Modern Linguistics 现代语言学, 2025, 13(12), 195-205
Published Online December 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.13121253

Hans汉斯

体裁与语向差异对人工与大型语言模型笔译评分行为的影响

刘玲燕，唐　青，王彦南

北京林业大学外语学院，北京

收稿日期：2025年11月2日；录用日期：2025年11月28日；发布日期：2025年12月11日

摘　要

本文基于多面Rasch模型(Many-Facet Rasch Model, MFRM)，探讨体裁与语向差异对人工评分员与大型语言模型(LLMs)在笔译评分中的影响。研究以北京市某211高校商务英语专业30名本科生完成的八项平行翻译任务为语料，涵盖科技、商务、新闻与议论文四类文本的汉译英与英译汉方向。根据权威测评报告选取四个主流大语言模型(ChatGPT-4o、DeepSeek、通义Qwen-2.5与腾讯元宝)，并与两位专家评分员共同参与评分。结果表明，人工评分整体偏宽，而大型语言模型普遍偏严；体裁差异显著影响评分严厉度，反映出评分者对文本功能与语言密度的敏感性。语向效应分析显示，人工评分在英译汉方向更为严格，腾讯元宝在汉译英方向偏严，而ChatGPT-4o与通义Qwen-2.5在双向评分中保持较高一致性。研究结果表明，大语言模型已具备初步的体裁识别与语向适配能力，但仍存在一定的严厉度偏移。本文为智能笔译测评系统的校准机制、体裁化教学反馈及跨语向评分公平性优化提供了实证参考。

关键词

体裁敏感性，语向效应，多面Rasch模型，大型语言模型，翻译测评

The Impact of Genre and Translation Direction on Human and Large Language Model Scoring Behaviors in Translation Assessment

Lingyan Liu, Qing Tang, Yannan Wang

School of Foreign Languages, Beijing Forestry University, Beijing

Received: November 2, 2025; accepted: November 28, 2025; published: December 11, 2025

文章引用: 刘玲燕, 唐青, 王彦南. 体裁与语向差异对人工与大型语言模型笔译评分行为的影响[J]. 现代语言学, 2025, 13(12): 195-205. DOI: 10.12677/ml.2025.13121253

Abstract

Drawing on the Many-Facet Rasch Model (MFRM), this study investigates how genre and translation direction shape the rating behavior of human raters and large language models (LLMs) in translation assessment. The dataset comprises eight parallel translation tasks completed by 30 Business English undergraduates at a 211 university in Beijing, covering scientific, business, news, and argumentative texts in both Chinese-English and English-Chinese directions. Four mainstream LLMs (ChatGPT-4o, DeepSeek, Tongyi Qwen-2.5, and Tencent Yuanbao) were selected based on authoritative evaluation reports and, together with two expert human raters, evaluated the translations using a unified scoring rubric. The findings show that human raters were generally more lenient, whereas all LLMs exhibited a consistent tendency toward stricter scoring. Genre exerted a significant influence on rating severity, indicating raters' sensitivity to textual function and information density. With respect to translation direction, human raters were stricter in the English-Chinese tasks, while Tencent Yuanbao demonstrated higher severity in the Chinese-English direction. In contrast, ChatGPT-4o and Tongyi Qwen-2.5 maintained relatively high consistency across both directions. Overall, the results suggest that LLMs have begun to develop initial capacities for genre recognition and direction adaptation, although noticeable severity biases remain. These findings offer empirical support for the calibration of intelligent translation assessment systems and provide pedagogical implications for genre-based instructional feedback and the enhancement of cross-directional fairness in translation evaluation.

Keywords

Genre Sensitivity, Directional Effect, Many-Facet Rasch Model, Large Language Models, Translation Assessment

1. 引言

近年来，以 ChatGPT 为代表的大型语言模型凭借其在语义理解、篇章生成与上下文推理等方面的优势，逐渐被引入语言教学与测评实践[1]。语言测评正经历由人工主导向智能协同的转型[2]。在笔译教学中，自动化评分工具的应用被认为有助于减轻教师负担、提升反馈效率与客观性[3]。笔译测评不仅是检验学习者语言运用能力的手段，更是促进其跨语际思维与译后反思的重要环节。已有研究表明，大语言模型在写作与口语测评中能较好地再现人工评分模式，具有较高的一致性与信度[4] [5]。然而，在笔译质量评估领域，其评分准确性、稳定性及任务敏感性尚缺乏系统验证[6]。为此，本研究在前期信度分析的基础上，从任务体裁与翻译方向两维度探讨人工智能模型评分差异，旨在揭示大型语言模型评分行为规律，为高校翻译笔译中智能测评的公平性与信度提升提供实证依据。

2. 文献综述

2.1. 大型语言模型在语言测评中的应用

随着生成式人工智能的快速发展，大型语言模型在语言教育与教育评价领域的应用受到广泛关注[7] [8]。大语言模型具备强大的自然语言理解与生成能力，可在短时间内对大规模语料进行分析与评分，被

认为在自动化评估与智能反馈中具有较高潜力[9]。已有研究显示，GPT 系列模型在英语写作和口语评分中能与人工评分表现出较高的一致性[4] [5]。然而，尽管大语言模型在语言测评中的应用成果逐渐增多，其研究焦点却多集中于模型输出结果的整体信度与可靠性[10]，对其评分行为内部机制的探讨相对不足。部分研究在特定任务(如作文或口语)中验证了大型语言模型测评的可行性[4] [5]，但针对学习者译文的教育性评分研究仍较少，尤其在多任务、多文本类型条件下，模型评分的稳定性与公平性问题尚值得进一步分析。在前人研究的基础上，本研究关注大型语言模型在笔译教学场景中的表现，考察不同模型在体裁与语向条件下的评分差异，并结合人工评分结果，探讨模型评分的一致性与信度，为后续 AI 辅助笔译测评的应用提供参考。

2.2. 翻译测评与评分严厉度研究

翻译测评作为翻译教学的重要组成部分[11] [12]，这对于确保翻译学习者是否达到了预期的学习目标以及验证教学大纲的有效性都至关重要[13]。长期以来，人工评分在笔译课堂中发挥着关键作用，但其主观性强及操作成本高等问题，一直是制约翻译测评客观化与信度提升的难点[14]。为改进评分信度，语言测评领域引入多面 Rasch 模型，用于同时估计评分员严厉度、任务难度与被评对象能力等多维特征[15] [16]。该模型在揭示评分差异、校准评分标准及分析任务难度方面具有较高解释力[17]。已有研究表明，评分员的严厉程度并非稳定不变的个体特质，而可能因评分任务及评分标准的特征而有所差异[18]。不同类型文本在语言密度、逻辑组织与功能目的上的差异，往往会引起评分者在标准应用上的调整，即所谓“体裁敏感性”[19] [20]。目前，关于翻译测评中体裁与评分严厉度关系的定量研究仍较有限。基于此，本文将“评分员严厉度”与“文本体裁”结合考察，利用多面 Rasch 模型分析人工评分员与大型语言模型在不同体裁下的评分特征，为笔译教学中的体裁化测评提供实证参考。

2.3. 语向差异与模型偏差研究

翻译方向是影响译文质量与评估表现的重要变量。多项实证研究表明，翻译方向会系统性地影响认知加工负荷与产出质量，进而影响评估表现：在母语方向(L2→L1)通常表现出较低的认知负荷与更高的译文质量，而在外语方向(L1→L2)则相反[21] [22]。有研究亦发现，人类评分员在面对不同语向译文时会在宽严标准上有所调整[23]，类似现象同样存在于大语言模型之中，由于主流大语言模型的训练语料以英语为主，模型在汉译英与英译汉任务中的表现常呈不对称特征[24]。现有研究主要从生成层面讨论语向差异，即模型在不同语向任务中的输出质量差异[25]，但对其在评分阶段可能产生的系统性偏差关注较少。由于不同语言方向涉及的语义结构、文化信息及语言熟悉度差异较大，模型在评分时是否会出现倾向性判断，仍需结合量化数据加以分析。本研究以语言方向为切入点，利用多面 Rasch 模型分析不同大语言模型在汉译英与英译汉任务中的评分差异，关注其语向敏感性特征。通过与人工评分者的对比分析，旨在进一步了解模型评分偏差的可能来源，为构建更加公平、稳健的 AI 笔译测评体系提供参考。

3. 研究方法

3.1. 研究问题

本研究以人工评分员与大型语言模型在笔译评分中的表现为研究对象，考察体裁与翻译方向对评分严厉度和偏差的影响。研究采用多面 Rasch 模型分析框架，探讨人机评分的一致性及方向敏感性。具体包括两个问题：(1) 不同文本类型(科技、商务、新闻、议论文)下，人工评分员与大语言模型的评分严厉度是否存在显著差异；(2) 不同模型(ChatGPT-4o、DeepSeek、通义 Qwen-2.5、腾讯元宝)的评分偏差是否受翻译方向(汉译英与英译汉)影响。通过构建评分员 - 任务 - 受试者三维测量框架，研究旨在揭示人机评

分差异的系统特征，为智能评测在笔译教学中的校准与应用提供实证支持。

3.2. 研究对象

本研究被试为北京市某 211 高校商务英语专业 30 名本科生，均修读《基础笔译》课程。所有被试完成 8 项平行翻译任务，共 240 份译文。任务涵盖科技、商务、新闻和议论文四类体裁，每类包含中译英与英译中两个方向。研究文本选自 CATTI 三级考试、TEM-4 真题及教材《A Practical Course in Business English Translation》，并经前测在篇幅、难度和可译度上进行控制，以保持体裁平衡与任务难度一致，减少外部变量对评分结果的干扰。

3.3. 研究工具

本研究所用工具包括翻译任务、评分量表、分析软件与提示语模板四部分。依据多份权威评测报告(《2025 大模型意识商数评估报告》《2024 年度大模型综合评测报告》及《中国媒体智能能力模型测评报告》)，在综合性能与语言覆盖度基础上，选取 ChatGPT-4o (OpenAI)、腾讯元宝(Tencent Yuanbao)、通义 Qwen-2.5 (阿里云)与 DeepSeek (深度求索研究院)四个模型为研究对象。翻译任务采用分段式设计，共两轮译写任务，每轮含一篇英译中(约 150 词)与一篇中译英(约 120 词)文本，体裁涵盖科技、新闻、议论文与商务四类。文本来源于 CATTI 三级考试、2024 年 TEM-4 真题及教材《A Practical Course in Business English Translation》，经 Hemingway Editor 测评可读性与 DeepL 可译度测试，筛选语言密度与复杂度适中材料以保证体裁平衡。评分量表依据 TEM-8 与 CATTI-2 标准修订，涵盖语言准确性、译文完整性、语篇连贯性与体裁得体性四维，采用四级评分制(1~4 分)。量表经专家审定与预实验验证，四类体裁 Cronbach's α 系数分别为 0.850、0.877、0.793 与 0.708，均达可接受水平。数据分析使用 FACETS 4.3.2 软件进行多面 Rasch 建模，考察评分员、任务与受试者三维交互效应，显著性水平设定为 $p < 0.05$。

3.4. 研究过程

研究过程分为任务实施、评分操作与数据处理三个阶段。在任务实施阶段，研究者依据 CSE 框架和课程要求，向被试说明译文任务目标与评分标准，确保其理解任务要求与提交格式。所有任务均在线上进行，每次限时 30 分钟，提交后由研究者统一编号整理。评分阶段包括人工评分与模型评分。两名人工评分员依据评分量表独立打分，取平均值作为基准分；随后，研究者将相同译文输入四个大型语言模型，并使用统一提示语生成分项与总分。提示语基于 CSE 能力维度设计，涵盖语言准确性、译文完整性、语篇连贯性与体裁得体性四个方面，要求输出结构化评分表。为保证可比性，各模型评分独立进行，参数设置保持一致。数据处理阶段将所有评分结果经核验后导出为 CSV 文件，并导入 FACETS 软件进行多面 Rasch 分析。研究结合 Infit MnSq 指标对人机评分结果进行交叉验证，进一步检验翻译方向与体裁的交互效应。所有分析均在受控条件下完成，以确保结果的客观性与可复现性。

4. 研究结果

4.1. 不同文本类型下人机评分严厉度差异

为检验不同文本类型(科技类、商务类、新闻类与议论文类)下人工评分员与大型语言模型评分严厉程度的差异，本文基于多面 Rasch 模型对评分结果进行了估计与交互分析。分析结果如表 1~4 及图 1 所示，分别呈现了评分者总体严厉度、任务(文本类型)难度、评分者与体裁交互效应以及信度统计指标。

由表 1 可见，五个评分主体的严厉度存在显著差异(χ^2 (4) = 70.8, $p < 0.001$)，分离度为 3.59，信度为

0.93，说明模型能有效区分评分者的严厉水平。人工评分员严厉度最低(−0.71 logit)，表现较宽松；ChatGPT-4o (0.26 logit)与 DeepSeek (0.40 logit)偏严，腾讯元宝最严(0.47 logit)，通义 Qwen-2.5 较宽(−0.42 logit)。结果表明，人工评分与大型语言模型在评分严宽取向上存在系统性差异，前者更具包容性，后者整体呈现“偏严”趋势。

Table 1. Rater measurement report
表 1. 评分者测量结果报告

评分者	测量者	标准误差	内拟合均方	外拟合均方	解释
专家	−0.71	0.05	1.01	0.97	评分最宽
ChatGPT-4o	0.26	0.04	1.03	1.05	略严
DeepSeek	0.40	0.04	0.98	1.00	偏严
腾讯元宝	0.47	0.04	1.02	1.00	最严
通义 Qwen-2.5	−0.42	0.04	1.04	1.01	略宽

注：正 logit 值表示评分趋严，负 logit 值表示评分趋宽。评分者间差异显著(χ^2 (4)= 70.8, $p < 0.001$)。评分者分离指数(Separation)为 3.59，信度(Reliability)为 0.93。

任务维度分析结果(表 2)进一步显示，不同文本类型的任务难度存在显著差异(χ^2 (7) = 51.6, $p < 0.001$)，分离度为 2.19，信度为 0.84。商务类与科技类汉译英任务难度最高(0.69 与 0.60 logit)，而新闻类与科技类英译汉任务相对容易(−0.55 与−0.35 logit)。这一差异可能反映出体裁在语言密度、逻辑组织与信息负载方面的不同要求，进而影响评分者在评判过程中的尺度应用，体现出一定的“体裁敏感性”。

Table 2. Text type measurement report
表 2. 文本类型测量报告

文本类型	测量值	标准误差	内拟合均方	解释
新闻 汉译英	−0.45	0.05	1.03	较易
新闻 英译汉	−0.55	0.05	1.02	最易
科技 汉译英	0.60	0.06	0.98	较难
科技 英译汉	−0.35	0.05	1.00	较易
商务 汉译英	0.69	0.06	1.00	最难
商务 英译汉	−0.28	0.05	0.99	较易
议论文 汉译英	0.22	0.06	1.00	略难
议论文 英译汉	0.12	0.05	1.01	中等

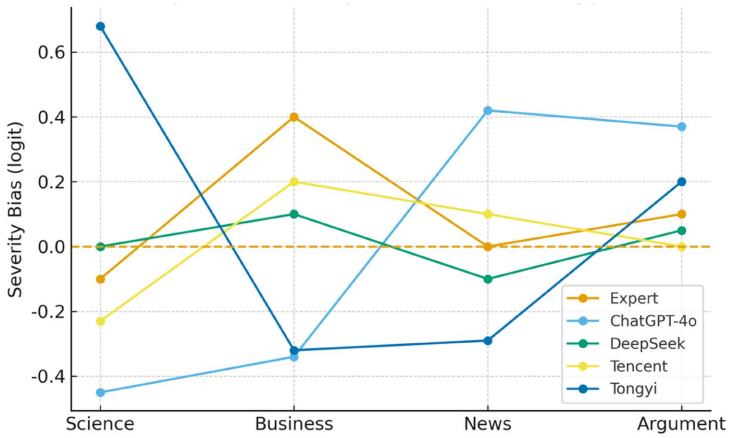
注：正 logit 表示更难。卡方检验结果为 χ^2 (7) = 51.6, $p < 0.001$；分离指数(Separation) = 2.19；信度(Reliability) = 0.84。

进一步的评分者 × 体裁交互分析(表 3 与图 1)显示，不同评分主体在体裁维度上存在显著的评分偏差。人工评分员在商务类汉译英任务中的严厉度最高(+0.41 logit, t=3.69)，说明其对专业术语与逻辑连贯性的要求较高；ChatGPT-4o 在科技类与商务类英译汉任务中表现为负偏差(−0.34 与−0.45 logit)，在新闻类与议论文体裁中则偏宽(+0.42 与+0.37 logit)，表明其在体裁识别与语用判断方面存在灵活调整；通义 Qwen-2.5 在科技与新闻体裁下也呈现方向相反的偏差($t > 3.0$)，而 DeepSeek 在各体裁间无显著差异，表现出较高的评分稳定性。总体来看，不同模型在面对不同语篇类型时呈现出差异化的响应模式，部分大型语言模型(如 ChatGPT-4o 与通义 Qwen-2.5)在体裁变换下的严宽差异更为明显。

Table 3. Rater × genre interaction table
表 3. 评分者 × 体裁交互表

评分者	文本类型	偏差(logit)	t	显著性
专家	商务 汉译英	0.41	3.69	显著
ChatGPT-4o	议论文 汉译英	0.37	3.40	显著
ChatGPT-4o	商务 汉译英	−0.34	−3.14	显著
ChatGPT-4o	新闻 英译汉	0.42	3.91	显著
ChatGPT-4o	科技 英译汉	−0.45	−4.16	显著
DeepSeek	全部体裁			不显著
腾讯元宝	科技 汉译英	−0.23	−2.12	显著
通义 Qwen-2.5	商务 汉译英	−0.32	−3.04	显著
通义 Qwen-2.5	新闻 英译汉	−0.29	−2.74	显著
通义 Qwen-2.5	科技 英译汉	0.68	6.24	显著

注：偏差正值表示更严，负值表示更宽；$|t| \geq 2$ 视为显著。



说明：横轴为文本类型，纵轴为严厉度偏差(logit)。折线展示各评分者体裁敏感性趋势。

Figure 1. Severity bias across text types
图 1. 评分者在不同体裁下的严厉度偏差趋势

信度与显著性统计(表 4)显示，评分者与任务两个维度的信度均较高(0.93 与 0.84)，说明模型估计结果稳定可靠。整体而言，人工评分员与大型语言模型在评分严宽标准上存在显著差异，不同体裁的任务特征对评分表现具有系统性影响。体裁敏感性在一定程度上调节了人机评分差异，显示出评分行为受语篇类型与任务难度的双重作用影响。

Table 4. Reliability and significance summary
表 4. 信度与显著性汇总

因素	分离指数	信度	卡方值 χ^2 (df)	显著性 p
评分者	3.59	0.93	70.8(4)	<0.001
任务	2.19	0.84	51.6(7)	<0.001

综上可见，人工评分员与大型语言模型在各类文本体裁下的评分严宽程度存在系统性差异，评分表

现受语篇类型与任务特征的共同作用影响。部分评分主体(如 ChatGPT-4o 与通义 Qwen-2.5)在体裁变化下呈现出更明显的评分幅度波动，而人工评分则在部分任务中展现出更高的体裁区分度。整体而言，体裁因素在一定程度上影响了评分者的尺度运用，形成了不同评分主体在体裁维度上的响应模式。

4.2. 翻译方向对大语言模型评分偏差的影响

为检验不同大型语言模型的评分偏差是否受到翻译方向(Direction)的影响，本文基于多面 Rasch 模型对评分者与语向之间的交互效应进行了统计分析。分析结果如表 5~8 所示，分别呈现了评分者在两种语向下的严厉度测量值、相对严厉度偏差、显著性检验结果以及平均得分情况。

从表 5 可以看出，不同评分者在汉译英与英译汉两种任务中的严厉度存在一定差异。ChatGPT-4o 和腾讯元宝在汉译英任务中表现出较高的评分严厉度(0.33 与 0.76)，而 DeepSeek 与人工评分员在英译汉方向的严厉度值相对更高(0.64 与−0.35)，说明部分评分者在面对不同语向的译文时，其评分标准存在方向性偏移。通义 Qwen-2.5 在两方向间的严厉度差异最小(−0.27 与−0.62)，表明其在跨语向评分中保持了较好的稳定性。

Table 5. Absolute severity of rater × language direction (AM-3-4)
表 5. 评分者 × 语向的绝对严厉度(AM-3-4)

评分者	汉译英严厉度(logit)	英译汉严厉度(logit)
ChatGPT-4o	0.33	0.14
DeepSeek	0.13	0.64
专家	−1.09	−0.35
腾讯元宝	0.76	0.08
通义 Qwen-2.5	−0.27	−0.62

注：正 logit 值表示评分较严，负值表示评分较宽。

相对严厉度指标(表 6)进一步验证了这一趋势。ChatGPT-4o (+0.07 vs. −0.11)、腾讯元宝(+0.30 vs. −0.38)及通义 Qwen-2.5 (+0.14 vs. −0.21)在汉译英方向上更为严格，而 DeepSeek (−0.26 vs. +0.26)与人工评分员(−0.40 vs. +0.34)则在英译汉方向上表现出更高的严厉度。结果表明，不同类型评分主体在两种语向任务中的评分取向存在差异，反映出一定程度的“语向效应”。

Table 6. Absolute severity of rater × language direction (RM-3-4)
表 6. 评分者 × 语向的相对严厉度(RM-3-4)

评分者	汉译英相对严厉度 Δ	英译汉相对严厉度 Δ
ChatGPT-4o	+0.07	−0.11
DeepSeek	−0.26	+0.26
专家	−0.40	+0.34
腾讯元宝	+0.30	−0.38
通义 Qwen-2.5	+0.14	−0.21

注：Δ 为去中心化后的相对偏差，正值表示该语向更严。

表 7 给出了评分者严厉度差异的 t 值检验结果。人工评分员($t = 2.18$)和腾讯元宝($t = 2.11$)的语向差异接近显著水平($|t| \geq 2$)，说明其在不同语向下评分严厉度的变化具有统计意义；而 ChatGPT-4o ($t = -0.42$)、

DeepSeek (t = 1.53)与通义 Qwen-2.5 (t = −0.80)的 t 值均未达到显著阈值，表明这些模型在两种语向任务中的评分较为稳定。总体而言，大语言模型的语向敏感性低于人工评分员，呈现出较高的评分一致性。

Table 7. Absolute severity of rater × language direction (TV-3-4)
表 7. 评分者 × 语向差异的显著性检验(TV-3-4)

评分者	汉译英 t 值	英译汉 t 值	显著性(参考阈值)
ChatGPT-4o	−0.42	0.63	
DeepSeek	1.53	−1.48	
专家	2.18	−1.84	≈显著(|t| ≥ 2)
腾讯元宝	−1.86	2.11	≈显著(|t| ≥ 2)
通义 Qwen-2.5	−0.80	1.14	

注：|t| ≥ 2 视为显著差异。

从平均得分情况(表 8)看，多数评分者在英译汉方向给出的分数略高，如 ChatGPT-4o (2.96 > 2.76)与腾讯元宝(2.97 > 2.63)，说明其在评估以中文为目标语的译文时倾向于较宽松的评分标准。该结果与相对严厉度指标相吻合，进一步印证了模型在语向变化下的轻微宽严偏差。

Table 8. Absolute severity of rater × language direction (AO-3-4)
表 8. 评分者 × 语向的观测平均分(AO-3-4)

评分者	汉译英平均得分	英译汉平均得分
ChatGPT-4o	2.76	2.96
DeepSeek	2.82	2.83
专家	3.13	3.08
腾讯元宝	2.63	2.97
通义 Qwen-2.5	2.92	3.15

注：多数评分者在英译汉方向得分略高(相对宽松)，与相对严厉度表结果一致。

综上所述，语向效应分析结果表明，不同大型语言模型在评分过程中确实受到翻译方向的影响，但总体差异幅度有限。其中，人工评分员和腾讯元宝在不同语向下表现出更显著的严厉度差异，而 ChatGPT-4o 与通义 Qwen-2.5 的评分表现相对稳定。该结果说明，大型语言模型在跨语向翻译质量评估中已具备较高的一致性与客观性，但个别模型仍存在一定的方向性偏差。

5. 讨论

5.1. 不同体裁任务下评分严厉度的差异机制与适配基础

本研究发现，不同文本体裁下人工评分员与大型语言模型的评分严厉度存在显著差异。人工评分整体偏宽，而 ChatGPT-4o、DeepSeek 与腾讯元宝偏严，通义 Qwen-2.5 相对宽容。该结果表明，模型评分在数值输出上更具稳定性，而人工评分更易受语篇解读与尺度应用风格的影响，反映出评分主体在评估行为上的系统差异。体裁因素对评分行为具有显著影响，科技与商务类汉译英任务的评分难度显著高于新闻与议论文体裁。人工评分员在商务类任务中表现出最高严厉度，显示其对术语规范性与逻辑一致性的高度敏感。语言密度高、信息负载大的任务易诱发更严格的评判标准，体现了体裁敏感性机制在评分过程中的作用，与 Wiseman (2012)关于语言任务复杂度影响评分严厉度的研究观点相一致[26]。

值得注意的是，ChatGPT-4o 与通义 Qwen-2.5 在体裁维度上呈现评分变动趋势，说明部分大型语言模型已具备语篇识别与评分调节能力，有潜力用于体裁适应型智能评分系统构建。相比之下，DeepSeek 在不同任务间评分相对平稳，反映其算法更偏向标准化路径而非语境调节机制。总体而言，体裁对评分者行为具有显著调节效应，不仅影响评分稳定性，也决定模型在多任务环境下的适配潜力。未来可通过引入体裁识别模块与任务调节策略，进一步提升模型在复杂译评场景中的适应性与信度。教师亦可利用“人机评分差异比对”引导学生关注体裁特征，促进译后反思与能力提升。

5.2. 翻译方向对评分偏差的一致性表现与调节能力分析

研究结果显示，不同翻译方向下人工评分员与部分大型语言模型的评分存在显著差异。人工评分员在英译汉任务中偏严，而腾讯元宝在汉译英任务中表现更严格，反映出一定的语向敏感性。该结果表明，评分者在处理母语输出与非母语产出时，对语言自然度和表达规范的预期存在方向性差异。这一结论与 McNamara (1996)关于“语向偏移影响评估表现”的观点相一致[15]，亦延续了后续研究关于评分者语言背景与语向熟悉度会显著影响其严厉度与语言自然度判断的实证观察[27]-[29]。与此同时，ChatGPT-4o 与通义 Qwen-2.5 在语向维度上保持较高一致性，评分严厉度差异未达显著水平，平均得分稳定，表明其在语向变化条件下具备较好的迁移能力。该结果提示部分模型已具备支持双向任务的评分一致性，为跨语向自动评分系统的构建提供了实证依据。

总体来看，不同评分主体在语向任务中呈现系统性差异，部分模型虽具稳定性，但在特定语向下仍存在偏差。教学实践应结合模型评分的方向性特征，引导学生理解中英互译中的语言规范差异，提升双语迁移与目标语表达能力。从系统优化角度看，未来模型应在保证一致性的同时完善语向适配机制，通过引入语向变量调节功能，增强其在多语向环境中的评分灵活性与稳定性。

6. 结论

本研究基于多面 Rasch 模型，从评分严厉度与评分偏差两个维度系统分析了人工评分员与四款大型语言模型(ChatGPT-4o、DeepSeek、通义 Qwen-2.5、腾讯元宝)在多体裁、多语向翻译任务中的表现差异。结果显示，人工评分整体偏宽，而模型评分普遍偏严，评分主体间存在稳定的尺度使用差异。体裁特征显著调节评分表现，科技与商务类汉译英任务因语言负载高、逻辑复杂而更具挑战性，评分严厉度随之上升。人工评分员在商务类文本中最严，体现出较强体裁敏感性；部分模型(如 ChatGPT-4o 与通义 Qwen-2.5)在体裁变化下出现方向性偏差，表明其已具备初步语篇识别与调节能力。在翻译方向上，人工评分员与部分模型存在“语向效应”：人工评分在英译汉中偏严，腾讯元宝在汉译英中更严格；而 ChatGPT-4o 与通义 Qwen-2.5 在两种语向下表现稳定，体现出较好的跨语向适配能力，为智能测评系统的标准迁移提供支持。

总体而言，本研究为理解人机评分机制与智能评测应用提供了实证依据。教学中，模型评分可作为译后反馈工具，辅助学生发现语言问题；教师可借助人机差异设计比较型任务，促进学生体裁意识与批判性思维。未来研究可扩大语料覆盖范围，并结合纵向追踪设计，探讨模型评分反馈在教学中的动态作用，为智能评估系统的优化与应用提供理论支撑。

致　谢

本研究得到北京林业大学 2025 年度大学生创新训练项目(项目编号：202510022116)的资助，在此谨致谢意。作者衷心感谢吕晓轩教授在本研究全过程中给予的悉心指导与宝贵意见。

基金项目

本研究系 2025 年大学生创新训练项目(项目编号：202510022116)阶段性成果，项目名称为“大语言

模型赋能英语笔译教学评价的实证研究”，依托单位为北京林业大学。

参考文献

[1] 何莲珍. 大语言模型在语言测评中的应用[J]. 外语教学与研究, 2024, 56(6): 903-912+960.

[2] 刘建达. 人工智能时代的语言测评: 机遇与挑战[J]. 现代外语, 2024, 47(6): 859-869.

[3] Hao, J., von Davier, A.A., Yaneva, V., Lottridge, S., von Davier, M. and Harris, D.J. (2024) Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. *Educational Measurement*: *Issues and Practice*, **43**, 16-29. https://doi.org/10.1111/emip.12602

[4] Mizumoto, A. and Eguchi, M. (2023) Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics*, **2**, Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

[5] Kwako, A., Wan, Y., Zhao, J., Hansen, M., Chang, K. and Cai, L. (2023) Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment? In: Kochmar, E., Burstein, J., Horbach, A., *et al*., Eds., *Proceedings of the* 18*th Workshop on Innovative Use of NLP for Building Educational Applications* (*BEA* 2023), Association for Computational Linguistics, 122-131. https://doi.org/10.18653/v1/2023.bea-1.54

[6] Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., *et al.* (2025) Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Applied Sciences*, **15**, Article 671. https://doi.org/10.3390/app15020671

[7] 苏祺. 大语言模型在二语教学中的应用效能解析[J]. 外语界, 2024(3): 35-42.

[8] Kirwan, A. (2023) ChatGPT and University Teaching, Learning and Assessment: Some Initial Reflections on Teaching Academic Integrity in the Age of Large Language Models. *Irish Educational Studies*, **43**, 1389-1406. https://doi.org/10.1080/03323315.2023.2284901

[9] Qin, H. and Lu, Y. (2024) The Application of Large Language Models in Foreign Language Education: An Exploration Based on Language Abilities. *Foreign Language World*, **6**, 37-44.

[10] Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y. and Wang, R. (2025) Evaluating Large Language Models as Raters in Large-Scale Writing Assessments: A Psychometric Framework for Reliability and Validity. *Computers and Education*: *Artificial Intelligence*, **9**, Article 100481. https://doi.org/10.1016/j.caeai.2025.100481

[11] Kelly, D. (2005) A Handbook for Translator Trainers. St. Jerome Publishing.

[12] Colina, S. (2015) Fundamentals of Translation. Cambridge University Press. https://doi.org/10.1017/cbo9781139548854

[13] Hu, W. (2018) Revisiting Translation Quality Assurance: A Comparative Analysis of Evaluation Principles between Student Translators and the Professional Trans-Editor. *World Journal of Education*, **8**, 176-184. https://doi.org/10.5430/wje.v8n6p176

[14] Abanomey, A.A. and Almossa, S.Y. (2023) Translation Quality Assessment Practices of Faculty Members of Colleges of Languages and Translation in Arab Countries: An Exploratory Study. *Humanities and Social Sciences Communications*, **10**, Article No. 835. https://doi.org/10.1057/s41599-023-02352-z

[15] McNamara, T.F. (1996) Measuring Second Language Performance. Longman.

[16] Myford, C.M. and Wolfe, E.W. (2004) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, **5**, 189-227.

[17] Eckes, T. (2015) Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments. 2nd Edition, Peter Lang.

[18] Erguvan, I.D. and Aksu Dunya, B. (2020) Analyzing Rater Severity in a Freshman Composition Course Using Many Facet Rasch Measurement. *Language Testing in Asia*, **10**, Article No. 1. https://doi.org/10.1186/s40468-020-0098-3

[19] Bouwer, R., Béguin, A., Sanders, T. and van den Bergh, H. (2015) Effect of Genre on the Generalizability of Writing Scores. *Language Testing*, **32**, 83-100. https://doi.org/10.1177/0265532214542994

[20] Jeong, H. (2017) Narrative and Expository Genre Effects on Students, Raters, and Performance Criteria. *Assessing Writing*, **31**, 113-125. https://doi.org/10.1016/j.asw.2016.08.006

[21] Jia, J., Wei, Z., Cheng, H. and Wang, X. (2023) Translation Directionality and Translator Anxiety: Evidence from Eye Movements in L1-L2 Translation. *Frontiers in Psychology*, **14**, Article ID: 1120140. https://doi.org/10.3389/fpsyg.2023.1120140

[22] 王湘玲, 王律, 郑冰寒. 翻译方向对信息加工过程及质量的影响——基于眼动和屏幕记录等数据的多元互证[J]. 外语教学与研究, 2022, 54(1): 128-139.

[23] Han, C., Hu, J. and Deng, Y. (2023) Effects of Language Background and Directionality on Raters' Assessments of Spoken-Language Interpreting. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, **36**, 556-584. https://doi.org/10.1075/resla.21009.han

[24] Qu, Y. and Wang, J. (2024) Performance and Biases of Large Language Models in Public Opinion Simulation. *Humanities and Social Sciences Communications*, **11**, Article No. 1095. https://doi.org/10.1057/s41599-024-03609-x

[25] Chang, V.C. and Chen, I. (2023) Translation Directionality and the Inhibitory Control Model: A Machine Learning Approach to an Eye-Tracking Study. *Frontiers in Psychology*, **14**, Article ID: 1196910. https://doi.org/10.3389/fpsyg.2023.1196910

[26] Wiseman, C.S. (2012) Rater Effects: Ego Engagement in Rater Decision-Making. *Assessing Writing*, **17**, 150-173. https://doi.org/10.1016/j.asw.2011.12.001

[27] Eckes, T. (2012) Operational Rater Types in Writing Assessment: Moving toward a Theory of Rater Cognition. *Language Testing*, **29**, 381-402.

[28] Kim, H.J. (2015) Rater Effects in L2 Writing Assessment: The Role of Rating Experience and L1 Background. *Assessing Writing*, **26**, 1-15.

[29] Winke, P., Gass, S. and Myford, C. (2013) Raters' L2 Background as a Potential Source of Bias in Rating Oral Performance. *Language Testing*, **30**, 231-252. https://doi.org/10.1177/0265532212456968