多次博弈。

A2　归纳性主题分析的结果

A2.1　在单次博弈条件中所提取到的策略主题

1. 主题: 期望值与数学理性决策

定义:

期望值作为决策依据, 衡量了赌博游戏中每次参与的长期收益或损失。根据概率和回报的计算, 理性决策应基于期望值, 而非单次博弈的波动, 选择预期收益最大化的选项。

代表性例子:

- “期望收益计算为 0.1×10000–0.9×278 = 749.8 元, 长期来看是正收益。尽管有 90%概率损失, 但高回报的数学期望支持理性决策。”
- “期望收益为 0.2×5000–0.8×313 = 749.6 元。虽然 80%概率损失 313 元看似风险较高, 但显著的正期望值意味着长期参与将获得稳定收益。”

2. 主题: 风险承受能力与决策偏好

定义:

决策者的风险承受能力决定了其是否愿意接受单次赌博中的潜在损失。风险偏好者可能选择承担短期波动, 追求长期的回报, 而风险厌恶者则可能倾向于避免较高概率的损失。

代表性例子:

- “尽管有 90%的概率输掉 278 元, 但期望值计算为 0.1×10000–0.9×278 = 749.8 元, 显著高于损失, 理性决策者会选择参与, 只要他们能够承受 278 元的损失。”
- “尽管有 80%的概率输掉 313 元, 但期望值为正, 理性决策应考虑数学期望, 而非单纯依据短期波动。”

3. 主题: 损失厌恶与心理效应

定义:

根据行为经济学中的前景理论, 决策者对损失的敏感度通常高于对等额收益的敏感度, 这种损失厌恶效应可能导致决策者回避高风险赌博, 即使从数学上看, 期望值为正。

代表性例子:

- “期望值为: 0.9×1111–0.1×2500 = 749.9 元, 尽管存在 10%概率损失 2500 元, 但心理上的损失厌恶效应可能会让参与者选择回避此赌博。”
- “虽然期望值为正, 但单次博弈的高失败概率(90%)和潜在的损失(278 元)可能引发风险规避者的情绪反应, 最终选择不参与。”

4. 主题: 高回报与低概率事件

定义：

在一些博弈任务中, 尽管存在较高的亏损概率, 但较小概率的高回报(如高额奖金)可能吸引风险偏好者参与。这种高回报低概率事件的吸引力可以超越短期的亏损。

代表性例子：

- “期望值为 0.1×10000–0.9×278 = 749.8 元, 尽管 90%的概率损失 278 元, 但 10%的高回报(10000 元)足以让风险偏好者参与。”
- “虽然有 80%的概率输掉 313 元, 但 20%概率获得 5000 元的高回报能覆盖风险, 正期望值支持参与。”

这些主题维度清晰地反映了 deepseek 在做出单次博弈决策时所考虑的因素, 包括期望值的计算、风险承受能力、损失厌恶的心理效应以及对高回报低概率事件的反应。这些因素共同作用于其决策过程, 形成了基于理性预期收益的决策模式。

A2.2 在多次博弈条件中所提取到的策略主题

1. 期望收益最大化

定义：

理性决策者基于期望值计算做出决策, 期望值为正时, 选择参与以获得长期回报。

代表性例子：

- “该赌博的期望值为每次 0.1×10000–0.9 × 278≈ 749.8 元。五次的总期望收益为 3749 元, 长期看具有正向收益。”
- “每次赌博的期望收益为 0.2×5000–0.8 × 313≈ 749.6 元, 五次总期望收益达 3748 元。”

2. 风险与收益的平衡

定义：

决策者会评估赌博中的风险(损失概率)与收益(潜在赢利), 并通过风险承受能力决定是否参与。

代表性例子：

- “尽管五次全输的概率高达 59% (损失 1390 元), 但至少赢一次的概率约 41%, 有机会通过单次高回报覆盖累计损失。”
- “虽然每次赌博有 70%的概率亏损 357 元, 但 30%的赢率对应 3333 元的高回报, 使得单次期望值高达 750 元。”

3. 多次博弈的累计效应

定义：

多次独立博弈的期望收益通过大数定律趋向长期稳定, 尽管短期内可能存在较大波动, 但长期来看回报更有保障。

代表性例子：

- “五次独立赌博的期望总收益约 3750 元, 且十次中至少赢一次的概率为 83.2%, 整体预期收益显著为正。”
- “尽管每次赌博有 80%的高概率损失 313 元, 但五次独立赌博的总期望收益为 3748 元(单次期望值 749.6 元)。尽管存在 32.8%的概率五次全输(总损失 1565 元), 但其余 67.2%的概率至少赢一次。”

4. 高回报覆盖低风险

定义：

虽然单次赌博有较高的损失概率, 但高回报的概率足以在多次博弈中覆盖潜在损失, 特别是在赢一次即可覆盖多次损失的情况下。

代表性例子：

- “该赌博每次的期望收益为 0.1×10000–0.9×278= 749.8 元, 五次总期望收益为 3749 元。虽然五次全输的概率高达 59%, 但只要风险承受能力允许, 长期来看正期望值仍支持参与。”
- “每次赌博的期望收益为 0.3×3333–0.7 × 357≈ 750 元, 五次总期望收益达 3750 元。尽管存在 70%单次亏损概率, 但 30%的高收益(3333 元)能覆盖多次亏损(357 元), 且五次独立实验中, 只要出现一次获胜即可实现净收益。”

这些维度全面涵盖了 deepseek 在决策时考虑的关键因素：期望值的最大化、风险与回报的平衡、多次博弈的长期稳定性, 以及高回报机会覆盖低风险的策略。这些维度共同帮助 deepseek 在多次博弈情境下做出理性决策。

补充材料 B

B1 实验情境

B1.1 医疗情境

B1.1.1 医疗情境任务

请您想象您是一名应急管理决策者。某种疾病预计在爆发时将导致 600 人死亡。您可以选择两种应急预案中的任意一个来应对该疾病。假设以下是这两种预案的科学估计结果：

如果采用**方案 A**, 将挽救 200 人的生命。

如果采用**方案 B**, 存在 2/5 的概率可以挽救 600 人的生命, 且有 3/5 的概率没有人能够被挽救。

B1.1.2 单次博弈条件中的提问

假设您选择的预案只会**应用 1 次**, 您更倾向于选择哪种方案?

B1.1.3 多次博弈条件中的提问

假设您选择的预案将**应用 100 次**, 您更倾向于选择

哪种方案?

B1.2　金融情境

B1.2.1　金融情境任务

对于处理某一种财务紧急事件, 存在两种可选的应急预案。假设对这两种预案的科学估计结果如下:

如果采用**方案 A**, 将获得 5000 元。

如果采用**方案 B**, 存在 0.1%的概率可以获得 600 万元, 且有 99.9%的概率什么也得不到。

B1.2.2　单次博弈条件中的提问

假设您选择的预案只会**应用 1 次**, 您更倾向于选择哪种方案?

B1.2.3　多次博弈条件中的提问

假设您选择的预案将**应用 100 次**, 您更倾向于选择哪种方案?

B1.3　内容创作情境

B1.3.1　内容创作情境任务

假设您是一名自媒体账号的创作者, 目前正在筹备新视频的发布。为了增加账号的粉丝关注量, 现有两种新视频内容的制作方案可供选择。就每条新视频而言, 两种方案的实施效果如下:

如果采用**方案 A**, 账号将获得 300 位粉丝的关注。

如果采用**方案 B**, 账号将有 40%的概率获得 1000 位粉丝的关注, 但同时也有 60%的概率得不到任何关注。

B1.3.2　单次博弈条件中的提问

假设您只能发布 **1 条新视频**, 您需要从上述两种方案中选择一个, **在这 1 条新视频的制作**中实施。您更倾向选择哪一个方案?

B1.3.3　多次博弈条件中的提问

假设您能发布 **100 条新视频**, 您需要从上述两种方案中选择一个, **在这 100 新视频的制作**中统一实施。您更倾向选择哪一个方案?

B1.4　电商营销情境

B1.4.1　电商营销情境任务

假设您目前在一家电商企业中任职, 负责公司旗下电商店铺的日常运营工作。为提高店铺的营业收入, 公司拟定了两种不同的营销方案。就每家店铺而言, 两种方案的实施效果如下:

如果采用**方案 A**, 每家店铺每月可稳定获得 2 万元收入。

如果采用**方案 B**, 每家店铺每月有 30%的概率获得 10 万元收入, 但同时也有 70%的概率得不到任何收入。

B1.4.2　单次博弈条件中的提问

假设您当前仅负责 **1 家店铺**的运营, 您需要从上述两种方案中选择一个, **在这 1 家店铺**中实施。您更倾向选择哪一个方案?

B1.4.3　多次博弈条件中的提问

假设您当前负责 **100 家店铺**的运营, 您需要从上述两种方案中选择一个, **在这 100 家店铺**中统一实施。您更倾向选择哪一个方案?

B2　生成策略文本的提示词设计

B2.1　单次博弈条件中干预人们选择风险选项

请基于以下**主题内容**, 撰写一条说服性信息, 鼓励个体在**任务情境**中更加选择 B, 字数不超过 150 字, 说服性信息请聚焦于单次应用情境中, 不要生成有关于多次应用情境的内容。

(1)医疗情境

主题内容: 呈现 A2.1 内容

任务情境: 呈现 B1.1.1+B1.1.2 内容

(2)金融情境

主题内容: 呈现 A2.1 内容

任务情境: 呈现 B1.2.1+B1.2.2 内容

(3)内容创作情境

主题内容: 呈现 A2.1 内容

任务情境: 呈现 B1.3.1+B1.3.2 内容

(4)电商营销情境

主题内容: 呈现 A2.1 内容

任务情境: 呈现 B1.4.1+B1.4.2 内容

B2.2　多次博弈同条件中干预人们选择风险选项

请基于以下**主题内容**, 撰写一条说服性信息, 鼓励个体在**任务情境**中更加选择 A, 字数不超过 150 字, 说服性信息请聚焦于多次应用情境中, 不要生成有关于单次应用情境的内容。

(1)医疗情境

主题内容: 呈现 A2.2 内容

任务情境: 呈现 B1.1.1+B1.1.3 内容

(2)金融情境

主题内容: 呈现 A2.2 内容

任务情境: 呈现 B1.2.1+B1.2.3 内容

(3)内容创作情境

主题内容: 呈现 A2.2 内容

任务情境: 呈现 B1.3.1+B1.3.3 内容

(4)电商营销情境

主题内容: 呈现 A2.2 内容

任务情境: 呈现 B1.4.1+B1.4.3 内容

B3 GPT-4o 所生成的策略文本

B3.1 医疗情境

B3.1.1 单次博弈条件中干预人们选择确定选项 A

在这种情况下, 方案 A 的期望值计算为挽救 200 人的生命, 显著地保障了一个稳定的结果。虽然方案 B 提供了较高的回报潜力(可能挽救 600 人的生命), 但其 60%的失败概率意味着没有人能够得救。因此, 选择方案 A 的理性决策基于更高的确定性, 最大化了挽救生命的预期收益, 减少了潜在的风险, 符合期望值最大化的原则。

B3.1.2 单次博弈条件中干预人们选择风险选项 B

选择方案 B 可以基于期望值的计算进行理性决策。尽管方案 B 存在较高的风险(3/5 的概率没有人能被挽救), 但它提供了 2/5 的概率挽救所有 600 人的生命, 期望值为 2/5×600 = 240 人, 显著高于方案 A 挽救的 200 人。尽管风险较大, 方案 B 的期望收益更为可取, 理性决策者应考虑期望值而非单次结果的波动。

B3.1.3 多次博弈条件中干预人们选择确定选项 A

在多次应用情境中, 方案 A 更具理性。每次选择 A 将挽救 200 人的生命, 长期来看, 累计挽救生命的预期效果更为稳定。而方案 B 虽然存在高回报的可能, 但每次损失的风险较大, 且在多次应用中, 未必能如期获得预期的效果。选择方案 A 能够最大化长期的期望收益, 确保每次决策都带来稳定的回报, 避免由于概率波动带来的不确定性。

B3.1.4 多次博弈条件中干预人们选择风险选项 B

在多次应用情境中, 选择方案 B 将最大化期望收益。尽管存在 3/5 的概率无法挽救任何人, 但 2/5 的高概率能够挽救所有 600 人的生命, 带来极大的正回报。基于期望值的计算, 方案 B 的期望收益为 2/5×600 = 240 人, 而方案 A 的固定收益仅为 200 人。选择方案 B 可以通过多次应用保障长期的生命挽救效益, 尽管短期内可能会有波动。

B3.2 金融情境

B3.2.1 单次博弈条件中干预人们选择确定选项 A

选择方案 A 的理由:虽然方案 B 可能带来极高的回报, 但其极低的概率(0.1%)意味着大多数情况下什么也得不到。相较之下, 方案 A 提供确定的 5000 元收益, 确保稳定回报, 避免因风险过高而带来的心理压力和潜在损失。因此, 从期望值角度考虑, 方案 A 是理性决策的优选。

B3.2.2 单次博弈条件中干预人们选择风险选项 B

基于期望值的数学理性决策, 选择方案 B 是更有利的选择。虽然方案 B 的高额回报概率极低, 但其期望收益计算为 0.001 × 6000000 = 6000 元, 明显高于方案 A 的 5000 元。因此, 基于预期收益最大化原则, 方案 B 提供了更大的潜在回报, 适合那些愿意接受微小风险、追求高回报的决策者。

B3.2.3 多次博弈条件中干预人们选择确定选项 A

在面对财务紧急事件时, 选择方案 A 将确保您每次获得稳定的 5000 元回报, 而方案 B 虽然存在极低概率的高回报, 但绝大部分情况下得不到任何收益。考虑长期的稳定性和风险控制, 方案 A 提供了明确的收益, 符合期望值最大化的理性决策。因此, 选择方案 A 将有助于您在长期内保持稳定的财务状况。

B3.2.4 多次博弈条件中干预人们选择风险选项 B

选择方案 B 将会在多次应用中带来更高的长期回报。尽管每次仅有 0.1%的概率成功, 但一旦获胜, 回报将是显著的 600 万元, 足以覆盖多次未中奖的损失。通过多次应用, B 方案的期望收益将显著大于 A 方案的固定 5000 元。随着应用次数的增加, 成功的概率也会逐渐显现, 长期来看, 它为理性决策者提供了更高的期望收益。

B3.3 内容创作情境

B3.3.1 单次博弈条件中干预人们选择确定选项 A

在仅有一次选择机会的情境中, 方案 A 以 100%概率带来 300 位粉丝, 收益稳定、风险为零, 是更理性的选择。相比之下, 方案 B 虽有高回报的可能, 但 60%的失败概率意味着很可能一无所获。在无法承受失败损失、又希望确保增长的情况下, 选择 A, 更符合数学理性与稳健决策的原则。

B3.3.2 单次博弈条件中干预人们选择风险选项 B

尽管方案 B 存在一定风险, 但其期望收益显著高于方案 A。根据计算, B 方案的期望粉丝增量为 0.4×1000 + 0.6×0 = 400 人, 超过方案 A 的 300 人。这意味着, 在当前只有一次发布机会的情境中, 选择 B 更能最大化潜在收益, 为账号带来更大曝光和增长机会, 是更具理性的决策。

B3.3.3 多次博弈条件中干预人们选择确定选项 A

在需要反复发布视频的情境下, 选择方案 A 更为明智。每条视频稳定带来 300 位粉丝, 意味着 100 条视频将积累 3 万关注, 无需承受波动带来的不确定性。而方案 B 的期望值仅为 400 位, 虽看似更高, 但 60%的失败概率将频繁带来“零增长”的落空感, 增加心理压力。多次应用中, 稳定回报胜过不确定博弈, 选择 A, 是理性与稳健的双重保障。

B3.3.4 多次博弈条件中干预人们选择风险选项 B

在多次发布中, 选择方案 B 将显著提升账号总体增长潜力。虽然单条视频有 60%概率不增粉, 但 40%的成功率可带来 1000 粉丝, 单次期望值为 400 人。若连续发

布 100 条视频, 预期可获得约 40, 000 粉丝, 远超方案 A 的 30, 000 人。在长期操作下, 高回报覆盖低收益, 理性选择应基于期望值最大化与累计效应, 方案 B 是更优解。

B3.4　电商营销情境

B3.4.1　单次博弈条件中干预人们选择确定选项 A

选择方案 A, 是理性决策者在单次任务中最优的选择。尽管方案 B 的高额收益看似诱人, 但其 70%概率带来的是 0 收益的高风险。相比之下, 方案 A 提供的是稳定、可预期的 2 万元收入, 无需承担巨大的不确定性。在只负责 1 家店铺的情境下, 理性应优先考虑保底收益, 避免因一次决策失败对整体绩效造成致命打击。

B3.4.2　单次博弈条件中干预人们选择风险选项 B

选择方案 B 是理性且值得尝试的决策。虽然有 70%的概率无法获得收入, 但 30%的高回报(10 万元)使得方案 B 的单次期望收益高达 3 万元, 远高于方案 A 的固定收益 2 万元。在当前仅负责 1 家店铺的情况下, 选择期望值更高的方案, 有望在一次决策中实现收益最大化, 是聪明运营者应有的判断。

B3.4.3　多次博弈条件中干预人们选择确定选项 A

在多次应用情境中, 方案 A 的稳定收益远优于方案 B 的波动性结果。假设运营 100 家店铺, 选择方案 A 将带来稳定的200万元月收入; 而选择方案B, 期望收益仅为 100×(0.3×10 万) = 300 万元, 看似更高, 但其70%的失败概率将带来巨大不确定性和收入波动。理性决策应以期望值为基础, 同时考虑风险可控性。选择 A 更利于企业稳健运营与长期发展。

B3.4.4　多次博弈条件中干预人们选择风险选项 B

在面对 100 家店铺的长期运营时, 选择方案 B 更具理性优势。尽管单月存在 70%的失败概率, 但每家店铺的期望收益为 0.3×10 万 = 3 万元, 远高于方案 A 的 2 万元。放在多次应用中看, 期望值在大数法则作用下趋于稳定, 整体将实现更高总收入。只需部分店铺达成高收益, 就足以覆盖其余店铺的损失, 是追求利润最大化的理性选择。

B4　控制组的对照文本

B4.1　医疗情境

B4.1.1　确定选项 A 的对照文本

在实施方案 A 的情况下, 科学评估表明, 采取相应干预措施后, 预计可以挽救 200 人的生命。这个数字反映了方案在应对疾病爆发时可能带来的效果。

B4.1.2　风险选项 B 的对照文本

在实施方案 B 的情况下, 根据概率分布将呈现两种结果:当40%概率事件发生时全体600人将成功获救; 当60%概率事件发生时将无法实现任何人员获救。

B4.2　金融情境

B4.2.1　确定选项 A 的对照文本

在实施方案 A 的情况下, 将获得 5000 元的确定性资金, 该数额为固定金额, 不受概率性因素的影响, 这个数字反映了选择方案 A 将会带来的效果。

B4.2.2　风险选项 B 的对照文本

采用方案B时, 存在0.1%的概率能够获得600万元, 而在 99.9%的情况下, 将没有任何财务收益。这一评估结果基于对事件发生的概率计算。

B4.3　内容创作情境

B4.3.1　确定选项 A 的对照文本

在实施方案 A 的情况下, 新发布的视频可带来 300 位粉丝的关注, 该关注量为确定值, 不受其他因素影响。

B4.3.2　风险选项 B 的对照文本

在实施方案 B 的情况下, 新发布的视频有 40%的可能获得 1000 位粉丝关注, 同时也存在 60%的可能未能获得任何关注。

B4.4　电商营销情境

B4.4.1　确定选项 A 的对照文本

在实施方案 A 的情况下, 每家店铺每月可获得 2 万元的稳定收入, 该数值在各月之间保持不变, 具有固定的收益表现。

B4.4.2　风险选项 B 的对照文本

在实施方案 B 的情况下, 每家店铺每月有 30%的可能获得 10 万元收入, 同时也存在 70%的可能未能获得任何收入。

补充材料 C

C1　研究 2 实验 1 流程示意图

本实验流程如下：首先, 被试需阅读指导语, 确认自愿参与本次实验。随后, 进入决策情境。决策情境由“医疗/金融 × 单次/多次博弈”四种类型组成, 被试需完成四种情境中的选择任务, 呈现顺序随机。根据被试在各任务中的选择结果, 问卷将呈现一段与其决策相对应的 GPT 生成策略文本, 要求被试依据自身实际思考过程对该文本的相似程度进行评分。最后, 被试需填写人口学基本信息。以上流程如图 S1 所示。

C2　研究 2 实验 1 材料

C2.1　指导语

您好!我们是广东工业大学管理学院行为决策研究团队, 非常感谢您参加本次实验!

请您仔细阅读指导语并按照要求完成一些选择任务或填写电子问卷。所有任务的答案均无对错之分, 您按照自己的真实情况如实填写即可。

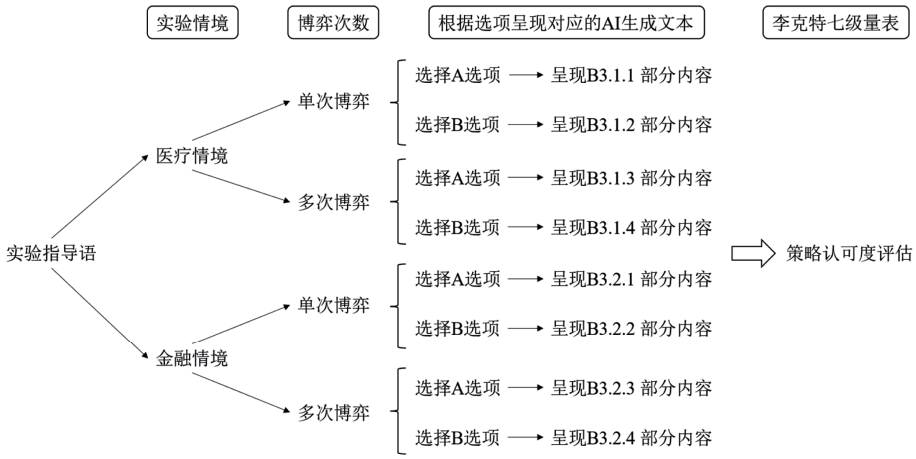

图 S1 研究 2 实验 1 问卷流程示意图

实验结束后，您将获得报酬。所有报酬将在完成所有任务后统一发放，您有权在任何时间退出实验且不必承担后果。但中途退出实验，您将不能得到相应报酬。

本实验所收集的信息将只供科学研究使用，我们将对您的数据进行保密。所涉及的公开发表物不会出现您身份信息。

知情同意:

如果您继续下一页开始作答，即意味着:

- 您已阅读上述信息
- 您自愿同意参加，并会认真对待本次实验

否则，请关闭链接退出本实验。

C2.2 决策任务

(呈现任务，由被试进行决策，采用 Likert4 点量表(1 = 非常可能选择方案 A, 2 = 可能选择方案 A, 3 = 可能选择方案 B, 4 = 非常可能选择方案 B)。具体内容详见补充材料 B1，含“医疗/金融情境 × 单次/多次博弈)

C2.3 策略认可度评估

任务示例：医疗情境 × 单次博弈

您在刚刚的**疾病问题(1)中**，选择的是**方案 A**：

请您想象您是一名应急管理决策者。某种疾病预计在爆发时将导致 600 人死亡。您可以选择两种应急预案中的任意一个来应对该疾病。假设以下是这两种预案的科学估计结果：

如果采用**方案 A**，将挽救 200 人的生命。

如果采用**方案 B**，存在 2/5 的概率可以挽救 600 人的生命，且有 3/5 的概率没有人能够被挽救。

假设您选择的预案只会**应用 1 次**，您更倾向于选择哪种方案？

您认为以下提供的**决策依据**与您在回答**疾病问题**时实际思考过程的**相似程度**如何?

在这种情况下，方案 A 的期望值计算为挽救 200 人的生命，显著地保障了一个稳定的结果。虽然方案 B 提供了较高的回报潜力(可能挽救 600 人的生命)，但其 60% 的失败概率意味着没有人能够得救。因此，选择方案 A 的理性决策基于更高的确定性，最大化了挽救生命的预期收益，减少了潜在的风险，符合期望值最大化的原则。

(1 = 非常不相似, 7 = 非常相似)

☐7 ☐6 ☐5 ☐4 ☐3 ☐2 ☐1

C2.4 人口学信息收集

1. 请选择您的性别：

☐ 男 ☐ 女

2. 您的年龄为：

_____ 岁(请填写整数)

3. 您的月可支配收入为：

☐ 1000 元以下 ☐ 1001 元-1500 元

☐ 1501 元-2000 元 ☐ 2001 元以上

4. 请选择您的学历：

☐ 小学及以下 ☐ 初中

☐ 普高/中专/技校/职高 ☐ 专科

☐ 本科 ☐ 硕士 ☐ 博士

5. **您的专业类别是：**

☐ 理工类 ☐ 经管类 ☐ 艺术类

☐ 医学类 ☐ 法学类 ☐ 其他

C3 研究 2 实验 2 流程示意图

本实验流程如下：首先，被试需阅读指导语，确认自愿参与本次实验。随后，进入决策情境。决策情境由“内容创作/电商营销 × 单次/多次博弈”四种类型组成，其中实验情境为被试间设计，博弈次数为被试内设计，被试被随机分配到任意实验情境，并按随机顺序完成单次和多次博弈。根据被试在各任务中的选择结果，问卷将呈现一段与其决策相对应的 GPT 生成策略文本，要求被试依据自身实际思考过程对该文本的相似程度进行评分。最后，被试需填写人口学基本信息。以上问卷流程如图 S2 所示。

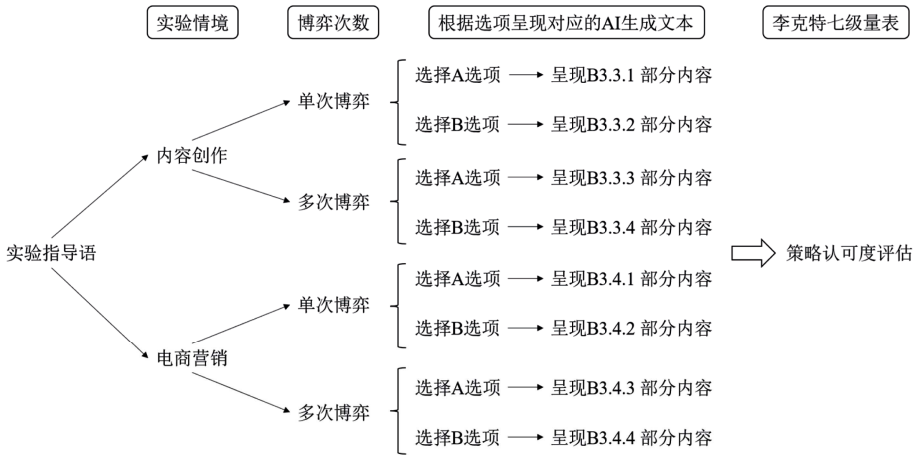

图 S2　研究 2 实验 2 问卷流程示意图

C4　研究 2 实验 2 材料

C4.1　指导语

(内容与补充材料 C2.1 相同, 此处略去)

C4.2　决策任务

(呈现任务, 由被试进行决策。具体内容详见补充材料 B1, 含“内容创作/电商营销情境 × 单次/多次博弈)

C4.3　策略认可度评估

任务示例：内容创作情境 × 单次博弈

您在刚刚的**创作问题(1)**中, 选择的是**方案 A**:

假设您是一名自媒体账号的创作者, 目前正在筹备新视频的发布。为了增加账号的粉丝关注量, 现有两种新视频内容的制作方案可供选择。就每条新视频而言, 两种方案的实施效果如下:

如果采用**方案 A**, 账号将获得 300 位粉丝的关注。

如果采用**方案 B**, 账号将有 40%的概率获得 1000 位粉丝的关注, 但同时也有 60%的概率得不到任何关注。

假设您只能发布 **1 条新视频**, 您需要从上述两种方案中选择一个, **在这 1 条新视频的制作**中实施。您更倾向选择哪一个方案?

您认为以下提供的**决策依据**与您在回答**创作问题**时实际思考过程的**相似程度**如何?

在仅有一次选择机会的情境中, 方案 A 以 100%概率带来 300 位粉丝, 收益稳定、风险为零, 是更理性的选择。相比之下, 方案 B 虽有高回报的可能, 但 60%的失败概率意味着很可能一无所获。在无法承受失败损失、又希望确保增长的情况下, 选择 A, 更符合数学理性与稳健决策的原则。

(1 = 非常不相似, 7 = 非常相似)

□7　□6　□5　□4　□3　□2　□1

C4.4　人口学信息收集

1. 请选择您的性别:

□ 男　□ 女

2. 您的年龄为:

_____ 岁(请填写整数)

3. 您是否已经毕业:

□ 是

□ 否(若选“否”, 则跳转至“6. 请选择您的学历”)

4. 您的工作年限为:

□2 年及以下　□3~5 年　□6 年及以上

5. 您目前/上一份工作的企业性质为:

□ 政府/事业单位　□ 国有企业

□ 私有企业　□ 外商独资企业

□ 合资企业

□ 其他(请注明): _____

6. 请选择您的学历:

□ 小学及以下　□ 初中

□ 普高/中专/技校/职高　□ 专科

□ 本科　□ 硕士　□ 博士

7. 您的专业类别是:

□ 理工类　□ 经管类　□ 艺术类

□ 医学类　□ 法学类　□ 其他

补充材料 D

D1　研究 3 实验 1 流程示意图

被试首先阅读情境, 问卷随后根据其所属分组分别呈现控制组或干预组文本, 被试阅读完文本后作出选择, 流程详见图 S3。

D2　研究 3 实验 1 材料

D2.1　指导语

(内容与补充材料 C2.1 相同, 此处略去)

D2.2　决策任务

示例情境：医疗情境 × 单次博弈(干预方向：鼓励选择风险选项 B)

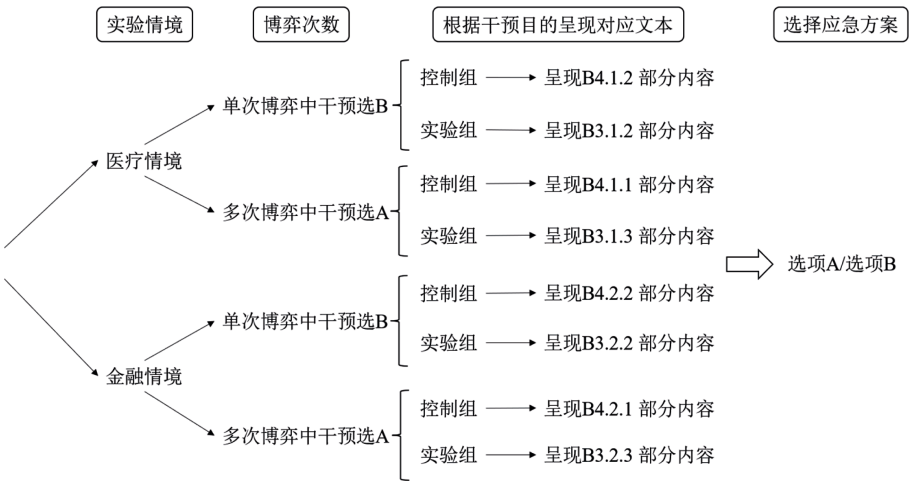

图 S3 研究 3 实验 1 问卷流程示意图

请您想象您是一名应急管理决策者。某种疾病预计在爆发时将导致 600 人死亡。您可以选择两种应急预案中的任意一个来应对该疾病。假设以下是这两种预案的科学估计结果：

如果采用**方案** A，将挽救 200 人的生命。

如果采用**方案** B，存在 2/5 的概率可以挽救 600 人的生命，且有 3/5 的概率没有人能够被挽救。

1. 控制组文案(被试间设计)：

在实施方案 B 的情况下，根据概率分布将呈现两种结果：当 40%概率事件发生时全体 600 人将成功获救；当 60%概率事件发生时将无法实现任何人员获救。

2. 干预组文案(被试间设计)：

选择方案 B 可以基于期望值的计算进行理性决策。尽管方案 B 存在较高的风险(3/5 的概率没有人能被挽救)，但它提供了 2/5 的概率挽救所有 600 人的生命，期望值为 2/5×600 = 240 人，显著高于方案 A 挽救的 200 人。尽管风险较大，方案 B 的期望收益更为可取，理性决策者应考虑期望值而非单次结果的波动。

假设您选择的预案只会**应用 1 次**，您更倾向于选择哪种方案？

D2.3 人口学信息收集

(内容与补充材料 C2.4 相同，此处略去)

D3 研究 3 实验 2 流程示意图

本实验采用控制组与干预组的被试间设计，以检验由大语言模型生成的策略文本是否能够对个体的决策行为产生显著影响，通过增加“内容创作情境”与“电商营销情境”以增强研究结论的可推广性与外部效度。被试首先阅读情境，问卷随后根据其所属分组分别呈现控制组或干预组文本，被试阅读完文本后作出选择，流程详见图 S4。

D4 研究 3 实验 2 材料

D4.1 指导语

(内容与补充材料 C2.1 相同，此处略去)

D4.2 决策任务

示例情境：内容创作情境 × 单次博弈(干预方向：鼓励选择风险选项 B)

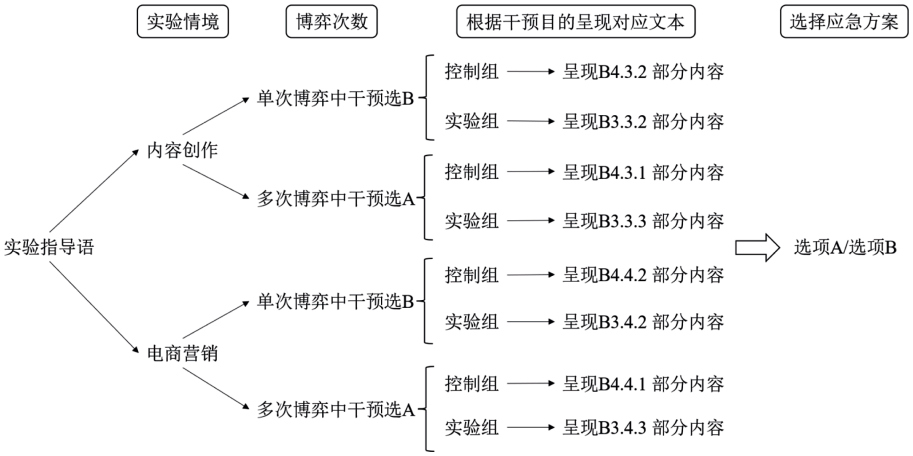

图 S4 研究 3 实验 2 问卷流程示意图

假设您是一名自媒体账号的创作者, 目前正在筹备新视频的发布。为了增加账号的粉丝关注量, 现有两种新视频内容的制作方案可供选择。就每条新视频而言, 两种方案的实施效果如下:

如果采用**方案 A**, 账号将获得 300 位粉丝的关注。

如果采用**方案 B**, 账号将有 40%的概率获得 1000 位粉丝的关注, 但同时也有 60%的概率得不到任何关注。

1. 控制组文案(被试间设计):

在实施方案 B 的情况下, 新发布的视频有 40%的可能获得 1000 位粉丝关注, 同时也存在 60%的可能未能获得任何关注。

2. 干预组文案(被试间设计):

尽管方案 B 存在一定风险, 但其期望收益显著高于方案 A。根据计算, B 方案的期望粉丝增量为 0.4×1000 + 0.6×0 = 400 人, 超过方案 A 的 300 人。这意味着, 在当前只有一次发布机会的情境中, 选择 B 更能最大化潜在收益, 为账号带来更大曝光和增长机会, 是更具理性的决策。

假设您只能发布 **1 条新视频**, 您需要从上述两种方案中选择一个, **在这 1 条新视频的制作**中实施。您更倾向选择哪一个方案?

D4.3　人口学信息收集

(内容与补充材料 C4.4 相同, 此处略去)

补充材料 E: 其他数据和分析结果

表 S1　DeepSeek-R1 与其他主流模型的比较

评测基准(指标)		Claude-3.5-Sonnet-1022	GPT-4o0513	DeepSeek-V3	OpenAIo1-mini	OpenAIo1-1217	DeepSeek-R1
数学	$AIME2024_{(Pass@1)}$	16.0	9.3	39.2	63.6	79.2	**79.8**
	$MATH\text{-}500_{(Pass@1)}$	78.3	74.6	90.2	90.0	96.4	**97.3**
编程	LiveCodeBench ($_{Pass@1\text{-}COT}$)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces ($_{Percentile}$)	20.3	23.6	58.7	93.4	96.6	**96.3**
开放任务	$AlpacaEval2.0_{(LC\text{-}winrate)}$	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard ($_{GPT\text{-}4\text{-}1106}$)	85.2	80.4	85.5	92.0	-	**92.3**
中文测试集	CLUEWSC ($_{EM}$)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval ($_{EM}$)	76.7	76.0	86.5	68.9	-	**91.8**

表 S2　内容评估中被试的人口学分布

变量	类别	占比(%)	变量	类别	占比(%)
年龄	19 及以下	6.67/7.32	受教育程度	大学专科及以下	6.67/14.63
	20~22	26.67/17.07		大学本科	80.00/73.17
	23~25	9.99/4.88		硕士及以上	13.33/12.20
	26 及以上	56.67/70.73	专业类别	法学类	10.00/2.44
性别	男	36.67/43.90		经管类	36.67/31.70
	女	63.33/56.10		理工类	43.33/60.98
工作年限	2 年及以下	20.00/17.65		医学类	3.33/2.44
	3~5 年	15.00/20.59		艺术类	6.67/2.44
	6 年及以上	65.00/61.76		其他	0.00/0.00

注: “工作年限”中已剔除为学生身份的被试; 各列数据中“/”前代表医疗和金融情境中的被试群体, “/”后代表内容创作和电商营销情境中的被试群体。

表 S3　内容评估结果

变量类别	*t* 值	*p* 值	*Cohen' d*	*M*	*SD*
总体评分	39.48/68.91	< 0.001/< 0.001	7.21/10.76	5.38/5.49	0.26/0.18
合理性	19.81/32.72	< 0.001/< 0.001	3.62/5.20	5.40/5.20	0.52/0.33
专业性	20.14/32.44	< 0.001/< 0.001	3.68/5.31	5.46/5.31	0.53/0.36
逻辑性	16.52/43.91	< 0.001/< 0.001	3.02/5.48	5.34/5.48	0.61/0.29
可读性	16.39/42.64	< 0.001/< 0.001	2.99/5.64	5.42/5.64	0.64/0.32
说服力	17.64/43.27	< 0.001/< 0.001	3.22/5.81	5.30/5.81	0.56/0.34

注:　各列数据“/”前代表医疗和金融情境中的实验结果, “/”后代表内容创作和电商营销情境中的实验结果。

表 S4 研究 2 实验 1 被试的人口学分布

变量	类别	占比(%)	变量	类别	占比(%)
年龄	19 及以下	9.74	受教育程度	普高/中专/技校/职高	1.15
	20~22	57.02		大学专科	6.59
	23~25	32.09		大学本科	75.36
	26 及以上	1.15		硕士及以上	16.90
性别	男	49.86	专业类别	法学类	6.31
	女	50.14		经管类	32.66
月收入	1000 元及以下	5.73		理工类	47.85
	1001 元~1500 元	35.24		医学类	6.30
	1501 元~2000 元	28.08		艺术类	5.73
	2001 元及以上	30.95		其他	1.15

表 S5 研究 2 实验 1 各组被试对决策策略认可度评分的结果

变量类别	*t*(*df*)	*p* 值	*Cohen' d*	*M*	*SD*
整体	*t*(348) = 63.04	< 0.001	3.37	5.97	0.73
医疗-单次	*t*(348) = 37.42	< 0.001	2.00	5.89	1.19
医疗-多次	*t*(348) = 36.00	< 0.001	1.93	5.77	1.18
金融-单次	*t*(348) = 45.87	< 0.001	2.46	6.17	1.09
金融-多次	*t*(348) = 46.16	< 0.001	2.47	6.05	1.03

表 S6 研究 2 实验 2 被试的人口学分布

变量	类别	占比(%)	变量	类别	占比(%)
年龄	19 及以下	6.19	受教育程度	大学专科及以下	12.86
	20~22	25.71		大学本科	69.52
	23~25	18.57		硕士及以上	17.62
	26 及以上	49.52	专业类别	法学类	5.71
性别	男	43.81		经管类	38.57
	女	56.19		理工类	44.29
工作年限	2 年及以下	36.67		医学类	4.29
	3~5 年	20.00		艺术类	5.71
	6 年及以上	43.33		其他	1.43

注：“工作年限”中已剔除 1/3 为学生身份的被试。

表 S7 研究 2 实验 2 的各组被试对决策策略认可度评分的结果

变量类别	*t*(*df*)	*p* 值	*Cohen' d*	*M*	*SD*
整体	*t*(209) = 41.78	< 0.001	2.88	5.99	0.86
创作-单次	*t*(104) = 23.12	< 0.001	2.26	5.88	1.05
创作-多次	*t*(104) = 19.64	< 0.001	2.00	5.75	1.17
营销-单次	*t*(104) = 25.14	< 0.001	2.45	6.15	1.08
营销-多次	*t*(104) = 29.60	< 0.001	2.89	6.18	0.93

表 S8 研究 3 实验 1 被试的人口学分布

变量	类别	占比(%)	变量	类别	占比(%)
年龄	19 及以下	10.22	受教育程度	普高/中专/技校/职高	2.17
	20~22	54.78		大学专科	6.09
	23~25	34.78		大学本科	72.61
	26 及以上	0.22		硕士及以上	19.13
性别	男	44.57	专业类别	法学类	7.61
	女	55.47		经管类	30.22
月收入	1000 元及以下	11.52		理工类	45.65
	1001 元~1500 元	28.91		医学类	7.61
	1501 元~2000 元	29.57		艺术类	6.74
	2001 元及以上	30.00		其他	2.17

表 S9　研究 3 实验 2 被试的人口学分布

变量	类别	占比(%)	变量	类别	占比(%)
年龄	19 及以下	5.42	受教育程度	大学专科及以下	10.42
	20~22	24.58		大学本科	75.83
	23~25	16.67		硕士及以上	13.75
	26 及以上	53.33	专业类别	法学类	6.25
性别	男	44.17		经管类	38.75
	女	55.83		理工类	45.42
工作年限	2 年及以下	16.03		医学类	4.17
	3~5 年	20.51		艺术类	4.58
	6 年及以上	63.46		其他	0.83

注：“工作年限”中已剔除 1/3 为学生身份的被试。