国内外六款AI大语言模型英语写作文本特征对比研究

罗芷汀,柴省三
(北京语言大学 国际学生教育政策与评价研究院,北京 100083)

摘要:基于人工智能(AI)的大语言模型正在全方位重塑语言教育生态。大语言模型通过基于海量文本数据的预训练,能够捕捉到语言的复杂性,生成适应不同语境与主题的文本,并能够为学习者提供即时反馈、写作文本生成和逻辑框架,有望改变传统的二语写作教学模式。由于每种大语言模型在训练方式、底层架构和算法方面并不完全相同,因此在写作文本质量上也可能有差异。本文通过实验研究,基于Coh-Metrix文本分析工具,对ChatGPT、Gemini、Claude、DeepSeek、文心一言和Kimi等主流大语言模型在英语同题写作文本的词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性进行了对比分析,研究结果可为英语二语写作教学中的模型选择提供适配度参考。

关键词:大语言模型(LLMs);英语写作评价;Coh-Metrix;语言特征
中图分类号:H319.3 **文献标志码:**A **文章编号:**1001-5744(2025)04-0069-11

一 引言

(一)研究背景与意义

随着人工智能(Artificial Intelligence, AI)技术的发展,国内外若干大语言模型(Large Language Models,LLMs)如ChatGPT、Claude、DeepSeek等纷纷面世,LLMs在教育领域尤其是英语二语(L2)写作教学中的垂直应用已经越来越普遍[1]。LLMs可以为英语第二语言学习者提供实时反馈(Intermediate feedback)、写作示例(Exemplars)和自动评分(Automated writing evaluation,AWE)应用,从而帮助学习者提高写作能力。基于庞大的训练数据和复杂的深度学习(Deep learning)架构,LLMs在语言理解和内容生成方面展现了不俗的能力[2],而且可以通过提供有针对性的高效反馈、增强词汇使用广度和改善语法准确性等方式帮助英语二语学习者提升写作能力,辅助英语教师提高教学效果,使其成为英语二语写作的重要个性化辅助工具[3]。

针对LLMs生成文本的语言特征,不少研究人员进行了实证研究。有研究表明,ChatGPT生成的文本几乎没有语法错误,因此也被语言教师尝试用于二语写作教学,如生成写作大纲、提供写作思路与观点、对论文初稿进行修改、润色等[4]。ChatGPT生成的文本无论是在词形还是词种比率上,都显著高于人类作文[5]。Herbold等人针对英语二语学习者和ChatGPT分别产出的270篇议论文进行了对比研究,发现ChatGPT生成的作文在语言表达、主题完整性、逻辑结构和文本衔接性等方面明显优于英语二语学习者的作文[6]。还有研究表明,虽然ChatGPT和Gemini等大语言模型在生成连贯且语境

收稿日期:2025-06-06
作者简介:罗芷汀(2000—),女,贵州贵阳人,北京语言大学国际学生教育政策与评价研究院博士研究生,主要从事语言习得研究。
通信作者:柴省三(1964—),山东潍坊人,北京语言大学教授,博士研究生导师,主要从事语言习得、语言测试研究。
引用格式:罗芷汀,柴省三.国内外六款AI大语言模型英语写作文本特征对比研究[J].宁夏大学学报(社会科学版),2025,47(4):69-79.

适宜的句子方面表现出色，但它们在内容选择和批判性思维（critical thinking）的处理方面存在不足[7]。此外，基于汉语母语环境开发的AI大语言模型，比如文心一言和Kimi等可能在为中国英语二语学习者提供本地化语言支持方面具有独特优势，能够更好地契合中国学生英语写作学习的需求。虽然LLMs已引起了语言学界的关注，但不同大语言模型输出的文本在语言复杂性（complexity）、词汇多样性（lexical diversity）、语义衔接性（coherence）和句法结构（sentence structure）等方面可能存在差异，因此通过实验研究，全面、系统地考察不同大语言模型在英语写作文本方面的语言学特征，深入分析不同LLMs生成的文本的语言特征，并据此为英语二语学习者提供模型评价、选择和高适配度应用具有重要的实践价值。

基于上述考虑，本研究将使用Coh-Metrix英语写作自动分析工具，针对国内外六种主流大语言模型（ChatGPT 4.0、Gemini 1.5、Claude 3.0、DeepSeek、文心一言4.0和Kimi）进行英语同题写作文本的实验研究，通过比较不同大语言模型生成的文本在词汇多样性（lexical diversity）、语义衔接（coherence）、句法复杂度（syntactic complexity）和文本可读性（readability）等语言特征指标方面的差异，全面评价不同大语言模型生成的英语文本的相对优势，为英语作为第二语言的写作教学和学习提供借鉴，并为LLMs在语言教学和评估中的针对性应用提供实证依据。

（二）Coh-Metrix在英语文本分析中的应用

Coh-Metrix是由美国孟菲斯大学McNamera等人基于语料库语言学、计算语言学、自然语言处理等多学科的最新进展而研发的可以采集并量化写作文本的可读性、衔接性、句法复杂度等多项指标的英语文本自动化分析工具[8]。Coh-Metrix可以实现对英语文本的表层结构和深层特征的自动量化分析和有效评估。目前，Coh-Metrix已广泛应用于英语二语写作的评价研究中，它可以帮助英语教师分析学生作文中的语言特征，评估写作质量，并探索不同语言特征与写作能力之间的相关关系[9]。

Crossley和McNamara通过Coh-Metrix对文本的语法复杂度、衔接性和词汇多样性等特征进行了分析，结果表明上述指标可以有效预测英语二语者的写作熟练程度[10]。Ullmann在2019年使用Coh-Metrix分析了学生的写作文本，发现语言的复杂性、衔接性和词汇多样性与英语写作成绩之间存在着显著的相关关系[11]，表明Coh-Metrix文本分析指标在预测学习者的英语写作水平方面具有较高的效度。

Coh-Metrix文本分析工具可以提供多达108项文本评估指标，能够较准确地量化分析文本的语言特征。在这些指标中，最有价值的评估指标主要包括词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性，它们可以全面揭示文本的复杂性、连贯性和适宜性。

词汇多样性（LDTTRa）是一个反映文本语言丰富程度的重要评估指标。多项研究发现，词汇因素在作文整体评分中对分数的影响最大[12-14]，文章的质量在很大程度上取决于所使用的词汇[15]。然而，过度丰富的词汇使用有时也会导致文本主题表达不明确，并可能对低水平英语学习者的理解构成负担。通过Coh-Metrix对词汇多样性的评估，能够帮助英语教学者选择哪些模型生成的文本在词汇使用上更适合英语二语学习者。

句子衔接性（LSASS1）和段落衔接性（LSAPP1）是Coh-Metrix提供的评估文本衔接性的两个重要指标。衔接能力对于二语学习者而言尤为重要。语言学习者在写作中往往面临连接句子和段落的困难，而缺乏衔接性的文本会让读者难以理解作者的语义逻辑和表达意图。通过分析文本的LSASS1和LSAPP1指标，可以有效揭示不同LLMs生成的文本的衔接性差异，准确判断哪些模型生成的文本更具连贯性，更适合充当二语学习者的写作学习辅助工具。

在句法复杂度（SYNLE）方面，Coh-Metrix通过计算每个句子中的名词短语数量来衡量句法复杂度。早在1980年，Crowhurst的研究表明，10到12年级学生的议论文写作分数与句法复杂度之间存在显著相关[16]。过高的句法复杂度可能使二语学习者在理解和产生句子时遇到困难。

文本可读性指标（RDFKGL）是Coh-Metrix文本分析工具提供的可以评估文本可读性的重要指标，所谓可读性（readability）就是指文本的理解难易度。对于二语学习者来说，文本的可读性直接影响其理

解和产出能力。过高的RDFKGL值可能意味着文本的词汇句法等过于复杂，不适合英语低水平学习者。

Coh-Metrix作为一个精确评估写作质量的有效工具，不仅在评估学生写作质量方面具有重要的价值，而且在评估LLMs生成的文本语言学特征方面也是一个可靠的工具。它能够从多个维度对文本的语言特征进行量化分析，为二语学习者提供精准的认知诊断指导，也为英语教育工作者选择合适的模型提供了针对性建议。随着LLMs在二语教育中的广泛应用，结合Coh-Metrix进行文本分析的应用性研究颇为必要，研究成果将有助于英语教学者根据学生的实际需求选择最适合的写作辅助工具。

二　研究设计

（一）研究语料

本研究所使用的语料均是国内外六个主流大语言模型（LLMs）生成的英语文本。具体来说，这些模型包括ChatGPT、Gemini、Claude、DeepSeek、文心一言和Kimi，六个模型均是基于Transformer架构的生成式预训练通用模型，其核心技术包括自注意力机制、多层解码器结构以及预训练与微调相结合的训练策略，目前LLMs已在文本生成、分类、翻译、逻辑和语言对话—反馈等通用领域衍生出了广阔的应用场景，尤其在生成写作文本时表现出各自独特的特点和优势，作为研究对象的六个大语言模型的基本概况见表1。

为了深度分析不同LLMs生成的文本语言学特征，我们采用相同的提示语（prompts）和写作规则（rubrics），在相同的提示框下引导六个大模型分别生成9篇英文文本，共计获得54篇文本作为研究语料。文本生成指令的题目要求如下。

表1　六种LLMs基本特征

模型全称	开发公司	年份	国家	开发环境
ChatGPT	OpenAI	2022年11月	美国	英语
Claude	Anthropic	2023年3月	美国	英语
Gemini	Google	2023年12月	美国	英语
文心一言	百度	2023年3月	中国	汉语
Kimi	月之暗面	2023年1月	中国	汉语
DeepSeek	深度求索	2025年1月	中国	汉语

Based on the following topic, write 3 English essays. Write at least 250 words in each essay. Each essay should be divided into paragraphs and logically coherent.

Topic: In universities, people tend to concern more about the career prospect of the subjects they learn. Some people propose that students should learn subjects with a fast job growth including science, technology, and engineering even if these are not parts of their field of study. Do you agree or disagree this statement? Use specific reasons and examples to support your answer. Write at least 250 words.

为了避免题目（topic）因素对文本生成产生的偏差，我们的作文任务设计借鉴了大型国际标准化英语写作考试（改编自IELTS 8 Test 2）的题目要求，以统一的提示语引导六个大语言模型围绕给定的主题进行论证写作。写作题目要求模型生成不少于250字的英语作文，题目内容为开放性讨论类问题。

本研究将针对语言大模型生成的作文语料，使用Coh-Metrix英语文本分析工具对不同LLMs生成的文本的典型语言特征指标进行分析和对比研究。

（二）研究问题

1.六种大语言模型生成的文本在词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性指标方面是否存在显著差异？

2.国内与国外两类大语言模型生成的文本在词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性指标方面是否存在显著差异？

（三）研究思路

首先，针对六个大语言模型（LLMs）生成的54篇英语写作文本进行数据收集和预处理。为了确保数据的准确性和分析的可操作性，所有文本首先被输入Word文档进行存储，并分别转换成txt格式，形成一个微型语料库。同时，根据文本的来源和生成模型，对每篇作文进行编码处理，例如，ChatGPT-1表示由ChatGPT生成的第一篇文本，Claude -2表示由Claude生成的第二篇文本，以此类推，以确保每个文本都具有明确的来源标识。

其次，采用Coh-Metrix 3.0英文文本自动分析软件，分别针对每篇文本进行语言特征指标提取和分析。本研究中，Coh-Metrix提取的语言指标主要包括LDTTRa（词汇多样性）、LSASS1（句子衔接性）、LSAPP1（段落衔接性）、SYNLE（句法复杂度）和RDFKGL（文本可读性）等。

最后，基于Coh-Metrix提取的所有模型的文本的词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性等语言学特征指标，使用SPSS 27.0软件对每个模型生成的文本的上述指标进行描述性统计分析，同时，采用单因素方差分析（ANOVA），深入考察不同模型在不同语言特征指标方面的差异，探讨ChatGPT、Gemini、Claude、DeepSeek、文心一言和Kimi等模型生成的文本特征。此外，对六种大语言模型按照国内外两大类进行分类，对比国内LLMs（DeepSeek、文心一言和Kimi）与国外LLMs（ChatGPT、Gemini和Claude）生成的作文文本总体在五种文本指标上的差异。

（四）指标定义

针对六个大语言模型生成的文本，我们采用Coh-Metrix 3.0工具进行分析。Coh-Metrix可以针对文本提供若干指标，但对文本质量评价最有价值的指标主要关注的是文本词汇层面、句子层面、段落层面以及整体层面。具体的文本语言特征指标定义见表2。

表2　5项Coh-Metrix文本语言特征指标

指标缩写	中文名称	计算方法	指标含义
LDTTRa（Lexical Diversity, Type - Token Ratio, all words）	词汇多样性	所有词的类型标记比率	评估文本的词汇多样性，有助于判断文本的语言水平。数值越大，用词越丰富
LSASS1（Latent Semantic Analysis Sentence-to-Sentence Similarity）	句子衔接性	用潜在语义分析（LSA）为每个句子生成一个向量，再计算每一组相邻句子的余弦相似度平均值	衡量相邻句子之间的语义相似度，数值越高，句子间语义联系越强，句子衔接性越好
LSAPP1（Latent Semantic Analysis Paragraph-to-Paragraph Similarity）	段落衔接性	计算相邻段落之间LSA余弦的平均值	测量段落与段落之间的语义相似度，数值越大，段落间衔接性越好。反映文本的整体结构和连贯性
SYNLE（Syntactic Complexity: Mean Number of Modifiers per Noun Phrase）	句法复杂度	句子中主句的主动词之前的平均字数	评估文本的句法复杂度，反映语法结构的复杂性，数值越大，文本句法结构越复杂
RDFKGL（Readability: Flesch-Kincaid Grade Level）	文本可读性	READFKGL =（0.39 ×句平均词数）+（11.8 × 词平均音节数）- 15.59	评估文本难度，数字越大，文本越难读。该指标有助于评估文本适用不同二语水平等级的学习者

三　研究结果

（一）描述性统计

为了考察不同LLMs在相同提示语下生成的同题英语文本的宏观特征，我们首先针对六个LLMs生成的文本长度进行描述性统计，具体结果见表3。

描述性统计结果显示，文心一言和Claude生成的文本较长，9篇文本的平均长度分别是353个词和334个词；ChatGPT和DeepSeek生成的文本长度较短，分别是289个词和283个词。国内模型（文心一言、Kimi和DeepSeek）和国外模型（ChatGPT、Claude和Gemini）生成的文本平均长度分别是312个词和320个词。

为了从微观上探讨不同LLMs生成的文本的语言学特征变量，针对每个模型的9篇文本采用Coh-Metrix进行文本分析，并获得词汇多样性（LDTTRa）、句子衔接性（LSASS1）、段落衔接性（LSAPP1）、句法复杂度（SYNLE）和文本可读性（RDFKGL）五个指标的描述性统计结果（见表4）。

表3　六个LLMs描述统计

	ChatGPT	Claude	Gemini	文心一言	Kimi	DeepSeek
文本篇数	9	9	9	9	9	9
最长篇幅词数	298	386	340	380	356	322
最短篇幅词数	273	290	279	331	290	259
总体平均词数	289	334	312	353	323	283

表4　六个大语言模型生成文本的语言特征指标描述性统计

		ChatGPT	Claude	Gemini	文心一言	Kimi	DeepSeek
LDTTRa	平均值	0.628	0.614	0.582	0.578	0.558	0.626
	标准差	0.027	0.046	0.024	0.016	0.027	0.017
	最小值	0.598	0.549	0.533	0.544	0.528	0.610
	最大值	0.670	0.703	0.608	0.598	0.610	0.661
LSASS1	平均值	0.198	0.306	0.244	0.208	0.224	0.237
	标准差	0.025	0.045	0.058	0.023	0.035	0.025
	最小值	0.156	0.232	0.150	0.176	0.173	0.201
	最大值	0.227	0.367	0.352	0.242	0.273	0.283
LSAPP1	平均值	0.349	0.455	0.365	0.343	0.369	0.396
	标准差	0.058	0.055	0.037	0.042	0.035	0.059
	最小值	0.280	0.390	0.284	0.275	0.305	0.292
	最大值	0.457	0.568	0.401	0.411	0.410	0.503
SYNLE	平均值	6.239	5.562	6.153	6.032	7.031	5.829
	标准差	1.112	1.297	1.480	0.730	0.841	0.468
	最小值	4.556	3.882	3.813	4.944	5.750	5.056
	最大值	7.588	8.467	8.733	7.050	8.625	6.533
RDFKGL	平均值	13.958	16.333	14.107	13.693	14.065	13.507
	标准差	1.080	0.824	1.041	0.982	0.928	0.650
	最小值	12.252	15.225	12.528	12.378	12.789	12.176
	最大值	15.401	17.164	15.404	14.980	15.493	14.436

从大语言模型生成的文本指标的描述性统计数据来看，我们可初步发现以下结果：(1)从生成的词汇多样性指标来看，ChatGPT表现最突出，LDTTRa的平均值达到了0.628，表明其生成文本的词汇丰富度最高；文心一言和Kim的词汇多样性指标则相对较低。(2)在句子衔接性方面，Claude模型的表现最好，LSASS1的平均值是0.306，ChatGPT则最低，LSASS1的平均值仅为0.198。(3)在段落衔接

性方面，Claude模型的表现最好，LASPP1的平均值是0.455，文心一言模型的段落衔接指标最低，平均值是0.343。Claude生成的文本更连贯，衔接度更好。(4)在句法复杂度方面，Kimi的SYNLE值最高(7.031)，表明其生成的文本语法结构较为复杂，Claude模型的句法复杂度则最低，平均值仅为5.562。(5)在文本可读性方面，Claude模型的平均值得分最高，达到16.333，表明其生成的文本难度较大，而DeepSeek和文心一言的平均值得分比Claude低，分别为13.507和13.693，说明两者生成的文本更容易理解。

如果将六种大语言模型分成两大类，即国内模型（文心一言，Kimi，DeepSeek）和国外模型（ChatGPT，Claude，Gemini）两大类，我们可以得到国内与国外大语言模型在Coh-Metrix五种文本指标上的结果，见表5。根据表5的数据，对比国内与国外大语言模型在各项文本指标上的总体差异，见表6。

表5 国内与国外LLMs的五种语言特征指标统计结果

	分组	LDTTRa	LSASS1	LSAPP1	SYNLE	RDFKGL
1	国外(ChatGPT)	0.612	0.220	0.358	5.167	14.451
2	国外(ChatGPT)	0.628	0.156	0.294	7.158	13.561
3	国外(ChatGPT)	0.644	0.199	0.360	6.824	14.970
4	国外(ChatGPT)	0.598	0.218	0.299	5.412	12.252
5	国外(ChatGPT)	0.601	0.227	0.417	4.556	13.251
6	国外(ChatGPT)	0.659	0.167	0.280	5.333	12.757
7	国外(ChatGPT)	0.640	0.213	0.457	6.882	15.401
8	国外(ChatGPT)	0.670	0.203	0.331	7.588	14.938
9	国外(ChatGPT)	0.599	0.181	0.342	7.235	14.039
10	国外(Claude)	0.601	0.348	0.390	5.611	16.631
11	国外(Claude)	0.591	0.293	0.426	4.944	15.770
12	国外(Claude)	0.659	0.350	0.444	4.750	17.164
13	国外(Claude)	0.601	0.367	0.502	4.929	17.157
14	国外(Claude)	0.579	0.272	0.427	5.556	15.586
15	国外(Claude)	0.608	0.320	0.486	5.500	17.102
16	国外(Claude)	0.549	0.266	0.439	6.421	15.225
17	国外(Claude)	0.703	0.232	0.414	3.882	15.395
18	国外(Claude)	0.631	0.310	0.568	8.467	16.968
19	国外(Gemini)	0.604	0.189	0.349	6.789	13.651
20	国外(Gemini)	0.559	0.220	0.347	7.100	12.528
21	国外(Gemini)	0.590	0.279	0.385	4.294	14.443
22	国外(Gemini)	0.600	0.352	0.401	8.733	15.404
23	国外(Gemini)	0.582	0.225	0.360	3.813	14.174
24	国外(Gemini)	0.533	0.283	0.366	5.647	14.938
25	国外(Gemini)	0.582	0.254	0.400	6.353	15.338
26	国外(Gemini)	0.584	0.241	0.394	5.944	13.728
27	国外(Gemini)	0.608	0.150	0.284	6.700	12.759
28	国内(文心一言)	0.591	0.193	0.275	6.700	12.400
29	国内(文心一言)	0.584	0.234	0.325	5.316	14.689
30	国内(文心一言)	0.598	0.193	0.367	6.944	14.537
31	国内(文心一言)	0.576	0.195	0.322	5.632	12.378

续表5

	分组	LDTTRa	LSASS1	LSAPP1	SYNLE	RDFKGL
32	国内(文心一言)	0.573	0.176	0.313	5.950	12.847
33	国内(文心一言)	0.584	0.215	0.345	4.944	14.131
34	国内(文心一言)	0.564	0.231	0.393	6.000	13.511
35	国内(文心一言)	0.544	0.242	0.411	5.750	13.762
36	国内(文心一言)	0.586	0.194	0.333	7.050	14.980
37	国内(Kimi)	0.528	0.257	0.382	6.750	15.493
38	国内(Kimi)	0.610	0.236	0.401	6.800	15.199
39	国内(Kimi)	0.592	0.230	0.391	6.353	14.479
40	国内(Kimi)	0.540	0.211	0.360	7.444	13.648
41	国内(Kimi)	0.548	0.179	0.305	7.467	13.021
42	国内(Kimi)	0.547	0.204	0.410	8.625	14.205
43	国内(Kimi)	0.563	0.251	0.367	5.750	13.420
44	国内(Kimi)	0.552	0.173	0.323	6.533	12.789
45	国内(Kimi)	0.541	0.273	0.384	7.556	14.331
46	国内(DeepSeek)	0.615	0.225	0.419	5.875	13.507
47	国内(DeepSeek)	0.610	0.227	0.503	5.500	13.960
48	国内(DeepSeek)	0.616	0.251	0.375	5.056	13.680
49	国内(DeepSeek)	0.631	0.264	0.385	5.667	13.291
50	国内(DeepSeek)	0.616	0.283	0.356	6.533	13.546
51	国内(DeepSeek)	0.613	0.215	0.412	5.500	13.002
52	国内(DeepSeek)	0.661	0.201	0.375	5.789	14.436
53	国内(DeepSeek)	0.638	0.228	0.292	6.375	12.176
54	国内(DeepSeek)	0.634	0.240	0.447	6.167	13.967

表6　国内和国外大模型生成文本的语言特征指标描述性统计

指标	类别	样本	平均值	标准差	标准误差	最小值	最大值
LDTTRa	国内	27	0. 58722	0. 03510	0. 00675	0. 528	0. 661
	国外	27	0. 60796	0. 03754	0. 00722	0. 533	0. 703
LSASS1	国内	27	0. 22300	0. 02961	0. 005700	0. 173	0. 283
	国外	27	0. 24944	0. 06248	0. 01202	0. 150	0. 367
LSAPP1	国内	27	0. 36930	0. 05001	0. 00962	0. 275	0. 503
	国外	27	0. 38963	0. 06823	0. 01313	0. 280	0. 568
SYNLE	国内	27	6. 29726	0. 85752	0. 16503	4. 944	8. 625
	国外	27	5. 98474	1. 29077	0. 24841	3. 813	8. 733
RDFKGL	国内	27	13. 75500	0. 86467	0. 16641	12. 176	15. 493
	国外	27	14. 79930	1. 45826	0. 28064	12. 252	17. 164

在国内外LLMs生成的文本中，两类文本在五项指标上均存在一定程度的差异。结果表明，国外LLMs生成的文本在词汇多样性上表现更优，其中词汇多样性（LDTTRa）指标平均值为0.60796，优于国内文本（平均值为0.58722）。在句子衔接性指标（LSASS1）上，国外LLMs文本的优势更为明显，平均值达到0.24944，国内文本的平均值为0.223。在段落衔接性指标（LSAPP1）上，国外LLMs文本表现优于国内LLMs文本。从句法复杂度（SYNLE）指标来看，国内LLMs生成的文本句法复杂度平均值为6.29726，但国外大模型生成的句法复杂度标准差（SD=1.29077）比国内（SD=0.85752）模型要大。最后，国外模型在文本可读性指标上大于国内模型，国内外文本的对比分析进一步揭示了文本生成模型的来源差异性。

（二）单因素方差分析

为了进一步探讨ChatGPT、Claude、Gemini、文心一言、Kimi与DeepSeek六个大语言模型生成的英文文本的语言特征指标是否存在显著差异，本研究针对Coh-Metrix提取的五项文本指标（LSASS1、LSAPP1、LDTTRa、SYNLE、RDFKGL）分别进行了单因素方差分析（ANOVA），方差分析结果汇总见表7。

在方差分析前，首先需要进行球形假设检验（Mauchly's Test of Sphericity），检验结果显示：六个LLMs在词汇多样性、句子衔接性、段落衔接性、句法复杂度和文本可读性五个文本指标的球形假设检验结果的P值均大于0.05，即全部满足方差齐性要求，符合方差分析的条件（见表8）。

从单因素重复测量方差分析结果来看，六个LLMs在词汇多样性、句子衔接性、段落衔接性和文本可读性四个指标上的P值全部小于0.01的显著水平，只有句法复杂度指标上的差异没有达到显著水平（F=2.101，P=0.085>0.01）。分析结果表明ChatGPT、Claude、Gemini、文心一言、Kimi和DeepSeek六个语言大模型生成的英语文本除了在句法复杂度方面没有显著差异外，在其他四个指标上均存在极其显著差异（P<0.01）。

由于单因素重复测量方差分析结果表明，六个LLMs在词汇多样性、句子衔接性、段落衔接性和文本可读性四个指标上的P值全部小于0.01的显著水平，为了进一步分析六个模型在这四个文本指标上的具体差异，我们需要采用事后多重比较法对模型之间的差异进行显著性检验。

（三）Tukey事后检验

将六个LLMs的生成文本在词汇多样性、句子衔接性、段落衔接性和文本可读性四个指标的结果采用Tukey后续检验后，检验结果显示：（1）在词汇

表7　六种LLMs在五种文本指标的方差分析结果

变量	球形检验	III类平方和	自由度	均方	F	显著性（P值）
LDTTRa	满足假设	0.038	5	0.008	9.793	0.000
LSASS1	满足假设	0.066	5	0.013	9.364	0.000
LSAPP1	满足假设	0.077	5	0.015	6.013	0.000
SYNLE	满足假设	11.213	5	2.243	2.101	0.085
RDFKGL	满足假设	48.035	5	9.607	11.020	0.000

表8　六种LLMs在五种文本指标的球形假设检验结果

	主体内效应	Mauchly W	近似卡方	自由度	显著性（P值）
LDTTRa	a	0.043	19.178	14	0.187
LSASS1	a	0.068	16.420	14	0.324
LSAPP1	a	0.068	16.360	14	0.328
SYNLE	a	0.244	8.600	14	0.870
RDFKGL	a	0.144	11.812	14	0.652

多样性（LDTTRa）指标上，ChatGPT与DeepSeek均显著高于Gemini、文心一言和Kimi，P值均小于0.01；Claude与ChatGPT、DeepSeek之间无显著差异，但Claude显著高于Kimi。ChatGPT与DeepSeek之间无显著差异。说明ChatGPT与DeepSeek生成的英语文本具备更丰富的词汇分布，Kimi在词汇多样性方面相较于其他五个模型表现较弱。（2）在句子间衔接指标（LSASS1）方面，Claude显著高于其余五个模型，P值均小于等于0.01，其余五个模型之间的差异并不显著，表明在句子层面的连贯性上，Claude明显优于所有同类模型。（3）在段落间衔接（LSAPP1）指标上，Claude显著高于ChatGPT、Gemini、文心一言和Kimi。DeepSeek略低于Claude，但两者无显著差异，且DeepSeek与其他模型的差异未达到显著水平。在段落衔接层面，Claude保持着显著优势。（4）在文本可读性方面（RDFKGL），Claude与其余五个模型差异显著，P值均小于0.01，且Claude在此指标上的值显著高于其余五个模型，表明Claude生成文本阅读难度最大。

总体而言，Claude在句子衔接性、段落衔接性和文本可读性三个指标上显著高于多数同类模型。

（四）国内外LLMs文本指标差异的T检验

为了从整体上考察国内大语言模型和国外大语言模型在生成的文本指标之间是否存在差异，我们将六种LLMs分为国内（DeepSeek、文心一言、Kimi）和国外（ChatGPT、Claude、Gemini）两类模型，然后基于Coh-Metrix提取的语言特征指标，进行了独立样本T检验，检验结果见表9。

在进行T检验以前，首先需要进行方差齐性检验（Test of Homogeneity of Variance），并根据方差齐性假设是否满足分别以T值为依据进行决策。独立样本T检验结果显示，国内大语言模型和国外大语言模型只在词汇多样性（LDTTRa）（t=2.097，P=0.041<0.05）和文本可读性（RDFKGL）（t=3.201，P=0.003<0.05）两个指标上存在显著差异，其他指标上则不存在显著差异。

四　讨论

针对国内外六个大语言模型生成的文本语言学特征指标的方差分析结果显示，六个LLMs在词汇多样性、句子及段落的衔接性和文本可读性方面显著不同。这一研究结果与以往的试验结果吻合，验证了不同模型在训练方式、数据处理和算法架构方面的多样性。具体而言，在词汇多样性方面，DeepSeek表现突出，其生成的文本在词汇丰富性方面明显高于其他模型；而Kimi模型则表现相对较弱。Claude在句子与段落衔接性指标中也表现突出，这显示其生成的文本具有较高的语义连贯性和

表9　国内外LLMs文本指标的T检验结果

		方差齐性检验		均值T检验结果		
		F	Sig.	t	df	Sig.（双侧）
LDTTRa	满足方差齐性时	0.043	0.837	2.097	52.000	0.041
	不满足方差齐性时			2.097	51.767	0.041
LSASS1	满足方差齐性时	15.061	0.000	1.987	52.000	0.052
	不满足方差齐性时			1.987	37.116	0.054
LSAPP1	满足方差齐性时	1.990	0.164	1.249	52.000	0.217
	不满足方差齐性时			1.249	47.680	0.218
SYNLE	满足方差齐性时	5.455	0.023	-1.048	52.000	0.300
	不满足方差齐性时			-1.048	45.209	0.300
RDFKGL	满足方差齐性时	7.128	0.010	3.201	52.000	0.002
	不满足方差齐性时			3.201	42.271	0.003

衔接流畅性。Claude模型在文本可读性方面得分最高，表明其生成的文本语言难度最大。ChatGPT和DeepSeek模型的文本难度较低，生成的英语文本更易于理解。在句法复杂度方面，Kimi模型的平均得分相对最高，表明其生成的句法结构相对复杂，但六个大语言模型之间在句法复杂度上没有显著差异，说明六个LLMs生成的英语文本在句法复杂度上比较相似。

Tukey事后检验结果显示，Claude模型在句子衔接性和段落衔接性方面表现尤其突出，其相对优势可能归因于更为精准的语言结构学习和语义衔接机制。DeepSeek在词汇多样性方面的突出表现，则提示了国内模型在特定方面已具备一定的竞争优势，但在整体语言表现上仍需更加优化和提高。

另外，国内与国外大语言模型在Coh-Metrix文本指标的对比分析揭示出显著的差异。整体而言，国外模型在词汇多样性和文本可读性方面表现明显优于国内模型。这一结果可能与国外大语言模型英语训练语料规模更大、语言多样性更丰富有关[17]，使得其生成文本在词汇使用方面更加丰富和多样。国内模型在此指标上表现相对不足，可能需要进一步优化训练策略，增加语料库规模，以提升文本生成的词汇多样性和丰富性。在句子衔接性方面，国外模型表现也较为优异，这意味着国外模型可能更擅长捕捉句子之间的上下文语义关系，从而生成更加连贯的文本。然而，在段落衔接性和句法复杂度方面，国内外模型差异未达到显著水平，表明国内外大语言模型在段落结构的连贯性和句法复杂性方面的表现趋于一致[18]。这一结果可能反映出大语言模型在特定语篇结构和句法生成方面已达到了相对成熟的技术水平。

五　结论

本研究基于Coh-Metrix文本分析工具，通过方差分析和独立样本T检验法，探讨了六个国内外大语言模型（LLMs）生成的英语文本，在词汇多样性（LDTTRa）、句子衔接性（LSASS1）、段落衔接性（LSAPP1）、句法复杂度（SYNLE）和文本可读性（RDFKGL）五项文本指标上的表现差异，同时进一步探讨了国内与国外LLMs在上述文本特征上的整体表现差异，试验研究的结论如下。

首先，从六个LLMs之间的对比研究来看，在句法复杂度方面模型之间不存在显著差异以外，但六个LLMs在词汇多样性、句子衔接性、段落衔接性以及文本可读性方面均存在显著差异。

其次，从国内和国外两组LLMs生成的文本指标来看，国外大语言模型在词汇多样性和文本可读性两个方面差异显著，即三个国外大语言模型在这两项指标上明显优于国内三个大语言模型，但在句子衔接性、段落衔接性和句法复杂度三个文本指标上则无显著差异。

研究结果表明，国外大语言模型在英语文本生成方面具有一定的整体优势，但同时也揭示了国内大语言模型在生成文本的某些方面具有一定潜力和提升空间。未来研究可进一步探索如何通过优化国内大语言模型的训练数据与技术架构，提升其语言生成质量，从而更有效地满足语言学习者的多样化需求。同时，后续研究可以更多地关注大语言模型在实际教学场景中的应用与评估，以充分发挥人工智能技术在语言学习和写作教育领域的辅助作用。

参考文献：

[1]MINDNER L, SCHLIPPE T, SCHAAFF K. Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT [M]. International Conference on Artificial Intelligence in Education Technology. Singapore: Springer Nature Singapore, 2023: 152-170.

[2]苏祺，杨佳野.语言智能的演进及其在新文科中的应用探析[J].中国外语，2023，(3)：4-11.

[3]BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of Artificial General Intelligence: Early Experiments with GPT-4 [EB/OL]. (2023-3-22) [2023-4-13]. https://arxiv.org/abs/2303.12712.

[4]郭茜，冯瑞玲，华远方.ChatGPT在英语学术论文写作与教学中的应用及潜在问题[J].外语电化教学，2023(2)：18-23+107.

[5]吴琼.汉语二语者、母语者及ChatGPT生成记叙文写作质量和词汇复杂度对比研究[J].世界汉语教学，2024(4)：517-532.

[6]HERBOLD S, HAUTLI-JANISZ A, HEUER U, et al. AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays [EB/OL]. (2023-4-22). https://arxiv.org/abs/2304.14276.

[7]SMITH G, FUNK J. When It Comes to Critical Thinking, AI

Flunks the Test [EB/OL]. (2024-3-13). https://www.chronicle.com/article/when-it-comes-to-critical-thinking-ai-flunks-the-test.
[8]GRAESSER A C,MCNAMARA D S,CAI Z,CONLEY M,LI H,PENNEBAKER J.Coh-Metrix measures text characteristics at multiple levels of language and discourse[J].The Elementary School Journal,2014,115(2):210-229.
[9]MCNAMARA D S.Coh-Metrix:Capturing linguistic features of cohesion[J].Discourse Processes,2010(47):292-330.
[10]CROSSLEY S A, MCNAMARA D S. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication[J].Journal of Research in Reading, 2012,35(2):115-135.
[11]ULLMANN T D. Automated analysis of reflection in writing: validating machine learning approaches[J]. International Journal of Artificial Intelligence in Education,2019,29(2):217-257.
[12]SANTOS T. Professors' reactions to the academic writing of nonnative-speaking students[J].TESOL Quarterly, 1988, 22(1):69-90.
[13]ASTIKA G G. Analytical assessments of foreign students' writing[J].RELC Journal,1993,24(1):61-70.
[14]NATION P. Learning vocabulary in another language[M]. Cambridge:Cambridge University Press,2001.
[15]KYLE C,CROSSLEY S A.The relationship between lexical sophistication and independent- and source-based writing[J].Journal of Second Language Writing,2016,34:12-24.
[16]CROWHURST M. Syntactic complexity and teachers' quality ratings of narrations and arguments[J].Research in the Teaching of English,1980(14):223-231.
[17]BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[M]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook, NY:Curran Associates Inc.,2020:1877-1901.
[18]YANG K, KLEIN D, PENG N, TIAN Y. Improving long story coherence with detailed outline control[M]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 3378-3465.

【责任编辑　王　涛】

(上接第58页)

[17]HOARD J E.The new phonological paradigm[J].Glossa, 1971(5):222-268.
[18]端木三.对立、特征和发音动作[J].语言学论,2009(40):120-153.
[19]端木三.复合音和"无序原则"[J].语言科学,2018,17(1):1-17.
[20]林茂灿,颜景助.普通话轻声与轻重音[J].语言教学与研究,1990(3):88-104.
[21]WOO N.Prosody and phonology[D].Cambridge:Massachusetts Institute of Technology,1969.
[22]BOROWSKY T. Structure preservation and the syllable coda in English[J].Natural Language and Linguistic Theory,1989,7(2):145-166.
[23]DUANMU S.Syllable structure:The limits of variation[M]. Oxford:Oxford University Press,2008.
[24]DUANMU S.Rime length,stress,and association domains[J].Journal of East Asian Linguistics,1993,2(1):1-44.
[25]TURK A, NAKAI S, SUGAHARA M. Acoustic segment durations in prosodic research: A practical guide[M]// SUDHOFF S, LENERTOVA D, MEYER R, et al. Methods in empirical prosody research. Berlin: Mouton de Gruyter, 2006:1-27.
[26]林焘,王理嘉.语音学教程[M].北京:北京大学出版社,2013.
[27]LIU L(刘恋).Syllable structure in Standard Chinese:A durational perspective(从时长角度看普通话的音节结构)[D].成都:西南民族大学,2024.
[28]端木三.音位分析的"多解论"和最佳答案[J].语言科学,2019,18(2):113-131.
[29]马秋武,王平.何为底层形式?[J].当代语言学,2024,26(2):246-258.

【责任编辑　芮　芳】