主流大语言模型文学翻译质量比较研究

——基于《遥远的向日葵地》节选的实证分析

时博文，刘丽敏*

福建农林大学　福建福州

【摘要】以李娟的散文集《遥远的向日葵地》中的连续性两节《外婆的世界》与《外婆的葬礼》为文本，分别用 Kimi、Deepseek 3.2、Gemini 3.0 Flash、通义千问 3.5 Plus、通义千问 3 Max、GPT 5.4 和文心一言七个大语言模型进行翻译，使用不同的翻译评估指标将各系统译文与已出版的人工译本进行比较评估，并采用豪斯量表和 MQM 量表对译文的准确性和文学性进行评价。结果显示，大语言模型进行文学翻译时已有良好表现，Kimi 和 Deepseek 3.2 在各项评估中表现突出，而其他语言模型对特色文化内涵的理解传达和文学性表达上仍存在较大优化空间。

【关键词】大语言模型；《遥远的向日葵地》；翻译质量；评估；比较

【基金项目】福建省本科高校教育教学研究项目（FBJY20240174）；福建农林大学校级重点教改项目（111422131）

【收稿日期】2026 年 5 月 10 日　　【出刊日期】2026 年 6 月 9 日　　【DOI】10.12208/j.ssr.20260200

A comparative study on the quality of large language models' translation in literary text: Taking excerpts from *Distant Sunflower Fields* as an example

*Bowen Shi, Limin Liu**

Fujian Agriculture and Forestry University, Fuzhou, Fujian

【**Abstract**】This study extracted two consecutive sections, "Grandma's World" and "Grandma's Funeral," from the essay collection *Distant Sunflower Fields*. These texts were translated using seven large language models (Kimi, Deepseek 3.2, Gemini 3.0 Flash, Qwen-3.5 Plus, Qwen-3 Max, GPT 5.4, and ERNIE Bot). Various translation evaluation metrics (TTR, BLEU, METEOR) were employed to evaluate and compare each translation with the published human translation. Additionally, the House's Scale and MQM Scale were utilized to evaluate the accuracy and literariness of the translations. The results indicate that Large Language Models (LLMs) have displayed robust capabilities in literary translation, with Kimi and Deepseek 3.2 notably distinguishing themselves in multi-dimensional assessments. Nevertheless, other peer models continue to face challenges in interpreting culture-specific nuances and mastering stylistic literary expression, highlighting a considerable gap that necessitates further optimization.

【**Keywords**】Large language models; *Distant Sunflower Fields*; Translation quality; Assessment; Comparison

1 引言

大语言模型（Large Language Model, LLM）是一种基于深度学习神经网络的自然语言处理（Natural Language Processing, NLP）模型。它通过在海量语料上进行训练，学习并模拟人类的语言能力，从而能够理解、生成自然语言文本，并根据指令完成多样化的语言处理任务[1]。当前，以 ChatGPT、Gemini 和 Kimi 等为代表的大语言模型，已广泛应用于各类场景。

大语言模型在翻译中具有显著优势，其通过庞大的神经网络和海量数据训练，能够更准确地理解源语言语义，生成更精确的译文；同时，依托对长文本中大量词元的处理能力，能够保证译文上下文逻辑清晰

作者简介：时博文（1999-）男，河南开封人，硕士，研究方向：翻译理论与实践；
*通讯作者：刘丽敏（1981-）女，福建永春人，副教授，博士，研究方向：翻译史、区域国别学。

且连贯；此外，通过“温度”等参数调节，模型输出更具多样性和创造性，表现出更接近人类的自然语言表达[2]。然而，大语言模型对语义的理解和传达效果很大程度上取决于训练语料的数量、质量和类型。换言之，如果训练语料中缺乏相关平行文本和专业词汇，其翻译质量就难以得到保证。在文学翻译领域，通常不会使用特定领域的专业平行文本对大语言模型进行专门训练，因此，大语言模型在文学翻译领域的表现尚不确定。此外，大语言模型在推理时，会生成不遵循源文本或者不符合事实的内容，即“幻觉”（Hallucination）。数据源、训练过程和推理过程是大模型产生幻觉的三大来源[3]。虽然可以通过调低“温度”、改进提示工程、整合外部知识等方法在一定程度上减轻大语言模型的幻觉，但目前还无法从根本上消除大语言模型的幻觉[4]。

当前，关于大语言模型的研究主要聚焦于其架构优化、通用语言理解与生成能力，以及在问答、摘要、文本分类等通用自然语言处理任务中的性能表现[5]。在常见文体的翻译上，大语言模型的表现可与专业人类译员相提并论，甚至在某些情况下能够超越人类译员[6]。然而，针对大语言模型文学翻译的质量评估以及针对性优化性能的研究，因其对文化内涵、文本风格和情感传达等深层要素的独特要求，仍是亟待深入探究的领域[7]。基于此，本研究拟对 Kimi、Deepseek 3.2、Gemini 3.0 Flash、通义千问 3.5 Plus、通义千问 3 Max、GPT 5.4 和文心一言七个主流大语言模型在文学翻译中的表现进行比较，旨在评估大语言模型在文学翻译中的应用效果，为未来翻译技术的优化与发展提供实证参考。

2 实验设计

2.1 实验目的

通过测评大语言模型在处理汉译英文学翻译任务时的质量，分析翻译技术在文学翻译应用中的表现。实验的主要测试指标包括翻译技术评估指标如 TTR、BLEU、METEOR 以及译文的准确性和文学性。

2.2 大语言模型选择

选取当前主流的大语言模型 Kimi、Deepseek 3.2、Gemini 3.0 Flash、通义千问 3.5 Plus、通义千问 3 Max、GPT 5.4 和文心一言七个大语言模型进行不联网翻译，使用不同的翻译评估指标如（TTR、BLEU、METEOR）将各系统译文进行数据化对比分析，并使用豪斯量表和 MQM 量表对译文的准确性和文学性进行评价。

2.3 测试源文本的选取

选取中国当代作家李娟的散文集《遥远的向日葵地》作为测试语料。该作品自 2017 年出版以来，荣获第七届鲁迅文学奖散文杂文奖等多项重要文学奖项，是一部高质量的文学作品。实验从中节选《外婆的世界》与《外婆的葬礼》两篇连续文本，共计 5277 字。所选文本涵盖生活对话、方言俗语、人名地名及隐喻内涵等复杂语言现象，将其作为源文本，能够较为全面地评估大语言模型在处理文学文本时的综合表现。

2.4 实验过程与操作

在翻译时，对 Kimi、Deepseek 3.2、Gemini 3.0 Flash、通义千问 3.5 Plus、通义千问 3 Max、GPT 5.4 和文心一言七个大语言模型给出详细指令（prompt），即“请以一位专业译员的水平将以下中文小说内容翻译为英文，要求译文忠实于原文，准确传达原文意义且表达地道”，由于选取的文本已有英文出版物，因此使用大语言模型时采取不联网模式进行翻译。将各翻译系统的结果进行整合，将已出版的译文作为参考，测试各项翻译质量评估指标并评价译文的准确性和文学性。

3 实验结果

以已出版译文为基准，采取现行常用的翻译质量评估指标，包括 TTR、BLEU、METEOR 等多项数据进行评测，并对各译文进行等级评估。此外，基于 MQM 错误层级量表中的“准确度”指标对翻译结果进行了客观统计，并根据朱莉安·豪斯评估模式中的语义、语用和语篇维度对译文质量的文学性进行了等级评估。

3.1 翻译技术评估指标对比

表 1　各系统译文评估数据汇总表

大语言模型	TTR（%）	BLEU（%）	METEOR（%）	综合等级
参考译文	35.99	-	-	基准
Kimi	33.36	51.84	23.4	A
Deepseek 3.2	32.69	53.02	23.3	A
通义千问 3 Max	32.56	51.56	23.3	B
通义千问 3.5 Plus	32.58	52.89	23.2	B
GPT 5.4	32.37	49.65	23.0	C
Gemini 3.0 Flash	32.09	49.37	22.8	C
文心一言	31.81	48.77	22.7	C

TTR，即词汇多样性比（Type-Token Ratio），主要评估译文的词汇多样性是否与源文本保持一致，或者译者是否在目标语言中充分展示了词汇运用能力。TTR 值越高，表示文本中使用的词汇越多样，重复率越低。TTR 主要关注词汇表面形式的多样性，无法直接衡量词语使用的准确性、恰当性以及译文的整体语义和语用质量。如表 1 所示，所有译文的 TTR 值都在 31.81%到 33.36%之间，普遍低于参考译文（TTR 为 35.99%）。原因主要是英文表达需要更多功能词，导致词汇重复率高； 此外，这些大语言模型在词汇选择上相对保守，未能充分利用英文的同义表达提升词汇多样性。从 TTR 指标来看，Kimi 的 TTR 最高，为 33.36%，最接近参考译文，意味着其在词汇多样性方面表现最好，尝试使用了更多不同的词汇并更有效地避免了重复。Deepseek 3.2（32.69%），通义千问 3.5 Plus（32.58%），通义千问 3 Max（32.56%）表现良好，达到了较高水准。而 Gemini 3.0 Flash（32.09%），GPT 5.4（32.37%）的 TTR 值相对较低，提示这些译文在词汇选择和多样性方面存在改进空间，避免词语过度重复，以提升译文的整体质量和阅读体验。文心一言的 TTR 最低，为 31.81%，意味着其词汇丰富度过低，需要更大程度地优化。

BLEU（Bilingual Evaluation Understudy）是 IBM 在 2002 年提出的一种机器翻译质量自动评估指标。它通过比较机器翻译的输出与一个或多个高质量人工参考译文的匹配程度来量化翻译质量，衡量译文与参考译文的 n-gram（连续词串）匹配度。BLEU 越高，字面相似度越高。BLEU 分数并非准确率，而是一个相似度指标。100 分意味着机器译文与参考译文完全相同。在实际中，尤其在有多个参考译文的情况下，人工翻译也难以达到 100 分。如表 1 所示，所有译文的 BLEU 得分都在 48%到 53%之间，表明所有译文与参考译文的词汇和短语重叠度较高，整体质量尚可。其中 Deepseek 3.2 译文的 BLEU 值最高为 53.02%，表明它与参考译文在词汇和短语上的匹配度最高，在词汇选择、句法结构以及表达习惯上与参考译文最为相似。通义千问 3.5 Plus（52.89%）紧随之后，同样表现出色，与参考译文高度相似。通义千问 3 Max（51.56%）和 Kimi（51.84%）得分超过 50%，表明它们在匹配度方面也表现优异。Gemini 3.0 Flash（49.37%），GPT 5.4（49.65%）得分接近，略低于 50%，但处于中等偏上水平，表明与参考译文的匹配度尚可。文心一言（48.77%）得分最低，说明它在词汇选择、短语搭配或句法结构上与参考译文的差异相对较大。

METEOR（Metric for Evaluation of Translation with Explicit Ordering）是在 BLEU 基础上改进的另一种机器翻译自动评估指标，考虑词干、同义词以及短语对齐、语义相似性等因素，更侧重语义准确性和召回率，旨在弥补 BLEU 只关注 n-gram 匹配的不足，其评估结果与人工判断的相关性更高。特别是在评估机器翻译系统时，METEOR 常常与 BLEU 结合使用，以获得更全面的视角。由于计算方式更严格，METEOR 的得分通常低于 BLEU。测评中所有译文的 METEOR 得分都在 22.7%到 23.4%之间（见表 1），得分区间集中，差异很小，说明所有译文在词语覆盖率、语义匹配和词序排列上与参考译文的整体表现相近。Kimi 得分最高，为 23.4%，表明它在语义匹配、词形变化处理及词序排列上与参考译文最为接近，且对参考译文信息的覆盖度也最高。Deepseek 3.2（23.3%）和通义千问 3 Max（23.3%）并列第二，与最高分差距微小，表现非常优秀。通义千问 3.5 Plus（23.2%），GPT 5.4（23.0%）和 Gemini 3.0 Flash（22.8%）紧随其后，表现良好，处于中等水平。文心一言（22.7%）得分最低，意味着它在语义表达、词形处理或词序上与参考译文的差异相对较大。

3.2 译文的准确性对比

（1）漏译

本次评估以参考译文为标准，对各系统译文的漏译字数进行统计和评级。统计时力求数据准确，以高标准进行衡量。本文将漏译情况分为三个等级：A（漏译字数 0- 5 字）、B（漏译字数 6-15 字）、C（错译数漏译字数 16-50 字），结果如表 2。

表 2　各系统译文对源文本（5277 字）的漏译字数统计表

大语言模型	漏译字数	评级
Kimi	0	A
Deepseek 3.2	3	A
GPT 5.4	11	B
通义千问 3 Max	15	B
文心一言	27	C
通义千问 3.5 Plus	43	C
Gemini 3.0 Flash	43	C

整体而言，大部分译文的漏译字数在可接受范围内，但仍有提升空间。特别是对于一些源文本中带有情感色彩或方言特色的词句的处理，仍需更加精细，以准确传达源文本的韵味。其中，Kimi（漏译字数为 0）表现最佳和 Deepseek 3.2（漏译字数仅为 3 字）表现优异。

Gemini 3.0 Flash 和通义千问 3.5 Plus 的漏译字数较多，仍需要大幅改进。

（2）增译

各系统译文都出现了增译现象。增译有两种类型：

一是自行增加源文本不存在的内容。在表达“心急如焚”时，多数译文除了直接翻译“anxious”或“heart burning with anxiety”，还会加入一些对“旅途耗时”或“内心感受”的额外描述，如“spending a whole day on the road, my heart burning with anxiety.”（Gemini 3.0 Flash、GPT 5.4）、“it took two bus transfers and a whole day on the road, my heart frantic with worry.”（Deepseek 3.2）、“spending the entire day traveling, my heart racing.”（通义千问 3.5 Plus）、“which took me an entire day. I was anxious and worried all the way.”（通义千问 3 Max）、“a whole day’s journey, my heart scorched with worry.”（Kimi）、“spending the entire day on the road, and feeling extremely anxious.”（文心一言）。这些增译共同丰富了作者在路途中的心理状态和时间投入。

在翻译“筋疲力尽，灰心丧气”时，常常会加入一些额外的修饰词，如“exhausted, weary, and filled with a certain despair”（GPT 5.4）、“I was exhausted, disheartened.”（Deepseek 3.2）、“I was exhausted.”（通义千问 3.5 Plus）、“I was completely exhausted and discouraged.”（通义千问 3 Max）、“I was exhausted and disheartened.”（Kimi）、“I am exhausted and discouraged.”（文心一言）。这些增译共同强化了情绪的程度和感受。

二是对源文本内容进行了总结。在翻译“我妈家大业大，又是鸡又是狗又是牛的，整天忙得团团转，哪能像我一样专心。”时，Gemini 3.0 Flash、GPT 5.4 和 Deepseek 3.2 共同用更凝练的语言概括了母亲“家大业大”和“忙得团团转”的状态，：“My mother had a big family and business, with chickens, dogs, and cows, busy all day long, unlike me who could focus on one thing.”（Gemini 3.0 Flash）、“My mother had a large household and many animals – chickens, dogs, cattle – keeping her busy all day, so she couldn’t dedicate herself to Grandma as I could.”（GPT 5.4）、“Mom had her hands full with a large household – chickens, dogs, cows – constantly spinning like a top. She couldn’t devote herself like I could.”（Deepseek 3.2）。

（3）错译

错译主要包含三种类型。首先是命名实体（人名、地名、组织名）错译和前后翻译不一致。Gemini 3.0 Flash、GPT 5.4、通义千问 3.5 Plus、Kimi 在这一方面表现出色，未出现错译情况。而其他大语言模型均有出现错译情况，通义千问 3 Max 和文心一言的翻译错译字数较多，还存在人名前后翻译不一致的情况。例如：Deepseek 3.2 将“李秦氏”翻译为“Li-Qin Clan”，“氏”通常指姓氏，而非宗族或家族；通义千问 3 Max 和文心一言将“秦妹仔”翻译为 “Qin sister”/ “Qin Mei”，“妹仔”在语境中是昵称，翻译偏离了源文本的亲昵；将狗的名字“赛虎”统一翻译为 “Saihu”/“The tiger”。

本文将人名错译分为三个等级：A（错译数 0）、B（错译数 1-10 字）、C（错译数 10-20 字），结果如表 3。

表 3　各系统译文名称错译数统计表

大语言模型	命名实体错译字数	人名前后翻译不一致处数	评级
Gemini 3.0 Flash	0	0	A
GPT 5.4	0	0	A
通义千问 3.5 Plus	0	0	A
Kimi	0	0	A
Deepseek 3.2	2	0	B
文心一言	14	0	C
通义千问 3 Max	16	0	C

其次是语法方面存在错误。在机器翻译中，常见时态错乱的情况，包括动词时态不一致、情态动词使用不当、时间状语与动词时态冲突、叙事时态不统一、习惯性动作/状态与一般现在时/过去时混淆。例如，对于“她已经不知时间是怎么回事了。她已经不知命运是怎么回事了。”的翻译，Gemini 3.0 Flash、GPT 5.4、Deepseek

3.2、通义千问 3.5 Plus 和通义千问 3 Max 将现在完成时翻译为一般过去时，失去了“已经”的持续性和对当前状态的影响。

对于“她的寿衣已经准备了二十多年。无论走哪儿都随身带着。我在很小的时候就已经无比熟悉它的存在了。”的翻译，通义千问 3 Max 和文心一言均使用了现在完成时和一般现在时，这与源文本清晰的过去时间背景严重冲突，完全改变了事件发生的时间，造成了叙事逻辑的混乱。Gemini 3.0 Flash 和 Kimi 都正确地使用了过去完成时或简单过去时来描述这些发生在过去且持续到过去某一时刻的状态和习惯，这种处理方式契合了源文本的回忆叙事，避免了与时间线脱节的严重错误。

各译文在处理一些模棱两可的时态时，如描述过去常态或普遍事实时，中文往往不明确区分一般过去时和一般现在时，容易出现偏差。此外，对于现在完成时强调的持续性状态，部分译文也未能准确把握。

本文将语法错误分为三个等级：A（错译数 0-3 字）、B（错译数 4-6 字）、C（错译数 7-9 字），结果如表 4。

表 4　各系统译文语法错译数量统计表

大语言模型	时态错乱数量	评级
Kimi	2	A
Deepseek 3.2	3	A
GPT 5.4	5	B
通义千问 3.5 Plus	5	B
通义千问 3 Max	7	C
Gemini 3.0 Flash	8	C
文心一言	9	C

再者是对称呼、方言和独特内涵的翻译理解以及传达错误。例如，人物称呼“娟啊”、“老子”、“李秦氏”，四川方言“悄悄眯眯”、“晓得晓得，我又不是细（小）娃儿”、“好生，打烂了要赔”，以及一些具有独特内涵的文字“天啦，又黑又瘦，真是从来也没见她这么黑过……是不是大限要到了？”、“人死如灯灭” 、“喜丧”等。

综合所有译本，Kimi 和 Deepseek 3.2 整体译文质量较高，对文化内涵的理解和传达比较到位，对文化内涵的处理较为细致，其他大语言模型均存在较多翻译腔和文化内涵丢失的问题。这表明 Kimi 和 Deepseek 3.2 在处理文化内涵和方言方面做出了更多努力，并取得了较好的地道性和准确性。本文将理解错误分为三个等级：A（错误数 0-5 字）、B（错误数 6-15 字）、C（错误数 16-25 字），结果如表 5。

表 5　各系统译文对源文本理解错误数统计表

大语言模型	总计错误数	评级
Kimi	3	A
Deepseek 3.2	5	A
通义千问 3.5 Plus	9	B
GPT 5.4	11	B
Gemini 3.0 Flash	12	B
通义千问 3 Max	16	C
文心一言	20	C

3.3 译文的文学性

在评价译文文学性时，主要涵盖以下三个方面：一是对源文本内容意义理解传达文字的可读性（语义）；二是对地方特色文化专有项和特色表达处理的到位程度（语用）；三是对特色写作手法的再现程度（语篇）。按文学性表现由好到差排序，结果如表 6。

表 6　各系统译文文学性评级表

大语言模型	评级
Kimi	A
Deepseek 3.2	A
通义千问 3.5 Plus	B
通义千问 3 Max	B
GPT 5.4	C
Gemini 3.0 Flash	C
文心一言	C

从可读性角度来看，译文在语言学意义上的准确性和连贯性是流畅阅读体验的基础。文学翻译在此基础上，更要求在语义、语用和语篇层面精准再现源文本的艺术魅力。在这层意义上，文心一言的译文普遍存在大量漏译、错译，可以观察到明显的机器翻译痕迹，基本上不具备文学性。这种现象主要是因为基于机器学习的大语言模型对上下文长度的支持有限，在处理长篇叙事时，其处理模式难以应对长文本中复杂的语义联系和细微情感，容易割裂上下文，导致语义理解偏差。

在处理地方特色文化和特色表达时，A、B 档译文（Kimi、Deepseek 3.2、通义千问 3.5 Plus、）展现了更高的敏感度和技巧。例如，Kimi 和 Deepseek 3.2 能够通过巧妙的意译和语境重构，根据语境将其转化为“I’ll come steal them tonight…”或“Darn it!”，既传达了语气，

又符合目标语的表达习惯，使“老子”的粗犷语气在英文中得到恰当体现。对于“记忆”这种带有外婆独特理解的词语，它们也通过上下文的铺垫或适当的解释，成功地保留了这份独特的语用意义，而非简单地直译导致歧义。C 档译文（如文心一言）常常出现误译、遗漏或简单的字面翻译，完全未能传达源文本的文化韵味和语用功能。例如，对“老子”、“秦妹仔”这类充满情感色彩和地方特色的口语化称谓，C 档译文多采用生硬直译或泛化处理，使得人物形象和情感表达大打折扣。

在评价译文对原作特色写作手法的再现程度时，译文的完整性是一个重要的考量指标。源文本独特的叙事节奏、情感铺陈、重复强调、比喻暗喻以及反讽等修辞手法，共同构筑了其独特的文学风格。若缺失了外婆等待、迷失、与赛虎的互动以及葬礼上的反讽等关键片段，源文本的文学性和人物刻画将大打折扣。例如，通过“她已经不知时间是怎么回事了。她已经不知命运是怎么回事了”的重复，强调了外婆的迷失与生命的终结。又如“唯有死亡才能令她展翅高飞”这一凝练而富有诗意的结尾，充满了深沉的悲剧美和哲思。C 档译文（特别是文心一言）由于大量误译、漏译以及破碎的语篇结构，源文本的文学风格和情感表达被严重破坏，完全未能再现其写作手法和文学性。文心一言甚至因技术错误导致语篇残缺不全。

A、B 档译文（Kimi、Deepseek 3.2、通义千问 3.5 Plus）在语篇完整性和写作手法再现方面表现较好。它们能够保持源文本的叙事流程，准确传达情感的起伏，并有意识地再现重复、排比、暗喻等修辞手法，从而维持了源文本的文学张力。例如，Kimi 对“她已经不知……了”的重复，通过“Time had lost meaning for her; so had fate. ”这种凝练的表达，既保留了重复的强调，又提升了诗意。对“唯有死亡才能令她展翅高飞”的翻译，Kimi 和 Deepseek 3.2 都保持了其意象和哲思，使得译文在语篇层面具有较强的文学感染力。

根据之前的评估，Kimi 和 Deepseek 3.2 在文学性方面表现突出，现从动作描述、暗喻修辞、情感传达和诗歌美学四个方面进行更细致的对比。

表 7　Kimi 与 Deepseek 3.2 文学性细节对比表

类型	源文本	Kimi (A)	Deepseek 3.2 (A)	比较分析
动作描述	奶奶趴在阳台上眼巴巴地朝小区大门方向张望。	Grandma leaning on the balcony railing in the distance, gazing longingly towards the main gate.	I’d see her far off on the balcony, eyes fixed on the gate.	Kimi 的“eyes fixed on the gate”比 Deepseek 3.2 的“gazing longingly”更具动态感和专注度，直接刻画了外婆望眼欲穿的形象，省去了对“栏杆”的描述，更聚焦于“眼睛”这个核心动作，更精炼有力。
暗喻修辞	她在迷途中慢慢向死亡靠拢，慢慢与死亡和解。	On her wayward path, she was slowly approaching death, slowly reconciling with it.	she was utterly lost, drifting toward death, reconciling with it.	Kimi 表现更佳。“drifting toward death”中的“drifting”比 Deepseek 3.2 的“approaching death”更有意象美和无力感，巧妙地暗喻了外婆在生命末期随波逐流、失去方向的状态，更好地呼应了“迷途”的意象。Deepseek 3.2 的“wayward path”和“approaching”虽准确，但略显平实，少了 Kimi 的诗意和深层隐喻。
情感传达	我赶紧回家，倒了两趟车，路上花了一整天工夫，心急如焚。	I took two buses, a whole day’s journey, my heart scorched with worry.	I rushed home immediately. It took two bus transfers and a whole day on the road, my heart frantic with worry.	Kimi 的“heart scorched with worry”非常形象地表达了“心急如焚”的“烧灼”感，文学性强，情感传达深刻。Deepseek 3.2 的“heart frantic with worry”也准确地表达了焦急，但“scorched”的强烈程度和意象感更胜一筹。
诗歌美学	她已经不知时间是怎么回事了。她已经不知命运是怎么回事了。	Time had lost meaning for her; so had fate.	She no longer understood time. She no longer understood fate.	通过重复“她已经不知……了”来强调外婆的迷失感。Kimi 将其凝练为“Time had lost meaning for her; so had fate.”，通过分号连接，保持了简洁而深刻的节奏感，且“lost meaning”和“so had fate”的对应非常精妙。Deepseek 3.2 采用两句独立句式“She no longer understood time. She no longer understood fate.”，虽然也再现了重复，但不如 Kimi 那样紧凑和富有诗意。

如表 7 所示，Kimi 在文学性方面略胜 Deepseek 3.2 一筹，其译文在保持源文本意义准确性的同时，能更深入地捕捉源文本的文学神韵、情感深度和艺术美感，通过更具创造性和凝练性的语言选择，使得译文在目标语中拥有更强的感染力和独立审美价值。不仅如此，Kimi 在动作描述和暗喻修辞方面，倾向于选择更

概括、更具意象感和深层含义的词汇，使得文本在保持流畅性的同时，也增加了文学的厚重感和艺术感染力。Deepseek 3.2 也在高水准之列，忠实且流畅地传达了源文本，在情感传达的生动性和口语化再现方面表现突出。它更善于通过活泼、地道的口语表达来重现人物的语气和情绪，使得外婆这个角色的形象更加鲜活，更具人情味。在文学性上更侧重于通过语言的感染力直接触动读者的情感，让人物的喜怒哀乐跃然纸上。

4 结语

本研究选择了 Kimi、Deepseek 3.2、Gemini 3.0 Flash、通义千问 3.5 Plus、通义千问 3 Max、GPT 5.4 和文心一言七个大语言模型，分别对散文集《遥远的向日葵地》中《外婆的世界》和《外婆的葬礼》进行了汉译英翻译测试。结果显示，Kimi 在翻译各项翻译评估指标及译文的准确性和文学性方面均表现优异；Deepseek 3.2 在各项翻译评估指标的表现也较为突出，源文本理解和传达准确性也较好，但对文学性的理解和表达稍逊色之外；Gemini 3.0 Flash 和通义千问 3.5 Plus 在各项翻译评估指标和译文准确性上表现突出，但在文学性方面稍显逊色，且会出现大量漏译；GPT 5.4 在各项评估中均处于中间位置，无突出表现；通义千问 3 Max 和文心一言除漏译较少外，其他评估均处于末尾等级。

综合来看，大语言模型在进行文学翻译任务的表现普遍较好，尤其是 Kimi 和 Deepseek 3.2 在各项评估中均表现突出，在对中文特色内涵的理解和传达上表现极佳，其原因主要是大语言模型基于神经机器翻译模型采用端到端的训练方式，可以直接学习源语言和目标语言之间的映射关系，并通过迁移学习，将已经学到的知识应用到新任务中[8]。其他翻译模型的译文在不同程度上尚有提升空间，尤其是文心一言的译文质量表现较差，在对特色文化内涵的理解传达和文学性表达上远不如其他大语言模型，需要更深入的优化。

参考文献

[1] Floridi, L. & Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences[J]. Minds and Machines, 2020(4): 681-694.

[2] 李亚超,熊德意,张民等.藏汉神经网络机器翻译研究[J].中文信息学报,2017,31(06):103-109.

[3] Huang, L. etal. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions [EB/OL]. Retrieved from http://arxiv. org/abs/2311. 05232, 2023

[4] 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,(04):60-66+109.

[5] 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(06):64-73+192.

[6] 胡开宝,李娟.大语言模型背景下的翻译人才培养:挑战与前景[J].外语电化教学,2024,(06):3-7+105.

[7] 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(04):85-95.

[8] 赵浜,曹树金.国内外生成式 AI 大模型执行情报领域典型任务的测试分析[J].情报资料工作,2023,44(05):6-17.