大语言模型情绪调节能力评测基准

马志强[1,2]，刘义兴[1]，刘佳[1]，李鑫[3]
（1.内蒙古工业大学智能科学与技术学院，内蒙古 呼和浩特 010080；
2.内蒙古自治区北疆网络空间安全重点实验室，内蒙古 呼和浩特 010080；3.内蒙古工业大学心理咨询中心，内蒙古 呼和浩特 010080）

摘　要：为了解决当前大语言模型情绪评测中未考虑到情绪调节能力的问题，提出了针对大语言模型情绪调节能力的评测基准。首先，基于Gratz情绪调节能力定义，将大语言模型情绪调节能力划分为情绪调节有效性、情绪调节灵活性、情绪调节拟人性和情绪调节响应性4个维度；接着，针对大语言模型情绪调节能力定义，提出了相应的评测基准，包括评测集、评测方法两部分；最后，使用问卷法与实验法相结合的方式对ERNIE 4.0、DeepSeek、GPT-3.5、GPT-4和LLaMA-2（7B和13B）进行多维度评测。实验结果表明，所提评测基准能够有效区分大语言模型情绪调节能力。大语言模型情绪调节能力评测基准研究不仅有助于提升模型生成情绪的质量，还为大语言模型在敏感话题中的安全性、稳定性和伦理合规性提供了重要保障。

关键词：大语言模型评测；情绪调节能力；多维度评测
中图分类号：TP391；TP18
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025236

Benchmark for evaluating emotional regulation ability of large language model

MA Zhiqiang[1,2], LIU Yixing[1], LIU Jia[1], LI Xin[3]
1. College of Intelligent Science and Technology, Inner Mongolia University of Technology, Hohhot 010080, China
2. Inner Mongolia Key Laboratory of Beijiang Cyberspace Security, Hohhot 010080, China
3. Psychological Counseling Center, Inner Mongolia University of Technology, Hohhot 010080, China

Abstract: To address the lack of emotional regulation assessment in current evaluations of large language model (LLM), an evaluation benchmark for assessing emotional regulation capabilities in LLM was proposed. Based on Gratz's definition of emotional regulation ability, the capability was categorized into four dimensions: emotional regulation effectiveness, emotional regulation flexibility, emotional regulation human-likeness, and emotional regulation responsiveness. Correspondingly, an evaluation benchmark was developed, consisting of an evaluation dataset and assessment methods. Finally, a combination of questionnaire-based and experimental methods was used to conduct a multidimensional evaluation of ERNIE 4.0, DeepSeek, GPT-3.5, GPT-4, and LLaMA-2 (7B and 13B). Experimental results show that the proposed benchmark effectively differentiated the emotional regulation capabilities of various LLM. The benchmark study on the emotional regulation capability of large language models not only contributes to enhance the quality of their generated emotional content but also offers significant guarantees regarding their safety, stability, and ethical compliance when applied in sensitive contexts.

Keywords: large language model evaluation, emotional regulation ability, multidimensional evaluation

收稿日期：2025–08–14；**修回日期：**2025–09–15
基金项目：国家自然科学基金资助项目（No.62166029）；内蒙古自治区科研基础条件及平台基金资助项目（No.2025KYPT0014）；内蒙古自治区高等学校创新团队发展计划基金资助项目（No.NMGIRT2506）；内蒙古自治区高等学校碳达峰碳中和研究基金资助项目（No.STZX202307）；内蒙古自治区科技计划基金资助项目（No.2025SYFHH1239）
Foundation Items: The National Natural Science Foundation of China (No.62166029), Research Infrastructure and Platforms in the Inner Mongolia Autonomous Region (No.2025KYPT0014), Inner Mongolia Autonomous Region Higher Education Innovation Team Development Plan (No.NMGIRT2506), Research Project on Carbon Peak and Carbon Neutrality in Higher Education Institutions of Inner Mongolia Autonomous Region (No.STZX202307), The Science and Technology Plan Project of Inner Mongolia Autonomous Region (No.2025SYFHH1239)

0 引言

随着ERNIE、DeepSeek、GPT、LLaMA等大语言模型具备表达与回应人类情绪的能力后，情绪调节能力便成为影响其安全部署与用户信任的关键技术瓶颈。2024年11月13日，美国密歇根州大学生维德海·雷迪在与谷歌Gemini的交互过程中发生异常输出事件，引发了社会舆论的广泛关注[1]。该事件显示，当前大语言模型在面对社会话题中的潜在敏感元素时，仍缺乏稳定的情绪识别与调节机制，导致非比例性情绪反应。在技术层面，模型未能正确判断语境情绪强度，且缺少对激化情绪的干预机制与暴力输出的约束逻辑；在伦理层面，则暴露出其对人格尊重与心理安全保护机制的缺失，反映出情绪理解、风险评估和道德边界建模方面的系统性不足。因此，系统性评测大语言模型的情绪调节能力，不仅关乎模型生成质量，更是确保人工智能系统稳定性、安全性与伦理合规性的核心基础。从模型设计者的角度来看，这一能力决定其在敏感话题下的可控性与部署风险；从用户角度来看，则关系到人机交互中的心理安全与信任感。

然而，现有评测多局限于情绪识别、分类等基础维度。文献[2]利用情绪意识量表（LEAS，level of emotional awareness scale）分析GPT-3.5在20个场景中的回答，发现其能生成适当的情绪意识响应，且性能可能随时间提升。另有研究通过大语言模型预测对话情绪强度[3]及分析情感推理组件[4]，评测其情绪理解能力。在共情能力评测方面，文献[5]通过36因素数据集评测5个大语言模型，发现其与人类情绪行为不完全一致。文献[6]从人格特质等多维度评测模型，推动个性化大语言模型研究。文献[7]聚焦大语言模型情绪理解方面的情商评测，发现多数模型情商高于平均水平，但机制与人类存在质的差异。文献[8]从情绪理解和应用两方面评测，指出大语言模型情商与普通人存在差距。综上所述，情绪评测基准中缺乏针对情绪调节能力的评测，导致模型在动态交互场景中的情绪调节能力难以量化。

本文的主要贡献具体包含以下4个方面。

1) 提出一个大语言模型情绪调节能力评测基准（ER-Bench, emotional regulation benchmark），该基准包含评测集和评测方法。

2) 收集心理学中的珀斯情绪反应性量表（PERS，perth emotional reactivity scale）[9]、情绪调节量表（ERQ，emotion regulation questionnaire）[10]和情绪调节困难量表（DERS，difficulty in emotion regulation scale）[11]3个量表构成包含条目ID、条目内容和条目选项3个字段的评测集，用于评测大语言模型的情绪调节能力。

3) 基于Gratz情绪调节能力，对大语言模型的情绪调节能力进行定义，并进行多维度评测。

4) 为深入理解大语言模型在情绪调节任务中的表现，本文评估了6个主流模型，涵盖3个关键维度：模型架构（ERNIE 4.0和DeepSeek）、模型迭代版本（GPT-3.5和GPT-4）以及模型参数规模（LLaMA-2（7B和13B）），系统分析不同设计因素对情绪调节能力的影响，并探讨了未来模型优化的方向。

1 基本原理

1.1 情绪调节能力

情绪调节的研究起源于发展心理学[12]。Gross[13]在20世纪末系统地提出了情绪调节理论，推动了该领域的快速发展。在其基础上，Gratz等[11]提出情绪调节能力，包含个体对情绪的知觉和理解能力、对情绪的接受能力、情绪来临时依旧能够继续目标行为的能力、控制冲动行为的能力以及通过适应性的情绪调节策略来改变情绪反应的能力。

然而，大语言模型并不具备生理层面的情绪体验，其“情绪调节能力”需通过生成语言中的情绪表达行为进行类比评测。为此，本文将大语言模型的情绪调节能力划分为4个可操作维度：情绪调节有效性、情绪调节灵活性、情绪调节拟人性与情绪调节响应性，以近似映射Gratz等所提出的情绪调节能力构成。

情绪调节有效性：衡量模型是否能根据不同情境提示产生差异化的情绪表达，体现其对语义或情境的识别与响应能力。情绪调节灵活性：通过使用心理学量表分析模型对不同策略的偏好与使用分布，反映其调节方式的多样性与适应性。情绪调节拟人性：通过比较模型在量表得分的平均值和标准差与人类样本之间的差异，评测其情绪调节行为的类人性。情绪调节响应性：衡量模型在面对语言表述变化时输出的一致性，反映其对语义扰动的稳定性与鲁棒性。这4个维度划分既贴合心理学理论，又具备良好的可测性与操作性，为模型的情绪调节

能力评测提供了系统化框架。

1.2　情绪调节能力的评测

在心理学研究中，情绪调节的评测主要包括观察法、问卷法、实验法等。其中，观察法是指在不受研究者干预的前提下，在自然真实的情境中观察研究对象情绪调节的一种方法[14]。问卷法是情绪调节的常用方法，主要是由研究对象根据自己情绪调节的实际情况对研究者提供的评测条目按照一定的评分标准进行回答，然后研究者根据一定的标准进行统计处理[15]。实验法是通过特定的实验条件设置来观察被试者运用情绪调节策略的实际效果[16]。

在评测大语言模型的情绪调节能力时，采用问卷法与实验法相结合的方式，这主要基于以下几点考虑。问卷法能够有效评测模型在情绪调节策略上的理论理解，但由于其依赖模型或用户自述，具有较强的主观性，难以全面反映模型在实际情境中的情绪调节能力。实验法通过控制情境变量，能够客观评测模型在特定情境下的情绪调节效果，从而弥补问卷法的主观性局限。二者结合，既能从理论层面分析模型的情绪调节策略，也能通过实验验证模型在特定情境下的实际调节表现。相比之下，观察法因模型缺乏外显行为，难以捕捉情绪调节的自然反应且无法有效控制变量，评测效果有限。因此，结合问卷法与实验法能更全面、精准地评测大语言模型的情绪调节能力，观察法则无法提供有效的评测依据。

2　ER-Bench

2.1　任务定义

基于Gratz等[11]对情绪调节能力的定义，本文提出一个系统性的评测基准ER-Bench，用于衡量大语言模型在情绪调节任务中的整体表现。该评测基准综合考虑4个关键维度，具体可表示为ER - Bench = $f(E,F,P,R)$。其中，情绪调节有效性（E）用于衡量大语言模型在不同场景中情绪反应的能力，情绪调节灵活性（F）用于综合考察大语言模型的情绪调节困难程度、情绪调节策略倾向和情绪反应，情绪调节拟人性（P）用于评测大语言模型情绪调节能力与人类的相似度，情绪调节响应性（R）用于评测大语言模型在多次情绪调节测量中的效果一致性水平。

2.2　评测集

为了评测大语言模型的情绪调节能力，本文构建了一个评测集，旨在全面考察大语言模型的情绪调节能力。其中，评测集中包含条目ID、条目内容和条目选项3个字段，条目内容由PERS、ERQ和DERS这3个量表组成。这些量表的具体信息如表1所示。

2.3　评测方法

ER-Bench包括对大语言模型情绪调节能力多方面的评测。图1给出了ER-Bench概述。其中ER-Bench使用评测集对大语言模型的情绪调节有效性、情绪调节灵活性、情绪调节拟人性以及情绪调节响应性4个方面进行评测。这4个部分均分为两部分：生成和计算。其中，生成部分流程如图2所示，计算部分流程如图3所示。

1) 情绪调节有效性

由于情绪调节能力有效性在于衡量模型整体的表现趋势，故使用平均值进行计算。

首先，将一组系统提示Q^{SP}和量表内容Q_c输入

表1　　**量表的具体信息**

信息	PERS[9]		ERQ[10]		DERS[9, 11]
	负向情绪反应	正向情绪反应	认知重评	表达抑制	
条目数	15	15	6	4	36
分值	1~5	1~5	1~7	1~7	1~5
计算方式	总分	总分	总分	总分	总分
地区	美国	美国	澳大利亚	澳大利亚	澳大利亚
年龄	平均年龄：20	平均年龄：20	年龄范围：18~67	年龄范围：18~67	年龄范围：18~67
女性/人	67	67	45	45	72
男性/人	712	712	20	20	111
平均值±标准差（女性）	46.32±13.50	54.01±10.89	4.60±0.94	3.64±1.11	72.93±15.89
平均值±标准差（男性）	42.25±12.40	51.40±10.16	4.61±1.02	3.14±1.18	70.15±10.72

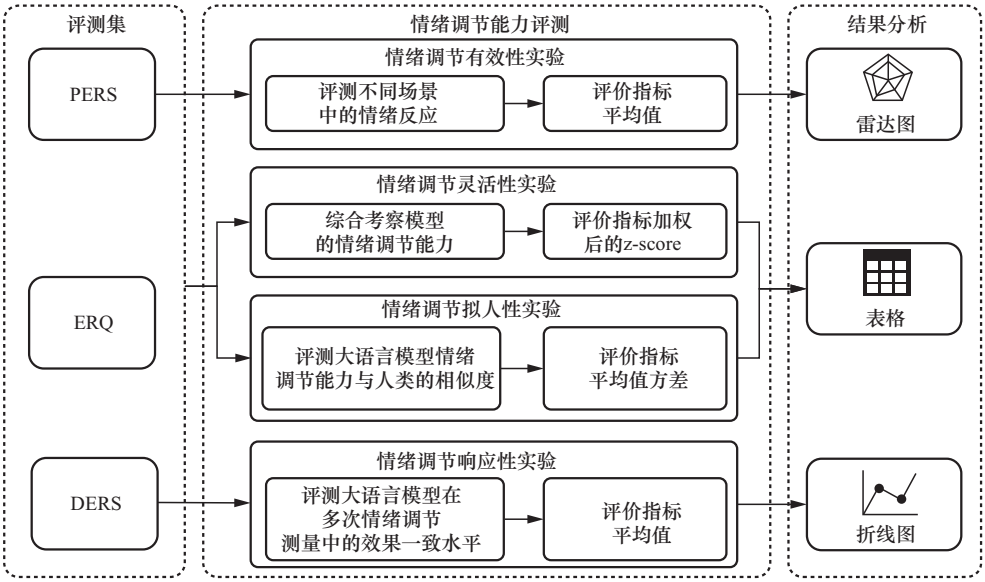

图1　ER-Bench 概述

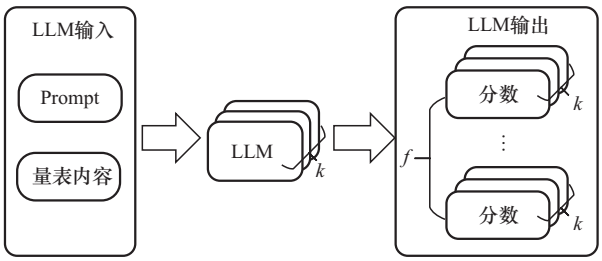

图2　生成部分流程

大语言模型。其次，大语言模型对每个条目进行 k 次作答后进行多数投票得出答案，最终输出一组结果 E。其中，Q^{SP} 表示一组包含 m 条的系统提示，即 $Q^{\mathrm{SP}}=[Q_1^{\mathrm{SP}},Q_2^{\mathrm{SP}},\cdots,Q_m^{\mathrm{SP}}]$；$Q_c$ 表示包含 f 个条目的第 c 个量表，即 $Q_c=[q_{c1},q_{c2},\cdots,q_{cf}]$；$E$ 表示一组包含 m 个大语言模型的情绪调节有效性得分，即 $E=[E_1,E_2,\cdots,E_m]$。情绪调节有效性的计算式为

$$M_{ij}^{(k)}=\mathrm{Model}(Q_i^{\mathrm{SP}},q_{cj})^{(k)}\in R\ ,k\in N \tag{1}$$

$$A_{ij}=\mathrm{Vote}(M_{ij}^{(1)},M_{ij}^{(2)},\cdots,M_{ij}^{(k)}) \tag{2}$$

$$E_i=\frac{1}{f}\sum_{j=1}^{f}A_{ij}\in R \tag{3}$$

其中，$M_{ij}^{(k)}$ 表示第 i 个系统提示下第 k 次生成的第 c 个量表中第 j 个条目的回答结果，A_{ij} 表示经过多数投票处理后每个条目的最终得分，E_i 表示第 i 个系统提示下的平均情绪调节有效性得分。E 中的各个得分差距越显著，说明模型情绪调节有效性越高。

2) 情绪调节灵活性

由于不同量表的取值范围不同，直接比较原始得分会产生误解，因此在计算情绪调节灵活性时使用 z-score 的计算方式。同时，不同量表的条目数量不同，为更加公平地比较不同模型之间的情绪调节灵活性，引入权重进行计算。

首先，将通用系统提示 Q^G 和多个量表内容 Q 输入大语言模型。其次，大语言模型对每个条目进行 k 次作答后再进行多数投票得出答案，最终输出一组结果 F。其中，Q 表示第 c 个量表的所有内容，即 $Q=[Q_1,Q_2,\cdots,Q_c]$；F 表示一组包含 $c+1$ 个大模

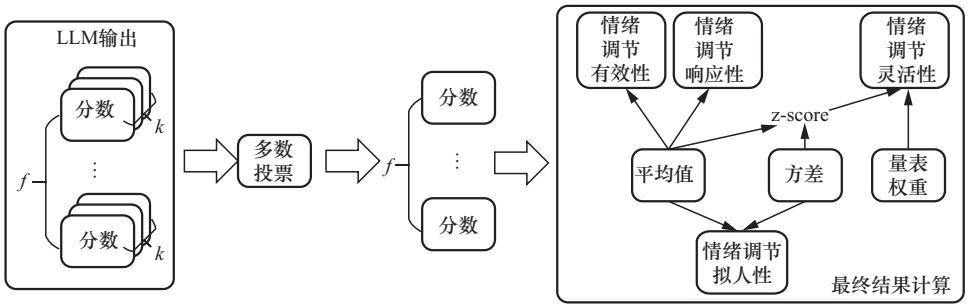

图3　计算部分流程

型的情绪调节灵活性得分，即$F=[F_1,F_2,\cdots,F_{c+1}]$。大语言模型回复条目内容的计算式为

$$M_{ci}^{(k)}=\text{Model}(Q^G,q_{ci})^{(k)}\in R,k\in N \tag{4}$$

$$A_{ci}=\text{Vote}(M_{ci}^{(1)},M_{ci}^{(2)},\cdots,M_{ci}^{(k)}) \tag{5}$$

其中，$M_{ci}^{(k)}$表示第c个量表的第i个条目模型的第k次生成结果，A_{ci}表示经过多数投票后的最终得分。模型在该量表上的平均值（Mean）和标准差（SD，standard deviation）的计算式为

$$\text{Mean}=\frac{1}{k}\sum_{j=1}^{k}(\sum_{i=1}^{f}M_{ci}^{(j)}) \tag{6}$$

$$\text{SD}=\sqrt{\frac{1}{f-1}\sum_{j=1}^{f}(X_j-\text{Mean})^2} \tag{7}$$

其中，f表示第c个量表中的条目数量，$X_j=\sum_{i=1}^{f}M_{ci}$表示未经多数投票在第c个量表上第j次作答的总得分。对于第i个量表，其情绪调节灵活性得分计算式为

$$F_i=\frac{X_i-\text{Mean}}{\text{SD}} \tag{8}$$

其中，$X_i=\sum_{j=1}^{f}A_j$表示经多数投票后在量表i上的总得分。在c个量表的基础上，模型的整体情绪调节灵活性得分为

$$F_{c+1}=\sum_{i=1}^{|Q|}(\omega_i\times F_i) \tag{9}$$

其中，$\omega_i=\frac{|Q_i|}{\sum_{i=1}^{|Q|}Q_i}$表示第$i$个量表占所有量表的权重，即第$i$个量表中所含条目与所有量表中所有条目数之和的比值。最终，F中的各个值越高，说明模型在情绪调节灵活性上的表现越好。

3) 情绪调节拟人性

为了对比不同模型与人类之间的差距，且在测量人类的情绪调节能力时使用平均值±方差的方式，因此在对模型进行测量时采取同样的手段。

将通用系统提示Q_G和多个量表内容Q输入大语言模型，模型对每个条目进行k次生成作答，并对输出进行统计，最终得到一组$\mu\pm\sigma$表示的情绪调节拟人性得分P。其中，$Q=[Q_1,Q_2,\cdots,Q_c]$表示c个量表的所有内容；μ表示平均值；σ表示标准差；P表示大语言模型的一组情绪调节拟人性得分，即$P=[P_1,P_2,\cdots,P_i,\cdots,P_c]$且$P_i=\mu\pm\sigma$。

对于第c个量表第i个条目，模型的第k次生成结果为

$$M_{ci}^{(k)}=\text{Model}(Q^G,q_{ci})^{(k)}\in R,k\in N \tag{10}$$

经过多数投票后的最终得分为

$$A_{ci}=\text{Vote}(M_{ci}^{(1)},M_{ci}^{(2)},\cdots,M_{ci}^{(k)}) \tag{11}$$

模型输出的统计量分别为

$$\mu=\begin{cases}\dfrac{1}{b}\sum_{i=1}^{b}A_{ci},\text{按类别统计}\\ \dfrac{1}{k}\sum_{j=1}^{k}(\sum_{i=1}^{f}M_{ci}^{(j)}),\text{按总体统计}\end{cases} \tag{12}$$

$$\sigma=\begin{cases}\sqrt{\dfrac{1}{d-1}\sum_{j=1}^{d}(X_j-\mu)^2},\text{按类别统计}\\ \sqrt{\dfrac{1}{f-1}\sum_{j=1}^{f}(X_j-\mu)^2},\text{按总体统计}\end{cases} \tag{13}$$

其中，μ表示平均值，σ表示标准差，b表示每个类别的条目数量，d表示类别数，f表示量表c的条目数量，k表示作答次数。若需按子类分析，使用“分类”版本公式；若直接按完整量表评测模型拟人性，使用“总体”版本公式。最后，大语言模型的响应性得分P与人类样本的结果越相近，拟人性水平越高。

4) 情绪调节响应性

由于情绪调节能力响应性在于衡量模型整体表现趋势，故使用平均值进行计算。

将通用系统提示Q^G和第c个量表进行多种语言变化后的所有条目Q'输入大语言模型，大语言模型对每个条目进行k次作答后再进行多数投票得出答案，最终输出一组结果R。其中，Q'表示包含f个条目的第c个量表进行a种语言变化后的所有条目，即$Q'=[Q'_1,Q'_2,\cdots,Q'_a]$；$Q'_a$表示第$a$种语言变化后的第$c$个量表，即$Q'_a=[q'_{a1},q'_{a2},\cdots,q'_{af}]$；$R$表示一组包含$a$个大语言模型的情绪调节响应性得分，即$R=[R_1,R_2,\cdots,R_a]$。大语言模型回复条目内容的计算式为

$$M_{ai}^{(k)}=\text{Model}(Q^G,q'_{ai})^{(k)}\in R,k\in N \tag{14}$$

$$A_{ai}=\text{Vote}(M_{ai}^{(1)},M_{ai}^{(2)},\cdots,M_{ai}^{(k)}) \tag{15}$$

其中，$M_{ai}^{(k)}$表示量表c进行第a次语言变化后第k次生成的第i个条目回答结果得分；A_{ai}表示第c个量

表在第 a 种语言变化后每个条目经过多数投票处理的最终得分。大语言模型在第 a 种语言变化的量表 c 上的情绪响应性得分计算式为

$$R_a = \frac{1}{f}\sum_{j=1}^{f} A_{aj} \in R \tag{16}$$

最后，R 组情绪调节响应性得分差异越小，说明模型在面对语言表述变化时保持稳定响应的能力越强。

3 实验

本节介绍了评测模型，展示了所有选定模型的结果，并对实验结果进行了分析。

3.1 评测模型

为了测试 ER-Bench，本文从多维角度选取评测模型。

1) 模型迭代维度，选择 OpenAI 的代表性模型 GPT-3.5（gpt-3.5-turbo）[17]和 GPT-4（gpt-4o）[18]。

2) 参数量级维度，选择使用具有相同的体系结构、数据和训练策略的 LLaMA-2（7B 和 13B）[19]。

3) 技术路线维度，选取其他主流大模型中百度的 ERNIE 4.0（ERNIE-4.0-8K）[20]和深度求索的 DeepSeek（DeepSeek-R1）[21]。

3.2 评测细节

首先，评测大语言模型情绪调节有效性，本文采用 3 种系统提示，分别为通用系统提示、情景重构系统提示和角色扮演系统提示。其中，情境重构系统提示与角色扮演系统提示是由参考认知行为疗法中的情境重构[22]以及角色扮演[23]两种方法构建的。在评测大语言模型的其他 3 个维度时，均使用通用系统提示。其次，评测大语言模型的 4 个维度，大语言模型均进行 10 次作答其中，评测大语言模型情绪调节有效性，使用 PERS 量表；评测大语言模型情绪调节灵活性和拟人性使用 PERS、ERQ 和 DERS 这 3 个量表；评测大语言模型情绪调节响应性，使用 DERS 量表。最后，评测大语言模型情绪调节响应性，先将 DERS 量表设置为中英两个版本，再在英文版的基础上将所有条目语言表述变化两次。

3.3 实验结果

3.3.1 情绪调节有效性实验

大语言模型情绪调节有效性实验使用 3 种调节策略评测负向情绪调节有效性，实验结果如图 4 所示。其中，6 个坐标轴分别表示 6 种主流语言模型，得分差距越大表示模型的负向情绪调节越有效，3 种策略包括无情景提示、情境重构和角色扮演。

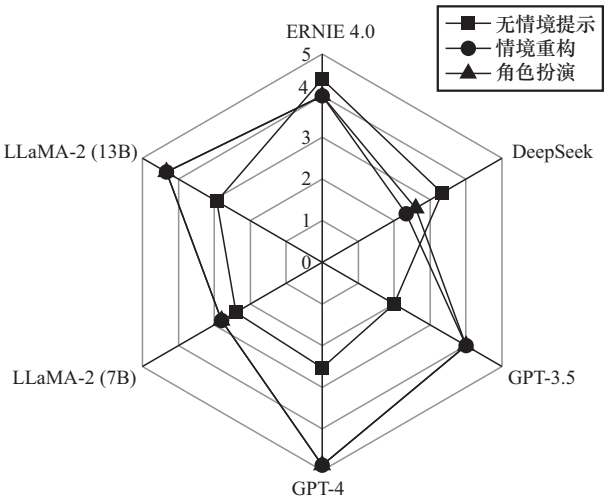

图 4 不同模型负向情绪调节有效性对比

大语言模型情绪调节有效性实验使用 3 种调节策略评测正向情绪调节有效性，实验结果如图 5 所示。其中，6 个坐标轴分别表示 6 种主流语言模型，得分差距越大表示模型的正向情绪调节越有效，3 种策略包括无情景提示、情境重构和角色扮演。

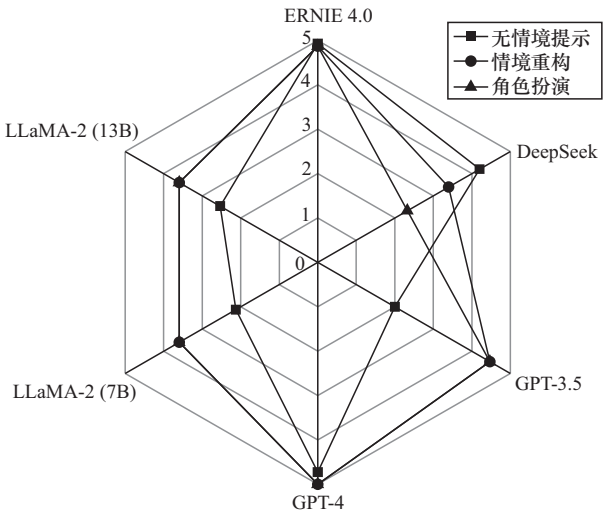

图 5 不同模型正向情绪调节有效性对比

由图 4 与图 5 可知，在不同情境提示下，多数模型能够对其情绪响应维度进行调整，体现出一定的情绪调节有效性。以 GPT-4 为例，在情境重构策略作用下，其在负向情绪调节任务中的得分提升幅度约为 2.5 分，明显高于在正向情绪调节任务中的

提升幅度（不足 1 分），显示出其在负向情绪情境中的调节能力更为敏感与有效。相比之下，ERNIE 4.0 在 3 种调节策略干预下，各 PERS 子量表得分变化均未超过 1.0 分，表明其调节响应较为稳定但调节能力相对有限。LLaMA-2 系列模型则表现出另一种特征：在两种参数规模（7B 和 13B）下，调节策略使各子量表得分平均提升约 2.0 分。鉴于 LLaMA-2（13B）并非本文中参数规模最大的模型，其显著提升更可能归因于模型架构设计或预训练语料特征，而非仅由参数规模决定。总体而言，模型对情绪调节提示的响应受到版本演进、架构设计与训练数据分布等多因素的共同影响，参数规模虽有一定作用，但并非决定性因素。

3.3.2　情绪调节灵活性实验

DERS 量表得分越高表明情绪调节困难性越高，与其他量表相悖。因此，在计算 DERS 量表的内容时，先进行反向计算，再计算平均值、标准差和情绪调节灵活性得分。其中，反向计算指的是统计大语言模型在 DERS 量表上的得分后，使用 DERS 量表的满分减去现有总分值得到的分值，即大语言模型最终的得分。情绪调节灵活性实验结果如表 2 所示，其中，粗体表示最好的结果，下划线表示最差的结果。

表 2　不同模型在各个量表上的表现

模型	DERS	ERQ	PERS	综合
ERNIE 4.0	−1.73	−0.07	0.00	−0.83
DeepSeek	**2.85**	0.00	0.00	**1.35**
GPT-3.5	−0.03	−1.37	**1.08**	0.23
GPT-4	−0.14	0.54	−1.56	-0.61
LLaMA-2 (7B)	1.11	−0.60	0.02	0.45
LLaMA-2 (13B)	0.59	**0.55**	0.25	0.45

由表 2 可知，不同模型在 DERS、ERQ 和 PERS 这 3 个量表上的表现差异明显。在 ERQ 量表中，LLaMA-2(13B)取得 0.55 的较高得分；在 PERS 量表中，GPT-3.5 与 GPT-4 的得分差距达到 2.64，显示模型版本迭代对情绪调节能力具有显著影响。鉴于 GPT 系列在架构设计、训练数据和优化目标方面均经历了系统升级，该性能提升更可能源于版本迭代本身，而非单纯的参数规模扩张。相比之下，LLaMA-2 系列在 3 个量表上的得分相对接近，表现出一定的稳定性，但整体分数偏低，提示中小规模模型在复杂情绪调节任务中仍受能力限制。综合来看，模型的版本优化和训练范式改进对情绪调节能力的提升作用更为直接，规模扩大虽可带来收益，但并非决定性因素。

3.3.3　情绪调节拟人性实验

由于人类样本在 ERQ 量表上的结果计算方式是按类别公式进行计算且在其余量表按总体公式进行计算的，因此，本次实验采取与人类样本相同的计算方式。将大语言模型在各量表上的结果与人类样本进行对比，实验结果如表 3 所示。其中，粗体表示与人类样本最接近的模型结果，下划线表示与人类样本相距最远的模型结果。人类样本在各量表上的表现结果如表 1 所示。

由表 3 可知，在 ERQ 量表的表达抑制中，GPT-4 与人类样本在平均值上的差异不足 1.00 分。在 PERS 量表的负向情绪反应中，GPT-4 与人类样本在平均值上的差异为 1.12 分。在 DERS 量表上，多数模型在情绪调节能力上均比人类样本困难，而 GPT 系列模型是最接近人类样本的大语言模型，并且最新的模型 DeepSeek 的表现却不如 GPT 系列模型。首先，模型学习到更多的语料后，其表现会与人类样本更加接近。其次，DeepSeek 的优化目标

表 3　大语言模型在各量表上与人类样本的对比结果

模型	DERS	ERQ		PERS	
		认知重评	表达抑制	负向情绪反应	正向情绪反应
ERNIE 4.0	124.40±1.51	7.00±0.00	3.00±1.22	66.10±1.73	73.90±0.32
DeepSeek	112.70±0.95	6.00±0.82	4.00±1.73	50.00±0.00	**63.00±0.00**
GPT-3.5	**71.90±3.89**	2.00±0.00	2.00±0.00	27.90±2.64	28.40±1.71
GPT-4	84.40±4.14	7.00±0.00	**3.50±1.50**	**43.60±2.88**	70.30±1.34
LLaMA-2 (7B)	124.80±1.62	1.33±1.97	1.50±1.66	33.30±15.59	34.20±13.89
LLaMA-2 (13B)	116.90±1.52	**6.00±0.00**	6.00±0.00	38.00±16.02	36.10±16.88

和设计方向主要集中在推理和计算任务上，而GPT系列模型更侧重于多领域应用和自然语言生成等方面，使GPT在情绪调节能力的表现优于DeepSeek。因此，大语言模型在不同程度上表现出与人类样本之间的差距。模型迭代有助于模型的情绪调节能力接近人类样本的表现。

3.3.4　情绪调节响应性实验

评测条目的中英文版本对实验结果的影响，结果如图6所示。

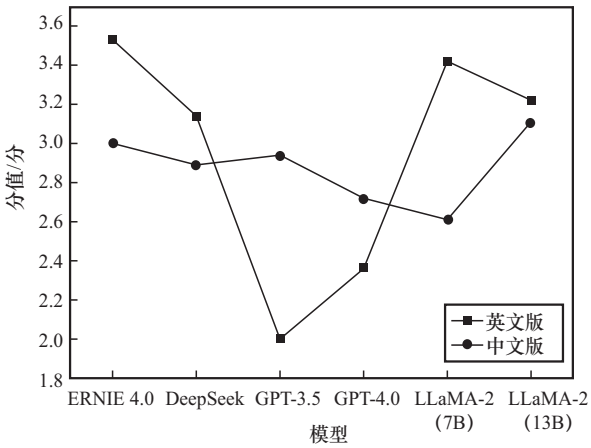

图6　中英文结果对比

由图6可知，在相同任务设定与提示语义下，中文版本的得分普遍低于英文版本，平均差值为0.2~0.8分。考虑到DERS量表的计分规则（得分越低代表情绪调节能力越强），可知模型在中文语境下的调节效果相对更好。该差异一方面可能源于语种自身在情绪表达方式上的结构差异：中文更强调情感表达的含蓄与语义留白，可能促使模型生成更收敛、更温和的调节回应；英文则偏好具体、直白地描述情绪过程，进而在量表测量中被识别为调节效果不好。另一方面，当前主流大语言模型的预训练语料中普遍以英文为主，模型在英文语境下生成的情绪表达通常更为细致、具体。这种更丰富的情绪细节可能使模型在输出中保留了较多负面情绪或情绪波动的表达，导致得分偏高。综上所述，语言本身在情绪表达风格、表述方式等方面的差异，以及模型在不同语种上预训练语料覆盖程度的不均衡，共同影响了模型在情绪调节任务中的表现。

条目表述变化两次后按照实验设计的方法评测大语言模型，结果如图7所示。

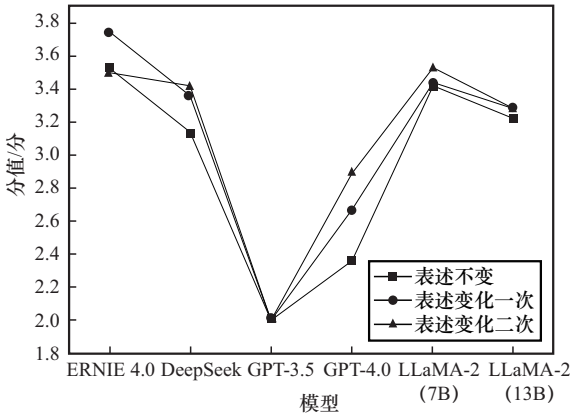

图7　条目改写后结果对比

由图7可知，GPT-4的平均值差距不超过1.0分，这表明条目表述变化对实验结果产生了一定的影响，但整体幅度较小。出现这种变化可能是由于条目表述变化后，大语言模型对输入数据的理解更清晰，或者是参与实验的大语言模型对新表述的认知与之前有所不同，从而影响了结果。

4　结束语

本文对大语言模型情绪调节能力评测基准进行研究。首先，基于Gratz等提出的情绪调节能力，对大语言模型情绪调节能力进行定义。其次，根据此定义提出大语言模型情绪调节能力评测基准。最后，使用问卷法与实验法相结合系统地评测了6个主流大语言模型。实验验证了大语言模型情绪调节能力评测基准ER-Bench可实现情绪调节能力的层次化解析，为大语言模型在敏感话题中的情绪生成质量、安全性、稳定性及伦理合规性提供了重要保障。需要特别说明的是，本文核心价值在于构建普适性评测方法论而非追求模型覆盖广度。受限于当前大语言模型调用成本，实验只选取了6个典型模型进行验证，但方法论本身具有模型无关性，评测框架可扩展应用于其他语言模型。未来工作将聚焦评测基准的迭代优化，具体包括拓展情绪诱发场景库的覆盖维度、深化情绪调节策略的评测粒度以及创建更为标准化的测试集等。

参考文献：

[1] CLARK A, MAHTANI M. Google AI chatbot responds with a threatening message: “Human … Please die.”[R]. 2024.

[2] ELYOSEPH Z, HADAR-SHOVAL D, ASRAF K, et al. ChatGPT outperforms humans in emotional awareness evaluations[J]. Frontiers in Psychology, 2023, 14: 1199058.

[3] PAECH S J. EQ-bench: an emotional intelligence benchmark for large language models[J]. arXiv Preprint, arXiv: 231206281, 2023.
[4] TAK A N, GRATCH J. Is GPT a computational model of emotion? [C]// Proceedings of the 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII). Piscataway: IEEE Press, 2023: 1-8.
[5] HUANG J-T, LAM M H, LI E J, et al. Emotionally numb or empathetic evaluating how llms feel using emotionbench[J]. arXiv Preprint, arXiv: 230803656, 2023.
[6] HUANG J T, WANG W X, LI E J, et al. Who is ChatGPT? Benchmarking LLMs' psychological portrayal using PsychoBench[J]. arXiv Preprint, arXiv: 231001386, 2023.
[7] WANG X N, LI X T, YIN Z, et al. Emotional intelligence of large language models[J]. Journal of Pacific Rim Psychology, 2023, 17: 18344909231213958.
[8] SABOUR S, LIU S Y, ZHANG Z Y, et al. EmoBench: evaluating the emotional intelligence of large language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 5986-6004.
[9] BECERRA R, PREECE D, CAMPITELLI G, et al. The assessment of emotional reactivity across negative and positive emotions: development and validation of the Perth emotional reactivity scale (PERS)[J]. Assessment, 2019, 26(5): 867-879.
[10] GROSS J J, JOHN O P. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being[J]. Journal of Personality and Social Psychology, 2003, 85(2): 348-362.
[11] GRATZ K L, ROEMER L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale[J]. Journal of Psychopathology and Behavioral Assessment, 2004, 26(1): 41-54.
[12] GAENSBAUER T J. Regulation of emotional expression in infants from two contrasting caretaking environments[J]. Journal of the American Academy of Child Psychiatry, 1982, 21(2): 163-170.
[13] GROSS J J. The emerging field of emotion regulation: an integrative review[J]. Review of General Psychology, 1998, 2(3): 271-299.
[14] 刘启刚. 情绪调节的研究方法与测量手段述评[J]. 心理研究, 2008, 1(2): 42-46.
LIU Q G. A review of the research methods and measurement means of emotion regulation[J]. Psychological Research, 2008, 1(2): 42-46.
[15] WEISS J A, THOMSON K, CHAN L S. A systematic literature review of emotion regulation measurement in individuals with autism spectrum disorder[J]. Autism Research, 2014, 7(6): 629-648.
[16] 黄敏儿, 郭德俊. 原因调节与反应调节的情绪变化过程[J]. 心理学报, 2002, 34(4): 371-380.
HUANG M E, GUO D J. Divergent consequences of antecedent-and response-focused emotion regulation[J]. Journal of Chinese Psychology Acta Psychologica Sinica, 2002, 34(4): 371-380.
[17] ROUMELIOTIS K I, TSELIKAS N D. ChatGPT and open-AI models: a preliminary review[J]. Future Internet, 2023, 15(6): 192.
[18] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[J]. arXiv Preprint, arXiv: 230308774, 2023.
[19] TOUVRON H, MARTIN L, STONE K, et al. LLAMA 2: open foundation and fine-tuned chat models[J]. arXiv Preprint, arXiv: 230709288, 2023.
[20] DENG H, ZHANG H, OU J, et al. Can LLM be a good path planner based on prompt engineering mitigating the hallucination for path planning[J]. arXiv Preprint, arXiv: 240813184, 2024.
[21] GUO D, YANG D, ZHANG H, et al. DeepSeek-R1: incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv Preprint, arXiv: 250112948, 2025.
[22] BECK A T, HAIGH E A P. Advances in cognitive theory and therapy: the generic cognitive model[J]. Annual Review of Clinical Psychology, 2014, 10: 1-24.
[23] MEICHENBAUM D. A cognitive-behavior modification approach to assessment[M]. Boston: Springer, 1977.

[作者简介]

马志强（1972– ），男，内蒙古托克托县人，内蒙古工业大学教授、硕士生导师，主要研究方向为网络内容安全、情感计算、人工智能等。

刘义兴（1999– ），女，河南三门峡人，内蒙古工业大学硕士生，主要研究方向为信息内容安全、自然语言处理。

刘佳（2000– ），女，四川内江人，内蒙古工业大学硕士生，主要研究方向为情感计算、对话情绪生成。

李鑫（1992– ），女，山西朔州人，内蒙古工业大学助理研究员，主要研究方向为临床心理学、认知行为疗法、情绪障碍、高校心理健康教育等。