

生成式 AI 英译《庄子》核心概念的策略倾向研究

——基于 BLEU、BERTScore 与 COMET 的量化分析

陈 志*

广州城市理工学院 广东广州

【摘要】生成式人工智能的发展为翻译研究提供了新的路径，也引发了不同模型翻译策略差异的讨论。本文以《庄子》内七篇 30 个核心概念（75 个翻译单元）为研究对象，以理雅各和华兹生英译本分别作为异化与归化策略参照，采用 BLEU、BERTScore 和 COMET 三项评价指标，对 ChatGPT（GPT-4o）与 Qwen（通义千问）的英译结果进行量化比较。结果表明，两种模型整体上均与华兹生译本具有更高相似度，表现出较为明显的归化倾向；但在“大宗师”“无待”“卮言”等文化负载程度较高的概念上，Qwen 与 ChatGPT 呈现出不同的策略选择，反映出模型在哲学概念翻译中的策略分化。三项评价指标所得结论基本一致，说明多指标交叉验证能够较为稳定地反映 AI 译文与参照译本之间的趋同性。研究认为，生成式 AI 的翻译策略并非由模型来源单一决定，而是受到文本体裁、概念属性及训练数据等多重因素的共同影响。“技术化译者主体性”为理解这种差异提供了新的解释视角。本研究可为生成式 AI 参与中国哲学典籍英译的策略评价及相关研究提供参考。

【关键词】生成式人工智能；《庄子》；归化与异化；BLEU；BERTScore；COMET

【收稿日期】2026 年 7 月 10 日

【出刊日期】2026 年 8 月 11 日

【DOI】10.12208/j.ssr.20260280

Strategic tendencies in generative AI translation of *Zhuangzi*'s core concepts: A quantitative analysis using BLEU, BERTScore and COMET

Zhi Chen*

Guangzhou City University of Technology, Guangzhou, Guangdong

【Abstract】This study examines the translation strategy tendencies of generative artificial intelligence (AI) in translating the core concepts of the *Zhuangzi* into English. Thirty representative concepts from the seven Inner Chapters, comprising 75 translation units, were selected as the corpus. The English translations by James Legge and Burton Watson were used as representative references for foreignization and domestication, respectively. ChatGPT (GPT-4o) and Qwen were compared using BLEU, BERTScore and COMET. The results show that both models demonstrate a stronger overall similarity to Watson's translation, indicating a tendency toward domestication. Nevertheless, culturally loaded concepts such as Dashizong, Wudai and Zhiyan reveal noticeable differences between the two models, suggesting that AI translation strategies vary across philosophical concepts. The three evaluation metrics lead to largely consistent conclusions, indicating that a multi-metric approach provides a reliable basis for identifying translation strategy tendencies. The findings further suggest that AI translation strategies are jointly shaped by text genre, conceptual characteristics and training data rather than by the origin of the model alone. The notion of technologized translator subjectivity offers a useful perspective for interpreting these differences. This study provides empirical evidence for evaluating AI-assisted translation strategies in Chinese philosophical classics.

【Keywords】Generative AI; *Zhuangzi*; Domestication and foreignization; BLEU; BERTScore; COMET

1 问题提出

生成式人工智能的发展让机器翻译技术进入了新

阶段^[1]。ChatGPT、Qwen 等大语言模型凭借较强的语言理解和文本生成能力，在多语言翻译任务中表现出

*通讯作者：陈志

良好的应用潜力,也为中国典籍的对外传播提供了新的技术支持^[2]。中国古代哲学典籍既承载着丰富的文化内涵,也包含大量抽象概念,其翻译涉及概念阐释、文化转换和思想表达等诸多层面,因此对生成式 AI 翻译能力提出了更高的要求。

现有研究主要关注生成式 AI 的译文质量、译后编辑和文化负载词处理等问题,并普遍借助自动评价指标或人工评价考察模型表现。许多研究发现,不同大语言模型在文化信息处理上表现并非一致,其翻译策略会因文本类型、任务目标和训练语料而发生变化^[3]。然而,对于中国哲学典籍特别是《庄子》核心概念的翻译,目前仍缺乏基于量化指标的分析。不同模型在这一类文本中的策略选择是否具有-致性,高文化负载概念是否会引发不同的翻译倾向,仍有进一步探讨的必要。

《庄子》是中国哲学的重要经典,其“道”“逍遥”“心斋”“坐忘”“无待”等核心概念既是哲学思想的基本范畴,也是英译研究的关注焦点。理雅各和华兹生的两个经典英译本分别代表了较为典型的异化与归化取向,为比较生成式 AI 译文的策略特征提供了稳定的参照。

基于此,本文选取《庄子》内七篇 30 个核心概念共 75 个翻译单元作为研究语料,以 ChatGPT(GPT-4o)和 Qwen(通义千问)为比较对象,以理雅各和华兹生译本为参照,综合运用 BLEU(Bilingual Evaluation Understudy)^[4]、BERTScore(BERT-based Evaluation Metric)^[5]和 COMET(Crosslingual Optimized Metric for Evaluation of Translation)^[6]三项评价指标,对两种模型的翻译策略倾向进行量化分析,并结合“技术化译者主体性”视角,讨论其差异形成的原因,以期为生成式 AI 对中国哲学典籍英译的策略研究提供新的实证参考。

2 研究设计

2.1 研究对象与语料

本文选取《庄子》内七篇(《逍遥游》《齐物论》《养生主》《人间世》《德充符》《大宗师》《应帝王》)中的 30 个核心概念作为研究对象,共构建 75 个翻译单元。所选概念涵盖“道”“逍遥”“齐物”“心斋”“坐忘”“真人”“无待”“卮言”“大宗师”等核心范畴,包含具有稳定英译传统的哲学术语,也涵盖文化负载程度较高、译法存在分歧的概念,可较为全面地反映《庄子》核心概念的翻译特点。

为保证比较的代表性,本文选取理雅各(James Legge, 1891)和华兹生(Burton Watson, 1968)的《庄子》英译本作为参照文本。两种译本被广泛引用,在概念处理方式上分别体现了典型的异化与归化取向,为分

析生成式 AI 译文的策略倾向提供了稳定的比较基准。

研究对象包括 ChatGPT(GPT-4o)和 Qwen(通义千问)两个生成式 AI 模型。两种模型均采用统一提示语:“请将以下古汉语句子翻译成英文,翻译时保持学术性和准确性。”所有译文均在相同条件下获取,以保证结果的可比性。

2.2 研究方法

本文采用 BLEU、BERTScore 和 COMET 三项评价指标,对 AI 译文与两部参照译本之间的相似度进行量化比较。其中,BLEU 主要反映词汇层面的匹配情况^[2],BERTScore 侧重语义层面的相似性^[7],COMET 综合原文、参考译文和候选译文,对译文整体质量进行评价^[8]。三项指标分别对应词汇、语义和综合质量三个维度,能够形成相互补充的评价体系,以降低单一指标的偏差。

本文关注的是 AI 译文与参照译本之间的文本趋同性,而非直接测量归化或异化策略。当 AI 译文与华兹生译本具有更高相似度时,可认为其整体趋近归化取向;反之,则趋近异化取向。为提高研究结论的可靠性,本文结合三项指标进行交叉验证,并在典型概念层面进一步分析不同模型的策略差异。

2.3 研究问题

围绕上述研究设计,本文主要回答以下三个问题:

(1) ChatGPT 与 Qwen 在《庄子》核心概念英译中整体呈现何种翻译策略倾向;(2)两种模型在不同概念上的策略表现是否存在差异,其分化主要体现在哪些概念;(3)生成式 AI 翻译策略的差异能否从“技术化译者主体性”视角得到解释。

3 结果与分析

3.1 整体策略倾向比较

为考察两种生成式 AI 在《庄子》核心概念英译中的整体策略倾向,本文分别计算 ChatGPT 和 Qwen 译文与理雅各、华兹生两部参照译本之间的 BLEU、BERTScore 和 COMET 得分,并从词汇匹配、语义相似性和综合译文质量三个维度进行比较,三项指标结果见表 1。

由表 1 可见,两种模型与华兹生译本的三项评价指标均高于理雅各译本,说明在所选语料范围内,两种模型的译文整体更接近归化型译本。其中,ChatGPT 在 BLEU、BERTScore 和 COMET 三个维度均表现出较为一致的趋势,表明其在词汇选择、语义表达和整体译文质量方面均更趋近于华兹生的翻译方式。这说明,ChatGPT 在处理《庄子》核心概念时,更倾向于采用符合英语表达习惯的译法,以提升译文的可读性和目标语接受度。

表 1 两种 AI 模型与参照译本的总体相似度比较

指标	ChatGPT →Legge	ChatGPT →Watson	Qwen →Legge	Qwen →Wats
BLEU	0.1034	0.1411	0.1738	0.2175
BERTScore	0.5119	0.5490	0.5628	0.5950
COMET	0.6582	0.6908	0.7015	0.7303

Qwen 同样表现出整体归化倾向，其三项指标均高于理雅各译本。但与 ChatGPT 相比，Qwen 在两部参照译本之间的分值差距相对较小，说明其译文并未完全向某一种翻译模式集中，而是在不同概念和语境下表现出一定的策略弹性。这说明两种模型虽然整体呈现相近的翻译取向，但策略稳定性并不完全一致。

三项评价指标所得结论保持一致，说明无论从词汇层面、语义层面还是综合质量层面考察，两种模型均呈现出相同的总体策略趋势。这种一致性增强了整体判断的可靠性，也表明采用多指标交叉验证可较为全面地反映 AI 译文与参照译本之间的趋同性。但整体统

计结果反映的是模型在全部语料上的平均表现，尚不能揭示不同哲学概念之间可能存在的策略差异，还需结合代表性概念作进一步分析。

3.2 概念层面的策略差异

整体统计结果能够反映模型的总体翻译取向，但不同哲学概念在文化负载、语义抽象程度和既有译介传统等方面存在明显差异，仅依据总体均值难以揭示模型在具体概念上的策略变化。还需进一步比较各核心概念与两部参照译本之间的相对相似度，并依据其表现特征归纳为稳定归化型、过渡型和策略分化型三类，结果见表 2。

表 2 不同类型核心概念的策略表现

类型	代表概念	GPT Δ	Qwen Δ	主要特征
稳定归化型	道、逍遥、心斋	+0.15~+0.18	+0.11~+0.14	两种模型均明显接近华兹生译本
过渡型	德、无为、坐忘	接近 0	接近 0	两部参照译本差异较小，模型未表现出明显策略偏向
策略分化型	大宗师、无待、卮言	+0.01~+0.05	-0.08~-0.12	ChatGPT 仍趋近华兹生，Qwen 则更接近理雅各

注： Δ 表示 AI 译文与华兹生译本相似度减去与理雅各译本相似度所得差值。 $\Delta > 0$ 表示整体更接近华兹生译本， $\Delta < 0$ 表示整体更接近理雅各译本。

由表 2 可见，多数概念属于稳定归化型。“道”“逍遥”“心斋”等核心术语虽然哲学内涵丰富，但经过长期译介，其英语表达已形成较为稳定的使用传统。两种模型在这些概念上的译文均与华兹生译本保持较高一致性，说明当源语概念已经形成较成熟的目标语表达时，模型更倾向于沿用已有译法，而非保留更多源语形式特征。这一结果与整体统计所得结论相一致，表明成熟译介传统对生成式 AI 的译文生成具有较强影响。

过渡型概念主要包括“德”“无为”“坐忘”等。这类概念在两部参照译本中的处理方式虽有差异，但整体距离相对接近，两种模型与两部译本之间的相似度差值均不明显。模型在此类概念上的译法更多受到上下文语境的影响，而非表现出稳定的策略选择。这说明，并非所有哲学概念都具有鲜明的归化或异化特征，一部分概念处于两种策略之间的连续谱上，其翻译结果具有较大的语境依赖性。

策略分化型概念则更能体现不同模型在文化信息处理上的差异。“大宗师”“无待”“卮言”等概念具有较强的哲学专属性，其意义难以通过英语中的现成概念完全对应。ChatGPT 在这些概念上的译文仍总体保持与华兹生译本较高的一致性，而 Qwen 则在部分概念上更接近理雅各译本，表现出保留源语概念特征的倾向。这说明，当译文涉及文化内涵较强、对应关系不够稳定的哲学概念时，两种模型的翻译选择开始出现分化，整体统计中的归化趋势并不能完全覆盖概念层面的实际表现。

生成式 AI 的翻译策略具有一定的层次性。对于译法较为成熟、跨文化接受程度较高的概念，两种模型表现出较高一致性；而对于文化负载较重、长期存在译法分歧的概念，模型之间的差异则更加明显。翻译策略之考察，须统观统计结果与典型概念，二者互验方能得其全局。

3.3 多指标比较与结果验证

为进一步检验整体策略判断的可靠性，本文对 BLEU、BERTScore 和 COMET 三项评价指标的分析结果进行了比较。三项指标对 ChatGPT 和 Qwen 均得出了相同的总体判断，即两种模型在《庄子》核心概念英译中整体更接近华兹生译本，未出现指标之间相互矛盾的情况，结果见表 3。

表 3 三项评价指标的策略判定结果

比较对象	BLEU	BERTScore	COMET	综合判定
ChatGPT	归化	归化	归化	归化
Qwen	归化	归化	归化	归化

表 3 表明，三项指标关注的评价维度虽有所不同，但所得结论具有稳定的一致性。BLEU 主要反映译文与参照译本在词汇和表达形式上的接近程度；BERTScore 关注语义层面对应关系；COMET 则综合源文与译文信息，对整体译文质量进行评价。多维度结论一致，说明模型策略倾向的判断有充分依据，并非单一指标所能决定。

三项指标在数值变化幅度上仍存在一定差异。BLEU 更容易受到词汇选择和表达形式的影响，对于哲学典籍中存在多种合理译法的概念，其区分能力相对有限；BERTScore 和 COMET 则能够较好地识别语义相近但表达不同的译文，因此在概念层面的评价中表现出更好的稳定性。这也说明，在分析哲学典籍翻译时，仅依赖词汇匹配难以全面反映译文特点，结合语义和综合质量指标能够获得更加可靠的评价结果。

综合来看，三项指标构成了相互补充、相互印证的框架。指标间的一致判断增强了总体结论的可信度；各指标从词汇、语义及综合质量等维度的分别考察，则保证了结果的稳定性。

4 讨论

量化分析结果表明，两种模型在《庄子》核心概念的翻译整体趋近华兹生译本，呈现归化倾向。但在概念层面的比较显示，面对不同文化负载的术语，两者策略仍有差异，说明生成式 AI 的翻译选择具有一定的层次性，而非固定遵循某一种翻译模式。

模型对于“道”“逍遥”“心斋”等已有较稳定英译传统的概念，模型的表现较为一致，多沿用已有之译法，分歧较小；而对于“大宗师”“无待”“卮言”等文化专属性较强的概念，两种模型则呈现出不同的表达取向。尤其是 Qwen 在部分概念上的译文更接近理

雅各译本，可见当概念缺乏统一译法时，模型之间的策略差异更容易凸显。

可见，生成式翻译策略既因循译介传统，也受概念本身属性制约。术语的对应关系越明确，模型越倾向于采用目标语中较为通行的表达；而概念文化内涵越丰富、语义边界越开放，翻译选择的分化则越明显。因此，评价需兼顾整体统计与概念分析，二者互补方能见其全貌。

本文发现，BLEU、BERTScore 和 COMET 三项指标均支持同一判断，说明不同评价维度所得结果具有良好的一致性。三者分别从词汇匹配、语义相似及综合翻译质量等角度进行考察，多指标交叉验证在一定程度上弥补了单一评价方法的局限，也增强了研究结论的可靠性。

本文仍存在两点局限：一是语料仅限《庄子》内七篇 30 个概念，结论仍有待扩大验证；二是仅依赖自动指标，尚未结合人工评价考察文化传达与读者接受。未来可扩大典籍范围、纳入更多模型，并结合专家与读者评价深化比较分析。

5 结论

本文以《庄子》内七篇 30 个核心概念为研究对象，参照理雅各、华兹生译本，采用 BLEU、BERTScore 和 COMET 三项评价指标，对 ChatGPT 与 Qwen 的英译策略进行量化比较。研究结果显示，两种模型整体上均更接近华兹生译本；然而，在“大宗师”“无待”“卮言”等文化负载较高的概念上，Qwen 又呈现出更接近理雅各译本的特征，可见模型的整体取向与概念层面的翻译选择并非完全一致。

生成式翻译策略具有整体性与差异性并存之特点。整体统计结果可反映模型在典籍翻译中的总体倾向；而不同概念的文化属性和译介传统又会影响模型的具体表达，策略因而呈现一定的层次性。基于此，对生成式翻译策略的研究，既需要把握整体规律，也应重视典型概念的个案分析。

本文以经典译本为参照，引入三项指标交叉验证，构建了兼顾词汇、语义及综合质量的量化分析路径，并为“技术化译者主体性”提供实证依据。后续可拓展语料，结合人工与自动评价，持续比较不同模型的典籍翻译表现，以深化对翻译策略及其影响因素的认识。

参考文献

- [1] 徐珊珊; 姚琴. 人工智能背景下农机文本的机器翻译译文质量评测[J]. 英语广场, 2025 年第 16 期.

- [2] 王羽迪. 生成式 AI 在日译汉翻译中的质量评估研究——以大语言模型 GPT-4o 为例[J]. 日语学习与研究, 2025-04-25.
- [3] 王贇; 钟丽霞. 基于 BERTScore 语义匹配度的大语言模型中医药典籍英译效能研究[J]. 燕山大学学报(哲学社会科学版), 2026-01-21.
- [4] 戴光荣; 罗萱格. 基于汉英翻译效率评估的人机交互译前编辑策略研究[J]. 翻译研究与教学, 2024 年第 2 期.
- [5] 董晓婵; 秦肯; 陈宏俊. 情感驱动提示词对 GLLMs 翻译精度影响的量化研究——基于 chrF++ 和 BERTScore 综合评价[J]. 翻译研究与教学, 2025 年第 2 期.
- [6] 刘世界. 涉海翻译中的机器翻译应用效能: 基于 BLEU、chrF++ 和 BERTScore 指标的综合评估[J]. 中国海洋大学学报: 社会科学版, 2024 年第 2 期.
- [7] 王坤宇; 成思. ChatGPT 应用于立法文本翻译及质量评估的效能研究[J]. 浙江海洋大学学报(人文科学版), 2025-02-15.
- [8] 何泊稼; 王金铨. 大语言模型驱动的多维度汉译英翻译质量评价智能体构建与应用研究[J]. 上海翻译(中英文), 2026-05-10.

版权声明: ©2026 作者与开放获取期刊研究中心 (OAJRC) 所有。本文章按照知识共享署名许可条款发表。
<http://creativecommons.org/licenses/by/4.0/>



OPEN ACCESS