Modern Linguistics 现代语言学, 2026, 14(1), 140-146
Published Online January 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.141020

Hans汉斯

海事文本的生成式AI翻译质量对比研究

黄圣珠

上海海事大学外国语学院，上海

收稿日期：2025年12月1日；录用日期：2025年12月26日；发布日期：2026年1月6日

摘　要

本研究针对生成式AI在海事领域的翻译质量展开系统评估。选取DeepSeek、豆包、ChatGPT和Gemini四款主流模型，以海事专业文本为基础，采用BLEU指标对英中/中英双向翻译质量进行量化分析。研究发现：四个模型在海事文本汉译英任务中表现均优于英译汉，Gemini综合表现最优且稳定，DeepSeek在技术文本中可靠，ChatGPT输出波动较大，豆包更适用于通用文本。研究成果为海事相关文本的翻译选择AI工具提供了实证依据。

关键词

生成式人工智能，机器翻译，译文质量评估，对比研究，海事文本

A Comparative Study on the Translation Quality of Generative AI on Maritime Texts

Shengzhu Huang

College of Foreign Languages, Shanghai Maritime University, Shanghai

Received: December 1, 2025; accepted: December 26, 2025; published: January 6, 2026

Abstract

The study systematically evaluates the translation quality of generative AI models in the maritime domain. We selected four mainstream models—DeepSeek, Doubao, ChatGPT, and Gemini—and used specialized maritime texts as the basis for quantitative analysis of English-to-Chinese (E-C) and Chinese-to-English (C-E) bidirectional translation quality using the BLEU metric. The findings indicate that all four models perform better in the maritime C-E translation task than in the E-C task. Gemini demonstrates the best overall and most stable performance, DeepSeek proves reliable for technical documents, ChatGPT's output shows significant variability, and Doubao is more suitable for general texts. The research results provide empirical evidence for selecting AI tools for translating maritime-

文章引用: 黄圣珠. 海事文本的生成式 AI 翻译质量对比研究[J]. 现代语言学, 2026, 14(1): 140-146.
DOI: 10.12677/ml.2026.141020

related texts.

Keywords

Generative AI, Machine Translation, Translation Quality Assessment, Comparative Study, Maritime Texts

1. 引言

近年来，生成式人工智能(Generative AI)浪潮席卷全球，其核心代表大语言模型(Large Language Models, LLMs)在自然语言处理领域展现出强大潜力。机器翻译技术历经从早期规则与统计方法到当前大规模预训练的演进，如今，以大语言模型为核心的生成式 AI 凭借其在跨语言知识与深度语境理解上的突破，正深刻重塑着翻译行业(王华树、刘世界，2025) [1]。

然而，现有翻译评估研究多集中于医学、法律与文学等领域，如钱晓彤、张弓(2025)等人对文学文本的翻译质量进行了系统评估[2]，任禹昕等(2025)则针对医学文本展开了深入分析[3]，针对海事文本的翻译质量评估则尚显匮乏。这一研究空白与海事文本自身的语言特点密切相关，如古体词语、使用间接指称(隋桂岚、张毅，2006) [4]、缩略词、被动语态(汪云婷、禹一奇，2015) [5]等，对翻译的准确性、专业性和一致性有着极高要求。尽管生成式 AI 潜力巨大，但学术界对主流模型在海事这一垂直领域的翻译性能仍缺乏系统性的对比评估，其优劣、特点及共性缺陷尚不明确。

为此，本研究选取 DeepSeek、豆包、ChatGPT 及 Gemini 四款代表性生成式 AI 模型，通过严谨的对比实验，系统评估其在海事文本英汉互译中的综合表现，旨在揭示各模型在该专业场景下的能力差异与特点，为海事领域的专业译员、相关企业与研究机构在选择与应用 AI 翻译工具时，提供关键的数据参考与实践指导。

2. 文献综述

翻译质量评估(Translation Quality Assessment, TQA)长期居于翻译学研究的核心。据王均松等人(2024)梳理，在人工翻译领域，国外形成了诸如文本类型评估原则、豪斯翻译质量评估模式等具有广泛影响力的成果；国内则以司显柱的翻译质量评估模式等为代表[6]。在机器翻译领域，评估方法主要分为人工与自动评估两类。人工评估方面，MQM 多维质量指标体系是目前应用最为广泛的错误扣分框架(田朋，2020)，该体系从准确性、流利度等八个维度对译文错误进行归类[7]。自动评估方面，主流指标包括 BLEU、TER、WER、METEOR 等，其中以 BLEU (Bilingual Evaluation Understudy)的应用最为普及。该指标由 IBM 团队于 2002 年提出，其核心思想是通过 N-gram 匹配度并引入惩罚因子，计算机器译文与参考译文之间的相似度，相似度越高则代表译文质量越佳(Papineni *et al*., 2002) [8]。综上所述，TQA 领域已发展出从人工到自动、从理论模式到操作指标的多元方法体系。这些成熟的评估范式为各类翻译实践与研究提供了关键工具；然而，如何将其有效应用于海事专业文本等垂直领域，以检验与引导新兴生成式 AI 的翻译质量，仍是当前值得深入探索的方向。

梁硕、张鹏蓉(2024)指出，当前海事翻译研究的主流仍聚焦于翻译策略、方法及技巧等传统议题[9]。

具体而言，李诚(2024)对海事英语被动句的翻译策略进行了专项研究[10]；王甜甜与程昕(2018)关注于海事英语中字母“A”的翻译方法[11]；吴海宁(2013)则对国际海事公约的翻译技巧进行了分析[12]。相较之下，海事文本的机器翻译研究则尚处起步阶段，成果有限。周忠良(2024)初步阐述了生成式人工智能在涉海翻译中于提升准确性及效率方面的潜力[13]；刘世界(2024)则对 DeepL 等主流翻译引擎的翻译质量开展了定性与定量评估[14]。综上所述，现有研究清晰地表明：尽管传统海事翻译研究已积累了丰富的策略与技巧，但面向机器翻译，尤其是生成式 AI 在此垂直领域的性能评估与应用的系统性研究仍存在明显不足，亟待深入探索。

3. 研究设计

3.1. 研究问题

为系统评估所选生成式 AI 模型在海事文本中的翻译性能，本研究围绕以下两个核心问题展开：

1、在英译中和中译英两个方向上，不同的生成式 AI 模型翻译海事文本的 BLEU 得分是否存在显著差异？

2、哪种类型的生成式 AI 模型在海事文本翻译中表现更优？

3.2. 研究方法

本研究采用的实验流程为：(1) 构建海事领域双语数据集；(2) 使用所选主流生成式 AI 模型进行双向翻译任务；(3) BLEU 分数计算；(4) 结合人工审校对典型错误进行归类分析；(5) 结果分析与对比。

3.3. 数据集

本数据集的源文本选自《国际防止船舶造成污染公约》(MARPOL)与《国际海上人命安全公约》(SOLAS)等国际海事核心规范文件，这些文本具有高度的专业性和规范性。所使用的参考译文均源自官方或权威海事组织发布的译本，并经过海事领域专业人士的校验与确认，从而确保其在专业术语、句式结构与行业规范上的高度准确性，为后续的评估提供了可靠依据。

数据集由中英和英中两个双向子集构成，每个子集各包含 50 个具有代表性的典型例句。这种平衡的设计旨在全面考察生成式 AI 模型在理解源语言和生成目标语言两个方向上的综合翻译能力。为确保例句能全面反映海事文本的语言特征，依据以下三个维度进行筛选与构建：(1) 句法复杂度：覆盖简单句、复合句及长难句(如含多重修饰或被动语态嵌套的句式)；(2) 文本功能：涵盖公约中定义条款、操作要求、禁止性规定、程序描述及事实陈述等主要功能类型；(3) 句子长度：按英文单词数将句子分为短句(<20 词)、中句(20~40 词)与长句(>40 词)三类，并在数据集中保持三类数量比例均衡。

此外，研究前已对所有文本进行了统一的预处理工作，包括清除无关格式标记、统一中英文标点符号、规范专业术语的书面表达形式，并确保每个句子都是独立完整的语义单位，旨在消除噪声干扰，为后续的公平评估奠定基础。

3.4. 机器翻译选择

本研究选取了四款具有代表性的生成式 AI 模型进行翻译质量评估，包括 DeepSeek-V3、豆包、GPT-5 以及 Gemini 3 Pro。

这四种模型在市场上具有高度代表性，涵盖了国内外主流技术阵营，DeepSeek 具备独特推理架构，豆包专注于翻译垂直领域，ChatGPT 擅长对话理解，而 Gemini 在复杂推理上具有领先优势。通过对比这些各具特色的模型，可以更全面地评估生成式 AI 翻译海事相关文本的能力。

3.5. BLEU 评估指标

本研究采用 BLEU 作为机器翻译质量的主要自动评估指标。BLEU 的基本原理是通过计算机器翻译输出与专业人工参考译文之间的 n-gram 匹配程度来衡量翻译质量，综合考虑了从 1-gram 到 4-gram 的精确度加权几何平均值，同时对过短译文施加长度惩罚机制，从而在词汇和短语层面上对翻译结果的准确性进行量化评估。

在具体实施过程中，我们选用目前学术界广泛认可的 Sacre BLEU 计算工具，在 Google Colab 平台上完成所有 BLEU 分数的计算。中文文本在计算 BLEU 之前使用 jieba 分词，以满足 BLEU 的 token 对齐要求。同时，为了衡量 BLEU 的稳定性，采用 bootstrap 采样方法对句子级译文进行 2000 次重采样，计算 BLEU 的均值、方差和 95%置信区间。

4. 研究结果

本研究对四种生成式 AI 模型在海事文本翻译中的表现进行了系统评估，得到了各模型在双向翻译任务中的 BLEU 得分情况，见表 1 和图 1。

Table 1. Scores of translation systems on the BLEU metric
表 1. 各翻译系统在 BLEU 指标上的得分情况

翻译方向	指标	DeepSeek	豆包	ChatGPT	Gemini
英译汉	BLEU	34.332	27.786	22.319	37.558
	方差	4.718	4.102	5.176	4.709
	标准差	2.172	2.025	2.275	2.170
汉译英	BLEU	61.496	42.269	54.983	67.278
	方差	5.150	6.319	6.904	4.189
	标准差	2.692	2.514	2.627	2.047

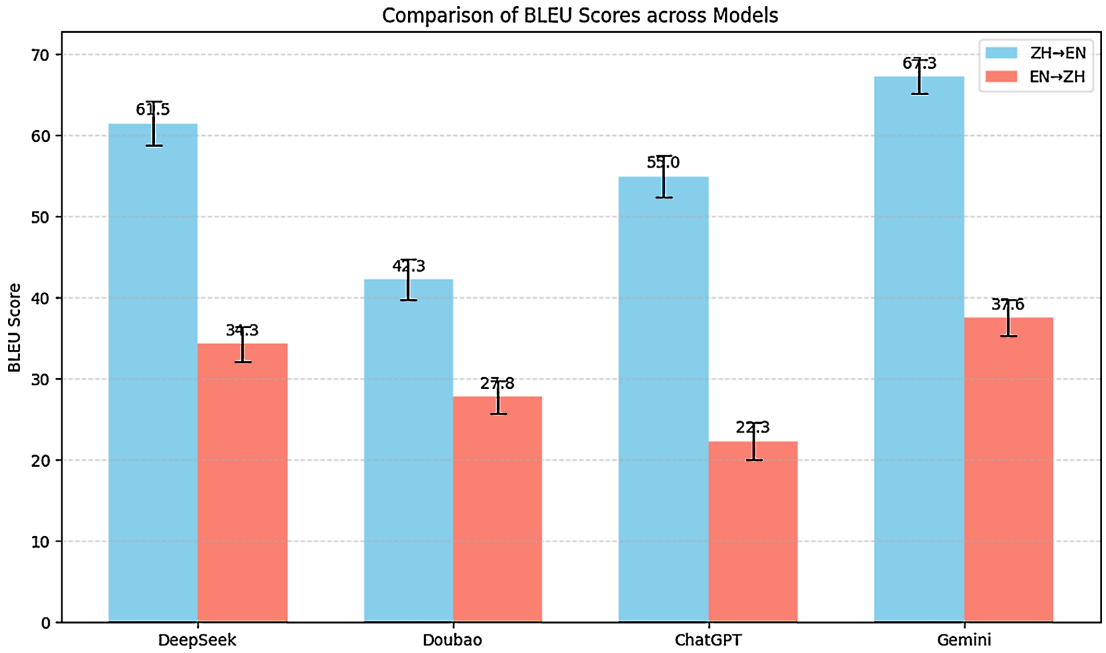

Figure 1. Comparison of BLEU scores across models
图 1. 各翻译系统在 BLEU 指标上的得分情况

4.1. BLEU 分数总体表现

从总体 BLEU 得分来看，四种模型在两个翻译方向上均表现出明显差异。Gemini 和 DeepSeek 在英译汉与汉译英任务中均取得领先表现，其中 Gemini 的 BLEU 得分最高，说明其在专业海事文本上的语义理解与生成能力最为突出。DeepSeek 得分紧随其后，整体准确性较高，在技术类文本翻译中表现出良好的稳健性。相比之下，豆包和 ChatGPT 的得分均低于前两者，其中 ChatGPT 在英译汉任务中的得分最低，说明其在处理复杂海事英语结构时存在一定困难。

4.2. 稳定性分析

为检验模型翻译质量的稳定性，本研究采用 bootstrap 重采样方法计算 BLEU 的均值、方差和标准差。

在英译汉任务中，四个模型的方差均在 4~5 区间，标准差均处于 2.0~2.3 之间，整体表明四个系统的英译汉输出较为稳定。值得注意的是，ChatGPT 虽然 BLEU 得分较低，但其方差(5.176)和标准差(2.275)均为最高，说明其输出波动较大，在不同语段之间表现不够一致。相较之下，DeepSeek 与 Gemini 均表现出“高分且稳定”的特征，不仅 BLEU 得分领先，方差和标准差也较低。

在汉译英任务中，Gemini 的方差(4.189)和标准差(2.047)最低，显示其不但得分最高，输出也最为一致。豆包与 ChatGPT 的方差和标准差明显偏高，反映了其在不同句段之间的性能波动较大。DeepSeek 的 BLEU 处于高位，但稳定性略低于 Gemini，不过仍处于较理想区间。

总体而言，Gemini 是四个系统中稳定性最好的模型，DeepSeek 次之。

4.3. 翻译方向差异的影响

从双向得分趋势来看，所有模型在汉译英任务中的 BLEU 均明显高于英译汉。这一方向性差异可能由以下因素共同造成：(1) 语言结构复杂度不同。海事英语原文句式长且嵌套多，逻辑层次复杂，增加了英译汉的难度；相对而言，中文的专业文本更线性，便于模型生成符合语法的英文表达；(2) 中文句法的隐性逻辑关系。中文高度依赖语境，隐含关系丰富，进行汉译英时需要模型进行较多显性逻辑补全与语法重构，难度大于英译汉；(3) 训练语料的方向偏置。多数大模型在英译中任务上的训练数据覆盖度更高，使其在该方向上表现更优，从而导致方向性差异。

这一结果也印证了技术文本翻译中较为普遍的现象：英译汉通常比汉译英更具挑战性。

4.4. 错误分析

为进一步验证 BLEU 得分反映的问题，本研究对低分样本进行了人工审校，归纳出以下几类常见错误模式。

(1) 专业术语翻译不准确

术语误译是导致 BLEU 评分下降的主要因素，在科技、海事及法律类文本中尤为突出。例如，在翻译句子“Ships’ routeing systems contribute to safety of life at sea, safety and efficiency of navigation and/or protection of the marine environment”时，其中“Ships’ routeing systems”应为“船舶定线制”。尽管 DeepSeek 与 Gemini 均正确译出，豆包将其译为“船舶航线系统”，ChatGPT 则译为“船舶航行路线系统”，均未能准确传达该海事术语的标准译法。

(2) 语序调整能力不足

部分模型在处理含有嵌套结构或复杂逻辑关系的英文长句时，未能有效重构汉语语序，影响了译文的通顺度与语义准确性。例如，在翻译“Ships’ routeing systems are recommended for use by, and may be made mandatory for, all ships...”一句时，DeepSeek、豆包与 ChatGPT 均将句末条件状语“with the guidelines

and criteria developed by the Organization”前置至句首，译为“经本组织制定的指南和准则采纳与实施后……”，造成主句信息后置，句子结构失衡，也模糊了条件状语与核心动词之间的逻辑关联。这反映出当前生成式 AI 对海事文本中特有的复杂句式在句法解析与语序重构方面仍存在局限。

(3) AI 生成幻觉问题

生成式 AI 在翻译过程中容易出现虚构或过度生成内容的现象，即“幻觉”。这种现象在海事文本的翻译中尤为突出。以“距最近陆地一词，系指……”的翻译为例，豆包在译文中自行添加了原文中并不存在的详细地理坐标序列，如“in latitude 11˚00'S, longitude 142˚08'E to a point in latitude 10˚35'S, longitude 141˚55'E…”，并重复使用“thence to a point…”结构，导致译文严重偏离原意，信息冗余且逻辑混乱。该案例说明，部分模型在处理专业规范性文本时，对内容生成边界的控制能力仍然不足，易产生无关或虚构信息，影响译文的严谨性与可靠性。

上述几类错误模式在不同模型中的出现频率与严重程度存在差异，这也从一个侧面解释了各模型 BLEU 得分方差较大的现象。

综合 BLEU、方差与标准差三项指标，从翻译方向来看，所有模型在汉译英任务中的表现均优于英译汉，进一步说明海事领域的英译汉仍是当前生成式 AI 模型的难点。总体而言，Gemini 在两个方向均取得最佳成绩，是整体性能最强、最稳定的翻译系统。DeepSeek 次之，尤其在技术类文本中表现稳定可靠。ChatGPT 中译英能力明显优于英译汉，但输出波动较大。豆包整体处于中等偏低水平，适用于一般文本但不适用于高专业领域。从方向上看，所有模型在中译英任务中表现均明显优于英译汉，说明海事领域的英译汉翻译复杂度更高。

5. 结论

本研究基于 BLEU 指标系统比较了四种生成式 AI 模型在海事文本双向翻译中的表现。结果表明，Gemini 和 DeepSeek 整体准确性与稳定性最佳，而豆包和 ChatGPT 波动较大，适用性有限。各模型在汉译英方向的表现均优于英译汉，反映出海事英语结构复杂度更高、英译汉仍是当前模型的主要挑战。研究结果可为海事文本大模型翻译优化和专业机器翻译系统的构建提供参考，帮助专业译员和航运企业基于实证数据选择适合的 AI 翻译工具，避免试错成本，提升翻译效率和质量。

本研究尚存一些局限性，主要包括评估指标主要依赖 BLEU、数据集规模有限，且未探索模型混合使用策略。后续研究可引入专业译员进行人工评估，结合更丰富的自动化指标，扩大数据集规模和文本类型多样性，并探索针对海事领域的生成式 AI 模型微调，以获取更全面、深入的研究结论。

参考文献

[1] 王华树, 刘世界. 从 MTPE 到 AIPE: GenAI 时代翻译模式演变及其对翻译教育的启示[J]. 山东外语教学, 2025, 46(3): 111-121.

[2] 钱晓彤, 张弓. 国产 AI 大模型应用于文学翻译的有效性评估——基于 DeepSeek、豆包和文心一言的译文对比研究[J]. 科技传播, 2025, 17(16): 26-32.

[3] 任禹昕, 陈子杰, 林咏臻, 等. 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/OL]. 中医药导报, 1-10. https://doi.org/10.13862/j.cn43-1446/r.20250822.001, 2025-08-22.

[4] 隋桂岚, 张毅. 海事法律英语的文体特征及其翻译策略[J]. 中国翻译, 2006, 27(6): 63-67.

[5] 汪云婷, 禹一奇. 海事英语的语言特点与翻译策略[J]. 安徽文学(下半月), 2015(8): 115-116.

[6] 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.

[7] 田朋. 翻译多维质量标准 MQM 模型介评与启示[J]. 东方翻译, 2020(3): 23-30.

[8] Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2001) BLEU: A Method for Automatic Evaluation of Machine

Translation. *Proceedings of the* 40*th Annual Meeting on Association for Computational Linguistics*, Philadelphia, 7-12 July 2001, 311-318. https://doi.org/10.3115/1073083.1073135

[9] 梁硕, 张鹏蓉. 海事术语翻译研究热点与趋势(1996-2023)——基于 CiteSpace 的文献计量分析[J]. 上海理工大学学报(社会科学版), 2024, 46(5): 395-402.

[10] 李诚. 海事英语中被动句的翻译对策研究[J]. 公关世界, 2024(6): 44-46.

[11] 王甜甜, 程昕. 海事英语中字母“A”的翻译策略[J]. 中国科技翻译, 2018, 31(3): 6-8.

[12] 吴海宁. 基于目的论分析国际海事公约的翻译技巧[J]. 中国海事, 2013(3): 62-64.

[13] 周忠良. 基于生成式人工智能的涉海翻译: 优势、挑战与前景[J]. 中国海洋大学学报(社会科学版), 2024(2): 12-20.

[14] 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(2): 21-31.