Modern Linguistics 现代语言学, 2026, 14(6), 177-187
Published Online June 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.146513

Hans汉斯

基于《生死疲劳》语料库的大语言模型文学翻译质量评估

吴佳佳，庞宝坤*

哈尔滨理工大学外国语学院，黑龙江 哈尔滨

收稿日期：2026年4月27日；录用日期：2026年5月26日；发布日期：2026年6月8日

摘　要

中华优秀传统文化“走出去”战略的深入实施，对翻译产出的质量标准提出了更高要求。与此同时，数智技术的快速发展为大语言模型(LLMs)介入文学翻译提供了技术支撑。基于此，本研究以莫言《生死疲劳》为例，构建“1对3”汉英平行语料库(源文本、葛浩文译本以及DeepSeek-V3.2与文心一言4.5译本)。研究运用BERTScore定量评估译文质量，并进一步引入豪斯(House)翻译质量评估模型从语场、语旨、语式三维度对大语言模型的翻译表现展开定性分析。研究结果表明，大语言模型译文的整体语义保真度接近人工译本，但是在低分例句中仍普遍存在文化信息缺失、情感表达弱化、文本风格趋同等共性问题。研究为数字化语境下中华文化的国际传播提供了实证支撑，并指出未来可以通过构建结构化提示词(Structured Prompting)等策略进一步优化大语言模型的翻译质量。

关键词

《生死疲劳》英译，翻译质量评估，大语言模型，BERTScore，豪斯模型

Translation Quality Assessment of Literary Translation by Large Language Models Based on the Corpus of *Life and Death Are Wearing Me Out*

Jiajia Wu, Baokun Pang*

School of Foreign Languages, Harbin University of Science and Technology, Harbin Heilongjiang

Received: April 27, 2026; accepted: May 26, 2026; published: June 8, 2026

*通讯作者。

文章引用: 吴佳佳, 庞宝坤. 基于《生死疲劳》语料库的大语言模型文学翻译质量评估[J]. 现代语言学, 2026, 14(6): 177-187. DOI: 10.12677/ml.2026.146513

Abstract

The strategy of promoting the excellent traditional Chinese culture to go global has imposed higher requirements on translation quality, while the rapid advancement in digital intelligence technologies has provided technical support for applying large language models (LLM) into literary translation. On this basis, this study takes Mo Yan's *Life and Death Are Wearing Me Out* as a case, constructing a "one-to-three" Chinese-English parallel corpus consisting of the source text, Howard Goldblatt's translation, and translations generated by DeepSeek-V3.2 and ERNIE 4.5. The study employs BERTScore for quantitative assessment of translation quality, and further introduces House's Translation Quality Assessment model to conduct a qualitative analysis of the LLMs' translation performance from the three dimensions of field, tenor, and mode. The findings indicate that the overall semantic fidelity of large language model translations is close to that of human translations, yet low-scoring sentences reveal common problems such as loss of cultural information, weakening of emotional coloring, and stylistic convergence. This study provides empirical evidence for the digitally-driven international dissemination of Chinese culture, and suggests that the translation efficacy of LLMs can be further enhanced through structured prompt optimization in the future.

Keywords

English Translation of *Life and Death Are Wearing Me Out*, Translation Quality Assessment, Large Language Model (LLM), BERTScore, House Model

1. 引言

当前，在中华文化“走出去”战略深入推进的背景下，如何高质量、高效率地将中华优秀传统文化翻译至海外市场，已经成为提升国家文化软实力的重要议题之一。近年来，大语言模型(Large Language Model, LLM)发展势头迅猛，应用领域持续拓展，也开始作为译者登上人类翻译实践的舞台[1]，在文本处理效率方面展现出显著优势。然而，其在复杂语境中的翻译可信度与语言准确性仍有待进一步的实证检验。

相较于一般应用型文本，文学翻译不仅涉及语义层面的转换，更强调文化意蕴、审美风格与话语表征的综合再现，这对译者的语言驾驭能力与跨文化适应能力提出了极高要求。正因如此，文学翻译常被视为检验翻译质量与译者综合能力的“试金石”[2]，亦为评估大语言模型在复杂语境下生成语义的性能提供了极具挑战性与解释力的实证场域。

鉴于莫言文学创作中鲜明的乡土意象与独特的叙事复调特征，本研究选取长篇小说《生死疲劳》作为研究对象，通过构建融合定量指标(BERTScore)与定性框架(豪斯模型)的多维评估体系，对国内主流大语言模型的文学翻译表现进行系统考察，旨在为后续的结构化提示词优化提供实证依据，进而赋能数智时代中国文学的海外传播。

2. 文献综述

近年来，学界围绕大语言模型在翻译中的应用及翻译评估方法开展了大量探索，同时，对《生死疲

劳》英译的研究持续深化。本文从上述三方面对相关文献进行梳理，旨在为将前沿翻译技术引入经典文学译本质量评估提供参考。

2.1. 大语言模型在文学翻译中的应用研究

近年来，大语言模型在翻译领域的应用研究大量涌现，包括法律、民航、科技、中医药等各个领域[3]-[6]。这些研究结果普遍表明，大语言模型在基本传意以及句式规范性方面有一定的优势，但在处理专业术语、文化特有概念时仍存在不足之处。

相比之下，大语言模型在文学翻译上的应用研究相对较少。文学翻译由于其对文本风格、情感色彩、文化内容有着极高的要求，被认为是机器翻译中最具挑战性的方向之一。赵衍等[7]以《繁花》为例，对比了大语言模型与人工译本的翻译质量，发现大语言模型在方言处理和文化专有项传达方面存在明显欠缺。张曙康等[8]对沈从文的《边城》进行了翻译质量的对比研究，得出不同大语言模型存在显著的性能差距，且认为在文学翻译中提示词工程优化的效果不如非文学领域显著。

2.2. 翻译质量评估方法研究

翻译质量评估是翻译研究的核心话题之一。“评估”最主要的出发点在于追求客观性，方法上力求科学，尽量避免主观性，特指基于特定目标收集信息并做出定量、定性分析的某种过程[9]。在各类评估手段中，自动化评估指标因其效率高、可重复而被广泛应用。传统的自动化评估指标如 BLEU、TER 等依赖 n-gram 匹配机制，虽然计算逻辑简单直接，但对于语义等价且表层形式迥异的表达不够敏感[10] [11]。针对上述局限，Zhang 等[12]提出了 BERTScore 评估指标，该指标借助预训练模型计算候选译文与参考译文之间的词语级上下文嵌入相似度。张曙康等[8]的研究证实了 BERTScore 能够有效捕捉译文的语义保真度，并能敏锐反映不同模型之间的质量差异。这一发现为本研究选取 BERTScore 作为定量分析工具提供了直接的方法论支撑。

除了定量指标外，定性维度的评估在文学翻译中同样不可或缺。豪斯的翻译质量评估模型[13]以系统功能语言学为理论基础，丰玉芳等[14]、孙玲等[15]、罗茜等[16]均已在宋词、散文、小说等多种文学体裁的翻译分析中得到应用与检验，证明了其在文学文本翻译中的适用性，这也为本研究将豪斯模型引入长篇小说翻译评估提供了可循的方法参照。

2.3. 《生死疲劳》英译研究

《生死疲劳》是莫言的代表作之一，入选中国改革开放四十周年 40 部最有影响力小说，曾荣获美国“纽曼华语文学奖”、日本“第 17 届福冈亚洲文化大奖”、中国香港第二届“红楼梦奖”，而葛浩文译本被公认为是高质量文学翻译的典范。目前，欧阳珊等[17]、杨莎莎等[18]、刘庚等[19]分别从成语翻译、文化负载词处理、翻译策略等角度对葛浩文译本进行了探讨。然而，这些研究均聚焦于人工译本本身，尚未有大语言模型译本与人工译本的系统对比。

综上所述，大语言模型在文学翻译中的应用研究仍处于起步阶段，研究文本的数量与种类有待进一步丰富与拓展。在评估方法方面，将自动化评估指标与定性分析模型相结合的混合研究仍较为有限。基于上述空白，本研究以莫言《生死疲劳》为例，构建汉英平行语料库，采用定量与定性分析相结合的混合框架，对比分析大语言模型译文与人工译本的翻译质量差异。

3. 研究设计与方法

本章详细阐述本研究的整体设计逻辑与具体实施路径，研究首先通过语料库构建与模型选取，为实验提供标准化的数据基础；在此基础上，构建一套定量与定性相结合的混合评估框架，以期对大语言模

型的文学翻译质量进行多维度的对比分析与深度评价。

3.1. 语料库构建与模型选取

研究选取莫言长篇小说《生死疲劳》为中文源文本[20]，选取葛浩文译本 *Life and Death Are Wearing Me Out* 为权威英文参考译本[21]。对齐工作基于在线对齐平台 Tmxmall 完成，首先将中英文文本按段落自动切分并进行初步句级对齐，随后由人工进行精细调整。完成初步对齐后，利用平台内置的去重功能剔除完全相同的句对，并随机抽取 200~300 个句对进行人工抽样校验。最终得到 9772 对有效句子，用于后续模型训练与质量评估。

鉴于国外大语言模型对于中文训练语料占比极低(0.16%)且部分不向中国提供 API 服务，本研究选取国内大语言模型为翻译工具。黄协安等[22]的研究表明，DeepSeek R1、文心一言在中译英方面的表现相对稳定，因此本研究选取 DeepSeek-V3.2 和文心一言 4.5 (ERNIE Bot 4.5)两个大语言模型生成机器译文。在完成中文原文与葛浩文英译本的句级对齐后，本研究通过 Python 脚本分别通过两个模型接口，生成两套大语言模型译本。两个模型均使用相同的基础提示词：“你是一个中英翻译专家。请将以下中文文本翻译成英文，只输出译文”，后接待译句子。提示词中未加入任何关于文化保留、风格或归化、异化的指导，便于后续分析模型自身的翻译能力。批量翻译完成后，将 DeepSeek 和文心一言的译文与葛浩文译本并列，形成包含“1 对 3”的汉英平行语料库，用于后续评估。

3.2. 混合评估框架构建

3.2.1. BERTScore 的计算与应用：定量评估

为量化不同模型译文的质量差异，本研究采用 Zhang 等[12]提出 BERTScore 作为定量评估指标。不同于传统 n-gram 匹配方法，BERTScore 基于上下文语义嵌入，计算机器译文(候选文本)与人工参考译文的语义相似度。

Bertscore 的具体计算过程如下：首先，将机器译文与参考译文分别输入 BERT 模型，获取每词元的上下文嵌入向量。其次，对于译文中的每个词，算法会在参考译文中匹配语义最相似的词，并计算二者嵌入向量的余弦相似度。最后，基于语义相似度矩阵，分别计算精确率(Precision)、召回率(Recall)与 F1 分数。其中，精确率衡量机器译文在参考译文中可找到语义对应词的比例，用于衡量译文的准确性；召回率反映参考译文信息被机器译文覆盖的程度，用于衡量完整性；F1 分数为二者的调和平均值，用于综合评估翻译的整体质量。

本研究以葛浩文英译本为参考译文，分别计算 DeepSeek 和文心一言两大语言模型译文的 BERTScore。计算基于 bert-score Python 库实现，选用多语言预训练模型 distilbert-base-multilingual-cased，并设置 re-scale_with_baseline = False 以保持原始分数，逐句获取精确率、召回率及 F1 值。

为客观比较两模型译文的整体质量差异，对逐句 F1 得分进行配对样本 t 检验，并计算 Cohen’s d 效应量以衡量差异幅度；同时绘制箱线图以直观呈现两译本 F1 得分的分布特征与离散程度。上述量化结果为后续目的性抽样选取典型案例提供依据。

3.2.2. 豪斯质量评估模型的应用：定性评估

为弥补定量评估在捕捉文化信息、情感色彩及风格特征方面的局限，本研究进一步引入豪斯翻译质量评估模型[13]进行定性分析。该模型以系统功能语言学为理论基础，通过对比原文与译文在语场(field)、语旨(tenor)、语式(mode)三个语域维度的特征，识别翻译过程中的误配与偏差，从而对翻译质量作出系统评估。

本研究的定性分析分为三步：第一，基于 BERTScore 定量评估结果，采用目的性抽样选取典型例句：

优先抽取两模型 F1 差异显著或双低分的句子，并兼顾文本类型与句子长度分布，确保样本具备典型性与代表性。第二，将所选例句依据语场、语旨、语式三维度分别归类，对比葛浩文译本与大语言模型译文在各维度上的差异，识别并归类翻译偏差与功能误配类型。第三，归纳大语言模型在翻译中的共性问题与典型偏差，为未来探索提示词优化与翻译质量提升提供实证依据。

4. 研究结果与讨论

4.1. Bertscore 定量评估结果与分析

4.1.1. 描述性统计

本研究以葛浩文英译本为参考，分别计算 DeepSeek-V3.2 和文心一言 4.5 两个大语言模型译文的 BERTScore F1 值。描述性统计结果如表 1 所示。

Table 1. Descriptive statistics of BERTScore F1 for large language model translations
表 1. 大语言模型译文的 BERTScore F1 值描述性统计结果

	平均 F1	标准差(SD)	最大值	最小值	中位数
DeepSeek-V3.2	0.7965	0.0953	1.0000	0.2822	0.8141
文心一言 4.5	0.6630	0.0905	1.0000	0.2584	0.6661

平均 F1 得分是衡量译文质量的核心指标，F1 值越高，表明译文与参考译文(葛浩文英译本)的语义相似度越高，翻译准确性与完整性越强。由表中数据可知，DeepSeek 译文的平均 F1 为 0.7965，显著高于文心一言译文的 0.6630，表明 DeepSeek 在整体语义保真度方面更接近葛浩文译本。

标准差(SD)反映 F1 得分的离散程度，标准差越小，说明模型在不同句对上的翻译表现越稳定、波动越小。两个模型的标准差数值相近，且均处于较低水平，表明两大模型的翻译稳定性整体较好。

F1 得分的最小值与最大值反映模型翻译性能的极端表现，能够体现模型在复杂句、简单句翻译中的极限能力，进一步补充整体表现与稳定性分析的不足。两大模型的 F1 最大值均为 1.0000，说明在部分简单句、语义清晰、无文化负载词的句子中，两者都能生成与参考译文完全一致的译文。DeepSeek-V3.2 最小值为 0.2822，高于文心一言 4.5 的最小值为 0.2584，表明其在最差情况下仍略优于文心一言，整体下限更高，表现出更强的稳健性。

中位数能排除极端值的干扰，更客观地反映数据的集中趋势。从本研究的数据来看，DeepSeek-V3.2 的中位数 F1 (0.8141)高于其均值(0.7965)，这表明其得分分布略呈负偏态(左偏)，即高分段译文所占比例较大。反观文心一言，中位数(0.6661)与均值(0.6630)基本一致，说明其译文质量集中在中等水平，分布较为对称。

综合各项统计指标来看，DeepSeek-V3.2 在平均表现、中位水平以及最低表现等方面均优于文心一言 4.5，显示出更高的整体翻译质量与更强的性能稳健性，而文心一言虽具备一定翻译能力，但整体表现相对较弱，在复杂语境下仍存在提升空间。

4.1.2. 统计检验与分布可视化

为检验两个模型在翻译质量上的差异是否具有统计学意义，本研究基于逐句 BERTScore F1 得分差值进行了配对样本 t 检验。结果表明，差值均值显著大于零($M\_diff = 0.1335$, $SD = 0.0045$)，$t(9720) = 182.58$，$p < 0.001$，说明两种模型在译文质量上存在高度显著差异。进一步计算效应量得到 Cohen's $d = 1.852$，达到大效应水平($d \geq 0.8$)，表明该差异不仅在统计意义上显著，在实际应用层面亦具有较强意义。

同时，为更直观地呈现数据分布特征，采用云雨图(Raincloud Plot)对两组数据进行可视化展示，如图1所示。

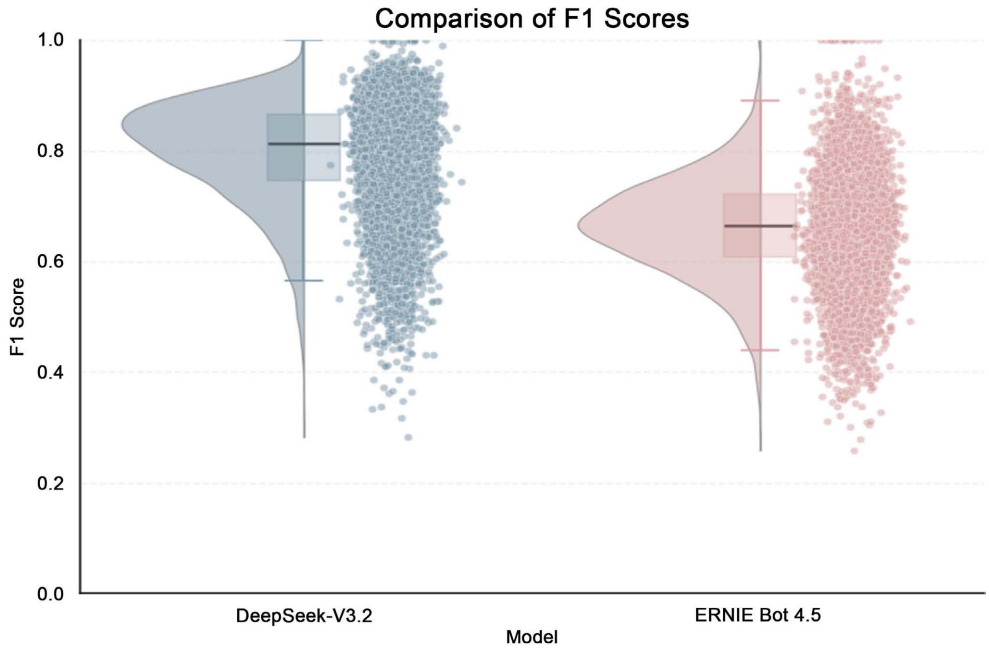

Figure 1. Comparison of BERTScore F1 scores
图 1. BERTScore F1 分布对比

云雨图融合了核密度估计、小提琴图与箱线图的优势，不仅能够反映数据的整体分布形态，还能同时呈现中位数、四分位数及离散程度，从而弥补单一统计指标在分布表达上的局限性。

从核分布密度(小提琴图形状)来看，DeepSeek-V3.2 的分布整体明显偏向高分区间，密度主要集中在0.75 至 0.90 之间，并在 0.80 附近形成显著峰值，表明其多数译文能稳定保持较高的语义相似度水平。相比之下，文心一言的分布整体向左偏移，主峰位于 0.65 附近，说明其整体质量集中于中等水平。

在箱线图结构上，DeepSeek-V3.2 的中位数 DeepSeek-V3.2 的中位数明显高于文心一言，且箱体整体位置更高，上四分位数接近 0.85 以上，而文心一言的上四分位数约位于 0.70 左右。这表明 DeepSeek 不仅整体翻译质量更高，其高质量译文所占比例也更大。同时，两者箱体高度(四分位距)差异不大，说明两组数据在各自区间内的离散程度相近，这一点与前文标准差分析结果一致。

散点分布(雨滴部分)表明 DeepSeek-V3.2 的数据点大多密集分布在 0.70 以上，且在 0.80 附近呈现明显聚集趋势，低于 0.60 的点相对较少。而文心一言的数据点分布更为分散，在 0.50 至 0.75 之间呈现较大密度，同时在 0.45 附近仍存在一定数量的低分点。这一现象进一步表明，文心一言在部分句对上的翻译质量波动较大，存在一定比例的低质量输出。

综合以上数据可以看出，DeepSeek-V3.2 在整体分布位置、高分集中程度以及低分控制能力方面均优于文心一言，表现出更高的翻译质量与更强的稳健性。后续章节将把上述低分异常的句子纳入豪斯模型的定性分析考察范围，以此探究大语言模型在翻译薄弱环节背后的深层原因。

4.2. 豪斯翻译质量定性评估与分析

本节以豪斯的翻译质量评估模型为理论依托，从语场、语旨、语式三大维度，对典型翻译案例展开

对比分析。受篇幅限制，本节不再单独罗列原文语域分析、译文文本分类等步骤，但全程遵循豪斯评估模式“原文分析－译文比对－质量判断”的核心逻辑，并将其贯穿于各案例分析之中。下文将直接切入典型案例的对照分析。

4.2.1. 语场对比与分析

语场对应文本所涉及的主题内容和社会活动。在翻译质量评估中，该维度重点考查文化专有项、专业术语及背景信息的跨语言传递是否准确。

例 1：

原文	葛浩文译文	DeepSeek (F1: 0.687)	文心一言(F1: 0.643)
阎王与身边的判官低声交谈几句，然后一拍惊堂木，说：“好了，西门闹，知道你是冤枉的。”	One of the judges leaned over and whispered something in Lord Yama’s ear. He banged his gavel to silence the hall. “All right, Ximen Nao, we accept your claim of innocence.”	The King of Hell exchanged a few low words with the judge beside him. Then he struck the gavel and said: “Alright, Ximen Nao, we know you’ve been wronged.”	The King of Hell and the judge beside him whispered a few words. Then he beat the court gong and said: “Okay, Simen Niao, I know you are innocent.”

本句语场聚焦中式冥界司法场景，包含“阎王”、“判官”、“惊堂木”等极具浓厚本土宗教与传统司法属性的文化专有词汇，构成独特的民间冥界叙事语境。

“阎王”源自印度佛教中的“阎魔罗阇”(Yamaraja)，后融入中国民间信仰，演化为阴曹地府的主宰神祇。葛浩文将其译为“Lord Yama”，采用音译加身份敬称的策略，在保留源语文化专名异质性的同时，通过“Lord”凸显其在冥界的统领地位。相比之下，DeepSeek 与文心一言均采用归化译法，将其译为“King of Hell”。其中“Hell”自带西方基督教地狱的语义色彩，与东方冥界“阴曹地府”的意象存在偏差。该译法虽利于通俗理解，却流失了“Yama”专属的佛教文化渊源，在文化信息传递的准确性低于葛浩文译本。

“惊堂木”是中国古代司法器具，用于拍击案桌以维持公堂秩序。英美法系中虽无完全对应物，但法官木槌(gavel)功能与其相近。葛浩文采用功能对等策略，译为“gavel”并增补“to silence the hall”，完整还原了其司法功能[23]。DeepSeek 仅译为“struck the gavel”，未加说明，文化补偿不足。文心一言误译为“court gong”(锣)，在形态、材质、使用方式及文化语境上均与惊堂木不符，属于概念误译。

整体而言，葛浩文对各类文化意象的处理手法颇为灵活，在传达文化意蕴的同时亦顾及目标受众的理解门槛。DeepSeek 产出的译文虽句意流畅、表层无误，但文化信息补偿不足，致使部分历史及民俗含义难免被冲淡。至于文心一言，则出现了关键文化概念的错译，语场层面的适配亦存在明显纰漏。由此可见，面对相对小众的历史文化专有项，大语言模型在语义把控与产出稳定性上仍有欠缺，其译文质量与人工译本之间的差距尚未消弭。

4.2.2. 语旨对比与分析

语旨关乎参与者之间的角色关系、作者的社会背景和交际立场、社会角色关系和社会态度。在翻译质量评估中，该维度重点考察译文能否忠实复刻原文的情感色彩、语气强度与人际态度，以保障交际关系与情感表达的对等性。

例 2：

原文	葛浩文译文	DeepSeek (F1: 0.641)	文心一言(F1: 0.545)
我西门闹何止救过一条命？大灾荒那年春天我平价粜出二十石高粱，免除了所有佃户的租子，使多少人得以活命。可我却落了个何等凄惨的下场，天和地，人和神，还有公道吗？还有良心吗？	Me, Ximen Nao, and not just one life. During the famine one spring I sold twenty bushels of sorghum at a low price and exempted my tenant farmers from paying rent. That kept many people alive. And look at my miserable fate. Is there no justice in heaven or on earth, in the world of men or the realm of spirits? Any sense of conscience?	How many lives have I, Ximen Nao, saved? In the spring of the great famine, I sold twenty dan of sorghum at a fair price and waived the rent for all my tenants, saving many lives. Yet what a miserable end I have come to! Is there no justice left in heaven and earth, among men and gods? Do you still have a conscience?	Have I not saved a life, Xi Men Nao? That year of great famine, I sold 20 shi of sorghum at a fair price and exempted all the tenants from rent, so that many people could survive. But what a miserable ending I have got. Is there any justice in the world between heaven and earth, between human and god? Have you got any conscience?

例 2 以第一人称“我”展开内心控诉与悲愤诘问，语旨特征鲜明。叙事主体为蒙冤的西门闹，交际立场兼具自我申辩、委屈愤懑与绝望质问，语气由行善自述的沉缓逐步转向对命运的控诉，并以连续反问宣泄对天地、神人、公道与良心的质疑，主观情感浓烈。

三版译文在语气、情感强度及交际立场的还原上存在明显差异。葛浩文译本通过短句切分与口语化独白贴合人物倾诉语气。开篇以碎片化短句强化自我申诉情绪，结尾将反问拆分为两个无主语短句，保留了对天地神明的泛化质问，整体语气沉郁悲凉，语旨还原充分。DeepSeek 译本以反问句再现开篇语气，行文紧凑，情感层次完整，但在结尾反问中引入了第二人称“you”，将原文对抽象良心的普泛性质问转向具体对象，偏离了原有的交际立场，语旨有所削弱。文心一言译本同样存在人称偏移，且开篇将“何止救过一条命”弱化为“是否救过一条命”，语气平铺直叙，情感张力不足，语旨弱化更为严重。

总的来说，葛浩文最能贴合人物独白的表达逻辑，语旨共情性最强；DeepSeek 虽保留了情感层次与反问语气，但人称处理影响语旨呈现；文心一言因句式平淡、人称错位及语义弱化，语旨适配度最低。

4.2.3. 语式对比与分析

语式关注文本的媒介(口语/书面语)及参与程度(独白/对话)。在翻译质量评估中，该维度重点考察句式结构、衔接手段及修辞风格是否与原文一致。

例 3

原文	葛浩文译文	DeepSeek (F1: 0.835)	文心一言(F1: 0.650)
在这危急时刻，说时迟，那时快，捕狼队中年纪最小的队员吕小坡，瞄准狼头开了一枪——因为狼是运动目标，击中的正是狼腹——狼从空中跌落，在地上翻滚，肠子流出来，拖出好长，其状凄惨，虽是凶残野兽，也让我们心中不忍。	At that desperate moment, in less time than it takes to tell, Lü Xiaopo, the youngest member of the hunting team, took aim and fired at the animal’s head. But since it was a moving target, the shot hit her in the abdomen. When she fell, she rolled on the ground, spilling her guts all over the place, a terrifying sight. Though she was a vile predator, it was too terrible to see.	At this critical moment, in the blink of an eye, Lü Xiaopo, the youngest member of the wolf-hunting team, aimed at the wolf’s head and fired a shot.—Because the wolf was a moving target, the shot hit precisely its abdomen. The wolf fell from the sky, tumbled on the ground, its intestines spilling out and dragging along for a long distance—a pitiful sight. Though it was a fierce beast, we couldn’t help but feel a pang of sympathy in our hearts.	At this critical moment, the youngest member of the wolf-catching team, Lv Xiaopo, quickly aimed at the wolf’s head and fired —Because the wolf is a moving target, it hit the wolf’s belly—The wolf fell from the sky and rolled on the ground, its intestines flowing out and dragging for a long distance. It was so pitiful that even as a fierce wild animal, it made us feel sorry.

原文采用“说时迟，那时快”这一口语化短句推进情节，营造紧迫节奏；中间以破折号插入补充说明，打破线性叙事节奏，形成书面语标记；随后将动作描写、场景刻画与主观抒情层层衔接，长短句错落有致，叙事张弛有度，形成极具画面感的文学叙事形态。

葛浩文译本将该口语套语意译处理，既保留节奏功能，又符合英语表达习惯。同时将破折号插入语转化为独立分句，通过简单连词实现衔接，逻辑清晰。后续将长句拆分，动作描写与情感评论分句呈现，整体节奏舒缓自然。DeepSeek 译本以“in the blink of an eye”对应原句，简洁传神，同时保留破折号插叙的篇章结构，形式上贴近原文。对后续长句采用复合句处理，信息密度较高，但整体仍具有。文心一言译本同样保留破折号格式，但句式多为简单句堆砌，衔接手段单一，语言平铺直叙，削弱了原文烘托的紧张感与情绪张力。此外，将原文的过去时态误用为现在时，影响语义准确性。

总体来看，葛浩文在语式层面兼顾口语节奏与书面逻辑，衔接自然；DeepSeek 形式贴近原文，语式还原度较高；文心一言在时态一致性与传情达意上存在不足，语式匹配度最低。

综合语场、语旨、语式三个层面的分析可以看出，大语言模型译文虽能实现基础语义的传递，但在文化负载项解读、情感语气还原、文学句式风格把控等方面存在共性局限，相较资深人工译者的文学化表达，在文学翻译场景中仍存在不可忽视的短板。

5. 研究结论与启示

本研究以《生死疲劳》为语料构建汉英平行语料库，结合 BERTScore 和豪斯模型构建混合评估框架，对葛浩文译本同国内主流大语言模型 DeepSeek-V3.2、文心一言 4.5 的翻译质量进行了系统比较。定量分析表明，DeepSeek-V3.2 的平均 BERTScore F1 值(0.7965)显著高于文心一言 4.5 (0.6630)，两者差异达到大效应量(Cohen’s d = 1.852)，说明当前国内不同大语言模型的文学翻译能力已存在明显的性能分层。定性分析进一步揭示了两模型共同的薄弱环节：在语场维度，文化专有项常被归化甚至误译，导致文化信息缺失；在语旨维度，第一人称情感控诉中的反问强度、人称立场会发生偏移或弱化；在语式维度，原文的口语节奏、破折号插叙等句法风格被简化或时态错配。

上述发现一方面验证了“大语言模型在文学翻译中语义保真度尚可但文化情感风格存在短板”这一已有共识，另一方面提供了新的实证贡献：① 首次对比了 DeepSeek-V3.2 与文心一言 4.5 在长篇乡土小说翻译中的量化差异及效应量；② 将笼统的“文化情感短板”细化为豪斯三维度下的具体误配类型及语言表征；③ 为后续提示词优化指明了精准的靶点，即通过构建“语域增强型”结构化提示词，针对性地实现文化信息补偿与语旨张力重构。

结构化提示词的核心思想是将翻译指令分解为若干独立、可组合的模块，每个模块分别针对特定的误配类型进行干预。具体而言，可围绕角色设定(你是一位精通中文乡土文学与英文文学翻译的专家，尤其擅长处理莫言作品中的民间信仰、历史典故和情感独白)、语场规则(显式规定文化负载词的异化策略与功能说明要求)、语旨规则(限定情感极性、语调强度及人称指称范围)以及语式规则(约束特定句法衔接手段与时态逻辑)构建多层提示结构。该框架具有良好的可迁移性，研究者可根据不同文本类型或误配类型增删相应模块。未来研究可在此基础上开展对照实验，验证结构化提示词相较于基础提示词在文化保真度、情感还原度和风格适配度上的提升效果。

本研究亦存在一定局限。由于实验测试次数有限，结果的稳定性与可重复性可能受大语言模型“幻觉”影响，仍需进一步检验。此外，研究基于单一文本进行，尚不足以全面反映模型在不同类型文学文本中的翻译表现。本文提出的结构化提示词框架尚未经过实证检验，其有效性有待后续实验验证。未来研究可在扩大语料范围与实验规模的基础上，根据已甄别的共性问题引入多轮测试与提示词干预机制，进一步探讨大语言模型在文学翻译中的优化路径。

基金项目

1) 2025 年度黑龙江省艺术科学规划项目，全感元宇宙视域下金源文化跨语言艺术重构与多模态传播途径研究，编号：2025B056。

2) 2025 黑龙江省文化和旅游科研课题，全感元宇宙视域下金源文化跨语言智能传播业态创新研究，编号：2025WL040。

3) 2026 年度黑龙江省语委语言文字科研项目，大语言模型在东北方言跨语言传播中效度与优化策略研究，编号：2026Y042。

4) 2023 年度黑龙江省艺术科学规划项目，数字赋能视角下黑龙江民俗短视频的多模态话语研究，编号：2023B093。

参考文献

[1] 胡开宝, 李晓倩. 大语言模型背景下翻译研究的发展: 问题与前景[J]. 中国翻译, 2023, 44(6): 64-73+192.

[2] 刘云虹, 许钧. 文学翻译模式与中国文学对外译介——关于葛浩文的翻译[J]. 外国语(上海外国语大学学报), 2014, 37(3): 6-17.

[3] 李奉栖, 张云, 丁丽杰. 大语言模型与神经网络机器翻译系统专业文本翻译质量对比——以法律汉英翻译为例[J]. 上海翻译, 2025(6): 62-67.

[4] 刘畅, 陈双双, 吴云涛. 基于语料库和 AI 大语言模型的科技翻译研究——以民航文本为例[J]. 中国科技翻译, 2025, 38(3): 25-28.

[5] 李银玲. 大语言模型赋能科技文本翻译质量研究[J]. 中国科技翻译, 2025, 38(4): 21-24.

[6] Xia, M., Wu, S., Yang, Y. and Yang, Y. (2025) Traditional Chinese Medicine Terminology Translation via Large Language Model: A Deepseek-Based Study. *English Language Teaching and Linguistics Studies*, **7**, 7-14. https://doi.org/10.22158/eltls.v7n3p7

[7] 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.

[8] 张曙康, 赵朝永. 大语言模型之于文学翻译的适切性研究——基于多指标评估的《边城》多模型译文质量对比[J]. 中国外语, 2025, 22(4): 85-95.

[9] 孙琳. 关于翻译质量评估的思考[J]. 上海翻译, 2023(5): 37-41.

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

[11] Snover, M., Dorr, B., Schwartz, R., *et al*. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the* 7*th Conference of the Association for Machine Translation in the Americas*, Cambridge, August 8-12 2006, 223-231.

[12] Zhang, T., Kishore, V., Wu, F., *et al*. (2019) Bertscore: Evaluating Text Generation with Bert. arXiv: 1904. 09675.

[13] House, J. (2015) Translation Quality Assessment: Past and Present. Routledge. https://doi.org/10.1057/9781137025487_13

[14] 丰玉芳, 王菲菲. 从豪斯的翻译质量评估模式看宋词翻译——以宋词《声声慢》和许渊冲英译本为例[J]. 扬州大学学报(人文社会科学版), 2015, 19(3): 114-121.

[15] 孙玲. 基于朱莉安·豪斯翻译质量评估模式评张培基英译散文诗《匆匆》[J]. 海外英语, 2018(17): 140-141.

[16] 罗茜. 基于豪斯翻译质量评估模式的《活着》英译本研究[J]. 英语广场, 2022(13): 3-7.

[17] 欧阳珊. 跨文化视角下葛浩文的四字成语英译策略研究——以莫言小说《生死疲劳》为例[J]. 湘潭大学学报(哲学社会科学版), 2023, 47(5): 153-157+172.

[18] 杨莎莎. 基于语料库的《生死疲劳》中文化负载词英译研究[J]. 重庆理工大学学报(社会科学), 2016, 30(10): 126-130.

[19] 刘庚, 卢卫中. 汉语熟语的转喻迁移及其英译策略——以《生死疲劳》的葛浩文英译为例[J]. 外语教学, 2016, 37(5): 91-95.

[20] 莫言. 生死疲劳[M]. 北京: 作家出版社, 2006.
[21] Goldblatt, H. (2008) Life and Death Are Wearing Me Out. Arcade Publishing Inc.
[22] 黄协安, 赵善江. 国内外主流大语言模型中译英翻译质量的实证对比分析[J]. 语言与翻译, 2026(1): 50-59+66.
[23] 郑银芬, 高歌. 体认翻译学视角下《生死疲劳》中文化专有项的英译策略研究[J]. 英语广场, 2026(4): 13-16.