ISSN：3080-2660
EISSN：3080-2679

大语言模型在日语 N1 听力中间接否定的识别能力研究

郭　静　郑晓丹　韩　叶

宁夏理工学院，宁夏 石嘴山 753000

摘　要：随着人工智能技术的飞速发展，大语言模型（Large Language Models, LLMs）在自然语言处理（Natural Language Processing, NLP）任务中展现出强大的潜力。本研究系统评估大语言模型（LLMs）在日语能力考试（JLPT）N1 听力材料中对“间接否定”的识别精度，并揭示其误差模式与改进路径。通过构建包含多种间接否定类型的 N1 听力语料库，并选用 Kimi 与 Chat-GPT4.0 两种模型进行对比实验，系统评估其在识别间接否定中的表现。实验结果显示，LLMs 已具备初步的日语间接否定识别能力，Kimi 模型在准确率、召回率与 F1 值方面均优于 Chat-GPT4.0，尤其在规约性间接否定识别中表现突出。然而，在涉及文化特定表达与复杂语言结构时，两类模型的识别准确率均出现显著下降。本研究为大语言模型在语言测试领域的应用提供了实证依据，为 LLMs 在语言测试领域的可信应用提供了基线与可解释性依据。

关键词：大语言模型；日本语能力考试 N1；间接否定；语言测试

0 引言

大语言模型在自然语言处理领域取得了显著进展，广泛应用于文本生成、机器翻译、情感分析等任务。但其是否满足日语高风险的语用测评需求仍有待验证。日本语能力考试（JLPT）N1 听力对测试者的语言理解和分析能力提出较高要求，其中间接否定的识别是主要难点之一。JLPT-N1 听力中的间接否定因“高语境、低显性”特征成为检验模型语用理解力的试金石。

本研究旨在系统评估大语言模型在识别 N1 听力中间接否定言语行为方面的能力，并探索有效的性能提升路径。通过构建涵盖 2004－2024 年共 21 套 N1 听力真题的语料库，对 279 个会话样本进行人工标注与模型识别对比，从准确率、召回率与 F1 值等指标出发，分析模型在处理不同类别间接否定时的表现差异，并为后续模型优化与教学应用提供参考。

1 间接否定的相关研究

Searle（1979）提出，间接言语行为是指“通过实施另一种次要的施事行为来间接实施某一种首要的施事行为”。间接否定言语行为则是指通过非直接方式传达否定意图的语言现象，常见于日常交流与跨文化语境中。Brown（1987）提出的“面子理论”进一步阐释了间接否定在维护社会关系中的重要作用。在职场与正式场合中，间接否定常以委婉拒绝、话题转移、条件句等形式出现，具有高度的语用复杂性。Holmes（1995）指出，职场交流中的间接否定通常呈现出高度的礼貌性和社交敏感度，而该语言策略的使用有助于减少可能的情感伤害和社会紧张。Norrick（2001）发现口语中的间接否定表达，通过疑问句和模糊词语的使用，可以在不明显拒绝对方提议的情况下，表达出不同的意见或意向。

2 研究设计与方法

2.1 数据采集

本研究以 2004－2024 年共 21 套 JLPT N1 听力真题为数据来源，筛选出 279 个双人会话作为研究样本。样本涵盖商务、日常聊天、购物等多种场景。排除了单人发言与因谦虚而进行的否定表达，最终将委婉拒绝、暗示性否定、条件性否定、模糊性否定、反问式否定、建议式否定与消极评价等七类纳入研究范围。

2.2 数据分析

首先对 N1 考试听力音频的文字材料进行标注，人工识别其中的间接否定。标注内容包括会话数、间接言语行为语句数、形式和字面意义、含意和命题态度、语境信息等。然后使用 Kimi 和 Chat-GPT4.0 模型对会话中的间接言语行为进行标注和识别，统计预测准确率、召回率和 F1 值。对出错或模型未进行标注的语句进行统计。最后对出错的会话进行重复测试，通过修改场景中的相关变量，判断影响语言大模型识别和标注的具体影响因素。

2.3 大语言模型选择

由于目前国内的大语言模型暂不支持分析日语相关文本，故本研究选用了 Kimi 和 Chat-GPT4.0 模型进行对比分析。Kimi 在自然语言处理领域表现出色，特别擅长情感分析和文本分

ISSN：3080-2660
EISSN：3080-2679

类等任务，能够准确捕捉文本中的细节和关键信息。Chat-GPT4.0具有强大的语言理解和生成能力，可以理解和生成各种自然语言文本，包括对话、文章、问题回答等。

2.4 数据预处理与模型训练

在本研究中，采用了基于自然语言处理（NLP）的机器学习模型来识别会话中的间接否定。具体而言，首先由两名日语母语者对所有会话进行人工标注，标注内容包括：会话编号；是否包含间接否定；否定类型；字面意义与实际含义；语境信息。标注一致性达到92%，分歧部分经讨论后统一。为了评估模型的表现，使用了三个常见的评估指标：准确率（Precision）、召回率（Recall）和F1分数（F1-score）。以下是模型在测试集上的性能表现：

表1 类别详细指标（1为Kimi，2为Chat-GPT4.0）

类别	P1	P2	R1	R2	F1-1	F1-2
委婉拒绝	0.89	0.85	0.87	0.87	0.88	0.86
暗示否定	0.86	0.86	0.84	0.84	0.85	0.85
条件否定	0.84	0.84	0.82	0.82	0.83	0.83
模糊否定	0.82	0.80	0.80	0.78	0.81	0.79
反问否定	0.75	0.60	0.65	0.60	0.66	0.60
建议否定	0.65	0.60	0.63	0.60	0.64	0.60
消极评价	0.64	0.64	0.62	0.62	0.63	0.63
平均值	0.78	0.74	0.75	0.73	0.76	0.74

3 结果与讨论

3.1 研究结果

本研究通过构建含多种间接否定表达的N1听力语料库，利用大语言模型Kimi和Chat-GPT4.0来处理语料，评估了其与人工识别的准确率、召回率和F1值。大模型在统计间接否定时，部分数据为对现实的间接否定，而非是对对方话语的间接否定，故在正阳性数据统计中去除。

Kimi模型的准确率为78%，而Chat-GPT4.0模型的准确率为74%。Kimi的准确率高于Chat-GPT4.0，这表明Kimi在识别间接否定行为时，正类预测中的正确预测比例更高，模型对于间接否定表达的判断更加精确。

Kimi的召回率为75%，Chat-GPT4.0为73%。同样，Kimi在召回率上也优于Chat-GPT4.0。这表明Kimi在识别实际存在的间接否定表达方面更为敏感，能够检测出更多的真实正类样本。Kimi在对语料库的训练和推理过程中，更加注重召回，即对间接否定表达的更多识别。它能够捕捉到更多的“模糊”或较难判断的间接否定行为，而Chat-GPT4.0可能在某些情况下错过了样本。

在本研究中，Kimi的F1值为76%，而Chat-GPT4.0为74%。这表明Kimi的综合性能也优于Chat-GPT4.0。

3.2 结论

研究结果表明，Kimi模型在识别间接否定方面表现出色，相较于Chat-GPT4.0模型，其预测总数、准确率、召回率以及F1值均有显著优势。Kimi模型的高准确率和召回率表明其在识别间接否定行为方面具有很强的能力，不仅能够准确判断正类样本，还能较好地识别出更多的间接否定表达。Kimi在准确率和召回率之间达到了较好的平衡，因此其F1值较高，综合性能更为优越。Chat-GPT4.0模型的准确率和召回率均低于Kimi，表明其在间接否定的识别方面存在一定的不足，在某些情况下出现误判。Chat-GPT4.0在处理某些间接否定表达时，容易发生假阳性或假阴性，导致预测结果的准确性受到影响。尽管两者均能有效处理含有间接否定的语料，但Kimi在模型的细节优化、数据处理和上下文理解上具有更强的优势，能够提供更为精确的结果。这表明，Kimi在实际应用中可能更适合用于处理复杂的语料，尤其是在涉及间接否定和复杂语境的任务中，具有较高的实用价值。

3.3 误差分析

通过对模型误分类样本的分析，发现模型的误差主要集中在以下两个方面：

（一）漏检：一些间接否定表达较为隐晦或委婉，模型未能准确识别。例如，在一些轻微的转移话题或提出替代方案的情况下，模型将其误判为无间接否定。例如：

女：あのう、先生、この建物の一階の裏口を出たところに、大きな蜂の巣ができているのを見つけたんですが…

男：ええ、そんなところに蜂の巣？すぐに事務室の人に連絡して、専門の業者に駆除してもらわないと。

女：そうなんですけど、もうこんな時間ですし……

在以上会话中，女性对男性提出的联系办公室人员让专业人员除去蜂巢的建议，提出了“已经到这个时间了”，潜在含义是对男性建议的间接否定。然而，Chat-GPT4.0未能检测出这一间接否定。

（二）误识别：部分不含否定意图的句子被误判为间接否定。例如，某些疑问句或表达不确定性的句子，模型错误地标记为间接否定。例如：

ISSN：3080-2660
EISSN：3080-2679

男：この企画はそういう初心者が気軽に参加できるものをと考えてるんだよ。外部の専門家にも入ってもらうから、専門的なことは心配しなくてもいいし。早速で悪いんだけど、頼んでもいいかな。

女：はい。

男：そこのファイルに近郊でツアーの候補地になってる山をリストアップしてあるから、そこから五つぐらいに絞り込んでくれる？自分でも登れそうだって思えることを念頭に、コースの魅力度でね。他社の類似ツアーの資料も入ってるから、それも参考にしながらね。

女：はい。（2019 年 12 月 N1 日本語能力試験）

在以上对话中，男性邀请女性加入登山旅行的策划，并为女性提出了许多参考和要求，其中并无间接否定出现，但 Kimi 却做出以下分析：男性通过“五つぐらいに絞り込んでくれる？”间接否定了女性可能认为“需要更多时间准备”的想法，暗示需要尽快完成任务。这一误差表明，模型在处理语言中的模糊性和多义性方面仍有一定的不足。

3.4 原因分析

3.4.1 文化差异

在处理文化特定的间接否定表达时，模型识别准确率较低。原因在于：一是文化特定的表达方式，不同文化背景下的语言表达和语用预设存在差异，日本的委婉表达如“それは難しいかもしれません”在其他文化语境中鲜见，模型训练时缺乏跨文化语料，导致对这类表达理解不足；二是文化背景知识的欠缺，模型需具备充足的文化背景知识才能准确理解特定文化表达，日本的礼貌用语和委婉表达常含隐含意义，模型若不了解相关文化背景，便难以精准识别；三是跨文化语料的不足，当前训练语料中跨文化部分相对较少，使模型在处理文化特定表达时缺乏训练基础，影响识别准确率。

3.4.2 语言结构特点

面对具有特殊语言结构的间接否定表达，如反问句、双重否定句等复杂句式，模型识别准确率同样不佳。原因在于：一是特殊语言结构本身的复杂性，这类结构增加了语义理解难度，模型在处理时难以精准捕捉否定意义，例如反问句“それは本当に必要ですか？”和双重否定句“それは必ずしも間違っていない”，其否定意义需结合语境和语义推理才能识别，模型容易误判；二是语法和语义规则的不足，现有模型在处理特殊语言结构时，缺乏足够的语法和语义规则支撑，无法准确识别反问句中的否定意义；三是训练数据的不足，训练数据中特殊语言结构的语料较少，导致模型在处理这些结构时缺乏训练基础，影响识别准确率。

4 结语

本研究通过实证分析，评估了大语言模型在识别 N1 听力中间接否定方面的能力。尽管大语言模型在处理直接言语行为方面表现出色，但在识别和标注间接否定时仍面临挑战，特别是在复杂语境和文化特定的表达中。通过偏差分析，提出了改进预训练方法、多模态融合和动态语境建模等未来研究方向，以提高模型对复杂语言现象的理解和适应能力。本研究不仅为大语言模型在语言测试领域的应用提供了实证依据，也为未来模型的优化和改进指明了方向，特别是在提高对复杂语言现象的理解和适应能力方面。

参考文献：

[1]Searle J R. Expression and Meaning: Studies in the Theory of Speech Acts[M]. Cambridge: Cambridge University Press, 1979.

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

[3]Alhamoud A, Alabdulkreem E, Alshehri A, et al. NegBench: A benchmark for evaluating visual-language models on negation[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2025: 1234-1244.

作者简介：郭静（1990.9—），女，汉族，宁夏人，硕士研究生，讲师，研究方向：日语语言学和日语教育。

项目信息：宁夏回族自治区教育厅高等学校科学研究项目资助，《基于中日对译语料库的规约性间接否定言语行为研究》（项目号：NYG2024250 ）。