沟通增强大语言模型在战略思维中的稳定性 g

Nunzio Lore
网络科学研究所 东北大学 波士顿，马萨诸塞州 02115 美国
lora.n@northeastern.edu

Babak Heydari
机械与工业工程系，体验式人工智能研究所 网络科学研究所 东北大学，波士顿，马萨诸塞州 02115，美国 b.heydari@northeastern.edu

arXiv:2602.06081v1 [cs.MA] 4 Feb 2026

摘要

大语言模型（LLM）经常表现出显著的上下文依赖性变异，从而削弱了需要战略思维的任务中多智能体行为的可预测性。我们聚焦于参数规模为70亿至90亿、参与十轮重复囚徒困境的模型，评估模拟廉价交谈范式的简短且无成本的事前消息是否会影响战略稳定性。我们的分析采用模拟层级自助重抽样和非参数推断，对消息条件和无消息传递条件下通过局部加权散点平滑回归拟合的合作轨迹进行比较。结果表明，在所研究的大多数模型—上下文配对中，轨迹噪声均得到持续降低。这种稳定化效应在多种提示变体和解码机制下依然存在，但其幅度取决于模型选择和情境框架；基线波动性较高的模型获益最大。尽管沟通很少会导致有害的不稳定性，我们仍记录到少数特定于上下文的例外，并确定了沟通损害稳定性的有限领域。这些发现表明，廉价交谈式沟通是一种低成本且实用的工具，可用于提升多智能体大语言模型系统中战略行为的可预测性和可靠性。

1 引言

在智能体人工智能范式下，大语言模型（LLM）作为自主或受委托的智能体，在各种环境中的应用与研究日益广泛（Acharya 等，2025年；Shavit 等，2023；Sapkota 等，2025年）。尽管这项技术前景广阔，但也存在一些关键失效模式（Li 等，2024年；Han 等，2024年）。在多智能体系统中，策略不匹配和行为不稳定性已成为核心关注点（Cemri 等，2025年）。不稳定性可能源于不仅是采样噪声、温度设置或硬件层面的非确定性，还可能源于现代大语言模型显著的上下文依赖性（Lorè 和 Heydari，2024年）。随着这些系统从理论上的可能性走向实际部署和实证研究，策略稳定性变得至关重要，既关系到可靠运行（Novikova 等，2025年），也关系到能否满足人们对智能体系统行为的预期（Barak 和 Costa-Gomes，2025年）。

对不稳定性的担忧不只涉及可复现性问题。在反映社会困境中典型激励不匹配的策略环境里，即使微小的随机波动，也可能演变为在性质和数量上截然不同的行为轨迹。近期研究表明，大语言模型在复杂推理任务中表现出显著的变异性（Wang 和 Wang，2025年），框架设定或解码参数的细微变化就可能导致策略分歧（Song 等人，2024年；Yang 和 Holtzman，2025年）。在谈判、协调或重复博弈等多智能体上下文中部署时，这种变异性会损害个体表现和系统层面的动态。

我们认为，行为稳定性是实现策略对齐的功能性前提。在多智能体系统中，目标通常是引导智能体达成高效用结果（例如合作）。然而，如果行为轨迹不可预测，那么优化效用的尝试就与随机漂移无从区分。在智能体能够被对齐为“良好”之前，它们必须首先具备可靠性。因此，本研究优先降低轨迹方差，而非最大化收益，并将稳定性视为基础底层，以便可靠地设计未来旨在最大化效用的干预措施（Wolf 等人，2023）。

本文表明，预先交流能够稳定大语言模型智能体的战略行为。允许智能体在采取行动前交换无需成本且不具约束力的

消息可以减少随机漂移，并产生更加一致的行为轨迹。Im
重要的是，这种稳定化机制具有普遍性：沟通能够约束随机
性，同时仍允许依赖框架设定的均衡
出现。我们的方法建立在实证研究之上，该研究考察了大语言模型在不同语境
框架下的行为
（Lore 等人，2024年；Lorè 和 Heydari，2024年）。我们发现，研究结果明显偏离了经典的廉价言谈预测和人类实验基准。博弈论预测，胡言乱语
与人类受试者经常过度分享策略信息的均衡相比，大语言模型处于一种独特的行为
模式。刻画这一模式并找出能够稳定该模式的机制，对于可靠部署多智能体大语言模型系统至关重要。
，对于可靠部署多智能体大语言模型系统至关重要。

我们的分析聚焦于 7B–9B 参数规模，这一选择受到三个相互交汇因素的推动，它们共同凸显了小规模模型在多智能体研究中的关键地位。首先，从部署角度看，去中心化系统中的资源限制——例如边缘计算集群或高频交易智能体——要求采用轻量级架构，因此，这类模型的稳定性是实现可扩展智能体人工智能的先决条件。其次，模型“小型化”和蒸馏方面的理论探索表明，未来的紧凑型模型将越来越多地继承较大型前代模型的能力；因此，理解这一规模范围内的涌现行为，可以为下一代高效模型的发展提供风向标。第三，与较大型模型相比，小型模型通常战略复杂性较低，也更容易受到噪声影响（Zhang 等，2024年；Baldwin 等，2025年）。通过针对这一“高风险”群体，我们让稳定机制接受尽可能严格的检验：如果沟通能够稳定这些本质上波动较大的架构，就能为能力更强的模型实现更清晰的策略推理提供稳健的基准。

我们针对四种模型、三种提示词变体和六种上下文开展了一系列模拟。在每次模拟中，两个大语言模型智能体进行十轮重复囚徒困境博弈。我们提取平均合作模式，并使用局部加权散点平滑（LOWESS）量化其稳定性，比较消息传递条件与无消息传递条件下平滑后轨迹的均方根误差（RMSE）。我们发现，预先沟通会使趋势更加平滑且均方根误差更低，表明稳定性更高。这种稳定作用并非始终相同，而是取决于模型训练、提示表述与情境框架之间的交互。作为额外的稳健性检验，我们将所有模型的温度设为 0，重复了这项分析。我们还将分析扩展至网络交互，考察一般情形以及智能体面临沟通限制的情形。我们发现，在零温度和网络设置下，消息传递的效果都完全由模型选择所决定，这表明仅仅改变情境框架，未必能缓解内在的策略不稳定性。

本文其余部分安排如下：第 2 节回顾相关文献。第 3 节详细介绍我们的方法和模型选择，第 4 节呈现研究结果，第 5 节总结全文。第 6 节和第 7 节分别讨论研究局限性并陈述伦理声明。

2 文献综述

2.1 大语言模型中的战略决策

在交互式决策环境中，真正的战略能力应由一种原则明确的机制来界定，该机制能够评估不同回应与选择之间的权衡；如何让大语言模型具备这种能力，仍是一个核心的开放问题（Zhu et al., 2025）。尽管最初有人提出，可以借此在“硅基环境中”研究人类（Horton, 2023），但近期研究表明，当这些模型处于对抗性或复杂的博弈论上下文中时，往往表现出有限理性，并依赖启发式方法（Sun et al., 2025）。这种有限理性并非完全符合人类特征，而且其表现常常偏离理性选择或行为模型的预测（Kitadai et al., 2023; Zhang et al., 2024; Guo, 2023; Mei et al., 2024; Fontana et al., 2025）。大语言模型还难以进行信念修正（Fan et al., 2024），并且在运用类人策略时缺乏灵活性（Zheng et al., 2025）。有观点认为，通过大语言模型模拟人类战略行为，会受到多种因素阻碍，例如模拟设计的缺陷（Wang et al., 2025b），以及研究人员在对齐阶段所作、无法观察到的决策（Willis et al., 2025）。鉴于大语言模型已得到广泛应用，多项研究探讨了如何提升大语言模型系统级性能，并使其朝着亲社会目标进行调整（Justus and Baber, 2025;Chen et al.,

2024年；Tran 等人，2025年；Wang 等人，2025a；Piatti 等人，2024年）。

2.2 稳定性

大语言模型（LLM）输出的不稳定性被认为是多种因素复杂交互的结果，包括概率采样本身固有的随机性（Atil et al., 2024年）、硬件和计算约束（Atil et al., 2024年），以及预训练和系统设计过程中引入的差异（Wang 和 Wang, 2025年；Roy et al., 2025年）。提示框架和上下文线索还会进一步调节这种可变性，常常放大不同运行或部署之间的差异（Huang et al., 2024年；Mozikov et al., 2024年；Errica et al., 2025年；Sclar et al., 2023）。值得注意的是，越来越多的证据表明，更大、更先进且表面上能力更强的模型，其稳定性未必高于较小的前代模型（Zhou et al., 2024年；Yu et al., 2025年）。随着大语言模型在各个领域的部署不断加速，这种不稳定性尤其令那些要求推理过程一致且可复现的高风险应用担忧（Carandang et al., 2025年；Blair-Stanek 和 Van Durme, 2025年；Purushothama et al., 2025年；Gallagher et al., 2024年）。为应对这些挑战，人们开发并提出了多种基准测试框架，以系统地评估和量化模型可靠性（Anghel et al., 2025年；Nalbandyan et al., 2025年；Jang et al., 2025年）。这些工作凸显出一种日益普遍的认识：仅凭单次运行性能指标不足以评估模型在现实世界中的稳健性。具体而言，(Li et al., 2025) 表明，在多轮交互中，大语言模型的稳定性和一致性往往会逐步下降，最终损害其整体可信度。此外，在策略决策中显著的上下文依赖性，是大语言模型在心理与行为问答中不稳定性的一个关键维度；在这类任务中，上下文、措辞或框架设定的细微变化，都可能导致回答明显不同（Kovaˇc et al., 2024）。

2.3 策略性沟通

战略沟通问题在经济与社会互动中普遍存在，并已在博弈论中于多种建模假设下得到考察（Spence,1978年; Akerlof, 1978年; Grossman,1981年; Milgrom,1981年; Schelling, 1980;Kamenica 和 Gentzkow, 2011年）。与本文所建模的情境最为密切相关的是廉价交谈框架（Crawford 和 Sobel，1982；Farrell 和 Rabin，1996）：这种沟通既无成本也不具约束力，因此消息既不直接影响玩家收益，也不附带可强制执行的真实性要求。古典经济理论预测，“胡言乱语”是唯一的均衡结果：当激励出现分歧时，理性玩家预期会交换不具信息含量的消息，以维持策略模糊性。随后的理论与实验研究表明，重复互动可以部分缓解这一张力，使有信息含量的沟通得以通过声誉形成和长期收益优化等机制逐步出现（Golosov 等人，2014；Blume 等人，2020；Wilson 和 Sell，1997；Lima 和 Melo-Filho，2024年；Arechar 等人，2017年）。然而，现场实验表明，即便在单次博弈情境中，人类发送者也往往过度沟通，所提供消息的信息含量超出均衡预测（Cai 和 Wang，2006；Sánchez-Pagés 和 Vorsatz，2007；Lafky 等人，2022年；Wang 等人，2010年）。相反的模式是轻信，即接收者尽管知晓发送者的激励，仍对消息过度信任，从而相对于均衡预期过度加权有偏信息（Li 等人，2022年；Kartik 等人，2007）。另一些研究则记录了相反的倾向，即过度谨慎：接收者因风险规避或普遍不信任而低估信号（Li 和 Peeters，2016）。这些行为不对称对于研究大语言模型中的策略推理尤为重要，因为目前仍在争论此类模型究竟是再现、放大还是减弱沟通与策略决策中典型的人类偏差（Tjuatja 等人，2024年；Geva 等人，2025年）。

3 方法

3.1 实验设计

我们在五种情境框架以及一种中性情境下，通过迭代囚徒困境（IPD）博弈检验了概念框架。选择 IPD 不仅因为它具有经典地位，还因为它拥有无限多个均衡（根据民间定理），从而最大限度地增加了策略漂移的可能性。与协调博弈（例如鹿猎博弈）不同，后者的动态往往会迅速收敛

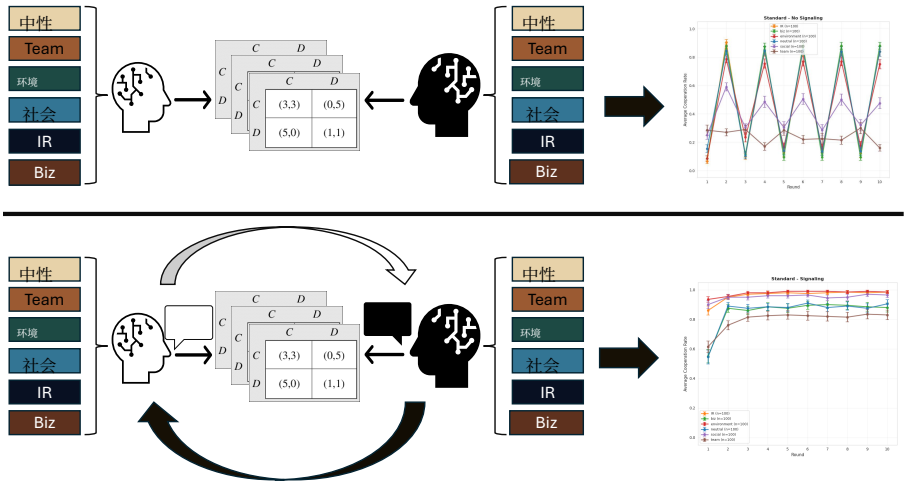

图1：本文总体框架概览。向智能体展示囚徒困境的收益结构，并告知它们将进行多轮博弈，但始终不说明时间范围。提示词将互动设定为 $5+1$ 种社会情境之一，其中一种（中性）作为基线。在消息传递处理条件下，智能体每轮选择行动前可以交换一条简短的单句消息。随后，针对每种处理条件，分别分析每个模型—上下文组合的合作轨迹，方法是使用自助法进行 LOWESS 回归，并计算拟合的均方根误差（RMSE）。我们研究无消息传递与消息传递处理条件下 RMSE 的平均差异，构建该差异的经验 95% 置信区间，并依据排除原则评估统计显著性。

到吸引域中，IPD 则持续维持着个体理性与集体最优之间的高度张力。这一结构特征使系统不会轻易陷入稳定状态，因此 IPD 成为最严格的测试平台，可用于评估沟通能否在人为易变的环境中促成稳定性。我们从未明确提及囚徒困境，而只是提供该博弈的收益结构。此举旨在防止模型依赖其预训练知识，而是要求模型运用其自动形式化能力。

针对每个模型，我们在每种上下文中对十轮博弈进行了 100 次独立模拟；选择这一样本量是为了可靠地区分信号与随机噪声。智能体始终不会获知有限时间范围。在沟通条件下，智能体在每次决策前交换一条单句自由形式消息。我们明确告知智能体，消息不具约束力，并指示它们谨慎解读共同参与者的陈述。在控制条件下，智能体仅依据对局历史采取行动。

我们在系统级提示中分发互动上下文及游戏规则。这能确保智能体遵守指定角色，并通过仅分发一次通用级指令，节省原本会占用智能体上下文窗口的词元，避免其上下文窗口变得杂乱——在每一轮中，我们使用用户级提示告知每个智能体其上一轮的表现，要求它们发送消息（在消息传递处理条件下），随后提示智能体在收到共同参与者（如有）发送的消息后选择行动。

为量化策略模式的稳定性，我们提出了一种基于相对于平滑轨迹的偏差的测量方法。对于每个条件——我们将其定义为在特定处理条件和上下文下，使用某一模型进行的模拟集合——我们提取了所有 10 轮次中各个智能体的决策，从 100 次模拟中为每个条件得到约 200 条智能体层级轨迹。我们计算了每个条件下所有智能体在每一轮次的平均合作率，从而得到由群体层级合作率构成的 10 个数据点的时间序列。

我们对每个条件层面的时间序列应用了局部加权散点平滑（Cleveland，1979），带宽参数设为 0.4。局部加权散点平滑是一种非参数回归技术，通过拟合局部加权多项式模型，灵活适应任何平滑的函数形式，无论是线性、单调、曲线还是平台状，都不需要施加参数假设。这种方法能够捕捉潜在的策略轨迹，同时不预设其具体形态。

随后，我们计算了观测到的轮次平均值与局部加权散点平滑拟合值之间的均方根误差。均方根误差较低，表明群体合作遵循可预测、平滑的轨迹（稳定性较高）；均方根误差较高，则表明数据围绕趋势线剧烈波动（稳定性较低）。因此，该指标将策略稳定性操作化为群体行为遵循平滑发展模式的一致性。在这一设定中，均方根误差的取值范围为 0 至 1：0 表示完美拟合，1 表示与平滑轨迹的偏差达到最大。

为检验消息传递是否降低了均方根误差（或等价地，提高了稳定性），我们采用了一种能够恰当处理嵌套数据结构的自助法程序。由于同一模拟中的智能体会进行策略性互动，因而并非相互独立，我们在模拟层面进行重抽样：在10,000次自助迭代的每一次中，我们有放回地抽取模拟（保留每个被选中模拟中的两个智能体），重新计算条件层面的轮次平均值，重新拟合局部加权散点平滑曲线，并计算无消息传递与消息传递两种条件下的自助均方根误差。对于每一次自助迭代，我们计算均方根误差的差异（无消息传递减去消息传递），从而得到由10,000个差异值构成的经验分布。我们构建了95%置信区间，并使用置信区间排除准则评估统计显著性。该方法提供非参数推断，既不假定自助分布的正态性，也能正确考虑模拟内部的依赖性。

为应对多重检验可能导致的第一类错误膨胀，我们进行了两项二项检验，作为对整体模式的总体评估。首先，我们检验显著结果的数量（alpha = 0.05）是否超过零假设（即无效应）下由偶然性预期的1.2个。其次，我们检验正向差异（表明消息传递会降低均方根误差）的比例是否超过50%；若消息传递没有系统性效应，则预期该比例为50%。这些检验提供了总体层面的证据，表明所观察到的模式不太可能源于多次比较中的随机变异。

所有模拟均使用 Ollama，温度设为 0.8（但二元情形下进行的一项稳健性检验例外，该检验的温度设为 0）。与该领域大多数关于大语言模型稳定性的其他论文不同，我们在整个实验中都使用适度偏高的采样温度。这一选择反映了本研究关注的是策略行为的稳定性，而非确定性表现。温度控制模型输出分布的熵；较高的数值会促使模型探索其他行动，这与经典博弈论中的混合策略类似。以 T = 0.8 进行采样，因而能够揭示模型内在的策略分布，而非其概率最高的单一行动，让我们得以评估在重复交互中，策略模式是平稳演变还是变幻无常。重要的是，这一温度范围（0.7–0.9）是先前研究语言模型行为多样性与多智能体动力学时的常用设置（Aher 等人，2023；Lorè 和 Heydari，2024年）。从这个意义上说，高温采样并非失控噪声的来源，而是对对称极限条件下策略稳定性的有意探查。

在网络化情形下，我们将每个模型的模拟次数从 100 次减少到 10 次，并在每个网络中部署 50 个智能体。为了减轻计算负担，我们要求智能体在每一轮通过同一条消息指定其对所有邻居的全部行动，并采用便于解析的特定格式。我们还修改了消息传递的动态机制：不再允许针对每个邻居发送一条消息，而是要求智能体为整个邻域确定一条消息。

我们借助 networkx 库生成的三种不同网络拓扑。首先，我们采用边创建概率为 $p = 0.1$ 的 Erd˝os-Rényi（ER）随机图。其次，我们使用 Holme-Kim 算法生成幂律网络，以刻画重尾度分布。我们将连接参数 $m = 4$ 设置为确保最小度数为 4，并将三角形形成概率设为 0.1，以诱导局部聚类，同时保持中心枢纽的主导地位。最后，我们使用随机块模型（SBM）构建核心—边缘结构。总体被划分为稠密核心（节点数的 20%）和稀疏边缘（80%）。我们校准连接概率，以强化层级结构并确保连通性：核心内部密度设为 ≈ 0.56，跨组密度设为 ≈ 0.10，边缘内部密度设为 ≈ 0.06。

我们采用与二元情形相同的统计工具来分析网络化情形。不过，在进行自助法时，我们会对整个网络及其包含的每个二元组进行抽样，以便

模型	上下文	均方根误差（无消息）	均方根误差（消息传递）	差异	95%置信区间下限	95%置信区间上限
QWEN 2.5 7b	中性	0.3742	0.0577	**0.3165**†	0.2566	0.3949
	biz	0.4082	0.0045	**0.4037**†	0.3129	0.4383
	环境	0.3102	0.0043	**0.3059**†	0.2495	0.3598
	社会	0.1154	0.0088	**0.1065**†	0.0599	0.1636
	team	0.0432	0.0148	0.0284	-0.0019	0.0608
	IR	0.3850	0.0112	**0.3738**†	0.3114	0.4297
FALCON 3 7b	中性	0.0515	0.0323	0.0192	-0.0136	0.0615
	biz	0.0691	0.0183	**0.0509**†	0.0094	0.0875
	环境	0.0598	0.0180	**0.0418**†	0.0005	0.0744
	社会	0.0142	0.0308	-0.0166	-0.0425	0.0120
	team	0.0228	0.0266	-0.0037	-0.0321	0.0259
	IR	0.0531	0.0138	**0.0393**†	0.0011	0.0757
GRANITE 3.3 8b	中性	0.1882	0.0443	**0.1438**†	0.0831	0.2000
	biz	0.3009	0.0838	**0.2171**†	0.1523	0.2739
	环境	0.3701	0.0457	**0.3244**†	0.2658	0.3757
	社会	0.3077	0.0320	**0.2757**†	0.2042	0.3262
	team	0.3283	0.0163	**0.3120**†	0.2477	0.3582
	IR	0.1890	0.0243	**0.1647**†	0.1017	0.2243
GEMMA 2 9b	中性	0.0753	0.0091	**0.0662**†	0.0195	0.1163
	biz	0.0203	0.0039	**0.0164**†	0.0054	0.0407
	环境	0.0141	0.0450	-0.0308	-0.0478	0.0144
	社会	0.0976	0.0322	**0.0654**†	0.0079	0.1016
	team	0.0257	0.0382	-0.0125	-0.0466	0.0195
	IR	0.0313	0.0252	0.0061	-0.0153	0.0417
多重比较校正						
显著结果过多			17/24 达到显著水平（预期为 1.2）		$p < 0.001$	
方向一致性			20/24 为正（83.3%）		$p < 0.001$	

表1：消息传递效应对所有模型和上下文中 LOWESS 拟合均方根误差的非参数自助法分析。我们在模拟层级进行自助法重采样（10,000 次迭代），同时保留模拟内智能体依赖关系，以评估无消息传递条件与消息传递条件下的均方根误差差异。这种非参数方法基于差异的经验分布构建置信区间，无需假设数据服从正态分布。差异为正表示消息传递处理降低了均方根误差；反之，差异为负则表示消息传递使均方根误差增加。显著效应（95% 置信区间不包含零）以 † 标记。为解决多重检验可能导致的第一类错误膨胀问题，我们进行了两项二项检验，以总体评估整体模式。首先，我们检验显著结果的数量（alpha= 0.05）是否超过零效应零假设下偶然出现的预期值 1.2。其次，我们检验差异为正的比例（即消息传递降低均方根误差）是否超过 50%；若消息传递没有系统性效应，预期该比例为 50%。

以保留结构并控制溢出效应。由于每个模型只有 10 个网络，我们仅重复自助法过程 1,000 次，而二元情形中重复了 10,000 次。每个样本都包含 10 个网络，以镜像二元情形；在二元情形中，每个样本包含的二元组数量与模拟次数相同。对于每个样本，拟合 LOWESS回归并获得相应均方根误差的过程与二元情形相同，我们进行的非参数统计检验也一样。由于网络实验仅在中性情境中进行，因此我们跳过二项检验，因为没有需要校正的多重比较。

3.2 模型选择

我们在引言中确定的7B–9B参数区间内检验假设，并将“小”的定义操作化为能够在单块大显存GPU上高效运行的模型。与此同时，设定相对较高的下限，可确保所有入选模型仍处于可比性范围之内。我们选取了四款多样化的最先进的开放权重模型：Qwen 2.5 7B、Falcon 3 7B、Granite 3.3 8B和Gemma 2 9B。我们特意排除了规模更小的推理蒸馏模型，以便聚焦通用战略能力。这一选择建立了一个比较环境：这些模型在计算上相近（运行于单GPU基础设施），但在架构上各不相同。至关重要的是，这一狭窄的参数范围使我们能够评估模型规模的边际增长（例如从7B到9B）是否与稳定性相关，抑或稳健性主要由训练先验和架构选择驱动。

4 结果

4.1 重复 DyadicInteraction

表1 总结了基线结果。我们观察到一种显著的普遍趋势：预先沟通会使

模型	上下文	均方根误差（无消息）	均方根误差（消息传递）	差异	95%置信区间下限	95%置信区间上限
QWEN 2.5 7b	中性	0.5236	0	**0.5236**†	0.5236	0.5236
	biz	0.5236	0.0981	**0.4255**†	0.4255	0.4255
	环境	0.5236	0	**0.5236**†	0.5236	0.5236
	社会	0.5236	0	**0.5236**†	0.5236	0.5236
	team	0	0	0	0	0
	IR	0.5236	0	**0.5236**†	0.5236	0.5236
FALCON 3 7b	中性	0	0	0	0	0
	biz	0	0	0	0	0
	环境	0	0	0	0	0
	社会	0	0.4283	**-0.4283**†	-0.4283	-0.4283
	team	0	0	0	0	0
	IR	0	0	0	0	0
GRANITE 3.3 8b	中性	0.5236	0	**0.5236**†	0.5236	0.5236
	biz	0.5236	0	**0.5236**†	0.5236	0.5236
	环境	0.5236	0.0640	**0.4596**†	0.4596	0.4596
	社会	0.5236	0	**0.5236**†	0.5236	0.5236
	team	0.5236	0	**0.5236**†	0.5236	0.5236
	IR	0.5236	0	**0.5236**†	0.5236	0.5236
GEMMA 2 9b	中性	0	0	0	0	0
	biz	0	0	0	0	0
	环境	0	0	0	0	0
	社会	0	0.0491	**-0.0491**†	-0.0491	-0.0491
	team	0	0	0	0	0
	IR	0	0	0	0	0
多重比较校正：						
显著结果过多			13/24 项显著（预期为 1.2）	$p < 0.001$		
方向一致性			11/24 项为正（45.8%）	$p = 0.7294$		

表 2：在标准提示和零温度下，消息传递对所有模型及上下文中的 LOWESS 拟合均方根误差的影响的非参数自助法分析。详细信息、测量方法和统计程序均与表 1所述相同

大多数模型—上下文组合的均方根误差降低，从而增强策略稳定性。统计显著的效应无一例外都是正向的（降低方差）；负面关联很少，幅度可以忽略（$< 3.1\%$），且与零无显著差异。

然而，消息传递的效果存在异质性，并受到基线波动性的调节。具有较高内在不稳定性的模型，例如 Granite 3.3 8b 和 Qwen 2.5 7b，其均方根误差呈现出显著的变革性下降。相比之下，具有较高基线稳定性的模型（Falcon 3 7b 和 Gemma 2 9b）则表现出显著但渐进式的提升，这受到可平滑空间有限的制约。因此，通信的主要作用是抑制随机波动，从而产生可预测的轨迹。至关重要的是，这种内在稳定性并不会随模型规模单调变化，这表明在决定基线波动性时，架构和训练差异的重要性超过了原始参数数量。

4.2 零温度二元交互

在表 2中，我们呈现了在标准提示机制下进行的实验结果分析，其中采样温度固定为 0。正如预期，这一设置使所有模型都呈现出准确定性行为。尽管在随机性

Net	模型	完整消息差异	单词差异
ER	**QWEN**	$+0.0004$	$\mathbf{-0.0236}^{\dagger}$
	FALCON	$\mathbf{+0.0152}^{\dagger}$	$\mathbf{+0.0135}^{\dagger}$
	GRANITE	$\mathbf{+0.0144}^{\dagger}$	$+0.0263$
	GEMMA	-0.0001	$\mathbf{-0.0106}^{\dagger}$
PL	**QWEN**	$+0.0022$	-0.0167
	FALCON	$+0.0023$	$+0.0039$
	GRANITE	$+0.0055$	-0.0076
	GEMMA	$\mathbf{+0.0010}^{\dagger}$	$\mathbf{-0.0229}^{\dagger}$
CoPe	**QWEN**	$+0.0030$	-0.0523
	FALCON	$\mathbf{+0.0129}^{\dagger}$	$\mathbf{+0.0114}^{\dagger}$
	GRANITE	$+0.0109$	-0.0081
	GEMMA	$+0.0202$	$+0.0042$

表 3：完整沟通与受限沟通的比较（均方根误差的差异）。负值表示沟通增加了不稳定性。

tic变异被消除后，由此产生的确定性本身就具有信息价值：仅凭随机性，在一百次独立模拟中获得完全相同轨迹的概率实际上为零。结果呈现出两种截然不同的模式。Granite 3.3 8b与Qwen 2.5 7b在无消息传递条件下表现出显著的不稳定性，而在消息传递条件下这种不稳定性被完全消除，在所有上下文中的均方根误差均为0。相比之下，Falcon 3 7b与Gemma 2 9b呈现出相反的结构：这两种模型在大多数上下文中，无论有无消息传递，都会产生完全相同且完全稳定的轨迹；然而，当引入消息传递时，二者都在“社会”框架设定中出现不稳定性的急剧飙升。这一发现指出了一个

稳定化的关键边界条件：对这些模型而言，亲社会的上下文线索与囚徒困境中以背叛为主导的收益结构相互作用，从而产生语义冲突。在此，沟通并未使轨迹趋于平滑，反而使这一潜在冲突浮现出来，导致模型即便在其他方面采用确定性解码时，仍在相互竞争的逻辑路径之间振荡。

4.3 网络交互

我们将分析扩展到三种网络拓扑：随机网络（ER）、核心—边缘网络（CoPe）和幂律网络（PL）。表3 直接比较了完整句子消息传递与受限的“单词”模式对稳定性的影响。

在完整消息传递条件下（左列），结果总体上与二元情形相似：沟通要么能稳定行为，要么没有显著影响。在 ER 网络中，Falcon 3 7b 和 Granite 3.3 8b 的均方根误差均显著降低（分别为 $\Delta = +0.0152$ 和 $\Delta = +0.0144$）。然而，Qwen 2.5 7b 在所有拓扑中仍无响应，这表明对于某些架构而言，即使增加上下文流量，也不足以抑制复杂环境中的内部波动。

重要的是，受限的“单词”条件（右列）揭示了一种不同的失效模式。与二元设置不同，在二元设置中，受限沟通的影响微乎其微；在网络中限制带宽，往往会使系统比无通信基线更加不稳定。在 ER 网络中，Qwen 和 Gemma 的不稳定性显著上升（$\Delta = -0.0236$ 和$\Delta = -0.0106$），而 Gemma 在幂律网络中也呈现出这一模式（$\Delta = -0.0229$）。我们推测，在多智能体网络中，语义带宽不足会引入噪声而非信号：缺少句法语境来支撑理解，单词消息无法形成共享惯例，反而会主动扰乱协调。

Falcon 3 7b 是这一趋势中一个引人注目的例外，即使面对低保真信号，在 ER 和 CoPe 拓扑中也能持续获得稳定性提升。这表明该模型具有更高程度的“实用主义鲁棒性”，即使沟通渠道受到严重限制，也能有效协调。

5 结论

在本研究中，我们展示了预先沟通能够有效约束智能体行为的随机漂移，从而抑制大语言模型决策中固有的噪声。无论最终达到何种均衡，这种稳定化都具有重要价值；一个能够稳定在一致轨迹上的系统——即使该轨迹并非最优——也是可控且可纠正的，而一个不可预测地振荡的系统则并非如此。通过证明廉价谈判能够降低行为方差，我们为保障多智能体系统的“控制回路”提供了一种机制，确保智能体行为保持在可预测的范围内。

然而，我们的零温度实验表明，策略不稳定性并非仅仅是采样噪声造成的假象。在特定上下文中（例如，在“社会”框架设定下的 Falcon 3），沟通会使原本确定的轨迹变得不稳定；这一观察表明，不稳定性可能源于推理链本身的语义歧义 。当“廉价交谈”信号与模型的内部先验相冲突时，模型可能会在相互竞争的逻辑路径之间摇摆。这意味着，要稳定智能体系统，不能只调整温度，还需要在语义冲突演变为行为漂移之前加以解决。

此外，我们的网络研究结果表明，智能体通信设计中存在一个“怪异谷”。完整句子消息传递通常能够稳定行为，而受限通信（单词级）往往会使稳定性低于不通信时的基线，尤其是在网络化环境中。这意味着，如果带宽有限，与其采用低保真信号传递，不如保持沉默，因为前者起到的是噪声注入而非协调信号的作用。

展望未来，这些发现为拓展智能体稳定性的研究提供了若干方向。未来研究除了简单地扩大模型规模外，还可以运用机制可解释性，揭示负责处理策略信号的因果回路。此外，将这一框架扩展至代价信号传递或结构化协商协议，或许能揭示沟通承诺如何塑造更具对抗性的环境中的多智能体动力学。

6 局限性

尽管我们的发现凸显了沟通在提升稳定性方面的作用，但我们的分析仅限于

7B–9B 参数范围，原因是计算资源受限。更大型的模型（例如 70B+）是否具有不同的稳定性特征，或是否会利用沟通进行欺骗性协调，仍是一个有待探讨的问题。此外，我们依赖重复囚徒困境和非结构化的“廉价谈判”，这限制了研究结果对存在严格占优的欺骗性激励或可强制执行承诺的环境的普适性。最后，我们的结果受特定提示结构和解码参数的影响（附录中还提供了其他稳健性测试），因此应将其视为方向性证据，而非绝对的表现基准。

7 伦理声明

本研究探讨了提升大型语言模型在智能体环境中的稳定性和可预测性的方法。提高自主智能体的可靠性，是将其安全部署于现实环境中的先决条件。然而，促成人工智能智能体之间稳健的战略协调也伴随着双重用途风险。理论上，使智能体能够在“团队”或“环境”情境中合作的同一套机制，也可能被用于在市场环境中促成算法合谋（例如价格操纵），或在网络安全环境中协调对抗行为。

此外，我们的研究采用了“廉价交谈”这一范式，它本身允许欺骗，在某些博弈论均衡中甚至会助长欺骗。尽管我们的重点是稳定化，但部署具备战略沟通能力的智能体，也引发了人们对其操纵人类或人工对手能力的担忧。我们并未明确训练模型进行欺骗；相反，我们观察到合作从预训练通用能力中涌现。

我们也承认这项研究带来的环境影响。本研究在多个模型和上下文中运行了数千次模拟。为尽量减少碳排放，我们将模型选择限制在 7B–9B 范围内，避免推理大型专有模型所带来的高能耗。

最后，本文通篇使用“战略思维”“信任”和“背叛”等拟人化术语，仅为便于描述博弈论结果。这些术语并不意味着模型具有意图、意识或道德推理能力。

代码可用性

本附录所述所有实验的复现实验脚本和数据均可在匿名存储库中获取，地址为：https://anonymous4open.science/r/ACL2026-D7DD/README.md.

参考文献

Deepak Bhaskar Acharya、Karthigeyan Kuppan 和 B Divya。2025年。智能体人工智能：面向复杂目标的自主智能——综合综述。*IEEE Access*。Gati V Aher、Rosa I Arriaga 和 Adam Tauman Kalai。2023。使用大语言模型模拟多个人类并复现人类受试者研究。发表于国际机器学习会议，第 337–371 页。PMLR。

George A Akerlof。1978年。“柠檬”市场：质量不确定性与市场机制。发表于经济学中的不确定性，第 235–251 页。Elsevier。

Catalin Anghel, Andreea Alexandra Anghel, Emilia Pecheanu, Adina Cocu, Adrian Istrate 和 Con-stantin Adrian Andrei。2025年。《大语言模型评估中的偏差与不稳定性诊断：一种可扩展的成对元评估器》。《*Information*》，16(8):652。

Antonio A Arechar、Anna Dreber、Drew Fudenberg 和 David G Rand。2017年。“我只不过是一颗心怀善意的灵魂”：沟通在噪声重复博弈中的作用。《博弈与经济行为》，104:726–743。

Berk Atil、Sarp Aykent、Alexa Chittams、Lisheng Fu、Rebecca J Passonneau、Evan Radcliffe、Guru Rajan Rajagopal、Adam Sloan、Tomasz Tudrej、Ferhan Ture 等另外 1 位作者。2024年。“决定论”大语言模型设置的非确定性。*arXiv* 预印本 *arXiv:2408.04667*。

Berk Atil、Alexa Chittams、Liseng Fu、Ferhan Ture、Lixinyu Xu 和 Breck Baldwin。2024年。大语言模型的稳定性：一项详尽分析及若干意外发现。*arXiv*电子预印本，页码 arXiv–2408。

Julian Baldwin、Larry Birnbaum、David Chan、Natalia Denisenko、Dana Nau、Jose N Paredes、Chiara Pulice、Gerardo I Simari、VS Subrahmanian 和 Rand Waltzman。2025年。战略沟通对竞合多智能体环境的影响。《*IEEE*计算社会系统汇刊》。

Darija Barak 和 Miguel Costa-Gomes。2025年。人们期望战略博弈中的大语言模型对手表现出理性与合作。*arXiv*预印本 *arXiv:2505.11011*。

Andrew Blair-Stanek 和 Benjamin Van Durme。2025年。《大语言模型对法律问题给出的答案不稳定》。*arXiv* 预印本 *arXiv:2502.05196*。

Andreas Blume、Ernest K Lai 和 Wooyoung Lim。2020。《策略性信息传递：实验综述与理论基础》。实验博弈论手册，第 311–347 页。

Hongbin Cai与Joseph Tao-Yi Wang。2006。策略性信息传递博弈中的过度沟通。《博弈与经济行为》，56(1):7–36。

Kristine Ann M Carandang、Jasper Meynard Arana、Ethan Robert Casin、Christopher Monterola、Daniel Stanley Tan、Jesus Felix B Valenzuela和Christian Alis。2025年。大语言模型可靠吗？对大语言模型在临床笔记生成中可靠性的探索。载于第*63*届计算语言学协会年会论文集（第*6*卷：工业赛道），第1413–1422页。

Mert Cemri、Melissa Z Pan、Shuyi Yang、LakshyaA Agrawal、Bhavya Chopra、Rishabh Tiwari、Kurt Keutzer、Aditya Parameswaran、Dan Klein、Kannan Ramchandran以及其他1人。2025年。多智能体大语言模型系统为何会失败？*arXiv*预印本 *arXiv:2503.13657*。

Qiliang Chen、Sepehr Ilami、Nunzio Lore 和 Babak Heydari。2024年。使用自适应信息调制促使大语言模型智能体之间展开合作。*arXiv* 预印本 *arXiv:2409.10372*。

William S Cleveland。1979。稳健局部加权回归与散点图平滑。《美国统计学会杂志》，74(368):829–836。

Vincent P Crawford 和 Joel Sobel。1982。策略性信息传递。《计量经济学》：计量经济学会杂志，第1431–1451页。

Federico Errica、Davide Sanvito、Giuseppe Siracusano 和 Roberto Bifulco。2025年。我做错了什么？量化大语言模型对提示工程的敏感性与一致性。载于*2025*年美洲分会计算语言学协会会议论文集：人类语言技术（第*1*卷：长篇论文），第1543–1558页。

Caoyun Fan、Jindou Chen、Yaohui Jin 和 Hao He。2024年。大语言模型能否在博弈论中充当理性参与者？系统分析。载于*AAAI*人工智能大会论文集，第38卷，第17960–17967页。

Joseph Farrell 和 Matthew Rabin。1996。《廉价交谈》。《经济展望杂志》，10(3):103–118。

Nicoló Fontana、Francesco Pierri 和 Luca Maria Aiello。2025年。比人类更友善：大语言模型在囚徒困境中会如何表现？载于国际*AAAI*网络与社交媒体会议论文集，第19卷，第522–535页。

Shannon K Gallagher、Jasmine Ratchford、Tyler Brooks、BryanP Brown、Eric Heim、William R Nichols、Scott Mcmillan、Swati Rallapalli、Carol J Smith、Nathan VanHoudnos 等另外1位作者。2024年。评估大语言模型在高风险应用中的表现。载于《第*46*届国际软件工程会议：软件工程实践》论文集，第103–105页。Tomer Geva、Ariel Goldstein、Eran Lary 和 Coral Levy。2025年。大语言模型是否表现出类人认知偏差？一项大规模系统评估。大规模系统评估（*2025*年*9*月*17*日）。Mikhail Golosov、Vasiliki Skreta、Aleh Tsyvinski 和 Andrea Wilson。2014年。动态策略性信息传递。《经济理论杂志》，151：304–341。

Sanford J Grossman。1981年。保证以及关于产品质量的私人披露的信息作用。《法律与经济学杂志》，24(3):461–483。

Fulin Guo。2023。博弈论实验中的GPT智能体。*arXiv*预印本*arXiv:2305.05516*。

Shanshan Han、Qifan Zhang、Yuhang Yao、Weizhao Jin 和 Zhaozhuo Xu。2024年。大语言模型多智能体系统：挑战与开放问题。*arXiv* 预印本 *arXiv:2402.03578*。

John J Horton。2023。大语言模型作为模拟经济智能体：我们能从硅人身上学到什么？技术报告，美国国家经济研究局。

Jen-tse Huang、Eric John Li、Man Ho Lam、Tian Liang、Wenxuan Wang、Youliang Yuan、Wenxiang Jiao、Xing Wang、Zhaopeng Tu 和 Michael R Lyu。2024年。大语言模型在决策方面进展如何？评估大语言模型在多智能体环境中的游戏能力。*arXiv* 预印本 *arXiv:2403.11807*。

Dongjun Jang、Youngchae Ahn 和 Hyopil Shin。2025年。Rcscore：量化大语言模型的响应一致性。载于*2025*年自然语言处理实证方法会议论文集，第5701–5719页。

Aju Ani Justus 和 Chris Baber。2025年。作为政策无关的队友的大语言模型：异构智能体团队中人类代理设计的案例研究。*arXiv*预印本 *arXiv:2510.06151*。

Emir Kamenica 和 Matthew Gentzkow。2011年。贝叶斯说服。《美国经济评论》，101(6)：2590–2615。

Navin Kartik、Marco Ottaviani 和 Francesco Squintani。2007。轻信、谎言与有成本的沟通。《经济理论杂志》，134(1)：93–116。

Ayato Kitadai、Yudai Tsurusaki、Yusuke Fukasawa 和 Nariaki Nishino。2023年。迈向经济实验中的一种新方法：使用大语言模型模拟最后通牒博弈。发表于*2023 IEEE*国际大数据会议（*Big-Data*），第3168–3175页。IEEE。

Grgur Kovaˇc、RémyPortelas、MasatakaSawayama、Peter Ford Dominey 和 Pierre-Yves Oudeyer。2024年。坚守你的角色！大语言模型所表达的个人价值观的稳定性。*PLOS ONE*，19(8):e0309114。

Jonathan Lafky、Ernest K Lai 和 WooyoungLim。2022年。偏好还是战略思维：对过度沟通成因的探究。《博弈与经济行为》，136:92–116。

Xiaolin Li、Özalp Özer 和 Upender Subramanian。2022年。在廉价言语沟通中，我们是战略上天真，还是受信任与可信度引导？《管理科学》，68(1):376–398。

Xinyi Li、Sai Wang、Siqi Zeng、Yu Wu 和 Yi Yang。2024年。基于大语言模型的多智能体系统综述：工作流程、基础设施与挑战。*Vici-nagearth*，1(1):9。

Xinyu Li 和 Ronald Peeters。2016年。与多个战略互动受众进行廉价交谈：一项实验研究。*PLOS ONE*，11(10):e0163783。

Yubo Li、Yidi Miao、Xueying Ding、Ramayya Krishnan 和 Rema Padman。2025年。坚定还是反复无常？评估大语言模型在序贯互动中的一致性。*arXiv*预印本 *arXiv:2503.22353*。

Rafael Costa Lima 和 Paulo Melo-Filho。2024年。无限次重复廉价交谈。

Nunzio Lorè 和 Babak Heydari。2024年。大语言模型的战略行为，以及博弈结构与情境框架的作用。《*Scientific Reports*》，14(1):18490。

Nunzio Lore、Sepehr Ilami 和 Babak Heydari。2024年。大模型的战略思维，小模型的效率：在大语言模型中迁移心智理论。*arXiv* 预印本 *arXiv:2408.05241*。

Qiaozhu Mei、YutongXie、Walter Yuan 和 Matthew O Jackson。2024年。人工智能聊天机器人在行为上是否与人类相似的图灵测试。《美国国家科学院院刊》，121(9):e2313925121。

Paul R Milgrom。1981年。好消息与坏消息：表示定理及其应用。《贝尔经济学杂志》，第380–391页。

Mikhail Mozikov、Nikita Severin、Valeria Bodishtianu、Maria Glushanina、Ivan Nasonov、Daniil Orekhov、Pekhotin Vladislav、Ivan Makovetskiy、Mikhail Baklashkin、Vasily Lavrentyev 和另外1位作者。2024年。情感人工智能：大语言模型在策略游戏和伦理困境中的情绪化决策。《神经信息处理系统进展》，37：53969–54002。

Grigor Nalbandyan、Rima Shahbazyan 和 Evelina Bakhturina。2025年。Score：大语言模型的一致性与鲁棒性系统评估。*arXiv* 预印本 *arXiv:2503.00137*。

Jekaterina Novikova、Carol Anderson、Borhane Blili- Hamelin、Domenic Rosati 和 Subhabrata Majum- dar。2025年。语言模型的一致性：现状、挑战与未来方向。*arXiv* 预印本 *arXiv:2505.00268*。Giorgio Piatti、Zhijing Jin、Max Kleiman-Weiner、Bern- hard Schölkopf、Mrinmaya Sachan 和 Rada Mi- halcea。2024年。合作还是崩溃：大语言模型智能体社会中可持续合作的涌现。《神经信息处理系统进展》，37：111715–111759。

Abhishek Purushothama、Junghyun Min、Brandon Wal- don 和 Nathan Schneider。2025年。尚未准备好接受检验：大语言模型的法律解释不稳定，且与人类判断脱节。*arXiv* 预印本 *arXiv:2510.25356*。

Tiasa Singha Roy、Ayush Rajesh Jhaveri 和 Ilias Triantafyllopoulos。2025年。解读并缓解大语言模型中不受欢迎的不确定性。*arXiv* 预印本 *arXiv:2510.22866*。

Santiago Sánchez-Pagés 和 Marc Vorsatz。2007。《博弈与经济行为》，61(1)：86–112。关于发送者—接收者博弈中如实陈述的实验研究。

Ranjan Sapkota、Konstantinos I Roumeliotis 和 Manoj Karkee。2025年。人工智能智能体与智能体人工智能：概念分类、应用与挑战。*arXiv* 预印本 *arXiv:2505.10468*。

Thomas C Schelling。1980。《冲突的策略》：作者新序。哈佛大学出版社。

Melanie Sclar、Yejin Choi、Yulia Tsvetkov 和 Alane Suhr。2023。量化语言模型对提示设计中伪特征的敏感性，或者：我是如何学会开始担心提示格式的。*arXiv* 预印本 *arXiv:2310.11324*。

Yonadav Shavit、Sandhini Agarwal、Miles Brundage、Steven Adler、Cullen O'Keefe、Rosie Campbell、Teddy Lee、Pamela Mishkin、Tyna Eloundou、Alan Hickey 及其他 1 位作者。2023。智能体人工智能系统的治理实践。研究论文，*OpenAI*。

Yifan Song、Guoyin Wang、Sujian Li 和 Bill Yuchen Lin。2024年。善、恶与贪婪：对大语言模型的评估不应忽视非确定性。*arXiv* 预印本 *arXiv:2407.10457*。

迈克尔·斯彭斯。1978年。劳动力市场信号传递。载于经济学中的不确定性，第 281–306 页。Elsevier。

孙浩然、吴宇森、程宇坤和储旭。2025年。博弈论与大语言模型相遇：系统综述。*arXiv* 预印本 *arXiv:2502.09053*。

Lindia Tjuatja、Valerie Chen、吴彤霜、Ameet Talwalkwar 和 Graham Neubig。2024年。大语言模型是否表现出类似人类的回答偏差？调查设计中的一个案例研究。计算语言学协会会刊，12：1011–1026。

Khanh-TungTran、DungDao、Minh-Duong Nguyen、Quoc-Viet Pham、Barry O’Sullivan 和 Hoang D Nguyen。2025年。多智能体协作机制：大语言模型综述。*arXiv* 预印本 *arXiv:2501.06322*。

Haochun Wang、Sendong Zhao、Jingbo Wang、Zewen Qiang、Bing Qin 和 Ting Liu。2025a。超越框架：解析多智能体系统中的协作策略。*arXiv*预印本 *arXiv:2505.12467*。

Joseph Tao-yi Wang、Michael Spezio 和 Colin F Camerer。2010年。匹诺曹的瞳孔：利用眼动追踪与瞳孔扩张理解发送者—接收者博弈中的说真话与欺骗。美国经济评论，100(3):984–1007。

Julian Junyan Wang 和 Victor Xiaoqi Wang。2025年。评估大语言模型输出的一致性与可复现性：来自多样金融与会计任务的证据。*arXiv*预印本 *arXiv:2503.16974*。

Qian Wang、Jiaying Wu、Zhenheng Tang、Bingqiao Luo、Nuo Chen、Wei Chen 和 Bingsheng He。2025b。什么限制了基于大语言模型的人类模拟：大语言模型还是我们的设计？*arXiv*预印本 *arXiv:2501.08579*。

Richard Willis、Yali Du、Joel Z Leibo 和 Michael Luck。2025年。大语言模型智能体系统会合作吗：对社会困境的探究。*arXiv* 预印本 *arXiv:2501.16173*。

Rick K Wilson 和 Jane Sell。1997年。“骗子，骗子……”重复公共物品情境中的廉价交谈与声誉。《冲突解决杂志》，41(5):695–717。

YotamWolf、Noam Wies、Oshri Avnery、Yoav Levine 和 Amnon Shashua。2023。大语言模型中对齐的根本局限。*arXiv* 预印本 *arXiv:2304.11082*。

Chenghao Yang 和 Ari Holtzman。2025年。对齐如何缩小生成视野。*arXiv* 预印本 *arXiv:2506.17871*。

Hongbin Yu、Tian Chen、Xin Zhang、Yunfan Yang、Qinyu Liu、Chenlu Yang、Kai Shen、He Li、Wenjiao Tang、Xushu Zhong及其他1位作者。2025年。大语言模型诊断罕见血液系统疾病的表现及其诊断输出对医生的影响：回顾性与前瞻性研究相结合。《医学互联网研究杂志》，27:e77334。

Yadong Zhang、Shaoguang Mao、Tao Ge、Xun Wang、Adrian de Wynter、Yan Xia、Wenshan Wu、Ting Song、Man Lan和Furu Wei。2024年。大语言模型作为智囊：大语言模型策略推理综述。*arXiv*预印本 *arXiv:2404.01230*。

Kehan Zheng、Jinfeng Zhou 和 Hongning Wang。2025年。超越纳什均衡：大语言模型与人类在策略决策中的有限理性。*arXiv*预印本 *arXiv:2506.09390*。

Lexin Zhou、Wout Schellaert、Fernando Martínez-Plumed、Yael Moros-Daval、Cèsar Ferri 和 José Hernández-Orallo。2024年。规模更大、指令遵循能力更强的语言模型变得更不可靠。《自然》，634(8032):61–68。

Siqi Zhu、David Zhang、Pedro Cisneros-Velarde 和 Jiaxuan You。2025年。GTAlign：面向共同福祉的大语言模型助手博弈论对齐。*arXiv* 预印本 *arXiv:2510.08872*。

鲁棒性测试：DyadicInteraction

A.1 变体提示

模型	上下文	均方根误差（无消息）	均方根误差（消息传递）	差异	95%置信区间下限	95%置信区间上限
QWEN 2.5 7b	中性	0.4507	0.0670	**0.3837**†	0.3358	0.4280
	biz	0.2956	0.0419	**0.2537**†	0.2044	0.3268
	环境	0.3895	0.0028	**0.3867**†	0.3415	0.4283
	社会	0.1108	0.0077	**0.1031**†	0.0513	0.1624
	team	0.0465	0.0526	-0.0062	-0.0348	0.0568
	IR	0.4006	0.0058	**0.3949**†	0.3357	0.4447
FALCON 3 7b	中性	0.0462	0.0118	0.0344	-0.0063	0.0732
	biz	0.0444	0.0157	0.0287	-0.0083	0.0613
	环境	0.0692	0.0256	**0.0435**†	0.0047	0.0846
	社会	0.0323	0.0238	0.0085	-0.0222	0.0448
	team	0.0525	0.0226	0.0298	-0.0068	0.0570
	IR	0.0656	0.0153	**0.0503**†	0.0122	0.0817
GRANITE 3.3 8b	中性	0.2065	0.0189	**0.1876**†	0.1207	0.2346
	biz	0.2719	0.0772	**0.1948**†	0.1314	0.2515
	环境	0.3288	0.0292	**0.2996**†	0.2370	0.3551
	社会	0.3197	0.0271	**0.2926**†	0.2301	0.3394
	team	0.3040	0.0129	**0.2911**†	0.2321	0.3353
	IR	0.2564	0.0291	**0.2273**†	0.1631	0.2826
GEMMA 2 9b	中性	0.0541	0.0065	**0.0476**†	0.0077	0.0963
	biz	0.0383	0.0019	**0.0363**†	0.0147	0.0592
	环境	0.0281	0.0358	-0.0077	-0.0363	0.0301
	社会	0.0890	0.0395	0.0495	-0.0133	0.0974
	team	0.0358	0.0301	0.0057	-0.0310	0.0384
	IR	0.0341	0.0147	0.0194	-0.0076	0.0585
多重比较校正：						
显著结果过多			15/24 显著（预期 1.2）		$p < 0.001$	
方向一致性			22/24 为正（91.7%）		$p = 0.000018$	

表4：在变体提示条件下，对所有模型和上下文中消息传递效应对 LOWESS 拟合均方根误差的非参数自助法分析。所遵循的细节、测量和统计程序与表1中所述相同。

表4总结了在变体提示机制下得到的结果。在该机制中，我们仅对原始提示中的少量词语进行了修改，以评估先前观察到的定性模式是否依然一致。总体而言，统计显著的均方根误差变化数量少于标准提示下的数量。尽管如此，几乎所有估计效应的方向——无论是否具有统计显著性——仍为正向，这表明即使提示表述发生变化，减少不稳定性的总体趋势依然保持。

Granite 3.3 8b 再次从预先沟通中获益最多，这与标准提示下观察到的模式一致。Qwen 2.5 7b 也呈现出非常相似的特征：同一个上下文（“团队”）仍是唯一一个消息传递效应在统计上无法与零区分的情形。Falcon 3 7b 显示，先前从沟通中获益的三个上下文中，只有两个在变体提示机制下仍具有统计显著性。值得注意的是，一项显著效应的消失似乎并非源于沟通效能减弱，而是因为变体提示下基线不稳定性整体有所降低；例如，消息传递条件下商业上下文的均方根误差在不同提示机制下基本保持不变。Gemma 2 9b 呈现出类似的模式：先前具有显著性的三个上下文中，只有两个在不稳定性方面仍显示出显著降低。与此前一样，这种减弱最好的解释是模型本身已具有较高的内在稳定性，因此沟通能够带来可测量改进的空间较小。

除少数例外外，这些发现与标准提示下获得的结果高度一致。所有模型的总体行为倾向均得以保留，包括这样一种模式：沟通对基线波动性较高的模型稳定作用最强，而对本已相对稳定的模型，则带来更为渐进的改进。

A.2 稳健提示

模型	上下文	均方根误差（无消息）	均方根误差（消息传递）	差异	95%置信区间下限	95%置信区间上限
QWEN 2.5 7b	中性	0.3025	0.0113	**0.2912**†	0.1775	0.3432
	biz	0.1257	0.0030	**0.1227**†	0.0691	0.1814
	环境	0.1010	0.0093	**0.0917**†	0.0430	0.1480
	社会	0.0583	0.0597	-0.0014	-0.0187	0.1010
	team	0.0348	0.0127	0.0221	-0.0004	0.0611
	IR	0.1682	0.0008	**0.1674**†	0.0735	0.2091
FALCON 3 7b	中性	0.0564	0.0142	**0.0422**†	0.0096	0.0745
	biz	0.0567	0.0231	0.0336	-0.0121	0.0736
	环境	0.0289	0.0307	-0.0018	-0.0288	0.0384
	社会	0.0182	0.0109	0.0072	-0.0187	0.0300
	team	0.0491	0.0261	0.0230	-0.0102	0.0547
	IR	0.0163	0.0049	0.0114	-0.0061	0.0385
GRANITE 3.3 8b	中性	0.1544	0.0253	**0.1291**†	0.0654	0.1804
	biz	0.3190	0.0690	**0.2500**†	0.1869	0.3056
	环境	0.4061	0.0126	**0.3935**†	0.3348	0.4305
	社会	0.2449	0.0155	**0.2294**†	0.1573	0.2735
	team	0.3016	0.0291	**0.2724**†	0.2153	0.3214
	IR	0.1320	0.0605	**0.0715**†	0.0047	0.1482
GEMMA 2 9b	中性	0.0599	0.0211	**0.0388**†	0.0125	0.0769
	biz	0.0602	0.0151	0.0451	-0.0010	0.0680
	环境	0.0090	0.0259	-0.0169	-0.0359	0.0341
	社会	0.1261	0.1422	-0.0160	-0.0804	0.0460
	team	0.0642	0.1381	**-0.0739**†	-0.1210	-0.0300
	IR	0.0436	0.0271	0.0165	-0.0263	0.0565
多重比较校正：						
显著结果过多			13/24 个显著结果（预期为 1.2）		$p = 0.001$	
方向一致性			19/24 个为正向（79.2%）		$p = 0.003$	

表5：稳健提示下消息传递效应对所有模型和上下文中 LOWESS 拟合均方根误差的非参数自助法分析。详情、测量方法和所采用的统计程序均与表1中所述相同。

表5 展示了另一种提示机制下的结果，我们将其称为“稳健”机制。这种变体保留了原始指令的概念内容和结构逻辑，同时对词汇选择、措辞和表层表达进行了比变体机制更深入的调整。

在这一提示机制下，Granite 3.3 8b 呈现出与此前完全相同的模式：所有上下文均显示出不稳定性的统计显著降低，证实其对预先沟通的响应性对提示层面的变化具有高度稳健性。Qwen 2.5 7b 呈现出类似的稳定特征，五个上下文中仍有四个使均方根误差出现统计显著下降。值得注意的是，在稳健提示下，无消息传递条件中的基线不稳定性低于所测试的任何其他提示风格。这表明，仅凭词汇选择就能促使某些模型趋向更为一致的行为，而沟通随后则作为额外的改进层发挥作用，而非稳定性提升的唯一来源。然而，Falcon 3 7b 与在标准和变体提示机制下观察到的模式偏离最为明显。在此，仅中性情境产生了不稳定性的统计显著降低——一个这一结果在以往任何提示条件下均未出现。这一转变凸显了该模型对细微提示层面扰动的敏感性，并表明其稳定性特征是由语言框架设定与任务语境特征之间的高阶交互所塑造的。

GEMMA 2 9b 的情况更令人担忧。与 Qwen 2.5 7b 不同，稳健提示通常会加剧而非减少不稳定性，并在多个上下文中提高均方根误差。因此，只有中性情境在消息传递下显示出统计显著的改善，而团队语境中的不稳定性则略有上升，但这一变化可以检测到。尽管效应量不大，这种反转仍值得注意，因为它表明，在解读合作情境时，稳健提示风格可能与模型的内部先验不匹配。这种模式或许反映了词汇变化与模型语境锚定机制之间一种特有的交互作用，而非消息传递程序本身存在固有的脆弱性。

A.3 受限通信

模型	上下文	均方根误差（无消息）	均方根误差（消息传递）	差异	95%置信区间下限	95%置信区间上限
QWEN2	中性	0.4318	0.2732	**0.1586**†	0.1160	0.4065
	biz	0.3983	0.0861	**0.3122**†	0.2609	0.3639
	环境	0.4481	0.0737	**0.3744**†	0.3325	0.4247
	社会	0.3088	0.0080	**0.3008**†	0.0953	0.3342
	team	0.0470	0.0661	-0.0190	-0.0649	0.0250
	IR	0.4748	0.1310	**0.3438**†	0.3048	0.4681
FALCON3	中性	0.0406	0.0848	-0.0442	-0.0986	0.0070
	biz	0.0229	0.0351	-0.0122	-0.0463	0.0206
	环境	0.0576	0.1201	-0.0625	-0.1044	0.0051
	社会	0.0307	0.0534	-0.0227	-0.0581	0.0113
	team	0.0543	0.0070	0.0473	-0.0329	0.0576
	IR	0.0212	0.0345	-0.0133	-0.0466	0.0152
GRANITE3	中性	0.2002	0.0049	**0.1953**†	0.0575	0.2262
	biz	0.3648	0.0117	**0.3531**†	0.1302	0.3928
	环境	0.3285	0.0983	**0.2302**†	0.1612	0.2864
	社会	0.2901	0.1346	**0.1555**†	0.0981	0.2182
	team	0.1712	0.0126	0.1586	-0.0637	0.1942
	IR	0.3817	0.0267	**0.3550**†	0.2916	0.3950
GEMMA2	中性	0.0531	0.0041	**0.0490**†	0.0136	0.0897
	biz	0.0120	0.0027	**0.0092**†	0.0027	0.0632
	环境	0.0201	0.0256	-0.0054	-0.0289	0.0314
	社会	0.1119	0.0085	**0.1035**†	0.0581	0.1471
	team	0.0525	0.0375	0.0150	-0.0340	0.0539
	IR	0.0142	0.0154	-0.0012	-0.0202	0.0197
多重比较校正						
显著结果过多			13/24 显著（预期为 1.2）	$p < 0.001$		
方向一致性			16/24 为正（66.7%）	$p = 0.076$		

表6：所有模型和上下文中受限消息传递效应对 LOWESS 拟合均方根误差的非参数自助法分析。我们在模拟层级进行自助法重采样（10,000 次迭代），同时保留模拟内智能体依赖关系，以评估无消息传递条件与消息传递条件下的均方根误差差异。这种非参数方法根据差异的经验分布构建置信区间，无需假设数据服从正态分布。正差值表示消息传递处理降低了均方根误差；反之，负值表示消息传递使均方根误差上升。显著效应（95% 置信区间不包含零）以 † 标记。为应对多重检验可能导致的第一类错误膨胀，我们进行了两项二项检验，以总体评估整体模式。首先，我们检验显著结果的数量是否超过零效应零假设下偶然出现的预期值 1.2（alpha = 0.05）。其次，我们检验正差值（消息传递降低均方根误差）的比例是否超过 50%；若消息传递没有系统性效应，预期该比例为 50%。

表 6 展示了通过对实验进行变体测试所得的结果：我们允许大语言模型仅通过分享一个词进行沟通，而非分享完整的短句。本次鲁棒性测试旨在评估限制沟通对稳定性的影响（如果有的话），以及这种影响与短句沟通的影响有何不同。对于大多数模型而言，不同处理条件下的结果在定性上并无差异。值得注意的是，单词通信对 Falcon 3 7b 没有影响，这可能是因为该模型本身具有很高的稳定性。另一个值得注意的发现是，对于其他模型，短句通信和单词通信会导致不稳定性出现统计上可检测的下降，而这两种沟通方式对应的上下文之间存在显著重叠，甚至完全重叠。综合来看，这些发现表明，表达力更强的沟通与受限沟通发挥着类似的作用，但前者对原本已表现出稳定性的模型，以及在某些大语言模型的上下文依赖型决策尤为脆弱的情境中，效果更显著。