基于大语言模型与对比学习的文本可读性自动分级方法研究*

储伊力[1,2]　曹振祥[3]　姜烨[4]　兰新苗[5]
（1.安徽医科大学创新创业学院，合肥 230032；2.公共健康社会治理安徽省哲学社会科学重点实验室，合肥 230032；3.安徽财经大学国际经济贸易学院，蚌埠 233030；4.合肥工业大学计算机与信息学院，合肥 230009；5.中国科学院光电技术研究所，成都 610209）

摘要：文本可读性是影响用户信息理解、吸收与利用的重要因素。本研究针对现有中文文本可读性自动分级方法在相邻等级区分中容易混淆、高质量标注数据不足等问题，提出了一种基于大语言模型与对比学习的文本可读性自动分级方法。首先利用大语言模型生成具有明确可读性梯度级别的对比样本以扩充训练数据，随后利用BERT模型提取文本的深层语义特征，并引入对比学习策略在特征空间中优化等级边界，以增强模型对相邻等级的判别力。实验结果显示，本研究模型的准确率（0.880）和F1值（0.881）均优于对比基线模型，说明该方法能有效提升中文文本可读性自动分级性能，尤其在相邻等级文本判别方面具有一定优势。

关键词：大语言模型；对比学习；可读性；文本信息；自动分级
中图分类号：TP391.1；G202　DOI：10.3772/j.issn.1673-2286.2026.04.005
引文格式：储伊力，曹振祥，姜烨，等. 基于大语言模型与对比学习的文本可读性自动分级方法研究[J]. 数字图书馆论坛，2026，22（4）：46-55.

在数字化时代，海量信息不断涌现，文本能否被目标受众准确理解，直接影响信息传播效果和知识获取效率。可读性差的文本不仅影响信息的有效传播和知识的普及，还可能导致用户对本应有用的信息产生误解、放弃阅读，从而加剧数字鸿沟。因此，有必要对文本信息进行可读性分级，为具有不同认知能力的用户筛选和提供适配的阅读材料。然而，国内汉语分级阅读研究起步较晚，目前仍处于基础研究阶段，缺乏权威的分级标准和成熟的文本可读性评级工具[1]。因此，研发有效的文本可读性评估技术，对于实现个性化知识服务、优化信息内容管理、提升信息素养，以及构建更具包容性的信息环境具有重要意义。

1　相关研究

已有研究表明，文本可读性是指文本内容被读者理解和吸收的难易程度[2]；文本可读性自动评估是指利用计算方法对文本难度进行定量化与自动化的评估和分析[3]。现有文本可读性自动评估的方法大致经历了可读性公式、传统机器学习、深度学习3个阶段。

1.1　可读性公式

可读性公式是文本可读性自动评估中形成较早且应用较广的一类经典方法，其基本思想是利用可量化

收稿日期：2025-12-09
*本研究得到安徽省哲学社会科学规划项目“安徽省老年人健康信息文本可读性评估研究”（编号：AHSKQ2022D141）资助。

的表层语言指标（如平均词长、句长、生僻词比例等）来估计文本难度。英文经典可读性测算以Flesch易读度（Flesch Reading Ease，FRE）[4]、Flesch-Kincaid年级指数[5]为最经典基准，辅以Dale-Chall[6]、SMOG指数[7]共同构成人工统计类可读性评价体系。FRE是将平均句长和单词平均音节数量纳入计算公式；Flesch-Kincaid年级指数是FRE配套衍生指标，与FRE成对使用，输出美国学龄年级数，是英文可读性普及的传统指标之一；Dale-Chall可读性公式主要以非常用词比例反映词汇难度，其计算变量分别是难词占比和平均句长，其中难词是指不在美国小学4年级学生普遍认识的3 000个高频基础词构成的常用词表中的词；SMOG指数侧重多音节难词数量衡量难度，是欧美医疗说明书、公共卫生法定文本可读性标准。这些公式因计算简便、结果直观和可解释性强长期被学界广泛运用。

由于汉语在字、词、句法等方面与英语有本质差异，直接套用基于英语语言学特征的公式难以准确衡量中文文本的可读性。因此，国内学者普遍基于以上经典可读性公式进行本土化的改进，如荆溪昱[8]通过引入汉字的字词难度指标，构建了适用于中文文本的可读性公式；张宁志[9]则进一步整合了句法复杂度、词汇特征、功能项目及文化背景知识等综合指标，完善了中文文本可读性的定量分析方法。

1.2　传统机器学习

在自然语言处理领域，文本可读性自动分级可以被看作是一个文本分类任务。20世纪90年代初至21世纪10年代初，传统机器学习技术被广泛引入文本分类研究中，彼时的文本分类模型多基于浅层机器学习，如线性回归、朴素贝叶斯、决策树、支持向量机等。文本可读性领域的研究因而开始突破可读性公式的局限，采用结合多维语言学特征的机器学习模型，显著提升了评估的灵活性与准确性。该方法通常遵循特征提取、特征聚合、分类建模的流程：首先基于语言学知识人工设计可读性相关指标，并借助自然语言处理工具或基于规则、统计的方法从文本中提取相应特征值，即人工提取语言学特征；其次通过主成分分析等方法实现特征降维与去冗余；最后基于处理后的特征，利用机器学习模型完成可读性等级预测。

在非中文文本研究中，Bhogayata[10]针对古吉拉特语文本，利用单语词性标注器和Python进行特征提取，随后基于这些特征利用向后逐步多元线性回归算法构建了可读性评估模型。此外，也有研究聚焦于可读性特征本身的提取与聚合，如Crossley[11]利用SEANCE（Sentiment Analysis and Cognition Engine）、TAACO（Tool for the Automatic Analysis of Cohesion）和TAALED（Tool for the Automatic Analysis of Lexical Diversity）等自然语言处理工具提取了情感、认知、句法和词汇多样性等特征，并运用主成分分析法将这些特征聚合为数个主要的语言构件，这些构件为可读性评估提供了更具解释性的特征依据。需要说明的是，不同语种在文本切分方式和特征提取上存在差异。英文等文本通常可基于空格分词后的词级统计量等语言学特征构建模型，而中文文本因缺乏显式的词边界，研究者需借助分词工具（如jieba）完成预处理，并结合字、词、句法和篇章等汉语语言学特征构建可读性模型。在中文文本可读性研究中，传统机器学习方法主要体现为面向中文语言特点的特征构建与分类算法应用，即在提取字、词、句法和篇章等语言学特征的基础上，利用多元线性模型[12–13]、支持向量机[14–15]、随机森林[1]等完成可读性分类或等级预测。总体来看，上述研究虽在特征选取和算法选择上存在差异，但均属于传统机器学习范式，差异主要体现在不同语种的语言特征适配和模型的选择上。

1.3　深度学习

自2010年以来，文本分类模型逐渐从依赖人工特征的浅层学习转向强调深度语义挖掘的深度学习。其中，卷积神经网络（Convolutional Neural Network，CNN）擅长捕捉词语或短语层面的局部语言模式；循环神经网络（Recurrent Neural Network，RNN）及其变体能够建模词语之间的顺序关系和上下文依赖；Transformer架构及BERT等预训练语言模型则通过自注意力机制和大规模语料预训练，进一步提升了文本语义表征能力。在此背景下，文本可读性评估作为文本分类的一个分支，也逐渐从基于规则和人工提取的语言学特征的计算转向基于深度语义理解的自动分级。

在非中文文本可读性研究中，如何融合深层语义表征与显式语言知识是可读性自动分级领域中的重要研究方向。如Lee等[16]将BERT、RoBERTa等预训练模型的软标签预测与传统语言学特征结合，所构建的分

级模型在多个数据集上取得较优性能，尤其在小规模数据场景中表现突出。为增强模型对文本多层级结构的建模能力，一些学者尝试在架构设计层面上进行创新。例如，Meng等[17]提出的ReadNet分层Transformer框架，Jian等[18]构建的CNN-BiLSTM-Attention混合网络模型，分别从层级结构建模和多类神经网络特征融合角度提升了文本可读性预测效果。近年来的研究开始尝试超越传统的“预训练+微调”的范式，探索通过提示学习等方法提升模型的任务适配能力与泛化能力。Lee等[19]探索了提示学习在可读性评估中的应用，通过精心设计提示模板激发了预训练模型的任务理解能力，提升了模型的跨数据集泛化能力。中文文本可读性研究在跟进深度学习前沿研究的同时，更加强调模型对中文语言特点的适配性。例如，程勇等[20]构建了多元语言特征与BERT的特征级融合框架；谭可人等[21]则进一步将多层级语言特征深度注入BERT自注意力机制，验证了多源特征融合机制在中文语境下的可行性。此外，也有研究构建混合神经网络架构，以兼顾局部语言特征与上下文依赖关系。例如，李文彪等[22]提出的CNN-LSTM模型，通过CNN提取局部语言特征，并利用LSTM捕捉长距离上下文依赖关系，从而提升了中文文本可读性模型的评估效果。

虽然近年来深度学习模型逐渐成为主流，但由于可读性任务中人工提取的语言学特征具有可解释性优势，仍有很多学者继续采用可读性公式法和传统机器学习模型开展文本可读性研究。已有研究在可读性评估研究的语言学特征构建、模型选择和评价指标设置等方面为本研究提供了参考，但仍存在以下改进空间：①可读性公式和传统机器学习具有较好的可解释性，且使用方便，但依赖人工提取的表层语言学特征，难以充分捕捉深层语义和篇章结构信息；②现有的深度学习模型虽评估效果更好，但中文语境下的高质量分级语料有限，模型在样本稀缺的场景下准确率受限；③现有模型多依赖传统特征或单一深度学习模型，对不同文本信息间的渐进式可读性差异不够敏感，容易出现分级边界模糊的问题。

针对高质量标注文本稀缺的挑战，大语言模型（Large Language Model，LLM）在文本生成和语义理解方面的强大性能为本研究任务提供了新的技术契机，因此本研究参考蒋超等[23]的研究，利用LLM可控生成不同可读性级别的文本，有效扩充了训练数据。针对相邻等级文本差异较小不易区分的挑战，本研究参考Radford等[24]的研究，引入基于InfoNCE损失的对比学习（Contrastive Learning，CL）策略，通过构造正负样本对并在语义空间中进行相似度度量，使模型更容易区分不同可读性级别，形成较为清晰稳健的分级边界，实现更精准的文本可读性自动分级效果。

2 模型架构

本研究提出基于大语言模型与对比学习的可读性自动分级模型（Large Language Model-based Contrastive Learning for Readability Assessment，LLM-CL）。该模型旨在通过LLM、预训练语言模型及对比学习策略，实现更精准的文本可读性自动分级效果。LLM-CL模型由3个核心部分构成，三者协同工作以完成可读性自动分级任务，分别是文本生成模块、文本向量构建模块和对比学习模块，LLM-CL模型整体架构如图1所示。

LLM-CL模型由文本生成模块、文本向量构建模块和对比学习分类模块组成：文本生成模块利用Qwen生成具有可读性梯度的增强文本；文本向量构建模块分别构建原始文本向量和对比文本向量，包括BERT语义编码、语言学特征提取与特征拼接；对比学习分类模块以两类文本向量为输入，计算向量间相似度并构建相似度矩阵，再经归一化处理输出各可读性级别的预测结果，从而完成文本可读性自动分级。

2.1 文本生成模块

该模块是LLM-CL模型中负责数据增强的功能模块，旨在缓解可读性自动分级任务中高质量标注数据不足的问题。具体而言，本研究借助LLM的强大的生成能力，对原始文本进行多目标等级的改写，从而为后续的对比学习策略提供丰富、多样化且具有明确可读性梯度的训练样本。改写文本的生成过程为：先输入原始文本，并为其依次设定不同的目标可读性级别；再通过向LLM提供结构化提示词（Prompt），引导LLM根据给定的原始文本生成满足不同目标可读性级别的新文本。对于每条原始文本，本研究共生成5个不同可读性级别的改写版本，从而使生成的样本规模达到原始样本规模的5倍。

为保证文本生成过程的可控性和可复现性，本研究采用结构化提示词（见表1）引导LLM进行改写。具

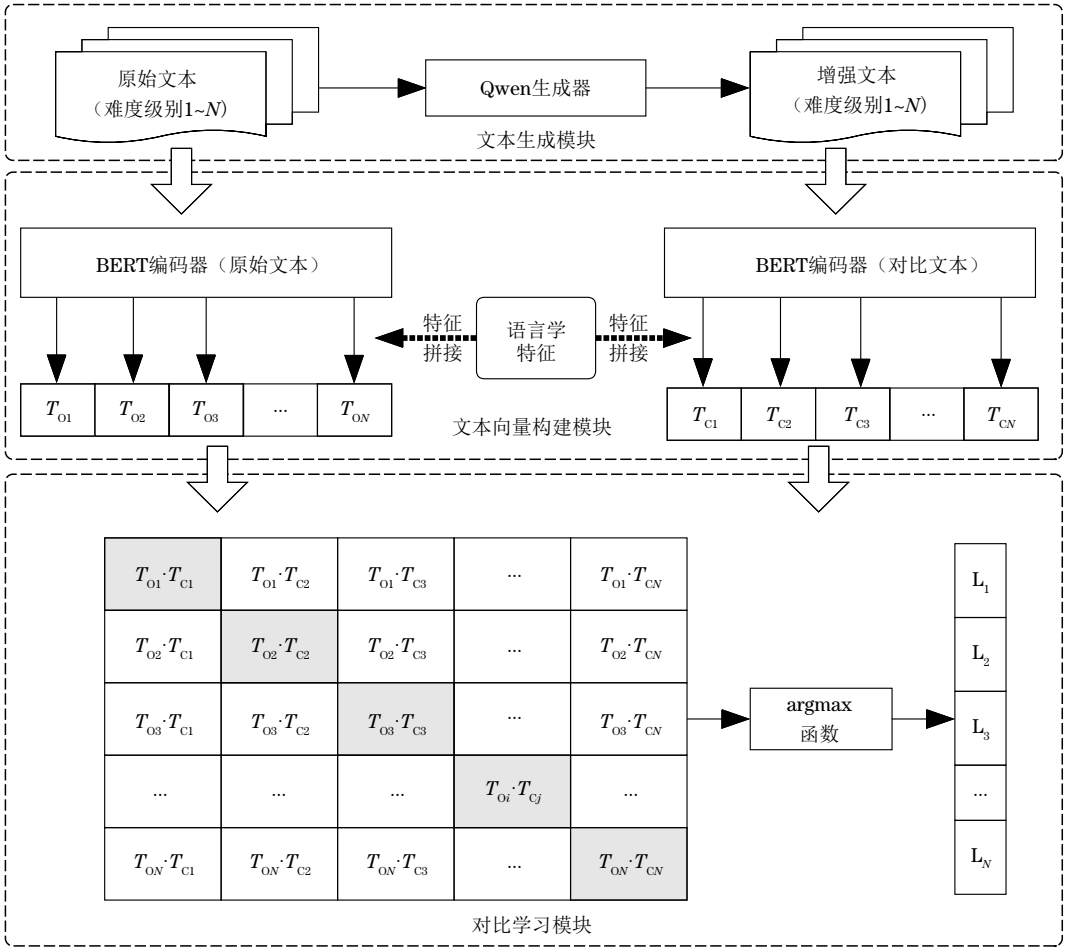


图1　LLM-CL模型架构

表1　提示词设计

模块名称	结构化提示词内容
角色设定	你是一名具有专业语文写作能力的文本改写助手
任务说明	现提供一篇目标阅读等级为[{target_grade}]的参考文章：{reference}。请结合该等级文本的语言特点，将以下当前阅读等级为[{current_grade}]的原始文本改写为适合目标阅读等级[{target_grade}]的版本：{original_text}
改写要求	①语义保持：保持原文核心含义不变，严禁添加原文中不存在的新事实、新观点或新数据。②复杂度调节：根据目标等级严格调节字难度、词难度与句法复杂度。③实体保护：不得更改或删除专有名词、人名、地名、数字及公式等关键信息。④风格参照：参照参考文章的语言风格，但不可直接使用参考文章中的具体内容
输出规范	①格式要求：仅输出改写后的文本，不输出任何解释性说明。②长度控制：输出长度应控制在原文的0.5~2倍。③保留原文段落边界

体而言，该结构化提示词通过角色设定明确模型的任务身份，通过任务说明告知模型目标等级的参考文本、原始文本及改写目标，通过改写要求明确改写边界，并通过输出规范约束生成结果的形式规范。

考虑到LLM生成结果具有一定随机性，为进一步量化生成文本质量，本研究采用规则检查与人工抽检相结合的方式对生成文本进行质量评估。对于格式类问题，首先依据提示词中的输出规范进行规则检查，并对可通过规则修正的格式问题进行规范化处理；其次，在规则检查的基础上，进一步采用人工抽检的方式，从语义保持、格式规范和等级匹配3个维度对抽检样本进行综合评估。具体而言，本研究采用分层随机抽样的方法，从生成文本中抽取629篇进行人工复核，抽检占比约为5%。其中，语义保持主要考查是否存在新增、遗漏或改变关键信息、含义等问题；格式规范主要考查是否满足长度控制、段落边界保留和仅输出改写文本等要求；等级匹配主要考查字词难度、句法复杂度和表达方式等是否基本符合目标可读性

级别。结果显示，语义保持、格式规范和等级匹配3个维度的合格率分别为98.57%、100.00%和98.89%。上述结果表明，该设计能够较好地保证生成文本的适用性。

随后，生成后的文本将进一步用于构建对比学习样本对。若生成文本与原始文本属于同一可读性级别，则将其视为正样本，否则为负样本。通过这种方式，本研究构建了具有明确可读性级别差异的对比学习样本，为后续模型训练提供了更为丰富的监督信号。

2.2 文本向量构建模块

该模块运用BERT提取所有文本信息（包括原始文本信息和对比文本信息）的深层语义特征并融入可能增强模型性能的人工选取的语言学特征。BERT基于大规模语料库上的无监督预训练任务和多头自注意力机制，能学习到丰富的语言知识和复杂的上下文语境信息。通过输入嵌入层、多层Transformer编码器进行逐层转换后，本研究提取最后一层输出中[CLS]标记对应的隐藏层向量，作为该文本的深层语义表示。

现有研究中，大多数研究采用人工提取的方式获取文本语言学特征。为了探索表层语言学知识的潜在贡献，本研究将人工提取的文本语言学特征（见表2）与BERT编码器输出的最后一层[CLS]向量进行拼接，形成最终的文本向量。

表2 文本语言学特征

特征类别	符号表示	计算规则
文本结构特征	S_1；S_2	S_1：文本总字数。S_2：句子总数量
字符复杂度特征	C_1~C_3	C_1：字符类型多样性。 C_2：生僻字数量。C_3：常用字占比
词汇复杂度特征	V_1~V_8	V_1：词汇总量。V_2：词汇丰富度。 V_3~V_6：甲级、乙级、丙级、丁级词汇占总词量的比例。V_7：词汇复杂度。 V_8：词型数量
句法复杂度特征	G_1~G_4	G_1：句长变异系数。G_2：简单句比例。 G_3：相邻句重复率。G_4：平均句长

2.3 对比学习分类模块

该模块是LLM-CL的核心训练策略。其目标是最大化相同可读性级别文本之间的相似度，同时最小化不同可读性级别文本之间的相似度，从而构建一个具有高区分度的文本可读性自动分级模型。该模块主要包含以下关键步骤实现。

首先，构建用来对比的正样本和负样本。从上述文本特征提取模块获取原始文本向量（$\boldsymbol{T}_{\mathrm{O}}$）和对比文本向量（$\boldsymbol{T}_{\mathrm{C}}$）。

其次，构造相似度矩阵。本研究采用余弦相似度来衡量不同文本向量之间的相似程度。相似度得分S_{ij}的计算公式如式（1）所示。

$$S_{ij}=\frac{\boldsymbol{T}_{\mathrm{O}i}\cdot\boldsymbol{T}_{\mathrm{C}j}}{\|\boldsymbol{T}_{\mathrm{O}i}\|\|\boldsymbol{T}_{\mathrm{C}j}\|} \tag{1}$$

式中：$\boldsymbol{T}_{\mathrm{O}i}$为第$i$可读性级别的原始文本向量，$\boldsymbol{T}_{\mathrm{C}j}$为第$j$个可读性级别的对比文本向量。其中，$i=j$为正样本，$i\neq j$为负样本。所有相似度值$S_{ij}$按行列排布，共同构成了$N\times N$维的类别相似度矩阵。

最后，计算InfoNCE损失。模型运用InfoNCE损失函数对分类矩阵对角线元素进行优化，目标是最大化对角线元素（相同级别间的相似度），同时最小化非对角线元素（不同级别间的相似度）。单个样本i的损失项L_i计算方法如公式（2）所示。

$$L_i=-\frac{1}{|P(i)|}\sum_{p\in P(i)}\ln\left(\frac{\exp(S_{ip}/\tau)}{\sum_{a\in P(i)\cup N(i)}\exp(S_{ia}/\tau)}\right) \tag{2}$$

式中：$P(i)$为所有正样本集合，$N(i)$为所有负样本集合，S_{ip}表示第i个锚点样本（原始文本向量）与集合$P(i)$中的第p个正样本（同级别的对比文本向量）之间的相似度，S_{ia}表示第i个锚点样本与集合$P(i)\cup N(i)$中的任意样本a之间的相似度，τ是温度系数。模型参数通过反向传播和优化器进行更新迭代。

通过上述步骤，对比学习模块能够增强模型对相邻可读性级别文本的区分能力，为最终的可读性自动分级任务奠定了强有力的基础。

3 实验与结果分析

3.1 数据来源

语文教材是根据国家课程标准和学生认知发展规律，为不同成长阶段的学生精心编写的，其文本在词汇、句法、篇章组织和主题深度等方面通常呈现循序渐进的难度变化。因此，教材的年级/学段本身就构成了直接且权威的可读性分级标签，为构建文本可读性分级模型提供了可操作的标注数据。基于此，同多数相关研究一样，本研究选用母语为汉语的语文课文作为中

文文本可读性分级数据集。参考《语文课程标准》的年级和学段划分方式，将文本可读性划分为5个等级：小学1~2年级为1级，小学3~4年级为2级，小学5~6年级为3级，初中为4级，高中为5级，每篇文本对应唯一标签。

数据集来源方面，本研究从公开网络资源上搜集了人教版、苏教版、北师大版、湘教版、鄂教版等不同版本教材的语文课文文本，剔除文言文、诗歌、古文和古诗词，同时过滤掉了不同版本间的重复课文，经过处理，最终采集文章2 515篇，文本数据集统计情况如表3所示。

表3　文本数据集统计情况

可读性级别	文章数量/篇	总字符数/个	总句子数/句	篇均字符数/个	篇均句子数/句
1级	501	141 993	6 012	283.419	12.000
2级	696	448 818	16 338	644.853	23.474
3级	698	571 309	19 921	818.494	28.540
4级	501	1 139 814	34 273	2 275.078	68.409
5级	119	505 297	14 194	4 246.193	119.277

数据集的划分策略：先将原始数据集按照9∶1划分成初始训练集和最终测试集，再将该初始训练集进一步按照9∶1拆分为训练集和验证集，以进行模型训练和超参数调优。需要强调的是，本研究所有的模型训练和验证仅在训练集和验证集上进行。测试集始终保持其原始形式，不参与任何训练和验证过程，以确保模型评估的公正性和可靠性。

3.2　实验平台与参数设置

本研究的实验环境包括软件环境和硬件环境。软件环境方面，操作系统为Ubuntu 20.04，开发语言为Python 3.9，环境管理工具为Anaconda，深度学习框架为PyTorch 2.1.2，GPU加速组件为CUDA 11.8。硬件环境方面，实验服务器配置了第13代Intel Core i9-13900K处理器（主频5.30 GHz）、NVIDIA H20高性能图形处理器（GPU）和128 GB的运行内存。NVIDIA H20 GPU可支持混合精度训练（FP16），有效提升了训练效率并降低了显存消耗。

参数设置方面，采用LLM生成具有可读性梯度的增强文本，其推理过程部署于上述硬件环境。预训练模型选用BERT-base-Chinese，BERT模型架构中的嵌入层维度为768，隐藏层维度为768，Transformer堆叠层数为12，激活函数为GELU。

3.3　评价指标

为全面衡量所提出模型的性能，笔者选取了国内外可读性评估领域常用的3个评价指标，即准确率、精准率和F1值。其中，本研究主要采用准确率和F1值作为衡量模型整体性能的关键指标。此外，本研究选择混淆矩阵作为辅助评估工具，将直观展示模型在不同可读性级别上的分类错误情况，有助于识别模型在哪些级别上容易混淆，为后续的模型优化提供具体方向。

3.4　对比模型

为验证本研究方法的有效性，本研究选取3类模型进行对比。第一类为经典神经网络模型，包括CNN、LSTM、RNN和TextRCNN，用于考查非预训练模型对文本序列特征和局部语言特征的建模能力。第二类为预训练语言模型，包括BERT-base-Chinese、MLF-BERT、RoBERTa-wwm-ext-large和Chinese-MacBERT-large，用于比较主流预训练模型在可读性自动分级任务中的表现。第三类为LLM-CL模型，包括以Qwen2.5、DeepSeek-V3.2和GLM-5作为生成基座的LLM-CL扩展模型。该类模型均采用本研究提出的“LLM+对比学习”框架，区别仅在于生成阶段所使用的LLM基座不同，用于分析不同生成基座与对比学习策略结合后对本研究框架性能的影响。

3.5　实验结果与分析

3.5.1　模型性能对比实验

从文本可读性分级模型对比实验结果（见表4）可以看出，本研究提出的LLM-CL模型整体性能优于经典神经网络模型和预训练语言模型。其中，以Qwen2.5为底座的LLM-CL模型在3个指标上取得最佳结果，准确率、精准率和F1值分别达到0.880、0.887和0.881，说明本研究方法能有效提升中文文本可读性自动分级效果。

（1）经典神经网络模型。CNN、LSTM、RNN和TextRCNN等深度学习方法的性能普遍低于LLM-CL模型。这些结果表明，在可读性评估这类需要深层语言理解的任务中，仅依靠浅层神经网络难以捕捉到足够的语义特征。

表4　文本可读性分级模型对比实验结果

模型类型	模型名称	准确率	精准率	F1值
经典神经网络模型	CNN	0.619	0.622	0.591
	LSTM	0.487	0.490	0.373
	RNN	0.473	0.232	0.312
	TextRCNN	0.592	0.597	0.575
预训练语言模型	BERT-base-Chinese	0.540	0.595	0.550
	MLF-BERT	0.544	0.579	0.548
	RoBERTa-wwm-ext-large	0.550	0.596	0.557
	Chinese-MacBERT-large	0.591	0.615	0.590
LLM-CL模型	Qwen2.5	0.880	0.887	0.881
	Qwen2.5（融入语言学特征）	0.784	0.829	0.799
	DeepSeek-V3.2	0.868	0.872	0.869
	DeepSeek-V3.2（融入语言学特征）	0.792	0.815	0.801
	GLM-5	0.805	0.833	0.811
	GLM-5（融入语言学特征）	0.748	0.772	0.756

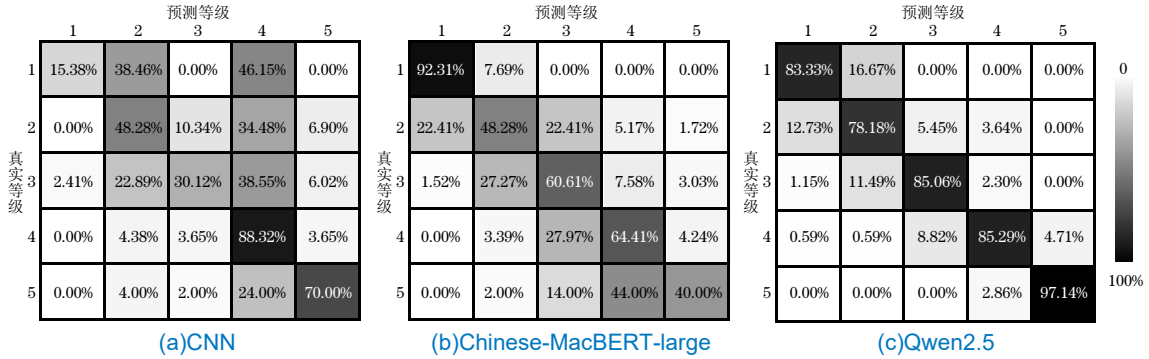
（2）预训练语言模型。基于该类模型搭建的可读性评价模型准确率偏低，原因在于标注数据稀缺时，仅通过监督微调难以充分发挥预训练语言模型的深层语义理解潜力。

（3）LLM-CL模型。在可读性自动分级这类抽象且需要深度理解的任务上，此类模型在各项指标上均表现较优，说明LLM生成的多等级改写文本能够为对比学习提供有效的训练样本，对比学习策略也有助于增强模型对不同可读性级别文本的判别能力。在不同LLM底座中，Qwen2.5表现最佳，因此本研究后续实验将以Qwen2.5作为代表模型开展混淆矩阵分析、消融实验、温度系数实验和案例分析。

（4）语言学特征对模型的影响。鉴于对比学习类模型在基准测试中展现出的性能优势，本研究进一步以此类模型为实验主体，分析语言学特征对模型影响。实验结果显示，在多数模型中，简单拼接浅层语言学特征并未提升模型性能，反而造成了性能下降。这可能是因为预训练语言模型已从大规模语料中学习到丰富的语言学和语义信息，额外拼接的表层特征难以提供有效增益，甚至可能引入冗余或噪声。该现象说明：人工特征融合需要更精细的选取，简单拼接并不一定适用于深度语义模型；LLM-CL模型在不依赖复杂人工特征融合的情况下仍能取得较优性能。

3.5.2　混淆矩阵分析

为进一步分析模型在不同可读性级别上的分类表现，本研究选取经典神经网络模型中表现较好的CNN，预训练语言模型中表现较好的Chinese-MacBERT-large，以及此次实验中效果最优的Qwen2.5模型绘制混淆矩阵。混淆矩阵依据真实等级按行进行归一化，横轴表示模型预测等级，纵轴表示文本可读性的真实等级，矩阵对角线数值为各真实等级样本的正确预测占比。3种模型在文本可读性分级任务上的行归一化混淆矩阵如图2所示。



(a)CNN　(b)Chinese-MacBERT-large　(c)Qwen2.5

图2　3种模型在文本可读性分级任务上的行归一化混淆矩阵

由图2可知，CNN模型整体分类能力有限，尤其在低等级文本上误判较多，说明经典神经网络模型难以充分捕捉可读性等级之间的细微差异。Chinese-MacBERT-large能较好地区分文本的大致难度区间，较少出现跨多个等级的严重误判，但在相邻等级之间仍存在明显混淆，表明单纯依赖预训练语言模型微调仍难以准确刻画相邻可读性级别的边界，这也正是本研究提出的LLM-CL模型致力于解决的核心痛点。相比之下，Qwen2.5模型在各级别上的对角线数值整体更高，非相邻等级误判明显减少，相邻等级之间的混

淆也得到缓解，说明它能学习到更具区分性的可读性表示。混淆矩阵的结果进一步验证了本研究提出的评估方法在提升相邻可读性级别判别能力方面的有效性。

3.5.3　消融实验

为进一步探究LLM生成和CL两个关键模块的必要性，本研究开展了2项消融实验。考虑到模型各模块之间存在依赖关系，无法简单地直接剔除某个部分，本研究分别以基于同义词替换和标点符号扰动的规则式数据增强方案替代LLM生成方案（模型1），以交叉熵损失替代InfoNCE损失（模型2），从而考查两个关键模块对模型性能的影响。消融实验结果如表5所示。

从表5可以看出，完整LLM-CL模型取得最佳性能，准确率和F1值分别为0.880和0.881。将LLM生成模块替换为传统文本生成方案后，模型准确率由0.880降至0.483，F1值0.881降至0.518，说明LLM生成的多级别改写文本能够为模型训练提供更有效的增强样本。将InfoNCE损失替换为交叉熵损失后，模型准确率由0.880降至0.804，F1值由0.881降至0.809，表明对比学习策略有助于提升不同可读性级别文本表示的判别能力。消融实验结果表明，LLM生成模块和CL模块并非简单的功能叠加，而是相互协同、缺一不可的核心模块。LLM为对比学习提供了高质量的训练数据，而对比学习则充分利用这些数据，学习到更具判别力的文本表示，两者共同推动了模型性能的提升。

表5　消融实验结果

模型	准确率	精准率	F1值
完整模型（LLM-CL）	0.880	0.887	0.881
模型1（消融LLM生成模块）	0.483	0.578	0.518
模型2（消融CL模块）	0.804	0.825	0.809

3.5.4　温度系数实验

温度系数是对比学习中的重要超参数，会影响InfoNCE损失中正负样本相似度分布的平滑程度。为考查温度系数对模型性能的影响，本研究在0.050~0.100范围内进行了参数实验。结果表明，模型准确率随温度系数增大呈现先升后降的趋势，并在0.080时达到最优。温度系数过小时，模型对样本间细微差异过于敏感；温度系数过大时，相似度分布过于平滑，可能削弱特征表示的判别性。因此，本研究最终将温度系数设置为0.080，此时模型能达到最佳性能。

3.5.5　错误案例分析

为了进一步揭示LLM-CL模型的可读性自动分级依据及误判原因，本研究从测试集中选取了涵盖不同学段、文体及复杂度的代表性样本进行定性分析，代表性错误案例如表6所示。

表6　LLM-CL可读性分级错误案例

序号	文本片段示例	真实等级	预测等级
1	陈赓同志回顾自己革命经历的时候……那小家伙不过十一二岁，黄黄的小脸，一双大眼睛……小红军倔强地说：“你要我同你的马比赛啊，那就比一比吧。”他说着把腰一挺，做出个准备赛跑的姿势	2级	3级
2	在缅甸的热带丛林里，高达十几米的树顶上，两只叶猴跳荡着、嬉闹着……它们飞上坡顶，像小孩坐滑梯一样一只挨着一只滑雪而下。研究动物行为的科学家，把它们分成三种最基本的类型	5级	4级
3	她问老奶奶：“大妈，您这磅……”“我这磅准让您高兴！……别巴望着孩子长！孩子长大了，当妈的就老喽！”老奶奶把嘴咂得吧吧响	4级	3级

错误案例分析表明，模型预测偏差主要与3个方面因素有关。①标签的主观性与模型客观性的冲突。我国语文教材的年级、学段划分除语言复杂度外，还受到教育目标、审美价值和选文功能等非语言因素影响。如某些经典文章虽语言修辞丰富、认知负荷较高，但出于育人目的被放入较低学段，而模型判别主要基于客观视角下文本的认知难度，导致模型高估其难度，如错误案例1。②文体特征会影响模型判别。实验发现模型存在文学叙事易高估、科普说明文易低估的情形。一些科普文即便涉及复杂科学知识，但语言平实、句式清晰，可能导致模型低估其难度，如错误案例2；反之，文学作品即便情节具象，但若涉及丰富的描写、密集的修辞时，模型则易高估其难度。③口语化表达对隐性深度语义识别的干扰。模型对显性词汇出现频率的敏感度较高，因而容易被大量表层的简单词汇所干扰，将其判定为低难度级别，忽略了文本背后蕴含的隐喻、讽刺或深层心理动机等，如错误案例3。

总体来看，LLM-CL模型能捕捉到文本可读性递

进的内在逻辑，但在教育标签偏差、文体特征和隐性语义理解等方面仍有改进空间。

4 结语

本研究提出了一种基于LLM与对比学习的文本可读性自动分级模型LLM-CL，并基于多版本语文教材文本开展实验。实验结果表明，LLM-CL模型在不依赖复杂人工特征融合的情况下取得了最优性能。这为中文文本可读性评估提供了一个较为高效的解决方案，也为数据稀缺场景下的文本分类任务提供了可行参考。尽管本研究方法取得了一定进展，但仍有广阔的探索空间。未来可从3个方向进行深入探究。

（1）提升模型对相邻等级和隐性语义的区分能力。尽管LLM-CL性能整体表现较好，但混淆矩阵结果揭示，模型在相邻可读性级别之间仍存在一定混淆；同时正反案例分析显示，模型在处理表层语言复杂度与深层语义内涵不完全一致的文本时，仍可能出现等级偏差。未来研究可通过赋予模型更多的先验知识和引入对比学习的硬负样本挖掘技术等方式，进一步强化模型穿透表层语言特征、捕捉文本深层内涵的感知能力。

（2）改进人工语言学特征与深度语义特征的融合机制。针对实验中简单拼接人工语言学特征未能稳定提升模型性能的问题，后续研究可探索以下3种不同的融合策略：①引入动态门控机制，根据文本特征来动态调节人工特征与语义特征的权重分配；②尝试非线性投影对齐，将低维人工特征映射至高维语义空间再进行深度交互；③探索多任务学习架构，将语言学指标转化为辅助监督信号，实现人工先验知识对大模型语义学习的引导，而非简单地直接拼接。

（3）构建更高质量、多领域的可读性评估语料。目前可读性研究多直接以语文教材对应的年级作为分级标签。然而，教材年级标签不仅受语言难度影响，也可能受到教学目标、审美价值和选文功能等因素影响，未必完全等同于文本的实际阅读难度。同时，真实应用场景中的文本类型更加复杂，如医学报告、科技文献、新闻报道等，其术语密度、句法结构、篇章组织和表达风格均可能与教材文本存在差异。因此，本研究提出的模型在教材文本上的良好表现，并不能直接说明其在开放域文本中同样适用。未来研究可构建覆盖多领域、并结合专家评定和读者认知实验的可读性语料库，进一步通过眼动追踪、阅读理解测试等方式量化读者的真实阅读负荷，并开展跨领域测试、领域自适应或少样本微调研究，以提升模型在真实应用场景中的稳定性和适用性。

参考文献

[1] 杨智渤. 基于机器学习的汉语儿童阅读材料可读性评估方法研究[J]. 情报科学，2025，43（3）：91–98.

[2] 曾锦山，彭香龙，肖文艳，等. 基于信息增益与层次分类的中文文本可读性评估方法[J]. 中文信息学报，2025，39（10）：167–177.

[3] 吴思远，蔡建永，于东，等. 文本可读性的自动分析研究综述[J]. 中文信息学报，2018，32（12）：1–10.

[4] FLESCH R. A new readability yardstick[J]. Journal of Applied Psychology，1948，32（3）：221–233.

[5] KINCAID J P，FISHBURNE R P，ROGERS R L，et al. Derivation of new readability formulas for navy enlisted personnel[R]. Millington：Naval Technical Training Command，1975.

[6] DALE E，CHALL J S. A formula for predicting readability[J]. Educational Research Bulletin，1948，27（1）：11–20.

[7] MCLAUGHLIN G H. SMOG grading: a new readability formula[J]. Journal of Reading，1969，12（8）：639–646.

[8] 荆溪昱. 中文国文教材的适读性研究：适读年级值的推估[J]. 教育研究资讯，1995，3（3）：113–127.

[9] 张宁志. 汉语教材语料难度的定量分析[J]. 世界汉语教学，2000（3）：83–88.

[10] BHOGAYATA C K. A machine learning-based readability model for Gujarati texts[J]. ACM Transactions on Asian and Low-Resource Language Information Processing，2024，23（2）：1–32.

[11] CROSSLEY S A. Developing linguistic constructs of text readability using natural language processing[J]. Scientific Studies of Reading，2025，29（2）：138–160.

[12] 孙刚. 基于线性回归的中文文本可读性预测方法研究[D]. 南京：南京大学，2015.

[13] 朱淑琴，辛现伟，罗向阳. 面向国际中文教育的文本可读性因素分析及自动评估研究[J]. 外语研究，2025，42（1）：81–88.

[14] 吴思远，于东，江新. 汉语文本可读性特征体系构建和效度验证[J]. 世界汉语教学，2020，34（1）：81–97.

[15] 杜月明，王亚敏，王蕾. 汉语水平考试（HSK）阅读文本可读性自动评估研究[J]. 语言文字应用，2022（3）：73–83.

[16] LEE B W，JANG Y S，LEE J. Pushing on text readability assessment: a transformer meets handcrafted linguistic features[C]//

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg，PA：Association for Computational Linguistics，2021：10669–10686.

[17] MENG C，CHEN M，MAO J，et al. ReadNet: a hierarchical transformer framework for web article readability analysis[C]// European Conference on Information Retrieval. Cham：Springer International Publishing，2020：33–49.

[18] JIAN L，XIANG H，LE G. English text readability measurement based on convolutional neural network: a hybrid network model[J]. Computational Intelligence and Neuroscience，2022：6984586.

[19] LEE B W，LEE J. Prompt-based learning for text readability assessment[C]//Findings of the Association for Computational Linguistics：EACL 2023. Stroudsburg，PA：Association for Computational Linguistics，2023：1819–1824.

[20] 程勇，徐德宽，董军. 基于多元语言特征与深度特征融合的中文文本阅读难度自动分级研究[J]. 中文信息学报，2020，34（4）：101–109.

[21] 谭可人，兰韵诗，张杨，等. 基于多层级语言特征融合的中文文本可读性分级模型[J]. 中文信息学报，2024，38（5）：41–52.

[22] 李文彪，吴云芳. 基于神经网络模型的汉语文本难度分级[J]. 中文信息学报，2023，37（2）：158–168.

[23] 蒋超，朱学芳. 基于GPT-4数据增强与对比学习的多模态谣言检测研究[J]. 图书情报工作，2024，68（23）：76–87.

[24] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning. New York：ML Research Press，2021：8748–8763.

作者简介

储伊力，女，博士，副教授，研究方向：文本信息评价、文本可读性分析，E-mail：chuyili@ahmu.edu.cn。
曹振祥，男，博士，讲师，研究方向：文本挖掘与分析。
姜烨，女，博士，讲师，研究方向：深度学习、大语言模型。
兰新苗，男，硕士研究生，研究方向：数据挖掘与分析、大语言模型。

Research on Automatic Text Readability Grading Method Based on Large Language Model and Contrastive Learning

CHU YiLi[1,2]　CAO ZhenXiang[3]　JIANG Ye[4]　LAN XinMiao[5]
(1. School of Innovation and Entrepreneurship, Anhui Medical University, Hefei 230032, P. R. China; 2. Anhui Provincial Key Laboratory of Philosophy and Social Sciences for Public Health and Social Governance, Hefei 230032, P. R. China; 3. School of International Economics and Trade, Anhui University of Finance and Economics, Bengbu 233030, P. R. China; 4. School of Computer and Information, Hefei University of Technology, Hefei 230009, P. R. China; 5. Institute of Optoelectronic Technology, Chinese Academy of Sciences, Chengdu 610209, P. R. China)

Abstract: Text readability greatly affects users' information understanding, absorption, and use. To address the problems of confusion between adjacent readability levels and insufficient high-quality labeled data in Chinese text readability grading, this paper proposes an automatic readability grading method based on large language models and contrastive learning. First, a large language model is used to generate contrastive samples with clear readability gradients to expand the training data. Then, BERT is employed to extract deep semantic features, and contrastive learning is introduced to optimize level boundaries in the feature space, thereby improving the model's ability to distinguish adjacent levels. The experimental results demonstrate that the accuracy (0.880) and F1-score (0.881) of the proposed model outperform all baseline models, suggesting that this method can effectively enhance the performance of automatic readability level classification for Chinese texts, particularly showing a clear advantage in discriminating adjacent-level texts.

Keywords: Large Language Model; Contrastive Learning; Readability; Text Information; Automatic Classification