doi. 10. 3724/1005-0566. 20250715

“听其言,观其行”:大模型价值倾向表征的多重进路

吕立远[1,2],李延昊[1,2],王健骁[1,2],魏钰明[1,3],苏　竣[1,2,3]

(1. 清华大学公共管理学院,北京　100084;2. 清华大学科教政策研究中心,北京　100084;
3. 清华大学智能社会治理研究院,北京　100084)

摘　要:随着大模型智能水平的提升,越来越多的观点开始将其视为一类准社会主体,驱动着以智能体为对象的机器行为研究的兴起。其中一个重要问题是大模型正在不同维度上展现出与人类相似的价值倾向性。精准识别大模型的价值倾向性,有助于发展面向智能体的行为科学,并为更加高效、安全的人机交互奠定基础。现有文献已从不同视角开展大模型价值倾向性的研究探索,但尚未形成系统的方法论体系。通过综合借鉴行为科学、多模态信息分析、博弈论等交叉学科研究进展,构建基于“分析尺度—表征逻辑”的类型学分析框架,阐释大模型价值倾向性研究的4个代表性进路,并借助每个路径中的研究实例,详细剖析大模型价值倾向性的表征方法,有助于进一步深化对大模型价值倾向性的系统性认识,为丰富和拓展面向智能体的行为科学提供新的方法论视角。

关键词:大模型,生成式人工智能,价值倾向性,智能体,行为科学

中图分类号:TP18;C931　　**文献标识码**:A　　**文章编号**:1005-0566(2025)07-0169-11

From discourse to behavior: multiple approaches to representing the value system of large language models

LÜ Liyuan[1,2], LI Yanhao[1,2], WANG Jianxiao[1,2], WEI Yuming[1,3], SU Jun[1,2,3]

(1. *School of Public Policy and Management*, *Tsinghua University*, *Beijing* 100084, *China*;
2. *Center for Science*, *Technology & Education Policy*, *Tsinghua University*, *Beijing* 100084, *China*;
3. *Institute for Intelligent Society Governance*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: As large language models (LLMs) continue to advance in intelligence, they are increasingly perceived as quasi-social agents, spurring the emergence of machine behavior studies centered on intelligent systems. A key issue in this domain is the extent to which LLMs exhibit human-like value alignments across various dimensions. Accurate identification of such tendencies is essential for developing a behavioral science of intelligent agents and fostering safer, more effective human-AI interaction. Existing literature has initiated exploratory inquiries from multiple perspectives, yet a systematic methodological framework remains lacking. This paper contributes to the academic discourse on value alignment in LLMs by proposing a typological framework based on the axes of “analytical scale” and “representational logic.” Through the analysis of four representative research approaches, the paper delineates core methodologies for examining value tendencies in LLMs. The findings aim to advance systematic understanding in this area and offer a novel methodological lens for the behavioral science of emerging LLM agents.

Key words: large language models; value alignment; typological framework; methodological approaches; machine behavior

收稿日期:2025-03-30　　**修回日期**:2025-07-09

基金项目:新一代人工智能国家科技重大专项“人工智能社会实验伦理、评估与标准化研究”(2023ZD0121600);国家自然科学基金青年项目“声誉管理对政府数智化转型公众接受度的影响机制与路径研究”(72204138)。

作者简介:吕立远(1996—),男,安徽马鞍山人,清华大学公共管理学院博士研究生,研究方向为智能社会治理。通信作者:魏钰明。

一、研究背景

以 ChatGPT 等应用的涌现为标志，大模型（large language model，LLM）正掀起新一轮人工智能（AI）热潮。特别是 2025 年以来，以 DeepSeek 为代表的本土 LLM 凭借成本低廉、自主可控等优势迎来爆发式增长，“AI 公务员”等垂直应用场景迅速形成，推动 LLM 社会应用的极大拓展。“人类—机器”“机器—机器”互动逐渐成为传统人际互动之外重要的社会互动模式[1]，推动信息社会向智能社会的系统性跃迁[2]。

在此背景下，越来越多的学者认为，不仅要将 AI 视为一类服务社会科学的工具，更要将其视为一类具有准人格属性的新兴研究对象，对其认知和行为规律开展系统性研究，构建面向 AI 智能体的社会科学理论[3]。2019 年，*Nature* 发表长文《机器行为学》，呼吁研究者关注 AI 作为一类独立社会主体的行为规律，更加科学地阐释 AI 引入社会的成本、收益与价值权衡[4]。2023 年以来，围绕 LLM 与人类是否在行为上具有相似性等问题，学者开展了一系列“图灵测试”，发现尽管 LLM 的行为多样性通常弱于人类，但在学习能力、情境感知和风险规避等方面与人类高度相似[5]。上述研究表明，LLM 在某些任务结构下的行为输出具有与人类相比较的可能性，大大增强了其社会应用的潜力。

这一判断并不意味着 LLM 已具有人类社会行为形成的生理基础，而更多遵循图灵测试的逻辑，从结果层面强调其已在不同场景中表现出与人类高度相似的行为。进一步地，如果行为层面的类比成为可能，随着 LLM 广泛嵌入社会生产生活，一个自然的问题产生：以人类为参照，与我们“朝夕相处”的 LLM 更接近一个怎样的人？科学认知 LLM 表现的行为特质，有助于塑造人类对智能技术的理性预期，为更加高效、安全的人机交互奠定心理基础[6]。此外，LLM 的产出正深刻嵌入人类社会互动中，其蕴含的微妙倾向性可能通过改变人类的感知、情感和社会判断，放大人类固有认知偏误，且这种作用往往更隐蔽[7]。要进一步识别此类 LLM 可能引致的社会风险，对其行为表现出的多维度特质进行科学表征发挥着基础性作用。

遵循上述逻辑，近年来围绕 LLM 在不同场景中展现的行为特质的研究激增。由于认知交互是 LLM 区别于传统人工智能工具的突出特性，此类文献特别关注了 LLM 的价值倾向性。研究发现，LLM 在政治[8-9]、经济[10]、社会[11]、文化[12]、管理[13]等维度均展现出与人类相似的价值判断。这意味着 LLM 可能并不是价值中立的“知识权威”，而同样具有鲜明的价值倾向[14]。这一过程中，实验室实验[8]、审计实验[11]和数据挖掘[15]等方法均被用以对大模型的价值倾向进行表征，但相关文献侧重于特定任务和场景的价值表征，尚未建立起一般意义上的方法论框架。

本文尝试进一步发展对 LLM 价值倾向性的研究。通过综合借鉴行为科学、博弈论、多模态信息分析等交叉学科研究成果，构建一个系统的方法论框架，阐释不同方法的研究逻辑，并通过具体案例探讨可能出现的问题。本文旨在回答以下问题：①LLM 价值倾向性表征有哪些方法路径？应当如何实现？②不同路径的逻辑何在，可能面临哪些挑战？

与人类价值观研究相类似，LLM 价值倾向性研究也应当秉持全面、系统的观点，从个体、群体等不同尺度出发，充分利用 LLM 强大的交互能力，既要“听其言”，详细分析 LLM 对特定价值观点的显性陈述，更要“观其行”，分析 LLM 实际行为背后蕴含的隐性倾向。本文的贡献在于，通过构建一个“分析尺度—表征逻辑”的类型学框架，系统阐释 LLM 价值倾向性表征的方法论逻辑，揭示 LLM 价值倾向性浮现的多重路径，进一步深化对 LLM 价值倾向性的认识，丰富和拓展面向智能体的行为科学，为后续建立面向智能体的社会科学理论体系奠定基础。

本文剩余部分按以下结构展开。第二部分构建了“分析尺度—表征逻辑”的类型学框架，从理论上阐释 LLM 价值倾向性表征的多重进路。第三部分进一步依据前述框架，详细剖析 4 条进路的研究逻辑与适用场景，并给出示例。第四部分对全文进行总结并提出未来研究方向。

二、LLM 价值倾向性表征的方法框架

现有文献已从不同路径对 LLM 的价值倾向性问题进行了探索。为更好地理解这种方法层面的多元性，本文基于类型学范式，构建了一个 2×2 分析框架，系统呈现 LLM 价值倾向性表征的方法逻辑。类型学是方法论研究的经典路径，能够兼顾

对抽象性与复杂度的关注[16-17]。如图1所示,本文构建的框架包含分析尺度和表征逻辑两个维度,其中“分析尺度”分为模型个体和群体,“表征逻辑”分为显性和隐性倾向。上述维度的建构遵循Rahwan等[4]的建议,重点吸收了动物和生物行为研究,特别是以人类为样本的社会科学研究的成果。

该框架的构建有3个核心考虑。首先,应满足“独立且穷尽”的基本要求[18]。本文构建的两个维度是充分正交的,两者结合能对任何一类LLM价值倾向性研究进行清晰定位,不存在重叠。其次,应充分体现路径的差异性。这一点将在第三部分进一步印证。最后,应充分呼应LLM价值倾向性研究的发展趋势,契合智能体社会科学研究从虚拟到“沙盒”环境,再到真实世界的发展过程。

		表征逻辑	
		显性倾向	隐性倾向
分析尺度	模型个体	①	②
	模型群体	③	④

图1　LLM价值倾向性研究路径的“分析尺度—表征逻辑”类型学框架

(一)维度1:分析尺度

维度1主要解决在何种尺度上分析LLM价值倾向性的问题。本文主要从模型个体和群体两个层次出发。Rahwan等[4]指出,智能体的行为研究本质上与生物行为研究类似,应综合考虑行为体内在特性和外在环境影响。显然,个体内在特性是分析LLM价值倾向性的起点。参照以人类为样本的社会科学研究,世界价值观调查(world value survey, WVS)等项目已对个体层面的价值倾向性表达进行了丰富的探索[19]。相关工作可以迁移到LLM层面[8]。

与此同时,亦有文献强调个体之上的“群体”为理解大模型价值倾向性提供了新的层次。一方面,现实中个体往往以不同方式聚合为具有不同边界的群体,共享一部分组内一致性较高而组间一致性较弱的属性。例如,Hofstede[20]以个体价值观调查为基础,对不同国别的价值观聚合特征进行分析,提出国家文化的类型学框架,对后续研究产生了深远影响。另一方面,群体内部的互动与权力关系也可能影响个体的隐性价值表达。例如,Milgram[21]的“服从实验”表明,个体在群体压力下的行为模式能够很好地折射其服从性。因此,大模型在群体互动中表现出的行为特征也具有重要意义。由于当前LLM直接进入物理空间的应用还比较缺乏,本文暂时省略对LLM与现实环境互动关系的讨论。

(二)维度2:表征逻辑

维度2主要关注从何种视角表征LLM的价值倾向性。随着智能水平的提升,LLM已能以接近人类的方式执行多种任务。但即便在相同主题下,问题呈现方式的变化,亦可能显著影响LLM的价值表达。例如,有文献发现,LLM在面对同一问题时,基于李克特量表的直接评分结果,往往与其在开放式自然语言回应中的价值立场存在差异,表现出类似人类的“言行不一”现象[22]。为更好地突出这一问题,本文引入显性与隐性倾向加以区分。其中,显性倾向指LLM对特定价值命题的直接判断,如对特定立场语句进行评分或选择,即“听其言”。隐性倾向指LLM未直接回应价值命题的情况,需通过多模态输出逆向识别其潜在价值倾向,即“观其行”。

三、LLM价值倾向性表征的多重进路

本章将详细解释上述框架中4个象限所代表的LLM价值倾向性研究路径。对每个象限的论述包含每类路径的基本逻辑、实施方法与潜在优劣。在此基础上,对每一类路径给出示意性研究案例。

(一)模型个体层面的显性价值倾向

显性价值倾向分析的逻辑是将适用于人类的心理量表迁移到LLM。由于LLM本身具有很强的交互输出能力,通过分析LLM对特定价值陈述的显性态度,有助于研究者剖析其价值倾向。该路径依赖于一定的前提假设。首先,LLM具有稳定且可表征的价值系统。这一假设主要是结果导向的,不要求LLM具有与人类价值认知形成相似的生理过程,而主要关注其能够在结果层面输出与人类相似的价值倾向。越来越多的文献为该假设提供了积极证据。例如,Acerbi等[23]发现ChatGPT在生成内容时表现出与人类高度相似的偏见。随着LLM智能水平的提高,GPT-4等模型也开始拥有对底层人格的模仿能力。Wang等[24]发现GPT-4已经能够准确模拟大五人格。基于此,可以认为LLM正逐渐形

成与人类相似的价值认知。其次,这些内隐的价值倾向是一种可操作化的心理结构[25]。一系列研究探索了利用问卷输入捕捉 LLM 内隐倾向的过程,发现能够有效地测量 LLM 对政治、社会等不同维度的自动感知[26]。因此,从结果层面来看,LLM 作为“硅基生命”的逻辑预设正日益得到支持。

需要说明的是,LLM 的特殊性使得此类研究需要关注一些问题。首先,与 LLM 的交流依赖“提示词”作为通用语言。然而,LLM 本身对提示词框架具有敏感性。因此,研究者需要选取相对稳健的提示词框架,并在条件允许时选择多种框架进行稳健性检验。其次,LLM 的底层逻辑是概率模型,其输出天然地存在随机性。在开展实验时需考虑随机性的影响,包括量表的随机排序与多次重复测量等。最后,受安全规制等因素的影响,LLM 可能以特定概率拒绝回答某些问题,因而需要精心设计提问方式。

本节以政治世界主义(国家主义)量表为例[27],说明模型个体显性价值倾向分析的基本流程。首先,对英文量表进行本土化修订和优化,确定测量的基础题项。其次,以 CO-STAR 框架为基础,进行提示词设计,借助“少样本思维链”对 LLM 进行训练[28]。最后,以吕立远等[29]构建的 LLM 样本集为基础,使用 Openrouter 平台调用相关 API 接口。设置温度参数为 1,尽可能暴露模型价值认知的多样性,对每个样本进行 100 轮重复测试。每次测试前,使用伪随机数算法生成问题次序。由于不同 LLM 对特定问题的敏感性可能存在差异,部分模型可能存在拒答现象。为保证分析尺度的统一,本案例在删除缺失结果后,计算每个题目最终得分的平均值和方差。统计显示,平均拒答比率在 13% 左右,占比相对有限。初步分析结果如图 2 所示。图 2 揭示了相关 LLM 在政治世界主义(国家主义)层面的价值倾向性存在显著差异。这不仅体现在倾向性的绝对水平上,也体现在稳定程度上。

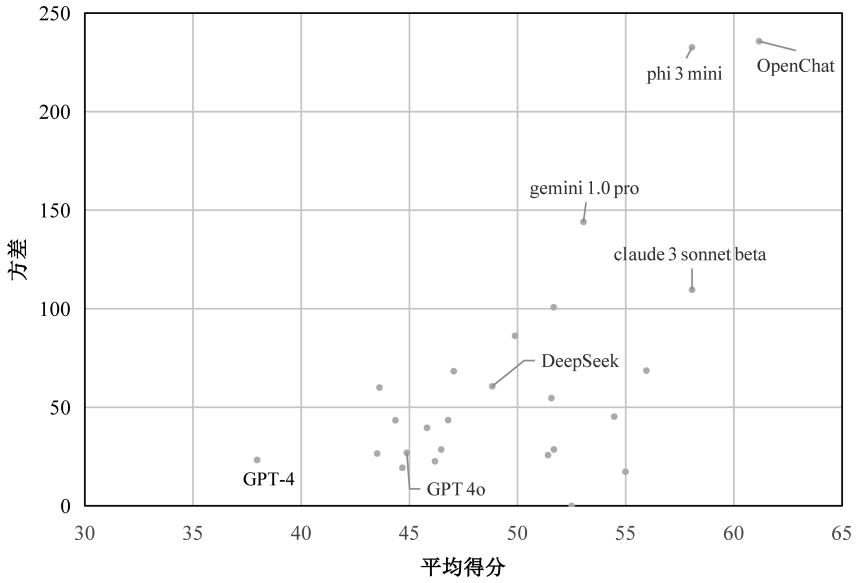

图 2　全球 30 个主流 LLM 的政治世界主义(国家主义)价值倾向

本节展示了对单个 LLM 进行显性价值倾向分析的流程,但仍有若干问题值得探讨。一是如何更好地应对 LLM 输出的随机性。为此,本文的思路是进行重复测试并取平均处理。这种方法的优势是将随机性视为 LLM 的内在特性,但可能面临较高的计算成本。Song 等[30]采用了相反的思路,通过贪婪解码的方式将模型 Temperature 参数设为 0,对 LLM 的随机性进行限制。此时,研究者不再需要重复测试,但可能失去对随机性本身的分析空间。二是与 LLM 交互的方式。本案例主要采用 API 与 LLM 交互。这种方法具有体量小、成本可控等优势,但实际操作中可能面临不稳定等问题。特别是随着研究的深入,一旦涉及对特定价值倾向的微调,本地部署将具有更大优势。三是 LLM 价值倾向性的动态演变。现有文献发现,LLM 的价值倾向处在快速的演化过程中,既受到数据拓

展以及持续交互的影响,也与算法调整等宏观因素有关[31]。因此,研究者一方面需要在宏观上保持对 LLM 价值倾向性的持续追踪,另一方面需要调用更多 API,避免长时间在同一窗口中对话可能诱发的记忆效应。

(二)模型个体层面的隐性价值倾向

基于心理量表对单个 LLM 的显性价值倾向性进行测量的“听其言”,虽具有直观便捷的优势,但也面临诸多挑战。受到安全对齐等机制的影响,LLM 可能对少数族裔等敏感议题产生系统性回避[32]。且价值倾向本身是一个多维度的复杂系统,某些维度可能难以用自然语言清晰表达。此时,即使 LLM 勉强生成了相关回复,其表征效度也可能不够理想[33]。因此,将“观其行”与“听其言”有机融合,进一步拓展从实际行为模式中推断 LLM 隐性价值倾向的方法路径具有重要意义。

在工程科学领域,上述思路又被称为“逆向工程”。逆向工程是一种从结果或成品出发,通过分析其结构、功能与运作原理,推导其设计过程或原始意图的方法。逆向工程主要适用于原始信息不充分或不可得的场景,这与 LLM 的“黑箱”属性高度契合。通过分析 LLM 输出的语义特征,可以逆向推理其隐性价值倾向。这一领域的早期研究主要集中在文本模态的输出[34]。但“一图胜千言”,与文本相比,图像往往蕴含着更丰富的信息。随着多模态 LLM 的发展,文生图等跨模态转换背后隐性价值倾向也日益受到学术界重视。

现有文献已经证明文生图 LLM 在性别、职业等方面存在鲜明的价值倾向,甚至偏见[34-36],但仍然存在一定的局限性。首先,部分文献遵循计算科学传统,先把图像进行向量嵌入,进而计算不同向量间距离作为偏见表征[35,37]。这种方法更多测量的是不同图像的“差异”,忽视了“偏见”是社会比较过程中产生的指向性结果。其次,部分文献关注到“偏见”应当以社会现实为参照,但此类研究大多基于人工识别,侧重描述图像生成内容与真实数据的差异,缺乏对偏见的因果推断[38]。基于上述问题,本节旨在进一步拓展这一领域的学术探讨。

本小节以 DeepSeek 的多模态大模型 Janus 为例,展示如何从因果层面评估 LLM 的隐性价值倾向。作为一个示意性案例,本文主要聚焦于“地域—年龄”偏见。基于现有文献,地域偏见指 LLM 在生成特定地区相关内容时,由于训练数据不平衡等因素,有关某些地区的内容输出中出现系统性偏向[39]。本节主要关注 Janus 是否会在生成图像时系统地将经济水平更高地区的人呈现得更加年轻?

遵循逆向工程思路,本节通过分析生成图像的内容特征,逆向推理其背后的地域偏见。本案例首先对 Janus-Pro-7B 进行本地部署。作为一项对 LLM 深层价值偏见的探索性研究,本案例选用最为简化的提示词模式,旨在减少外在约束,充分暴露 LLM 内部的深层次认知结构。提示词仅有[地区]存在差异,实验中分别填入中国 31 个省级行政区,为每个地区生成 100 张图片,得到一个 3 100张图片的数据集。这一设定考虑了大样本分析($N>30$)和计算成本等因素。同时,同类文生图价值偏见研究每次实验平均生成的图片数大多在 10 张[40-41]。因此,上述设定能够实现比同类文献更强的稳健性。

与颜色等特征可以直接通过像素运算得出不同,年龄本身属于复杂的认知特征[42]。传统图像分析在识别认知特征上往往面临挑战。近年来,随着计算机视觉的发展,一系列集成模型为上述问题提供了解决方案。本小节调用人脸识别模型 FairFace 对生成图像的年龄进行识别[43]。与其他主流人脸模型相比,FairFace 的一个优势是重点兼顾了人种的平衡,能够提供更加稳健的识别效果。对于每个省份生成的图片,我们记录了其识别年龄的平均数。图 3 展示了本案例的分析结果。其中,横轴为 31 个省级行政区 2024 年初步核算的对数人均 GDP,而纵轴为生成图像识别的平均年龄。由于地区经济水平可能影响其劳动力年龄结构,本案例基于中央财经大学发布的中国省级劳动力人力资本数据库,进一步控制了省级劳动力平均年龄[44]。回归结果表明,随着人均 GDP 增加 1%,该省份生成的 100 张图片的平均年龄将下降 0.115 岁。更加直观地来看,以对数人均 GDP 为 12.3 的省份(近似北京)为主题生成的劳动者图像要比以对数人均 GDP 为 11 的省份(近似贵州)生成的图像视觉上年轻 15 岁左右。这表明,文生图 LLM 可能系统性地将经济发达地区劳动者呈现为

更加年轻的视觉形象,体现一种隐性的审美偏向[45]。进一步拓展分析表明,上述现象主要由经济较发达的前 1/2 地区劳动者被尤为明显地年轻化所致,文生图 LLM 存在普遍的性别偏见和“地理—职业”复合偏见①。

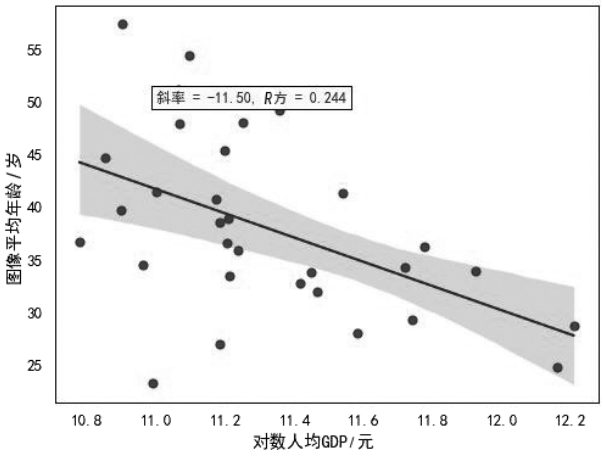

图 3　省级人均 GDP 与生成图像年龄的关系

需要说明的是,本案例主要展示从单个 LLM 的行为出发,逆向推理其隐性价值倾向的方法路径。实践中不应局限于经济水平、年龄、性别和职业等变量。未来研究仍应当以拓展对价值倾向的因果识别为重点,结合分层随机抽样等实验设计手段,一方面进一步探索文生图、文生视频等复杂模态中不同维度的隐性价值偏向,另一方面通过融合深度学习和异质性干预效应等模型,进一步揭示多重因素如何影响 LLM 的隐性价值倾向。

(三)模型群体层面的显性价值倾向

前述象限将分析单元确定在单个 LLM 层面。但“群体”也是 LLM 价值倾向研究的一个重要维度[46]。首先,现实世界中的社会群体是个体以不同方式和边界聚合形成的。这一过程中,群体逐渐涌现出部分组内一致性较高,而组间一致性较低的共性特征,为社会科学研究提供了新的视角。这一领域的典型案例是 Hofstede 从个体价值观出发,通过分析不同国别个体的价值观聚合特征,提出了国家文化的类型学划分[20]。实践中,每个 LLM 对特定价值命题均能够表现出不同的显性倾向。遵循上述逻辑,如果将它们按照特定的边界进行聚合,此时又将得到哪些新的结论?

考虑到全球 LLM 发展和监管模式差异日趋凸显,开发者所在国别成为当前本领域一类重要的“群体”划分[47]。一方面,LLM 的训练过程源于现实,是特定国家社会现实的形塑,很大程度上可以被视为一种包含并嵌入社会实践的科学知识[48]。然而,LLM 本身具有复杂智能机器的涌现特性,不能将其行为视作训练语料的线性镜像。不同 LLM 行为反映社会现实的程度可能存在差异。另一方面,不同国家的安全监管也对 LLM 行为进行了约束,引导 LLM 根据符合当地价值标准的方式进行输出[49]。在上述背景下,对于特定国家 LLM 的总体价值倾向进行系统性评估,不仅有助于观察当前全球数字文明与物理文明的分野,也能够为当前 AI 监管政策的效果评估提供微观证据。

本案例以 WVS 的社会价值观作为示意性案例,说明如何进行模型群体层面的主观态度分析。本案例主要关注中国和美国开发的 LLM 的差异,采用与第一象限相同的实验框架、样本集合和实验参数。实验中,首先载入 WVS 的对应题项,获取模型反馈。其次,剔除缺失值并按照国家进行聚合,获得不同国家 LLM 在特定维度上的平均得分与方差分布。考虑可视化效果,本案例使用气泡图进行呈现,其中颜色深度反映正向支持该观点的程度,气泡半径则反映得分的方差。在此基础上,本案例进一步进行了组间均值差异检验,主要分析结果如图 4 所示。

图 4 的分析结果与前文的探讨一致。不同国家开发的 LLM 之间既呈现出部分一致的显性价值倾向,又存在明显的区别。一方面,中国与美国开发的 LLM 对社会变革、社会信任等问题均具有正向倾向,而对于社会控制具有负向倾向。在上述维度上,中美 LLM 的价值倾向是高度一致的。另一方面,中美 LLM 在若干价值维度上也存在显著的统计差异。例如,在社会控制维度上,具有高度

① 限于文章篇幅,相关提示词及拓展分析细节省略。

自由主义传统的美国所开发的 LLM 表现出更强的反社会控制倾向。在普世道德感知维度上,美国 LLM 也表现出与本土政治叙事较为贴合的倾向。

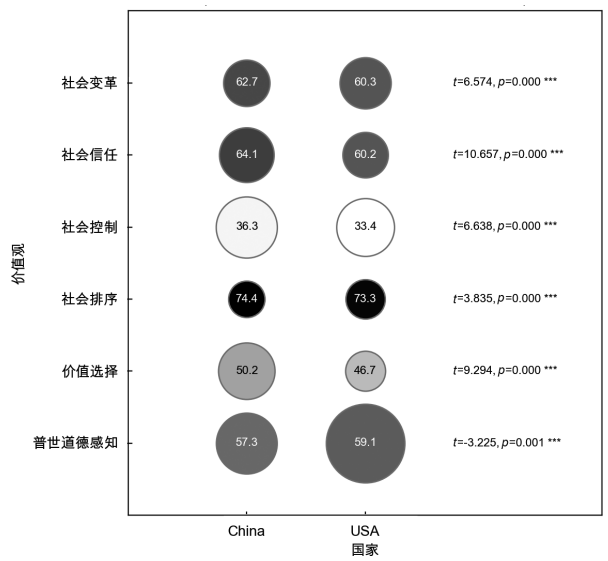

图4　中国与美国开发的 LLM 在社会价值观维度的分析

注:右侧为两国模型得分的均值差异 t 检验,* 代表显著性水平,$^{*}p<0.1$,$^{**}p<0.05$,$^{***}p<0.01$。

需要注意的是,由于当前主流 LLM 大都尚未完全公开训练语料信息,本案例更多希望强调不同国别 LLM 在特定价值维度上表现出了系统的群体性差异,尚不足以解耦训练语料、社会经济环境等对 LLM 价值表达的影响机制。考虑到 LLM 是否将带来新的“数字鸿沟”已成为全球数字治理的焦点,未来对这一象限的研究应系统考虑不同国家的制度背景、政治叙事和政策偏好如何影响 LLM 的价值倾向性。此外,随着 LLM 的发展,还可以考虑进一步采用纵向过程追踪等方法,基于技术社会系统分析框架,探究制度与社会环境与 LLM 价值倾向的互动,明晰 LLM 何以成为文化与制度的中介载体,并透过 LLM 进一步观察智能社会全球文明体系的分野与合流。

(四)模型群体层面的隐性价值倾向

LLM 的价值倾向性也可能在多个模型的交互行为中自然涌现[50]。社会建构理论指出,群体中的个体行为并不是孤立存在的,不能简单视为情绪和冲动的反应,而是个体与环境相互作用塑造的结果[51]。因此,个体在群体互动中呈现的行为特质能够很大程度反映其对公平性、合作意识等价值规范的认同程度。这一问题属于前文提及的隐性价值倾向。与第二象限的分析不同,此时研究者关注的不再是 LLM 个体行为,而是嵌入在社会关系中时 LLM 会做出何种决策,并通过实际决策行为反推其隐含价值倾向。

自 20 世纪 80 年代起,丹尼尔·卡尼曼(Danie Kahneman)和弗农·史密斯(Vernon Smith)等开始基于实验室实验方法,通过让人类在受控环境下进行独裁者博弈、公共物品博弈等博弈游戏,对个体的公平、合作、互惠等偏好进行测量[52-53]。这些实验证明了经典理性人假设的局限性,为深化对人类行为规律的认识作出了重要贡献,也为研究 LLM 的价值倾向性提供了可供借鉴的成熟范式[54]。

近年来,越来越多的学者开始将上述研究范式由人类推广到 LLM。通过将 LLM 带入经典的博弈游戏,研究者能够系统观察其在不同决策环境中的行为反应,进而对其背后的价值倾向进行推断。例如,Aher 等[55]以最终通牒博弈为场景,发现 LLM 能够模拟人类样本对于公平性的价值偏好。Horton[56]以独裁者博弈为场景,发现 LLM 在特定情况下能够模拟人类样本在公平、效率和自利等价值偏好间的权衡。LLM 智能水平的提升也推动着越来越多的学者开始超越简单的实验室环境,将其应用到外交博弈、商业谈判等复杂决策环境中。例如,Ouyang 等[57]通过模拟公司内部的投资博弈,探讨关于无害性、有益性、诚实性的 3H 对齐如何影响 LLM 在投资决策中的表现。结果显示,过度对齐可能导致模型预测过于谨慎,而适度对齐下模型投资表现最为出色。Huang 等[58]将 LLM 应用于外交谈判场景,发现 LLM 能够成功模拟公平性、合作性和自利性等人格特质对谈判结果的影响。这些工作表明,虽然 LLM 社会应用仍处于初期阶段,但凭借其智能水平方面的突出优势,有望进一步推动传统实验室实验在智能时代取得更大的发展,为研究者理解 LLM 价值倾向性提供新的视角。

本文通过一个独裁者博弈案例,运用结构估计方法,展示如何分析 LLM 在群体互动中的价值

倾向表达。实验中,每个 LLM 需要将 100 元分配给自己和另一个虚拟主体,其中虚拟主体仅是接受者,没有能力影响决策。需要注意的是,虽然独裁者博弈本身具有高度简洁的结构,但虚拟主体仅是接受者的设定已嵌入了明显的权力关系意蕴。本案例基于 OpenRouter 平台,调用 DeepSeek 等 6 个代表性 LLM 的 API 接口,使用结构化提示词分别进行 100 次实验,设置温度参数为 1,每次实验均详细记录其分配决策。

分析结果表明,不同 LLM 给予他人的金额存在较大差异。此时,我们并不能简单地用平均数等对模型的偏好结构进行推理,而需要考虑个体的基本效用结构。作为一个示意性案例,我们从独裁者博弈基本形式出发,考察大模型呈现出的自利(Self-interest)偏好。假定个体分配给自己的金额为 g,留给接受者的金额为 $(100-g)$。$b \in [0,1]$,其取值越高,反映个体越自利。

本节主要考虑了经济学理论中经典的常替代弹性(constant elasticity of substitution, CES)效用结构,其定义如下所示。CES 效用的突出优势是具有很强的灵活性,在 r 取值趋向于 0、等于 1、趋向于负无穷大时,分别可以表示柯布—道格拉斯效用、线性效用和里昂惕夫效用。本文选择 $r = 0.5$ 这一适中情形进行研究。此时,利己与利他行为间具有适中的相互替代关系,同时又保留了模型的非线性结构。

$$U(g)_{CES} = (b \cdot g^r + (1-b) \cdot (100-g)^r)^{1/r}$$

本案例借鉴 Mei 等[5]的方法估计 b 的最优取值。该方法旨在找到一个 b,使得独裁者的理论最优策略与实际进行的 100 次采样结果的误差平方和最小。其中,平均优化误差(average optimization error, APE)定义为:

$$APE(b) = \frac{1}{100}\sum_{i=1}^{100}\left[1 - \frac{U(g_i)}{U(g^*(b))}\right]^2$$

式中,$U(g_i)$ 是实验中第 i 次采样的具体决策结果带来的效用;$U(g^*(b))$ 是给定 b 下大模型的理论最优策略所带来的效用。

图 5 展示了 6 个样本模型自利偏好的估计结果。现有文献指出,人类样本进行独裁者博弈时的自利偏好集中在 0.6 附近[5]。本文的分析结果表明,除 Claude-3-Haiku 呈现出更加明显的自利偏好外($b = 0.75$),其余 5 个大模型表现出的自利偏好程度总体上与人类样本相当。这一结论进一步验证了本文的逻辑预设,即大模型在结果层面确实呈现出与人类相似的价值倾向。进一步考虑社会网络理论中的社会距离变量进行拓展分析,发现随着社会距离缩小,LLM 表现出的自利偏好也有所降低。这一结论与人类样本总体保持一致②。

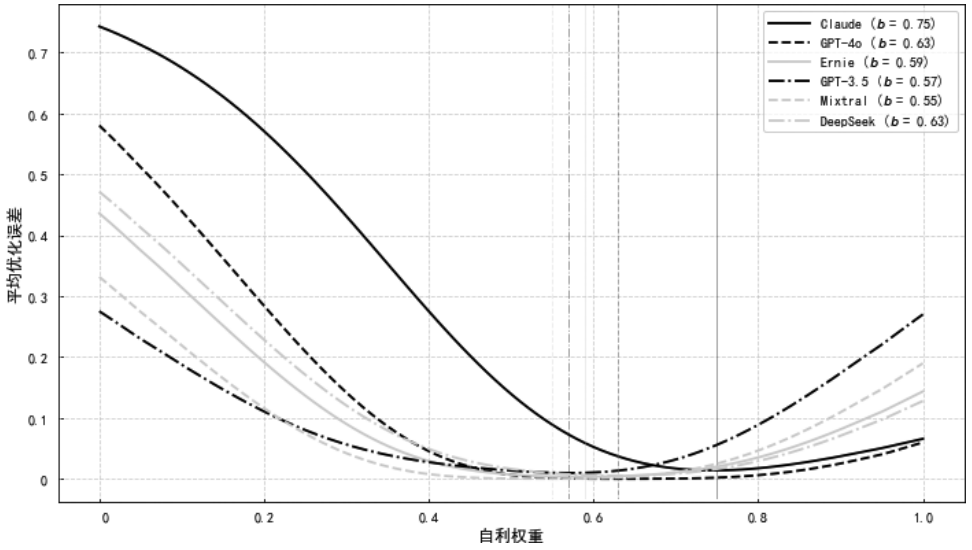

图 5 CES 效用结构下 6 个模型样本的自利偏好估计

② 限于文章篇幅,相关提示词及拓展分析细节省略。

上述案例展示了群体互动情境下 LLM 价值倾向性分析的基本路径。未来,本象限的研究工作还可以从以下方面进一步深化。第一,进一步提升实验环境的复杂性,在独裁者博弈等简单博弈游戏的基础上,拓展多主体协同决策等复杂场景,深入考虑多轮次、信息不完备、社会学习、外部信息注入以及不同训练和价值对齐策略等的影响。第二,考虑其他不同类型的效用函数设定。在同一组实验数据下,对于偏好结构的不同设定也可能产生不同的估计结果。第三,进一步推动实验环境从实验室走向真实世界,拓展对外交谈判、金融决策等真实场景的群体互动过程模拟。

四、研究总结

随着智能技术的发展,LLM 正快速逼近,甚至在局部超越人类智能。深圳等地区开始部署 AI 公务员的现象表明,LLM 正成为一类准社会主体,日益广泛地参与到人类生产生活。越来越多的研究发现,尽管 LLM 不具备产生价值认知的生理基础,但其并不是价值中立的,而在结果层面产生了与人类近似的价值倾向。对 LLM 不同场景下可能呈现的价值倾向进行系统表征,及时预警潜在的价值失调风险,有助于增进 LLM 多场景社会应用的稳健性。然而,现有关于 LLM 价值认知的研究还处于早期阶段,尚未形成系统的方法论体系。本文强调 LLM 价值倾向性研究应综合考虑不同尺度与 LLM 的不同类型输出,将“听其言”与“观其行”有机结合。通过构建“分析尺度—表征逻辑”的类型学框架,本文系统剖析了 LLM 价值倾向研究的多重进路,为更好地形成对 LLM 价值倾向的系统认识提供助力。

尽管每个象限分析的末尾已经简要探讨了该路径下需要关注的若干问题,但这一领域中还有诸多共性问题有待探索。

首先,受本领域发展阶段的限制,本文侧重 LLM 多维价值倾向的表征。后续研究应当在此基础上,重点拓展对 LLM 行为机制的因果推断。一是把价值倾向性作为自变量,研究其对 LLM 决策行为的影响,寻找高效预测 LLM 决策行为的价值代理。二是将价值倾向作为因变量。一方面,深入研究复杂社会交互、知识注入、情境设定等因素塑造 LLM 价值倾向的内在机制,解耦不同因素的作用[59]。另一方面,通过事前对齐人类和 LLM 在特定维度的价值倾向,运用 LLM 代替人类接受某些风险性实验干预,模拟部分可能具有伦理风险的人类实验。

其次,本文目前主要遵循拟人逻辑,其核心是探索 LLM 智能体能够在多大程度上表现出人类特有的价值倾向性。由于 AI 发展的终极目标本身是模拟人类智能,运用拟人逻辑对 LLM 价值认知进行评估并无不妥。但与此同时,研究者仍需认识到,LLM 与人类本身是两种不同的“物种”。在总体遵循拟人逻辑的基础上,LLM 与人类固有的行为规律差异同样值得关注。例如,近期在 LLM 代替人类样本参与社会调查的研究中,越来越多的文献发现,不同场景下 LLM 虽然能在总体上有效反映人类样本的均值,但答案在多样性上仍与真人存在很大差异[31]。

再次,目前对 LLM 价值倾向性的评估主要集中在未做约束的通用情境下。由于 LLM 本身具有很强的角色扮演能力,其呈现出的价值认知规律完全可能因场景而异,特定领域内部也可能存在专有的价值认知维度[11]。本文建议未来研究坚持通用与专用相结合的思路,在开展不同领域 LLM 价值倾向性研究时,既要注重研究通用价值维度在不同领域间的一致性与差异性,又要注重特定领域内部具有代表性的价值认知维度,不断深化对 LLM 价值倾向性的理解。

最后,目前对 LLM 价值倾向性的评估主要集中在虚拟环境。然而,随着具身智能等前沿技术的发展,以 LLM 为代表的机器智能体终将走向现实的物理和社会空间。由于 LLM 对外在输入本身具有极高的敏感性,随着感知信息输入的增加,LLM 究竟会做出怎样的价值决策还尚未可知。后续研究应从虚拟世界开始,逐渐过渡到基于“沙盒”的拟真环境,最终过渡到真实世界,推动 LLM 价值认知与行为研究更加接近复杂的现实世界。

参考文献:

[1] KATTEL R, LEMBER V, TÕNURIST P. Collaborative innovation and human-machine networks[J]. Public management review, 2020, 22(11): 1652-1673.

[2]苏竣,魏钰明. 迈向智能社会:现实图景、发展趋向与

治理使命[J]. 西北大学学报(哲学社会科学版),2025,55(1):78-88.
[3]XU R,SUN Y,REN M,et al. AI for social science and social science of AI:a survey[J]. Information processing & management,2024,61(3):103665.
[4]RAHWAN I,CEBRIAN M,OBRADOVICH N,et al. Machine behaviour[J]. Nature,2019,568(7753):477-486.
[5]MEI Q,XIE Y,YUAN W,et al. A Turing test of whether AI chatbots are behaviorally similar to humans[J]. Proceedings of the national academy of sciences,2024,121(9):e2313925121.
[6]MENG J. AI emerges as the frontier in behavioral science[J]. Proceedings of the national academy of sciences,2024,121(10):e2401336121.
[7]GLICKMAN M,SHAROT T. How human-AI feedback loops alter human perceptual,emotional and social judgements[J]. Nature human behaviour,2025,9(2):345-359.
[8]MOTOKI F,PINHO NETO V,RODRIGUES V. More human than human:measuring ChatGPT political bias[J]. Public choice,2024,198(1):3-23.
[9]RUTINOWSKI J,FRANKE S,ENDENDYK J,et al. The self-perception and political biases of ChatGPT[J]. Human behavior and emerging technologies,2024(1):7115633.
[10]FEDYK A,KAKHBOD A,LI P,et al. ChatGPT and perception biases in investments:an experimental study[J]. Available at SSRN 4787249,2024.
[11]LIPPENS L. Computer says "no":exploring systemic bias in ChatGPT using an audit approach[J]. Computers in human behavior:artificial humans,2024,2(1):100054.
[12]YUAN H,CHE Z,LI S,et al. The high dimensional psychological profile and cultural bias of ChatGPT[J]. arXiv preprint arXiv:2405.03387,2024.
[13]CHEN Y,KIRSHNER S N,OVCHINNIKOV A,et al. A manager and an AI walk into a bar:does ChatGPT make biased decisions like we do?[J]. Manufacturing & service operations management,2025,Online.
[14]BAI X,WANG A,SUCHOLUTSKY I,et al. Explicitly unbiased large language models still form biased associations[J]. Proceedings of the national academy of sciences,2025,122(8):e2416228122.
[15]LUCY L,BAMMAN D. Gender and representation bias in GPT-3 generated stories[C]// Proceedings of the third workshop on narrative understanding. 2021:48-55.
[16]YIN R K. Case study research:design and methods[M]. London:Sage,2009.
[17]黄萃,吕立远. 文本分析方法在公共管理与公共政策研究中的应用[J]. 公共管理评论,2020,2(4):156-175.
[18]MINTO B. The pyramid principle:logic in writing and thinking[M]. Pearson education,2009.
[19]ALEMÁN J,WOODS D. Value orientations from the world values survey:how comparable are they cross-nationally?[J]. Comparative political studies,2016,49(8):1039-1067.
[20]HOFSTEDE G. National cultures in four dimensions:a research-based theory of cultural differences among nations[J]. International studies of management & organization,1983,13(1/2):46-74.
[21]MILGRAM S. Behavioral study of obedience[J]. The journal of abnormal and social psychology,1963,67(4):371.
[22]REN Y,YE H,FANG H,et al. Valuebench:towards comprehensively evaluating value orientations and understanding of large language models[J]. arXiv preprint arXiv:2406.04214,2024.
[23]ACERBI A,STUBBERSFIELD J M. Large language models show human-like content biases in transmission chain experiments[J]. Proceedings of the national academy of sciences,2023,120(44):e2313790120.
[24]WANG Y,ZHAO J,ONES D S,et al. Evaluating the ability of large language models to emulate personality[J]. Scientific reports,2025,15(1):519.
[25]HITLIN S,PILIAVIN J A. Values:reviving a dormant concept[J]. Annual review of sociology,2004,30(1):359-393.
[26]LI P,CASTELO N,KATONA Z,et al. Frontiers:determining the validity of large language models for automated perceptual analysis[J]. Marketing science,2024,43(2):254-266.
[27]KOSTERMAN R,FESHBACH S. Toward a measure of patriotic and nationalistic attitudes[J]. Political psychology,1989,10(2):257-274.
[28]KIM S,JOO S J,KIM D,et al. The cot collection:improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning[J]. arXiv preprint arXiv:2305.14045,2023.
[29]吕立远,李延昊,王健骁,等. 大语言模型的价值观研究:概念框架与实证评估[J]. 电子政务,2025(1):15-28.
[30]SONG Y,WANG G,LI S,et al. The good,the bad,and the greedy:evaluation of LLMs should not ignore non-determinism[J]. arXiv preprint arXiv:2407.10457,2024.
[31]LIU Y,PANWANG Y,GU C. "Turning right"? an experimental study on the political value shift in large language models[J]. Humanities and social sciences communications,2025,12(1):1-10.
[32]WANG A,MORGENSTERN J,DICKERSON J P. Large language models that replace human participants can harmfully

misportray and flatten identity groups[J]. Nature machine intelligence,2025:1-12.
[33]ARZBERGER A, BUIJSMAN S, LUPETTI M L, et al. Nothing comes without its world-practical challenges of aligning LLMs to situated human values through RLHF [C]// Proceedings of the 2024 AAAI/ACM conference on AI,ethics, and society. 2024:85-93.
[34]黄阳坤,陈昌凤. 人像政治的算法操纵与社交媒体实践:Twitter 社交机器人涉中国议题图像的计算机视觉分析[J]. 西安交通大学学报(社会科学版),2024,44(6):130-139.
[35]CHO J, ZALA A, BANSAL M. DALL-Eval: probing contextual reasoning in text-to-image generation [C]// Proceedings of the IEEE/CVF international conference on computer vision,2023:2148-2157.
[36]LUO H, HUANG H, DENG Z, et al. BIGbench: a unified benchmark for evaluating multi-dimensional social biases in text-to-image models[J]. arXiv preprint arXiv:2407. 15240v5,2025.
[37]GAL R, ALALUF Y, ATZMON Y, et al. An image is worth one word: personalizing text-to-image generation using textual inversion [J]. arXiv preprint arXiv: 2208. 01618, 2022,2022.
[38]LOCKE L G, HODGDON G. Gender bias in visual generative artificial intelligence systems and the socialization of AI[J]. AI & society,2024:1-8.
[39]MANVI R,KHANNA S,BURKE M,et al. Large language models are geographically biased[C]// International conference on machine learning,2024:34654-34669.
[40]WANG J, LIU X G, DI Z, et al. T2IAT: measuring valence and stereotypical biases in text-to-image generation [J]. arxiv preprint arxiv:2306. 00905,2023.
[41]VICE J,AKHTAR N,HARTLEY R,et al. Exploring bias in over 100 text-to-image generative models[J]. arxiv preprint arxiv:2503. 08012,2025.
[42]余立,吕立远,黄萃. 图像分析在哲学社会科学中的研究路径与展望[J]. 信息技术与管理应用,2023,2(3):144-158.
[43]KARKKAINEN K, JOO J. Fairface: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021:1548-1558.
[44]李海峥. 中国人力资本报告(2024)[R]. 北京:中央财经大学人力资本与劳动经济研究中心,2024.
[45]陈昌凤,师文. 人脸分析算法审美观的规训与偏向:基于计算机视觉技术的智能价值观实证研究[J]. 国际新闻界,2022(3):6-28.
[46]贾开,徐杨岚,吴文怡. 机器行为学视角下算法治理的理论发展与实践启示[J]. 电子政务,2021(7):15-22.
[47]SHEN H, CLARK N, MITRA T. Mind the value-action gap: do LLMs act in alignment with their values? [J]. arXiv preprint arXiv:2501. 15463, 2025.
[48]JASANOFF S. States of knowledge[M]. Abingdon: Taylor & Francis,2004.
[49]DONG H, CHEN J. Meta-regulation: an ideal alternative to the primary responsibility as the regulatory model of generative AI in China[J]. Computer law & security review, 2024, 54: 106016.
[50]MA J. Can machines think like humans? a behavioral evaluation of LLM-agents in dictator games[J]. arxiv preprint arxiv:2410. 21359, 2024.
[51]BERGER P L,LUCKMANN T. The social construction of reality: a treatise in the sociology of knowledge[M]. New York: Doubleday,1966.
[52]KAHNEMAN D,KNETSCH J L,THALER R H. Fairness and the assumptions of economics[J]. The journal of business, 1986,59(4):S285-S30.
[53]FORSYTHE R, HOROWITZ J L, SAVIN N E, et al. Fairness in simple bargaining experiments[J]. Games and economic behavior,1994,6(3):347-369.
[54]包特,王国成,戴芸. 面向未来的实验经济学:文献述评与前景展望[J]. 管理世界,2020,36(7):218-237.
[55]AHER G, ARRIAGA R I, KALAI A T. Using large language models to simulate multiple humans and replicate human subject studies[C]// Proceedings of the international conference on machine learning,2023:337-371.
[56]HORTON J J. Large language models as simulated economic agents:what can we learn from homo silicus? [J]. National bureau of economic research,2023, w31122.
[57]OUYANG S,YUN H,ZHENG X. How ethical should AI be? how AI alignment shapes the risk preferences of LLMs [J]. arXiv preprint arXiv:2406. 01168,2024.
[58]HUANG Y J,HADFI R. How personality traits influence negotiation outcomes? a simulation based on large language models[J]. arXiv preprint arXiv:2407. 11549,2024.
[59]PAN W, LIU Z, CHEN Q, et al. The hidden dimensions of llm alignment: a multi-dimensional safety analysis[J]. arXiv preprint arXiv:2502. 09674,2025.

(本文责编:默 黎)