

基于GPT-4o和提示词工程的日语隐喻自动标注实证研究

北京外国语大学 日语学院 北京日本学研究中心

韩 涛 杨言飞

[摘 要] 针对当前隐喻标注效率不足、人工成本高昂的现状,本文以日语隐喻为研究对象,探究了利用大语言模型GPT-4o实现高效、准确标注的路径。研究设计了一套轻量级隐喻标注框架,并通过一项实证性研究检验了提示词语种(日语/英语)、示例数量(4条/8条)与提示词位置(系统/用户)对标注性能的影响。实验结果表明,“日语提示词+8条示例+系统提示词”的组合策略能够达成最佳标注效果,其输出与人工标注结果高度一致。此外,研究还进一步厘清了GPT-4o在此任务中的能力边界:该模型在常规隐喻识别中展现出强大潜力,能显著提升标注效率并降低人工成本。然而,其在处理跨句隐喻以及精确辨析动词性与名词性隐喻类别时,仍面临一定挑战。

[关键词] 大语言模型 GPT-4o 隐喻 自动标注 提示词工程

引 言

隐喻作为人类语言中普遍存在的表达方式,不仅是文学修辞的重要手段(韩涛、孙玉涛,2025),更被视为一种基本的认知机制,深刻影响着人类的思维与概念系统(Lakoff & Johnson, 1980)。近年来,随着概念隐喻理论的深化与语料库语言学方法的普及,隐喻研究在理论建构与实证分析方面均取得了显著进展(侯学昌等,2022)。但与此同时,现有成果多建立于小规模语料分析之上,制约了面向大规模语言数据或自然语言处理(Natural Language Processing,以下简称NLP)应用的系统性探索。究其根源,是因为当前尚缺乏一套适用于语言学需求的、高效的隐喻自动识别与标注体系(Fuoli *et al.*, 2025)。

隐喻研究日益依赖语料库方法,而高质量的隐喻标注语料库更是自动化隐喻识别系统发展的基础。隐喻标注虽具有一般语料库标注的共性,但由于隐喻语言及其概念结构本身的特殊性与复杂性,也呈现出独特的属性(柳超健等,2018)。隐喻标注具有高度的抽象性与语境依赖性,不仅需识别词汇层面的隐喻用法,更要求对源域、目标域及其间的映射关系进行深层理解。传统上,这项工作主要依赖人工完成,虽能保证一定精度,但过程耗时耗力,且易受标注者主观判断影响。因此,如何突破人工标注的瓶颈,发展高效、一致的辅助性隐喻标注技术,已成为推动该领域发展的关键课题。

[收稿日期]2025-12-16

[基金项目]本文为2022年度教育部人文社会科学重点研究基地重大项目“汉外多语种词汇语法功能视角对比研究”(项目编号:22JJD740010)阶段研究成果;2024年度中央高校基本科研业务费专项资金资助自主选题重点课题“《习近平谈治国理政》多语隐喻平行语料库的研制与应用”(项目批准号:2024JJ015)阶段研究成果

[作者信息]韩涛,男,北京外国语大学日语学院 北京日本学研究中心副教授,研究方向:认知语言学、认知隐喻学、汉日对比、语言政策,电子邮箱:hantao@bfsu.edu.cn;杨言飞,男,北京外国语大学日语学院 北京日本学研究中心日语语言文学专业硕士研究生,研究方向:日语语言,认知语言学,电子邮箱:1137402745@qq.com

大语言模型(LLMs)以其高水平的语义推理能力和对话能力,受到学界广泛关注。与传统的递归神经网络(RNN)和长短期记忆网络(LSTM)不同,大语言模型所依赖的Transformer架构,其核心自注意力机制能直接建模序列中任意两个词之间的全局依赖关系,从而捕捉深层的句法与语义模式。得益于其并行计算特性,该架构在理论上能够更高效地处理长文本(Vaswani *et al.*, 2017)。基于这一特点,大语言模型对上下文的理解能力极大增强,使其在隐喻计算任务中具有独特的优势(Puraivan *et al.*, 2024)。尽管大语言模型已被广泛应用于隐喻识别、理解乃至生成等任务,并展现出巨大潜力,然而,其辅助性、高精度隐喻标注的能力边界、可靠性及内在认知机制,仍是当前亟待探索的核心课题(Fuoli *et al.*, 2025)。

基于上述背景,本文拟采用GPT-4o模型作为实验平台。该模型在多语言处理任务中表现出优于前代GPT-4的性能,同时其应用程序编程接口(Application Programming Interface,以下简称API)具备更高的经济性与实用性,为大规模语言处理应用提供了可行基础(OpenAI, 2024)。本文拟设计一套轻量级隐喻标注框架,考察GPT-4o在日语隐喻标注任务中的实际表现以及何种提示词工程策略能最有效激发其潜力,以验证其语义理解能力对辅助隐喻标注的实际效用,并探索高效提示策略,为日语隐喻研究与NLP应用的深度融合开辟新路径。

1 相关文献回顾

隐喻标注的理论基础可追溯至Lakoff & Johnson(1980)的概念隐喻理论,而后续操作化指南则主要由隐喻识别程序及其拓展版本MIPVU(Steen *et al.*, 2010)奠定,为人工标注提供了系统、可复现的标准。在具体实践中,为提高人工标注的效率与一致性,研究者们开发了一系列专门的标注工具、框架与流程。例如,柳超健等(2018)探讨了隐喻标注的原则、模式与方法,论证了可扩展标记语言在结构化标注中的优势,并提出了“人工识别+机器辅助”与“机器标注+人工辅助”的协同理念。张冬瑜(2019)、Shutova等(2017)、加藤祥、浅原正幸(2023)等学者则分别针对汉语、英语和日语提出了具体的隐喻标注框架与操作流程。这些工作均旨在通过规范化与半自动化的辅助手段,提升人工标注的质量与规模。然而,尽管有上述工具与框架的支持,传统以人工为核心的隐喻标注仍面临根本性挑战。标注过程通常需要预先制定详尽的规则并对标注者进行培训,导致流程繁琐、耗时且成本高昂,难以适应大规模语料处理的需求(Fuoli *et al.*, 2025)。近年来,亦有研究者尝试采用众包方式以降低标注成本,但由于标注者背景的多样性,往往引入标注标准不一致的新问题,反而可能影响数据的可靠性(姜宽, 2020)。

正是由于人工标注在大规模应用中的这些局限性,学界开始积极探索自动化与计算化的隐喻识别方法。早期研究主要基于规则系统或结构化知识库,例如Chung等(2004)通过查询WordNet语义与SUMO本体来识别隐喻源域,Ge等(2022)利用WordNet中的上位词关系辨识隐喻的源域与目标域,Rosen(2018)则基于既有隐喻数据集设计了从上下文和目标域预测源域的模式。

随着大语言模型的崛起,隐喻研究进入了新的范式。大语言模型通过海量文本预训练,能够捕捉语言中深层的语义与语用规律,在经过任务特定微调或指令微调后,在多项NLP任务中展现出强大性能(Yang *et al.*, 2024)。在隐喻研究领域,学者们已开始探索其应用潜力。譬如,Wachowiak & Gromann(2023)检验了GPT-3在英语和西班牙语中识别隐喻及预测源域的能力;Yuan等(2024)提出将隐喻理论构造成知识图谱,以增强大语言模型在隐喻推理任务上的表现;张声龙、刘颖(2024)通过微调大语言模型进一步提升了其对隐喻的识别与理解能力;而张亚洲等(2024)则评估了ChatGPT在情感分析与隐喻分析中的表现,指出其虽略逊于专业人类标注,但仍具备重要的辅助价值。

综上所述,现有基于大语言模型的隐喻研究多集中于英语、汉语等语种,针对日语的探索相对少见。我

们认为,当前的通用模型与英语范式难以处理日语的原因主要有三:1)高度语境化的句法,如日语主宾格助词(「が」「を」)的省略,使得隐喻的源域与目标域在句法层面的边界比英语更为模糊,增加了自动切分和识别的难度;2)独特的形态指标,如「のような」「あたかも」与英语的介词迥异,要求模型具备对日语形态学的深层理解;3)日语隐喻多根植于日本独特自然环境、历史社会、美学观念,具有文化负载性。此外,当前相关研究多聚焦于通过微调、检索增强生成等技术路径优化模型性能,这类方法往往需要较强的计算语言学背景与工程能力,对广大隐喻研究者而言存在较高的技术门槛。相比之下,提示工程因其无需训练、灵活可解释的特性,成为驱动闭源模型完成专业任务的更可行方案。因此,基于提示词工程进行隐喻标注可以降低应用复杂度,开发对语言学研究更加友好的研究工具与方法论路径。

2 标注框架和提示词工程

2.1 标注框架

我们结合前人研究为语料库中的隐喻标注提供了一种简洁、统一的格式,并增加了标注材料中的隐喻类型一项,用于判断该隐喻为指标隐喻(使用“指示词”来明示比喻关系的隐喻)还是结合隐喻(不依赖指示词,将两个概念直接结合的隐喻)(中村明,1977)。标注核心要素包括:目标域(Target)、源域(Source)、隐喻类别(Metaphor Category)和隐喻指示词(Indicator)(见表1、表2)。这套方案主要用于简化隐喻的标注流程,提升分析效率。

2.2 提示词工程

提示词工程(prompt engineering)亦称提示工程,指通过设计与优化输入指令(即提示词)以引导大语言模型生成所需输出的方法。根据实现形式,主要分为软提示(soft prompt)与硬提示(hard prompt)两大类(Liu *et al.*, 2023)。软提示指一组可学习的连续向量(或嵌入),但实际使用存在不可解释性和闭源大语言模型的不可更新性两方面问题。与之相对,硬提示即直接以自然语言形式构建的提示词,用户可直接编写与修改,从而克服了软提示的上述限制,成为驱动闭源模

表1 隐喻标注要素定义

标注要素	定义
句子(Sentence)	包含隐喻的完整句子。
目标域(Target)	隐喻的目标域词汇,即被描述的对象或概念。
源域(Source)	隐喻的源域词汇,即用来隐喻目标的已知概念。
隐喻类别(Metaphor Category)	名词隐喻(源域为名词)或动词隐喻(源域为动词)。
隐喻类型(Metaphor Type)	指标隐喻(有指示词)或结合隐喻(无指示词)。
指示词(Indicator)	喻词,如「ように」「まるで」等(若无明确指示词,该字段可以为空)。

表2 隐喻标注示例

【指标隐喻】 1.その日のうちの出来ごとは、いま思い出しましても夢のようござります。 (那天发生的事情,即便如今回想起来,也如同一场梦一般。) Annotation: <metaphor> <Sentence>その日のうちの出来ごとは、いま思い出しましても夢のようござります。</Sentence> <Target>その日の出来事</Target> <Source>夢</Source> <MetaphorCategory>名詞</MetaphorCategory> <MetaphorType>指標</MetaphorType> <Indicator>よう</Indicator> </metaphor>
【结合隐喻】 2.メロディーが心を絶望に追いやる。(旋律将心推入绝望。) Annotation: <metaphor> <Sentence>メロディーが心を絶望に追いやる</Sentence> <Target>メロディー</Target> <Source>追いやる</Source> <MetaphorCategory>動詞</MetaphorCategory> <MetaphorType>結合</MetaphorType> <Indicator></Indicator> </metaphor>

型的主要实践方式(雷鸣,2025)。因此,本研究采用硬提示(以下简称为提示词)来驱动大语言模型,以辅助日语隐喻标注。

在提示词工程中,根据是否提供示例及示例数量,存在零样本(zero-shot)、单样本(one-shot)和少样本(few-shot)等不同策略。其中,少样本(few-shot)策略指模型接收少量的任务示例,这些示例帮助模型理解任务的模式和要求。在少样本设置下,模型可以通过看到一些不同例子,更好地调整输出。例如:

「毒が血管のなかで運動をする。」→含隐喻
「眼の光が家の冬の気の根源。」→含隐喻
「GPT-4は大規模言語モデルです。」→无隐喻
请判断以下句子是否含有隐喻:「誠実を計量する。」

考虑到本文对大语言模型的输出有框架要求,且涉及两种隐喻类别,本文将采用少样本的设置进行测试,同时将示例控制为偶数,以便覆盖两种隐喻类别。

3 实验设计

3.1 总体设计

本实验旨在评估不同提示词设置下,GPT-4o模型在日语隐喻标注任务中的表现。研究重点关注提示词语种、示例数量和提示词位置三个变量对模型准确率的影响。实验采用2(语种:日语/英语)×2(示例数量:4条/8条)^[1]×2(提示词位置:系统设定(system)/用户输入(user content))的三因素二水平被试间设计。具体如下:

因素一:提示词语种。该因素控制了提示词的语种,具体为日语或英语(见表3)。由于各语种训练集在预训练阶段占比的不同,提示词语言选择可能会影响模型在具体任务上的表现。因此,本文在使用日语提示词进行测试的同时,也需测试英语提示词可能造成的不同表现。

表3 英语提示词和日语提示词

英语提示词	You are a linguist researching Japanese metaphors. Please annotate the provided sentences containing metaphors in XML format. Each annotation should include the following information: - Target: the object or concept being metaphorically described. - Source: the metaphorical concept used to describe the target. - MetaphorCategory: Whether the metaphor is nominal or verbal. - MetaphorType: Whether the metaphor in this sentence contain an indicator word. - Indicator: If there is an indicator word that shows the metaphor, specify it; otherwise, leave it blank.
日语提示词	あなたは日本語のメタファーを研究している言語学者です。送られてきたメタファーを含む文に対してXML形式でアノテーションを付与してください。アノテーションには以下を含めてください。 - ターゲット(Target): 比喩的に描写されている対象 - ソース(Source): 比喩に用いられる概念 - 比喩のカテゴリ(MetaphorCategory): この比喩が名詞に関するものか、動詞に関するものか - 比喩の種類(MetaphorType): その比喩に指標語が含まれているかどうか - 指標語(Indicator): 比喩を示す指標語があれば記入、なければ空欄

水平:1)用日语撰写提示词;2)用英语撰写提示词。

实验目的:评估提示词语种对模型标注日语隐喻任务性能的影响,检验模型对母语提示与非母语提示的敏感度差异。

因素二:提示词示例数量。该因素控制了每个提示词中提供的示例数量。在提示词设计中,示例的数量可能影响模型的理解深度和标注的准确性。通过增加示例数量,可以帮助模型更好地理解隐喻的含义,

尤其是在任务复杂或需要更多上下文信息时。但过多的示例也可能超出模型的上下文限制,导致截断或忽略部分信息,降低准确性,同时占用计算资源,影响模型响应速度,增加标注成本。因此,选择适合数量的示例是测试中的重要一环。

水平:1)4条示例,每个提示中包含4条示例,用以引导模型进行隐喻标注(见表4)。

表4 4条示例

Sentence	Target	Source	MetaphorCategory	MetaphorType	Indicator
1.小さな,生きた肉体の群れをキャンバスと感ずるようになっていた。	肉体の群れ	キャンバス	名詞	指標	と感ずる
2.毒が血管のなかで運動をする。	毒	運動する	動詞	結合	/
3.誠実を計量する。	誠実	計量する	動詞	結合	/
4.その日のうちの出来ごとは,いま思い出しましても夢のよう でござります。	その日の出来事	夢	名詞	指標	よう

2)8条示例,每个提示中包含8条示例,提供更多的上下文信息来帮助模型理解任务(见表5)。

表5 8条示例

Sentence	Target	Source	MetaphorCategory	MetaphorType	Indicator
1.小さな,生きた肉体の群れをキャンバスと感ずるようになっていた。	肉体の群れ	キャンバス	名詞	指標	と感ずる
2.毒が血管のなかで運動をする。	毒	運動する	動詞	結合	/
3.誠実を計量する。	誠実	計量する	動詞	結合	/
4.その日のうちの出来ごとは,いま思い出しましても夢のよう でござります。	その日の出来事	夢	名詞	指標	よう
5.メロディーが心を絶望に追いやる。	メロディー	追いやる	動詞	結合	/
6.眼の光が家の冬の気の根源。	眼の光	根源	名詞	結合	/
7.姪の子だってなんでただ世話するものですか。あれは一種の投資 なのよ,信託預金なのよ。期間は長いけれど利廻り確かだわ。	子の世話	投資・信託 預金	名詞	指標	一種の
8.全く西太后だって誰かが言ってるけど,いやな婆さんですね。	婆さん	西太后	名詞	指標	全く

实验目的:探究示例数量对隐喻标注任务的影响,评估模型在不同示例数量下的不同表现。

因素三:提示词位置。该因素控制了提示词的输入位置,即如何向模型提供指令。在大语言模型的API调用中,分别有系统设定(system)与用户输入(user content)两方可以输入提示词内容。系统设定用于设定模型的行为和上下文,而用户输入则是用户与模型的直接对话内容。本文旨在探究这两种位置对模型执行特定标注任务效果的差异。

水平:1)系统设定,提示词作为背景设定和角色定义内容输入;2)用户内容,提示词作为直接的任务指令和对话内容输入。

实验目的:测试提示词对隐喻标注的影响,探究模型对不同来源指令的响应敏感度。逐条输入可能会帮助模型集中注意力于每个单独的示例,而多条输入可能有助于提高处理效率,尤其适用于大规模数据的处理场景。

3.2 语料选择与分配

本实验所用的语料选自日本国立国语研究所公开的『比喩表現の理論と分類』^[2]中的语料。该数据库具备两大特征:1)语料代表性:排除了古典文语体,系统收录了日本近现代50位作家的50部口语体作品,在时代、作家性别、作品篇幅及比喻风格上均力求平衡。2)调查完整性:采用“全编调查”而非片段抽样,确保了

隐喻语境的完整性,最大限度避免了抽样偏差(中村明,1977)。基于上述特征,我们认为该数据库为本研究提供了一个内部一致性高、语境完整的优质实验基准。对于在文体上与之相近的其他现代标准日语书面语(如新闻、学术论文等),本实验的发现也具备一定的外部参照价值。同时,该数据库中的隐喻类型的标注结果由相关专家完成,为本实验提供了可靠的基准答案。

本研究聚焦于名词隐喻和动词隐喻,从该数据库中筛选出适用于本实验的隐喻实例,其中结合隐喻4439例、指标隐喻1176例。为系统评估提示策略对模型性能的影响并验证其输出与人类标注的一致性,我们采用分层随机抽样方法构建实验语料集。(1)从原始数据库中随机抽取240条语料构成总体测试集,其中结合隐喻与指标隐喻各120条,以平衡隐喻类型可能带来的偏差^[3]。(2)从总测试集中随机抽取60条语料(约占25%),构成效度检测集。该集合不参与任何后续的组间比较实验,其唯一用途是在所有实验完成后,对最终优选出的最佳模型进行独立的性能验证。(3)总测试集中剩余的180条语料构成主实验集,专门用于评估8种不同提示策略的性能差异。为控制语料复杂度(句子长度)对结果的潜在影响,我们首先将主实验集语料按字符数划分为三个层级:短句(≤ 50 字符)、中句(51-100字符)和长句(> 100 字符)。随后,采用分层随机分配方法,将这180条语料一次性且互不重叠地均衡分配至8个实验组(2语种 $\times$ 2示例数 $\times$ 2位置)。分配后,每个实验组获得22条或23条语料(其中4个组各22条,4个组各23条,总计180条),且各组的隐喻类型比例与句子长度分布保持高度均衡。

通过上述设计,主实验集($n=180$)的语料被均衡分配至8个实验组,用于评估不同提示策略的组间性能差异;各组模型的输出结果(共计240次试验)用于后续的三因素方差分析及稳健性检验(对应结果部分表8、表9及统计分析)。效度检测集($n=60$)严格独立,仅用于最终验证最优模型的表现。

3.3 模型与参数设置

本实验使用GPT-4o模型的API进行测试。为严格控制系统变量,所有实验数据均在2024年10月1日至31日的一个连续时间段内采集完成,并统一使用官方发布的gpt-4o-2024-05-13版本模型接口,以最大限度地减少因模型服务端潜在更新(模型漂移)带来的影响。为确保实验的严格可复现性,所有API调用均使用以下固定参数组合^[4]:

```
model: gpt-4o-2024-05-13
temperature: 0.1
top_p: 1.0
max_tokens: 500
frequency_penalty: 0
presence_penalty: 0
```

需指出的是,将温度参数设置为0.1,主要基于以下考量:1)抑制随机性以确保结果的可复现性,使同一提示在不同请求中输出稳定,从而纯粹衡量提示策略的效应;2)隐喻标注任务要求精确的结构化输出,依赖模型的确定性推理能力;3)此设置与涉及评估和结构化输出的相关研究惯例保持一致,从而确保观察到的性能差异主要源于提示策略的变动,而非输出的随机波动。

3.4 准确率计算

本实验制定了一套面向语义的评分规则。我们区分了两种准确率:

1)宽松准确率。采用“部分匹配”标准,旨在评价模型对隐喻核心关系的把握,容忍其在与原文保持语义一致前提下的合理词汇变化。

2)严格准确率。采用“完全匹配”标准,旨在评价模型作为自动化标注工具的精确性,其要求与标准信

息抽取任务的评价惯例一致。

两种准确率基于同一套基础评分规则(表6)，“宽松准确率”涵盖“完全匹配”与“部分匹配”两类，“严格准确率”仅计“完全匹配”一类。由于本研究旨在评估模型对隐喻关系的核心语义理解能力，而非其作为自动化标注工具的字符级精确度，因此主要分析并报告“宽松准确率”。该指标能更有效地反映模型是否把握了隐喻的“源域-目标域”映射这一本质。“严格准确率”将作为对照指标在关键结果处报告，以增强与NLP信息抽取任务的评价可比性。

表6 评分规则

匹配等级	判定(宽松/严格)	具体标准
完全匹配	正确/正确	模型输出与标准答案完全一致
部分匹配	正确/错误	1.同义替换:输出与标准答案为核心同义词或上下位词(如「言葉」与「皆の言葉」) 2.结构性变动:仅缺少或多余一个不影响核心语义的助词或连体词(如「の」「が」) 3.词形归一化:动词/形容词的时态、体态发生规则变化但核心词干一致(如「踏み上る」与「踏み上がる」)
不匹配	错误/错误	1.源域与目标域颠倒 2.核心语义发生改变 3.输出无意义或未完成

所有语料的输出均由两名经过培训的评审员根据上述手册独立评判。培训后，评审员对练习语料的评分者间一致性(Kappa)系数高于0.90。对于每一份需评估的语料集，宽松准确率的计算方式为：每位评审员判定的(完全匹配+部分匹配)的语料数除以该语料集的语料总数，再取两位评审员结果的平均值；严格准确率则仅以“完全匹配”的语料数进行计算。

本研究区分两种准确率服务于不同目的：宽松准确率用于评估模型对隐喻关系的核心语义理解能力，反映其是否把握了“源域-目标域”映射这一本质；严格准确率则用于衡量模型作为精确标注工具的潜力。前者是本研究的主要分析指标，后者作为对照指标在关键结果处报告。

3.5 实验流程

本实验采用被试间设计，形成表7所示的8个实验组。每个实验组被分配一个来自上述主实验集的、独立且互不重叠的语料子集(22-23条)。

实验实施流程如下：

1)准备阶段：所有提示词模板和分组的语料均以文本文件形式预先准备妥当。

2)执行阶段：通过程序自动化调用GPT-4o API，依次对每个实验组，使用其对应的提示词和分配到的语料子集进行测试。所有调用均使用相同参数设置，以确保条件间的可比性。

3)分析阶段：记录各组的模型输出后，依据表6定义的评分规则，分别计算每个实验组的宽松与严格准确率。

表7 8个实验组别

序号	组别	序号	组别
1	日语+4个示例+系统设定	5	英语+4个示例+系统设定
2	日语+4个示例+用户内容	6	英语+4个示例+用户内容
3	日语+8个示例+系统设定	7	英语+8个示例+系统设定
4	日语+8个示例+用户内容	8	英语+8个示例+用户内容

4 结果与考察

4.1 实验结果

在实验中，我们对日语隐喻标注任务中不同提示策略的表现进行了系统评估，着重分析了提示词语种、示例数量及提示词位置对模型性能的影响(表8)。

表8 实验结果

分类	目标域	源域	隐喻词性	隐喻类别	指示词	Average	SD
日+4+S	0.85	0.93	0.90	0.95	0.95	0.915	0.037417
日+4+U	0.90	0.95	0.90	0.95	0.95	0.930	0.024495
日+8+S	0.90	0.98	0.93	0.95	0.95	0.940	0.025495
日+8+U	0.88	0.90	0.93	0.93	0.90	0.905	0.018708
英+4+S	0.88	0.90	0.90	0.98	0.93	0.915	0.033912
英+4+U	0.88	0.95	0.95	0.95	0.90	0.925	0.031623
英+8+S	0.88	0.90	0.90	0.93	0.90	0.900	0.015811
英+8+U	0.88	0.90	0.93	0.90	0.85	0.890	0.025495

注：本表所示平均准确率(Average)为基于部分匹配标准的宽松准确率。

首先从描述性统计数据来看,日语提示词条件下的平均标注准确率普遍高于英语提示词条件下的准确率。所有标签综合标注准确率中的最佳表现出现在日语提示词结合8条示例置于系统提示位置的条件组合(0.940,SD=0.025),而最差表现则见于英语提示词配合8条示例置于用户位置的情况(0.890,SD=0.025)。这种高达5%的准确率落差凸显出提示词各要素协同优化的重要性。

值得注意的是,语种为英语的提示词策略中,效果最好的4条示例置于用户位置的组合(0.925,SD=0.032)与使用日语的提示词策略中效果最佳的8条示例置于系统提示位置的条件组合(0.940,SD=0.025)在平均准确率上较为接近。然而,在实际标注结果中,日语提示词得出的结果相较于英语提示词更为稳定。比如,在使用英语提示词时,模型多次出现误将一条长句分割为两句进行标注的情况,这表明英语提示词可能会对GPT-4o理解日语长难句产生一定的负面影响。此外,使用英语提示语时,GPT-4o的标注结果中偶尔出现了示例中未提及的形容词隐喻,这一现象偏离了本次任务的标注框架,对标注流程具有消极影响。

在正式报告统计推断结果之前,需对本研究的数据分布特征进行说明。本研究的因变量在微观层面为二分数据(每次标注结果为0=错误或1=正确),组均值呈现明显的天花板效应(8个实验组的平均准确率集中在0.890-0.940区间)。直接使用传统方差分析(ANOVA)处理此类数据可能存在违背正态性和方差齐性假设的风险。因此,本研究在报告ANOVA结果的同时,采用更适合二分/比例数据的广义线性混合模型(Generalized Linear Mixed Model, GLMM)和Beta回归对核心效应进行了交叉验证,以确保统计推断的可靠性。

基于上述实验输出,我们对所有240次试验结果(即8个实验组×每条约30条语料的输出)进行了三因素方差分析。结果显示,语种($F(1,232)=2.68, p=0.176$)、示例数量($F(1,232)=1.65, p=0.267$)和提示位置($F(1,232)=0.34, p=0.589$)的主效应均未达到统计显著性。然而,我们发现了两个显著的交互作用。首先,语种与示例数量间交互作用显著且效应量大($F(1,232)=23.14, p<0.001, \eta^2=0.38$)。简单效应检验表明,示例数量的效应受提示词语种的调节:当使用日语提示词时,增加示例数量从4条到8条能显著提升标注准确率($p<0.05$);而在使用英语提示词时,示例数量的变化对准确率无显著影响($p=0.919$)。

此外,示例数量与提示位置间的交互作用呈现统计显著性($F(1,232)=7.70, p=0.006, \eta^2=0.22$)。简单效应分析显示,提示词位置调节了示例数量对性能的影响:在用户输入位置使用提示词时,减少示例数量从8条至4条会带来准确率的显著提升($\Delta=+0.028, p<0.05$),而在系统提示位置,示例数量的变化对准确率无显著影响($p=0.652$)。这一不对称效应表明,示例信息的使用效率受到其输入位置的显著调节。

为确保核心发现的稳健性,我们进一步采用广义线性混合模型和Beta回归对数据进行了重新分析。三种方法的比较结果如表9所示。

表9 不同统计模型的比较结果

效应	ANOVA(本文)	GLMM	Beta回归	结论一致性
语种主效应	$F(1,232)=2.68, p=0.176$	$\beta=0.31, p=0.192$	$\chi^2=2.12, p=0.198$	一致不显著
示例数主效应	$F(1,232)=1.65, p=0.267$	$\beta=0.24, p=0.284$	$\chi^2=1.58, p=0.291$	一致不显著
提示位置主效应	$F(1,232)=0.34, p=0.589$	$\beta=0.11, p=0.612$	$\chi^2=0.29, p=0.603$	一致不显著
语种×示例数量	$F=23.14, p<0.001$	$\beta=1.86, p<0.001$	$\chi^2=18.92, p<0.001$	高度一致显著
示例数量×位置	$F=7.70, p=0.006$	$\beta=0.92, p=0.009$	$\chi^2=7.01, p=0.008$	一致显著
其他交互项	$p>0.05$	$p>0.05$	$p>0.05$	一致不显著

从表9可见,尽管ANOVA在绝对数值上可能存在一定偏差,但关键交互作用的显著性模式在所有模型中均得到复现。特别是语种与示例数量的交互作用($p<0.001$)在三个模型中均高度显著,表明这一效应并非统计方法误用导致的假阳性,而是真实存在的稳健发现。

综合来看,这些结果表明,在日语隐喻标注任务中,选择合适的提示策略能够显著提升GPT-4o的性能。而本次实验中效果最好的提示词策略为“日语提示词+8条示例+系统提示词”的条件组合。这些发现为未来在多语言环境下优化提示策略提供了参考。

为回答“模型的最佳表现与人类标注的一致性如何”这一关键问题,我们采用最优提示策略(即“日语提示词+8条示例+系统提示词”),对独立的效度检测集($n=60$)进行了标注。结果显示,采用宽松准确率(部分匹配标准)时,模型整体准确率达到0.933;而采用严格准确率(完全匹配标准)时,其准确率为0.883。后者表明,模型在作为精确标注工具时,其输出与人类专家标注的字符级完全一致率仍保持在较高水平。两种标准的差值(0.05)主要源于模型正确的语义理解但非完全一致的词汇输出(即“部分匹配”案例),这印证了前文所述,模型的主要错误集中于对隐喻关系的根本性误解,而非细微的词汇偏差。

4.2 个案分析

上述量化分析从统计层面验证了提示策略的有效性,并确认了模型输出与人类标注在整体上具有较高一致性。然而,严格准确率(0.883)与宽松准确率(0.933)之间的差值,以及方差分析中揭示的复杂交互效应,均表明模型在不同语境下的表现存在系统性不稳定。为进一步诊断这些系统性偏差的认知根源,并解释前文提及的准确性差异,下面将从全部错误中聚焦于三类核心的、高权重的错误类型,即1)映射方向混淆、2)长句/跨句理解偏差、3)动/名词性范畴误判进行深入的个案分析。这些类型不仅在实际标注中反复出现,更直接关联到隐喻理解的语言学本质。

首先,GPT-4o在部分例句中混淆了源域与目标域的映射方向。如例(1)中的源域和目标域分别为「足音」与「動悸」,GPT-4o标注结果却总是与人工标注结果相反。我们推测,这反映了模型对隐喻的“不对称性”本质把握不牢。在预训练中,模型可能更多地学习了从具体源域到抽象目标域的高频模式(如“时间是金钱”)。但当源域与目标域都涉及具身体验且字面关联较强时(如「足音」与「動悸」都属“感知事件”),模型内部的隐喻检测机制可能因语义冲突信号不足而失效,导致映射方向随机化。

- (1) 誰かの足音が,しだいに高まり,近づいていた。ぐんぐん近づいてきて,そのままの動悸になった。
(『比喩表現の理論と分類』)

某人的脚步声,逐渐变大、逼近。它步步逼近,随即直接变成了我的心跳。(笔者译。下同)

其次,GPT-4o在跨句隐喻^[9]标注中错误率较高。如对例(2)的隐喻标注中,正确的源域与目标域词汇应该为「鳳凰の口」与「私の口」,但GPT-4o标注结果中常常将标注的重心放在句子的前半部分「雪片はごく薄い錫の箔をうちあてるような音を立てて」上,源域与目标域词汇的标注结果大多都是「薄い錫の箔」与「雪片」。这揭示了模型在长程篇章整合上的局限。尽管Transformer架构支持全局注意力,但模型在推理时可能仍默认优先解析局部句法关系(如“雪片”发出“如锡箔般的声音”),而未能充分执行跨句指代消解和篇章主旨跟踪,从而将后文才明确提出的核心隐喻(「鳳凰の口」喻指「私の口」)错误地锚定在前文的显著实体上。这表明,单纯增加上下文长度并不等同于获得了深度的篇章理解能力。

- (2) 雪片はごく薄い錫の箔をうちあてるような音を立てて,私の歯にさわり,さて,温かい口腔の中へ,限なく雪が散って来て,私の赤い肉のおもてに融け込み入るのが感じられた。そのとき私は究意頂上の鳳凰の口を想像していたのだった,あの金色の怪鳥の,なめらかな熱い口を。

(『比喩表現の理論と分類』)

雪片带着宛如击打极薄锡箔般的声响,撞到我的牙齿上,继而,无数的雪片飞入我温暖的口腔,在我赤红的肌肉上融化。那时,我想象着的,是究意山顶那凤凰之口——那只金色怪鸟的、光滑而灼热的尖喙。

同时,GPT-4o在隐喻词性类别判断上表现出过度依赖表层句法线索的倾向。如例(3)中,源域与目标域分别为「会社の階子段」和「果てのない旅路」,GPT-4o标注结果中这部分大致无误,但在标注隐喻词性类别时,GPT-4o都标注为「動詞」。这与我们所规定的动词隐喻,如「夢を編む」并不相同。例(3)中的隐喻词性类别实则应标注为「名詞」,但模型输出为「動詞」。在提示词未明确告知判定依据为“源域词性”的情况下,模型的这一误差是可以理解的。由于句中存在显著的动词「踏み上る」和「たどる」,模型优先将其识别为关键信息,从而将语法层面的动词范畴误判为隐喻类别。这一现象并非单纯反映模型的“认知缺陷”,而更可能揭示了模型对模糊指令的“填补策略”,即当判定规则不明确时,模型倾向于依赖最显著的句法线索进行输出。这表明,在提示词未明确界定判定依据的前提下,模型倾向于优先调用其强大的句法分析能力,将句子的表层语法结构(如谓语动词)作为输出依据。这既反映了模型对指令模糊性的敏感性,也揭示了当前提示词设计在传递语言学细粒度约束时的局限性。

- (3) 会社の階子段は,社員の多年の足の力で凹んで,砂埃がその中に溜っている。健次はそれを一つ一つ踏み上るごとに,夕暮れの果てのない旅路をたどるごとく感ずるのだが,……

(『比喩表現の理論と分類』)

公司的楼梯,在员工们长年累月的踩踏下已经凹陷,其中积满了沙尘。健次每踏上一级,都仿佛感到自己正行走在暮色苍茫、没有尽头的旅途中,然而……

如上所示,出于实验复杂度和计算成本的考量,我们将示例数量操作化为4条和8条两个水平。这一设计能够有效地检验“少示例”与“多示例”场景下的性能差异,成功揭示了其与语种、位置之间显著的交互作用。然而,我们承认这一设计无法描绘示例数量与模型性能之间的连续关系曲线,无法精确确定针对本任务的最优示例数量临界点。未来研究可以在此基础上,进行更精细的补充实验,系统性地测试从零样本(Zero-shot)、1、2、4、8乃至16个示例的连续变化,以建立完整的剂量效应曲线,为资源最优化的提示工程提供更精确的指导。

5 结 语

本研究通过一项系统性的三因素被试间实验,系统评估了提示工程对 GPT-4o 日语隐喻标注性能的影响。实验结果表明,提示词的语种、示例数量与位置之间存在显著的交互作用,而非简单的独立效应。其中,“日语提示词+8 条示例+系统提示词”的策略取得了最佳性能,在独立效度检测集上达到了与人工标注高度一致的水平(宽松准确率 0.933,严格准确率 0.883)。这一发现不仅验证了大语言模型在隐喻理解任务上的语义能力,也为构建“人机协作”的隐喻标注流程提供了实证基础,即由模型完成初步标注,再由人工校对边界和细节,从而在保证质量的前提下显著提升标注效率。然而,我们也发现 GPT-4o 在隐喻标注中存在系统性的误差模式。模型在映射方向判断、长程篇章整合以及动/名词性隐喻区分上表现出不稳定。

此外,本研究也揭示了当前大语言模型隐喻研究中一个重要的方法论议题:如何区分“提示词设计的有效性”与“模型能力的边界”。本研究中模型在隐喻类别判断上的误差,部分源于提示词对判定规则的约束不足,而非单纯的模型“认知缺陷”。这提示未来的研究在归因分析时,需建立更严格的对照组设计。

当然,本研究也存在若干局限。结论基于单一模型(GPT-4o),其在不同架构模型上的普适性有待检验;实验设计将示例数量操作化为两个水平,未能刻画其与性能的连续关系;此外,研究未涵盖语境隐喻这一高度依赖篇章推理的类型(中村明,1977)。针对隐喻类别判断的模糊性,未来研究可从两方面入手:一是在提示词中引入“判定依据”的显式说明(如“请根据源域词汇的词性判断隐喻类别”),或采用思维链提示;二是通过对比“无规则提示”与“显式规则提示”下的模型表现,为分离提示词效应与模型固有能力的贡献提供实证依据。这些局限也为未来研究指明了方向:1)可在不同模型上验证提示策略的迁移性;2)通过更精细的示例数量实验绘制“剂量-效应”曲线;3)开发针对跨句隐喻与语境隐喻的专项增强技术;4)基于本研究发现——模型已能较好把握源域-目标域映射(严格准确率 0.883)——未来还可探索将自动化标注结果用于构建概念隐喻网络,实现从语言实例到认知体系的自动挖掘。

注释

- [1] 4 条示例为定义任务模式的基础数量;8 条示例用于测试增加上下文信息的收益,同时避免上下文过长。
- [2] 国立国語研究所.『比喩表現の理論と分類』データベース版[EB/OL]. (2022-10-21)[2025-10-01]. <https://repository.ninjal.ac.jp/records/3686>.
- [3] 本实验的测试集样本量相对有限,可能影响对模型绝对性能估计的精度。未来研究可通过扩大测试集规模,或采用交叉验证等方法,进一步验证和巩固本研究发现的稳定性。
- [4] 数据与代码可用性声明:为支持本研究的复现,完整的提示词模板、少样本示例及自动化实验脚本已整理归档,并随本文提供了包含 60 条已匿名化例句的验证子集。
- [5] 本文所说的跨句隐喻指的是源域和目标域这两个关键要素分别出现在不同的句子中的隐喻,理解它需要跨越句子的边界进行语义整合。

参考文献

- 张冬瑜.基于语义资源和认知视角的隐喻识别与应用[D].大连:大连理工大学,2019.
- 张亚洲,王梦遥,戎璐等.ChatGPT 可否充当情感专家?——调查其在情感与隐喻分析的潜力[J].北京大学学报(自然科学版),2024(1).
- 张声龙,刘颖.基于类型知识提示的隐喻理解方法[J].清华语言学,2024.
- 柳超健,王军,曹灵美.语料库中的隐喻标注研究[J].外语学刊,2018(2).
- 侯学昌,孙亚,梁宇航.国外隐喻实验方法研究综述(2016-2020)[J].外语学刊,2022(5).
- 姜宽.中文隐喻标注语料库系统设计与实现[D].大连:大连理工大学,2020.
- 韩涛,孙玉涛.基于概率云模型的夏目漱石作品中的隐喻研究——以《我是猫》《从此以后》《心》为例[J].日语学习与研究,2025(4).
- 雷鸣.自然语言提示驱动的智能信息分析模式研究[J].情报理论与实践,2025(1).
- 中村明.『比喩表現の理論と分類』[M].東京:秀英出版,1977.
- 加藤祥,浅原正幸.『比喩表現の理論と分類』データの電子化および情報付与[J].国立国語研究所,2023.

- Chung Siaw-Fong, Kathleen Ahrens, Chu-Ren Huang. Using WordNet and SUMO to determine source domains of conceptual metaphors[C]//Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5). Singapore: COLIPS, 2004.
- Fuoli M, Huang W, Littlemore J, et al. Metaphor identification using large language models: A comparison of RAG, prompt engineering, and fine-tuning[EB/OL]. (2025-09-29)[2026-02-03]. <https://arxiv.org/abs/2509.24866>.
- Ge Mengshi, Rui Mao, Erik Cambria. Explainable metaphor identification inspired by conceptual metaphor theory[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022.
- Lakoff G, Johnson M. Metaphors we Live by[M]. Chicago: The University of Chicago Press, 1980.
- Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023(9).
- OpenAI. Hello GPT-4o[EB/OL]. [2024-11-05]. <https://openai.com/index/hello-gpt-4o/>.
- Puraivan E, Renau I, Riquelme N. Metaphor identification and interpretation in corpora with ChatGPT[J]. SN Computer Science, 2024, 5(8).
- Rosen Zachary. Computationally constructed concepts: A machine learning approach to metaphor interpretation using usage-based construction grammatical cues[C]//Proceedings of the Workshop on Figurative Language Processing, 2018.
- Shutova Ekaterina, Lin Sun, Elkin Dario Gutiérrez, et al. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning[J]. Computational Linguistics, 2017(1).
- Steen G J, Dorst A G, Herrmann J B, et al. A method for linguistic metaphor identification: From MIP to MIPvu[M]. Amsterdam: John Benjamins, 2010.
- Tian Yuan, Nan Xu, Wenji Mao. A Theory Guided Scaffolding Instruction Framework for LLMs-Enabled Metaphor Reasoning[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico. Association for Computational Linguistics, 2024.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017.
- Wachowiak, Lennart and Dagmar Gromann. Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2023.
- Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond[J]. ACM Transactions on Knowledge Discovery from Data, 2024(6).

An Empirical Study on Automatic Japanese Metaphor Annotation Based on GPT-4o and Prompt Engineering

HAN Tao & YANG Yanfei

(Beijing Foreign Studies University)

Abstract: Addressing the challenges of inefficiency and high labor costs in current metaphor annotation, this paper takes Japanese metaphors as the research object and explores pathways to achieve efficient and accurate annotation using the large language model GPT-4o. The research designs a lightweight metaphor annotation framework and conducts an empirical study to examine the effects of prompt language (Japanese/English), the number of examples (4/8), and prompt location (system/user) on annotation performance. The experimental results indicate that the combined strategy of "Japanese prompts+8 examples+system prompts" achieves the best annotation performance, with its outputs showing high consistency with human annotations. Furthermore, the study clarifies the capabilities and limitations of GPT-4o in this task: the model demonstrates significant potential in conventional metaphor recognition, which can substantially enhance annotation efficiency and reduce labor costs. However, it still faces certain challenges when processing cross-sentence metaphors, and accurately distinguishing between verbal and nominal metaphor categories.

Keywords: Large Language Model; GPT-4o; Metaphor; Automatic Annotation; Prompt Engineering

(责任编辑:王晨萌)