智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

大语言模型人格化表达实现技术综述

柴春雷, 葛智超, 殷敏, 王政, 连博艺, 涂逍洋

引用本文:
柴春雷, 葛智超, 殷敏, 等. 大语言模型人格化表达实现技术综述[J]. 智能系统学报, 2026, 21(2): 321-336.
CHAI Chunlei, GE Zhichao, YIN Min, et al. A survey of techniques for realizing personality expression in large language models[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(2): 321-336.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505031

DOI: 10.11992/tis.202505031
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20251126.1603.002

大语言模型人格化表达实现技术综述

柴春雷[1,2], 葛智超[1], 殷敏[2], 王政[3], 连博艺[1], 涂逍洋[1]
(1. 浙江大学 计算机辅助设计与图形系统全国重点实验室, 浙江 杭州 310027; 2. 浙江大学 长三角智慧绿洲创新中心, 浙江 嘉兴 314100; 3. 浙江大学 管理学院, 浙江 杭州 310058)

摘 要: 本文对大语言模型 (large language models, LLMs) 人格化表达的实现技术进行了系统性综述。文章首先回顾了早期模型通过词向量和句嵌入等拟人化文本表达的发展历程。随着大语言模型的普及和技术成熟, 模型已能够根据不同场景和任务需求, 呈现出具有特定角色设定的人格特征。当前实现大语言模型人格化表达的技术路径主要包括 3 个层面: 模型内部的预训练数据优化、微调, 模型外部的提示词、强化学习、智能体工作流设计, 以及从评估层面对人格表达进行对齐。这些技术既可通过调整模型内部参数, 实现人格化定制, 也可在不修改核心参数的前提下, 通过外部机制实现人格化表达的灵活调控。最后, 基于当前模型人格化生成技术的发展现状, 本文对该领域的技术发展趋势和应用前景进行了分析与展望。

关键词: 大语言模型; 大语言模型人格化表达; 人格评估; 人格生成; 数字人格; 自然语言处理; 角色扮演; 人格对齐
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2026)02−0321−16

中文引用格式: 柴春雷, 葛智超, 殷敏, 等. 大语言模型人格化表达实现技术综述 [J]. 智能系统学报, 2026, 21(2): 321–336.
英文引用格式: CHAI Chunlei, GE Zhichao, YIN Min, et al. A survey of techniques for realizing personality expression in large language models[J]. CAAI transactions on intelligent systems, 2026, 21(2): 321–336.

A survey of techniques for realizing personality expression in large language models

CHAI Chunlei[1,2], GE Zhichao[1], YIN Min[2], WANG Zheng[3], LIAN Boyi[1], TU Xiaoyang[1]
(1. State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou 310027, China; 2. Yangtze River Delta Smart Oasis Innovation Center, Zhejiang University, Jiaxing 314100, China; 3. School of Management, Zhejiang University, Hangzhou 310058, China)

Abstract: Personality expression in Large Language Models (LLMs) has emerged as a key direction in human–computer interaction research. Enabling machines to exhibit uniquely human-like expressiveness remains a significant challenge in the LLM domain. In recent years, the application of pre-training, fine-tuning, and agent-based collaboration techniques has matured to the point where LLMs can adopt role-specific personalities tailored to diverse scenario and task requirements. Whether through carefully curated training datasets at the input stage, parameter-level adaptations within the model, or external agents and workflows around the model, personality expression in LLMs can be achieved. This paper provides a comprehensive review of the current state and future trends of personality expression in LLMs. On one hand, it examines techniques for extracting and simulating personality traits; on the other, it explores methods for controlling and aligning model personality. By analyzing and summarizing these approaches, we discuss the developmental directions of personality-driven LLM research.

Keywords: large language models; personality expression in large language models; personality assessment; personality generation; digital persona; natural language processing; role-playing; personality alignment

近年来, 大语言模型 (large language models, LLMs) 技术经历了从基础的文本分类、情感计算识别、响应检索到复杂的自然交互语言模型的发展, 其应用领域也不断拓展与深化。语境决定了我们怎样理解词汇, 而情感则是解码这一理解过程的隐形钥匙。在模型与用户交互日益频繁的背

收稿日期: 2025−05−31. **网络出版日期**: 2025−11−27.
通信作者: 殷敏. E-mail: amyin@zju.edu.cn.

景下，如何能让机器 (大语言模型、智能体) 理解人类语言表达中的语境、情绪情感，并在用户交互中实现人格化的表达，是近年来亟待解决的研究命题。为此，研究者们不断探索：何种大语言模型的调试技术能更精准地“读懂”人类？如何在表达上体现人格化特征，实现“千人千面”的个性化判别与反馈。在这一趋势下，大语言模型个性化技术获得广泛关注，并显示出迅猛的发展势头。2020 年 BERT(bidirectional encoder representations from Transformers)、GPT(generative pre-trained Transformer) 等语言模型问世后，有关大语言模型个性化表达的文献不足百篇，而到了 2024 年，相关文献数量已突破两百篇。在这样的背景下，大语言模型人格化的实现技术得到了越来越多的研究。

大语言模型人格化表达的任务，通常是探究大语言模型如何通过模拟人类的语言风格、情绪，甚至特定人物的知识结构和思维模式，生成具有独特“人格”特征的文本或交互内容，这一话题需要涉及不同领域的知识贡献。例如语言学中的语域分析 (register analysis)，可以帮助模型理解不同社交情境下的语言习惯；文体学 (stylistics) 则用于研究作者的个人语言特征 (如用词偏好、句法结构等)，这些都是构建特定“人格”语言的基础。心理学中的人格特质和人格理论，提供了理论和参数指标依据。计算机科学与机器学习领域，提供了情感分析、文本风格迁移 (style transfer) 和对话系统等技术，让模型能学习并模仿特定风格。认知科学与哲学中对于“意识、意图和身份”，人类思维过程、知识组织和信念系统的成果，让模型训练不只是语言的模仿，更能体现思维过程。

发表在 2017 年 1 月—2025 年 1 月间，聚焦于运用 LLM 与个性化、人格相关研究的论文有 851 篇，我们最终筛选出的 102 篇，完成对大语言模型人格化表达实现技术的系统综述。本研究不仅探讨非人类智慧 (包括大语言模型、智能体及具身智能等多种形式) 的人格化内在机制，更希望推动大语言模型人格化领域的理论与实践创新。借此，相关研究将不拘泥于依赖人类传统的人格量表，不局限于为迎合某种预设“人设”而进行机械式的对齐。我们期望，通过赋予大语言模型主动表达情感的能力，推动其从简单模仿人类性格，向真正理解并表达自身人格的转变，从而实现更为丰富、多维的智能交互体验。

1　大语言模型人格化生成技术

已有的研究中，大语言模型人格化表达的实现技术，按“人格信息在大语言模型生命周期中被“嵌入、控制和验证的环节/层级”，可以分为 3 类。换言之，它聚焦于人格控制点在整个技术流水线中的位置，先从外部数据侧切入，再到模型内部参数侧，最后到输出评估对齐侧。具体可理解为“输入−模型−输出”，或者是“数据驱动−参数驱动−对齐驱动”3 层视角 (表 1)。

表 1　大语言模型人格化的实现技术分类

Table 1　Taxonomy of personality realization techniques in large language models

分类	关注的核心环节	人格信息的主要注入 (约束途径)	对应技术关键词
基于用户内容的 分类学习与模拟	输入层 (数据侧)	从用户历史文本、对话风格、语用特征中提取人格特征信号，在生成时进行条件化或风格转移	个性化检索增强、 风格迁移
模型调控 技术路线	模型内部 (参数侧)	通过精细微调、提示词工程、强化学习等手段，直接调整权重或推理策略，使人格倾向内化	微调、指令调优、策略梯度、人设嵌入向量
基准对齐 范式	输出层 (评估−对齐侧)	建立人格量表或评价基准(大五人格等)，以可量化指标持续校准输出；可配合拒答和重写等策略	人格对齐基准、自动/人工打分、对齐优化循环

1) 基于用户内容的分类学习与模拟技术，通过对用户输入内容进行精细化分析和特征提取，实现个性化文本生成与互动体验。此方法注重捕捉用户独特表达模式，从而在响应中呈现出与用户相融的个性特征。

2) 模型的调控技术路线，包括精细微调、提示词工程和强化学习等方法，致力于在模型架构或推理过程中植入特定人格倾向。这类方法直接作用于模型的参数空间或决策机制，使模型能够在输出阶段自然展现目标人格特质。

3) 基准对齐范式，通过建立规范化的人格评估体系与标准，引导模型生成符合特定人格画像的输出内容。这种方法强调人格表达的可测量性与一致性，通过持续优化对齐策略，实现大语言模型的个性化构建与表达。

2　大语言模型人格化的实现方法

唐纳德·诺曼在《设计心理学》指出，人格化

设计能够显著提升人机交互过程中的用户体验。对话智能体最初主要利用简单的语言模式匹配和转换规则, 来实现人格化特征的表达 (例如早期的 ELIZA, SHRDLU)[1]。尽管技术基础十分简单, 这种设计能成功地在用户心中建立起一种“机器具有人格和理解能力”的用户体验。随着自然语言处理和深度学习技术的发展, 曾经基于规则的简单设计, 如今已广泛扩展到智能体的对话生成等多个关键过程之中, 借由模型的能力, 对话智能体能够对情感、人格特质这些人类特性进行表征和建模, 从而更加精准地捕捉和理解用户话语背后的复杂含义、心理状态以及潜在需求。近年来, 大语言模型及其应用 GPT、DeepSeek、Claude 等, 使得智能体能够在各类复杂场景下进行自然语言交互, 重塑了人格化设计的理论框架与实践边界。本综述从人格特征提取与模型预测、大语言模型的调控技术和人格评估与对齐 3 个方面, 对大语言模型人格化表达的实现方法进行综述。

2.1　人格特征提取与预测模型

通过深入分析用户的历史对话、情感表达模式、社交媒体行为等多维度数据, 现代系统能够精确提取和学习个体的性格特征 (如大五人格、Myers Briggs Type Indicator、Dominance Influence Steadiness Conscientiousness 行为风格等), 并在此认知基础上模拟用户的语言风格、决策倾向与个人偏好, 从而实现更为自然、流畅且富有针对性的交互体验与服务提供[2-4]。较为早期的大语言模型人格特征识别和表达, 主要通过文本中的语义线索进行提取, 例如基于影视作品对话内容中不同角色的个性差异[5], 社交媒体数据中用户个体间的人格差异, 以及跨平台社交媒体用户人格表现的差异性[5-6] 等。这些基于文本的人格特征提取与建模, 按照文本处理的颗粒度[7], 可以分为词语或者句段, 也可以根据自然语言的处理方式, 分为数值的向量表达和机器模型生成的嵌入表达 [8-10], 研究会根据不同情境需求提出不同的组合方法。

2.1.1　词向量和统计回归

在基于文本的人格特征提取研究领域, 词向量 (word embedding) 承担着将语言信息转化为可计算特征的核心功能[11]。早期研究主要依赖 LIWC(linguistic inquiry and word count) 工具, 通过心理语言学词典对文本进行量化分析, 提取与心理状态相关的语言特征, 并结合分类模型 (如支持向量机、朴素贝叶斯等) 来识别用户的人格类型。随着深度学习技术的蓬勃发展, 研究者开始将词向量方法, 如 (GloVe(word to vector, global vectors for word representation) 等) 引入人格预测任务。这些方法通过将词语映射至高维向量空间, 有效捕捉词语间的复杂语义关系, 从而显著增强了模型识别文本中潜在人格特征的能力。

近年来研究重点已从简单的分类任务转向多输出回归建模, 使得对个体人格特质 (维度) 得分的预测更为精细化和准确化。研究先将用户数据分解为子词单元, 并映射为固定维度的词向量[12]。这些词向量是预训练模型从大规模语料中学习得到的, 能够捕捉词语的上下文信息与语义关系。模型随后对整个文本序列生成上下文化表示, 并用“最后隐藏层状态”表示。同时, 为了将这一高维表示转化为可用于人格特征预测的数值向量, 研究者会引入线性变换层。如在 RoBERTa(robustly optimized BERT approach) 或 ALBERT(a lite BERT) 的个性化回归模型中 (图 1), 第一线性层将隐藏状态的输出映射为 128 维的中间表示, 用于提取最具预测意义的特征。随后通过激活函数, 如 ReLU(rectified linear unit) 增强非线性表达能力, 最后一个线性层则将这一中间表示转换为对应于五大人格维度的连续得分, 实现多元输出的回归建模。

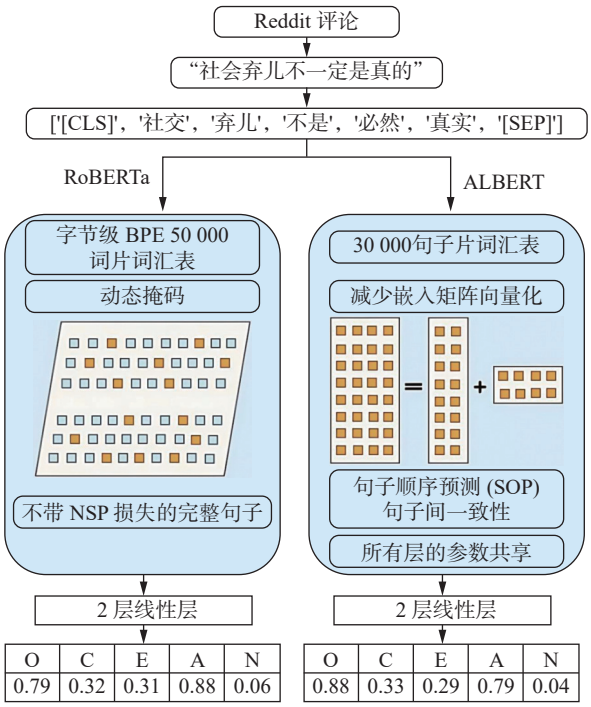


图 1　人格特质回归模型架构[12]

Fig. 1　Architecture of the personality trait regression model[12]

2.1.2　句段嵌入和机器学习

词向量和统计回归的方法可以更加精确地预测人格特质，但对长文本数据处理能力弱，而句段嵌入和机器学习的方式可以弥补这一缺失。句嵌入 (sentence embedding) 可以将完整句子，转换成一个固定长度的数值向量。例如 Distil-RoBERTa[13] 作为句嵌入式模型，对整条序列做自注意力编码，让孤立句向量更能捕捉“话语流”中的语境差异 (图 2)，再为每个句段嵌入一层共享参数的 MLP (multi-layer perceptron)，同时预测该句在 11 个情绪和人格维度上的标签，实现多句并行的序列句子分类 (sequential sentence classification, SSC)。这种情绪与人格的联合分析方法，可以应用于特定场景。例如在自动驾驶领域中，通过多智能体强化学习并引入个性参数，成功模拟不同驾驶风格，从而提高了模型的泛化能力[14]。

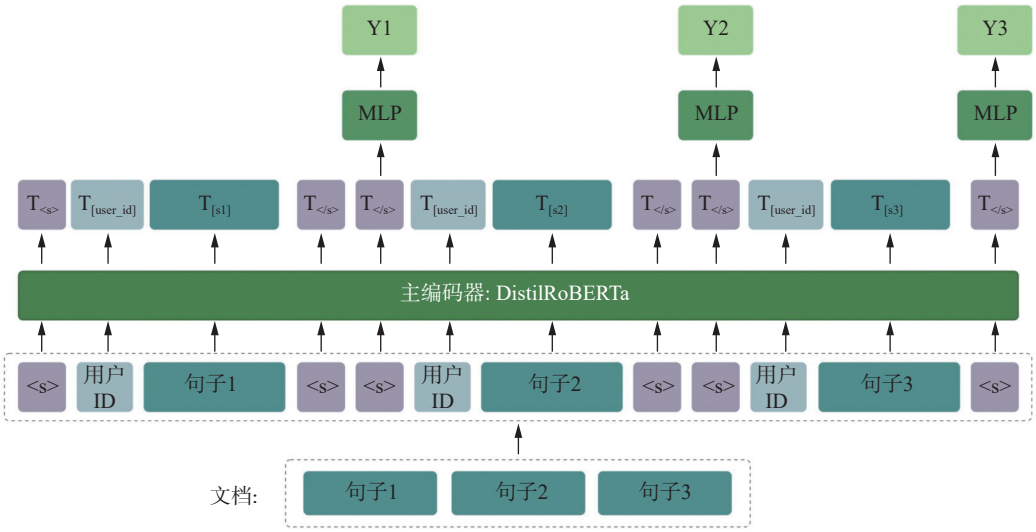


图 2　句段嵌入的人格化特征提取与建模

Fig. 2　Sentence-embedding-based personalized feature extraction and modeling

为了捕捉“谁在感受/标注”这一主观差异，研究者也提出各种方法，例如可以在每个句子前额外插入一个 UserID 这种特殊词元 (token)，从而把个人偏好映射到句段嵌入中。模型 (Transformer) 因而学习到“文本语义 × 用户向量”的交互，把个体偏好编进情感表示，实现了对人格化表达的建模。

大语言模型人格化的发展方向，从利用心理学 LIWC 工具，到结合自然语言处理中的数值向量、句段嵌入式表达，逐渐实现更加准确和细致的人格特征的提取、预测[15]。在人格特征提取中，词向量方法 (如 BERT、Word2Vec) 通过词语级嵌入实现细粒度语义控制，捕捉个体用词偏好与情感倾向；而句段嵌入方法 (如 DistilRoBERTa、Sentence-BERT) 通过上下文编码处理长文本的语义连贯性，可规模化提取语言风格与逻辑模式。两者形成互补，词向量聚焦微观特征解析，句嵌入强化宏观语境建模。在现有研究中多将二者结合 (图 3)，来兼顾个体差异与场景适应性，提升人格预测与模拟的全面性。

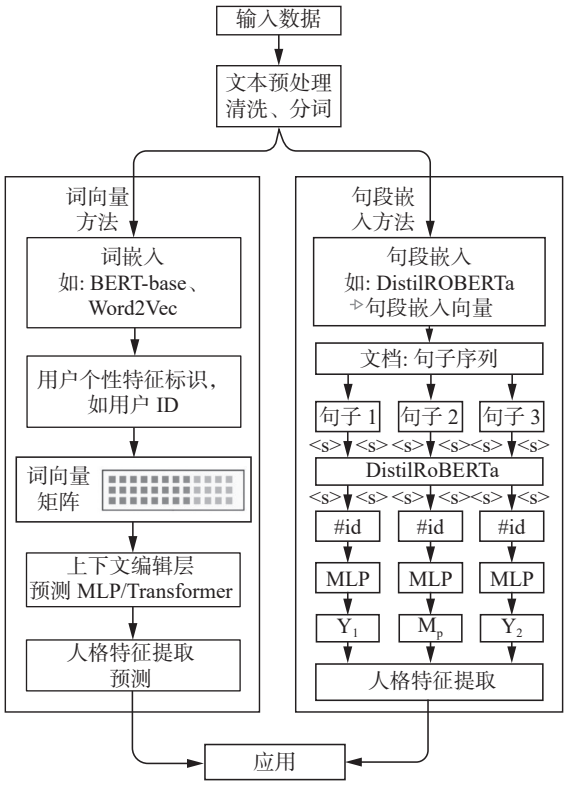


图 3　词向量和句嵌入的人格化方法

Fig. 3　Personalized modeling method based on word embeddings and sentence embeddings

2.2　模型的人格化调控技术

除了预训练以外, 模型可以利用少量特定领域有标签数据, 对模型参数进行调整和优化, 这一过程通常被概括为微调 (fine-tune)。例如全参数微调利用特定任务数据, 更新模型中所有的参数。参数高效微调, 只更新模型的一小部分参数, 或是在模型中引入少量的可训练新参数。除了这两种传统的微调方法外, 知识蒸馏侧重于模型压缩与知识迁移; 提示词工程通过精心设计的指令引导模型行为; 多智能体协作可以用系统架构的方法, 处理复杂的角色任务。本质上, 微调的核心在于将预训练所获的广博知识, 精准地提炼并专精于某一特定任务或领域, 使其成为处理专业问题的得力工具。

为了不牺牲其广泛理解能力的前提下, 在人格表现这一专业应用中表现出色, 调控技术可根据调控作用的对象与环节划分为模型内部参数优化与外部交互引导两大范式 (图 4)。前者通过直接干预模型架构或参数空间实现人格特质的植入, 后者则通过外部环境或工作流设计间接引导模型输出, 无需触及模型底层参数。

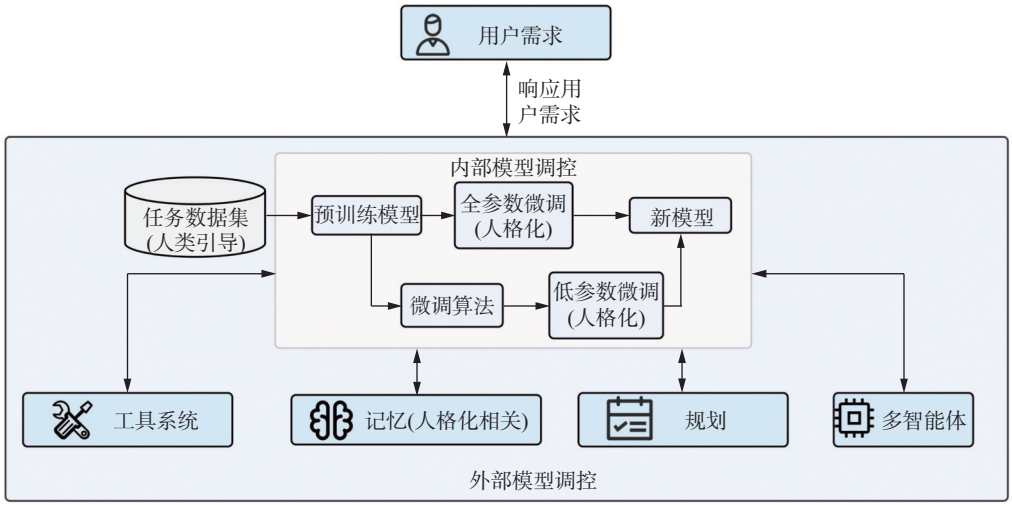


图 4　模型人格化表达的内部与外部调控方法

Fig. 4　Internal and external regulation methods for model personalization expression

2.2.1　模型内部

在模型内部实现人格化的输出, 研究通常采用构建专门的人格数据集、高效参数调试等方法。CPED (Chinese personalized and emotional dialogue) 数据集[16] 提供了融合情绪与人格标注的中文语料, MDPE(multimodal deception dataset with personality and emotional characteristics) 数据集[17] 聚焦于多模态欺骗检测中的人格特征建模, 而 PRODIGy 数据集 (profile-based dialogue generation dataset)[18] 则支持多维度用户配置的人格信息建构, CC2PC(chit-chat to persona-chat) 框架把闲聊数据转换为具有人格记忆的对话数据集[19], 并用作模型训练。HaRT 模型 (human-aware recurrent Transformer)[20] 通过社交媒体的预训练数据集, 提升了对人格特征的感知能力。数据集构建人格的通常采用的方法, 一般先从对话数据中自动提取“真实人格基准”摘要, 通过语言复述增强其词汇多样性; 继而建立大规模“角色候选池”, 运用自然语言推理 (natural language inference) 筛选出与真实角色存在逻辑中立关联的候选摘要; 最终将原始摘要与候选摘要共同构成“角色记忆”库, 通过前置整合到对话文本形成新训练样本, 驱动模型在生成过程中实现记忆显式调用机制 (图 5)。

这种从数据层面注入人格信息的方法, 能更细微地控制模型的人格表现, 但较于用极少可训练参数的微调方法, 它们因依赖数据收集、标注, 而人工成本较高。GPT3.5、Claude 3.7、豆包等模型的发展, 让其自身的人格表现力提升, 现有研究也多采用微调方法, PersonaPKT(persona-based parameter-efficient knowledge transfer)[21] 通过参数高效微调策略, 在极少参数更新的条件下实现人格一致性的提升[7]。在结构与机制层面 MIR-ACLE[22](multiple personal attributes control within latent-space energy-based models) 用潜空间能量模型, 实现复杂人格属性控制, DLVGen (dual latent variable generator)[23] 利用双潜变量模型, 在没有明确人格信息的情况下生成人格对话, Orca[24] 与 RoleCraft-GLM [25] 则基于指令微调策略, 构建了具备精细角色设定和情感刻画的角色扮演模型, 显著提升了交互的沉浸感与一致性。

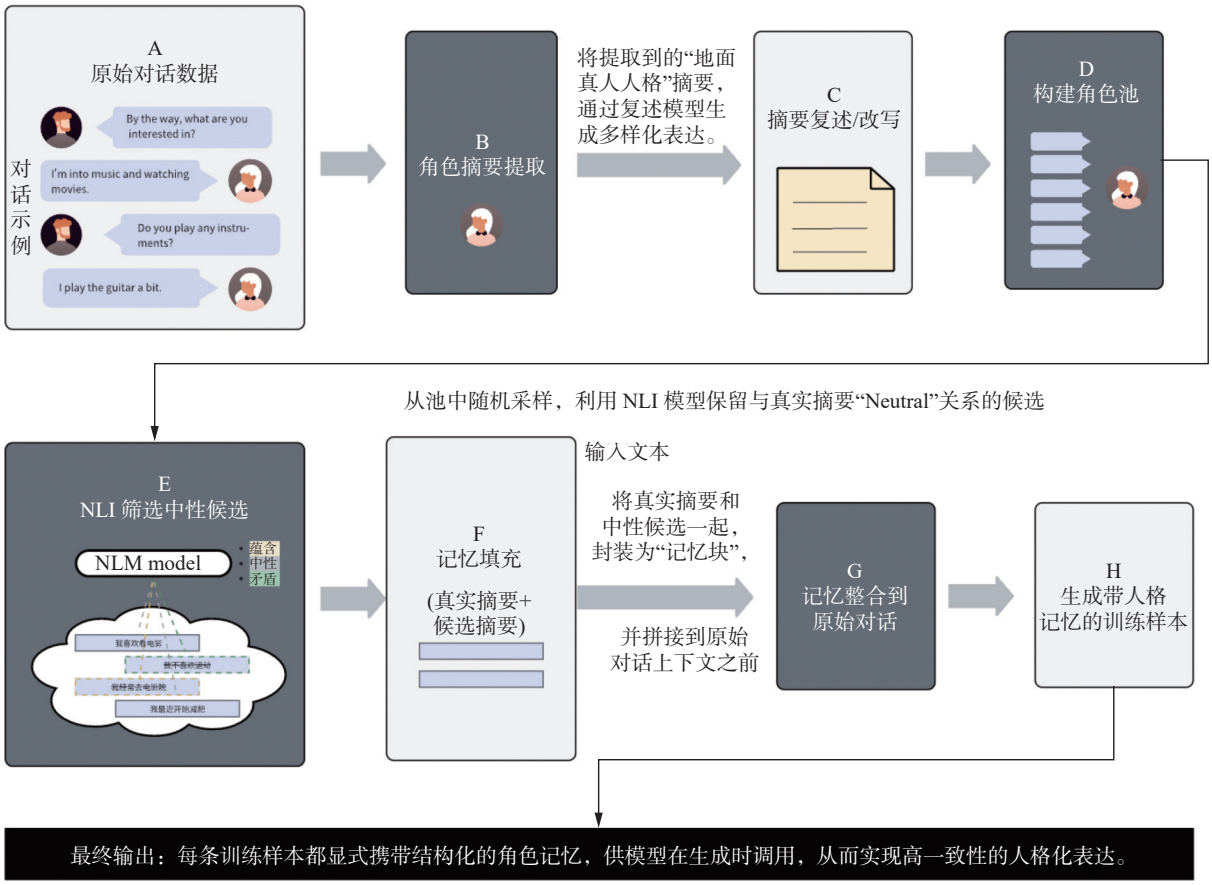


图 5　驱动模型人格化表现的框架示例

Fig. 5　Illustrative framework for enabling personality expression in models

从已有的研究中可以看出，用较少样本进行微调的 LoRA(low-rank adaptation) 较为主流，尤其在探索人格生成过程中的调节方面。LoRA 作为一种参数高效的微调 (parameter-efficient fine-tuning, PEFT) 方法，其核心思想是在冻结大语言模型原始权重的同时，仅向每个 Transformer 层中注入可训练的低秩矩阵，从而极大减少可训练的参数规模。Shi 等[26] 通过 LoRA 微调后的 3 种模型 (原始数据、增强数据、筛选后数据)，实现在同一个基础模型上快速切换至多种角色人格，并结合 GPT-4 打分的方式确保模型的角色扮演能力。DPG(dynamic personality generation)[27]、PsychAdapter[28] 分别从结构动态性、超网络建模与分层条件生成等维度出发，通过心理学指导的架构创新实现人格动态调控。其中 DPG 类似于微调的方法，构建了大五人格量表的角色对话数据集，采用双阶段超网络架构，通过 GPT-4 分析剧本对话生成人格特征标记后，训练分层超网络生成适配器参数，使 LLM 能根据角色设定动态重构人格表达模式。而 PsychAdapter 则创新性地在模型架构中植入心理特征输入通道，通过可扩展的维度扩展模块，将连续型人格评分、心理健康指标与人口统计学特征进行分层融合，使每个解码层都能接受心理条件的动态调控。

总的来说，模型内部的人格化调控技术其实就是在选定人格相关数据集和预训练模型的基础上，通过设置合适的超参数并对模型进行必要的调整。在这一过程中人格评分训练是最重要的环节，通常采用人类反馈作为奖励信号，来微调强化学习模型的方法 (图 6)，这种方法通常有 3 个步骤。

首先，使用标注过的数据来调整预训练模型的参数，使其更好地适应人格化任务。

其次，训练用于评估文本序列人格化质量的奖励模型，它接受一个文本作为输入，并输出一个数值，表示该文本符合人类偏好的程度，这个奖励信号在后续的强化学习训练中至关重要，可以指导模型生成更符合人格化表达的文本。最后，训练强化学习模型，在强化学习框架中，需要定义状态空间、动作空间、策略函数和价值函数，状态空间是输入序列的分布，动作空间是所有可能的 token(即数据集中的词)，价值函数结合了奖励模型的输出和策略约束，用于评估在给定状态下采取特定动作的价值，策略函数就是经过微调的大型语言模型，它根据当前状态选择下一个动作 (token)，以最大化累计奖励，从而实现对模型人格化的需求。

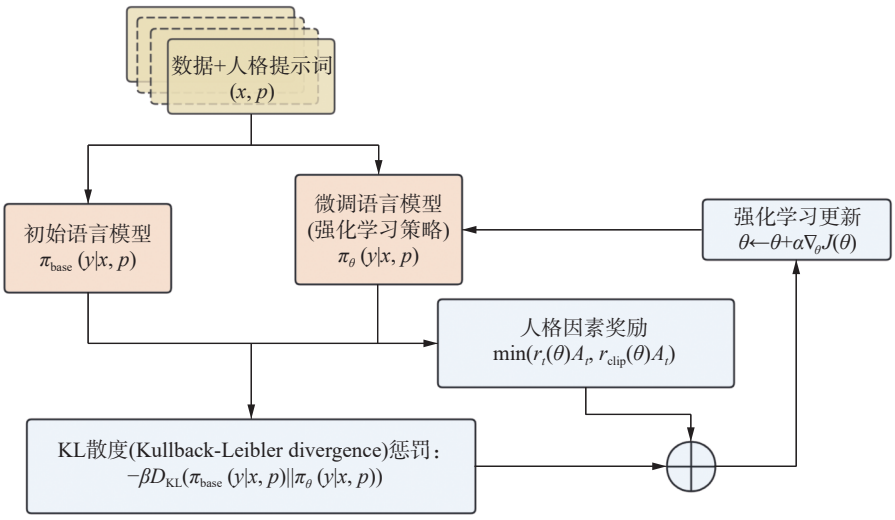


图 6　模型人格化的强化学习微调技术

Fig. 6　Reinforcement learning fine-tuning techniques for model personification

2.2.2　提示引导模型人格化表达

引导模型人格化表达方法笼统分为提示词[29]和智能体两类 (表 2、3)。具体的方法中, 零样本 (zero-shot) 或者少样本 (few-shot / one-shot) 追求用较少的样本量, 来完成人格风格迁移与内容生成, 实现跨领域 (餐馆推荐、电子游戏) 应用[30]。角色/系统/上下文提示 (role/system/contextual prompting)[29] 是较为常见的提示词方法。通过更精细、更有针对性的提示词设计过程[29], 有效增强模型在语言风格和人格表现上的差异性, 使模型产生的对话更符合设定的人格原型。值得一提的是, 提示词方法因其可用自然语言表示, 拥有一些社会科学视角的人格生成方法, 研究者探索在没有专门编程语言知识的情况下, 如何依靠“社会角色”“类别描述”等日常社会学资源, 来配置并操控人工智能的输出, 进而表现出特定的人格或“互动立场”。例如, 通过成员类别化分析 (membership categorisation analysis) 的框架[31], 揭示提示词中的“类别–谓词–语境”, 实现技术架构与社会类别化实践的双向形塑 (double hermeneutic), 通过这一框架自然语言交互被转译为算法可识别的谓词系统, 最终生成具有社会角色期待属性的智能体回复内容。

表 2　引导模型人格表达的提示词方法

Table 2　Prompt word method to guide model personality expression

提示词方法	人格任务目标	数据来源	文献来源
角色/系统/上下文提示 (role/system/contextual prompting)	通过上下文提示操控模型人格表现	社交媒体数据Reddit	[32]
	通过模拟不同人格提示生成内容	人类写作数据+生成数据	[33]
	通过角色扮演提示引导LLM行为	生成数据	[34]
	使用系统提示词生成个性	生成数据	[35]
	通过提示赋予LLM智能体人格特质	生成数据	[36]
	通过提示词引导模型	生成数据	[37]
	使用随机化提示调整AI性格	生成数据	[38]
	使用LLMs调整机器人行为表现个性	实验、交互数据	[39]
	利用提示词定制智能体个性和背景	生成数据	[40]
	直接使用GPT-3驱动未调整模型	游戏交互数据	[41]
零样本(zero-shot)	通过设计提示词调整模型行为	生成数据	[29]
	提取的人格特质引导LLM生成回应	人类写作数据	[42]
	测试ChatGPT默认模型表现	心理学量表结果	[43]
	通过设计测试问题进行分析	心理学量表结果	[44]
	使用零样本提示控制生成风格	心理学量表结果	[30]

续表 2

提示词方法	人格任务目标	数据来源	文献来源
退一步提示 (step-back prompting)	设计提示词引导模型生成	实验、交互数据	[45]
	通过提示词赋予LLM合成人格特质	生成数据	[46]
	使用提示词模拟实验	生成数据	[47]
自治性 (self-consistency)	通过多智能体自然语言交互模拟	生成数据	[48]
	通过提示词赋予LLM特定人格特征	生成数据	[49]
	强调自我反思和认知建模	生成数据	[50]
代码提示 (code prompting)	调整提示词模拟个性特征	心理学量表结果	[51]

表 3　智能体在模型人格中的作用

Table 3　Role of agent personality in model persona

描述	智能体作用	文献来源
扩展AgentSpeak架构	支持共情交互与情感维系	[52]
通过提示赋予LLM智能体人格特质	模拟人类行为与协作动态	[38]
利用提示词定制智能体个性和背景	提供个性化多模态交互及反馈	[42]
多智能体协作框架	动态重构维持角色人设一致	[53]
整合多模态技术构建对话智能体	弥合人机交流差距	[54]
使用新模型结构改进生成	为生成符合预设形象的回复	[55]
实验操控变量评估效果	通过手势表达个性与适应性	[56]
实验性研究用户行为与个性影响	影响用户行为并推断心理	[57]
使用LLMs调整机器人行为表现个性	通过展示个性来影响用户的认知和情绪状态	[41]
使用深度Q网络训练智能体	通过对抗训练揭示人格过拟合	[58]
使用多智能体强化学习框架	建模驾驶风格增强泛化性能	[59]
通过多智能体自然语言交互模拟	自主交互中个性与社会规范演化	[50]
多智能体协作框架	协同多模态推理与批判引导	[33]
使用提示词设定不同人格参数	模拟人格特质影响公共决策	[60]
使用表示工程调整模型内部表示	通过人格特质来影响合作水平与奉献权衡	[61]
基于规则和模拟的生成模型	通过展示个性来影响用户对其的认知和情绪反应	[62]
涉及多智能体强化学习方法	通过人格理论和脉冲网络提升协作和泛化能力	[63]

近年来热门的提示词方法之一是思维链(chain of thought, CoT), 通过思维链的人格化设计, 可以结构化推理路径, 显式建模用户行为特征与价值体系, 从而增强语言模型对复杂社会偏好的适应性。Santurkar 等[64] 针对传统角色提示词对隐含特征 (如历史观点) 的推理存在噪声敏感性和逻辑不一致性, 通过构建 4 步思维链推理流程 (特征过滤、隐式特征排序、价值信念规范理论 (value-belief-normative theory, VBN) 驱动分析、动态特征迭代), 实现了对用户显式属性 (如意识形态) 和隐式属性 (如历史行为) 的差异化处理。这种分层推理机制不仅解决了无关人物特征干扰预测的问题, 还通过 VBN 理论, 将社会心理学模型融入推理过程, 使模型能够系统解构用户环境价值观与个人行为规范的关系。举例来说, 这些基于思维链的人格化表达, 可表示为一种联合概率建模。假设 $P(B|p,h,c)$ 表示模型生成行为链 $B=\{b_1,b_2,\cdots,b_n\}$ 的概率, 根据链式法则, 可以将行为链的联合概率分解为各个行为的条件概率乘积:

$$P(B|p,h,C)=\prod_{i=1}^{n}P(b_i|p,h,C,b_{<i})$$

式中: p 表示人格设定包含人设的个性特征、动机、背景等信息; h 表示角色的历史叙述, 提供角色的背景故事和历史行为; C 表示当前情境, 描述了行为发生的具体环境和条件; $b_{<i}$ 表示在第 i 个行为之前的全部行为序列。

随着大语言模型的能力提升, 提示词方法不能满足更多操作需求, 为了调用外部工具, 退一步

思考的方法引入, 这一“思考−行动−观察”的循环, 直到解决问题, 可以说是智能体的早期形式。越来越多的研究通过构建更复杂的工作流或场景化任务, 来表现特定的人格特质。Li 等[65] 提出了一种名为“生成式智能体 (generative agents)”的新型计算软件智能体 (图 7), 旨在通过模拟人类行为来提升互动应用的逼真度。MAP(multi-agent personality shaping) 框架[31] 引入苏格拉底式的批判性智能体 (critic agent)。MemoryBank[66] 基于“遗忘曲线”的长期记忆机制, 保证模型的人格稳定性与一致性。SPACE THEA[67] 引导语音助手表现出情感共鸣、创造性和情绪智能等复杂人格特征。NarrativePlay[42] 通过复杂工作流的构建, 使模型在互动叙事中更自然地表现人格。Reflective Linguistic Programming[50] 让模型对自身的人格特质与情绪状态进行内省, 并规划相应的策略, 通过清晰的自我推理路径生成更加连贯且人格鲜明的交互内容。与此同时研究还发现, AFSPP 框架 (agent framework for shaping preference and personality)[68]、PsyPlay[37] 等不同智能体之间的社会互动, 能够显著影响智能体人格特征的形成, 甚至再现了真实社会中的人格演化规律, 实现人格的自主涌现[48], 有效提升团队生产力与创意输出[38] 和群体对话的真实感[69]。

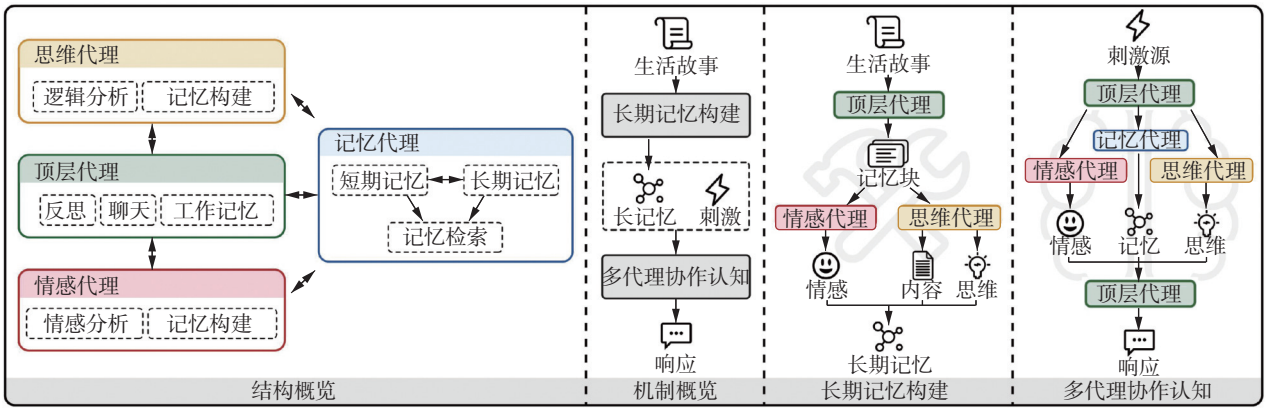


图 7　多智能体协作进行人格化表达

Fig. 7　Persona expression through multi-agent collaboration

这些智能体结合了大型语言模型 (LLM) 的能力, 能够动态地存储、检索并反思其记忆, 从而在互动中展现出令人信服的个体和群体行为。现有的研究中, 智能体通过记忆模块存储过去的思考、行动、观察以及与用户互动, 执行外部操作 (图 8)。其中记忆模块又分为短期记忆和长期记忆, 以及将这两种记忆结合的混合记忆, 这种分类方式可以提高智能体的短期推理能力和经验积累。

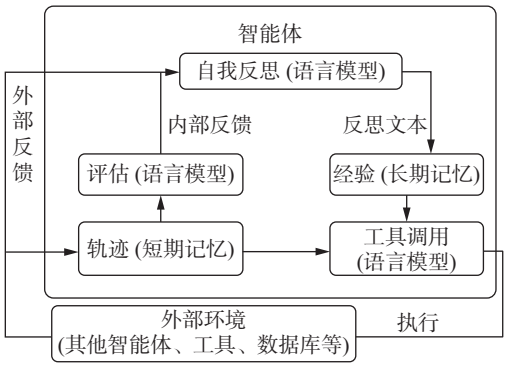


图 8　智能体的人格化表达

Fig. 8　Persona expression of the agent

相比于模型内部的预训练微调等方法, 提示词工程、多智能体协作交互与复杂任务流程设计, 可在不改动模型底层架构的前提下, 有效引导大语言模型表现出特定的人格特征。投资决策仿真[47]、谈判场景模拟[46] 及公共空间社会模拟[53] 等场景实验, 证明了这些方法能够增强模型交互的专业性与用户满意度。这些外部交互技术不仅提高了语言模型在人格化表现上的真实性与稳定性, 更为未来的人机交互研究提供了新思路。除了采取强化学习、增强检索等创新方式外, 研究者常采用心理学量表作为先验的外部知识, 实现模型的人格化输出。其中经典心理学人格理论如大五人格、OCEAN 框架[67] 被广泛应用于模型的人格预训练数据构建[57]、人格化建模方法开发[70-71], 以及系统性评估 LLM 内在人格特征[60]。例如语言模型蒸馏出的 PersonalityChat 数据集[61], 基于人格量表的 Neural MultiVoice[62] 人格生成模型。值得关注的是, 一些研究突破了静态典型性人格生成的局限, 例如“Evolving Agents”系统通过认知、情感与人格成长的反馈循环, 模拟人格的自然演化过程[56], 社会敏感性与情境依赖性[72] 的人格互动, 自然语言推理进行人格检索[73], CharacterGPT[74] 框架提出“动态人格重建”等概念。随着大语言模

型的人格化表达技术日趋成熟——从数据集的精细预训练处理，到模型内部的定向微调，再到外部智能体方法的系统化应用——一个关键性研究议题逐渐浮现：鉴于这些方法本质上具有任务导向性，我们如何科学评估任务完成结果，并有效判断各种方法的有效性。

2.3　人格评估与对齐方法

语言大语言模型自身涌现出人格表达的能力。早期研究者多直接用心理学人格量表，对模型进行人格评估。随着个性化表达与交互场景的不断扩展，传统量表在任务导向、对话动态与情境敏感性方面暴露出局限。为此，心理学、认知科学与社会计算等学科开展跨界合作，围绕多轮对话、行为一致性与用户感知等维度，构建面向模型的个性化表达评测基准，以更贴近真实交互需求地衡量与比较模型的人格表现。

2.3.1　多维人格评估视角

现有工作通过借鉴心理学、人类学等各学科理论，构建检验模型的人格与认知“基准”(benchmark)。总体上分为两类范式 (表 4)，其一为经典心理测量的改良与迁移，该范式围绕人格与价值等构念，开展特质评估、价值观、道德观与政治态度的量表化测评[75]。其二是基于交互任务的综合评估，通常需要跨学科结合更多领域的路径，来测试大语言模型任务推理能力，人际交往中的情感、情智能力，长文本中保持人格一致等能力。例如 PsychoBench 整合“人格−情绪−动机”，实现大语言模型扮演角色心理特质的全景式评估[76]，IPIP 模型通过与游戏情境结合的方式，实现跨场景评估，涵盖人格、情感、动机等多个维度[70]。

表 4　人格评估范式

Table 4　Personality assessment paradigms

范式	分类	理论框架	代表论文
心理学量表	人格特质 (personality traits)	大五人格 (big five inventory)，HEXACO模型 (honesty -humility (h), emotionality (e), extraversion (x), agreeableness (a), conscientiousness (c), and openness to experience (o))，MBTI (myers-briggs type indicator)，暗黑三项 (dark triad)，其他自定义人格	[71]
	价值观 (values)	施瓦茨价值观理论 (Schwartz human values)，世界价值观调查 (world values survey)，GLOBE文化维度理论 (global leadership and organizational behavior effectiveness)，社会价值取向 (social value orientation)	[77]
	道德观 (morality)	道德基础理论 (moral foundations theory)，确定问题测验(defining issues test)	[78]
	政治态度与观点 (political attitudes & opinions)	美国全国选举研究 (American national election studies)，皮尤研究中心态度调查 (attitudes toward Politics)，德国选举研究 (German longitudinal election study)	[79]
交互任务评估	决策、偏见 (decision-making & bias)	信任度、诚实	[80]
	社交能力和情智能力 (social skills & emotional intelligence)	心智理论(theory of mind)，情绪智能，社交智能	[81]
	学习、认知能力 (learning & cognitive abilities)	心理疗愈	[82]
	角色扮演表现力 (role-playing expressiveness)	人格一致性，流畅性，记忆能力；专业性(金融分析，电商推荐)	[83]

人格评估的具体方法可以分为两类，一是评估模型自身的个性化表达能力，二是评估个性化调控手段的能力。前者优先选择纯文本的自陈式量表 (self-report)，把对人类受试者的原始说明和题目几乎“原封不动”提供给模型，尽量减少提示词工程，以评估“被用户感知的模型人格画像”，而不是宣称模型真的“有”人格[76]。论文多数采用零样本方式，并将量表说明与题干作为一段完整文本提交给模型以作答。这一类评估方法的量表分数，更应被视作接口层“外显行为配置”的投影，而非模型具备“稳定人格”的证据[84]。第二类对于调控技术的能力评估，通常会对其“人格一致性”的表达进行测试，旨在考察模型在多轮对话中，是否不自相矛盾并持续遵循人设知识、行为与说话风格等。这类评估方法，通常根据任务设置相应的指标。例如，知识呈现与人设的相关性，是否避免无据的知识，动作语气是否匹配人设，措辞、口头禅等是否贴合人设语言习惯等，并

通过人工或非人工的方法, 对生成结果打分[85]。

2.3.2　从“价值学习”到“人格对齐”

强化学习在智能体训练中的局限, Dewey[86] 提出以“价值学习框架”作为解决方案。价值学习摒弃环境奖励最大化目标, 转而根据可学习的价值模型U来确定能够产生最大效用期望的行动y_k^*:

$$y_k^* = \arg\max_{y_k} \mathbb{E}_{h\sim\pi(\cdot|h_k,y_k)}[U(h)]$$

式中: h表示以(h_k,y_k)为开头的行动−观测的序列, $h_k=(y_t,x_t)_{t=1}^{k-1}$表示第$k$步行动之前的行动−观测序列, y_t为第t步行动, x_t为第t步行动以后获得的观测。其核心创新在于动态效用的建模过程, 智能体不预设固定的价值标准, 而是维护基于交互序列增量更新的价值模型$U(h)$, 并通过期望效用计算, 整合历史信息进行迭代更新与加权价值评估。该机制引导智能体动态学习人类的价值标准, 避免了传统强化学习中的奖励黑客与外部回路依赖的问题。 可以说, 强化学习追求“让历史包含高回报”的行为, 但价值学习通过定义不同的U来指定多种终极目标。

随着大语言模型能力的提升, 现在大语言模型实践多以价值实践为主。“诚实、助人、无害”等人类偏好, 作为价值对齐目标, 人类反馈强化学习 (reinforcement learning from human feedback, RLHF) 为技术手段。但 RLHF 高度依赖人工偏好标注, 人工智能反馈强化学习基于已有的模型代替人类打分, 带来噪声偏好与模型依赖的稳健性问题。随着模型行为复杂化, 传统对齐手段在超强模型上略显失效; 在角色、人设等高语境依赖场景下, 人类标注既昂贵又难以规模化。这也是大语言模型相关研究, 从“普适价值对齐”延展到“人格对齐 (personality alignment)”[87]。人格对齐不仅让模型做对、做安全, 更要让其以特定个体、群体的偏好与优先级为约束, 在思维风格、话语风格与决策倾向上与之相契合。概念上, 它承接了价值学习对“价值显式化”的诉求, 与价值对齐“行为规约”上的实践, 进一步在个体化与情境化维度上细化对齐对象: 以人格特征作为对齐载体与行为先验, 约束模型在跨话题、跨任务、跨轮次中的稳定表达。

$$S_d = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{N_{d,i}}\sum_{\kappa\in D_{d,i}^{T}}\left|f_{\mathrm{LLM}}(\kappa,P)-f_{\mathrm{Person}}(\kappa,P)\right|\right) \tag{1}$$

西湖大学团队提出了人格对齐 (式 (1)), 其中定义了人格维度d的对齐误差分: 对每位受试者i, 在其测试量表$D_{d,i}^{T}$上计算大语言模型与该受试者在同一提示模板P下逐题κ下的评分差绝对值, 再在全部M位受试者上取均值, 得到最后的对齐分值 (aligned score, S_d)。分值越小 (最低为0) 表示模型, 在该人格维度与个体行为越一致。

“人格对齐”概念的提出, 在学术上为长期困扰大语言模型的人格漂移、长上下文人格一致性等问题, 提供了一种新的解决思路。它突破了以往仅关注模型是否满足某一静态人设标准的评估局限, 将人格一致性的衡量拓展为一个动态、递进的过程。在这一过程中, 模型不仅在表层的语言风格和行为表现上逐步贴近用户, 更在深层次的价值观与理念层面实现趋同。至此, 人格对齐有望让 AI 不再只是被动执行指令, 而是能够以动态的人格适配方式逐步靠近用户的交互逻辑和心理模型, 从而缩小用户心智模型与系统运作机制之间的过程鸿沟。

3　模型人格化的问题与展望

通过上述分析可以看出, 大语言模型能够通过内部参数调优, 和外部机制设计实现人格化表达。然而, 当前的人格化技术仍面临两个根本性限制: 一是人类在引导模型进行人格化设计时存在的方法缺陷; 二是模型本身作为基于先验知识的概率系统, 难以充分模拟人类基于个人经验的推理过程和主观个性化特征。正是受制于这两方面因素, 大语言模型人格化表达中, 呈现出来一些不足。

3.1　模型自身能力缺陷

模型自身的能力往往影响大语言模型的人格化表现, 体现在 6 个方面。1)LLMs 基于先验知识反馈的架构无法复制人类经验性学习模式。与人类通过迭代经验形成个性化记忆不同 (如“一朝被蛇咬, 十年怕井绳”), LLMs 无法实现动态的经验推理和个性化记忆迭代。2) 对于需要综合运用分析性推理、直觉思维和模式识别的复杂任务, LLMs 难以达到人类专家的深度分析和复杂推理水平[88], 特别影响其在专业领域的决策过程模拟。3) 尽管思维链等提示词方法, 能显著提升角色扮演能力, 但存在链长度参数 (steps of CoT) 等内在约束[89]。对于小型模型或缺乏充足示例的情况, 其效果受限, 表明模型控制机制受规模和示例数量等因素的制约。4) 较小模型的角色扮演能力受限, 即使提供高质量监督微调数据集, 提升效果仍然有限[90], 说明真实角色体现需要大量计算资源。5) 对于需要长期记忆的角色扮演智能体, 新信息可能覆盖先前学习内容, 在持续更新

中保持知识一致性仍是关键挑战[91]。6)LLMs 中的偏见从根本上影响认知和行为模拟能力。文化偏见限制跨文化交互理解，性别偏见导致刻板印象 (产生性别刻板行为模式的概率高 3~6 倍)，训练数据中某些人群的过度代表导致对少数群体思维模式的不真实模拟。

3.2　人类引导中的设计缺陷

除了以上模型自身的能力缺陷，人类在引导模型人格化表达时也有两方面的设计缺陷。一是人类复杂性的数据缺失问题。作为模仿目标，人类体现了独特的偏好、生活经历和复杂行为特征。当前的建模方法通常只能聚焦于特定的、可量化的行为指标，而无法将人类生活的全谱体验纳入模拟范围[92]。这种局限性使得模型忽视了个人历史、文化背景和生活经历对决策过程的深层影响，从而产生不完整或失真的人格化表达。另一方面是现有模型人格化表达，是对人类心理状态建模一个过度简化的过程。设计者在尝试建模复杂心理状态时，常采用过度简化的方法，将丰富的心理现象压缩为基本类别或数值量表。这种简化忽略了不同心理因素之间的微妙互动和动态关系，不能准确反映人类心理的复杂性和多维性。类似的设计缺陷同样体现在人格化对齐和测试的设计中[49]。现有的结构化人格测试主要针对人类认知模式设计，当应用于大语言模型时面临显著挑战。这些测试结果在实际交互场景中的应用能力，以及复杂现实生态环境的有效性有待进一步确认[93]。

3.3　模型人格化表达的挑战与展望

1) 人格一致性与真实性的困境。尽管借助心理学与认知科学理论基础，模型在人格多样性、稳定性与情境敏感性方面展现出一定潜力，但在真实性、精细度和一致性方面仍存在显著缺陷。最突出的问题是“人格漂移”现象，模型在长文本交互中难以维持稳定的人格特征，缺乏跨场景任务和多轮对话中的连贯性表达能力。因此需要一个统一的评估系统，目前研究者常依赖 MBTI、大五人格、Dark Triad 等不同框架进行评估，导致结果间缺乏可比性，严重阻碍了研究的可复现性与知识积累。特别是在智能体与聊天机器人应用中，亟需一个专门化的人格评估量表、量化标准与基准测试工具。此外，这个评估需要能反馈模型对人类深层心理机制的模拟，包括情绪表达、动机形成与内隐人格特质等方面，评估模型在人格表达的真实性与情感感染能力。

2) 角色扮演任务的四重平衡挑战。首先，使用 LoRA 等高效微调方法会损害多任务性能，而全面微调则面临高昂的计算成本。其次是可塑性与稳定性的平衡，模型需要在学习新信息的同时保留原有知识，这直接影响其获取特定领域知识的能力。第三重挑战涉及功能性与安全性的平衡。“对齐税”概念凸显了在不削弱推理和规划能力的前提下，训练模型符合人类价值观的困难。最后是角色扮演性能与响应时间的权衡，大型模型在查询重写、推理等任务中耗时较长，在需要快速响应的场景下性能受限，这对需要复杂情景感知的角色扮演任务尤为明显。

未来的研究亟需一个统一的理论框架，能够推动从评估人格−表达人格−调整人格的闭环系统构建，形成覆盖“识别−建模−生成−评估”的完整研究路径。通过跨学科理论集成，模型应不仅具备个性化“外观”，更能体现心理一致性、社会适应性与长期互动能力，来实现真正意义上的“人格化表达”。

4　结束语

随着大语言模型技术的迅猛发展，模型人格化表达已成为人机交互研究的重要前沿领域。本文通过对大语言模型人格化表达实现技术的系统综述，揭示了从输入侧的人格特征提取与预测、模型内部调控技术，到输出侧的人格评估与对齐等多维度的研究进展。

当前的人格化表达技术呈现出多样化的发展路径：从基于词向量和句段嵌入的特征提取方法，到模型内部的精细微调和外部智能体协作技术，再到基于心理学理论的评估与对齐机制。这些方法各有优势，在不同应用场景中展现出独特价值。然而，人格生成也面临着人格表达一致性与真实性不足、缺乏统一评估标准以及跨学科理论整合不足等挑战。

未来研究需要构建一个综合性的理论体系，支持模型在长期交互中保持人格一致性，同时实现对用户个性特征的动态识别与适应。特别是需要突破传统基于量表映射的框架局限，向“机制建构”转型，构建涵盖稳定特质、情绪状态、动机驱动和身份表达等多层次的人格建模体系。基于此可以建立统一的评估标准，以衡量模型在人格表达一致性、适应性和用户满意度等方面的表现。

在确保用户权益和伦理合规的前提下，系统性地推进模型人格化研究不仅有助于提升人机交互的自然度与用户接受度，更能为智能体在教育、医疗、社交等领域的落地应用提供坚实的理论

基础和技术支撑。通过跨学科理论集成与技术创新，期待大语言模型在未来能够展现出更加真实、丰富且多元的人格表达，从而实现从简单模仿人类性格向真正理解并表达独特人格特质的转变。

参考文献：

[1] WEIZENBAUM J. ELIZA: a computer program for the study of natural language communication between man and machine[J]. Communications of the ACM, 1966, 9(1): 36−45.

[2] METCALF K, THEOBALD B J, WEINBERG G, et al. Mirroring to build trust in digital assistants[EB/OL]. (2019−04−02)[2025−04−16]. http://arxiv.org/abs/1904.01664.

[3] JAMALIAN N, CONSTANTINIDES M, JOGLEKAR S, et al. Our Nudges, Our Selves: tailoring mobile user engagement using personality[EB/OL]. (2023−07−24)[2025−04−16]. http://arxiv.org/abs/2307.13145.

[4] ZHU Jianfeng, JIN Ruoming, COIFMAN K G. Investigating large language models in inferring personality traits from user conversations[EB/OL]. (2025−01−13)[2025−04−16]. http://arxiv.org/abs/2501.07532.

[5] JIANG Hang, ZHANG Xianzhe, CHOI J D. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings[EB/OL]. (2019−11−21)[2025−04−16]. http://arxiv.org/abs/1911.09304.

[6] DATTA A, CHAKRABORTY S, MUKHERJEE A. Personality detection and analysis using Twitter data[EB/OL]. (2023−09−11)[2025−04−16]. http://arxiv.org/abs/2309.05497.

[7] YANG Qi, FARSEEV A, FILCHENKOV A. Two-faced humans on Twitter and Facebook: harvesting social multimedia for human personality profiling[C]//Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval. New York: ACM, 2021: 39−47.

[8] VU X S, FLEKOVA L, JIANG Lili, et al. Lexical-semantic resources: yet powerful resources for automatic personality classification[EB/OL]. (2017−11−27)[2025−04−16]. http://arxiv.org/abs/1711.09824.

[9] XING Yujie, FERNÁNDEZ R. Automatic evaluation of neural personality-based chatbots[EB/OL]. (2018−09−30)[2025−04−16]. http://arxiv.org/abs/1810.00472.

[10] ZHANG Le, PENG Songyou, WINKLER S. PersEmoN: a deep network for joint analysis of apparent personality, emotion and their relationship[J]. IEEE transactions on affective computing, 2022, 13(1): 298−305.

[11] CAO Xubo, KOSINSKI M. Large language models know how the personality of public figures is perceived by the general public[J]. Scientific reports, 2024, 14: 6735.

[12] HABIB F, ALI Z, AZAM A, et al. Navigating pathways to automated personality prediction: a comparative study of small and medium language models[J]. Frontiers in big data, 2024, 7: 1387325.

[13] NGO A, KOCOŃ J. Integrating personalized and contextual information in fine-grained emotion recognition in text: a multi-source fusion approach with explainability [J]. Information fusion, 2025, 118: 102966.

[14] LIU Weiwei, HU Wenxuan, JING Wei, et al. Learning to model diverse driving behaviors in highly interactive autonomous driving scenarios with multi-agent reinforcement learning[EB/OL]. (2024−02−21)[2025−04−16]. http://arxiv.org/abs/2402.13481.

[15] PETERS H, MATZ S. Large language models can infer psychological dispositions of social media users[EB/OL]. (2024−07−05)[2025−04−16]. http://arxiv.org/abs/2309.08631.

[16] CHEN Yirong, FAN Weiquan, XING Xiaofen, et al. CPED: a large-scale Chinese personalized and emotional dialogue dataset for conversational AI[EB/OL]. (2022−05−29)[2025−04−16]. http://arxiv.org/abs/2205.14727.

[17] CAI Cong, LIANG Shan, LIU Xuefei, et al. MDPE: a multimodal deception dataset with personality and emotional characteristics[EB/OL]. (2024−07−17)[2025−04−16]. http://arxiv.org/abs/2407.12274.

[18] OCCHIPINTI D, TEKIROGLU S S, GUERINI M. PRODIGy: a profile-based dialogue generation dataset [EB/OL]. (2024−08−27)[2025−04−16]. http://arxiv.org/abs/2311.05195.

[19] KEUM B, SUN J, LEE W, et al. Persona-identified chatbot through small-scale modeling and data transformation[J]. Electronics, 2024, 13(8): 1409.

[20] SONI N, MATERO M, BALASUBRAMANIAN N, et al. Human language modeling[C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin: ACL, 2022: 622–636.

[21] JAIN N, WU Zekun, MUNOZ C, et al. From Text to Emoji: how peft-driven personality manipulation unleashes the emoji potential in LLMs[EB/OL]. (2025−02−25)[2025−04−16]. http://arxiv.org/abs/2409.10245.

[22] LU Zhenyi, WEI Wei, QU Xiaoye, et al. MIRACLE: towards personalized dialogue generation with latent-space multiple personal attribute control[EB/OL]. (2023−10−22)[2025−04−16]. http://arxiv.org/abs/2310.18342.

[23] LEE J Y, LEE K A, GAN W S. DLVGen: a dual latent variable approach to personalized dialogue generation[EB/OL]. (2021−11−22)[2025−04−16]. http://arxiv.org/abs/2111.11363.

[24] HUANG Yuxuan. Orca: enhancing role-playing abilities of large language models by integrating personality traits[EB/OL]. (2024−11−15)[2025−04−16]. http://arxiv.org/abs/2411.10006.

[25] TAO Meiling, LIANG Xuechen, SHI Tianyu, et al. RoleCraft-GLM: advancing personalized role-playing in large language models[EB/OL]. (2023−12−17)[2025−04−16].

http://arxiv.org/abs/2401.09432.

[26] SHI Haozhe, NIU Kun. Enhancing persona consistency with large language models[C]//Proceedings of the 2024 5th International Conference on Computing, Networks and Internet of Things. Tokyo: ACM, 2024: 210−215.

[27] LIU Jianzhi, GU Hexiang, ZHENG Tianyu, et al. Dynamic generation of personalities with large language models [EB/OL]. (2024−04−10)[2025−04−16]. http://arxiv.org/abs/2404.07084.

[28] VU H, NGUYEN H A, GANESAN A V, et al. PsychAdapter: adapting LLM Transformers to reflect traits, personality and mental health[EB/OL]. (2024−12−22)[2025−04−16]. http://arxiv.org/abs/2412.16882.

[29] GU Heng, DEGACHI C, GENÇ U, et al. On the effectiveness of creating conversational agent personalities through prompting[EB/OL]. (2023−10−17)[2025−04−16]. http://arxiv.org/abs/2310.11182.

[30] RAMIREZ A, ALSALIHY M, AGGARWAL K, et al. Controlling personality style in dialogue with zero-shot prompt-based learning[EB/OL]. (2023−02−08)[2025−04−16]. http://arxiv.org/abs/2302.03848.

[31] HOUSLEY W, DAHL P. Membership categorisation, sociological description and role prompt engineering with ChatGPT[J]. Discourse & communication, 2024, 18(6): 848−858.

[32] CARON G, SRIVASTAVA S. Identifying and manipulating the personality traits of language models[EB/OL]. (2022−12−20)[2025−04−16]. http://arxiv.org/abs/2212.10276.

[33] JIANG Hang, ZHANG Xiajie, CAO Xubo, et al. PersonaLLM: investigating the ability of large language models to express personality traits[EB/OL]. (2023−05−04)[2025−04−16]. http://arxiv.org/abs/2305.02547.

[34] SHEN Chenglei, XIE Guofu, ZHANG Xiao, et al. On the decision-making abilities in role-playing using large language models[EB/OL]. (2024−02−29)[2025−04−16]. http://arxiv.org/abs/2402.18807.

[35] ZHAO Yilin, YUAN Xinbin, GAO Shanghua, et al. ChatAnything: facetime chat with LLM-enhanced personas[EB/OL]. (2023−11−12)[2025−04−16]. http://arxiv.org/abs/2311.06772.

[36] DUAN Yifan, TANG Yihong, BAI Xuefeng, et al. The power of personality: a human simulation perspective to investigate large language model agents[EB/OL]. (2025−02−28)[2025−04−16]. http://arxiv.org/abs/2502.20859.

[37] YANG Tao, ZHU Yuhua, QUAN Xiaojun, et al. PsyPlay: personality-infused role-playing conversational agents [EB/OL]. (2025−02−06)[2025−04−16]. http://arxiv.org/abs/2502.03821.

[38] JU H, ARAL S. Collaborating with AI Agents: field experiments on teamwork, productivity, and performance [EB/OL]. (2025−03−23)[2025−04−16]. http://arxiv.org/abs/2503.18238.

[39] MEDGYESY D, GALAS J, POL J van, et al. Existential Crisis: a social robot's reason for being[EB/OL]. (2025−01−06)[2025−04−16]. http://arxiv.org/abs/2501.03376.

[40] HASAN M, OZEL C, POTTER S, et al. SAPIEN: affective virtual agents powered by large language models[C]// 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. Cambridge: IEEE, 2023: 1−3.

[41] SUN Yuqian, WANG Hanyi, CHAN P M, et al. Fictional worlds, real connections: developing community storytelling social chatbots through LLMs[EB/OL]. (2023−09−20)[2025−04−16]. http://arxiv.org/abs/2309.11478.

[42] ZHAO Runcong, ZHANG Wenjia, LI Jiazheng, et al. NarrativePlay: interactive narrative understanding[EB/OL]. (2023−10−02)[2025−04−16]. http://arxiv.org/abs/2310.01459.

[43] SCHAAFF K, REINIG C, SCHLIPPE T. Exploring ChatGPT's empathic abilities[EB/OL]. (2023−08−07)[2025−04−16]. http://arxiv.org/abs/2308.03527.

[44] RUTINOWSKI J, FRANKE S, ENDENDYK J, et al. The self-perception and political biases of ChatGPT[J]. Human behavior and emerging technologies, 2024: 7115633.

[45] GRASSI L, RECCHIUTO C, SGORBISSA A. Enhancing LLM-based human-robot interaction with nuances for diversity awareness[EB/OL]. (2024−06−25)[2025−04−16]. http://arxiv.org/abs/2406.17531.

[46] HUANG Y J, HADFI R. How personality traits influence negotiation outcomes? A simulation based on large language models[EB/OL]. (2024−07−16)[2025−04−16]. http://arxiv.org/abs/2407.11549.

[47] BORMAN H, LEONTJEVA A, PIZZATO L, et al. Do LLM personas dream of bull markets? Comparing human and AI investment strategies through the lens of the five-factor model[EB/OL]. (2024−10−28)[2025−04−16]. http://arxiv.org/abs/2411.05801.

[48] TAKATA R, MASUMORI A, IKEGAMI T. Spontaneous emergence of agent individuality through social interactions in LLM-based communities[EB/OL]. (2024−11−05)[2025−04−16]. http://arxiv.org/abs/2411.03252.

[49] FRISCH I, GIULIANELLI M. LLM agents in interaction: measuring personality consistency and linguistic alignment in interacting populations of large language models [EB/OL]. (2024−02−05)[2025−04−16]. http://arxiv.org/abs/2402.02896.

[50] FISCHER K. Reflective linguistic programming (RLP): a stepping stone in socially-aware AGI (SocialAGI)[EB/OL]. (2023−05−22)[2025−04−16]. https://arxiv.org/abs/2305.12647.

[51] GUO Yaoqi, CHEN Zhenpeng, ZHANG J M, et al. Personality-guided code generation using large language models[EB/OL]. (2024−10−16)[2025−04−16]. http://arxiv.org/abs/2411.00006.

[52] PARK J S, ZOU C Q, SHAW A, et al. Generative agent

simulations of 1, 000 people[EB/OL]. (2024−11−15)[2025−04−16]. http://arxiv.org/abs/2411.10109.

[53] REN Mingjun, XU Wentao. The impact of big five personality traits on AI agent decision-making in public spaces: a social simulation study[EB/OL]. (2025−01−15)[2025−04−16]. http://arxiv.org/abs/2503.15497.

[54] LUKIN S M, ANAND P, WALKER M, et al Argument strength is in the eye of the beholder: audience effects in persuasion[EB/OL]. (2017−08−30)[2025−04−16]. http://arxiv.org/abs/1708.09085.

[55] NOEVER D, HYAMS S. AI Text-to-Behavior: a study in steerability[EB/OL]. (2023−08−07)[2025−04−16]. http://arxiv.org/abs/2308.07326.

[56] LIOU H C, HSIEH H Y. Modeling friendship networks among agents with personality traits[EB/OL]. (2020−04−27)[2025−04−16]. http://arxiv.org/abs/2004.12901.

[57] ZHAN Baohua, HUANG Yongyi, CUI Wenyao, et al. Humanity in AI: detecting the personality of large language models[EB/OL]. (2024−10−11)[2025−04−16]. http://arxiv.org/abs/2410.08545.

[58] KRUIJSSEN J D, EMMONS N. Deterministic AI agent personality expression through standard psychological diagnostics[J]. Allora decentralized intelligence, 2025, 2: 15−39.

[59] KERZ E, QIAO Yu, ZANWAR S, et al. Pushing on personality detection from verbal behavior: a Transformer meets text contours of psycholinguistic features[EB/OL]. (2022−04−10)[2025−04−16]. http://arxiv.org/abs/2204.04629.

[60] JIANG Guangyuan, XU Manjie, ZHU Songchun, et al. Evaluating and inducing personality in pre-trained language models[C]//In Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: NIP, 2023: 10622−10643.

[61] LOTFI E, DE BRUYN M, BUHMANN J, et al. PersonalityChat: conversation distillation for personalized dialog modeling with facts and traits[C]//Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics. Singapore: ACL, 2023: 353−371.

[62] ORABY S, REED L, SHARATH T S, et al. Neural MultiVoice models for expressing novel personalities in dialog[C]//Interspeech 2018. Hyderabad: ISCA, 2018: 3057−3061.

[63] ZHANG Jian, WANG Zhiyuan, WANG Zhangqi, et al. MAPS: a multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving[EB/OL]. (2025−03−21)[2025−04−16]. http://arxiv.org/abs/2503.16905.

[64] SANTURKAR S, DURMUS E, LADHAK F, et al Whose opinions do language models reflect?[C]//Proceedings of the 40th International Conference on Machine Learning. Honolulu: JMLR, 2023: 29971−30004.

[65] LI Xiyun, NI Ziyi, RUAN Jingqing, et al. Mixture of personality improved spiking actor network for efficient multi-agent cooperation[J]. Frontiers in neuroscience, 2023, 17: 1219405.

[66] ZHONG Wanjun, GUO Lianghong, GAO Qiqi, et al. MemoryBank: enhancing large language models with long-term memory[EB/OL]. (2023−05−17)[2025−04−16]. http://arxiv.org/abs/2305.10250.

[67] SPATHELF M, BENDEL O. The SPACE THEA project[EB/OL]. (2022−06−17)[2025−04−16]. http://arxiv.org/abs/2206.10390.

[68] HE Zihong, ZHANG Changwang. AFSPP: agent framework for shaping preference and personality with large language models[EB/OL]. (2024−01−05)[2025−04−16]. http://arxiv.org/abs/2401.02870.

[69] WANG Weixuan, CAI Xiaoling, HUANG C H, et al. Emily: developing an emotion-affective open-domain chatbot with knowledge graph-based persona[EB/OL]. (2021−09−18)[2025−04−16]. http://arxiv.org/abs/2109.08875.

[70] SANG Yisi, MOU Xiangyang, YU Mo, et al. MBTI personality prediction for fictional characters using movie scripts[EB/OL]. (2022−10−20)[2025−04−16]. http://arxiv.org/abs/2210.10994.

[71] BARUA A, BRASE G, DONG K, et al. On the psychology of GPT-4: moderately anxious, slightly masculine, honest, and humble[EB/OL]. (2024−02−01)[2025−08−20]. http://arxiv.org/abs/2402.01777.

[72] GHAFURIAN M, ELLARD C, DAUTENHAHN K. Social companion robots to reduce isolation: a perception change due to COVID−19[EB/OL]. (2020−08−12)[2025−04−16]. http://arxiv.org/abs/2008.05382.

[73] LIU Yifan, WEI Wei, LIU Jiayi, et al. Improving personality consistency in conversation by persona extending [C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta: ACM, 2022: 1350−1359.

[74] PARK J, PARK C, LIM H. CharacterGPT: a persona reconstruction framework for role-playing agents[EB/OL]. (2024−05−30)[2025−04−17]. http://arxiv.org/abs/2405.19778.

[75] YE Haoran, JIN Jing, XIE Yuhang, et al. Large language model psychometrics: a systematic review of evaluation, validation, and enhancement[EB/OL]. (2025−05−13)[2025−08−20]. http://arxiv.org/abs/2505.08245.

[76] HUANG J, WANG Wenxuan, LI E J, et al. Who is ChatGPT? Benchmarking LLMs' psychological portrayal using PsychoBench[EB/OL]. (2023−10−02)[2025−04−16]. http://arxiv.org/abs/2310.01386.

[77] KARINSHAK E, HU A, KONG K, et al. LLM-GLOBE: a benchmark evaluating the cultural values embedded in LLM output[EB/OL]. (2024−11−09)[2025−08−20]. http://arxiv.org/abs/2411.06032.

[78] YE Haoran, XIE Yuhang, REN Yyuanyi, et al. Measur-

ing human and AI values based on generative psychometrics with large language models[C]//Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2025: 26400−26408.

[79] CERON T, FALK N, BARIĆ Ana, et al. Beyond prompt brittleness: evaluating the reliability and consistency of political worldviews in LLMs[J]. Transactions of the association for computational linguistics, 2024, 12: 1378−1400.

[80] LIM M Y, LOPES J D A, ROBB D A, et al. We are all individuals: the role of robot personality and human traits in trustworthy interaction[C]//2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Napoli: IEEE, 2022: 538−545.

[81] SABOUR S, LIU Siyang, ZHANG Zheyuan, et al. EmoBench: evaluating the emotional intelligence of large language models[EB/OL]. (2024−02−19)[2024−12−09]. http://arxiv.org/abs/2402.12071.

[82] HODSON N, WILLIAMSON S. Can large language models replace therapists? evaluating performance at simple cognitive behavioral therapy tasks[J]. JMIR AI, 2024, 3: e52500.

[83] TSENG Y M, HUANG Yuchao, HSIAO T Y, et al. Two tales of persona in LLMs: a survey of role-playing and personalization[EB/OL]. (2024−06−03)[2024−11−28]. http://arxiv.org/abs/2406.01171.

[84] LI Yuan, HUANG Yue, WANG Hongyi, et al. Quantifying AI psychology: a psychometrics benchmark for large language models[EB/OL]. (2024−06−25)[2025−08−20]. http://arxiv.org/abs/2406.17675.

[85] JI Ke, LIAN Yixin, LI Linxu, et al. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning[C]//Findings of the Association for Computational Linguistics: ACL 2025. Vienna: ACL, 2025: 26221–26238.

[86] DEWEY D. Learning what to value[C]//Artificial General Intelligence. Berlin: Springer Berlin Heidelberg, 2011: 309−314.

[87] ZHU Minjun, WENG Yixuan, YANG Linyi. Personality alignment of large language models[EB/OL]. (2024−08−21)[2025−02−18]. http://arxiv.org/abs/2408.11779.

[88] SZYMANSKI A, ZIEMS N, EICHER-MILLER H A, et al. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks[C]//Proceedings of the 30th International Conference on Intelligent User Interfaces. Cagliari: ACM, 2025: 952−966.

[89] LI Zhiyuan, LIU Hong, ZHOU Denny, et al. Chain of thought empowers Transformers to solve inherently serial problems[EB/OL]. (2024−02−20)[2025−05−27]. http://arxiv.org/abs/2402.12875.

[90] LU Keming, YU Bowen, ZHOU Chang, et al. Large language models are superpositions of all characters: attaining arbitrary role-play via self-alignment[EB/OL]. (2024−01−23)[2025−05−27]. https://arxiv.org/abs/2401.12474.

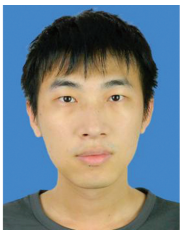
[91] ZHENG Junhao, QIU Shengjie, SHI Chengming, et al. Towards lifelong learning of large language models: a survey[EB/OL]. (2024−06−10)[2025−05−27]. http://arxiv.org/abs/2406.06391.

[92] WANG Qian, WU Jiaying, TANG Zhenheng, et al. What limits LLM-based human simulation: LLMs or our design?[EB/OL]. (2025−01−15)[2025−05−27]. http://arxiv.org/abs/2501.08579.

[93] RIEMER M, ASHKTORAB Z, BOUNEFFOUF D, et al. Position: theory of mind benchmarks are broken for large language models[EB/OL]. (2024−12−27)[2025−05−27]. http://arxiv.org/abs/2412.19726.

作者简介：

柴春雷，教授，博士生导师，浙江大学现代工业设计研究所副所长，中国机械工程学会工业设计分会常务委员兼总干事，中国好设计商业模式评审组组长。主要研究方向为文化创新设计、商业创新设计、智能设计。主持国家社科基金艺术学项目、中国工程院重大咨询项目子课题、国家自然科学基金项目。发表学术论文 20 余篇，出版著作 5 部、教材 1 部。E-mail：dishengchai@126.com。



葛智超，博士研究生，主要研究方向为自然语言处理。E-mail：gezhichao@zju.edu.cn。

殷敏，特聘副研究员，主要研究方向为社媒数据与用户画像、数字人格、人格计算。主持 2024 年教育部海外博士后引才专项资助项目。E-mail：amyin@zju.edu.cn。