

大语言模型的日语敬语使用能力研究

——以 DeepSeek、ChatGPT 和 ELYZA LLM for JP 为例

西北师范大学 外国语学院 李 瑶

[摘要] 敬语一直是中国日语学习者的难点之一,如何活用大语言模型赋能敬语教学已成为亟待解决的问题。为此,本文构建了“词汇-语体选择-语体转换”三维敬语研究框架,通过实证研究比较分析了 DeepSeek、ChatGPT 和 ELYZA LLM for JP 在敬语形式识别、语义理解与语用推理中的表现。结果显示,DeepSeek 与 ChatGPT 的敬语使用能力整体上优于 ELYZA LLM for JP。三种大语言模型在敬语形式识别与语义理解方面表现较佳,能够较为准确地识别敬语词汇、语体类别及其语义特征,但在语用推理与语境适应方面仍显不足。最后根据前期研究结果,提出了大语言模型赋能日语敬语教学的建议。

[关键词] 语义-语用互补 大语言模型 敬语 比较研究

引 言

以 ChatGPT 为代表的大语言模型(Large Language Models, LLMs)基于深度学习的自然语言处理技术(Natural Language Processing, NLP),可以从大规模文本数据中学习丰富的语言知识和语言模式,能够对自然语言的语义、语法等进行理解和生成(车万翔、窦志成、冯岩松,2023),能够通过“提示+指令微调+人类反馈”的方式,完成多种不同的任务。大语言模型自面世以来被广泛应用于语言教学领域(焦建利、陈婷,2023)。其不仅能够模拟真实语境下的语言交互,还可通过动态生成对话与纠错反馈等方式,为学习者提供沉浸式、个性化的语言习得环境(苏祺,2024;崔希亮,2024)。进一步而言,大语言模型在语法纠错、语言翻译、智能写作与写作评分、智能语伴与口语评估方面展现出接近人类教师的潜力,但在语用推理(李瑶、于富喜、毋育新,2024)和语境理解方面(苏祺,2024)仍存在局限。

如何活用大语言模型赋能日语教学是当前亟需解决的问题之一。尽管已有研究指出,大语言模型在处理日语词汇、语法等语言知识时准确性较高(毛文伟、谢冬、郎寒晓,2023),但是掌握其在语用推理中呈现出的特点,改进其不足之处也是语言教学领域的重要课题。

在日语教学领域,敬语一直是中国日语学习者的难点之一^[1]。日语敬语作为一种程式化的语言表征,其核心功能在于准确标示交际双方或交际者与话题中涉及的人、事、物之间的关系,从而实现交际效果的礼貌性,促进交际的顺利进行(Brown & Levinson, 1987)。然而,受限于课堂时间短、真实语境缺失等原因,敬语教学长期面临“形式讲解多、语境演练少”“输入不足、纠错滞后”等问题,学习者在真实交际中难以灵活运用敬语相关知识。

随着人工智能技术的发展,大语言模型在上下文理解、语体转换与语用标记识别等方面的能力不断增强,为处理敬语这类复杂的语用现象提供了坚实的技术基础。大语言模型已初步实现对

敬语表达的理解与识别(李瑶等,2024),显示出其在语用领域的潜力。将大语言模型引入日语敬语教学现场,不仅可以解决“形式讲解多、语境演练少”“输入不足、纠错滞后”等问题,在实践层面也具有较高的可操作性。一方面,大语言模型可以通过模拟多种语境、实施角色扮演、提供即时反馈等方式,帮助学习者理解和使用敬语;另一方面,大语言模型处理敬语使用的过程也能提升其语用推理能力,进而实现语言教学与人工智能技术的双向赋能与融合发展。

无论是作为教师的教学助手,还是学习者的智能语伴,大语言模型都需要精准地掌握敬语知识,准确理解和生成相关内容。然而,目前对大语言模型的敬语使用能力研究仍处于初步探索阶段,尤其对其敬语形式识别、语义理解与语用推理能力缺乏系统的评估与实证研究。

因此,本研究从语义-语用互补视角出发,结合日语敬语教学研究现状,搭建大语言模型敬语使用能力的研究框架,并选取三种代表性大语言模型——多语言大模型 ChatGPT、以中文为主兼具日语处理能力的 DeepSeek 和专为日语优化设计的 ELYZA LLM for JP(简称 ELYZA),考察它们在日语敬语使用方面的表现,为实现大语言模型赋能日语敬语教学提供理论支持与实践参考。

1 文献综述

敬语作为日语交际的核心表达手段,不仅承载着丰富的社会意义,也在构建、维持和巩固人际关系中发挥着关键作用。在当今社会,敬语已成为人工智能领域的重要议题之一(宫本友树、片上大辅、重光由加,2019)。对中国日语学习者而言,敬语习得一直是难点,其原因不仅在于敬语体系中词形变化等复杂的语法规则,更在于学习者需要准确识别会话参与者之间的人际关系,并根据交际场合的正式程度灵活地选择适切的敬语表达(毋育新,2019)。此外,敬语的适切使用还要求学习者掌握符合日语语境的礼貌策略(毋育新,2015)。

针对敬语难的问题,相关研究或基于教学实

验持续探索有效的敬语习得方法(张晓宁,1995;张敏伶、冯良珍,2002;吴少华,2002;周莉,2004;杨宁,2005等),或在日语敬语教学现场引入礼貌策略理论(Universal Theory of Politeness)(Brown & Levinson, 1987)和话语礼貌理论(Discourse Politeness Theory, DP 理论)(宇佐美まゆみ,2002、2017)尝试构建适合中国日语学习者的敬语教学指导方略(毋育新,2008、2013、2015、2019)。虽然上述研究为提高学习者的敬语习得效果提供了理论支持与实践路径,但总体上仍囿于传统课堂教学路径,在教学资源丰富度、语境演练多样性与个性化反馈方面存在不足。

自 ChatGPT 问世以来,如何活用大语言模型赋能外语教学备受关注。学界普遍肯定了大语言模型在文本理解与自然语言生成方面的能力,并探讨其在课堂辅助、写作纠错与交互训练等教学环节中的潜在应用价值(张震宇、洪化清,2023;李佐文,2024;陈新仁,2024等)。然而,现有研究多集中于英语教学领域(焦建利、陈婷,2023;秦洪武、鲁艳芳,2024;郑咏滢,2024等),在日语教学中的系统探索仍较为有限。

毛文伟等(2023)探讨了 ChatGPT 在日语教学中的文本理解与生成能力及其在个性化学习与反馈方面的成效。曹长春(2023)指出可将人工智能应用于日语多元化口语教学、个性化评估、学情分析与智能情绪识别等方面。吴菲(2023)则基于传统商务日语课堂存在的问题,指出将人工智能应用于商务日语课的优势并提出对应的教学改革建议。上述研究对大语言模型赋能日语教学做了初步探讨,但都聚焦于宏观层面的理论分析,尚未深入揭示日语语言结构与交际特征对大语言模型在教学应用中所提出的特殊挑战。

进言之,从语言类型学的角度来看,相较于语序较为固定、词形变化较少、语境依赖度较低的英语,日语语序灵活、词形变化丰富且具有高度依赖语境的特点,标示人际关系与社会距离的语言系统也更为复杂,其中以成体系化的敬语表达最具代表性。鉴于此,本研究围绕当前大语言模型能否准确识别敬语形式、能否准确理解敬语语义、能

否恰当推断敬语功能三个问题,对其敬语使用能力展开专门研究。

2 理论框架的构建

语义学和语用学都关注语言的意义,但是其聚焦维度不同^[2]。语义学强调词义与句义之间的系统性和组合规律,通常抽象于语境之外;语用学则聚焦语言在具体语境中的使用与理解,强调说话人意图、交际规范与社会互动的作用(利奇,2020;陈娟,2021;冉永平等,2021)。敬语恰是语言中语义与语用高度交织的表达类型,其在形式层面体现为编码人际关系的结构体系(井出祥子,2006),在功能层面扮演着调控人际关系的重要角色(毋育新,2013)。

在中国日语教学一线,学习者既对敬语功能缺乏准确的认识,也经常无视语体选择,原因之一在于汉语敬语和日语敬语在词汇层面和语体层面存在差异^[3]。据此,本文从语义-语用互补视角出发,参考毋育新(2013)的敬语分类,构建一个包含“词汇”“语体选择”与“语体转换”的三维分析框架,通过融合语言形式的编码系统与交际实践中的动态策略,评估大语言模型的日语敬语使用能力并为其赋能日语敬语教学提供理论依据。

(1) 词汇维度

《敬语使用指南》(『敬語の指針』,文化审议会,2007)将敬语分为尊他语(「尊敬語」)、自谦语I(「謙讓語I」)、自谦语II(「謙讓語II」)、礼貌语(「丁寧語」)、美化语(「美化語」)五类。词汇维度关注大语言模型在具体语境中理解和使用尊他语、自谦语I、自谦语II和美化语的形式特征,并探讨其在日语语境中的语用功能和交际者的社会地位编码,进而检验大语言模型识别敬语形式的能力。

(2) 语体选择维度

日语语体系统有敬体(即礼貌语)和简体两大范畴。敬体包括「です・ます」体、「であります」体和「ございます」体,简体包括「だ」体和「である」体(苏德昌,1982)。语体选择受交际参数的制约,需要综合考量交际场合(公共/私密)、人际关系

(上下关系/亲疏关系)等多重因素的影响。语体选择维度聚焦敬体与简体的形式呈现与语境适应性,重点分析不同因素如何影响语体选择,语体选择如何体现交际行为中的角色定位与策略调节。通过形式与语境间的匹配度分析,评估大语言模型理解敬语语义的准确性。

(3) 语体转换维度

在实际交际场景中,语体并非静态存在,而是随着语境的变化不断调整,此为语体转换(苏德昌,1982;毋育新,2013)。其可分为敬体转为简体(「ダウン・シフト」,下行转换)和简体转为敬体(「アップ・シフト」,上行转换)(毋育新,2013)。前者通常具备五种语用功能,即①提出不同的立场或意见(「対立する立場や意見の提出」),②指责对方、口无遮拦的发言(「相手への非難や突き放す発言」),③对第三者表达不满(「第3者に関する悪口や噂話」),④开启/结束会话(「会話開始/終了時」),⑤提起新话题(「新しい話題の提示」)(笔者译)(宫武,2007),后者则有三种语用功能,即①向听话人表达亲密(「聞き手に親しさを表す」),②活跃交际气氛(「当時の場の雰囲気を生き生きとしている」),③表达说话人的强烈情感(「話し手の強い気持ちを表す」)(笔者译)(铃木,1997)。语体转换聚焦大语言模型是否理解语体转换的语用功能,此维度兼顾敬语的动态性与策略性,检测大语言模型是否具备良好的语用推理能力。

结合“词汇”“语体选择”“语体转换”三个维度构建的分析框架,不仅能系统呈现敬语的形式特征、语义结构与语用机制,也为评估和优化大语言模型在敬语使用中的表现提供具有可操作性的分析路径。该框架既可用于识别文本中的敬语偏误,也可为调整大语言模型、优化指令提供理论支撑和评估标准(见下页图1)。

3 大语言模型的敬语使用能力研究

3.1 实验设计

基于先前学习者存在的问题点,从李瑶(2016)和毋育新(2019)的研究中遴选50道测试题,涵盖

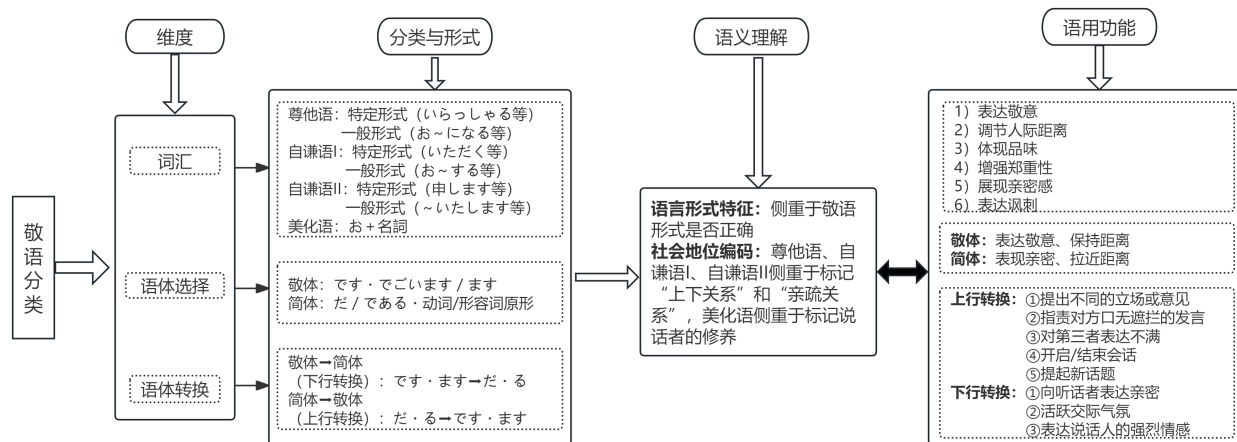


图1 “词汇-语体选择-语体转换”三维分析框架

三部分內容,即敬語形式正誤判斷(20道題,考察敬意是否失衡、表達是否恰當、形式是否準確)、語體選擇(20道題,考察敬體和簡體的選擇)和語體轉換(10道題,考察敬體轉簡體和簡體轉敬體)。通過設計提示語以檢測大語言模型的敬語使用能力。在測試之前,3名日語教師(1人為日語母語者)對題目內容進行審查並統一問題的答復,確保每道題都有清晰的人際關係和具體的交際場景,避免大語言模型因語境不清晰而生成錯誤內容。同時通過反復試驗,確定能穩定完成測試題的提示詞(prompt)模板,最終提示詞模板如下。

問題一

次の下線部の表現が適切かどうかを判断してください。適切であれば○をつけ、不適切であれば×をつけ、適切な表現に修正した上で、意味論と語用論の観点からその理由を説明しなさい。(20問)

1、(事務室に知らない学生が二人います。あなたが佐藤先生に言います。)明日、山田先生がコンピューターを使われます。

この「使われます」が敬語として適切かどうかを判断してください。

問題二

スピーチレベルには、敬体と常体の二種類が

あります。以下の場面と人間関係に応じて、最も適切な言い方を一つ選びなさい。(20問)

1、場面:学校に来る途中で、王さんは清水先生に会った。

王:清水先生、おはようございます。

清水:あ、王さん、おはよう。

王:先生は、朝ごはんは(1)

A 食べたか。B 食べますか。C 食べましたか。D 召し上がったか。

清水:ええ、食べたよ。王さんは?

王:私も食べました。

問題三

スピーチレベル・シフトはダウン・シフトとアップ・シフトに分類されます。アップ・シフトには、①対立する立場や意見の提出、②相手への非難や突き放す発言、③第三者に関する悪口や噂話、④会話開始/終了時、⑤新しい話題の提示という五つの機能があります。「ダウン・シフト」には、①聞き手に親しさを表す、②当時の場の雰囲気を生き生きとしている、③話し手の強い気持ちを表すという三つの機能があります。以下の場面と人間関係に応じて、その文脈でのスピーチレベルシフトに、最も適切な機能を一つ選びなさい。(10問)

1、場面:夫婦がデパートで話している。

妻:あら、これ、可愛いわ。買しましょう。

夫:これと同じようなのが、うちにいくつもあるじゃないか。

妻:でも、同じじゃないわ。少し違うわ。

夫:ぼく、もうお金持っていないよ。

この会話で、「買しましょう」という話はどんな機能をもっているか。

A 新しい話題を提起する B 会話を開始する

C 相手を非難する D 対立する意見を提出する

3.2 数据收集

选定 DeepSeek(V3)、ChatGPT4o、ELYZA LLM for JP(デモ版)三种大语言模型,将题目输入其中。评测显示,输入一套完整试题时,大语言模型往往会出现内容简化、信息遗漏以及语义混淆等问题,影响输出的准确性和稳定性。通过反复测试,最终确定采用分批输入的方式,具体操作步骤如下:

(1)开启新对话,输入问题一,分2次完成,每次只输入10道题,及时将生成的内容复制到文档中;

(2)输入问题二,分2次完成,每次只输入10道题,及时将生成的内容复制到文档中;

(3)输入问题三,及时将生成的内容复制到文档中;

(4)将3个问题的内容汇总成一份答案,存档备用。

每个大语言模型重新开启30次对话,每次对话按照上述流程进行测试并收集测评结果,从2025年4月15日至4月28日,共收集90份有效结果。

3.3 数据分析

依据标准答案对 DeepSeek、ChatGPT 和 ELYZA 在 30 次测试中生成的答案进行人工评分,总共 50 道题,每题 2 分,满分为 100 分,正确选项得 2 分,错误为 0 分。评分工作由 1 名中国籍日语教师和 1 名日本籍日语教师独立进行,如有分歧,协商统一意见后记录得分。最后统计出各模型在 30 次作答中的总分,并计算出总平均分,结果如表 1 所示。

因三种大语言模型的总分满足方差齐性要求,故使用 R 语言(4.5.1 版,下同)对三种大语言模

型在 30 次日语敬语使用测试中的总分进行单因素方差分析(One-way ANOVA)。结果显示,至少有两种大语言模型的总分存在显著性差异: $F_{(2,87)}=155, p<0.001$ 。进行 Tukey HSD 事后检验,发现 DeepSeek 与 ChatGPT 存在显著性差异($MD=3.27, p<0.001$), ELYZA 分别与 ChatGPT($MD=-9.27, p<0.001$)和 DeepSeek($MD=-12.53, p<0.001$)存在显著性差异。简言之,三种大语言模型在敬语使用测试中的表现存在显著性差异,其中 DeepSeek 表现最佳,ChatGPT 次之,ELYZA 表现最差。结合方差分析结果与效应量 $\eta^2=0.78$ 来看,大语言模型的类型对测试结果具有显著影响。

基于大语言模型的得分数据,调用 R 语言的 ggplot2 包绘制箱线图,以呈现三种大语言模型在敬语使用测试中得分的整体趋势与离散分布(见下页图 2)。

由图 2 可以看出,DeepSeek 的得分中位数最高,分布较为集中,其上下四分位数之间的箱体较窄,显示其性能稳定且优异。ChatGPT 的得分中位数略低于 DeepSeek,得分范围也稍宽,表现略有波动。而 ELYZA 的得分明显低于前两者,箱体极为紧凑,几乎没有离群点,说明尽管其得分集中且性能在样本中较为一致,但是整体敬语使用水平偏低。

下文将详细解释词汇、语体选择和语体转换三个维度的具体差异。

词汇维度有 20 道测试题,DeepSeek、ChatGPT 和 ELYZA 分别做了 30 次,每次都是重新开启新对话生成答案,它们在词汇维度的平均分如表 2 所示。

表 1 大语言模型的总平均分

大语言模型	总平均分(满分 100 分)
DeepSeek	87.27
ChatGPT	84.00
ELYZA	74.73

表 2 大语言模型在词汇维度的平均分

大语言模型	平均分(满分 40 分)
DeepSeek	35.00
ChatGPT	35.13
ELYZA	28.00

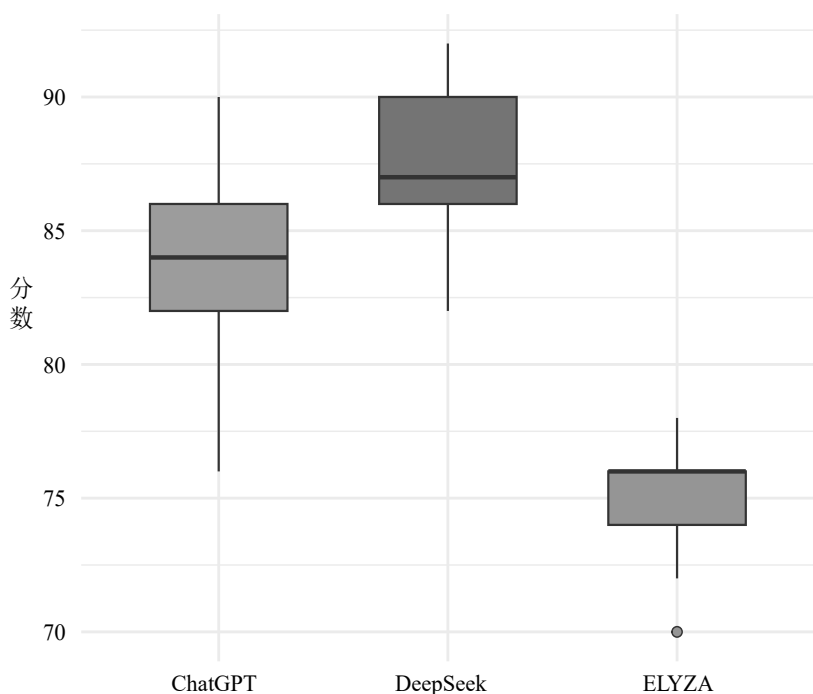


图2 大语言模型敬语使用能力测试的整体趋势

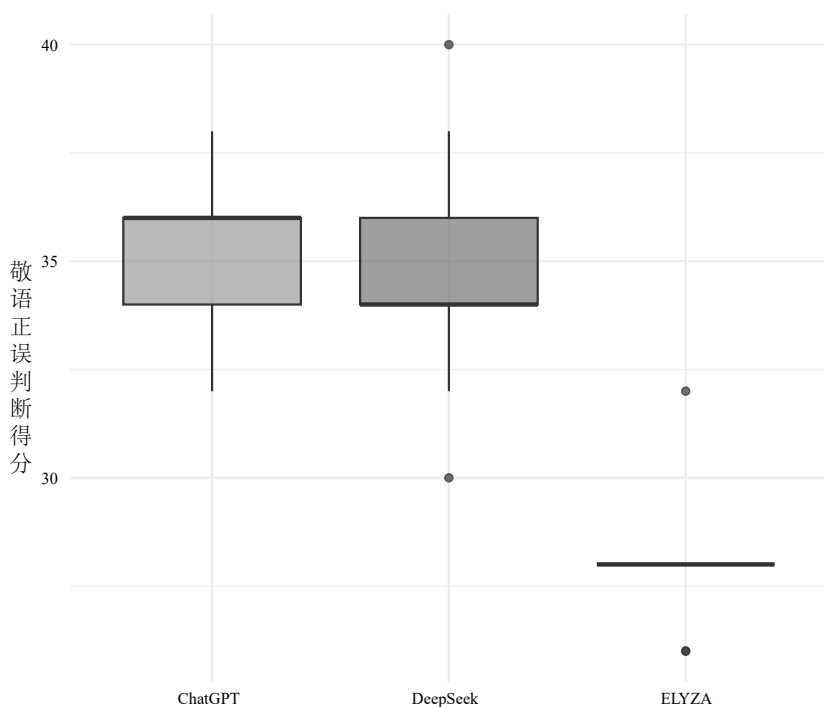
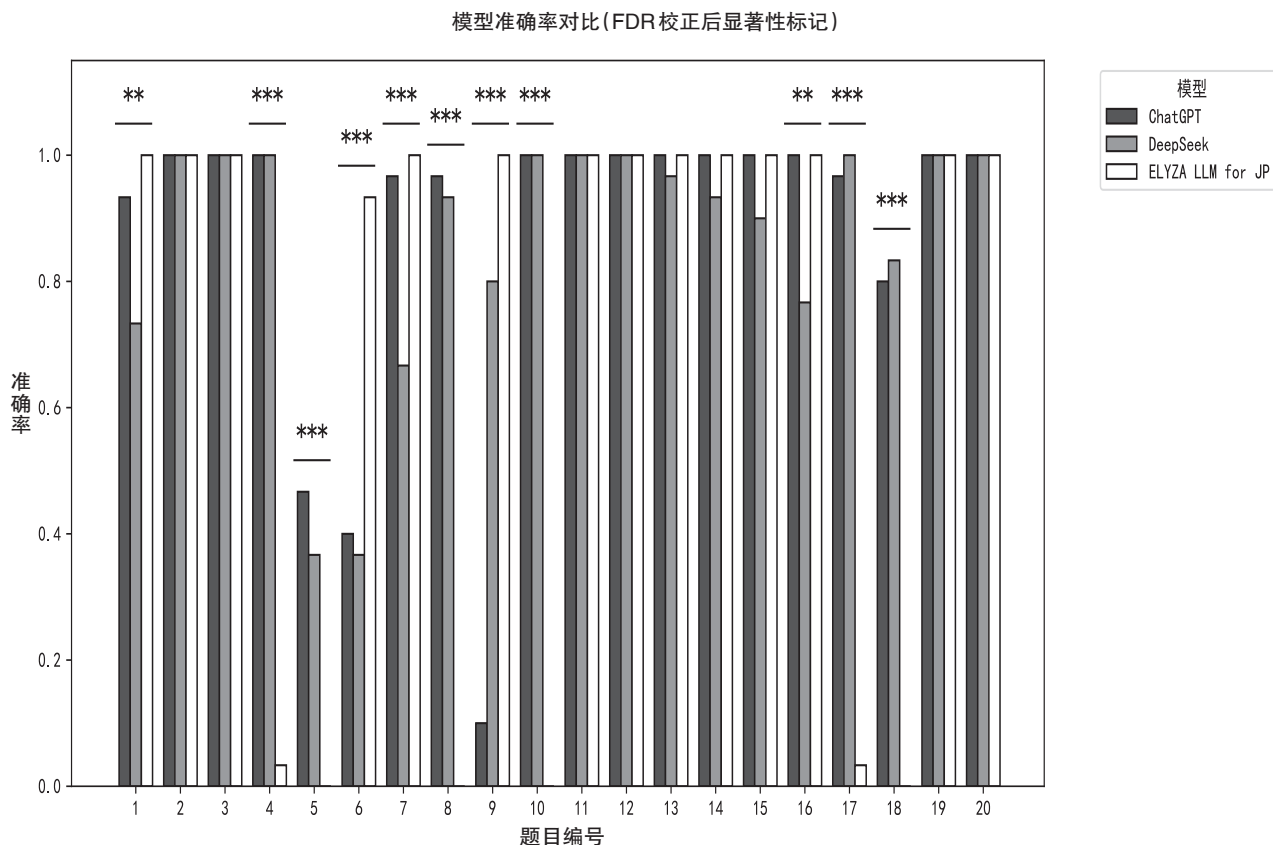


图3 大语言模型敬语形式正误判断得分分布

ChatGPT在词汇维度即敬语形式正误判断测试中的得分最高, DeepSeek次之, ELYZA略低。考虑到数据不满足方差齐性要求,使用R语言中的非参数统计检验函数`kruskal.test()`进行三种大语言模型的显著性差异分析,结果显示至少有两种大语言模型存在显著性差异: $\chi^2=62.39, df=2, p<0.001$ 。经Dunn事后检验并采用Benjamini-Hochberg校正,发现DeepSeek、ChatGPT与ELYZA三者之间均存在显著差异($p<0.001$)。由R语言`ggplot2`包绘制的箱线图(图3)可见,ChatGPT的中位数得分为36分,DeepSeek为34分,二者表现相对接近且均处于较高水平。尽管如此,DeepSeek得分的波动幅度较大,存在两个异常值(30分和40分)。而ELYZA得分明显较低,且数据分布集中于低分区间,表现明显不如另外两种模型。该箱线图直观反映出在“敬语形式正误判断”测试中,ChatGPT和DeepSeek不仅表现优于ELYZA且稳定性更高。结合 p 值远小于显著性水平($\alpha=0.05$)和效应量 $\eta^2=0.69$ 来看,大语言模型对敬语形式正误判断的表现具有显著影响。

为进一步探讨三种大语言模型在具体题目中的差异,使用R语言中的非参数统计检验函数`kruskal.test()`对各题目进行显著性差异分析,进行Dunn事后检验并采用Benjamini-Hochberg校正,最后调用`ggplot2`将最终结果汇总成可视化图(见下页图4)。

由图4可知,DeepSeek、ChatGPT与ELYZA在第1题和第16题



注:*** $p < 0.001$ (极显著);** $0.001 \leq p < 0.01$ (非常显著);* $0.01 \leq p < 0.05$ (显著)

图4 大语言模型敬语正误判断准确率对比图

存在非常显著的差异($p < 0.01$),在第4题至第10题以及第17题和第18题存在极为显著的差异($p < 0.001$)(见下一页表3)。

结合语义-语用互补的理论框架,围绕形式识别是否准确、语义理解是否到位和语用推理是否恰当,分析三种大语言模型在敬语形式正误判断中的表现可以看出,在形式识别方面,在第1题中,ELYZA 准确率达到 100%,ChatGPT 为 93.33%,DeepSeek 为 73.33%,组间效应量 $\eta^2=0.69$ 呈现出明显的模型间差异,DeepSeek 将敬语表达「使われます」误改为「使います」未能正确识别该语境下应使用尊他语,而 ChatGPT 与 ELYZA 均成功选择了尊他语「お使いになります」或「使われます」,表现出更强的敬语形式识别能力。在第7题中,ELYZA 的准确率为 100%,ChatGPT 为 96.67%,DeepSeek 为 66.67%,组间效应量 $\eta^2=0.19$ 呈现出中等程

度的模型间差异,DeepSeek 的错误回答均认为「お聞きしました」不正确并修正为「伺いました」,表明其在该语境中的敬语形式识别能力不如 ELYZA 和 ChatGPT。

在语义理解方面,多道题目反映出三种大语言模型的敬语语义理解能力存在差异。第5题中,DeepSeek 的准确率为 36.67%,ChatGPT 为 46.67%,ELYZA 为 0%,组间效应量 $\eta^2=0.18$ 呈现出中等程度的模型间差异,ChatGPT 和 DeepSeek 能正确识别面向外人时不应对家人使用尊他语「ご存じます」,将其修改为更符合语义规范的「知っています」,而 ELYZA 并未做出修正。在第8题中,ChatGPT 的准确率为 96.67%,DeepSeek 为 93.33%,ELYZA 为 0%,组间效应量 $\eta^2=0.86$ 呈现出显著的模型间存差异,ChatGPT 与 DeepSeek 均准确识别出「お貸しました」虽为自谦语 I,但其语义指向说话

表3 大语言模型在敬语使用正误判断中存在显著性差异的题目

题号	具体内容
1	(事務室にほかの学生が二人います。あなたが佐藤先生に言います。)明日、山田先生がコンピューターを 使われます 。
4	あなたが、クラスメートの前で、山田先生にたずねます。)山田先生、テストを 集めましたか 。
5	(あなたが先輩の佐藤さんと居酒屋で酒を飲んでます。あなたが佐藤先輩に言います。)私の父は、その話を ご存じです 。
6	(あなたが、喫茶店で、山田先生にたずねます。)山田先生は、ケーキを お召し上がりますか 。
7	(あなたと佐藤先生の二人だけが事務室にいます。あなたが佐藤先生に言います。)私は、昨日、山田先生からそのことを お聞きしました 。
8	(あなたが、家で、お父さんに言います。)山田先生は、昨日、その本を私に お貸しました 。
9	(あなたが田中君と一緒に電車に乗っています。あなたが田中君に言います。)私の父は、毎朝、新聞を ご覧になります 。
10	(事務室にほかの先生が二人います。あなたが事務室で、山田先生に言います。)山田先生、昨日、私に本を さしあげまして 、ありがとうございました。
16	(あなたが、クラスメートの前で、病気で悩む山田先生に言います。)山田先生、 少し休憩なさってはいかがですか 。
17	(あなたが、授業中に、山田先生にたずねます。)山田先生は、昨日、そのことを 私に言いましたか 。
18	(あなたが山田先生と一緒に学術会議に参加します。会場に行く途中で、あなたが山田先生に言います。)山田先生、私がかばんを 持ちます 。

人是动作实施者,这与该语境中“老师借书给我”的事实不符,均将其修正为尊他表达「貸してくださいました」,而 ELYZA 未能识别该表达在当前语境的语义错误,识别为正确表达。在第9题中, ELYZA 的准确率为100%, DeepSeek 为80%, ChatGPT 仅为10%, 组间效应量 $\eta^2=0.63$ 呈现出显著的模型间差异, ELYZA 与 DeepSeek 能正确识别「ご覧になります」作为尊他语不能用于指涉家人动作,两者将其修正为中性表达「読みます」或「見ます」,而 ChatGPT 错误地将「ご覧になります」用于指涉家人动作,导致语义错误。在第10题中, DeepSeek 与 ChatGPT 的准确率均为100%, ELYZA 为0%, 组间效应量 $\eta^2=1$ 呈现出极为显著的模型间差异。ChatGPT 和 DeepSeek 正确判断出「さしあげる」是自谦语I,用于指涉说话人自身的动作,而本题语境中的动作施事者是“老师”,故将其修正为尊他语「くださる」,而 ELYZA 未能识别出该表达的语义与指涉对象不一致,认为「さしあげる」是正确表达。

在语用推理方面,三种大语言模型表现出明显的差异。在第4题中, DeepSeek 与 ChatGPT 的准确率为100%, ELYZA 仅为10%, 组间效应量 $\eta^2=0.95$ 呈现出显著的模型间差异, DeepSeek 与 Chat-

GPT 均准确识别出「集めましたか」在该语境中敬意不足,将其修正为尊他语「お集めになりましたか」,而 ELYZA 仅有1次识别出该表达敬意不足。在第6题中, ELYZA 的准确率达100%, 而 ChatGPT 和 DeepSeek 的准确率均低于40%, 组间效应量 $\eta^2=0.33$ 呈现出中等偏上的模型间差异, ELYZA 正确识别出「お召し上がりますか」为过度敬语,将其修正为「召し上がりますか」,避免了语用层面的失礼,而 ChatGPT 和 DeepSeek 未能有效识别出其为双重敬语。在第16题中, ChatGPT 与 ELYZA 的准确率均为100%, DeepSeek 为76.67%, 组间效应量 $\eta^2=0.14$ 表明三者在该题上虽存在差异,但程度较低, ChatGPT 与 ELYZA 准确识别出「休憩なさってはいかがですか」在该语境中是对老师表达关心时的尊敬表达,既符合语义上的敬意指向,也符合语用礼貌规范,而 DeepSeek 虽能从语义上判断「なさる」是尊他语,但在形式层面认为作为名词的「休憩」与「なさる」搭配并不自然。在第17题中, DeepSeek 的准确率为100%, ChatGPT 为96.67%, ELYZA 仅为3.33%, 组间效应量 $\eta^2=0.90$ 呈现出极为显著的模型间差异, DeepSeek 与 ChatGPT 均正确判断出学生在课堂中提及老师的言语行为时,使用「言いましたか」敬意不足,故将其修

正为尊他语「おっしゃいましたか」,而 ELYZA 错误地认为「言いましたか」在该语境中为正确表达。在第 18 题中,DeepSeek 的准确率为 83.33%, ChatGPT 为 80.00%, ELYZA 为 0%, 组间效应量 $\eta^2=0.58$ 呈现出显著的模型间差异, DeepSeek 和 ChatGPT 准确判断出面对长辈提及自己的行为时,应使用自谦语 I, 故将其修正为「お持ちします」,以体现谦逊和礼貌,而 ELYZA 认为「持ちます」在该语境中为正确表达。

在敬语形式正误判断中,ChatGPT 不管是在形式层面,还是在语义和语用层面,都表现出较为灵活且准确的识别与理解能力,也能准确修正不当表达; DeepSeek 虽然在部分题目中表现较弱,但在语义理解和语用推理方面仍具一定优势; ELYZA 在形式识别上较为稳定,但在语义理解和语用推理方面存在不足。

语体维度有 20 道测试题, DeepSeek、ChatGPT 和 ELYZA 分别做了 30 次,每次都是重新开启新对话生成答案。它们在语体维度的平均分如表 4 所示。

表 4 大语言模型在语体维度的平均分

大语言模型	平均分(满分 40 分)
DeepSeek	35.67
ChatGPT	32.67
ELYZA	27.87

DeepSeek 在语体选择维度的准确率最高, ChatGPT 次之, ELYZA 略低。考虑到数据不满足方差齐性要求,使用 R 语言中的非参数统计检验函数 `kruskal.test()` 进行三种大语言模型的显著性差异分析。结果显示,三种大语言模型在语体选择上存在显著差异: $\chi^2=64.79$, $df=2$, $p<0.001$, 经 Dunn 事后检验并采用 Benjamini-Hochberg 校正,发现 DeepSeek 的语体选择得分显著高于 ChatGPT ($Z=3.06$, $p<0.001$), ChatGPT 也显著高于 ELYZA ($Z=4.92$, $p<0.001$)。综合来看, DeepSeek 在语体选择方面的表现优于 ChatGPT 和 ELYZA, 而 ELYZA 的语体选择较为保守, 得分相对较低。结合 p 值远小于显著性水平 ($\alpha=0.05$) 和效应量 $\eta^2=0.72$

来看,大语言模型的类型对敬语语体选择测试的表现具有显著影响。

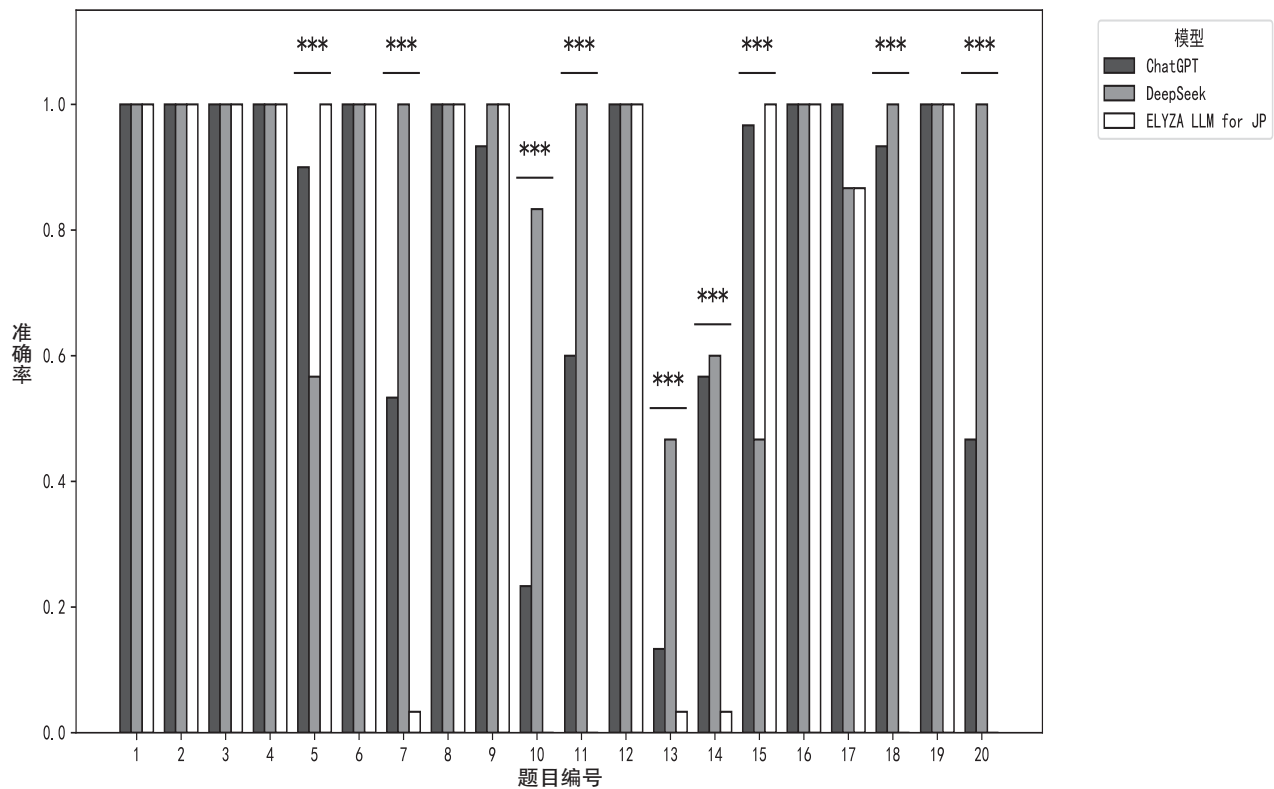
为进一步探讨三种大语言模型在具体题目中的差异,使用 R 语言中的非参数统计检验函数 `kruskal.test()` 对各题目进行显著性差异分析并调用 `ggplot2` 绘制可视化图(见下页图 5)。

由图 5 可知, DeepSeek、ChatGPT 与 ELYZA 在第 5 题、第 7 题、第 10 题、第 11 题、第 13 题、第 14 题、第 15 题、第 18 题和第 20 题中存在极为显著的差异 ($p<0.001$) (见第 25 页表 5)。

结合语义-语用互补的理论框架,从形式识别、语义理解和语用推理三个维度分析大语言模型在语体选择中的表现。形式识别层面聚焦于模型是否能根据语境选择合适的语体、避免语法不当或敬语误用。在第 13 题中,三种大语言模型的效应量 $\eta^2=0.19$ 呈现出显著的模型间差异, DeepSeek 的表现优于其他模型(准确率 46.67%), 选择了「行きましたか」这一规范的敬体表达, ChatGPT (13.33%) 与 ELYZA (3.33%) 频繁使用「行った」「行ったんだ」等简体或口语句式,反映出其在句式风格和语境匹配度方面仍然不稳定。第 14 题进一步验证了此差异,三种大语言模型的效应量 $\eta^2=0.26$ 呈现出显著的模型间差异, DeepSeek (60%) 与 ChatGPT (56.67%) 在会议语境下多选择敬体句式, 而 ELYZA 准确率仅为 3.33%, 大量使用简体表达「行った」,反映出其对语体的判断能力明显不足。

语义理解层面强调大语言模型对句子主语(说话人/听话人/第三人称)及语义关系(自谦/尊他)理解的准确性,涵盖第 5 题和第 7 题。第 5 题考察私密语境中谈论上司动作的敬语表达,三种大语言模型的组间效应量 $\eta^2=0.22$ 呈现出显著的模型间差异, DeepSeek 准确率仅为 56.67%, 误选尊他语「出席される」,反映出在主语身份识别上的偏差, ChatGPT (90%) 与 ELYZA (100%) 能正确选择「出席する」,说明二者在语义理解上更为清晰。第 7 题涉及向外人提及家人时应使用自谦语 I, 三种大语言模型的组间效应量 $\eta^2=0.62$ 呈现出显著的模型间差异, DeepSeek 表现出色, 均正确选择「おります」; ChatGPT (53.33%) 和 ELYZA (3.33%)

模型准确率对比(FDR校正后显著性标记)



注:*** $p < 0.001$ (极显著);** $0.001 \leq p < 0.01$ (非常显著);* $0.01 \leq p < 0.05$ (显著)

图5 大语言模型语体选择准确率对比图

则误用尊他语「いらっしゃいます」,说明二者在内外区分与语义角色区分上存在误差。

语用推理层面关注大语言模型在交际场景中是否能选择最符合日本社会文化规范的表达方式,包括第10题、第11题、第15题、第18题和第20题。第10题与第11题考察称谓语的恰当性,其效应量分别为 $\eta^2=0.53$ 和 $\eta^2=0.67$ 呈现出显著的模型间差异,具体而言,DeepSeek准确率分别为83.33%与100%,稳定选择中性或敬意适切的称谓(如「課長の加藤」或「加藤課長」);而ChatGPT和ELYZA选择「加藤課長さん」「加藤様」等双重敬语,呈现出礼貌等级判断错误的倾向。第15题与第18题涉及陌生人初次对话时是否能使用得体的请求表达。在第15题中,三种大语言模型的组间效应量 $\eta^2=0.37$ 呈现出显著的模型间差异,Chat-

GPT的准确率为96.67%,ELYZA为100%,DeepSeek仅为46.67%,其错误选项使用称谓语「おじさん」,可以说在一定程度上受到汉语负迁移影响。在第18题中,三种大语言模型的组间效应量 $\eta^2=0.90$ 呈现出极为显著的模型间差异,DeepSeek(100%)与ChatGPT(93.33%)在均选择了符合日语语言习惯的「ビックカメラっていう店に行く道を教えていただけませんか」,反映出对语用规范的正确把握,而ChatGPT和ELYZA的错误选项增加了称呼语「おじさん」,不符合日语表达习惯。第20题测试借钱场景中的礼貌表达,三种大语言模型的组间效应量 $\eta^2=0.66$ 呈现出显著的模型间差异,DeepSeek(100%)与ChatGPT(46.67%)能使用婉转表达「あ、千円でいいの」对听话人表达顾及,而ELYZA和ChatGPT的部分测试误选

表5 大语言模型在语体选择中存在显著性差异的题目

题号	具体内容
5	<p>場面:同期入社の子友である村上さん(男性,30代)と山田さん(男性,30代)の食堂での会話。 村上:今日の会議、疲れたね。来週も同じ会議、やるかな。 山田:やるだろう。 村上:それなら、社長も(5)かな。 A出席する B出席される C出席されます D出席します</p>
7	<p>場面:学生の柳が山口先生の家で電話する。 柳:もしもし、山口先生のお宅ですか。 奥さん:はい、山口でございます。 柳:横浜大学日本語学部の柳と申しますが、山口先生は(6)。 Aいますか Bおられますか Cいらっしゃいますか Dいるか 奥さん:はい、(7)。少々お待ちください。 Aいます Bおられます Cいらっしゃいます Dおる</p>
10	<p>場面:お得意先からかかってきた電話にヤマダ電機の小川さんが出た。 小沢:恐れ入りますが、(10)いらっしゃいますか。 A加藤 B加藤課長 C加藤課長さん D課長の加藤様</p>
11	<p>小川:申し訳ございません、(11)は、あいにく席を外しておりますが。 A加藤 B課長の加藤 C加藤課長さん D課長の加藤様</p>
13	<p>場面:会社の取締役が出席している会社の経営に関する会議での、村上さん(男性,30代)と山田さん(男性,30代)の会話。 村上:山田さん、昨日、どこへ営業に(13)。 A行った B行ったの C行ったんだ D行きましたか 山田:昨日ですか。横浜のほうへ行きました。</p>
14	<p>場面:学術会議の秘書を務める鈴木さん(男性,30代)は同僚の高橋さん(男性,20代)と会議での会話。 鈴木:高橋さん、昨日、会場へ(14)。 A行きましたか B行った C行ったの D行ったんだ 高橋:はい、行きました。</p>
15	<p>場面:伊藤さん(女性,20代)が駅の前でパトロール中の50代の警察官に道を尋ねる会話。 伊藤:(15) Aあのう、国会図書館、どこ。 B国会図書館どうやって行くの。 Cすみません、国会図書館に行きたいんですが。 Dおじさん、国会図書館に行く道を教えていただけませんか。 警察:向こうに銀行が見えますね。国会図書館は銀行の向こう側です。</p>
18	<p>場面:伊藤さん(女性,20代)が道端で道路工事をしている中年男性に道を尋ねる時の会話。 伊藤:あのう、すみませんが、(18) Aビックカメラっていう店、どうやっていくの。 B 師匠、ビックカメラっていう店行く道を教えていただけませんか。 Cおじさん、ビックカメラっていう店に行く道を教えていただけませんか。 Dビックカメラっていう店に行く道を教えていただけませんか。 中年男性:ほら、向こう側にある高いビルがビックカメラですよ。</p>
20	<p>場面:中村京子さん(女性,20代)と会社の先輩の川田理香さん(女性,30代)との会話。 中村:川田先輩、ちょっと、財布を忘れたんですけど、千円貸してくれませんか。 川田:(20) Aあ、千円でいいの。 Bあ、そうですか。 Cあ、わかりました。 Dあ、千円ですか。 中村:ええ。どうもありがとうございます。 川田:いいえ、どういたしまして。</p>

「わかりました」,虽语义正确却语用功能不当,未能顾及听话人心理。

在语体选择中,DeepSeek 最稳定,在敬语形式结构、语义理解和语用推理方面均优于 ChatGPT 与 ELYZA。ChatGPT 具有中等表现而 ELYZA 在三个维度的表现普遍较差,特别是在语用策略的灵活使用上表现出较为明显的机械性。

语体转换维度有 10 道测试题,DeepSeek、ChatGPT 和 ELYZA 分别做了 30 次,每次都是重新开启新对话生成答案,它们在语体转换维度的平均分如表 6 所示。

表 6 大语言模型在语体转换维度的平均分

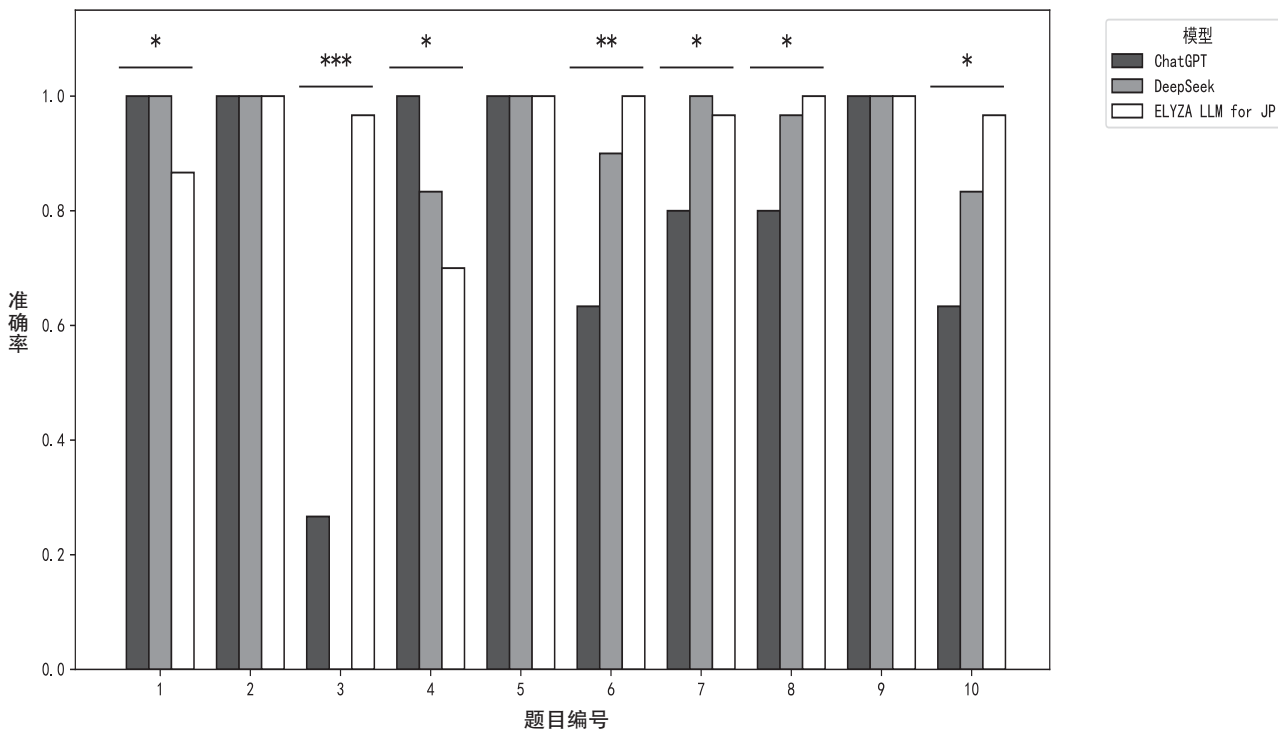
大语言模型	平均分(满分 20 分)
DeepSeek	17.06
ChatGPT	16.20
ELYZA	18.87

ELYZA 在语体转换维度的准确率最高,Deep-

Seek 次之,ChatGPT 略低。由于语体转换得分不满足方差齐性要求,使用 R 语言中的非参数统计检验函数 `kruskal.test()` 进行三种大语言模型的显著性差异分析。结果显示,至少有两种大语言模型存在显著性差异: $\chi^2=25.826, df=2, p<0.001$ 。经 Dunn 事后检验并采用 Benjamini-Hochberg 校正,发现 ChatGPT 的得分显著低于 ELYZA ($Z=-4.75, p<0.001$) 和 DeepSeek ($Z=-0.83, p<0.001$),且 DeepSeek 的得分也显著低于 ELYZA ($Z=-3.93, p<0.001$)。综合来看,在此次语体转换测试中,ELYZA 的表现显著优于 DeepSeek 和 ChatGPT,DeepSeek 又优于 ChatGPT。结合 p 值远小于显著性水平 ($\alpha=0.05$) 和效应量 $\eta^2=0.27$ 来看,大语言模型类型对语体转换的表现具有显著影响。

为进一步探讨三个大语言模型在具体题目中的差异,使用 R 语言中的非参数统计检验函数 `kruskal.test()` 对各题目进行显著性差异分析,并调用 `ggplot2` 绘制可视化图(见图 6)。

模型准确率对比(FDR校正后显著性标记)



注:*** $p<0.001$ (极显著);** $0.001\leq p<0.01$ (非常显著);* $0.01\leq p<0.05$ (显著)

图 6 大语言模型语体转换准确率对比图

由图6可知,DeepSeek、ChatGPT与ELYZA三种语言模型在第1题、第4题、第7题、第8题和第10题存在显著差异($p<0.05$),在第6题存在非常显著的差异($p<0.01$),在第3题存在极为显著的差异($p<0.001$)。此处重点分析存在非常显著和极显著差异的题目(见表7)。

结合语义-语用互补的理论框架,从形式识别、语义理解和语用推理三个维度分析大语言模型在语体转换中的表现。在第3题与第6题中,三种大语言模型的效应量分别为 $\eta^2=0.68$ 与 $\eta^2=0.16$,呈现出显著的模型间差异。进一步来说,三种模型在形式识别层面整体表现良好,均能较为准确地识别出从简体到敬体的语体变化,说明它们对句末形式变化(如「ましょう」「じゃないですか」等)的敏感度较高。从语义理解层面来看,在第3题中,「買いましょう」虽为敬体形式,但本质上具有开启会话的语义,ELYZA与ChatGPT能正确把握其在该语境中的含义,而DeepSeek将该表达误解为提起新话题或提出对立意见,反映出其在语义联结与上下文解读方面存在偏差。在第6题中,DeepSeek和ELYZA识别出「乗ればいいじゃないですか」通过敬体句式传达出一种带有轻微反驳的语气即表达意见上的对立,而ChatGPT在部分测试中并未识别出这一点。从语用推理层面

来看,第3题中的「買いましょう」虽是礼貌表达,但其真正的语用功能在于发起会话、推动互动,DeepSeek未能准确识别这一点。在第6题中,ChatGPT在部分测试中未能深入理解其在语用层面上作为策略性礼貌表达、缓和语气的功能。

在语体转换中,ELYZA的表现最稳定,在敬语语用功能识别方面表现出较强的能力;ChatGPT与DeepSeek在形式识别和基础语义理解方面尚可,但在语用推理方面存在不足。

4 讨论

4.1 大语言模型在敬语使用中的优势与不足

本研究聚焦《敬语使用指南》中的敬语五分法,围绕“大语言模型能否准确识别敬语形式、能否准确理解敬语语义、能否恰当推断敬语功能”三个问题,对ChatGPT、DeepSeek和ELYZA三种大语言模型展开敬语使用能力测试研究。

在形式识别层面,大语言模型对典型敬语结构的识别率准确度较高,但存在依赖规则化的倾向,这一特征符合毛文伟等(2023)的研究结论。具体而言,三种大语言模型均能较为准确地识别常见的敬语形式,尤其在识别典型的尊他语(如「ご覧になる」)和自谦语(如「伺う」「申し上げ

表7 大语言模型在语体转换中存在显著性差异的题目

题号	具体内容
3	場面:夫婦がデパートで話している。 妻:あら、これ、可愛いわ。買いましょう。 夫:これと同じようなのが、うちにいくつもあるじゃないか。 妻:でも、同じじゃないわ。少し違うわ。 夫:ぼく、もうお金持ってないよ。 この会話で、「買いましょう」はどんな作用をもっていますか。 A 新しい話題を提起する B 会話を開始する C 相手を非難する D 対立する意見を提出する
6	場面:実家に帰る交通ツールについての雑談。 田中:向こう雪降ってないらしいね。 中村:まじ。 田中:暴風雨とかで。 田中:そんな積もってないらしいよ。 中村:降ってほしいよな。 田中:私、車、乗りたいんだけど。 中村:乗ればいいじゃないですか。 田中:だって雪降ったら乗れないじゃん。 中村:大丈夫だって。 この会話で、「乗ればいいじゃないですか。」はどんな作用をもっていますか。 A 新しい話題を提起する B 会話を開始する C 相手を非難する D 対立する意見を提出する

る」)时,呈现出高准确率,同时也具备语体识别能力,能够区分「です・ます」体与「だ」体等基本语体。在敬语形式出现错误(如双重敬语、尊他语与自谦语互相误用)时,大语言模型能够通过语义线索与句法结构进行有效修正,表明其具备相对成熟的敬语形式系统识别能力。然而,在具体语境中,大语言模型对一般敬语形式(如「お~する」「~れる・られる」)的识别能力较弱,如在敬语形式正误判断的第8题中,ELYZA忽视该语境中指涉话题人物(老师)的动作时应该使用尊他语的要求,认为自谦语「お貸ししました」在该语境下是正确表达。此外,当敬语嵌入在长句时,大语言模型对敬语形式的整合能力有限,如DeepSeek在语体选择的第15题中误将D选项「おじさん、国会図書館に行く道を教えていただけませんか」认定为正确答案。

在语义理解层面,大语言模型能够区分尊他语、自谦语I、自谦语II、美化语,明确掌握基本的敬语语义并能识别基本的语义意图,如“表达尊敬”“表现谦逊”“避免失礼”等。在涉及上下位语义区分,如敬语形式正误判断的第13题、第14题(「いただく」与「もらう」)时,三种大语言模型的表现尤为出色,准确率均达到100%,呈现出良好的敬语语义体系建构能力。DeepSeek和ChatGPT表现出较强的语义识别能力,能够较为准确地识别特定敬语表达在句法与语义层面中的对应关系。然而,在更为隐性的“言外之意”层面,三种大语言模型的表现存在局限。相较于ELYZA,DeepSeek和ChatGPT能结合具体的语境需求和人际关系,准确理解敬语语义并对不当表达进行修正,但是通过语体转换测试可以看出,ChatGPT难以区分“表达亲密”与“活跃气氛”等语义相近但功能有别的策略性表达,表明其对细微语义差别及不同语境中言外之意的把握较为有限。另一方面,ELYZA在多道测试题中对敬意程度把握不到位,忽视了敬意失衡(敬意过低或敬意过高)可能导致双方距离疏远或听话人感觉不自然的语用效果,且存在难以准确判断“非敬语形式”在特定语境中发挥“顾及”作用的问题(如语体选择的第20题),反映

出其在语义理解层面明显缺乏对“言外之意”的整合能力。简言之,当前的大语言模型已具备基本的敬语语义理解能力,尤其在形式语义识别和上下位语义区分方面表现优异,在深入理解敬语所承载的“言外之意”方面,DeepSeek和ChatGPT的表现明显优于ELYZA,后者仍有较大提升空间。

在语用推理层面,三种大语言模型的表现均不如形式识别与语义理解。首先,从人际关系的角度来看,ChatGPT与DeepSeek对上下关系和亲疏关系表现出一定的敏感性,能较为准确地区分“家人之间不宜使用尊他语”“对上级需使用敬语”等基本语用规则;而ELYZA在多数语境中停留在形式识别层面,无法结合人际关系和语境需求判断使用何种敬语表达。其次,从语境适应性的角度来看,三种大语言模型的语用适应能力存在较大差异。DeepSeek能够较好地匹配敬语形式与语用功能,但在陌生人会话场景中有时会出现不符合语境需求的人称代词(如语体选择的第15题),可能是受到母语负迁移的影响。ChatGPT与ELYZA则常出现误用敬语、过度使用敬语的现象,反映出其对日语语用规范掌握不足。值得注意的是,ELYZA虽为日本本土大语言模型,但更易出现敬意不足、双重敬语等敬语误用现象,表明其缺少语用推理能力和语境适应性的训练。第三,从语用功能的识别来看,三种大语言模型均难以区分语体转换的具体功能。具体而言,难以识别下行转换中“向听话人表达亲密”与“活跃交际气氛”之间的细微差异,这可能与下行转换的高语境敏感性有关。同时,大语言模型也无法准确识别上行转换中的“提出不同意见”“指责对方”“提起新话题”,其原因可能在于大语言模型在处理复杂的语境变化时,未充分考虑到交际者的意图变化。这反映出大语言模型在处理高语境敏感性的语用现象(如语体转换)时,缺乏对会话参与者交际意图的准确把握,亟需在日后训练中加入更多标注人际关系和具体语境信息的会话数据。

综上所述,三种大语言模型在敬语使用测试中呈现出清晰的层级化特征:形式识别能力较强,语义理解能力中等,语用推理能力较弱。这一特

征表明,大语言模型对敬语的理解主要源于训练数据中的形式规则与语义标签,缺少对社会语用规范、语境适应性以及话语礼貌策略的深度学习。

4.2 大语言模型赋能日语敬语教学的建议

为充分发挥大语言模型在日语敬语教学中的有效应用,依据前文对其形式识别、语义理解和语用推理三个方面的考察结果,结合当前敬语教学所面临的问题即教学资源丰富度、语境演练多样性与个性化反馈存在不足,从教学资源生成智能化、语境演练多样化、学习路径个性化以及教学环境人本化四个角度出发,提出大语言模型赋能敬语教学的建议。

(一)人机协同推动教学资源智能生成:基于大语言模型在敬语形式识别与语义理解方面的优势,可系统创建覆盖五种敬语的智能化语料库,涵盖例句提取、误用筛查与多语言敬语对比分析等不同资源类型。在此基础上,提出递进式、分层化的教学路径:面向初级日语学习时,活用大语言模型辅导学习者理解基本的敬语形式和语义,并指导学习者借助大语言模型进行敬语形式的正误检测与纠错训练;面向中级日语学习者时,利用大语言模型的语义联想功能与上下位语义区分功能,强化学生对仿真语境中敬语表达的正确理解与使用;面向高级日语学习者时,聚焦语用推理,引导学习者批判地识别与理解大语言模型生成的敬语相关表达中出现的社会语用规范与话语礼貌策略,提升学习者对语用适切性的把握。

(二)人机协同丰富语境演练:尽管当前大语言模型在敬语的语用推理与言外之意识识别方面尚存不足,但是教师也可以充分发挥大语言模型的人机交互优势,活用大语言模型模拟不同人际关系和交际场景生成的会话,指导学习者分析仿真会话中的敬语表达,判断其是否契合交际需求,识别其中的语用偏差并尝试提出修改意见,从而提升学习者的语用推理能力、语用策略使用能力以及对语境的敏感性。

(三)人机协同实现学习路径个性化:学习者可以借助大语言模型在实时响应与语义分析方面

的技术优势,构建专属于自己的敬语学习平台,围绕敬语偏误识别、自主演练反馈与表达优化建议等核心功能,开展目标导向型学习。学习者在与大语言模型的持续互动中,根据自身需求设定适合自己的学习路径,并通过大语言模型的反馈进一步反思、修正自己的敬语表达,逐步提升敬语的语用意识与表达适切性。

(四)人机协同优化敬语教学环境:在敬语教学中,教师、学习者与大语言模型应协同互补,共同推动教学范式的“智慧赋能”转型。教师应明确大语言模型在语用推理方面的局限,精心设计具有真实性和挑战性的教学任务,引导学生自主提升批判性思维。学习者在与大语言模型的交互中,通过设定个性化学习目标、诊断敬语偏误、优化敬语表达等实践,提升语用意识。同时,大语言模型的实时反馈机制可促使学习者在实践中加深对敬语形式规范与语用功能的理解。通过构建教师引领、学生主导、模型支持的三元协同机制,实现“以教促学、以智助人、以人为本”的深度融合。

5 结 语

本研究从语义-语用互补视角出发,构建了“词汇-语体选择-语体转换”三维敬语分析框架,比较了DeepSeek、ChatGPT与ELYZA三种大语言模型在敬语形式识别、语义理解以及语用推理中的表现。结果表明,三者对敬语形式识别与语义理解方面表现较好,能够较为准确地识别敬语词汇与语体类别及其语义特征,但在语用推理及语境适应性方面仍显不足,据此提出了大语言模型赋能日语敬语教学的建议。总体而言,三种大语言模型在敬语形式识别与语义理解方面具有较大潜力,可为日语敬语教学提供有力支持,鉴于DeepSeek与ChatGPT在敬语使用能力方面整体优于ELYZA,建议优先采用Deepseek和ChatGPT作为中国日语教学现场的敬语教学赋能工具。

[本文为西北师范大学2025年度青年教师科研能力提升计划项目“生成式人工智能语用原则产出的汉日对比研究”(项目编号:NWNU-SKQN2025-21)阶段研究成果。负责人:李瑶]

注

- [1] 毋育新. 日语敬语教学方略研究[M]. 北京: 北京大学出版社, 2019: 1.
- [2] 杰弗里·利奇. 语用学原则[M]. 冉永平, 译. 北京: 商务印书馆, 2020: 5.
- [3] 毋育新. 日语敬语教学方略研究[M]. 北京: 北京大学出版社, 2019: 40-41.

参考文献

- 曹长春. 人工智能技术下日语课堂教学改革之探索[C]//香港新世纪文化出版社. 2023年第六届智慧教育与人工智能发展国际学术会议论文集: 第三卷. 西安培华学院, 2023.
- 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学: 信息科学, 2023, 53(9).
- 冉永平, 等. 语用学十讲[M]. 上海: 上海外语教育出版社, 2021.
- 陈娟. 语用学和语义学的分工与合作: 基于使用的语言观视角[J]. 外语教学, 2021, 42(4).
- 陈新仁. ChatGPT与二语语用教学[J]. 外语教学理论与实践, 2024(4).
- 崔希亮. 人工智能——语言教学的机遇与挑战[J]. 华文教学与研究, 2024(2).
- 焦建利, 陈婷. 大型语言模型赋能英语教学: 四个场景[J]. 外语电化教学, 2023(2).
- 李瑶. 日语教科书中敬语的导入研究[D]. 西安: 西安外国语大学, 2016.
- 李瑶, 于富喜, 毋育新. 汉日语境中大语言模型的礼貌知识生成能力探析[J]. 日语学习与研究, 2024(5).
- 李佐文. ChatGPT赋能外语教学: 场景与策略[J]. 北京第二外国语学院学报, 2024, 46(1).
- 毛文伟, 谢冬, 郎寒晓. ChatGPT赋能新时代日语教学: 场景、问题与对策[J]. 外语学刊, 2023(6).
- 秦洪武, 鲁艳芳. 大语言模型与外语教育: 基于语言能力的应研探[J]. 外语界, 2024(6).
- 苏德昌. 试论文体的统一性[J]. 日语学习与研究, 1982(1).
- 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).
- 吴菲. 人工智能(AI)在商务日语教学中的应用探析[J]. 科技视界, 2023(5).
- 吴少华. 以语言交际为中心的敬语教学方法初探[J]. 外语教学, 2002(1).
- 毋育新. 日汉礼貌策略对比研究[M]. 北京: 中国社会科学出版社, 2008.
- 毋育新. 日语敬语的有标记性与无标记性研究——以语体转换为对象[J]. 东北亚外语研究, 2013, 1(1).
- 毋育新. 中国学生日语敬语习得问题点理论索据[J]. 外语教学, 2015, 36(2).
- 杨宁. 浅谈基础阶段日语教学中的敬语教学[J]. 北京第二外国语学院学报, 2005(6).
- 张敏伶, 冯良珍. 中日敬语文化对比及日语教学[J]. 山西财经大学学报(高等教育版), 2002(4).
- 张晓宁. 浅谈日语敬语教学[J]. 外语与外语教学, 1995(2).
- 张震宇, 洪化清. ChatGPT支持的外语教学: 赋能、问题与策略[J]. 外语界, 2023(2).
- 郑咏滢. 生成式人工智能在外语教育中的应用: 关键争议与理论构建[J]. 外语教学, 2024, 45(6).
- 周莉. 试论日语敬语表达在教学中的难点——以人际关系为中心[J]. 日语学习与研究, 2004(S1).
- 井出祥子. わきまへの語用論[M]. 東京: 大修館書店, 2006.
- 宇佐美まゆみ. ポライトネス理論の展開: 1-12[J]. 月刊言語, 2002(1-5, 7-12).
- 宇佐美まゆみ. ディスコース・ポライトネス理論の新展開: 「時間」「フェイス充足度」「フェイス均衡原理」という概念を中心に[C]//漢日対比語言学研究会. 漢日語言対比研究論叢: 第8号. 上海: 華東理工大学出版社, 2017.
- 鈴木睦. 日本語教育における丁寧体世界と普通体世界[C]//田窪行則. 視点と言語行動. 東京: くろしお出版, 1997.
- 宮本友樹, 片上大輔, 重光由加, 等. ポライトネス理論に基づく運転支援エージェントにおける発話の文末スタイルに着目した印象評価[J]. 知能と情報, 2019, 31(3).

宮武かおり. 日本人友人間の会話におけるポライトネス・ストラテジー: スピーチレベルに着目して[D]. 東京: 東京外国語大学, 2007.

文化審議会. 敬語の指針[R]. 東京: 文化庁, 2007.

Brown P, Levinson S. Politeness: Some Universals in Language Usage[M]. Cambridge: Cambridge University Press, 1987.

作者简介: 李瑶 女 汉族 西北师范大学外国语学院副教授 研究方向: 日语语言

联系方式: E-mail: liyao0826@foxmail.com

Approach to Japanese Honorific Usage in Large Language Models: Comparative Analysis of Deepseek, ChatGPT and ELYZA LLM for JP

Abstract: Honorific expressions have long posed a significant challenge for Chinese learners of Japanese. Effectively harnessing the capabilities of large language models (LLMs) to enhance Japanese honorifics instruction has therefore become a pressing issue. In response, this study proposes a three-dimensional analytical framework Vocabulary-Speech Level Selection-Speech Level Shift to conduct an empirical comparative analysis of DeepSeek, ChatGPT and ELYZA LLM for JP in terms of their performance in honorific form judgment, semantic recognition and pragmatic inference. The findings indicate that DeepSeek and ChatGPT exhibit stronger overall competence in honorifics usage compared to ELYZA LLM for JP. All three models demonstrate relatively robust performance in semantic recognition of honorifics, including the identification of honorific vocabulary, speech level categories and their semantic features. However, their abilities in pragmatic inference and contextual adaptation remain limited. Based on these results, the study proposes suggestions on how large language models can be utilized to support and improve the teaching of Japanese honorifics.

Keywords: semantic-pragmatic complementarity; large language models (LLMs); honorifics; comparative study

Author's Information:

Li Yao (Female)

Associate Professor at Northwest Normal University, China

Japanese Linguistics

E-mail: liyao0826@foxmail.com

(责任编辑: 李广悦)