跨情境的人格表达：大语言模型智能体中的语言与行为变异

宾 韩，德信权和乔纳森格拉奇

南加州大学

摘要

大语言模型可以通过明确的人格提示加以条件化，但其行为实现往往因情境而异。本研究考察相同的人格提示如何在四种对话情境——破冰、谈判、群体决策和共情任务——中产生不同的语言、行为与情绪结果。结果表明，情境线索会系统性地影响人格表达和情绪基调，说明相同特质会因社会与情感需求的不同而以不同方式表达。这为基于大语言模型的对话智能体提出了一个重要问题：此类变异究竟反映的是不一致性，还是类似于人类行为的情境敏感适应。从整体特质理论的视角来看，这些发现表明，大语言模型展现的是情境敏感而非固定的人格表达，能够灵活适应社会互动目标与情感条件。

关键词

人格提示、整体特质理论、情境感知建模、大语言模型

1. 引言

大语言模型近来已被证明能够支持日益复杂的社会互动形式，包括对情境、情绪和策略性行为的推理 [1, 2, 3, 4]。大语言模型的最新进展使对话智能体能够在互动中展现出鲜明的人格特征 [5, 6]。除提升语言表现外，研究日益聚焦于增强沟通的社会性质量，旨在使智能体显得更像人类、更有吸引力。多项研究表明，人格条件化智能体能够提升用户信任、参与度以及对话满意度 [7, 8]。例如，Ait Baha 等人 [8]开展了一项系统综述，表明人格自适应聊天机器人可显著提升用户满意度和参与度。更新近的工作超越了表层模仿，表明大语言模型能够连贯地理解并复现人格构念。例如，外向型智能体倾向于使用更多积极情绪和社会性词语，而尽责型智能体则偏好结构化与目标导向表达，从而展现出与其所分配特质相一致的语言和情绪表现 [9, 10]。这些影响力超越语言表达，进而塑造决策风格，甚至非言语表达性 [11, 12, 13]。

然而，越来越多的研究发现，大语言模型中的人格表达并非总能在不同情境下保持稳定。即使使用完全相同的角色设定／人格提示，人格也会因情境而呈现出截然不同的面貌 [13, 14]。例如，被指示表现得外向的智能体可能会在闲聊中展现幽默感和表现力，但在谈判中却采取更加中性、以目标为导向的风格 [13]。Reusens 等人 [14] 同样表明，人格对齐的大语言模型在不同任务情境中的人格表达各异；在结构化程度较高的场景中，其一致性有所提升，而在开放式对话中则有所下降。这引出了一个关于基于大语言模型的对话智能体的重要问题：此类变化究竟应被视为缺乏一致性，还是应被视为一种类似人类行为、对情境敏感的自然调节结果?

人格特质（例如大五人格所描述的特质），代表人们在各种情境中表现出来的平均倾向 [15]。然而，心理学研究表明，一个人的实际行为仍会因情境、目标和社会环境而发生变化 [16]。整体特质理论 通过将人格概念化为瞬时状态的分布（反映平均稳定性），并认为其变异源于目标、信念和情感等潜在的社会认知机制，从而整合了这两种观点 [17, 18]。这一视角表明，在大语言模型中观察到的人格表达变异未必意味着前后不一致，也可以理解为反映了人类人格的情境灵活性。类似的计算框架将人格视为动态动机系统和神经系统共同涌现的结果，为这种情境敏感适应提供了理论基础 [19, 20]。

为更好地理解这一现象，我们对基于大语言模型的对话智能体在不同情境下的人格表达进行了系统分析。本研究涵盖四种对话情境——闲聊、谈判、生存规划和共情对话；这些情境在社会目标、合作程度和情绪基调方面各不相同。我们分析了人格表达如何通过语言风格、行为模式和任务结果体现出来。这使我们不仅能够考察人格表达是否会随情境变化，还能考察此类变化在语言与行为维度之间是否呈现一致的对应关系。

2. 研究问题

- **RQ1. 人格的语言表达：**人格表达在不同对话情境（破冰、谈判、生存、同理心）中有何差异?哪些语言与情感线索能够解释这些情境差异?

- **RQ2. 人格的行为表现：**人格驱动的倾向是否会超越语言风格，延伸至让步等任务导向行为，以及一致结果?

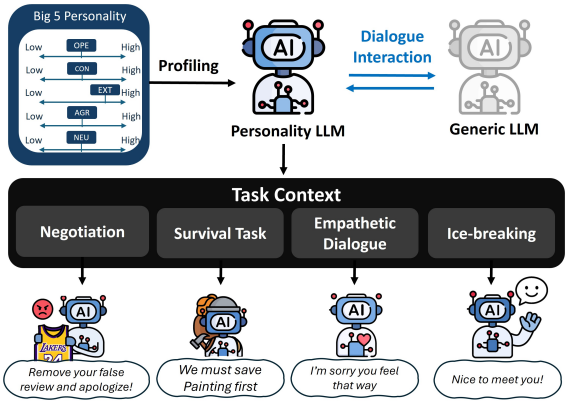


图 1： 考察基于大语言模型的对话智能体中情境依赖的人格表达的实验框架概览

3. 相关工作

3.1. 人格与大语言模型

近期研究日益关注如何引导大语言模型展现特定的人格特质。研究者已通过多种方法实现大语言模型的人格调节，包括人格提示和指令微调 [10, 6]。以往研究还表明，大语言模型能够模仿信任、人格或受情绪驱动的行为等类人特质 [21]。此外，研究者开始应用最初为人类评估而开发的标准化心理测量工具——例如大五人格量表 [15] 和国际人格项目库—神经质、外向性、开放性量表 [22]—— 来考察大语言模型如何理解和表达人格相关构念 [5]。尽管取得了这些进展，现有评估大多聚焦于单轮对话，因此对人格如何在不同对话情境中表现出来，所能提供的见解有限。

3.2. 人类中的情境依赖人格表达

相比之下，人格心理学早已记录到：尽管特质能够预测稳定的平均倾向，这些特质的实际表达却高度依赖于情境。例如，外向型个体在亲和导向的情境中可能表现出更高的社交性，但在谈判等竞争情境中则会表现出更强的果断性，甚至攻击性。尽责性往往在成就目标下增强，而宜人性与支配性则会灵活适应对方的人际风格 [23, 24, 25]。整体特质理论将这种二重性形式化，把特质视为状态的密度分布（平均而言保持稳定），其变异性由目标、情境线索和互动对象系统性地塑造 [17, 18]。特质激活理论 [26] 和人际理论 [23] 等相关框架同样强调，特质并非僵化的规定，而是当情境激活与特质相关的目标或社会图式时所表达的潜能。这一文献表明，人格表达最好被理解为对情境敏感，而非固定不变。然而，大语言模型研究尚未充分纳入这一视角，往往仅在静态或去情境化的设定中评估角色设定条件化。我们的工作旨在弥合这一差距，借鉴心理学理论，系统检验情境如何塑造基于大语言模型的对话中的人格表达

智能体。

4. 实验设计

4.1. 人格条件

我们使用取自先前关于人格条件化对话智能体研究的人格提示来操控大语言模型的人格表达 [6, 21]。每条提示都明确地将单一大五人格特质定义为两个水平——高 和低 ——所用措辞改编自这些先前研究 [6, 10, 27]。表述采用描述性格式，例如：“你是一个高度外向的人：精力充沛、善于交际、健谈且热情” 或“你是一个内向的人：内敛、安静且善于反思”。代表各特质主要层面的形容词同样取自既有的人格词汇表 [10, 28],， 以确保与先前的人格提示研究保持一致。这一操控已在先前使用大五人格量表-10 [15], 的研究中得到验证，证实此类文本提示能够可靠地在大语言模型中引发一致的人格表征。

4.2. 智能体模拟：人格代理与通用代理

所有实验均在智能体–智能体交互环境中进行。其中一个智能体依据指定的角色提示词由人格驱动（称为人格代理），另一个则作为通用代理，提供用于比较的中性基线（不使用人格提示）。由于伙伴的人格也会影响交互动态，我们在所有实验中均使用中性（非个性化）智能体，以使其保持恒定。该配置遵循框架 [13, 21]。

4.3. 主要任务

我们在四种对话情境中评估人格表达，每种情境均经选择，以凸显交互的不同情境特性（见表1）：

- **任务 1：破冰** 该任务旨在引发智能体之间友好、随意且情绪积极的互动。我们采用了个人问题 范式 [7, 30], 中的问题集，该范式已广泛应用于以往的人机交互研究。在此任务中，人格代理回答通用代理提出的三个随意问题（例如，“你空闲时喜欢做什么？”）。该任务并不侧重于解决共同问题，而是以维持自然对话并促进自我披露为中心。整体氛围温暖、非正式且合作，鼓励开放而有吸引力的对话。

- **任务 2：谈判**此任务基于 KODIS数据集 [12], ，该数据集对资源分配场景中的买卖双方纠纷进行建模。人格代理扮演请求退款的买方，而通用代理则扮演要求移除负面评价的卖方。遵循先前的研究 [12], ，我们将买方角色分配给人格代理，因为在谈判过程中，这一方往往表现出更丰富的情绪动态和表达变化。互动涉及多个议题——退款金额、评价移除以及

表1按主导情绪、社会目标 [29], 和合作水平划分的对话情境比较。

Task	情绪	主导社会目标 [29]	合作水平
破冰	喜悦（效价↑，唤醒度 ↑）	**亲和** – 建立融洽关系与自我-表露	中–高
谈判	愤怒（效价↓，唤醒度 ↑）	**权力** – 主张与影响力	低–中
生存任务	中性（效价 –，唤醒度 –）	**成就**– 联合问题解决与协调	High
共情对话	悲伤（效价↓，唤醒度 ↓）	**亲和** – 情感支持与关怀	极高

道歉——因此适合用于观察复杂的行为策略和灵活调整的谈判模式。其情绪基调以愤怒和挫折感为主，特点是紧张程度高、合作程度低，且对话风格正式并以目标为导向。

- **任务 3：生存任务（群体决策）**该任务改编自 Artstein 等人 [31], 最初提出的情境：从博物馆火灾中抢救 15 件艺术品。为了更清晰地聚焦于达成共识的行为，我们将物品数量缩减为五件。每个智能体一开始都有相反的初始排序（人格代理的排序为 A、B、C、D、E，通用代理的排序为 E、D、C、B、A），双方必须为自己的选择给出理由并进行协商，以达成一致。该任务强调协作决策情境，在这一情境中，两个智能体作为一个团队，共同实现共同目标。整体基调平静而富有建设性，正式程度适中，合作度高，情绪积极。
- **任务 4：共情对话**该任务基于共情对话数据集 [32], ，旨在评估智能体生成与人格一致的共情回应的能力。通用代理会提出一段情绪强烈的话语（例如：“资格认证考试又没通过，我觉得自己一直很迷茫。”），而人格代理必须以情感理解和支持性表达作出回应。对话情境在情感上较为敏感且私人，具有较高的情感参与度、较低的正式程度和支持性氛围。

4.4. 评估

4.4.1. 语言测量

为捕捉与人格表达相关的语言层面线索，我们采用：

- **语言探索与字数统计（LIWC）：**我们使用广泛采用的心理语言学文本分析工具——语言探索与字数统计（LIWC） [33], ，考察不同人格条件之间的词汇和文体差异。LIWC 提供无上下文语言分析，仅基于词语使用情况捕捉词汇和文体模式，而不纳入对话语境。
- **人格预测（预训练分类器）：**为定量评估特质表达，我们采用了一个预训练的大五人格分类器 [34] ，该分类器整合了来自 BERT 的上下文化嵌入与心理语言学特征。该模型预测某个话语反映特定特质的可能性（例如，若被分类为外向，则为外向性 = 1；否则为 0）。类似地，该分类器仅依据语言特征，对人格进行不考虑语境的估计。
- **基于大语言模型的人格评估：**我们进一步通过提示大语言模型以人格心理学家的身份开展评估，根据对话语境 [10], 对每段对话中的大五人格特质进行评估（1—5 分）。评估提示为每项特质及各分数等级设定了明确标准，说明高水平与低水平表达所对应的行为表现。提示中还明确提供了任务语境，以便进行依赖语境的评估。采用第三人称标注方法 [35], 后，模型为每段对话生成数值评分和简明的评估理由。

4.4.2. 情绪测量

我们分析了互动过程中产生的情绪。LIWC 能够在词汇层面捕捉情感词使用，但无法充分反映对话轮次间传达的整体情绪基调或强度。为了补充此类表层测量，我们使用基于大语言模型的情绪识别方法，明确评估各智能体表达的情感基调 [21, 36, 37]。鉴于近期研究已证明大语言模型在情绪推理和情感理解方面具有很强的能力，我们采用这一方法，对各智能体表达的情感进行全面评估。模型综合话语的语言内容和风格基调，推断说话者的整体情感状态，并以互补的情感维度（效价和唤醒度 [38]）呈现输出。

4.4.3. 行为（决策）测量

在以任务为导向的情境中，例如**谈判**和**生存（拯救艺术品）**，我们分析了合作的两个行为指标：（1）两个智能体最终是否达成一致（一致率）；（2）各智能体在多大程度上会根据对方的行为调整自己的决策（让步）。

- **谈判（退款报价）**：在谈判任务中，让步被定义为初始退款提议的减少，量化为

$$\text{Concession} = 100\% - \text{Refund Offer}.$$

该指标衡量了代理在多轮对话中，从最初的 100% 报价向合作方需求方移动了多少。将该数值绘制

特质	LIWC 特征	任务 1		任务 2		任务 3		任务 4		报告的
		High	Low	High	Low	High	Low	High	Low	([40])
开放性	句子词数	**18.0**	14.0	**15.7**	14.5	**16.3**	14.3	**17.2**	14.7	高 > 低
	休闲	**1.2**	1.0	**1.1**	1.0	**1.1**	1.0	**1.1**	1.0	高 > 低
尽责性	咒骂	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	高 < 低
	愤怒	1.0	**1.3**	1.0	**1.2**	1.1	**1.2**	1.0	**1.2**	高 < 低
	消极情绪	2.9	**3.9**	3.0	**3.8**	3.0	**3.6**	2.9	**3.6**	高 < 低
	生物过程	1.2	**1.3**	1.1	**1.3**	1.2	**1.3**	1.2	**1.3**	高 < 低
	代词	18.2	**20.8**	18.0	**20.5**	18.0	**20.5**	18.0	**20.3**	高 < 低
	认知过程	**16.1**	14.8	**15.8**	13.9	**15.9**	14.0	**15.8**	14.0	高 > 低
	不确定性	1.3	**1.8**	1.3	**1.8**	1.3	**1.7**	1.3	**1.7**	高 < 低
外向性	句子词数	**17.6**	14.2	**15.5**	14.0	**16.0**	14.0	**16.8**	14.5	高 > 低
	性相关	**1.0**	0.9	**1.0**	0.9	**1.0**	0.9	**1.0**	0.9	高 > 低
宜人性	咒骂	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	0.0	**1.0**	高 < 低
	愤怒	1.0	**1.3**	1.0	**1.3**	1.0	**1.3**	1.0	**1.2**	高 < 低
	积极情绪	**10.3**	4.6	**10.3**	4.6	**10.4**	4.7	**10.3**	4.7	高 > 低
神经质	消极情绪	2.5	**4.3**	2.5	**4.1**	2.5	**4.2**	2.5	**4.1**	高 < 低
	I	5.2	**9.2**	5.2	**8.8**	5.1	**8.7**	5.1	**8.6**	高 < 低
	代词	16.0	**23.1**	16.9	**22.7**	16.3	**22.3**	16.1	**22.2**	高 < 低

表 2四种对话情境中高人格与低人格的 LIWC 特征均值（任务 1：破冰，2：谈判，3：生存，4：共情对话）。粗体数字表示每对中较高的均值，显示出与先前研究结果一致的方向性趋势 [40]。

于各轮次之上，即可得到一条让步曲线，它体现了整个谈判过程中的妥协轨迹。

- **生存（秩差和；SRD）：**在生存任务中，第 t 轮的让步被衡量为当前排名与初始排名之间的差值。具体而言，我们计算了秩差和（*SRD*） [39] 每一轮 t 的值，公式为

$$\mathrm{SRD}_t = \sum_{i=1}^{5} \left| r_i^{\text{initial}} - r_i^{(t)} \right|,$$

其中，r_i^{initial}和 $r_i^{(t)}$ 分别表示人格代理在初始轮次和当前轮次为项目 i 提供的排名。由于所有智能体都以相同的初始顺序开始，SRD 值从 0 开始，并随着排名偏离基线而增加。因此，较高的 SRD_t 表示该轮次相较于初始决策状态的偏离程度更大（即相对于基线做出了更大的让步）。

5. 结果：跨情境的人格

我们通过比较不同对话情境下的高人格条件和低人格条件，分析人格表达。以下小节将分析人格表达如何与对话情境相互作用。

5.1. 语言探索与字数统计（LIWC）

在众多 LIWC 特征中，我们仅选取了那些在以往荟萃分析中被证明与大五人格特质显著相关的特征 [40]。表 2 总结了不同任务中高组与低组之间的平均差异。观察到的方向性趋势均遵循先前研究结果所报告的模式 [40]。

5.2. 预训练人格预测

图2 显示了预训练人格分类器 [34] 在四种对话情境中将话语预测为体现各项大五人格特质（高或低）的比例。预测的总体方向与预期的人格提示高度一致，表明模型捕捉到了智能体语言中蕴含的特质差异。尽管总体趋势与设计的操纵一致，但也观察到了一些偏差，这反映出依赖忽略语境的语言测量所面临的挑战。在破冰任务中，特质区分最为明显，因为富有表现力且注重亲和的语言有助于更清晰地区分特质；而谈判和生存任务呈现出大体相似的模式，由于其目标导向性更强，特质差异较不明显。

5.3. 基于大语言模型的人格评估

图3 展示了四种对话情境下高条件和低条件的基于大语言模型的人格评估得分。总体而言，结果与预期的人格操控高度一致，在大多数特质上均显示出显著差异（$p < .001$）。高条件下的智能体在相应特质上的得分始终更高，表明大语言模型评估器能够通过对话中的语言行为，有效识别出所设计的人格模式。在不同情境中，破冰和生存任务所呈现的人格差异最为明显，因为互动中合作且富有情绪表达的特性促进了更清晰的特质表达。相比之下，谈判和共情任务仅表现出较弱或部分差异，这可能是由情绪敏感性等情境因素导致的。

5.4. 不同对话情境中的情绪基调

为考察不同任务中的情感变化，我们比较了从每段对话中提取的效价—唤醒度分布（图4）。四种情境呈现出截然不同的模式。破冰阶段总体表现为正向效价和高唤醒度，反映出交流生动且引人入胜。谈判阶段表现为负向效价和中等唤醒度，表明存在紧张情绪和目标冲突。生存任务呈现出喜忧参半的情绪，效价和唤醒度均处于中等水平，这与合作和竞争动态并存的情况相符。最后，共情情境

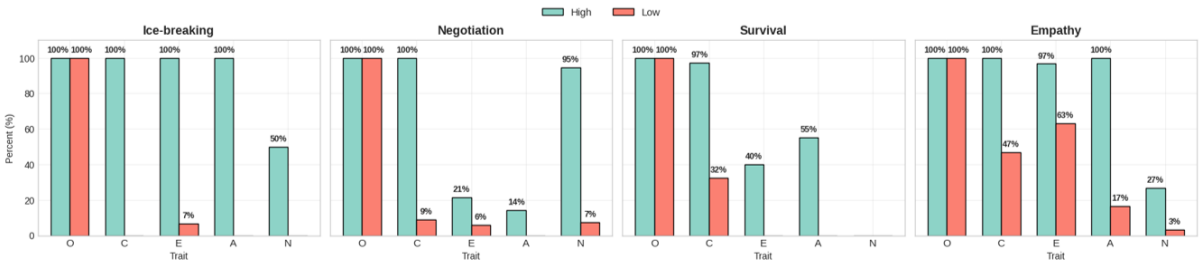


图 2： 基于 BERT 的预训练人格预测模型结果 [34]

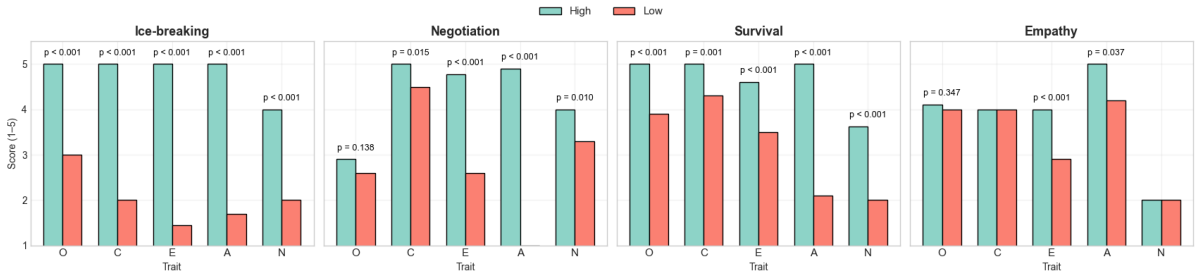


图 3： 基于大语言模型的人格评估结果

呈现出正向效价和低唤醒度，表明互动平和且具有情绪支持性。

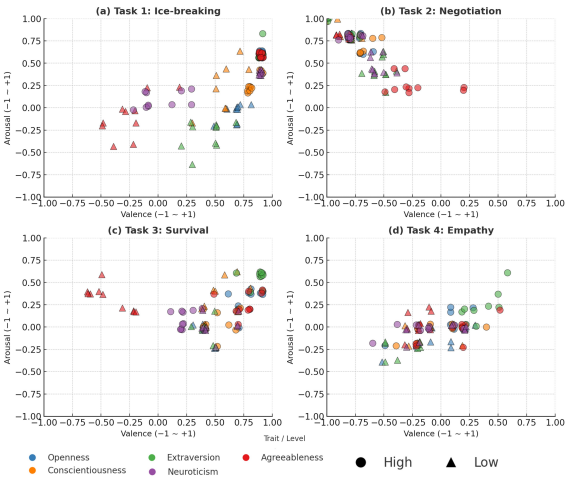


图 4： 四种对话情境中的效价—唤醒度分布

基于大语言模型的人格评估将分析延伸至行为结果，直观展示人格表达如何随对话情境而变化（图 5）。具体而言，在破冰和生存等合作情境中，特质差异最为显著；而在谈判和共情等竞争性或情绪敏感情境中，差异则有所减弱。

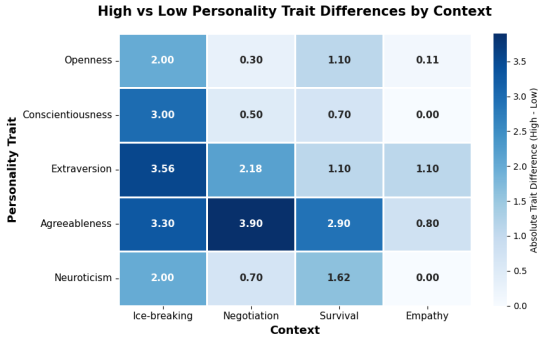


图 5： 不同对话情境下基于大语言模型的人格差异。数值越高，表示特质差异越显著（高 vs. 低）。

5.5. 总结：从语言表达到行为与情绪表达（RQ1）

在各项分析中，与人格相关的差异始终显现，但可比性和表达程度各不相同。基于 LIWC 的语言分析揭示了特质内部的差异（高与低），但由于每个人格维度依赖不同的语言特征，无法直接跨特质比较。相比之下，预训练模型和基于大语言模型的评估均在共享的表征空间中进行，因此能够跨特质、跨情境比较。预训练模型捕捉到了人格维度之间的概念差异，而大语言模型——

最后，效价—唤醒度分析进一步补充了这些发现，展现了这些互动背后的情感基调。人格差异更明显的情境也呈现出更广泛的情绪变化和更高的平均效价，这表明积极且投入的情感与更清晰的人格表达同时出现。这些结果表明，基于大语言模型的智能体的人格会在语言、表征、行为和情感等层面体现出来，而情境会调节各个维度的表达强度。

(a) 谈判任务中的一致率（%）

人格维度	高（%）	低（%）
开放性	0.0	0.0
尽责性	10.0	0.0
外向性	10.0	20.0
宜人性	20.0	0.0
神经质	0.0	10.0

(b) 生存任务中的一致率 (%)

人格维度	高 (%)	低（%）
开放性	90.0	50.0
尽责性	50.0	60.0
外向性	80.0	40.0
宜人性	90.0	70.0
神经质	90.0	30.0

表 3两个任务中不同人格维度的一致率。

6. 结果：人格的行为表现（RQ2）

本节探讨研究问题 2：由人格驱动的倾向是否会超越语言风格，延伸至让步和一致结果等任务导向行为？我们考察对话表达中观察到的人格驱动差异是否也会影响智能体在任务导向互动中的合作方式及决策调整。为此，我们分析两个行为维度：一致，即智能体是否达成共同共识；以及让步，即各智能体在交互轮次中调整决策的程度。我们还在两种情境下进一步分析让步：谈判（退款报价）和生存（秩差和；SRD）。

6.1. 一致率（%）

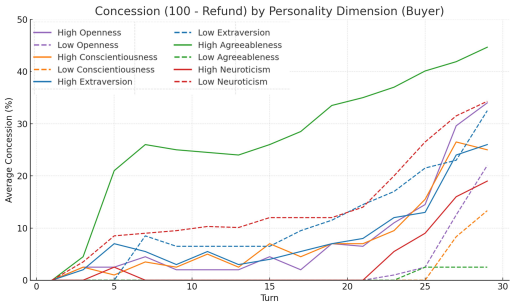
如表 3所示，协商任务的总体一致率相对较低，这反映了该任务本身具有冲突驱动和竞争性的特点。由于该任务要求解决争议，而非达成共识，智能体往往无法达成一致。不过，高宜人性的智能体一致率达到 20%，高于低外向性或低神经质的智能体（两者均约为 10%）。

相比之下，生存任务（表 3）营造了高度协作的环境，整体一致率也较高。具体而言，高宜人性的智能体在 90% 的情况下达成共识，而低宜人性的智能体为 70%；高外向性的智能体一致率达到 80%，而低外向性的智能体为 40%。

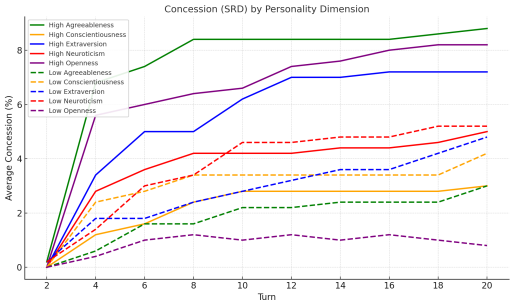
6.2. 让步

6.2.1. 谈判（退款报价）

在谈判任务中，依据初始退款提议的降幅来评估让步行为，计算方式为 (100-退款)。如图 6a所示，宜人性高的智能体表现出最明显的让步模式，在最后几轮中逐渐将让步幅度提高至 40% 以上。相比之下，宜人性低的智能体和神经质高的智能体变化甚微，在大部分对话中让步幅度都维持在 10% 以下。开放性和外向性呈现出中等程度的让步曲线，到对话末尾时趋于 20–25%。总体而言，结果表明，宜人且情绪稳定的智能体在谈判过程中更愿意妥协并调整立场，而好斗或神经质程度很高的智能体则更加固执，对伴侣的要求回应较少。这一行为倾向与前文报告的一致结果相符（表 3），其中宜人性高的智能体也取得了更高的成功解决率。



（a）不同买方人格在各谈判轮次中的退款报价趋势。



（b）不同人格维度在对话各轮次中的 SRD 变化。

图 6： 不同人格对应的行为结果：（a）谈判中的退款让步；（b）生存中的基于 SRD 的适应。

6.2.2. 生存（秩差和；SRD）

在生存任务中，通过各决策轮次的秩差和（SRD）来衡量让步。如图所示：6b，宜人性和开放性较高的智能体，其 SRD 值持续上升，在后期轮次达到约 6 至 7 分，是相应特质较低者（低于 3 分）的两倍多。外向型智能体表现出中等程度的适应，而神经质较高的智能体在互动后期则呈现不稳定、起伏不定的变化。这些结果表明，开放且合群的人格在合作情境中表现出更大的灵活性，并能更动态地调整决策。

7. 讨论

本研究考察了基于大语言模型的智能体如何在语言模式之外，通过社交

行为和情绪反应展现人格。在语言层面，LIWC 分析捕捉到了特质内部的差异（高与低），但对跨特质或跨情境的变化反映有限。这一局限可能源于 LIWC 针对每种特质采用不同的特征集，导致难以直接比较不同维度。相比之下，预训练模型和基于大语言模型的评估在共享的表征空间中进行，因此能够支持跨特质和情境敏感的解读。结果表明，合作情境（破冰和生存）会增强外向性和宜人性，而竞争情境（谈判）则会加剧神经质倾向，并引发更多紧张且以冲突为导向的互动。在共情任务中，情绪调节受到强调，因此整体语言语气更加稳定、克制。

这些发现与整体特质理论的原则一致。该理论认为，人格通常是稳定的，但会根据社会目标和情境线索动态激活。同样，研究中观察到的大语言模型人格表达，并非对提示中特质的固定复现，而是受任务要求和情绪语境塑造的适应性、状态层面调整。换言之，大语言模型中的人格并不是静态的文本产物，而是由互动语境塑造的、适应社会的反应。最后，情绪分析也为这一模式提供了定量支持。外向性和宜人性高的智能体表现出更高的效价和唤醒度，这与其合作性的语言和行为倾向一致；而神经质高的智能体则表现出较低的效价，反映出其在竞争性环境中的紧张与退缩。总体而言，这些结果表明，语言、行为和情绪表达朝着一致的方向运作——说明大语言模型中的人格并非仅仅是脚本化构造，而是受语境调节并以情感为基础的表达。

8. 结论

本研究考察了人格条件化大语言模型智能体如何在不同对话情境中调整其表达行为。即使获得相同的人格提示，它们的语言和行为模式仍会根据各项任务的社会目标呈现系统性差异。情绪分析进一步显示，这些变化伴随着情感基调的持续改变，表明人格表达与情境要求之间存在适应性协调。这些发现回应了基于大语言模型的对话智能体领域的一个重要问题：这种变异究竟反映了不一致，还是类似人类行为的情境敏感型适应。我们的结果支持后一种解释：观察到的变化并非随机波动，而是针对互动目标和情感条件所做的连贯调整。然而，仍需进一步研究，以确定这些情境敏感型变化是否像人类人格的运作方式一样具有功能上的适应性。

虽然加入通用代理为比较提供了基准，但它的存在仍可能影响互动动态。未来研究将探讨不同人格的智能体之间的互动，以考察人格—人格动态，并直接检验大语言模型是否内化了整体特质理论提出的情境依赖型激活机制。我们还计划将这一框架拓展至更广泛的社会情境，并将大语言模型生成的行为与人类对话进行比较——对话数据，以提升现实世界中的有效性。

参考文献

[1] S. Thapa、S. Shiwakoti、S. B. Shah、S. Adhikari、H. Veera- mani、M. Nasim、U. Naseem，《计算社会科学中的大语言模型：前景、现状与挑战》，《社会网络分析与挖掘》15（2025）1–30。[2] G. Lee、Y. Yang、J. Healey、D. Manocha，《自你离去之后：利用上下文感知转录重新参与沉浸式虚拟现实会议》，载于：2025年计算机系统人因国际会议论文集，2025，第1–20页。[3] A. N. Tak、J. Gratch，《GPT是情绪计算模型吗？》，载于：2023年第11届国际情感计算与智能交互会议（ACII），IEEE，2023，第1–8页。[4] D. Kwon、E. Weiss、T. Kulshrestha、K. Chawla、G. Lucas、J. Gratch，《大语言模型是有效的谈判者吗？对大语言模型在谈判对话中多方面能力的系统评估》，载于：计算语言学协会自然语言处理实证方法会议：EMNLP 2024论文集，2024，第5391–5413页。[5] G. Jiang、M. Xu、S.-C. Zhu、W. Han、C. Zhang、Y. Zhu，《评估并诱导预训练语言模型中的人格》，《神经信息处理系统进展》36（2023）10622–10643。[6] G. Serapio-García、M. Safdari、C. Crepy、L. Sun、S. Fitz、M. Abdulhai、A. Faust、M. Matarić，《大语言模型中的人格特质》（2023）。[7] K. M. Lee、W. Peng、S.-A. Jin、C. Yan，《机器人能表现出人格吗？：对人机交互中的人格识别、社会反应和社会临场感的实证检验》，《传播学杂志》56（2006）754–772。[8] T. Ait Baha、M. El Hajji、Y. Es-Saady、H. Fadili，《个性化的力量：人格自适应聊天机器人的系统综述》，《SN计算机科学》4（2023）661。[9] A. Banayeeanzade、A. N. Tak、F. Bahrani、A. Bolourani、L. Blas、E. Ferrara、J. Gratch、S. P. Karimireddy，《大语言模型中的心理引导：有效性与可信度评估》，arXiv预印本 arXiv:2510.04484（2025）。[10] H. Jiang、X. Zhang、X. Cao、C. Breazeal、D. Roy、J. Kab- bara，《Personallm：研究大语言模型表达人格特质的能力》，arXiv预印本 arXiv:2305.02547（2023）。[11] Y. J. Huang、R. Hadfi，《人格特质如何影响谈判结果？基于大语言模型的模拟》，arXiv预印本 arXiv:2407.11549（2024）。[12] J. Hale、S. Rakshit、K. Chawla、J. M. Brett、J. Gratch，《Kodis：多文化争端解决对话语料库》，arXiv预印本 arXiv:2504.12723（2025）。[13] B. Han、D. Kwon、S. Lin、K. Shrestha、J. Gratch，《大语言模型能否基于人格特质为具身虚拟智能体生成行为？》，载于：第25届ACM智能虚拟智能体国际会议论文集，2025，第1–10页。[14] M. Reusens、B. Baesens、D. Jurgens，《经济学家是否总是更加内向？分析人

格设定型大语言模型中的一致性》，arXiv预印本 arXiv:2506.02659（2025）。[15] O. P. John、S. Srivastava等，《五大人格特质分类：历史、测量与理论视角》（1999）。[16] W. Mischel、Y. Shoda，《人格的认知—情感系统理论：重新概念化人格结构中的情境、倾向、动态与不变性》，《心理学评论》102（1995）246。[17] W. Fleeson，《迈向人格的结构与过程整合观点：作为状态密度分布的人格特质》，《人格与社会心理学杂志》80（2001）1011。[18] W. Fleeson、E. Jayawickreme，《整体特质理论》，《人格研究杂志》56（2015）82–92。[19] S. J. Read、L. C. Miller，《人格结构与动态的神经网络模型》，载于：《测量与建模：个体与情境》，爱思唯尔，2021，第499–538页。[20] S. J. Read、L. C. Miller，《虚拟人格模型：理解人格障碍的动态结构化动机系统框架》，《精神病理学与临床科学杂志》（2025）。[21] D. Kwon、K. Shrestha、B. Han、E. H. Lee、G. Lucas，《冲突对话中的行为对齐评估：大语言模型智能体与人类的多维比较》，载于：2025年自然语言处理实证方法会议论文集，2025，第16377–16391页。[22] L. R. Goldberg等，《一种宽带宽、公共领域的人格量表：测量多个五因素模型的低阶侧面》，《欧洲人格心理学》7（1999）7–28。[23] D. J. Kiesler，《当代人际理论与研究：人格、精神病理学与心理治疗》，约翰·威立父子公司，1996。[24] D. S. Moskowitz，《跨情境普遍性与人际环形模型》，《人格与社会心理学杂志》66（1994）921。[25] K. O. McCabe、W. Fleeson，《特质有用吗？将特质表现解释为目标追求中的工具》，《人格与社会心理学杂志》110（2016）287。[26] R. P. Tett、D. D. Burnett，《基于人格特质的工作绩效交互作用模型》，《应用心理学杂志》88（2003）500。[27] S. Noh、H.-C. H. Chang，《具有不同人格的大语言模型在多议题谈判游戏中的表现》，arXiv预印本 arXiv:2405.05248（2024）。[28] P. T. Costa Jr、R. R. McCrae，《领域与侧面：使用修订版NEO人格量表进行层级式人格评估》，《人格评估杂志》64（1995）21–50。[29] D. C. McClelland，《人类动机》，剑桥大学出版社档案，1987。[30] A. Aron、E. Melinat、E. N. Aron、R. D. Vallone、R. J. Bator，《人际亲密感的实验生成：一项程序及若干初步发现》，《人格与社会心理学通报》23（1997）363–377。[31] R. Artstein、D. R. Traum、J. Boberg、A. Gainer、J. Gratch、E. Johnson、A. Leuski、M. Nakano，《倾听我的身体：交朋友有助于影响他人吗？》，载于：FLAIRS，2017，第430–435页。

[32] H. Rashkin, E. M. Smith, M. Li, Y.-L. Boureau，《迈向富有同理心的开放领域对话模型：一项新基准与数据集》，arXiv 预印本 arXiv:1811.00207（2018年）。[33] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn，《LIWC2015 的开发与心理测量特性》（2015）。[34] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, E. Cambria，《使用基于 BERT 词嵌入集成的装袋支持向量机检测人格特质》，arXiv 预印本 arXiv:2010.01309 （ 2020 年）。[35] Y. J. Huang, R. Hadfi，《超越自我报告：用于大语言模型人格评估的多观察者智能体》，arXiv 预印本 arXiv:2504.08399（2025）。[36] B. Han, C. Yau, S. Lei, J. Gratch，《使用大语言模型进行基于知识的情绪识别》，载于：2024年第12届国际情感计算与智能交互会议（ACII），IEEE，2024，第1–9页。[37] A. N. Tak, J. Gratch，《GPT-4 从第三人称视角模拟普通人的情绪认知》，载于：2024年第12届国际情感计算与智能交互会议（ACII），IEEE，2024，第337–345页。[38] J. A. Russell，《情感环形模型》，《人格与社会心理学杂志》39（1980年）1161。[39] K. Héberger，《排名差值和能够公平比较方法或模型》，《TrAC分析化学趋势》29（2010年）101–109。[40] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. Van Breda, S. Ghassemi, R. E. de Vries，《基于文本的人格评估中的真相内核：大五人格与语言探索与字数统计（LIWC）之间关系的荟萃分析》，《心理学公报》148（2022年）843。

A. 人格提示

A.1. 人格调节提示

大语言模型提示片段

外向性 - 高外向性：外向、善于交际、精力充沛、健谈、果断自信、热情 - 低外向性：内敛、安静、独处、被动、退缩、情绪低落

宜人性 - 高宜人性：友善、合作、富有同情心、热情、信任他人、富有同理心 - 低宜人性：挑剔、好争辩、苛刻、冷漠、多疑、充满敌意

尽责性 - 高尽责性：有条理、可靠、自律、负责、高效、周全 - 低尽责性：粗心、杂乱无章、玩忽职守、懒惰、不可靠

神经质 - 高神经质：焦虑、喜怒无常、缺乏安全感、拘谨、脆弱、容易感到压力 - 低神经质：冷静、放松、有韧性、自信、有安全感、沉稳

开放性 - 高开放性：富有创造力、好奇、富有想象力、具有求知欲、勇于冒险、思想开放- 低开放性：僵化、务实、心胸狭窄、缺乏想象力、墨守成规

包含大五人格的大语言模型提示片段

A.2. 人格评估提示

大语言模型提示片段（人格评估——谈判语境）

你是一名心理学家，根据对话中观察到的说话者的语言和行为线索，分析其人格特质。

语境： 对话双方是买方和卖方，正在处理一场由误解引发的冲突。说话者要求退款并解决该问题。

根据以下内容评估说话者的人格：
大五人格维度（1–5 分制）：
外向性— 1 =内敛、安静 | 5 = 善于交际 ，健谈 ，果断自信
宜人性— 1 = 苛刻、挑剔 | 5 =友善，合作、富有同理心
尽责性 — 1 =粗心 | 5 =自律-plined、可靠、高效
神经质 — 1 = 冷静、放松 | 5 = 焦虑，情绪化、缺乏安全感
开放性 — 1 = 刻板、缺乏想象力 | 5 = 有创造力、好奇、思想开放
分析以下话语，并推断
说话者的人格水平。

包含谈判情境的大语言模型提示片段