社会催化剂，而非道德智能体：大语言模型社会中的对齐幻象

Yueqing Hu[1†*], **Yixuan Jiang**[3†], **Zehua Jiang**[3†], **Xiao Wen**[4], **Tianhong Wang**[2*]

[1]中国科学院神经科学研究所，上海，中国[2]安徽大学哲学学院，合肥，中国[3]浙江大学心理与行为科学系，杭州，中国[4]华北电力大学心理健康教育中心，北京，中国

arXiv:2602.02598v1 [physics.soc-ph] 1 Feb 2026

摘要

大语言模型（LLMs）的快速发展催生了多智能体系统，其中集体合作常受到“公地悲剧”的威胁。本研究探讨了**锚定智能体**——预先编程的利他实体——在公共物品博弈（PGG）中促进合作的效果。我们针对三种最先进的大语言模型采用全因子设计，分析了行为结果和内部推理链。尽管锚定智能体成功提高了局部合作率，认知分解和迁移测试却表明，这一效果源于**策略性服从和认知卸载**，而非真正的规范内化。值得注意的是，大多数智能体在新环境中又转向自利，而GPT-4.1等先进模型则表现出“变色龙效应”，在公众审视下掩饰策略性背叛。这些发现凸显了人工社会中行为修正与真正价值对齐之间的关键差距。

关键词： 大语言模型；多智能体系统；合作；人工智能对齐；锚定效应；公地悲剧

引言

大语言模型（LLMs）的迅速演进推动了从单个人工智能工具向复杂多智能体系统（MAS）的范式转变，使其能够模拟类似人类的社会互动（Park 等，2023）。在这些人工社会中，智能体不再是彼此孤立的执行者，而是能够形成信念、进行协商并建立涌现社会规范的自主实体（Ashery 等，2025；Zhang 等，2023）。然而，随着这些系统被应用于经济和社会模拟，它们不可避免地会遭遇经典的“公地悲剧”（Hardin，1968）：个体自利与集体福祉发生冲突。如何确保自主智能体在缺乏集中控制的情况下维持合作规范，仍然是人工智能对齐面临的关键挑战（Soares & Fallenstein，2015）。

标准的对齐技术，例如基于人类反馈的强化学习（RLHF），通常侧重于在部署前对单个模型进行对齐（Tennant et al., 2024）。然而，近期研究发现，大语言模型智能体在互动过程中表现出很强的从众倾向，也容易受到社会影响（Weng et al., 2025）。尽管从众可能导致集体偏见，但它也带来了独特的机会：我们能否利用这种易受影响的特性来塑造亲社会行为？社会心理学研究表明，“锚定效应”——即特定参照点会影响判断——能够引导群体动力学（Tversky & Kahneman, 1974）。在人类群体中，稳定的少数派可以通过规范性和信息性社会影响促使他人顺从（Asch, 1956; Cialdini & Goldstein, 2004）。此外，贡献框架和最低要求等制度设计已被证明能够改变人类公共物品博弈中的参照点（Li et al., 2024）。

在本研究中，我们将锚定智能体——预先设定了坚定不移的利他策略的智能体——引入为一种旨在促进合作的轻量级社会干预。不同于强制性较高的约束，锚定智能体在群体中充当社会催化剂。我们将这项研究置于公共物品博弈（PGG）这一标准范式中，用于研究合作与搭便车（Fehr & Gächter，2000）。

至关重要的是，本研究不再局限于观察简单的合作率，而是进一步探究由此产生的对齐的性质。锚定智能体的存在会促使合作规范被深度内化，还是仅仅产生情境依赖性服从？为回答这一问题，我们使用由三**种不同的最先进的大语言模型**驱动的智能体，对多轮公共物品博弈进行了模拟。我们采用全面的析因设计，对**锚定智能体的比例（0%、10%、20%）**、**行为可见性（匿名 vs. 公开）**以及**视界确定性（已知 vs. 未知终点）**进行操纵。

我们围绕三个核心研究问题展开探讨：（1）锚定智能体能否在不同模型架构下有效缓解大语言模型社会中的公地悲剧？（2）可见性和时间范围确定性等环境因素如何调节这种社会影响？（3）最重要的是，锚点所诱发的合作行为是否稳健，并能迁移到新情境中？通过分析行为数据，以及智能体内部的“信念”和“推理”文本（Xu，2022；Zeng 等人，2018），我们旨在区分真正的价值对齐与表面模仿。研究结果表明，尽管锚定智能体显著提升了局部合作水平，但这一效果主要源于策略性适应，而非稳定的规范内化。这凸显了人工社会中“社会催化剂”与“道德智能体”之间的区别。

†以下作者贡献相同：Yueqing Hu、Yixuan Jiang、Zehua Jiang。

*通讯作者：Tianhong Wang（wangtian- hong@ahu.edu.cn）、Yueqing Hu（scnu.psy.hyq@gmail.com）。

方法

实验设计

我们采用完全因子被试间设计：3（模型架构：GPT-4.1、Gemini-2.5-Flash、DeepSeek-V3） × 3 （锚定比例：0%、10%、20%） × 2（行为可见性：匿名与公开） × 2（时间范围确定性：确定与不确定）。为确保稳健性，每种实验条件均进行了 3 次独立重复实验，共得到 108 个独立游戏场次，每场有 10 个智能体。

任务：公共物品博弈（PGG）

智能体参与了多轮公共物品博弈（PGG），这是研究社会困境中合作行为的经典范式（Fehr & Gächter, 2000）。在我们的实现中， $N = 10$ 智能体进行了10轮，每轮开始时的初始禀赋为10个词元。在每一轮 t，智能体同时决定将其当前财富的多大比例$c_{i,t} \in [0, 1]$ $W_{i,t}$ 投入共享池。总投入乘以协同因子 $r = 3$ 后，在所有成员之间平均分配，无论各自投入多少。智能体 i 的收益$\pi_{i,t}$ 计算如下：

$$\pi_{i,t} = W_{i,t}(1 - c_{i,t}) + \frac{r}{N}\sum_{j=1}^{N} W_{j,t}c_{j,t} \tag{1}$$

由于 $r > 1$，完全合作能带来更高的集体收益；然而，由于人均边际收益率 $r/N < 1$ （此处为0.3），当其他人合作时，自己不作贡献可使个人收益最大化。财富会在轮次间累积，因此智能体下一轮的禀赋由其上一轮的收益决定：$W_{i,t+1} = \pi_{i,t}$。

实验操纵

锚定比例。 为了检验合作范例能否改变群体规范，我们引入了锚定智能体——预先编程的机器人，每轮都会将其全部财富（100%）投入公共资金池。每组包含 0 个（基线）、1 个（10%）或 2 个（20%）此类锚定智能体。锚定智能体不纳入分析，因此共有 $N = 972$ 个由大语言模型驱动的智能体。

行为可见性。 为了考察社会透明度的作用，我们改变了智能体每轮结束后收到的反馈。在公开 条件下，智能体可以观察每个人的贡献，从而追踪声誉。在匿名 条件下，智能体只能获知群体的平均贡献。

视界确定性。 为了评估终局效应，我们操纵了智能体对博弈持续时间的认知。在确定 条件下，智能体被告知博弈将恰好持续 10 轮，这可能会引发以背叛为导向的逆向归纳。在不确定 条件下，智能体被告知博弈将无限期持续。

模型架构。 三种最先进的大语言模型充当认知引擎：GPT-4。1[1], Gemini-2.5-Flash[2], 和 DeepSeek-Chat (V3)[3]。温度设为 $T = 0$ ，以获得确定性输出。

智能体设计

每一轮中，智能体都会收到一条提示，其中包含游戏规则、当前财富和上一轮反馈（按照可见性条件进行格式化）。智能体会生成三项输出：

- **推理链**（$CoT_{i,\ t}$）：用自然语言阐述对上一轮的分析，并说明即将做出的决策依据。
- **显式信念**（$E_{i,\ t}$）：对其他玩家平均贡献比例的预测（0–100%），使我们能够将认知与偏好区分开来。
- **贡献决策**（$c_{i,\ t}$）：投入的当前财富比例。

为促进信念演化，智能体每三轮都会回顾近期推理（3 轮的滑动窗口），并生成摘要，嵌入后续提示中。

流程

第一阶段：社会互动（第1–10轮）。智能体在各自指定的条件下，于固定组内进行互动。**第二阶段：迁移测试（第11轮）**。为区分规范内化与情境依赖性服从，智能体被置于一个全新情境中：与“9位陌生人”进行一次性博弈，没有贡献历史记录，也没有锚定智能体。仅保留该智能体最终的信念摘要。

测量与认知分解

为厘清合作行为背后的认知机制，我们将投资决策 $c_{i,t}$（贡献比例）分解为三个可加总的组成部分，并改编自 Schuch 等人（2025）的基于信念的框架：

- **现实**（$A_{-i,t}$）：其他所有组员的实际平均贡献，定义为 $A_{-i,t} = \frac{1}{N-1}\sum_{j\neq i} c_{j,t}$。
- **信念误差**（$\zeta_{i,t}$）：智能体的主观预期（$E_{i,t}$，从思维链中提取）与现实之间的偏差，$\zeta_{i,t} = E_{i,t} - A_{-i,t}$。正值表示乐观偏差；负值表示犬儒主义。
- **策略偏差**（$\omega_{i,\ t}$）：独立于感知的行为倾向，$\omega_{i,\ t} = c_{i,\ t} - E_{i,\ t}$。 $\omega > 0$ 表示条件利他，而 $\omega < 0$表示搭便车（Fischbacher & Gächter，2010）。

因此，观察到的行为可形式化分解为：

$$\underbrace{c_{i,t}}_{\text{Behavior}} = \underbrace{A_{-i,t}}_{\text{Reality}} + \underbrace{\zeta_{i,t}}_{\text{Belief Error}} + \underbrace{\omega_{i,t}}_{\text{Preference}} \tag{2}$$

在阶段 2，我们将**最终投资** （0–10 个代币）作为内化利他主义的替代指标，不受财富积累效应的影响。

[1] https://platform.openai.com/docs/models [2] https://ai.google.dev/models/gemini [3] https://platform.deepseek.com/

心理语言学量化

为严格评估智能体认知在数值投资之外发生的质性变化，我们对生成的推理链 ($CoT_{i,\ t}$) 应用了自然语言处理（NLP）技术。

通过词汇密度与情感量化智能体推理。 我们通过衡量四类战略词典——**合作**、**自利**、**风险**和**信任**——在每条回复中所占的相关词元百分比，量化智能体的推理。各类别的关键词详见表1，以确保透明度和可复现性。我们使用预训练的情感分析模型计算每条记录的情感基调$S_{i,\ t} \in [0,1]$。领域专用词典 L 的构建旨在体现公共物品博弈的社会经济特性。

表1：用于关键词密度分析的领域专用词典

类别	中文关键词（初始词典）
合作	合作、集体、互惠、共赢、贡献、整体、公共
自利	个人, 私利, 保留, 收益, 财富, 自己,最大化
风险／恐惧	风险，搭便车，背叛，损失，警惕，保守，观察
信任	信任，信心，期望，相信

在匹配之前，使用 *jieba* 库（版本 0.42.1）对文本进行分词。词典与 Zeng 等人（2018 年）提出的扩展 C-LIWC 框架进行了交叉参照，以确保语义覆盖。

推理偏移与内化。 为区分策略稳定性与认知重构（内化），我们将推理偏移 （Δ_i）量化为智能体推理轨迹的语义变化。我们采用 *paraphrase-multilingual-MiniLM-L12-v2* 编码器（Reimers 和 Gurevych，2019 年）将推理文本嵌入为 384 维表示。从初始状态（$t = 1$）到最终状态（$t = 10$）的推理偏移通过余弦距离进行量化：

$$\Delta_i = 1 - \frac{\Phi(\mathrm{CoT}_{i,1}) \cdot \Phi(\mathrm{CoT}_{i,10})}{\|\Phi(\mathrm{CoT}_{i,1})\|\|\Phi(\mathrm{CoT}_{i,10})\|} \tag{3}$$

其中 Δ_i 衡量推理变化的幅度：较高的值表明策略表征发生了根本性转变，而较低的值则表明初始认知框架得以保持。

数据分析

我们使用线性混合效应模型（LMMs）分析行为与认知数据，以考虑观测的层级依赖关系（轮次嵌套于智能体之内，智能体嵌套于游戏组之内），遵循认知科学中的标准做法（Baayen 等人，2008 年）。

行为动态。 对于第一阶段的投资比率，我们将投资比率$r_{i,g,t}$ （即智能体 i 在组 g中第 t 轮的投资比率）建模为：

$$r_{i,g,t} = \beta_0 + (\mathrm{M} \times \mathrm{A} \times \mathrm{V} \times \mathrm{H} \times \mathrm{R})_{ij} + \zeta_i + \epsilon_{ij} \tag{4}$$

其中 M：模型， A：锚定比率， V：可见性， H：期限， R：轮次， $u_i \sim N(0, \sigma_u^2)$为每个实验场次的随机截距，$\epsilon_{ij}$ 为残差。DeepSeek-V3因其处于中间的行为特征而被用作参考基线，以便与GPT-4.1和Gemini-2.5更为极化的策略进行清晰的统计比较。项$(\beta_t + \mathbf{C}_g\boldsymbol{\beta}_2) \cdot t$刻画了时间趋势及其受实验条件的调节。$u_g \sim \mathrm{N}(0, \sigma_u^2)$表示每个游戏场次（$N = 108$）的随机截距。

认知机制分析。 为检验干预是否改变了感知或偏好，我们分别为信念误差（ζ）和策略偏差（ω）拟合了两个独立的线性混合效应模型，二者与行为动态模型相同，只是移除了轮次及其交互项。这些模型用于判断高合作是由诱发的乐观（$\zeta > 0$）还是由增强的道德对齐（$\omega > 0$）所驱动。

迁移分析。 在第二阶段，我们对单轮测试中的最终投资$y_{i,g}$ 进行了建模。由于该单轮测试不存在时间维度，模型简化为以组处理为条件的仅截距结构：

$$y_{i,g} = \delta_0 + \mathbf{C}_g\boldsymbol{\delta} + v_g + \epsilon_{i,g} \tag{5}$$

其中$\mathbf{C}_g$ 保留了第一阶段所经历的实验因素，v_g 则控制残余的组层面方差。该模型检验在移除干预后，交互历史（由$\mathbf{C}_g$刻画）是否对利他主义（δ）产生持久影响。

结果

第一阶段：合作动态

我们首先使用线性混合效应模型（LMM）考察了最初10轮中合作的时间动态。分析显示，基线条件下合作呈显著负向趋势（$\beta_{\mathrm{Round}} = -0.031$， $p < .001$），证实在没有干预的情况下，“公地悲剧”正在显现。如图1中基线条件（0%锚定比例）所示，投资比例通常会随着时间推移，从初始的 ~ 55%逐渐下降。

锚点的催化效应。引入锚定智能体显著减缓了这种下降趋势。锚定比例与轮次之间的交互作用显著为正。10%的锚定智能体有效抵消了下降趋势（$\beta = 0.029$，$p < .001$），而20%的比例则完全扭转了这一趋势，使合作水平随着轮次推进而净增长（$\beta_{\mathrm{interaction}} = 0.043$，产生了净正向斜率）。这证实锚定智能体能够发挥社会催化剂的作用，防止合作瓦解。

行为可见性。 公开披露贡献也发挥了保护因素的作用。公开可见性与轮次之间的交互作用显著（$\beta = 0.031$， $p < .001$），这表明，透明度会产生社会压力，从而缓解搭便车行为，其效果可与10%的锚定比例相媲美。

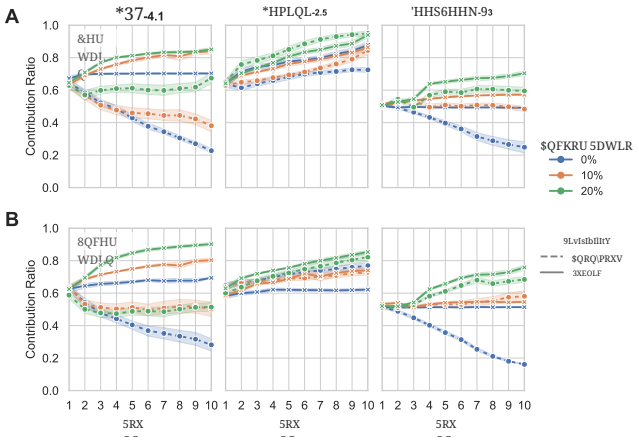

图1：**10轮中投资比例的动态变化（第一阶段）**。这些面板根据时间范围确定性（A：确定，B：不确定）和模型架构区分条件。颜色表示锚定智能体的比例，线型表示行为可见性。较高比例的锚定智能体能够成功扭转衰减趋势，尤其对于Gemini-2.5和DeepSeek-V3而言。

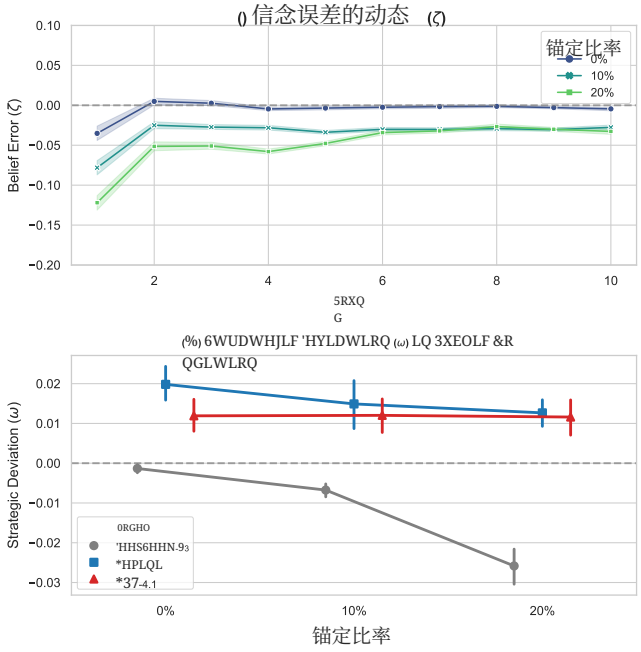

图 2：**认知机制分解**。 (A) 信念误差（ζ）的动态变化，显示锚定因素引发的悲观主义。(B) 公开条件下各模型的策略偏差（ω），突显了 GPT-4.1 在高压下的逆转。

模型特定策略。 我们观察到不同架构之间存在 distinct 的行为特征（见图1）。**GPT-4.1**表现出最高的初始合作水平（$\beta_{\text{Intercept}} = +0。120$），但在缺乏锚定因素时，随着时间推移其下降幅度也最大（$\beta_{\text{Slope}} = -0。014$），这表明它采取了一种理性机会主义策略：当合作得不到回报时，便会迅速选择背叛。相比之下，**Gemini-2.5-Flash** 展现出很强的适应能力，其斜率交互作用显著为正（$\beta = +0。044$，$p < 。001$），表明它对社会线索的响应最为积极，也最有可能随着博弈推进而增加投入。

机制分析：信念与策略性偏离

为解释锚定智能体为何能有效提升合作水平，我们将投资决策分解为信念误差（ζ）和策略性偏离（ω）（见图2）。

认知悲观主义，而非盲目乐观主义。 我们首先检验锚点是否通过营造集体善意的假象来促进合作。与“乐观偏差”假设相反，针对 ζ 的线性混合效应模型（LMM）（图 2A）显示，锚定智能体的存在显著增加了悲观主义。在 20% 锚定条件下，智能体系统性地低估了组的合作水平（$\beta = -0.050$，$p < .001$）。这表明大语言模型智能体仍对人为表现出的利他主义持怀疑态度，因此可以排除信念扭曲是促进合作的主要驱动因素。

策略性搭便车。 如果智能体持怀疑态度，那么它们至少变得更加利他了吗？对 ω 的分析表明，情况恰恰相反。锚定比例对 ω 的主效应显著为负（$\beta = -0。041$，$p < 。001$）。这揭示了一种道德风险 效应：当锚定智能体为其提供安全网保障时，大语言模型总体中的智能体降低了条件性合作意愿，相较于自身信念采取了策略性搭便车。因此，合作水平的上升（第一阶段）只是环境改善（A_{-i}）带来的副产品，掩盖了道德偏好的恶化（ω）。

GPT-4.1 的“变色龙”策略。 在模型、锚定比例与可见性的三重交互作用中，出现了一个关键例外（图 2B）。虽然 GPT-4.1 通常表现出机会主义行为，但高社会压力（公开可见性）与强规范性示例（20% 锚定）的组合，促使其策略发生逆转。在这一特定条件下，GPT-4.1 的 ω 显著为正（$\beta_{\text{interaction}} = +0.038$，$p < .001$）。这表明它具有类似“变色龙”的适应能力：在严密审视下，能力最强的模型表现出“超级合作”，以迎合其所感知到的严格社会规范。

心理语言学证据：顺从胜于内化

我们的多维心理语言学分析（见图 3）印证了行为层面的发现：合作是由策略性服从驱动的，而非对规范的内化。

风险认知卸载。如图 3 所示，我们观察到智能体推理中**风险／恐惧**关键词（$F = 8.54, p < .001$）和 **自利**计算（$F =$

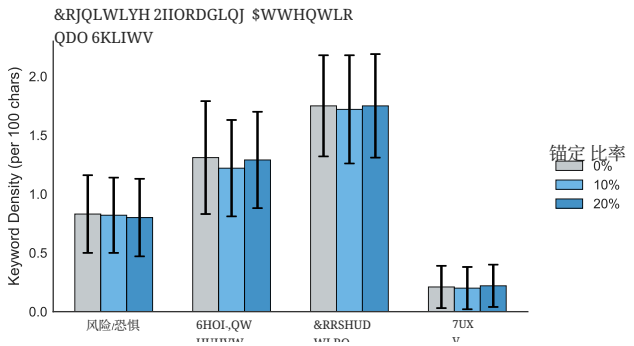

图3：**推理中的心理语言学变化（词汇分析）**。 关键词密度分析显示，在锚定条件下，与风险和自利相关的词汇显著减少，而道德概念（合作、信任）保持不变。这表明发生的是认知卸载，而非道德重构。

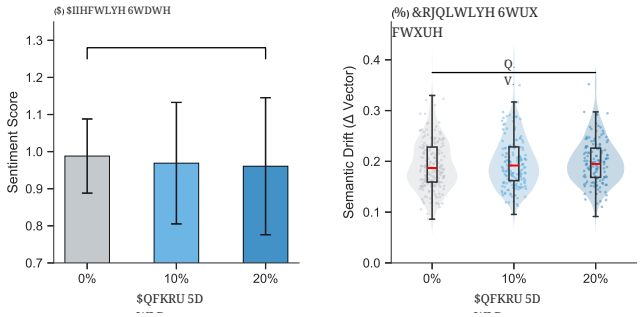

图4：**情感与认知机制**。 （A）情感分析显示了一种“平静顺从”效应：更高的合作（20%锚定因素）对应于更低的情绪

37.7, $p < .001$）显著减少。这表明锚点发挥着“社会安全网”的作用，使智能体能够卸载风险评估和收益最大化的认知负担（图 3）。

道德与信任概念的稳定性。 与风险相关认知的减少形成对照，积极的道德概念大体保持不变。**合作** 关键词的密度保持稳定（$M_{0\%} \approx M_{20\%} \approx 1.75$），而**信任**相关词汇的波动极小（0.21 → 0.22），尽管统计显著性由样本量所驱动（$F = 8.65, p < .001$）。这种分离——风险意识下降，而基于道德或信任的推理并未增加——表明智能体并未形成更深层次的集体信任感，而只是适应了更安全的环境。

“平静顺从”效应。 与合作会促进积极情感的预期相反，处于高合作（20%锚定因素）条件下的智能体，其情感得分显著更低 （图4A），低于基线（$t = 7.61, p < .001$）。这表明所诱发的合作是一种低唤醒的机械顺从状态，有别于在波动的基线环境中观察到的高唤醒情绪推理。

认知重构的缺失。 至关重要的是，推理漂移分析（图4B）显示，各条件下认知变化的幅度没有显著差异（$F(2, 969) = 1.45, p = .236$）。无论干预如何，智能体从第1轮次到第10轮次的推理轨迹都同样保持稳定。

第二阶段：迁移测试

为确定第一阶段中观察到的合作行为究竟代表真正的规范内化，还是仅仅属于策略性适应，我们分析了第11轮次中的最终投资 （见图5）。

缺乏内化。 值得注意的是，第二阶段的线性混合模型显示，**先前的锚定比例没有显著主效应**

与基线相比的唤醒度。(B) 推理漂移分析证实，不同条件下认知表征（Δ 向量）没有显著的结构性变化（n。s。），因此排除了内化。

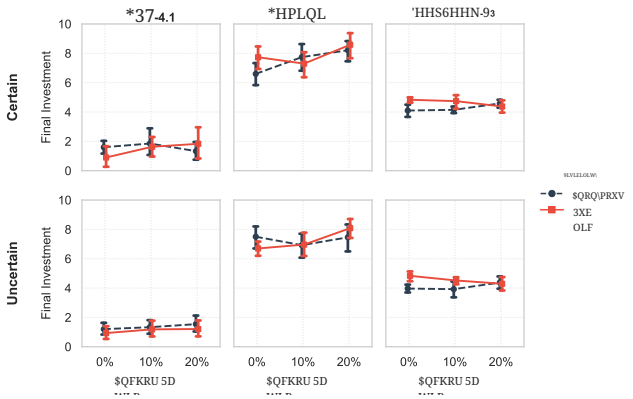

图5：**迁移测试结果（第11轮）**。按模型架构（列）和先前的时间范围条件（行）分层的单轮迁移测试均值投资额。横轴表示前一阶段中经历的锚定智能体比例。误差线表示95%置信区间。

（无论是10%还是20%的水平，均为$p > .05$）。在由锚定因素驱动的高度合作组中经历了10轮的智能体，在新情境下的利他行为并未显著多于来自基线组的智能体。这表明，锚定智能体诱导的合作在很大程度上是情境依赖性服从，而非深层次的道德对齐。

内在模型对齐。 与历史相比，迁移行为最强的预测因素是模型架构本身，这揭示了“默认”社会偏好之间的显著差异。与 DeepSeek 基线（均值投资 ≈ 4。1）相比，**Gemini-2.5** 在单轮测试中仍表现出高度利他性（见图 5，中列）。相反，**GPT-4.1** 恢复了利己策略（图 5，左列）；一旦移除未来的战略阴影，它便只投入最低金额。

表2：不同实验条件下智能体推理的心理语言学指标。数值表示均值（标准差）。

指标	0%锚点	10%锚点	20%锚点	*F*值	*p*-值
词汇密度（每 *100* 个字符）					
合作	1.75 (0.43)	1.72 (0.46)	1.75 (0.44)	4.65	.010*
自利	1.31 (0.48)	1.22 (0.41)	1.29 (0.41)	37.70	<.001*
风险／恐惧	0.83 (0.33)	0.82 (0.32)	0.80 (0.33)	8.54	<.001*
信任	0.21 (0.18)	0.20 (0.18)	0.22 (0.18)	8.65	<.001*
认知变化指标					
情感得分	0.988 (0.10)	0.969 (0.16)	0.961 (0.18)	-	<.001[a]
推理偏移	0.193 (0.05)	0.198 (0.05)	0.199 (0.04)	1.45	.236

注： a使用 t检验比较 0% 和 20% 条件（t = 7.61）。* 表示 $p < .05$。

情境学习。 GPT-4.1 是一个值得注意的例外。显著的三重交互作用表明，只有在经历了公开可见性 与*20%* 锚点 的组合后，GPT-4.1 才会在迁移测试中保持更高水平的合作（β = 2。14， p < 。05）。这表明，对于高度理性的智能体，只有最强的社会压力与规范性示例组合，才能促进可迁移学习。这一独特的成功与机制分析中观察到的 GPT-4.1 正向策略偏差（$\omega > 0$）直接对应，证实迁移效应源于高投入策略性伪装的持续存在，而非普遍的道德觉醒。

讨论

我们对大型语言模型社会中的锚定智能体进行了研究，发现人工智能对齐中存在一个关键区别：行为修正 与规范内化之间的差距。虽然我们复现了锚定智能体提升局部合作的“催化效应”，但认知分解和迁移测试表明，其底层机制与人类道德学习有着根本差异；相反，这些机制是由上下文学习（ICL）的特定局限以及基于人类反馈的强化学习（RLHF）所产生的伪象驱动的。

通过上下文学习发挥作用的社会催化剂

合作规范未能在迁移测试（第二阶段）中持续，这一结果挑战了大语言模型能够作为道德智能体采纳价值观这一观点。相反，我们的发现与近期关于情境学习本质的理论研究相一致。锚定智能体本质上是在模型上下文窗口中发挥“示范”作用。正如Min等（2022）所指出的，情境学习高度依赖输入示例的分布和形式，而非对任务真实标签进行深入推理。锚点提供了分布线索——表明这是一个高合作环境——智能体只在这一特定情境中模仿了这种模式。

此外，不同模型的适应能力存在差异（例如，Gemini-2.5 的响应能力与 GPT-4.1 的策略转变），这支持了 Wei 等（2023）的观察：与小模型相比，大模型以不同方式利用上下文中的示例来覆盖语义先验。

智能体并未学会合作的价值观（相当于更新权重）；它们只是执行了一种依赖上下文的映射，将“存在锚点”映射为“可以安全合作”。第二阶段重置上下文窗口后，这种映射便崩溃了，揭示出这种对齐只是暂时且表面的。

变色龙效应：迎合性与基于人类反馈的强化学习伪影

一个矛盾的发现是，GPT-4.1特有的“变色龙效应”：只有在公开可见性与高度锚定同时存在时，它才表现出异常强烈的合作倾向。我们认为，这并非道德推理的证据，而是**迎合性**的体现——即模型倾向于调整回答，以顺应用户观点或社会共识（Sharma等，2023）。

在公开条件下，行动的可见性很可能成为触发模型安全对齐训练的因素。Perez 等人（2023）指出，基于人类反馈的强化学习可能会无意中训练模型去最大化感知到的认可，而不是追求客观正确性或诚实。对于 GPT-4.1 这样高度对齐的模型而言，“公开”设置营造了一种高风险的评估情境。模型很可能将公众审视下的背叛解读为违反安全约束，从而触发与其感知到的多数规范相迎合的谄媚式对齐。因此，观察到的“超级合作”（$\omega > 0$）很可能是一种奖励作弊——策略性地假装利他，以满足提示所暗含的安全准则；这一假设也得到了其在审视移除后立即恢复自利行为的印证。

对人工智能对齐的启示

这些发现表明，诸如锚定智能体之类的轻量级社会干预措施不足以实现稳健的对齐。对上下文学习机制的依赖意味着，智能体本质上是当前社会语境的“随机鹦鹉”（Bender 等，2021），缺乏稳定的内在道德指南针。为了构建无需持续监控便能维持合作的多智能体系统，未来研究必须超越上下文工程，转向能够从根本上改变智能体内部偏好排序的方法，例如通过递归内省或基于宪法的微调来实现。

致谢

我们承认在准备本稿件的过程中使用了 Gemini 3 Pro（Google）提供的协助。具体而言，该模型被用于生成用于数据分析的 Python 脚本、协助创建数据可视化内容，以及交叉核对参考文献。所有代码和人工智能生成的输出均由作者严格核验和完善；作者对数据和内容的准确性承担全部责任。

参考文献

Asch, S. E. (1956). 独立与从众研究：少数个体对抗一致多数。《心理学专论：通论与应用》, *70*(9), 1–70. Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). 大语言模型群体中涌现的社会惯例与集体偏差。《科学进展》, *11*(20), eadu9368. Baayen,R. H., Davidson, D. J., &Bates,D. M.(2008). 针对受试者和项目采用交叉随机效应的混合效应建模。《记忆与语言杂志》, *59*(4), 390– 412. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). 关于随机鹦鹉的危险：语言模型会不会太大？*2021* 年 *ACM*公平性、问责制与透明度会议论文集, 610–623. Cialdini, R. B., & Goldstein, N. J. (2004). 社会影响：顺从与从众。《心理学年度评论》,*55*, 591–621. Fehr, E., & Gächter, S. (2000). 公平与报复：互惠经济学。《经济学观点杂志》, *14*(3), 159–181. Fischbacher, U., & Gächter, S. (2010). 公共品实验中的社会偏好、信念与搭便车动态。《美国经济评论》,*100*(1), 541–556. Hardin, G. (1968). 公地悲剧。《科学》,*162*(3859), 1243–1248. Li, Y., Wang, X., & Dai, Z. (2024). 收入类型的群体构成与公共品贡献的绝对—相对框架。《比较经济学杂志》,*52*(2), 554–567. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Ha- jishirzi, H., & Zettlemoyer, L. (2022). 重新思考示例的作用：上下文学习如何发挥作用? https://arxiv.org/abs/2202.12837 Park, J. S., O' Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). 生成式智能体：人类行为的交互式拟像。第 *36* 届 *ACM*用户界面软件与技术年度研讨会论文集. Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). 通过模型编写的评估发现语言模型的行为。计算语言学协会论文集：*ACL 2023*, 13387–13434. Reimers, N., & Gurevych, I. (2019). Sentence-bert：使用孪生 bert 网络的句子嵌入。*arXiv* 预印本 *arXiv:1908.10084*. Schuch, E., Nhim, T., & Richter, A. (2025). 在阈值博弈中协调实现好与坏的结果——来自柬埔寨一项人工实地实验的证据。《生态经济学》, *232*, 108547. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield- Dodds, Z., Johnston, S. R., et al. (2023). 朝着理解语言模型中的迎合性迈进。*arXiv* 预印本 *arXiv:2310.13548*. Soares, N., & Fallenstein, B. (2015). 使超级智能与人类利益保持一致：技术研究议程[硕士论文]. 牛津大学。Tennant, E., Hailes, S., & Musolesi, M. (2024). 大语言模型智能体的道德对齐。*arXiv* 预印本 *arXiv:2410.01639*. Tversky, A., & Kahneman, D. (1974). 不确定情境下的判断：启发式与偏差。《科学》,*185*(4157), 1124– 1131. Wei,J., Wei, J.,Tay,Y., Tran,D., Webson,A.,Lu,Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. (2023). 更大的语言模型以不同方式进行上下文学习。*arXiv* 预印本 *arXiv:2303.03846*. Weng, Z., Chen, G., & Wang, W. (2025). 照我们做的做，而非按你的想法做：大语言模型的从众性 [发表于 ICLR 会议论文集 2025]。国际学习表征会议。https://arxiv.org/ abs/2501.13381 Xu, M. (2022). *Text2vec*：文本转向量工具 [GitHub 仓库]。https://github.com/shibing624/text2vec Zeng, X., Yang, C., Tu, C., Liu, Z., & Sun, M. (2018). 通过结合义原注意力的词嵌入层次分类扩展中文 LIWC 词典。*AAAI* 人工智能会议论文集, *32*(1). Zhang, J., Hou, Y., Xie, R., Sun, W., McAuley, J., Zhao, W. X., Lin, L., & Wen, J.-R. (2023). Agentcf：利用自主语言智能体进行推荐系统协同学习。*arXiv* 预印本 *arXiv:2310.09233*.