Modern Linguistics 现代语言学, 2025, 13(9), 584-594
Published Online September 2025 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2025.1391015

Hans汉斯

海事文献人工智能翻译质量对比研究

刘　怡，王　宪

上海海事大学外国语学院，上海

收稿日期：2025年8月18日；录用日期：2025年9月5日；发布日期：2025年9月17日

摘　要

人工智能(AI)技术极大地促进了翻译效率，然而翻译质量良莠不齐，主要原因是翻译材料选取不一致。本文以海事绿色低碳领域的术语定义文件为基准材料，选用了BLEU、METEOR、TER和NIST等四种评价指标，对比分析DeepSeek v3、ChatGPT 4.0、DeepL、文心一言4.0和火山翻译等五种AI翻译平台，比较和分析了标准体系下不同平台的翻译质量。研究结果表明：ChatGPT 4.0在海事绿色低碳术语翻译中表现最佳，DeepSeek v3次之，两者在译文术语准确性，译文流畅性方面均优于其他AI翻译平台，可适配海事标准体系的术语翻译需求。本研究为海事标准体系术语翻译的质量评价提供了科学依据，并对AI翻译平台的选择与优化具有参考价值。

关键词

AI翻译平台，翻译质量评价，评价指标，海事技术类文本

A Comparative Study on the Translation Quality of Artificial Intelligence in Maritime Literature

Yi Liu, Xian Wang

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: Aug. 18th, 2025; accepted: Sep. 5th, 2025; published: Sep. 17th, 2025

Abstract

While AI-powered translation has significantly enhanced translation efficiency, the quality of translations varies greatly, primarily due to inconsistencies in the selection of source materials. This study takes a terminology definition document in the maritime green/low-carbon sector as the benchmark material, and adopts four assessment metrics, including BLEU, METEOR, TER and NIST.

文章引用: 刘怡, 王宪. 海事文献人工智能翻译质量对比研究[J]. 现代语言学, 2025, 13(9): 584-594.
DOI: 10.12677/ml.2025.1391015

It conducts a comparative analysis of five AI translation platforms (DeepSeek v3, ChatGPT 4.0, DeepL, ERNIE Bot 4.0, and Volctrans) with a view to comparing and analyzing the translation quality of different platforms within the framework of standard system documents. Findings reveal that ChatGPT 4.0 delivers the most accurate translations for maritime green/low-carbon terminology, followed closely by DeepSeek v3. Both platforms outperform others in terminological precision and linguistic fluency, demonstrating strong alignment with maritime standardization requirements. This study establishes a methodological foundation for assessing terminology translation quality in maritime standards documents, while providing practical insights for platform selection and algorithm optimization in domain-specific AI translation.

Keywords

AI Translation Platforms, Translation Quality Assessment, Assessment Metrics, Maritime Technical Texts

1. 引言

基于 Transformer 的 GPT 系列、Qwen、BERT、Llama 等大语言模型的人工智能(AI)技术，推动了金融，医疗，教育等行业的迅猛发展[1]，基于大语言模型的神经网络机器翻译工具可显著提高翻译效率和翻译质量[2]，然而这些平台在专业领域文本的翻译质量往往还不尽满意[3]，原因在于，翻译质量的评价使用的指标不同，有的指标关注词汇层面，有的关注句子层面，采用单一指标过于偏颇和武断；其次，源文本清洁度不够高，构建的语料库是专业领域的集合，包含了新闻报道、技术介绍等，没有从单一技术文件上考察翻译质量。为解决上述不足，本文基于单一海事术语文本，采用多个评价指标进行翻译质量评价，旨在分析 AI 翻译平台在标准类文本下的翻译质量，以期为翻译质量评价提供参考。

翻译质量评价(assessment metrics, AM)的指标较多，有 BLEU、TER、METEOR、NIST、BERTScore、COMET 和 BLEURT 等。通常多采用 BLEU 和 TER 两种评价指标。本研究梳理了相关评价指标(AM)，在前人研究成果的基础上加入了 NIST，METEOR，对 AI 翻译平台的翻译质量进行了系统评价与分析。本文采用的 4 种评价指标，增加了评价指标的综合衡量，可更全面、准确地评价翻译质量。

国际海事组织(IMO)积极应对气候变化，通过了《2023 年国际海事组织船舶温室气体减排战略》，明确在 2050 年前后实现温室气体净零排放。这表明绿色航运已经成为当前及以后一段时间内的主要领域。为了评价不同 AI 翻译平台对于海事文本的翻译质量，本研究选用《MEPC 81-7-3-Glossary of climate change definitions in relation to shipping》。该海事文本主要由海事温室气体减排的术语定义构成，属于气候减排的标准和规范文件。目前，学界的文本多是非单一的非术语型文件。雷鹏飞和张浮凌(2024)构建了两个外宣文本语料库进行翻译质量评价研究[4]，范新瑜(2023)则进行了海洋科技文本机器翻译译文质量的评测[5]，王坤宇和成思(2025)使用 ChatGPT 4.0 研究了立法文本翻译及质量评价[6]。对比上述翻译质量评价使用的文本，本研究所使用的海事绿色减排标准术语文本，具有单一性和标准性，能更加精确地评价不同 AI 翻译平台对标准体系文本的翻译质量。

本文的结构分为五部分。第一部分为引言，介绍了研究的背景，意义和目标，并概述了研究内容和创新点。第二部分为文献综述，梳理了国内外关于机器翻译质量评价的研究现状。第三部分为研究设计，介绍了本文的语料来源，选取的 AI 翻译平台，四种评价指标以及研究方法。第四章研究结果与讨论，呈

现了各项指标的评价结果，结合具体实例，从准确性、流畅性、稳定性三个方面对比不同 AI 翻译平台的表现，分析其优势与不足。第五部分为结语，总结了研究发现，不足和意义。

2. 文献综述

机器翻译普遍采用 BLEU、TER、METEOR、NIST、ROUGE、BERTScore、COMET 和 BLEURT 等指标衡量译文质量，其原理是衡量机器译文与参考译文的相似度，不同相似度的对比角度决定了不同评价指标。指标根据计算方法和语言特点可分为：基于词汇匹配的指标，BLEU、TER、NIST 和 ROUGE；基于语义相似性的指标，METEOR、BERTScore、COMET 和 BLEURT。

目前，多数学者采用多种评价指标研究翻译质量。有的学者使用 BLEU 一种指标进行评价，如张文煜和赵璧(2024)采用 BLEU 与人工评价的方式，对 6 种不同体裁(小说和散文等)的 ChatGPT 4.0、有道翻译、DeepL 翻译的机器翻译文本进行测评，发现 GPT 技术在文学翻译等方面已有质量提升，但未取代神经网络机器翻译[1]。郭望皓和胡富茂(2021)利用 BLEU 算法对比评测 5 个翻译系统在 1000 句军事文本及通用文本中的表现，发现当前神经机器翻译系统尚无法高质量翻译军事文本[3]。范新瑜(2023)结合 BLEU 值和人工评审，对有道、百度等四大机器翻译系统翻译的海洋科技文本进行质量评价。人工评审和 BLEU 值均显示有道翻译的平均质量最优[5]。有的学者采用了两种或三种评价指标。王子云、毛毳(2023)和文旭、田亚灵(2024)都运用了 BLEU 和 TER 值两个指标分别对淄博陶瓷琉璃博物馆中英介绍文本以及党的二十大报告中 10 个例句的译文质量进行了评价，前者发现 ChatGPT 4.0 的表现良好[7]；后者发现 ChatGPT 4.0 在处理意识形态、复杂结构、文化负载词、隐转喻等方面仍存在准确性局限[8]。Lavie A. (2011)运用 BLEU、METEOR 和 TER 指标，评价了阿英、中英等多语言对的 GOOGLE 等机器翻译系统的文本质量。评价材料有新闻、博客、对话等[9]。

目前，学者倾向于采用四种及以上的评价指标进行实验研究。雷鹏飞，张浮凌(2024)用 BLEU、METEOR、ROUGE 和 NIST 评价了百度翻译、谷歌翻译等 6 款机器翻译软件的翻译质量，发现这些翻译软件均不能满足外宣翻译质量要求[4]。Shweta Chauhan 等人(2022)以 BLEU、METEOR、NIST、TER、ROUGE 和 STD 等指标衡量了无监督神经机器翻译模型的翻译质量，发现该模型在处理英语、印地语和印地语-kangri 语等语言时，BLEU 得分均有所提高[10]。Hui Yu 等人(2019)运用 DPF、组合指标，以及 TER、BLEU、METEOR 等评价指标，评价了多种语言对的机器译文质量，发现 DPF 在系统级和句子级均取得最佳结果[11]。Snover M 等人(2006)运用 TER、BLEU、METEOR 及其人类标注变体 HTER、HBLEU 和 HMETEOR，评价了 MTEval 2004 阿拉伯语评价数据集中 100 个句子的机器译文质量，发现 HTER 的人类判断的相关性最高[12]。Agarwal A 和 Lavie A. (2008)运用 Meteor、BLEU、TER、m-bleu 和 m-ter 这些评价指标，对 WMT-07 中英语、德语等的机器译文进行翻译质量评价，发现重新调整参数后的 Meteor 与人类排名的相关性显著提高[13]。

综上所述，翻译质量评价的文本有来自文化与文学翻译领域、政治与科技翻译领域以及多语言与资源稀缺语言翻译等领域，而标准文件的翻译质量对比研究，尤其是海事领域绿色仍为空白。

3. 研究设计

3.1. 研究材料

本文的研究材料为国际海事组织(IMO)的海上环境保护委员会(MEPC)发布的《MEPC 81-7-3-Glossary of climate change definitions in relation to shipping》。它由 59 个英文气候变化相关术语及其定义构成，属于海事绿色低碳术语定义的标准文件，共 4027 个英文字符。本研究选择术语作为翻译材料，主要是考虑了(1) 海事绿色低碳领域的术语专业性强，既包含行业特定的专业术语，又包含逻辑严谨的定义句式，这

使得评价指标得以兼顾术语准确性和定义完整性。(2) 翻译材料属于气候减排的标准和规范文件，具有单一性和标准性。采用标准体系的术语文本避免了目前通用型翻译材料非单一的弊端，可更加客观地评价专业领域的翻译质量。

3.2. 研究工具

本研究选取 BLEU、TER、METEOR 和 NIST 四种评价指标对五个 AI 翻译平台：Deepseek v3、ChatGPT 4.0、文心一言 4.0 以及 DeepL 和火山翻译进行评价。为控制变量，实验采用各平台 2024 年发布的公开版本。

翻译质量评价指标

本研究选取 BLEU、TER、METEOR 和 NIST 四种指标从不同的视角对翻译质量进行评价。

BLEU 是一种基于 n-gram 的机器翻译评价指标，通过比较机器译文与参考译文的 n-gram 匹配度来评价翻译质量。它主要关注机器译文词汇层面(n 通常取 1~4)和参考译文的表面匹配率，并引入简短惩罚因子(brevity penalty, BP)以防止机器译文因过短而评分虚高[14]。BLEU 值越接近 1 表示匹配度越高，翻译质量越好[15]。但是，BLEU 只关注词汇的匹配，忽略语义和行文逻辑，难以识别同义词替换，难以体现译文的流畅性和语义连贯性[6]。因此 BLEU 在应用中常需结合其他评价指标共同使用。

TER 是一种基于编辑操作距离的机器翻译评价指标，通过计算将机器译文转化到参考译文所需的最少编辑操作次数(包括替换、删除、插入和移动)来衡量翻译质量。操作次数越少，机器译文与参考译文越接近，TER 值就越低，表明翻译质量越高[7]。编辑操作总次数除以参考译文平均词数得到 TER 值。由于 TER 仅关注词汇的匹配，忽略了同义词替换，不注重语法的自然度，会导致语法正确但表达生硬的译文获得高分，无法衡量译文的流畅性。

METEOR 是一种基于词汇匹配与语义相似性的机器翻译评价指标。METEOR 利用 WordNet 等外部语言资源识别同义词并纳入匹配范围[16]，计算 unigram 的精确率和召回率衡量词汇匹配度，并引入与词序差异相关的惩罚因子以降低语序混乱获得的高分。最终得分由匹配度与惩罚因子的共同决定，可以更好地识别机器译文与参考译文之间的语义相似性，反映译文的语义准确性和流畅性，分数越接近 1 表示翻译质量越好。但 METEOR 完全依赖外部语言资源，而这些资源库涵盖的专业术语和低频词汇不足，导致其在处理专业术语和低频词汇时的能力有限[14]。

NIST 是一种基于 n-gram 匹配进行加权计算相似性的机器翻译评价指标。与 BLEU 不同的是，NIST 根据 n-grams 在参考译文中出现的频率或其所包含的信息量进行加权，给予稀有词汇或短语更高的权重，从而提升了翻译质量评价的精确性[17]。NIST 对稀有词汇的重视有助于提升对技术性或专业性文本的机器翻译评价质量。此外，BLEU 在处理短句时容易给出过高的评分，而 NIST 对长度惩罚因子(BP)进行了优化，使长句和短句都能得到合理的评分[14]。NIST 分数越高，表明翻译质量越好。但 NIST 同样存在局限性，由于计算方式相对复杂，NIST 的实施难度高于 BLEU，一定程度上导致其在实际应用中未广泛普及。

由于本文的研究材料属于标准体系文件，是 IMO 绿色航运标准体系的术语规范，后续出台的技术文件或政策文本的术语表述都必须基于该文件的术语定义标准；该文件对专业术语及其技术性定义翻译的准确性要求也极高，而 NIST 凭借低频术语加权机制，给稀有词汇更高的权重，能够关注译文在专业术语准确性上的表现。因此，本研究引入 NIST，结合 BLEU、TER、METEOR 三个主流的机器翻译评价指标[16]，可以全面地评价 AI 翻译平台的翻译能力。参考译文由一位专家翻译并校对审核，确保了参考译文的专业性和可信度。在实验开始前，所有平台均在相同时间内输入统一指令，避免外部干扰影响翻译结果。

3.3. 研究过程

3.3.1. 数据预处理

笔者首先对研究材料进行标准化文本提取，使用 Python 中的 PyPDF 库将原始 PDF 文件转换为 TXT 格式文本文件，随后对文本进行多轮人工清洗，删除多余的空格、非常规的换行符以及无关元素。最终提取出 59 项术语及定义的完整英文，储存于 Excel 表格第一列。

3.3.2. 译文准备

为了控制输入指令对翻译质量的影响，本实验对于 AI 翻译平台(文心一言 4.0、Deepseek v3、ChatGPT 4.0)，使用统一查询指令“Translate the following maritime climate terminology into Chinese”；对于 DeepL 和火山翻译则采用英文原文直接输入翻译的方式，得到译文结果；再准备一份由专家审核的人工译文作为参考译文。最后，将五个 AI 翻译平台输出的译文与参考译文在 Excel 中进行对齐处理。

3.3.3. 评价指标计算

本研究选用 Python 作为编程工具，下载安装 Pandas、Jieba 和 Nltk 库来完成评价指标分数的计算。先通过 Pandas 导入 Excel 数据，并对数据进行清洗与结构化处理，再利用 Jieba 完成中文文本分词，再进行人工校验，最后结合 NLTK 库中的各种函数进行各评价指标分数的计算，将各 AI 翻译平台中 59 个术语的不同评价指标得分导入 Excel 表格，得到评价指标数据。部分程序代码示例如下，表 1 为各 AI 翻译平台 BLEU 指标的句子得分部分数据。

```
“"bleu_scores = {
    '序号': list(range(1, len(df) + 1)), 'Deepseek v3 译文': [], 'ChatGPT 4.0 译文': [],....."
.....
"for _, row in df.iterrows():
    ref_tokens = chinese_tokenizer(row['参考译文'])
    reference = [ref_tokens]
    for system in ['Deepseek v3 译文', ...]:
        hyp_tokens = chinese_tokenizer(row[system])
        smooth = SmoothingFunction().method1
        score = sentence_bleu(reference, hyp_tokens, smoothing_function=smooth)
        bleu_scores[system].append(score)"
.....
"result_df.to_excel('bleu_scores_detail.xlsx', index=False,
                    sheet_name='BLEU 明细',
                    float_format="%.4f")
print("\n 明细数据已保存至 bleu_scores_detail.xlsx")"”
```

Table 1. Partial sentence-level BLEU score data of AI translation platforms
表 1. 各 AI 翻译平台 BLEU 指标的句子得分部分数据

序号	Deepseek v3	ChatGPT 4.0	文心一言 4.0	DeepL	火山翻译
1	0.2567	0.1314	0.3008	0.1559	0.2173
2	0.3327	0.3017	0.6306	0.1419	0.2608
3	0.4553	0.4570	0.4527	0.3630	0.4809

续表

4	0.6491	0.6539	0.7076	0.5735	0.5700
5	0.5973	0.584	0.5622	0.4782	0.4932
6	0.4893	0.4835	0.5434	0.3004	0.3500
7	0.5683	0.6725	0.3569	0.3779	0.3540
8	0.7576	0.876	0.8214	0.5197	0.6896
9	0.4247	0.4334	0.4647	0.2590	0.3201
10	0.4407	0.6001	0.5614	0.4861	0.3281

为了验证统计有效性，将 Excel 中的指标数据导入编写好的 Python 程序中得到各 AI 翻译平台各指标的平均数、方差、中位数、和标准差；再将 Excel 中的指标数据导入 SPSS 数据分析软件得到各个指标的 Friedman 检验分析结果秩均值、统计量、P 值和 Cohen’s f 值。所有统计结果通过 Excel 生成可视化表 2，最终形成翻译质量数据对比的条形(图 1)。

Table 2. Descriptive statistics of evaluation metrics by AI translation platforms
表 2. 各 AI 翻译平台的评价指标统计量数据

指标类型	AI 翻译平台	平均数	方差	中位数	标准差	秩均值	统计量	P 值	Cohen’s f 值
NIST	Deepseek v3	4.382	0.547	4.411	0.740	3.47	39.176	0.000	0.446
	ChatGPT 4.0	4.436	0.719	4.588	0.848	3.74			
	文心一言 4.0	4.201	0.858	4.378	0.926	2.92			
	DeepL	3.755	0.865	3.701	0.930	2.08			
	火山翻译	4.148	0.728	4.216	0.853	2.79			
BLEU	Deepseek v3	0.478	0.011	0.493	0.106	3.34	32.967	0.000	0.403
	ChatGPT 4.0	0.504	0.027	0.492	0.165	3.66			
	文心一言 4.0	0.463	0.022	0.465	0.147	3.15			
	DeepL	0.387	0.030	0.363	0.174	2.14			
	火山翻译	0.426	0.024	0.434	0.154	2.70			
METEOR	Deepseek v3	0.735	0.008	0.747	0.088	3.36	36.014	0.000	0.424
	ChatGPT 4.0	0.742	0.014	0.740	0.120	3.51			
	文心一言 4.0	0.731	0.014	0.733	0.120	3.46			
	DeepL	0.668	0.018	0.678	0.135	2.17			
	火山翻译	0.683	0.014	0.689	0.119	2.51			
TER	Deepseek v3	0.392	0.012	0.388	0.110	2.36	34.654	0.000	0.415
	ChatGPT 4.0	0.404	0.027	0.414	0.163	2.49			
	文心一言 4.0	0.499	0.076	0.469	0.275	3.25			
	DeepL	0.553	0.049	0.552	0.220	3.84			
	火山翻译	0.448	0.025	0.441	0.159	3.07			

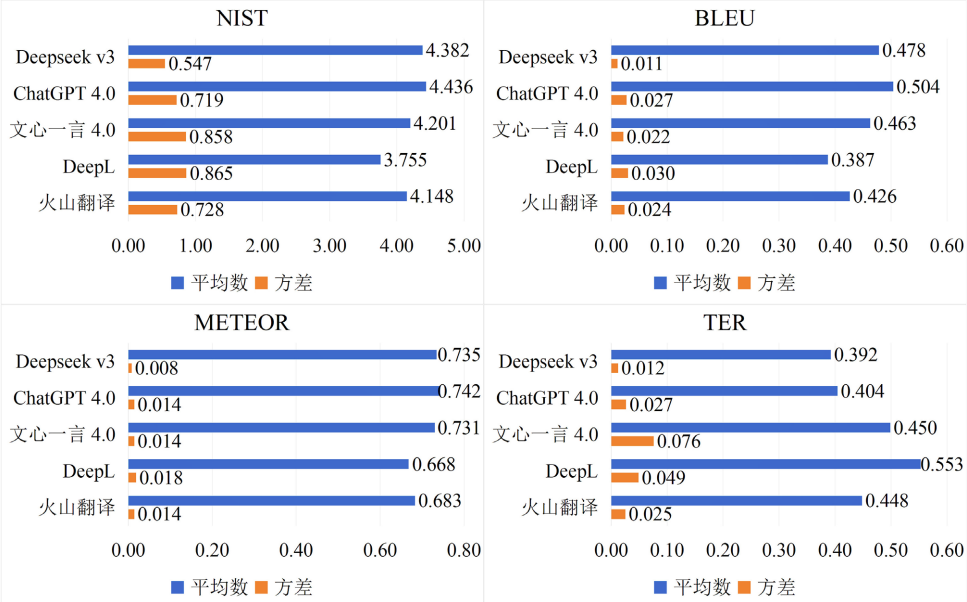

Figure 1. Translation quality data
图 1. 翻译质量数据

3.3.4. 数据有效性

为了验证实验数据的有效性，本研究通过 Friedman 非参数检验分析了五种 AI 翻译平台在以上四种指标上的得分差异。Friedman 检验显示，四种指标的 P 值均小于 0.001，这表明五个翻译系统的性能存在差异，且 Cohen’s f 值介于 0.40~0.45 之间，证明系统间存在显著差异；表 3 中的成对比较的调整后显著性检验也进一步支持了这一结论。例如，在 NIST 指标下，DeepL 译文与 Deepseek v3 译文和 ChatGPT 4.0 译文的 P 值均为 0.000，表明存在显著差异；而 DeepL 译文与火山翻译译文的 P 值为 0.157，大于 0.05，表明两者在 NIST 指标上无显著差异。此外，非参数检验的秩均值排序与描述性统计的平均数、中位数排序结果高度一致。这些结果表明，实验数据能够有效区分不同 AI 翻译平台的表现，可靠性较高，为后续分析奠定了基础。

Table 3. Paired comparison results with significance testing for each metric across AI translation platforms
表 3. 各 AI 翻译平台成对比较的各指标显著性数据

Sample 1-Sample 2	NIST 显著性	BLEU 显著性	METEOR 显著性	TER 显著性
DeepL-火山翻译	0.157	0.547	1.000	0.081
DeepL-文心一言 4.0	0.043	0.005	0.000	0.416
DeepL-Deepseek v3	0.000	0.000	0.000	0.000
DeepL-ChatGPT 4.0	0.000	0.000	0.000	0.000
火山翻译-文心一言 4.0	1.000	1.000	0.011	1.000
火山翻译-Deepseek v3	0.184	0.290	0.036	0.145
火山翻译译文-ChatGPT 4.0	0.011	0.010	0.006	0.478
文心一言 4.0-Deepseek v3	0.547	1.000	1.000	0.022
文心一言 4.0-ChatGPT 4.0	0.047	0.807	1.000	0.096
Deepseek v3-ChatGPT 4.0	1.000	1.000	1.000	1.000

注：每行都检验“样本 1 与样本 2 的分布相同”这一原假设。显示渐进显著性(双侧检验)。显著性水平为 0.050。

4. 研究结果与讨论

本文将五个 AI 翻译平台输出的译文分别和参考译文进行对比，并计算了它们的 NIST、BLEU、METEOR 和 TER 分数，各个指标的得分情况如上图 1 所示。

4.1. 准确性

准确性是衡量译文质量的重要标准之一。BLEU 通过 n-gram 的匹配程度衡量词汇的忠实度；NIST 指标给出现频率较低的词汇更高的权重，能反映专业术语的翻译准确性。TER 通过计算将机器译文转化到参考译文所需的最少编辑操作次数来评价机器译文中词汇和语法的准确性。因此，本研究将通过 BLEU、NIST 和 TER 三个指标衡量翻译系统的准确性。

综合图 1 来看，ChatGPT 4.0 译文在 NIST (均值 4.436)和 BLEU 指标(均值 0.504)上表现最好，而在 TER 指标(均值 0.404)的表现上仅次于 DeepSeek v3；Deepseek v3 在 TER 指标(均值 0.392)上表现最优且在 NIST 和 BLEU 指标上排名也靠前。文心一言 4.0 和火山翻译在 NIST、BLEU 和 TER 指标上的得分表明其翻译质量均落后于前两者； DeepL 生成的译文在三个指标上的得分均最低，译文准确性较差。

例如：原文“Carbon dioxide (CO_2) is a naturally occurring gas and is also a by-product of burning fossil fuels (such as oil, gas, and coal), of burning biomass, of land use changes (LUC) and of industrial processes (eg, cement production). It is the principal greenhouse gas (GHG) produced by, or resulting from, human activities that affects the earth’s radiative balance. It is the reference gas against which other GHGs are measured and therefore has a global warming potential (GWP) of 1. (IPCC, 2021)”

专业术语方面，ChatGPT 4.0 和 Deepseek v3 均准确地翻译了专业术语。例如，“land use changes (LUC)”译为“土地利用变化”，“global warming potential (GWP)”均准确译为“全球变暖潜力”。而其他 AI 翻译平台在术语翻译方面存在问题，例如火山翻译将“IPCC”误译为“气专委”，DeepL 在“industrial processes”部分出现重复翻译，这些错误直接影响了海事气候变化术语翻译的准确性，术语体系的严谨性、一致性与权威性。

句子结构方面，原文“It is the principal greenhouse gas (GHG) produced by, or resulting from, human activities…”这一复杂句式，ChatGPT 4.0 译为“它是人类活动产生的或由此产生的主要温室气体”，Deepseek v3 译为“它是主要的人类活动产生或导致的温室气体”，均准确传达了原文的含义及修饰关系。而 DeepL 的部分译文语序混乱(如“工业过程(如水泥生产)”的重复)，文心一言 4.0 的译文略显生硬(如“它是衡量其他温室气体的参考气体”)，两者在传达原文含义时均有欠缺，可能会导致后续各国家、组织及利益相关方基于该术语文件而出台技术文件或政策文本时，存在概念上的歧义，引发不必要的争议。

综上所述，在准确性方面，ChatGPT 4.0 和 DeepSeek v3 在翻译海事气候变化术语及其定义时的整体表现最好，DeepL 和火山翻译相对较差，尤其在翻译复杂术语时存在不足。ChatGPT 4.0 与 DeepSeek v3 在专业术语及复杂句式结构上的准确处理，为国际海事组织(IMO)气候术语框架提供了可靠的语言转换基础。

4.2. 流畅性

流畅性能反映出译文的自然度和语义连贯性，是衡量机器翻译质量的重要手段之一。本研究将通过 METEOR 指标评价译文的流畅性。METEOR 指标不仅关注词汇的匹配度，还考虑了语法准确性和语义一致性。METEOR 对同义词替换的关注度使其在评价译文流畅性方面更具优势。实证研究表明，在句子层面，METEOR 比 BLEU 更具有人工评价的相关性[9]。王坤宇和成思(2025)的研究指出，BLEU 指标存在

局限性，它关注词汇的匹配，忽略语义及上下文逻辑，难以衡量译文的流畅性和语义连贯性[6]。这说明METEOR 更能反映出译文的可读性。而文旭和田亚灵(2024)在进行 ChatGPT 4.0 对于中国特色话语翻译的有效性研究时，使用 BLEU 指标评价机器译文的流畅性显得不够专业[12]。

在表 2 中，METEOR 得分显示，ChatGPT 4.0 (均值 0.742)和 DeepSeek v3 (均值 0.735)在五个 AI 翻译平台中表现最佳，且两者之间无显著差异，说明其译文在语言自然度和语义连贯性上均接近参考译文。而 DeepL (均值 0.668)和火山翻译(均值 0.683)得分较低，表明两者存在译文表达不自然的问题。例 1，原文“Net zero carbon ship operations describes when the carbon dioxide (CO_2) emissions resulting from the operation of the ship are balanced by removals resulting from human activities over a specified period. Net zero carbon ship operations means reducing emissions and balancing the remaining residual emissions through removal rather than using offsets to other sectors. The emissions removal can be achieved during fuel production and/or after combustion.”。

在术语一致性方面，对于核心术语“Net zero carbon ship operations”，ChatGPT 4.0、DeepSeek v3 和 DeepL 都准确译为“净零碳船舶运营”，与参考译文一致；文心一言将其译为“净零碳排放船舶运营”，增译“排放”稍显冗余；火山翻译译为“船舶净零碳作业”，“作业”通常指具体的操作活动，不如“运营”涵盖范围广。文心一言 4.0 将“offsets”译为“转移”，与“抵消”概念不符，其他平台皆准确译为“抵消”。

在语义连贯性方面，对于核心结构：“describes when the carbon dioxide CO_2 emissions… are balanced by removals…”，ChatGPT 用“描述了……被……所平衡”，语序自然流畅，句子逻辑清晰，Deepseek v3 用“排放通过……移除达到平衡”，两者的句式避免了冗余表述，动词搭配既忠实于原文，连贯性又强，且符合中文表达习惯。而对比其他译文：DeepL 译文“通过清除而不是使用其他部门的抵消来平衡”，信息量大，略显紧凑，译文不够自然；火山翻译将“operations”译为“作业”，与其后文“运营”不一致，导致术语混淆，影响逻辑流畅性。

在表达自然度方面，中文倾向于简洁、避免重复，ChatGPT 4.0 和 Deepseek v3 的译文在措辞上更贴合这一特点，避免了不必要的冗余短语(如“的状态”、“的方式”)，动词(平衡、减少、移除/去除)和连接词(被……所/通过……)的使用使行文更连贯、一气呵成，词语之间的搭配更符合中文习惯(如“使用……措施”比“利用……抵消”或“进行……抵消”更地道)。而对比其他译文：文心一言“向其他部门转移碳排放的方式”中，“转移”与原文“offsets”(抵消)语义偏差，影响理解流畅性，“相平衡的状态”中“的状态”稍显累赘；火山翻译“利用对其他部门的抵消来减少排放和平衡剩余的剩余排放”中，“剩余的剩余排放”重复累赘，且“利用……抵消”表达生硬；DeepL“使用其他部门的抵消”省略了动作指向，语义不完整。

综上所述，通过 METEOR 指标验证，ChatGPT 4.0 和 DeepSeek v3 在专业术语翻译的自然度、语义连贯性和句式流畅性方面明显优于其他 AI 翻译平台，它们不仅能够将海事术语文本中的专业术语翻译准确，语义理解能力也更强，能准确传达原文信息，根据目的语习惯进行调整，使译文更加通顺自然。

4.3. 稳定性

在本文的翻译质量评价中，稳定性是指 AI 翻译平台在翻译多个海事气候术语及其定义时译文质量的一致性水平。稳定性通过五个 AI 翻译平台的各评价指标的方差和标准差进行衡量，数值越小，说明该平台的翻译质量波动越小、表现越稳定。

从方差和标准差进行分析发现，Deepseek v3 在四个指标上均保持最低水平，说明其翻译质量波动最小，在翻译多个海事气候术语及其定义时的稳定性较高；ChatGPT 4.0 仅在 BLEU 指标上的数值排名靠后

而在 NIST、METEOR 和 TER 三项指标上的数值相对较小，位列第二，因此其翻译质量的稳定性也较高；火山翻译四个指标的数值均位于第三位，稳定性表现中等；文心一言 4.0 在 NIST 和 TER 指标上的数值较高，稳定性较差，但在 METEOR 和 BLEU 指标上的数值排名第二，说明其在译文流畅性和词汇匹配度上相对稳定；DeepL 在四个指标上的数值均较高，说明其翻译质量波动最大，表现不稳定。

综上所述，通过对五个主流 AI 翻译平台在 NIST、BLEU、METEOR 和 TER 四项评价指标下的表现进行系统分析，可以发现 ChatGPT 4.0 与 DeepSeek v3 在翻译海事绿色低碳术语定义的标准文件时，准确性、流畅性和稳定性的表现都更好，这两个平台均表现出较高的翻译质量，尤其在翻译专业性强、句子结构复杂的气候变化相关术语及其定义时，其词汇匹配度、语义连贯性均优于其他翻译平台。而 DeepL 和火山翻译在多个指标中得分较低，表现出在术语准确性和语言自然度方面的不足，且整体稳定性相对较弱；文心一言 4.0 虽在部分指标中表现中等偏上，但与 ChatGPT 4.0 和 DeepSeek v3 相比仍存在一定差距，特别是在术语一致性和句法结构还原方面仍有提升空间。因此，ChatGPT 4.0 和 DeepSeek v3 更适合作为国际海事组织发布的专业气候术语翻译任务的机器翻译平台。

5. 结语

本研究围绕五个 AI 翻译平台在海事标准体系术语定义翻译中的表现进行对比分析。研究发现，在准确性方面，ChatGPT 4.0 与 DeepSeek v3 的译文与参考译文的术语一致性最高，能够更准确地传达原文的含义；在流畅性方面，ChatGPT 4.0 和 DeepSeek v3 在译文自然度和语义连贯性方面表现最优，其译文更加通顺自然，句子逻辑清晰，符合海事标准体系的术语翻译的表达规范；在稳定性方面，ChatGPT 4.0 和 DeepSeek v3 的翻译质量波动最小，稳定性较高。因此，ChatGPT 4.0 和 DeepSeek v3 更适配海事标准体系的术语翻译需求。

AI 正从大模型转化为专业能力突出的平台。同样，翻译平台应在模型规模、领域数据之外加强基于人类反馈的强化学习(RLHF) [18]。如果 RLHF 能够基于标准文件的训练，那么其准确性、流畅性和稳定性可以显著提高。目前，AI 大语言平台的翻译训练集中在规模和参数，而对基于标准文件的训练较少。可以推测，随着规模的进一步扩大，AI 大语言平台的下一个发展将集中在专业领域的训练。

本研究能够帮助从业者根据准确性和流畅性等具体需求选择适合的 AI 翻译平台。再者，BLEU、METEOR、TER 和 NIST 四种指标可从不同的角度评价翻译质量，特别是 METEOR 表现出来的流畅性和 NIST 代表的术语准确性，可作为专业领域翻译质量评价的基本要素，有助于提升翻译质量及行业标准化水平。

参考文献

[1] 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.

[2] 段田园. 人工智能时代机器翻译汉译英质量评测[J]. 数字技术与应用, 2025, 43(5): 9-11.

[3] 郭望皓, 胡富茂. 神经机器翻译译文评测及译后编辑研究[J]. 北京第二外国语学院学报, 2021, 43(5): 66-82.

[4] 雷鹏飞, 张浮凌. 基于机器翻译软件的外宣文本翻译质量评估研究[J]. 未来与发展, 2024, 48(6): 41-48.

[5] 范新瑜. 海洋科技文本机器翻译译文质量评测[J]. 武汉冶金管理干部学院学报, 2023, 33(4): 31-34.

[6] 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, 42(1): 86-95.

[7] 王子云, 毛毳. ChatGPT 译文质量的评估与提升——以陶瓷类文本汉英翻译为例[J]. 山东陶瓷, 2023, 46(4): 20-27.

[8] 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.

[9] Lavie, A. (2011) Evaluating the Output of Machine Translation Systems. *Proceedings of Machine Translation Summit*

XIII: *Tutorial Abstracts*, Xiamen, 19 September 2011, 4, 15.

[10] Chauhan, S., Saxena, S. and Daniel, P. (2022) Improved Unsupervised Neural Machine Translation with Semantically Weighted Back Translation for Morphologically Rich and Low Resource Languages. *Neural Processing Letters*, **54**, 1707-1726. https://doi.org/10.1007/s11063-021-10702-8

[11] Yu, H., Xu, W., Lin, S. and Liu, Q. (2019) Machine Translation Evaluation Metric Based on Dependency Parsing Model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, **18**, 1-15. https://doi.org/10.1145/3312573

[12] Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*: *Technical Papers*, Cambridge, 8-12 August 2006, 223-231.

[13] Agarwal, A. and Lavie, A. (2008) METEOR, M-Bleu and M-Ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. *Proceedings of the 3rd Workshop on Statistical Machine Translation*, Columbus, 19 June 2008, 115-118. https://doi.org/10.3115/1626394.1626406

[14] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., *et al*. (2023) A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, **11**, Article No. 1006. https://doi.org/10.3390/math11041006

[15] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2002, 311-318. https://doi.org/10.3115/1073083.1073135

[16] 王均松, 庄淙茜, 魏勇鹏. 机器翻译质量评估: 方法、应用及展望[J]. 外国语文, 2024, 40(3): 135-144.

[17] Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. *Proceedings of the 2nd International Conference on Human Language Technology Research*, San Diego, 24-27 March 2002, 138-145. https://doi.org/10.3115/1289189.1289273

[18] 冯志伟. 大语言模型时代的术语翻译[J]. 中国科技术语, 2024, 26(3): 93-96.