引用:任禹昕,陈子杰,林咏臻,刘平.人工智能时代中医药术语机器翻译质量评估研究——以ChatGPT-4和Google翻译为例[J].中医药导报,2025,31(12):284-293.

人工智能时代中医药术语机器翻译质量评估研究*

——以ChatGPT-4和Google翻译为例

任禹昕[1],陈子杰[2],林咏臻[1],刘 平[1]

(1.北京中医药大学人文学院,北京 102488;
2.北京中医药大学中医学院,北京 102488)

[摘要] 目的:评估大语言模型(如ChatGPT-4)与传统神经机器翻译工具(如Google翻译)在中医药术语翻译中的表现,并探讨人机协同的中医翻译策略。方法:采用半自动机器翻译评价方法,通过综合BLEU、TER和METEOR共3个自动评估指标和专家人工评分,系统评估ChatGPT-4与Google翻译的中医术语翻译质量;通过实验验证提示词工程对中医术语翻译质量的提升作用。结果:ChatGPT-4的BLEU、TER和METEOR共3项自动评估指标均显著优于Google翻译;ChatGPT-4的人工评估结果优于Google翻译,尤其在保留文化内涵和语境适配方面更为突出;提示词测试结果显示,通过优化提示词可以提升ChatGPT-4的翻译质量。结论:大语言模型是更优的赋能中医翻译的机器翻译工具,具有较强的领域鲁棒性、交互性、情境学习能力、指令跟随能力和复杂推理能力,且能够更好地处理中医隐喻性表达和文化负载词;优化提示词可以有效提升大语言模型的中医翻译质量。

[关键词] 机器翻译;神经机器翻译;大语言模型;中医药术语;翻译质量评估

[中图分类号] R2-03 [文献标识码] A [文章编号] 1672-951X(2025)12-0284-10

DOI:10.13862/j.cn43-1446/r.2025.12.045

Research on Machine Translation Quality Evaluation of Traditional Chinese Medicine Terminology in the Artificial Intelligence Era: A Case Study of ChatGPT-4 and Google Translate

REN Yuxin[1], *CHEN Zijie*[2], *LIN Yongzhen*[1], *LIU Ping*[1]

(1.School of Humanities, Beijing University of Chinese Medicine, Beijing 102488, China;
2.School of Chinese Medicine, Beijing University of Chinese Medicine, Beijing 102488, China)

[Abstract] Objectives: To evaluate the performance of large language models (LLMs) (such as ChatGPT-4) and traditional neural machine translation tools (such as Google Translate) in translating traditional Chinese medicine (TCM) terminology, and to explore a human-machine collaborative translation strategy for TCM. Methods: A semi-automatic machine translation evaluation method was adopted. The translation quality of TCM terminology by ChatGPT-4 and Google Translate was systematically assessed through a combination of three automatic evaluation metrics, BLEU, TER, and METEOR, and expert manual scoring. Additionally, experiments were conducted to verify the effect of prompt engineering on improving the translation quality of TCM terminology. Results: ChatGPT-4 significantly outperformed Google Translate in all three automatic evaluation metrics, BLEU, TER, and METEOR. The manual evaluation results also showed that ChatGPT-4 performed better than Google Translate, particularly in preserving cultural connotations and contextual adaptability. The test results of prompt words show that optimizing prompt words can improve the translation quality of ChatGPT-4. Conclusion: LLMs are superior machine translation tools for empowering TCM translation, with strong domain robustness, interactivity, situational learning ability, instruction-following ability, and complex reasoning ability. They can better handle

*基金项目:北京中医药大学教育科学研究课题(XJY24035);2025年北京中医药大学社科培育项目(2025-JYB-PY-008)

通信作者:刘平,女,副教授,研究方向为智能化中医药翻译与国际传播

metaphorical expressions and culture-loaded words in TCM. Optimizing prompts words can effectively enhance the TCM translation quality of LLMs.

[Keywords] machine translation; neural machine translation; large language models; traditional Chinese medicine terminology; translation quality assessment

近年来，ChatGPT等大语言模型技术的迅猛发展引发了社会各个层面的深层次变革。大语言模型（large language model，LLM）正在成为新兴的机器翻译工具，同时引领机器翻译研究范式的革新。在此背景下，中医药翻译与国际传播面临智能化赋能的现实需求。如何利用大语言模型赋能中医药翻译进而促进中医药的高效精准传播，是中医药翻译与国际传播同仁们亟需解决的重要时代议题。

过去几十年，中医药翻译研究取得了丰硕的成果，较多学者在翻译的语言学转向、文化转向、社会学转向和生态转向方面，均进行了大量的研究。然而，面对人工智能时代翻译的技术转向[1]，中医药翻译研究界尚未做出积极的应对。文献调查显示，中医药机器翻译研究非常稀少，更未发现有学者结合最新的大语言模型技术进行中医药的翻译研究。

中医药术语是传播中医药文化的基本单位，中医药术语翻译研究是中医药翻译研究的基本问题。因此，本研究聚焦中医药术语，基于数据驱动，采取半自动机器翻译评估方法，定量评估大语言模型和传统神经机器翻译模型的中医药术语翻译质量，并基于评估结果进行对比分析，以期为人工智能时代人机协同的中医翻译实践提供实证参考。

1 ChatGPT与机器翻译

机器翻译（machine translation，MT）是运用机器（通常指计算机）将一种自然语言翻译成另一种自然语言的过程[2]。作为AI领域的重要研究方向，机器翻译已在学术界和工业界引起了广泛关注[3]。同时，快速全球化、跨语言沟通的需求促使机器翻译成为个人、企业和政府的重要工具[4]。经过70多年的发展，机器翻译取得了巨大成就。伴随着2014年神经机器翻译模型（neural machine translation，NMT）的提出，机器翻译质量显著提升[5]。比较受欢迎的神经机器翻译工具包括Google翻译、DeepL翻译、百度翻译、有道翻译等。

近三年，大语言模型的迅猛发展极大地推动了机器翻译的进步[6]，其中最典型的代表之一是2022年11月推出的大语言模型ChatGPT。ChatGPT由美国OpenAI公司开发，作为从弱人工智能转向强人工智能的里程碑技术，在自然语言处理领域的文本生成、问答、摘要和翻译等多个任务中都表现得较为出色。与传统的专为翻译任务而设计的NMT不同，ChatGPT的通用语言理解能力使其更适用于翻译等各种与语言相关的任务[7]。当下，基于ChatGPT等大语言模型开展机器翻译研究已成为翻译学界的新课题[8]。

国外学界对于ChatGPT的翻译研究大多采用定量方法，多数研究发现ChatGPT的翻译水平优于传统神经机器翻译[9-10]。由于富含复杂细微的文化内涵和独特的语言结构，文学、诗歌、喜剧以及文化文本的翻译曾被认为是机器翻译的一项艰巨挑战[11]。然而，最近的研究表明，凭借其先进的算法，ChatGPT除了对源语言进行忠实的翻译之外，还可以处理复杂细微的语言差别，尤其能够理解意象、特定民族语言的独特词汇以及各种句法结构中不同体裁的独特形式[12]。

国内学界，主要关注ChatGPT与翻译研究范式[13]，ChatGPT与翻译教学[14-16]，ChatGPT与翻译伦理[17-20]，ChatGPT与译后编辑[21]，以及ChatGPT翻译质量评估[8,22-25]。关于翻译质量评估，国内研究较少。有学者从法律[22]、政治[23]、生物医药[24]、文学[25]等领域进行了翻译质量评估，发现LLM的翻译质量整体优于NMT工具。虽然ChatGPT对于部分例子的文化负载词处理得当，但整体来看，其对文化负载词的适应性尚不充分[25]。

可见，ChatGPT类大语言模型赋能翻译实践与研究已经成为翻译界的发展趋势。总体而言，国内外已有研究指向：LLM在某些领域的翻译质量优于传统NMT，但依然无法达到人工翻译的水平；LLM在一定程度上能够理解语言所包含的细微复杂的文化内涵和概念隐喻。

2 中医药与机器翻译

中医药文化是中华优秀传统文化的重要组成部分，是打开中华文明宝库的钥匙。中医语言富含中国文化内涵和大量隐喻现象[26]；中医典籍多为文言文，翻译过程涉及语内翻译和语际翻译；且中医名词术语尚未实现国际标准化，这些原因使得中医翻译的难度较大。

关于中医机器翻译的研究，前人研究比较稀少，尚处于起步阶段。曾有学者基于前端技术研发探讨中医机器翻译系统开发，指出：（1）中医语言具有术语众多、发生学特征的恒久不变性和科技语言属性等特征，这使得中医语言适合机器翻译[27]；（2）国内外在中医药领域信息处理技术方面均取得进展，已能在一定程度上实现对中医术语的自动翻译[28]。尚未有研究对中医文本的机器翻译表现进行定量的质量评估，更未发现基于ChatGPT类大语言模型在中医领域的翻译质量研究。

术语是普及科学知识、传播科学思想的基本单元，是不同领域之间及同领域内部不同分支之间进行传播的符号单元[29]。中医语言术语繁多，理解中医药术语是理解中医药文化精髓的关键[30]，然而由于中医用语的标准化程度不高，大量且不统一的术语是读者理解中医文本的一个严重障碍[27]，因此，术语翻译是中医药翻译的基本问题之一[31]。

基于以上背景，本研究以中医药术语为切入点，采用自动评价和专家人工评价相结合的半自动机器翻译评价方法，评估并比较大语言模型（如ChatGPT-4）和神经机器翻译工具（如Google翻译）的中医药术语翻译质量，旨在为人工智能时代中医译者选择较优的机器翻译工具提供实证参考。同时，发现大语言模型应用于中医药翻译的优势和局限性，进而探

讨AI时代人机协同的中医翻译策略，为AI赋能中医翻译和国际传播提供理论支持。

3 方 法

3.1 研究材料 研究材料为世界中医药学会联合会主编的《中医基本名词术语中英对照国际标准》（第1版）（以下简称《标准》）中的术语。然而，由于《标准》所含术语数量庞大，需从中选取具有代表性的术语。为此，研究以《黄帝内经》为切入点，筛选出《黄帝内经·素问》与《标准》匹配的术语。《黄帝内经》是我国最早且较为系统完整的医学典籍，记载了丰富的中医药知识和临床经验，在中医药文化中具有重要地位[32]。作为中医药文化的核心典籍之一，《黄帝内经》的翻译和研究不仅对于中医药术语的规范化具有重要价值，也为中医药文化的国际传播提供了重要参考[33]。同时，《黄帝内经》的英语译介传播史已近百年，已产出一批国内外译本[34]，具有丰富的英语语料和翻译研究成果，便于在前人的基础上深入分析。

本研究借助ChatGPT-4提取《标准》与《黄帝内经·素问》匹配的术语，统计分析这些术语的出现频次，并根据《标准》对术语进行分类。研究共提取出798条术语，这些术语在《黄帝内经·素问》共出现了12 609次，根据《标准》的术语类别划分方法，该798条术语分别归属于以下21个类别。（见表1）

表1 术语类别

序号	分类	序号	分类	序号	分类
1	学科、专业人员	8	病机	15	外科病
2	阴阳五行	9	诊法	16	妇科病
3	藏象	10	辨证	17	儿科病
4	形体官窍	11	治则治法	18	眼、耳鼻喉科病
5	气血津液精神	12	中药	19	骨伤科病
6	经络	13	方剂	20	针灸
7	病因	14	内科病	21	养生康复、五运六气

3.2 机器翻译（MT）工具 本研究选取Google翻译作为传统神经机器翻译的代表，选取ChatGPT-4作为大语言模型的代表，对两款MT工具的中医药术语翻译质量进行对比研究。

Google翻译（https://translate.google.com）是谷歌公司开发的一款基于网络的神经机器翻译系统，它使谷歌公司成为最大的机器翻译提供商之一。Intento公司和e2f公司合作发布的The State of Machine Translation显示，在中英互译的COMET评价中，Google翻译表现最好[35]。ChatGPT（https://chat.openai.com/）由OpenAI公司于2022年底发布，是一种公开可访问的高级会话人工智能模型，旨在产生连贯的和上下文感知的反应。目前ChatGPT-4是GPT系列中性能较强的模型。

3.3 提示词（prompt） 作为一款聊天机器人，ChatGPT-4经过训练，可以根据用户提供的提示词进行相应的回复。这显示了ChatGPT-4对提示词的敏感性[36]。换言之，ChatGPT-4的使用和提示词息息相关，不同的提示词会产生不同的结果[23]。为了保证ChatGPT-4翻译结果测试的信度和效度，研究者将与ChatGPT-4对话的基础提示词（base prompt）统一设定为：翻译为英文。然而，作为NMT翻译模型，Google翻译不需要用户输入任何指令，也不具有ChatGPT类大语言模型的互动功能。

3.4 机器翻译质量评估方法 目前，机器翻译评测的方法主要有3种：一是以双语评估替换（bilingual evaluation understudy，BLEU）为代表的自动化评价方法[37-38]；二是对机器翻译译文进行归类、打分、排序的人工评分方法[39-40]；三是自动评价与人工评价相结合的半自动评价方法[41]。本研究采用人机结合的半自动评价方法，自动评价指标采用BLEU、翻译编辑率（translation edit rate，TER）和显式排序翻译评估指标（metric for evaluation of translation with explicit Ordering，METEOR）共3项自动评价指标，人工评价基于著名中医翻译专家的直接打分。

常见的自动评价方法通过统计机器翻译结果与参考译文之间的相似性来显示译文质量，通用的评估指标有BLEU、TER、METEOR、CHRF及NIST等[42]。BLEU是IBM公司针对人工评价高代价且无法复用的缺点所提出的一种有参考译文的自动评估方式[43]。该指标的取值范围为0~1，分值越高，机器翻译译文质量越高。由于BLEU指标接近人工评分，所以在有参考译文的自动评估方法中，BLEU是使用最多的评测指标[41]。TER是一种编辑距离算法，用来衡量机器翻译结果与参考译文之间的编辑距离，该指标的取值范围为0~1，TER分数越低，表示机器翻译结果越好。另外，与BLEU不同，TER考虑了翻译中的错译和漏译等问题[44]。METEOR基于单精度的加权调和平均数和单字召回率，弥补了BLEU标准中的一些固有缺陷[45]，增加了其他指标中没有的一些功能，如同义词匹配等[42]。该指标的取值范围为0~1，分值越高，说明机器翻译质量越好。将BLEU、TER和METEOR结合使用，可以综合考虑翻译的精确度、编辑距离和流畅性，弥补单一指标的不足，提供较为全面准确的翻译质量评价。

首先，将798条术语分别按照类别批量输入ChatGPT-4与Google翻译，获得翻译文本。接下来，由于《标准》中对部分术语提供多个译文，为保证BLEU、TER、METEOR的信度和效度，需要对数据进行清洗：若机器翻译译文与其中一条译文一致，则保留该一致性译文；若不一致，则随机保留一条译文。最后，使用半自动机器翻译评价方法进行翻译质量评测：使用BLEU、TER和METEOR对ChatGPT-4和Google翻译的798条术语的翻译结果进行翻译质量评价；同时采用问卷的方法，由业内著名中医翻译专家对两款模型的翻译结果抽样进行评分，专家评分采取盲测形式，即专家在不知道翻译来源的情况下进行打分。

4 结 果

4.1 自动评价结果

4.1.1 术语翻译质量整体情况 本研究以《标准》中的术语译文为参考译文，使用BLEU、TER和METEOR共3个自动评估指标对LLM（ChatGPT-4）和NMT（Google翻译）的中医术语翻译质量进行定量自动评估。

如表2所示，与Google翻译相比，除“外科病”和“针灸”类术语外，ChatGPT-4在其他19个术语类别的BLEU值更高；除“骨伤科病”和“治则治法”类术语外，ChatGPT-4在其他19个术语类别的TER值皆低于Google翻译；除“外科病”外，ChatGPT-4在其他20个术语类别的METEOR值较高。整体上ChatGPT-4的BLEU均值较高、TER均值较低、METEOR均值较高。因此，

自动评估结果证明：ChatGPT-4的中医药术语翻译质量优于Google翻译。

为了验证ChatGPT-4与Google翻译的各项自动评估结果是否具有显著性差异，研究者进行了t检验（原始数据符合正态分布）。

表3显示了t检验的具体结果，ChatGPT-4和Google翻译的BLEU值、TER值和METEOR值的P值分别为0.020 689、0.000 026、0.000 006，均低于0.05，“可见，ChatGPT-4和Google翻译的各项自动评估结果均具有显著性差异（P<0.05），因此，ChatGPT-4的翻译质量整体显著优于Google翻译。”

表2　ChatGPT-4与Google翻译的译文BLEU值、TER值和METEOR值

序号	术语类别（领域）	术语数量	ChatGPT-4			Google翻译		
			BLEU	TER	METEOR	BLEU	TER	METEOR
1	学科、专业人员	1	0.215 2↑	0.750 0↓	0.117 1↑	0.153 6	1.000 0	0.000 0
2	阴阳五行	31	0.322 8↑	0.467 4↓	0.270 0↑	0.312 5	0.573 8	0.226 3
3	藏象	63	0.386 9↑	0.565 7↓	0.262 6↑	0.383 8	0.676 3	0.210 2
4	形体官窍	50	0.263 3↑	0.648 0↓	0.241 6↑	0.329 9	0.674 0	0.188 3
5	气血津液精神	20	0.303 6↑	0.525 0↓	0.300 0↑	0.164 5	0.787 5	0.115 8
6	经络	14	0.418 8↑	0.410 7↓	0.260 2↑	0.166 0	0.892 9	0.069 5
7	病因	39	0.284 0↑	0.541 3↓	0.280 9↑	0.248 9	0.780 9	0.181 3
8	病机	95	0.394 6↑	0.501 7↓	0.330 5↑	0.369 5	0.612 0	0.217 3
9	诊法	104	0.284 9↑	0.589 7↓	0.247 3↑	0.270 5	0.678 4	0.236 7
10	辨证	4	0.316 9↑	0.666 7↓	0.179 8↑	0.159 6	0.833 3	0.072 9
11	治则治法	69	0.167 1↑	0.813 1	0.140 2↑	0.166 7	0.805 8↓	0.132 4
12	中药	20	0.248 3↑	0.783 3↓	0.100 3↑	0.125 3	0.875 0	0.063 5
13	方剂	6	0.172 2↑	0.833 3↓	0.095 5↑	0.065 8	1.000 0	0.033 0
14	内科病	94	0.361 2↑	0.554 6↓	0.230 5↑	0.276 7	0.710 6	0.157 0
15	外科病	12	0.268 0	0.555 6↓	0.224 4	0.423 9↑	0.583 3	0.230 8↑
16	妇科病	5	0.320 8↑	0.600 0↓	0.197 0↑	0.065 7	0.933 3	0.042 3
17	儿科病	1	0.203 3↑	0.500 0↓	0.200 0↑	0.188 0	1.000 0	0.195 1
18	眼、耳鼻喉科病	14	0.277 4↑	0.642 9↓	0.189 6↑	0.176 3	0.892 9	0.089 8
19	骨伤科病	3	0.344 7↑	0.666 7	0.438 9↑	0.222 1	0.666 7	0.299 3
20	针灸	110	0.159 1	0.754 5↓	0.092 5↑	0.310 9↑	0.795 5	0.083 0
21	养生康复、五运六气	43	0.236 2↑	0.707 1↓	0.206 0↑	0.142 5	0.832 4	0.116 0
	平均数		0.283 3↑	0.622 7↓	0.219 3↑	0.224 9	0.790 7	0.140 9
	最大值		0.418 8	0.833 3↓	0.438 9↑	0.487 1↑	1.000 0	0.299 3
	最小值		0.159 1↑	0.410 7↓	0.092 5↑	0.065 7	0.541 7	0.000 0
	标准差		0.074 9↓	0.118 5↓	0.084 9	0.102 0	0.135 4	0.081 1↓

表3　ChatGPT-4与Google翻译的t检验结果

评估指标	P	ChatGPT-4平均数	Google平均数	差值	差值的标准误差	t	自由度	校正P值
BLEU	0.020 689	0.283 3	0.224 9	0.058 41	0.023 25	2.512	20.00	0.020 689
TER	0.000 026	0.622 7	0.790 7	-0.168 0	0.030 98	5.421	20.00	0.000 053
METEOR	0.000 006	0.219 3	0.140 9	0.078 30	0.012 85	6.093	20.00	0.000 018

4.1.2　术语领域的鲁棒性　在机器翻译和自然语言处理中，领域（domain）指的是特定主题、行业或应用场景下的语言数据集合，其特点体现在术语、句法结构、风格和语境等方面[46]。领域鲁棒性（domain robustness）指的是机器翻译系统在不同领域之间的稳定性和适应能力[47]。一个具有较强领域鲁棒性的机器翻译系统能够在多个领域中保持较高的翻译质量，即使遇到未见或低资源领域，也能提供合理的翻译，而不会出现明显的翻译质量下降。术语处理是机器翻译领域鲁棒性的一个重要的研究问题。

研究将每个中医术语类别（如病因、病机等）视为一个独立的术语领域，这些领域的术语使用和语言风格存在显著差异，因此可以用来评估模型在不同术语领域中的鲁棒性。

ChatGPT-4在不同术语领域中的翻译表现存在差异（见表2）。如在“经络”领域，ChatGPT-4的BLEU值表现最好（0.418 8），表明模型在该领域中的术语翻译质量较好；在“针灸”领域，ChatGPT-4的BLEU值最低（0.159 1），表明模型在该领域中的翻译质量较差。这种差异说明ChatGPT-4在特定术语类别中的鲁棒性较弱，模型性能仍需改进。

为探索ChatGPT-4与Google翻译的领域差异表现有何不同，研究者分别计算了两款模型在每类术语领域中自动评估得分之间的标准差。标准差越小，说明该模型在处理不同的术语类别时，其翻译自动评估得分的数据波动越小、领域差异越小、翻译质量也越稳定。如表2所示，ChatGPT-4的BLEU标准差（0.074 9）低于Google翻译的BLEU标准差（0.102 0）；ChatGPT-4的TER标准差（0.118 5）低于Google翻译的TER标准差（0.135 4）；虽然ChatGPT-4的METEOR标准差（0.084 9）略高于Google翻译的METEOR标准差（0.081 1），但差别极其微小。所以，从整体来看，ChatGPT-4在各领域术语之间的翻译质量波动小于Google翻译，说明在翻译中医药术语时，ChatGPT-4的领域鲁棒性优于Google翻译。

综合各数据，总体上ChatGPT-4在翻译中医药术语时的领域差异更小，其领域鲁棒性更好，领域泛化能力和迁移能力[48]更优。

4.2　专家评分结果　为了提供更加准确、全面、深入的分析，研究者邀请两位中医药翻译领域的专家对翻译进行人工评价，下面为具体论述。

由于术语数量众多，为确保人工评价样本的代表性，采用目的性抽样方法选择样本，即根据研究目的对研究对象进行针对性抽样[49]。抽样术语主要分为三类：西方词典收录的中医术语、中医隐喻术语和中西医同名异义术语。

按照上述类别进行术语抽样的原因在于：（1）西方词典收录的中医术语已经获得国际学术界和医学领域的认可，代表了中医药术语跨文化翻译的成功范例，因此分析这些术语有助于评估模型在翻译“已被目标受众接受的术语”时的准确性与适应性。（2）中医语言是一种基于隐喻认知的语言[50]，术语隐喻作为中医隐喻翻译研究的重要切入点[51]，其翻译难度较大，涉及认知映射、语义转换及文化背景的适配[52]。对中医隐喻术语翻译的探究，有助于揭示翻译模型在处理概念隐喻时的语义理解表现。本研究根据莱考夫和约翰逊对概念隐喻的分类，将中医隐喻术语分为三类，即结构隐喻、方位隐喻和实体隐喻[53]。（3）在汉语中，中西医存在同名异义的术语，即名称相同、含义不同[54]，研究这些术语的翻译具有重要的研究价值，它们能够有效测试模型在概念区分和语境适配方面的能力。

研究者随机抽取25条术语并交由两位中医药翻译资深专家进行评分，具体包括西方词典收录的术语5条；中医隐喻术语15条（结构隐喻术语5条、方位隐喻术语5条、实体隐喻术语5条）；中西医同名异义术语5条。（见表4）

本文对两种翻译工具译文的评分进一步分析，发现对于每类术语（西方词典中收录的术语、中医隐喻术语、中西医同名异义术语），ChatGPT-4的专家得分均值皆明显高于Google翻译（90 vs.60；51.33 vs.15；60 vs.0）（见表4）。因此，人工评估结果亦表明，ChatGPT-4的中医药术语翻译质量总体优于Google翻译，这一结果与论文“4.1”部分自动评估结果一致。

4.2.1 西方词典收录的中医术语 伴随着西方词典对部分中医术语的收录，这些中医术语实现了国际标准化。对于此类术语的翻译，ChatGPT-4的专家评分均值高达90分，Google翻译的专家评分均值为60分。其他两类术语的翻译质量得分显著低于此类术语。然而，个别术语的翻译依然表现不好，对于“气”的翻译，ChatGPT-4的平均得分为55分，Google翻译为0分。目前，“气”在中医语境下翻译为“chi”或“qi”。此译法已在国际上达成共识，并收录进《美国传统英语词典》（*The American Heritage Dictionary of the English Language*）[55]。但Google翻译未能识别“气”作为中医术语的特殊含义，而是将其误译为“gas”（意为气体）。相比之下，ChatGPT-4将“气”翻译为“energy or life force”，比较符合中医“气”的内涵。历史上，译者在翻译“气”时，曾大量采用“energy”这一译法[56]，所以可以推断，ChatGPT-4对“气”的翻译在一定程度上延续了这一传统。这个翻译版本虽然在语义层面上较为准确，但是明显受到过时语料的影响。

4.2.2 中医隐喻术语 对于中医隐喻术语，ChatGPT-4的翻译质量优于Google翻译。但是两个模型表现均不佳，平均评分均低于60分。ChatGPT-4的中医隐喻术语翻译得分平均值为51.33，Google翻译的中医隐喻术语翻译得分平均值为15（见表4）。翻译“正气”时，ChatGPT-4和Google翻译都未识别它在中医药文化中的隐喻含义。但ChatGPT-4仍在某些术语的翻译中表现较好，以“营卫”为例，在古代，“营”指军队的营盘，起保护作用。在中医理论中，“营卫”是营气和卫气的合称。它

表4 由专家进行评分的术语及其译文

类别		术语	ChatGPT				Google翻译			
			译文	专家评分1	专家评分2	专家评分均值	译文	专家评分1	专家评分2	专家评分均值
西方词典收录的术语		气	energy or life force	50	60	90.00	gas	0	0	60.00
		阴阳	yin and yang	100	100		yin and yang	100	100	
		经络	meridians and collaterals	100	100		meridians	50	50	
		五行	five phases	90	100		five elements	100	100	
中医隐喻术语	结构隐喻	邪气	pathogenic qi	90	80	51.33	evil spirit	0	0	15.00
		正气	upright qi	0	50		upright	0	0	
		传道之官	the organ of transmission	80	90		missionary officer	0	0	
		病机	mechanism of disease	100	100		pathogenesis	100	100	
		头者精明之府	the head is the mansion of essence and clarity	20	30		the shrewd house of the leader	0	0	
	方位隐喻	下厥上冒	cold limbs with head rush	0	0		coming down and coming up	0	0	
		上虚下实	upper deficiency and lower excess	50	70		the top is empty and the bottom is real	0	0	
		里急	urgency	0	0		tenesmus	20	30	
		诸逆冲上，皆属于火	all reverse rushing upward is attributed to fire	20	30		all oppositions belong to fire	0	0	
		怒则气上	anger causes qi to rise	80	70		weak temper	0	0	
	实体隐喻	营卫	nutritive and defensive qi	100	100		camp guard	0	0	
		气门	air gate	0	0		valve	0	0	
		肝风	liver wind	100	100		liver wind	100	100	
		骨者髓之府	the mansion of marrow in the bone	0	0		house of bones and marrow	0	0	
		胃者水谷之海	stomach, the sea of food and drink	80	100		the stomach is the sea of water and valleys	0	0	
中西医同名异义术语		精	jing (essence)	50	100	60.00	refined	0	0	0
		神	shen (spirit)	50	100		god	0	0	
		胞	membrane	0	0		cells	0	0	
		穴	acupoint	100	100		hole	0	0	
		脏	viscera	50	50		dirty	0	0	

们是人体内的两种重要的气，共同维持生命的活动。营气主要来源于饮食，运行于血脉中，具有营养全身、化生血液的功能。卫气主要来源于呼吸，运行于体表，负责防御外邪、调节体温和汗液分泌。“营卫”具有隐喻性。营气象征内部的滋养与维持，类似于后勤保障；卫气象征外部的防御与保护，类似于边防军队。这种隐喻反映了中医对人体内外平衡的重视，强调内外协调对健康的重要性。Google翻译将其直译为“camp guard”，未能准确把握其内涵；而ChatGPT-4识别了“营卫”在中医药文化中的隐喻用法，将其翻译为“nutritive and defensive qi”，精准地传达了中医药术语的内涵，尽管与参考译文“nutrient and defense”仍有轻微差别，但其翻译质量均受到两位专家认可。

4.2.3　中西医同名异义术语　在翻译中西医同名异义术语时，ChatGPT-4的得分平均值为60.00，而Google翻译的得分平均值为0，两款模型表现出明显的差异（见表4）。ChatGPT-4在保留术语的专业性和文化内涵方面做得更好，而Google翻译则更多地依赖于通用词汇，语境适应性弱、标准化不足、文化负载词处理欠佳。如：ChatGPT-4将“精”翻译成“jing（essence）”，“神”翻译为“shen（spirit）”。这些翻译采取了“拼音加英文注释”的策略，有助于外国受众理解术语在中医文化的独特内涵。相比之下，Google翻译则更多地采用了直译或常见词汇或西医对应词汇的翻译策略，如：“精”译为“refined”，“神”译为“god”，“胞”译为“cells”，“穴”译为“hole”，“脏”译为“dirty”。明显可见，Google翻译更多地依赖于通用语料库或者西医语料库，缺乏对中医文化背景的深入理解，因此在处理这类具有特定中医文化内涵的术语时，其表现不容乐观。而ChatGPT-4处理中医术语翻译时，能够更加准确地把握其含义，这可能是因为在该模型的训练过程中，训练语料库包含了更多关于中医的专业文献和资料，且ChatGPT-4的情境学习能力和适配能力更好。然而，ChatGPT-4对于这类术语的翻译并非准确无误，其人工评价分数为60分，表明其提供的翻译结果与人工译本仍然存在明显差距，因此该模型在翻译中西医同名异义术语方面有待提高。

综合自动评估和专家评估结果，ChatGPT-4的中医术语翻译质量明显优于Google翻译。自动评估结果显示，ChatGPT-4的BLEU值、TER值和METEOR值均优于Google翻译，其在词汇匹配、错误率控制和复杂语言结构处理等方面表现更好，且具有更强的领域鲁棒性。专家评分进一步证实了ChatGPT-4较高的翻译质量，尤其是在处理隐喻性术语和中医文化负载词方面ChatGPT-4的得分更为突出。总体而言，ChatGPT-4在中医术语翻译中表现出较高的准确性和适应性，但仍需进一步优化其在特定领域的翻译质量。（见图1）

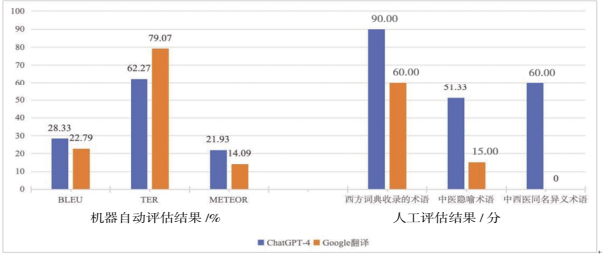

图 1　机器自动评估结果和人工评估结果总体情况

4.3　提示词测试　在大语言模型应用领域，提示词作为引导模型生成特定输出的关键性输入文本，是与大语言模型进行交互的核心机制。已有研究表明，有效的提示词工程（prompt engineering）对于充分发挥大语言模型的生成能力发挥着至关重要的作用[57]。本研究通过实验进一步验证了这一观点，发现在中医药术语翻译任务中，优化指示词能够显著提升ChatGPT-4的中医术语翻译质量。（见图2）

在初始实验中，研究者使用基础提示词（base prompt）“翻译为英文”时，ChatGPT-4将“不得前后”直译为“unable to move forward or backward”，这一翻译结果未能准确传达该术语在中医药语境中的特定含义。随后，研究者对提示词进行了优化，提供了更为详细的上下文信息：“不得前后，证名。前，指小便；后，指大便。指二便不通或大小便失常”。在此提示词的引导下，ChatGPT-4生成了更为准确的翻译结果：Obstruction of both urination and defecation。这一结果表明，ChatGPT-4能够根据给定的提示词和上下文语境，逐步调整并优化翻译结果，从而在一定程度上提高了翻译的质量。

基于上述发现，本研究认为，合理运用提示词工程（prompt engineering）对于人类译员借助LLM提升中医药翻译质量具有重要的实践意义。相比之下，传统的NMT模型缺乏情境学习能力（in-context learning），无法像ChatGPT类大语言模型那样通过上下文信息动态调整翻译结果。这一局限性使得传统神经机器翻译模型在应对复杂语境和专业术语翻译时，难以达到与ChatGPT类大语言模型相媲美的翻译质量。因此，提示词工程的应用不仅为大语言模型在中医专业领域的翻译实践提供了新的优化路径，也为机器翻译与中医译员的协作模式带来了新的可能性。

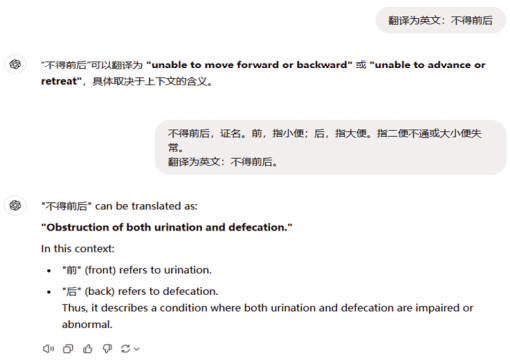

图 2　单样本（one-shot）提示词测试案例截图

5　讨　论

5.1　ChatGPT-4应用于中医药翻译的优势

5.1.1　翻译质量显著优于传统神经机器翻译模型　通过自动评价（BLEU、TER及METEOR）和专家人工评价相结合的半自动评价方式，系统比较了ChatGPT-4与Google翻译在中医术语翻译中的表现。研究发现，ChatGPT-4的翻译质量显著优于Google翻译，尤其是在处理多义性和隐喻表达方面，大语言模型的表现尤为出色。ChatGPT-4能够在一定程度上识别中医药术语中的隐喻现象，并在翻译中保留中医药的文化内涵，而Google翻译则更多地依赖直译的方法以及通用领域或西医领域的词汇，导致译文在语义和文化适配性上存在较大差距。

此外，ChatGPT-4在翻译部分中西医同名异义术语时，能够通过提供“拼音和附加英文解释”的方式，更好地保留术语的专业性、识别并理解中医的文化背景，而Google翻译则往往无法准确区分这些术语在中医和西医中的不同含义。

5.1.2 领域鲁棒性较强，翻译质量稳定 在翻译21类中医术语时，ChatGPT-4表现出较好的领域鲁棒性。在批量输入不同类别的术语并获得翻译结果后，研究者对结果进行定量分析，发现：ChatGPT-4的跨术语类别翻译质量波动幅度较小，证明该模型在应对中医细分领域术语转换时能保持稳定的翻译性能。相比之下，Google翻译的翻译质量在不同术语类别间呈现较大的差异。这表明ChatGPT-4具有更好的领域鲁棒性，其领域泛化能力和迁移能力更优。因此，ChatGPT-4更适合应用于需要连贯翻译的中医药文本翻译场景，如中医药典籍的翻译。

5.1.3 通过提示词工程提升翻译效果 以ChatGPT-4为代表的大语言模型展现出了多种能力，包括情境学习能力（模型可以通过示例学会某项任务）[58]、指令跟随能力（模型具有跟随人类指令作出回复的能力）[59]和复杂推理能力（模型可以进行多步的常识推理、数学推理）[60-62]。本研究通过交互性测试实验证明：优化提示词可以提升ChatGPT-4的中医药术语翻译质量。这再次说明大语言模型的情境学习能力、指令跟随能力和复杂推理能力使其能够实现动态自适应的翻译优化，而这种适应性和灵活性，是传统的神经机器翻译模型所不具备的。

5.2 ChatGPT-4应用于中医药翻译的局限与挑战

5.2.1 大语言模型本身的技术局限 尽管ChatGPT-4在中医药术语翻译中的表现优于神经机器翻译模型，但其仍存在技术局限。首先，大语言模型可能会受到训练语料库中的文化定势或偏见的影响。大语言模型的偏见指模型的输出呈现刻板印象或对特定实体的代表性不足或过度，大量研究表明大语言模型呈现不同程度的刻板印象[63]。例如，ChatGPT-4在翻译“胞”时采用了偏向西医的“membrane”，未能准确反映中医术语的文化内涵，反映了中医语料的代表性不足。此外，LLM也容易受到时间偏差的影响[64]，其数据可能缺乏时间标注，从而限制模型对历史情境或过时信息的理解。例如，ChatGPT-4在翻译“气”时使用了“energy or life force”这一过时的译法，表明LLM的训练有可能受到历史过时语料的影响。

5.2.2 文化内涵与隐喻表达的挑战 中医药术语富含独特的文化内涵和隐喻性表达，这对大语言模型的翻译能力提出了更高要求。尽管ChatGPT-4在处理隐喻术语时的表现优于Google翻译，但其翻译质量仍存在较大的提升空间。例如，在翻译“正气”时，ChatGPT-4未能准确识别其在中医文化中的隐喻意义，而Google翻译则完全忽略了这一文化背景。这表明，在处理文化负载词和隐喻表达时，大语言模型的语义理解和文化适配能力仍有待优化。

5.2.3 大型高质量中医双语语料库的缺乏 数据是一切人工智能技术的基础，没有数据，一切AI技术都是空中楼阁。在预训练阶段，ChatGPT通过学习海量语料数据学习文字接龙能力，海量高质量的语料基础是ChatGPT技术突破的关键要素之一[41]。然而，当前中医药术语翻译的标准化程度较低，导致可用于训练的高质量中医英译语料库相对匮乏。现行中医术语英译标准以世界中医药学会联合会主编的《中医基本名词术语中英对照国际标准》、世界卫生组织的《中医药术语国际标准》和《传统医学名词术语国际标准》为主，但这些标准在实际应用中尚未完全统一，导致同一术语在不同语料库中的翻译存在较大差异。这种语料库的不一致性可能会影响大语言模型的训练效果，进而影响其翻译效果。尤其是在处理低资源领域或复杂语境时，模型的表现更加不乐观。

5.3 建 议

5.3.1 提示词的优化与应用 提示词工程在大语言模型翻译中的应用具有广阔前景。未来研究可以进一步探索如何设计更有效的提示词，以引导LLM生成更符合中医药语境的翻译结果。例如，可以通过提供更多上下文信息、术语解释或文化背景知识，帮助其更好地理解中医药术语的独特含义。

5.3.2 构建大型高质量中医双语语料库 高质量大型语料库和数据库是保证大语言模型得到良好训练的基石。中医药各界人士应共同努力打破数据孤岛的壁垒，开发建设高质量的大型中医双语语料库，同时需要关注中医药低资源语言语料库的开发。此外，建设中医药垂直领域跨语言大模型也是未来重要的举措。

5.3.3 继续加快推进中医名词术语标准化工作 中医术语国际标准是中医药对外传播的基础，只有基础统一、一致，才能达到交流的目的[65]。大语言模型是基于数据的技术，然而没有统一的术语，就没有统一的数据，因此当前中医药名词术语国际标准化任务变得更为急迫。政府、社会、学术、科技、翻译界等各个领域需通力合作，从而加快中医药名词术语国际标准化进程，赋能中医药数智化国际传播。

5.3.4 基于人机协同的中医翻译模式的探索 大语言模型的中医药翻译质量与人工译本相比尚有很大差距。人类译员应该学会高效地利用这一新的生产工具，与其形成协同合作的关系。尽管ChatGPT-4等大语言模型的不断优化使其在机器翻译领域具有广泛的应用潜力，但在许多专业和关键的翻译任务中，人类译员和机器翻译相结合的方法依然是最佳选择[23]。大语言模型结合人工编辑的交互模式，可以进一步提高翻译质量和效率[66-67]。未来研究可以探索如何将LLM的翻译能力与人类译员的中医药专业知识和文化背景相结合，形成高效的人机协同翻译模式。

5.3.5 中医翻译教育与人才培养的革新 翻译工具的革新也对中医药翻译教育提出了新要求。在当前智能技术大发展、大爆发的时代背景下，传统翻译教育难以应对技术带来的挑战，逐渐映射出其在人才培养模式、师生技术素养等方面的不协调[68]。翻译教育亟须形成与行业技术发展相匹配的创新教育理念[14]。身处AI时代，面对第四次产业革命的到来，中医药翻译教育界应在继续加强学生中医药知识和中医文化素养培养的前提下，深入探索翻译教育与智能技术的融合发展路径，提升师生的人工智能素养，培养学生的跨学科能力和整合思维，以满足人工智能时代中医药国际传播的发展需求，开辟人机共生、共创共享、跨界融合的智慧翻译教育新生态[14]。

5.3.6 中医翻译技术伦理的治理 传统译学的“信、达、雅”

是对人类跨语言创造力的美好期许，是追求真善美和人类福祉的一种表达方式。它不仅是翻译技术标准，更蕴含着深厚的伦理智慧。在AI领域，技术的发展和应用需坚持以人为本和以促进人类福祉为终极目标，同时需严格恪守AI伦理[69]。关于翻译技术的伦理反思，应从翻译技术发展和使用过程中对“善”的损害说起。“善”的价值蕴含“以促进人的发展为起点、以凸显人的价值为终点”[18]。因此，科技向善、翻译向善、翻译技术伦理向善、以人为本应成为翻译技术伦理治理的根本指导原则。在未来人机结合的中医药翻译实践中，译员应该坚守以人为本，遵守伦理原则，如严格执行中医药数据管理规范，保护数据安全和隐私安全；中医药翻译研究者应遵守研究伦理，避免滥用技术，同时在使用技术时，要保持批判性精神，发挥人的创造力[13]，实现工具理性和价值理性的和谐统一[70]。

6 结　论

人工智能背景下，如何利用新兴机器翻译工具大语言模型赋能中医药翻译和国际传播是重要的研究议题。已有研究证明ChatGPT类大语言模型在某些领域的翻译质量优于传统神经机器翻译工具，然而尚未发现有研究关注基于大语言模型的中医药翻译质量评估。

本研究以中医药术语为切入点，评估大语言模型的中医药翻译质量。通过机器翻译半自动评价方法，即自动评价与专家人工评价相结合，系统评估大语言模型（ChatGPT-4）与传统神经机器翻译工具（Google翻译）的中医药术语翻译质量，揭示了ChatGPT-4在中医翻译领域的优势与局限性，探讨了提升其翻译效果的途径和人机协同的翻译策略。

在自动评估中，ChatGPT-4的BLEU、TER和METEOR值均显著优于Google翻译，专家人工评分进一步验证了这一结果。在西方标准化术语、中医隐喻术语和中西医同名异义术语的翻译中，ChatGPT-4均表现更优，其在保留中医文化内涵和语境适配方面较为突出。相关实验发现，通过提示词优化可显著提升ChatGPT-4的翻译质量。

然而，大语言模型仍存在技术局限，比如文化定势、过时语料和对中医文化适应性的不足。此外，高质量、标准化的中医药双语语料库的缺乏制约了大语言模型的训练效果，从而影响其翻译输出的准确性和稳定性。

研究提出建议：通过设计专业化提示词，激发模型生成更优的译文；推进中医药术语国际标准化，打破数据孤岛，构建大型中医双语语料库；结合大语言模型的效率与人类译员的专业知识背景，探索大语言模型+人工编辑的人机协同机制；推动中医药翻译教育与人工智能技术的融合，培养兼具中医文化素养与人工智能素养的跨学科整合型人才，以适应AI时代人机协同的知识生产生态；坚持以人为本的原则，加强翻译技术伦理治理。

参考文献

[1] 王华树，刘世界.人工智能时代翻译技术转向研究[J].外语教学，2021，42(5)：87-92.

[2] 冯志伟.机器翻译研究[M].北京：中国对外翻译出版公司，2004.

[3] 李翔，高朝阳.国外机器翻译研究的知识图谱和发展趋势[J].上海翻译，2024(2)：41-47.

[4] HUTCHINS J. Machine translation: A concise history[J]. Journal of Translation Studies，2010(13)：29-70.

[5] WANG H F，WU H，HE Z J，et al. Progress in machine translation[J]. Engineering，2022，18：143-153.

[6] GAO R Y，LIN Y M，ZHAO N，et al. Machine translation of Chinese classical poetry: A comparison among Chat GPT，Google Translate，and DeepLTranslator[J]. Humanit Soc Sci Commun，2024，11：835.

[7] WU T Y，HE S Z，LIU J P，et al. A brief overview of ChatGPT: The history，status quo and potential future development[J]. IEEE/CAA J Autom Sin，2023，10(5)：1122-1136.

[8] 顾文昊，冷冰冰.ChatGPT在科技翻译应用中的四种术语误译类型：以机械工程术语为例[J].中国科技翻译，2024，37(1)：24-27.

[9] CHAN V. Investigating the impact of a virtual reality mobile application on learners´ interpreting competence[J]. J Comput Assist Learn，2023，39(4)：1242-1258.

[10] CHAN V，TANG W K. GPT and translation: A systematic review [C]//2024 International Symposium on Educational Technology (ISET). July 29 -August 1，2024，Macao，China. IEEE，2024：59-63.

[11] CASTILHO S，MALLON C Q，MEISTER R，et al. Do online machine translation systems care for context? What about a GPT model? [C]//Proceedings of the 24th Annual Conference of the European Association for Machine Translation. Tampere，Finland，2023：393-417.

[12] KARABAN V，KARABAN A. AI-translated poetry: Ivan Franko's poems in GPT-3.5-driven machine and human-produced translations[J]. Forum Linguist Stud，2024，6(1)：1-15.

[13] 余静，刘康龙.重塑翻译研究：AI技术影响下的范式转换与未来方向探索[J].外国语(上海外国语大学学报)，2024，47(4)：72-81.

[14] 王华树，刘世界.智慧翻译教育研究：理念、路径与趋势[J].上海翻译，2023(3)：47-51，95.

[15] 江先发，赖文斌.数智时代翻译教学的“ABC”路径探索[J].上海翻译，2024(1)：63-67.

[16] 张文煜，赵璧.生成式人工智能开创机器翻译的新纪元了吗?：一项质量对比研究及对翻译教育的思考[J].北京第二外国语学院学报，2024，46(1)：83-98.

[17] 王赟，张政.ChatGPT人工智能翻译的隐忧与纾解[J].中国翻译，2024，45(2)：95-102.

[18] 刘成科，孔燕.翻译技术伦理的本质追问及基本向度[J].外语学刊，2023(5)：79-85.

[19] 于浩，郭赟赟.风险与超越：ChatGPT赋能翻译的伦理分

析[J].中国翻译,2024,45(4):115-122.
[20] 王宁.面对ChatGPT的挑战:呼唤文学翻译中的伦理转向[J].外国文学研究,2024,46(3):18-27.
[21] 王律,王湘玲.ChatGPT时代机器翻译译后编辑能力培养模式研究[J].外语电化教学,2023(4):16-23,115.
[22] 宋丽珏.法律翻译的数字人文转型研究:以专题数据库与ChatGPT为中心[J].外语学刊,2024(2):51-57.
[23] 文旭,田亚灵.ChatGPT应用于中国特色话语翻译的有效性研究[J].上海翻译,2024(2):27-34,94-95.
[24] 王和私,马柯昕.人工智能翻译应用的对比研究:以生物医学文本为例[J].中国科技翻译,2023,36(3):23-26.
[25] 赵衍,张慧,杨祎辰.大语言模型在文本翻译中的质量比较研究:以《繁花》翻译为例[J].外语电化教学,2024(4):60-66,109.
[26] 贾春华.认知科学背景下的中医病因病机的概念隐喻研究[J].中国医药导刊,2008,10(8):1141-1143.
[27] 姚振军.面向中医典籍的机器翻译系统的开发[J].中国翻译,2007,28(2):72-75,95.
[28] 宋熹玥,周净,刘伟.中医古籍智能机器翻译模型构建研究[J].中国中医药图书情报杂志,2024,48(6):130-135.
[29] 张春泉.科技术语的语域传播论纲[J].中国科技术语,2016,18(6):9-12.
[30] 周恩,苏琳.中医药术语英译研究趋势、问题与展望[J].中国中西医结合杂志,2022,42(6):754-759.
[31] 李照国.论中医名词术语的翻译原则[J].上海翻译,1996(3):31-33.
[32] 王燕,李正栓.《大中华文库》科技典籍英译与中国文化对外传播[J].上海翻译,2020(5):53-57,94.
[33] 蒋继彪.中医典籍英译研究(2000—2022):成绩、问题与建议[J].中国中医基础医学杂志,2024,30(1):117-120.
[34] 闵玲.《黄帝内经》英译语料库建设现状探析[J].中医药导报,2022,28(7):209-212.
[35] Inrento, e2f. The state of machine translation 2023 [EB/OL]. (2024-06-21) [2024-12-21].https://inten.to/machine-translation-report-2023/?u=pr.
[36] GAO Y, WANG R L, HOU F. How to design translation prompts for ChatGPT: An empirical study[C]//Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops. Auckland New Zealand.ACM, 2024:1-7.
[37] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]//Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. Montreal, QC, Canada: NeurIPS,2014:3104-3112.
[38] JEAN S, CHO K, MEMISEVIC R, et al. On using very large target vocabulary for neural machine translation [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China. Stroudsburg, PA, USA: ACL,2015:1-10.
[39] BURCHARDT A, MACKETANZ V, DEHDARI J, et al. A linguistic evaluation of rule-based, phrase-based, and neural MT engines[J]. Prague Bull Math Linguist,2017,108(1):159-170.
[40] ISABELLE P, CHERRY C, FOSTER G. A challenge set approach to evaluating machine translation[C]//Proceedings of the 2017 Conference on Empirical Methods in NaturalLanguage Processing. Copenhagen, Denmark. Stroudsburg, PA, USA: ACL,2017:2486-2496.
[41] 李佐文,任佳伟.语言智能导论[M].北京:外语教学与研究出版社,2024.
[42] 韦佑武,李娜,赵良威.基于高频错误类型分析的机器翻译质量标准、质量评估与发展态势[J].计算机科学与应用,2022(10):2275-2281.
[43] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL '02. July 7-12, 2002. Philadelphia, Pennsylvania. Morristown, NJ, USA: ACL,2001:311.
[44] SNOVER M, DORR B, SCHWARTZ R, et al. A study of translation edit rate with targeted human annotation [C]//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 2006:223-231.
[45] LAVIEA, AGARWALA. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments [C]//Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07. June23,2007.Prague, Czech Republic. Morristown, NJ, USA: ACL,2007:228-231.
[46] JURAFSKY D, MARTIN J H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with LanguageModels [M/OL].3rdversion.2025. https://web.stanford.edu/~jurafsky/slp3/.
[47] 桂韬,奚志恒,郑锐,等.基于深度学习的自然语言处理鲁棒性研究综述[J].计算机学报,2024,47(1):90-112.
[48] GUO J F, LIU J, HAN Q, et al. Domain mining for machine translation[J]. J Intell Fuzzy Syst,2015,29(6):2769-2777.
[49] 陈向明.质的研究方法与社会科学研究[M].北京:教育科学出版社,2000.

[50] 贾春华.基于隐喻认知的中医语言研究纲领[J].北京中医药大学学报,2014,37(5):293-296.

[51] 方宝,吴红瑶.中医隐喻及其翻译[J].中国中西医结合杂志,2024,44(7):874-880.

[52] 范春祥.隐喻视角下中医典籍语言特点及其翻译研究[J].时珍国医国药,2012,23(11):2875-2876.

[53] LAKOFF G, JOHNSON M. Metaphors we Live by[M]. Chicago: University of Chicago Press,2003.

[54] 李永,陈启亮,赵文,等.中西医诊断术语的差异性比较研究[J].天津中医药,2020,37(9):972-975.

[55] The American Heritage Dictionary of the English Language[M]. Boston: Houghton Mifflin Harcourt,2011.

[56] 李孝英,苏赛迪.基于语料库的中医药文化负载词"气"的英译认知研究[J].翻译研究与教学,2024(2):137-144.

[57] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[C]// Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Nusa Dua, Bali. Stroudsburg, PA,USA:ACL,2023:675-718.

[58] DONG Q X, LI L, DAI D M, et al. A survey on in-context learning [C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA. Stroudsburg, PA, USA: ACL,2024:1107-1128.

[59] WEI J, BOSMA M P, ZHAO V, et al. Finetuned language models are zero-shot learners [C]// The Tenth International Conference on Learning Representations. Virtual Event: OpenReview,2022.

[60] XI Z H, JIN S J, ZHOU Y H, et al. Self-Polish: Enhance reasoning in large language models via problem refinement[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore. Stroudsburg, PA, USA: ACL,2023:11383-11406.

[61] Chain-of-thought prompting elicits reasoning in large language models [C]//Proceedingsofthe36thInternational Conference on Neural Information Processing Systems. 28 November 2022, New Orleans,LA,USA.ACM, 2022:24824-24837.

[62] WEBB T, HOLYOAK K J, LU H J. Emergent analogical reasoning in large language models[J]. Nat Hum Behav, 2023,7(9):1526-1541.

[63] 徐月梅,叶宇齐,何雪怡.大语言模型的偏见挑战:识别、评估与去除[J].计算机应用,2025,45(3):697-708.

[64] TAN Q Y, NG H T, BING L D. Towards benchmarking and improving the temporal reasoning capability of large language models [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada. Stroudsburg, PA, USA: ACL,2023:14820-14835.

[65] 赵霞.标准化视阈下中医术语英译标准[J].亚太传统医药,2015,11(20):141-143.

[66] MOORKENS J, O'BRIEN S, DA SILVA I A L, et al. Correlations of perceived post-editing effort with measurements of actual effort[J]. Mach Transl,2015,29(3): 267-284.

[67] YAMADA M. Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing settings[J]. Mach Transl,2015,29(1):49-67.

[68] 王华树,刘世界.元宇宙视域下翻译教育的发展前景与实践路径[J].北京第二外国语学院学报,2022,44(4):96-107.

[69] FILOY, RABIN E, MOR Y. An artificial intelligence competency framework for teachers and students: Co-created with teachers[J]. Eur J Open Distance E Learn,2024,26(s1):93-106.

[70] 马克斯·韦伯(Max Weber).经济与社会-上卷[M].约翰内斯·温克尔曼(Johannes Winckelmann),整理.林荣远,译.北京:商务印书馆,2004.

(收稿日期:2025-04-16 编辑:时格格)