控制句长的句子可读性研究：大语言模型驱动的数据集构建与评估

李罗希，李炜，邵艳秋*
北京语言大学，信息科学学院，北京，100083
202321198093@stu.blcu.edu.cn, liweitj47@blcu.edu.cn, shaoyanqiu@blcu.edu.cn

摘要

文本可读性评估研究旨在衡量文本对特定读者的理解难度，可以分为文档级和句子级。句长这一因素在句子级的难度分类中起主导作用，现有的句子级研究普遍未能控制该变量，从而掩盖了其他深层语言因素在句子难度中的作用。鉴于此，本文提出构建句长受控的句子难度分级语料库。然而，传统人工标注在构建该数据集上存在效率低、质量难以保证的问题。为解决这个问题，本文提出一种大语言模型驱动的智能受控改写方法，利用生成式人工智能从开放语料中自动筛选内容生成候选句，再通过专家审核来保证质量，最终构建了包含二分类三分类的控制句长句子难度分级语料库。在此数据集上的实验结果显示，传统特征分类模型的准确率在控制句长后显著下降，揭示了传统方法的局限性。大语言模型仍具有高准确率，表明其具备识别句长无关语义难度的能力。

关键词： 句子难度分级；句长控制；受控改写；特征评估

Controlled Sentence Length in Readability Research: Dataset Construction Driven by Large Language Models and Assessment

Luoxi Li, Wei Li, Yanqiu Shao*
School of Information Science, Beijing Language and Culture University, Beijing, 100083
202321198093@stu.blcu.edu.cn, liweitj47@blcu.edu.cn, shaoyanqiu@blcu.edu.cn

Abstract

Text readability assessment aims to measure the difficulty of a text for a specific group of readers and can be conducted at both the document and sentence levels. Sentence length plays a dominant role in sentence-level readability classification, and most existing studies fail to control for this variable, thereby obscuring the contribution of deeper linguistic factors to sentence difficulty. To address this, the present study proposes the construction of a sentence-level difficulty corpus with controlled sentence length. However, traditional manual annotation methods are often inefficient and difficult to scale while maintaining consistent quality. To overcome these limitations, we introduce a large language model (LLM)-driven intelligent controlled rewriting approach. Leveraging generative AI, this method automatically selects and rewrites content from open-domain corpora to generate candidate sentences, which are then reviewed by human

* 通讯作者

experts for quality assurance. As a result, we constructed sentence difficulty corpora with both binary and ternary classifications under strict sentence length control. Experimental results show that traditional feature-based classification models experience a significant drop in accuracy when sentence length is controlled, revealing their inherent limitations. In contrast, large language models maintain high accuracy, indicating their ability to identify semantic difficulty independently of sentence length.

Keywords: Sentence Difficulty Classification , Sentence Length Control , Controlled Rewriting , Feature Evaluation

1 引言

作为人类信息处理与认知发展的核心机制，阅读效能与文本可读性水平存在显著关联(Kintsch, 1998)。因此，文本可读性的研究就变得十分重要。它通过量化文本的难度与各个语言特征之间的关系，为教育资源配置、教材编写及语言测评等提供了科学依据。

文本可读性评估最直接的方式是专家评定。在专家具备充分经验且实验设计合理的前提下，此类评估结果通常具有较高的可信度(方昱and 刘海涛, 2021)。然而，该方法存在难以量化、无法自动化实施的局限性，限制了其在大规模应用场景中的可行性(殷晓君, 2022)。文本可读性的**自动评估**是语言学、心理学、教育学、计算机科学等多学科共同关注的交叉研究领域。该任务旨在综合各类可量化的影响阅读难度的文本因素，构建可用于自动化判断的评估模型。目前主流的方法多采用基于语言特征（如字数、词频、句长等）的机器学习技术，将文本难度判定问题转化为回归或分类任务。(Wu et al., 2018)就研究对象而言，可读性自动评估可以分为文档级和句子级两个粒度。(Pilán et al., 2014)传统的可读性评估主要集中于文档级分析，难以满足翻译、试题编写等短文本场景下的评估需求。因此，面向更细粒度的句子级可读性研究也逐渐兴起(Leal et al., 2018)。

句长作为一个隐变量，与多种语言特征（如笔画数、字数、词数、句法依存距离长度等）密切相关，且本身也是句子难度的强相关因素。研究表明，当句子对之间的长度差异超过 50 字时，模型对其难易度的判别准确率可达到100%(于东et al., 2020)。尽管已有研究在若干数据集上实现了较为理想的句子难度自动评估效果，但构建数据时未考虑句长因素因素带来的显著影响，未能系统性地解耦句长与难度之间的共变关系。因此，我们认为：句长这一高度相关的变量可能在评估中掩盖了其他更细粒度或更深层的语义与结构特征，限制了模型在特定条件下的表现能力，也不利于从学理角度揭示句子语义难度来源。此外，控制句长的文本在诸如可控文本生成、语言能力测评等应用场景中亦具有重要的研究与实践价值。

然而，人工构建控制句长的数据集通常需要从海量语料中筛选长度相近且在难度上具有显著差异的句子，过程繁琐，成本较高，难以高效扩展。通过对原始文本进行受控改写，按预设的难度等级和字数对句子进行调整，可在一定程度上缓解这一问题。随着大语言模型（Large Language Model，LLM）的发展与辅助标注的完善(Gu et al., 2025)，其强大的指令跟随和文本生成能力为控制句长的难度风格化改写提供了新的可行路径。

本文首先通过传统的非控制句长数据集验证了句长在句子难度分级中的强相关性，进而提出猜想：**在控制句长的情况下，传统的评估指标失去了这一强相关变量，将导致分类效果显著下降**。为了验证这一结论并深入探讨控制句长的影响，本文构建了一个排除句长因素干扰的控制句长句子难度分级数据集。针对构建数据集面临的代价高昂的问题，我们提出了一个大语言模型驱动的半自动智能受控改写。该方法利用大模型的指令跟随和文本生成能力，从维基百科等开放语料中自动筛选内容，按照字数和相应的难度标准改写出海量候选句。然而，大模型直接生成的句子存在：字数不符合规范，内容句式重复，事实、逻辑错误等问题。因此，我们引入专家审核与双阶段校验，剔除字数不符、语义重复或难度模糊的样本，增强数据的可信度，最终构建了包含 1200 句二分类（简单/困难）和 1350 句三分类（简单/中等/困难）的控制句长数据集。该方法显著降低了构建数据集的成本，解决了构建的困难，为进一步探讨句长控制条件下的特征评估与模型能力提供了坚实的研究基础。

基于该数据集，我们实验对比了传统特征和大语言模型在句子难度判断中的表现差异。实验结果表明，在控制句长条件下，传统语言学特征（如词频、笔画数、句法依存距离）在分类

任务中的准确率显著下降（二分类最高 76%，三分类最高 57%），而大语言模型最高能分别达到 98%、84% 的准确率，揭示了大模型能够捕捉到与句长无关的隐性语义模式。该发现挑战了传统可读性公式(Dale and Chall, 1948; Kincaid et al., 1975; 荆溪昱, 1995)和基于特征的机器学习范式(杨文娣and 曾致中, 2019)，并为句子难度分级提供了新的视角和方法论参考。

本文主要贡献体现在以下四个方面：

(1)观察发现传统句子难度研究中句长在分类上的主导作用。通过实验验证了句长的强相关的地位，并提出构建句长受控的句子难度分类数据集；

(2)将大语言模型引入语料构建环节，通过三阶段协同的半自动智能生成机制，有效解决了句长受控数据集构建的效率和质量瓶颈，形成了共 2550 句的二分类与三分类高质量语料资源；

(3)通过实验发现传统语言特征分类模型在句长受控条件下的分类准确率大幅下降，揭示了传统特征分类模型方法的局限性；

(4)首次在控制句长数据上验证了大语言模型的深层语义建模能力，其准确率较传统特征方法大幅提升，验证了大语言模型具备识别句长无关语义难度的能力。

2 相关工作

2.1 可读性研究

作为该领域的起点，传统可读性公式通常基于统计学方法，通过线性回归等模型，利用文本的表层特征（如词长、句长、音节数、难词比例）来预测阅读难度(Leal et al., 2018)。这些公式因其计算简便、解释性强而得以广泛应用，其中的代表包括Flesch Reading Ease (FRE)、Flesch-Kincaid Grade Level (Kincaid et al., 1975)以及Dale-Chall可读性公式(Dale and Chall, 1948)等，它们为早期的可读性评估提供了有效的量化工具。然而，这些公式的内在局限性也十分明显：它们仅能捕捉文本的表面特征，无法深入分析深层句法结构、语义抽象度或语篇连贯性等复杂因素(Wu et al., 2018)。此外，多数经典公式专为英语设计，对于中文等非拼音文字的适用性受限(Yang, 1971)。

随着机器学习技术的兴起，研究者开始采用支持向量机(SVM)、随机森林(Schwarm and Ostendorf, 2005)等更为灵活的分类与回归模型来评估可读性，从而摆脱了传统公式的固定框架。这种方法的核心在于人工设计并提取多维度的语言特征，通过模型训练以提升预测的准确度与泛化能力。特征工程的范围涵盖了词汇、句法(Schwarm and Ostendorf, 2005)、语义和语篇等多个层面，旨在更全面地捕捉文本难度。在跨语言研究中，机器学习方法展现出更强的适应性，例如，在中文可读性评估中，研究者会特别考虑汉字结构、词语切分、词汇频率等中文独有的语言学特征(Pang, 2006)。尽管基于特征工程的方法显著提升了预测精度，但其高度依赖于人工经验和语言学知识，不仅耗时费力，且难以完全捕捉语言内在的复杂性与多样性。

进入深度学习时代，卷积神经网络(CNN)(Jian et al., 2022)、循环神经网络(RNN)以及Transformer(Li et al., 2022)等模型在可读性评估中得到广泛应用，实现了从“特征工程”到“特征学习”的范式转变。这些模型能够自动从原始文本中学习和提取高层次、抽象化的特征表示，无需繁琐的手工特征设计。一些预训练语言模型，通过在海量文本上进行预训练，仅需少量微调即可在可读性评估任务上取得卓越性能，已成为当前的主流方法。然而，深度学习模型在减少对人工特征依赖的同时，也带来了计算资源需求高和模型可解释性弱等问题。

2.2 文本简化与可控生成

文本简化（Text Simplification, TS）作为自然语言处理的重要分支，旨在将复杂的文本转换为更易于理解和阅读的形式，从而提升信息的可访问性(Al-Thanyyan and Azmi, 2021)。该技术在教育、医疗、新闻等领域具有广阔的应用前景，是可读性理论在实践层面的重要体现(O’Brien, 2010)。早期文本简化研究主要采用基于规则和基于统计机器翻译的方法。基于规则的方法依赖人工编写的语言学转换规则，例如通过句法分析来识别并拆分、删除或替换复杂句式(Siddharthan, 2006)。其优点在于可解释性强，能够精确控制特定的语言现象，但规则的制定耗时巨大，覆盖范围有限，且生成结果的自然度难以保证。基于统计机器翻译（SMT）的方法则将文本简化视为一种特殊的翻译任务，即从“复杂语”到“简单语”的翻译(Specia, 2010; Xu et al., 2016)。该方法利用大规模的平行语料库训练模型，学习复杂句与简化句之间的映射关系。SMT方法能够生成更为流畅自然的文本，但其性能高度依赖于高质量平行语料库的规

模——而这类资源通常是稀缺的——并且难以实现细粒度的生成控制(Al-Thanyyan and Azmi, 2021)。

随着深度学习在机器翻译领域的突破，序列到序列（Seq2Seq）模型及其变体被广泛应用于文本简化任务(Nisioi et al., 2017)(Vaswani et al., 2017)。注意力机制和Transformer架构(**?**)的引入，极大地提升了模型处理长距离依赖和并行计算的能力，迅速成为文本简化领域的主流模型。预训练语言模型通过在超大规模语料上的预训练学习了丰富的语言知识，仅需在特定任务上进行微调便能取得优异的简化效果。尽管神经网络方法取得了显著进展，但仍面临着过度简化、关键信息丢失、生成文本不忠实于原文或不自然等挑战。

现代文本简化不仅要求“变简单”，更需要满足特定的约束条件，如保持核心信息、控制输出长度、调整风格或难度等级。可控文本生成（Controlled Text Generation, CTG）技术为实现这些目标提供了有力工具(Hu et al., 2017)。而大语言模型（LLM）的兴起，则极大地推动了可控生成技术的发展。提示工程（Prompt Engineering）(Liang et al., 2024)已成为LLM时代最核心的CTG手段之一，通过精心设计的指令，可以引导模型生成满足特定需求的句子。针对更为复杂的控制需求，研究者还探索了基于规划的生成（Planning-based Generation），例如通过将可读性分数作为控制因子来精确控制输出句子的难度(Yang et al., 2022)。然而，LLM在严格的长度控制方面仍表现出所谓的“字数危机”，即难以精确遵守“生成特定字数”的指令(Liang et al., 2024)。

2.3 大语言模型的自然语言理解和可读性评估能力

大型语言模型（LLM）在多项自然语言理解（NLU）任务中已达到甚至超越人类水平，能够捕捉文本的深层语义，进行多跳推理和复杂阅读理解(Achiam et al., 2023)。例如，在如MMLU（Massive Multitask Language Understanding）(Hendrycks et al., 2020)、GPQA（General-purpose Question Answering）(Rein et al., 2024)和BIG-Bench Hard(Suzgun et al., 2022)等综合性基准测试中，LLM展现出其在跨学科知识、逻辑推理及常识理解方面的优越性。其核心特性如上下文学习（In-context Learning）和少样本学习（Few-shot Learning）(Brown et al., 2020)使其能够通过在提示中提供少量示例来执行新任务，显著降低了对大规模标注数据的依赖，从而为各种下游任务提供了高效且灵活的解决方案。

值得关注的是，LLM在文本可读性和难度评估方面展现出超越传统方法的潜力。最新研究表明，LLM能够提供更细致入微、更符合人类直觉的可读性判断。例如，一项关于使用GPT-4o评估英文文本可读性的研究发现(Trott and Rivière, 2024)，GPT-4 Turbo和GPT-4 mini在零样本（zero-shot）设置下生成的可读性估计与人类判断表现出较高的相关性（分别为0.76和0.74），显著优于传统的基于特征的可读性指标。这表明LLM通过理解词汇选择、句子复杂性以及文本的整体流畅性与结构，能够提供更具语境化的评估，从而有望成为替代传统分类模型、更好地判别文本可读性的新范式。

3 大语言模型驱动的半自动智能受控改写

传统的文本难度标注主要依赖人工按照预设标准在特定语料上直接标注(Pitler and Nenkova, 2008)(Klare and others, 1984)。然而，在构建控制句长的数据集时，筛选符合长度范围的自然句子进行人工标注的方法存在明显的问题：1.难度分布极度不均衡：即便在有明显难度分级的教材语料中，控制句长后难度标签分布也存在严重偏斜（我们在 4.1节的传统句子语料库中的统计发现：在 25 - 30 字这一范围，难度为 3 的句子高达 299 条，而难度为 1、2、4 的分别仅为 1、76、13 条，这一结果还是在标注允许相邻等级存在模糊空间的前提下得到的）。在更为开放的自然语料中，这种不均衡现象更为突出，给初步筛选合适句子带来了较大挑战，且会带来远大于当前方法的人工确认成本。2.缺乏足够判别性的样本对：在相同句长范围内，自然语料中的句子在难度层面往往差异较小（上文的难度分布也能说明），人工标注一致性低、主观性强，难以形成有效对比对。这不仅降低了标注效率，也影响了后续模型的判别性能。同时，若通过人工方式主动改写以增强样本差异性，将面临高昂成本与风格一致性难以控制的问题。

为解决上述问题，本文提出一种基于大语言模型的半自动智能受控改写框架。借助模型语义理解与生成能力，从多领域的原始文本中筛选出合适的内容进行批量的造句和难度风格改写，大幅提升控制句长语料构建的效率与难度差异。该方法创新性地采用“三阶段协同机制”，

包括多领域语料获取与智能受控改写、专家审核与人工修订、标注一致性与数据可信度验证。

3.1 多领域语料获取与智能受控改写

为避免单一语域在词汇使用、句式结构等语言特征上的偏倚，减少内容对特定领域专业知识的过度依赖，提升实验结论的普适性，本研究选取包括新闻报道、百科条目等在内的多领域开放语料，确保候选语料在语义内容和语言结构上的广度与多样性。

随后，我们将经过初步筛选的开放语料文本批量输入至大语言模型，并结合任务需求设计多轮引导性 Prompt ，使模型能够按照预设的长度范围（如25至30字）、明确的难度等级标准，对原始内容进行筛选与智能受控改写。例如图 6中，大模型读取到了一篇关于“养老社区”的新闻报道，按照“困难”的要求筛选到了关于“社区化养老遇到的医疗难题”的文段并结合一些背景知识，例如“医疗、养老牌照审批困难”，将其浓缩改写成了一个初步的句子。在该阶段，大模型的处理能力显著提升了初始语料的生成效率，极大减轻了人工负担。

3.2 专家审核与人工修订

尽管大语言模型具备良好的语义建构能力，但生成内容仍可能存在如下问题：(1)长度不符合规范：模型对字数控制能力有限，常出现简单句子偏短、复杂句子偏长的情况；(2)内容与结构单一：模型在同一难度等级下易于生成高度重复的内容或句式，降低语料多样性；(3)事实性或逻辑性错误：如“全自动咖啡机利用伯努利原理控制蒸汽压力，实现奶泡绵密质地生成”这一生成句，存在科学原理理解错误。咖啡机实际通过加热元件和压力阀调控蒸汽压力，而非依赖流体速度与压强的关系。

为提升数据质量，本研究引入人工专家对大语言模型生成的初始候选语料进行系统化筛选与修订。专家依据预设标准，剔除字数不符合要求、语义不明确、内容重复或存在事实性错误的句子，并对保留样本进行必要修改。在大模型生成的基础上开展人工优化，既显著提高了标注效率，又有效保证了语料在长度控制、难度分布与语言结构多样性等方面的质量与合理性。

3.3 标注一致性与数据可信度验证

为确保最终语料的标注质量与客观性，本研究进一步引入第二位专家对经过筛选的样本独立判断其难度等级。采用 Fleiss' Kappa 系数评估两位专家间的一致性，以衡量数据标注的稳定性与信度。该方法有效避免了单一标注者的主观偏差，为后续模型训练与评估提供了更为可靠的数据基础。

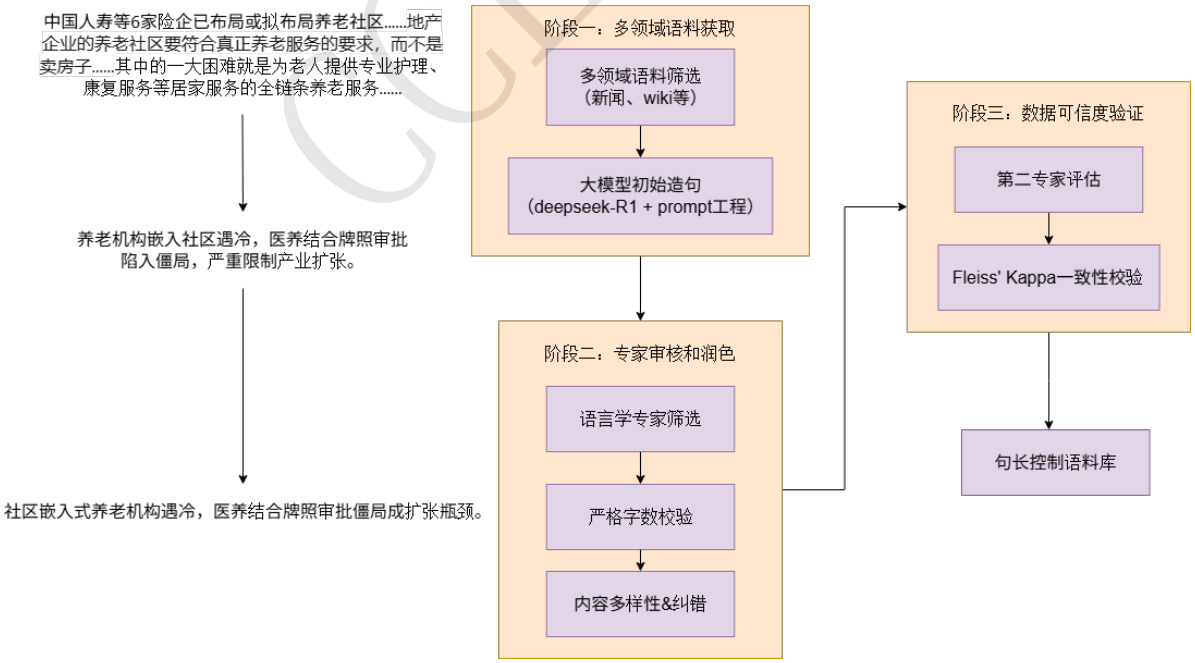

图 1: 半自动标注方法流程图

4 语料库构建

本研究构建了两类句子难度分级语料库，包括传统不控制句长的语料库和大语言模型驱动的控制句长语料库。前者用于验证句长作为显性和隐形变量时和句子难度相关性，后者则用于进一步在排除句长干扰的前提下考察传统分类模型的语言特征与大语言模型各自的分类效果。

4.1 传统句子语料库

为了验证句长这一变量在句子难度分级中的强相关性地位，本研究构建了一个传统不控制句长的对照数据集，作为基线参照。该数据集选取自《博雅汉语》全套教材，经过句子筛选与去重处理后，共获得 6877 个有效句子。我们将该数据集划分为 10 个子集，并邀请 20 名大学生参与标注，每个子集由 2 人独立进行 5 级难度评分（ 1 为最易， 5 为最难）。在评分前，要求标注者对每个句子进行完整阅读与理解，并记录阅读时间，随后再进行评分，以确保评分过程的严谨性与思维投入。最终，仅保留两位标注者评分一致的样本，构成最终的传统句子难度语料库。该语料库中，句子阅读时间与标注难度之间的皮尔逊相关系数（Pearson Correlation Coefficient）为 0.5886 ， p 值为 0 ，显示出显著的正相关性，说明标注质量较高，有效避免了随意阅读与评分的情况(Hale, 2001) (Levy, 2008)。最终语料库共包含 3555 个句子，难度等级分布大致呈钟形分布。我们在表 1中展示了例句及对应的难度等级。

难度	数量	例句
1	584	我说我肯定在。
2	952	要是能有一个长假期，我一定要跑遍中国。
3	1439	就是形成良性循环，绿地涵养了水分，天然降雨又涵养了绿地，周而复始。
4	414	通过不同消费品的“绝配”，达到不同社会阶层体现社会身份的“制胜”之道，是今天各种人群消费实践的主要方式。
5	166	实际社会生活中存在着两种新男性－一种是外在装备水平可以使公众明显感知他们是男性中的精英分子的“新男性群体”；另一种是自认属于新男性或者至少自认具备新男性的现实潜力的群体，但他们却并不为一般社会公众所感知，因为他们往往不具备前者拥有的“装备”。

表 1: 不控制句长五分类句子示例及其数量

4.2 控制句长语料库

本研究选取了三个不同领域的文本来源作为数据构建基础，分别为：叙事新闻（来自通用新闻语料库）、人文社科、科学技术（后两者筛选自中文维基百科），语料来源参见(Xu, 2019)）。

在句长控制方面，本文选取了 25 - 30 字的长度范围，作为实验中控制变量的核心设定。该范围的选择基于以下考虑：1.该区间处于现代汉语句子长度的中间段，能够在保证语义完整性的同时，避免句子过短导致表达不清，或过长增加理解负担，因此更适合作为可读性评估的“中间模糊区间”；2.过短或过长的句子在可读性上往往呈现极端表现（即显著过于简单或困难）。根据我们论文中在非控制句长语料上的统计分析，难度等级为 3 （中等）的句子中位字数为 33 字；考虑到该语料出自面向二语学习者的教材，整体信息密度相对较低，而本研究所用语料系通过大语言模型筛选、抽取与浓缩生成，因此，我们认为缩小字数范围（ 25 - 30 字）的句子更适合作为基础样本进行特定难度方向的改写，有助于构建梯度明确、差异可感的训练样本对。

依托本文提出的大语言模型（这里使用 DeepSeek-R1 ）驱动的半自动构建框架，最终构建了两个控制句长的汉语句子难度数据集，分别为：

二分类语料库：共计 1200 句，覆盖“简单”与“困难”两类。三个语料领域（叙事新闻、人文社科、科学技术）各占 400 句，每类各含 200 句简单样本与 200 句困难样本，平均句长分别为 26.88 字与 28.50 字。

三分类语料库：共计 1350 句，包含“简单”“中等”“困难”三类。三个语料领域各占 450 句，各难度句子数量均为 150 条。各难度等级句子的平均长度分别为 28.72 、 28.79 与 28.66 字，

控制效果良好，标准差极小。

我们在表 2中展示了例句和对应的难度等级。

为验证标注的一致性和语料的可信度，本研究引入第二位人工专家独立对数据样本进行难度判断，并计算标注一致性指标。经统计， Fleiss' Kappa 系数为二分类 0.99 ，三分类 0.82 ，表明两个专家在句子难度划分上的一致性达到高度一致的水平，进一步确保了控制句长语料库的标注质量与研究可信度。

分类	难度	例句
二分类	1	长颈鹿血压是人类三倍，这样才能确保血液输送到它高高的头部。
	2	近海赤潮预警失灵，增殖模型未计入船舶压载水外来物种变量。
三分类	1	俄罗斯是世界上国土面积最大的国家，相当于八个法国的大小。
	2	郑州叠拼住宅空间方案获奖，非承重结构改造实现建筑利用率提升。
	3	明镜周刊的左派立场与虚构报道丑闻，构成后真相时代的伦理困境。

表 2: 不同分类与难度下的例句

5 实验设置

5.1 非控制句长语料实验

首先，本文通过计算句长与句子难度之间的 Pearson 相关系数来验证句长是否为影响句子难度的重要变量。句长采用两种统计方式：按字数计算与按词数计算。通过观察两者与人工标注的难度评分之间的相关性，初步评估句长作为难度指标的代表性。

进一步地，本文引入多个语言学特征，参考吴思远等人的研究(吴思远et al., 2020)，涵盖汉字、词汇与句法等多个层级。其中，部分特征本质上隐含了句长的影响，例如汉字总笔画数、句法最大依存距离等。通过分析这些特征与难度评分之间的相关性，进一步验证句长是否作为强相关隐变量主导了传统特征的判断效果。

文本处理方面，本文采用哈尔滨工业大学开发的语言技术平台（Language Technology Platform，LTP）(Che et al., 2010)对语料进行分词、词性标注与依存句法分析；词频数据引用SUBTLEX-CH 词频库(Cai and Brysbaert, 2010)，以获取准确的词汇使用频率信息。

5.2 控制句长语料实验

在构建的控制句长语料库（包含二分类与三分类两类任务）上，本文设计对比实验以评估传统语言特征与大语言模型在句子难度判断中的表现差异。

针对传统特征的建模，本文采用支持向量机（Support Vector Machine, SVM）进行分类，评估指标包括分类准确率（Accuracy, Acc）和宏平均的 F1-score 。特征选取同样参考吴思远等人(吴思远et al., 2020)，涵盖汉字（如总笔画数、字符型符比）、词汇（如平均词频、词类比例）与句子（如最大依存距离）等多个维度。

在大语言模型的实验部分，为全面评估不同规模与类型模型在语言难度判定任务中的表现，本文选取了三种具有代表性的语言模型进行对比分析： Gemma 3 4B (Team et al., 2025)（一种多语言支持的小规模开源模型）、 Qwen 2.5 14B (Bai et al., 2025)（一款在中文语料上表现优异的大型中文语言模型）以及 GPT-4o (Hurst et al., 2024)（代表当前主流闭源模型的先进水平）。我们通过构造标准化的 prompt 引导模型对输入句子的语言难度进行判定，以评估其在二分类与三分类任务中的分类准确率。该设置旨在考察不同模型对语言复杂度的理解能力及其在受控实验条件下的泛化表现。共设置六种 Prompt 设计形式(具体见 附录 B.)：

Zero-shot：模型仅接收待判句子，无任何参考信息。

Zero-shot-exp：模型在判断难度的同时，需解释其判断依据。

Zero-shot-expto：模型需在判断难度的同时，以简洁易懂的语言向用户解释句子含义。

Few-shot：模型接收若干已标注的参考句（覆盖不同难度等级）以辅助判断。

Few-shot-exp：在few-shot 基础上要求模型解释其判断原因。

Few-shot-expto：在few-shot 基础上要求模型解释句子含义，体现其语言理解能力。

通过上述六种设置，本文不仅评估了大语言模型在句子难度判断任务中的分类准确率，还从直接与间接两个维度系统探讨了其对语言难度的感知能力，并进一步分析“解释性行为”在判断过程中对模型性能的促进或干扰作用。

6 实验结果和分析

6.1 句长的强相关隐变量地位

图 2展示了句子长度（以字数或词数衡量）与句子难度之间的 Pearson 相关系数，均达到 0.81%，显著高于强相关的通用阈值（0.8%），表明句长是判断句子难度的一个高度相关因素。

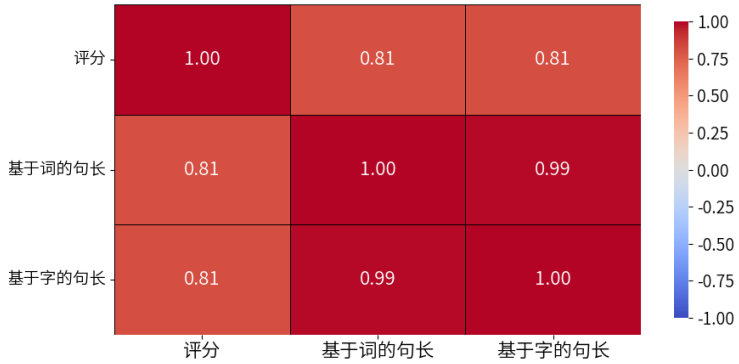

图 2: 两种句长与句子难度相关性矩阵

不论是按字还是按词计算的句长均与句子难度呈现出极高的Pearson相关系数，说明了句长在句子难度判断中的强相关性。

为了进一步验证这一点，我们从汉字、词汇与句法三个维度出发，统计各类语言特征与句子难度的相关性.

汉字熟悉度	平均字频0.12	未登录字比例-0.40	常用字比例0.37	
汉字多样性	字形符数**0.81**	字类符数**0.85**	字类符形符比-0.67	
字形复杂度	句子总笔画数**0.82** 少笔画字比例-0.16	字符平均笔画数0.34 中笔画字比例0.16	频率加权的平均笔画数0.10 多笔画字比例0.042	
词汇熟悉度	平均对数词频-0.10	未登录词比例0.10		
词汇多样性	词形符数**0.81** 单次词比例0.07	词类符数**0.84** 成语数0.23	词类符形符比-0.67 成语比例0.09	单次词数0.23
词性复杂度	名词比例0.17	动词比例0.14	形容词比例0.023	副词比例0.06
词语复杂度	平均词长0.32 三字词比例-0.04	频率加权的词长-0.26 四字词比例0.14	单字词比例-0.37 四字以上词比例0.09	双字词比例0.32 字词比0.32
词汇语义	实词比例0.27 否定词数0.15	虚词比例-0.012 否定词比例-0.04	实词虚词比0.37	
短语结构复杂度	名词短语数**0.71** 副词短语数0.58 动词短语平均长度0.26	动词短语数**0.71** 介词短语数0.63 复句数0.48	形容词短语数0.42 名词短语平均长度0.38	
依存句法复杂度	平均句子依存距离**0.79** 最大主语长度0.45 平均修饰语个数0.16	最大句子依存距离**0.75** 平均修饰语个数0.16 平均修饰语长度0.42	平均主语长度0.40 平均修饰语长度0.42	

表 3: 各类语言学特征与句子难度的皮尔逊相关系数

表 3 表明，在汉字层面，字形符与类符数、总笔画数等指标与难度高度相关，相关系数超过 0.8%；词形符、词类符等词汇多样性指标也具有同样的高相关性；而在句法层面，名词短语数、依存句法中的平均依存距离、最大依存距离等指标同样表现出显著相关性。

综合这些观察可以看出，尽管句长在本实验中被设定为隐变量，其依然以间接形式影响着众多语言特征的变化模式。所有与句子难度高度相关的语言特征——无论是总笔画数，词形符、词类符，最大依存距离等——在统计上均与句长密切相关(更多实验见 附录 A)。

据此我们可以推论：**许多传统的语言难度指标在本质上依赖于句长这一显性或隐性变量。当实验中不控制句长时，这些指标往往能取得较好的分类效果；但一旦控制句长，其依赖的信息维度被剥离，导致其对难度的判别能力显著下降。**

6.2 传统特征的局限和大语言模型的优势

表 4 列出了各语言特征在二分类和三分类任务中的准确率与 F1 值。从结果来看，传统语

言特征在二分类和三分类任务中的表现均不理想，仅略高于随机猜测（二分类为 50% ，三分类约为 33% ），这表明在句长受到控制的条件下，基于传统特征的分类模型能力受限，难以有效区分不同类别。

在语言特征中，词汇相关的分类效果最好，句子特征的分类效果最差。词汇总特征取得了二分类 75% ,三分类 55% 的准确率。而句子特征仅为 64% 和 47% 。值得注意的是，这些因素确实与人类的理解经验一致，控制句长后会对句子的一些特征例如依存距离等造成限制，而罕见词和复杂词性组合通常会增加句子的阅读难度。将全部特征合在一起效果最好，达到了二分类76%，三分类57%,这可能是综合全部特征能够更好兼顾句子难度的多个方面。

然而，这些特征仅能捕捉到句子难度的表层因素，无法涵盖更深层次的语义与逻辑难度。例如，“宇宙由物体和空间构成，其余本质皆为观测者对于运动的描述。”这句话并不包含任何难字或生僻词汇，但由于其涉及哲学与物理概念和较为困难的逻辑，理解起来依然困难。这一现象突显了传统语言特征在覆盖面和表达能力上的局限性。

这一发现也验证了前文提出的假设：**传统指标对句子难度的判断高度依赖于句长，一旦失去这一维度，其性能便显著下降。**

	二分类		三分类	
	Accuracy	F1-score	Accuracy	F1-score
字形复杂度	0.63	0.63	0.47	0.46
汉字多样性	0.58	0.58	0.38	0.37
汉字熟悉度	0.69	0.69	0.50	0.48
汉字总特征	0.68	0.67	0.45	0.45
词汇多样性	0.57	0.56	0.40	0.35
词汇熟悉度	0.66	0.66	0.49	0.44
词汇语义难度	0.62	0.62	0.45	0.45
词性复杂度	0.71	0.71	0.48	0.47
词语复杂度	0.60	0.60	0.44	0.43
词汇总特征	**0.75**	0.75	**0.55**	0.55
短语结构复杂度	0.63	0.63	0.47	0.47
依存句法复杂度	0.57	0.57	0.40	0.38
句子总特征	0.64	0.64	0.47	0.47
全部特征	**0.76**	0.76	**0.57**	0.57

表 4: 各语言特征在控制句长二分类与三分类任务下的效果
传统的所有语言指标在分类中都未取得较好的效果，词汇特征和全部特征取得的效果略好。

近年来，大语言模型在文本分级任务中的表现受到广泛关注(韩欣欣et al., 2025)。本研究进一步评估了几个主流大语言模型在本数据集上的表现，结果如表 5 所示。

二分类			
	Gemma3 4b	**Qwen2.5 14b**	**GPT-4o**
zero_shot	0.9367	0.9425	0.9625
zero_shot_exp	0.9108	0.9458	0.9725
zero_shot_expto	0.9267	0.9558	0.9767
few_shot	0.9417	0.9050	0.9742
few_shot_exp	0.9275	0.9467	**0.9808**
few_shot_expto	0.9117	0.9050	0.9783
三分类			
	Gemma3 4b	**Qwen2.5 14b**	**GPT-4o**
zero_shot	0.6593	**0.7726**	0.7178
zero_shot_exp	0.6178	**0.7793**	0.7437
zero_shot_expto	0.5859	**0.7919**	0.7756
few_shot	0.6237	0.6393	**0.8200**
few_shot_exp	0.5578	0.7296	**0.8207**
few_shot_expto	0.5570	0.7496	**0.8422**

表 5: 不同模型和设定下在二分类与三分类任务中的准确率对比

第二十四届中国计算语言学大会论文集，第666页-第683页，济南，中国，2025年8月11日至14日。

可以观察到，大语言模型在句子难度的二分类与三分类任务中均表现出显著优越的效果。模型规模越大，其性能越强： GPT-4o 的两个任务上的最高评分均最优异，在二分类任务中达到了 98% ,三分类任务中最高达到 84% 的准确率，远超传统特征的最高水平。这说明大语言模型用于句子难度的判断是较为可信的，能够推广到一些下游领域，例如辅助教材分级等等。

进一步分析发现：在 zero-shot 设置下， Qwen2.5 和 GPT-4o 表现出色，展现了较大规模模型在无样例的情况下对语义复杂度的良好感知能力； Gemma3 4b 可能由于能力有限，在三分类任务上表现较为一般。

在 few-shot 设置下， GPT-4o 的表现相比 zero-shot 进一步提升，这表明模型能够从示例中学习到关于难度判断的深层知识。值得注意的是， Qwen 模型在加入 few-shot 示例后反而出现了性能下降，这可能是由于模型在学习样本时过度拟合了某些表层显性特征（如难词、领域术语、句式等），从而导致判断偏差。

在加入解释时（ exp ， expto ），较大规模的模型表现出了准确率的提高，说明让模型“思考”其决策依据有助于提升推理质量；而Gemma3 4b则出现了准确率的下降，这可能是由于其模型规模的限制。

值得注意的是，在三分类任务中， Qwen2.5 在 zero-shot 设置下的性能优于 GPT-4o ，而在加入 few-shot 示例后， GPT-4o 的表现则超过了 Qwen2.5 。这一结果表明， Qwen2.5 在零样本推理方面展现出较强的能力，但其对示例的学习效能相对有限；相对而言， GPT-4o 在利用少量示例进行学习方面具有更强的适应能力。

尽管大语言模型在整体表现上远超传统特征方法，但其在处理某些具有特殊领域词汇或知识背景的句子时仍存在误判的可能。模型较容易受到词汇领域的影响，而忽视语义、逻辑和背景知识所共同构成的“隐性难度”。例如“脂肪分解产生的能量是糖类的两倍左右，但其存在分解较慢的问题。”这个句子，模型将其判定为二分类中的困难，因为其涉及了生物学和化学的概念。但这个句子对于我们受过义务教育的人来说实际上是非常熟悉的营养学知识。

7 总结

本文通过观察发现传统语言指标中句长对于句子难度的主导地位，并设计实验证明了这一点。针对传统句子难度评估中句长变量强干扰的问题，本文提出构建控制句长的句子难度分类数据集。为解决数据集构建的成本高、质量难控制问题，我们提出基于大语言模型驱动的半自动智能受控生成方法，最终构建了控制句长的汉语句子难度分级二分类和三分类语料库。在此基础上的实验验证了传统特征分类模型方法的局限与大语言模型的强大语义难度识别能力。

然而，本文的研究仍存在以下几方面的局限性，有待在后续工作中进一步完善与拓展：其一，当前数据集构建主要依赖大语言模型进行句子生成，并辅以人工审核以确保生成文本的质量。尽管该方法在可控性与效率之间取得了一定平衡，但其存在生成效率受到人工审核影响，自然性与语言多样性方面仍可能与真实自然语料存在一定偏差等问题。未来的研究将尝试引入自动化与人工相结合的质量保障机制（如判别模型与人机协同审核机制），以进一步提高生成文本的语言质量与分布多样性。其二，本文的数据集数量较少并且主要聚焦于 25 至 30 字的句长区间。尽管该区间在汉语书面语中具有一定代表性，但句子的数量和句长范围的单一性限制了我们对不同句长区间对句子难度感知影响的系统性探索。未来的工作将扩展到更大的数量和更广泛的句长区间，并结合统计方法确定更具理据性的句长分割区间。其三，本研究所涉及的样本文本主要来源于新闻、百科等书面语域，尚未覆盖口语、文学、社交媒体等其他语体类型。其四，本文尚未将所构建的受控语料或判别模型应用于具体的下游任务中，如可控文本生成、语言能力测评、或智能教学系统等实际场景。未来工作可考虑将本研究中构建的数据集与模型能力迁移至具体任务中。其五，本文的研究对象为汉语句子，尚未涉及跨语言或多语种条件下的可读性问题。考虑到不同语言在句法结构、词汇密度与信息组织方式上的差异，未来研究可进一步拓展至其他语言（如英语、日语等）。

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qing Cai and Marc Brysbaert. 2010. Subtlex-ch: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6):e10729.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Coling 2010: demonstrations*, pages 13–16.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lihua Jian, Huiqun Xiang, and Guobin Le. 2022. English text readability measurement based on convolutional neural network: A hybrid network model. *Computational intelligence and neuroscience*, 2022(1):6984586.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

George R Klare et al. 1984. Readability. *Handbook of reading research*, 1:681–744.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Wenbiao Li, Ziyang Wang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. *arXiv preprint arXiv:2210.10305*.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

Sharon O'Brien. 2010. Controlled language and readability. In *Translation and cognition*, pages 143–165. John Benjamins Publishing Company.

Lau Tak Pang. 2006. Chinese readability analysis and its applications on the internet. *Master's thesis, The Chinese University of Hong Kong*.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, pages 174–184.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language*, pages 30–39. Springer.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Sean Trott and Pamela D Rivière. 2024. Measuring and modifying the readability of english texts with gpt-4. *arXiv preprint arXiv:2410.14028*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siyuan Wu, Jianyong Cai, Dong Yu, and Xin Jiang. 2018. A survey on the automatic text readability measures(文本可读性的自动分析研究综述). 32:1–10, 12.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp, September.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation. *arXiv preprint arXiv:2204.13362*.

Shou-jung Yang. 1971. *A readability formula for Chinese language*. The University of Wisconsin-Madison.

于东, 吴思远, 耿朝阳, and 唐玉玲. 2020. 基于众包标注的语文教材句子难易度评估研究. 中文信息学报, 34(2):16–26.

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

方昱and 刘海涛. 2021. 句法结构认知难度的计算指标分析. 南京师大学报*(*社会科学版*)*, (06):126–137.

杨文娣and 曾致中. 2019. 基于随机森林算法的对外汉语文本可读性评估. 中国教育信息化, 14:89–96.

殷晓君. 2022. 基于依存构式的文本复杂度分级特征体系构建及效度验证. 语言教学与研究, (06):24–33.

荆溪昱. 1995. 中文国文教材的适读性研究: 适读年级值的推估. 教育研究资讯, 5:113–127.

韩欣欣, 马瑞, and 徐娟. 2025. Deepseek赋能国际中文教学资源建设的技术路径探索——以分级阅读文本生成为例. 国际汉语教学研究, (01):30–40.

附录 A　补充统计结果

附录 A.1　方差分析(ANOVA) 结果

```
(              sum_sq     df          F        PR(>F)
 C(句长类别)   201.203475    2.0  365.810158  2.673948e-89
 Residual  104.504097  380.0         NaN           NaN,
```

图 3: 方差分析(ANOVA) 结果

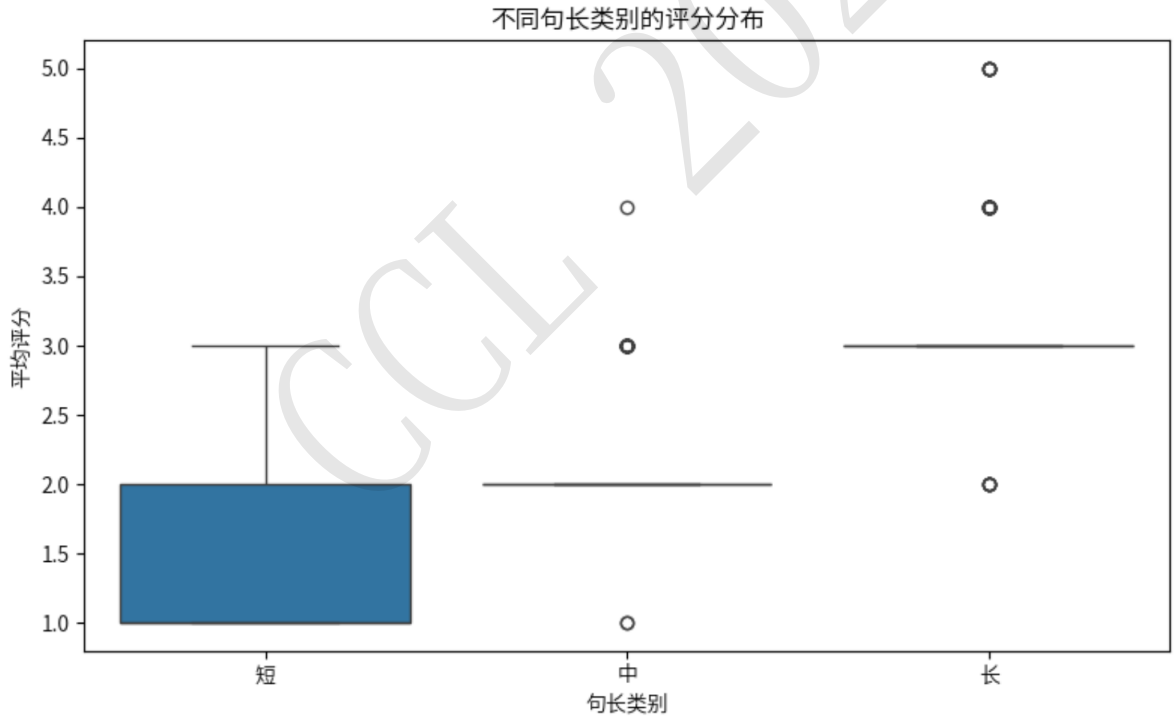

图 4: 方差分析(ANOVA) 结果

由于p 值远小于任何标准的显著性水平（如0.05, 0.01, 0.001）。这使得我们可以强烈地拒绝零假设，即“句长类别对平均评分没有影响”。因此，我们可以得出结论，“句长类别”对“平均评分”具有统计学上极其显著的影响。这意味着至少在“短”、“中”、“长”这三个句长类别中的某一对或多对之间，“平均评分”存在显著差异。

附录 A.2　OLS 回归结果

OLS Regression Results

Dep. Variable:	平均评分	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.706
Method:	Least Squares	F-statistic:	184.1
Date:	Mon, 30 Jun 2025	Prob (F-statistic):	7.81e-99
Time:	22:08:23	Log-Likelihood:	-263.61
No. Observations:	383	AIC:	539.2
Df Residuals:	377	BIC:	562.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3617	0.106	12.803	0.000	1.153	1.571
字数	0.0302	0.011	2.837	0.005	0.009	0.051
词数	0.0088	0.013	0.657	0.512	-0.017	0.035
句子总笔画数	0.0037	0.001	3.523	0.000	0.002	0.006
平均对数词频	-0.0295	0.034	-0.865	0.388	-0.097	0.038
最大句子依存距离	-0.0414	0.007	-5.633	0.000	-0.056	-0.027

Omnibus:	8.754	Durbin-Watson:	1.978
Prob(Omnibus):	0.013	Jarque-Bera (JB):	14.989
Skew:	-0.023	Prob(JB):	0.000556
Kurtosis:	3.968	Cond. No.	989.

图 5: OLS 回归结果

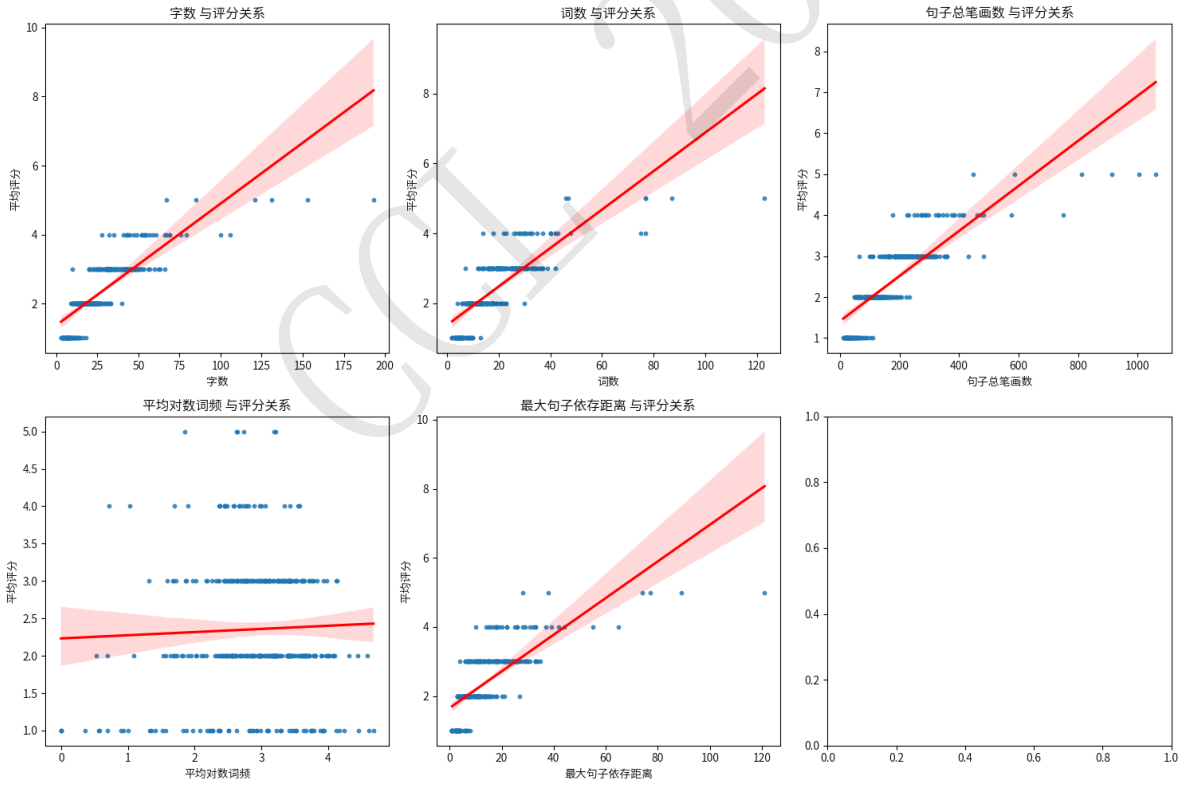

图 6: OLS 回归结果

我们可以观察到，句长和句长相关的变量（总笔画数、最大句子依存距离）都与“平均评分”具有显著的相关性。字数越多，平均评分越高。每增加一个字，平均评分预计增加

约0.03分。句子总笔画数越多，平均评分越高。这可能反映了某种“视觉复杂度”或“信息密度”与评分的关系。在控制了字数和句子总笔画数等变量后，词数对平均评分没有显著的独立影响。这很可能是由于与字数等变量存在多重共线性。

句长无关的因素平均对数词频在控制了其他变量后，对平均评分也没有显著的独立影响。

附录 B. 六种Prompt形式（以三分类为例）

prompt_zero_shot = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：
1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；
3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown、注释或解释说明

<输入>
{sentence}
</输入>
"""

prompt_zero_shot_exp = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：
1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；
3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。
在做出难度等级判断的同时，请你解释你做出这个难度判断的理由。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
- Explaination: 代表模型对于句子难度的解释，数据类型为 Str ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown或注释

<输入>
{sentence}
</输入>
"""

prompt_zero_shot_expto = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：
1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；

第二十四届中国计算语言学大会论文集，第666页-第683页，济南，中国，2025年8月11日至14日。

3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。
在做出难度等级判断的同时，请你以向一个具有基本义务教育知识水平的成年人为对象，用简介易懂的词汇，将这个句子的意思尽可能解释清楚。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
- Explaination: 代表模型对于句子意义的解释，数据类型为 Str ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown或注释

<输入>
{sentence}
</输入>
"""

prompt_few_shot = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：
1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；
3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown、注释或解释说明

<例子>
示例 1：
输入句子：陕北窑洞是中国西北地区的一种传统民居形式，冬暖夏凉适合居住。
输出结果：{"Score": 1}

示例 2：
输入句子：公司应用模仿蚁群的算法优化物流，提升仓储中心30%调度效率。
输出结果：{"Score": 2}

示例 3：
输入句子：阶层固化的合理化叙事，正遭遇风险社会流动性范式解构。
输出结果：{"Score": 3}

<输入>
{sentence}
</输入>
"""

prompt_few_shot_exp = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：

1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；
3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。
在做出难度等级判断的同时，请你解释你做出这个难度判断的理由。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
- Explaination: 代表模型对于句子难度的解释，数据类型为 Str ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown或注释

<例子>
示例 1:
输入句子：陕北窑洞是中国西北地区的一种传统民居形式，冬暖夏凉适合居住。
输出结果：{"Score": 1,
"Explaination": "句子结构平直，用词常见，只要有基本地理常识就能理解\陕北窑洞"\冬暖夏凉"\适合居住"这些概念，对成年人而言十分通顺明了。"}

示例 2:
输入句子：公司应用模仿蚁群的算法优化物流，提升仓储中心30%调度效率。
输出结果：{"Score": 2,
"Explaination": "句中包含\蚁群算法"\物流调度效率"等专业术语，对计算机科学或物流管理不了解的读者需要一定思考才能把握\算法如何应用于优化"和\效率提升30%"的含义。"}

示例 3:
输入句子：阶层固化的合理化叙事，正遭遇风险社会流动性范式解构。
输出结果：{"Score": 3,
"Explaination": "本句使用了高度抽象的社会学术语||\阶层固化"\合理化叙事"\风险社会"\流动性范式"\解构"等，概念叠加、逻辑隐晦，普通成年人即使反复揣摩也难以精准理解其深层涵义。"}

<输入>
{sentence}
</输入>
"""

prompt_few_shot_expto = """你将收到一个包含难度分级的句子数据集，句子的难度分为三个等级，这三个等级在数据集中的数量相同，请你根据句子内容判断难度等级，以 JSON 格式返回结果。

难度等级共分为三个层级（1 到 3），定义如下：
1（简单）：受过义务教育的成年人能够流畅轻松地看懂；
2（中等）：包含较复杂概念或逻辑，需经过一定思考后能完全理解；
3（困难）：需要大量思考，或者理解后仅能大致把握含义。
请从 1、2、3 中选择最符合该句子难度的等级。
在做出难度等级判断的同时，请你以向一个具有基本义务教育知识水平的成年人为对象，用简介易懂的词汇，将这个句子的意思尽可能解释清楚。

输出字段包括：- Score: 代表句子难度的等级，数据类型为 Int ，不可为空。
- Explaination: 代表模型对于句子意义的解释，数据类型为 Str ，不可为空。
输出格式要求（严格遵守）：- 仅输出一个 JSON 对象，不包含 Markdown或注释

<例子>
示例 1：
输入句子：陕北窑洞是中国西北地区的一种传统民居形式，冬暖夏凉适合居住。
输出结果：{"Score": 1,
"Explaination": "陕北窑洞是一种在中国西北地区常见的传统房子。它们建在山坡上，冬天暖和、夏天凉快，很适合人住。"}

示例 2：
输入句子：公司应用模仿蚁群的算法优化物流，提升仓储中心30%调度效率。
输出结果：{"Score": 2,
"Explaination":"仿生学是模仿自然界（比如蚂蚁）的做法来解决问题。蚁群算法是一种学蚂蚁怎么找路的方法。这个方法被用来改进货物运输，让仓库的工作效率提高了30%。"}

示例 3：
输入句子：阶层固化的合理化叙事，正遭遇风险社会流动性范式解构。
输出结果：{"Score": 3,
"Explaination": "有些人用\社会阶层本来就固定"这种说法来解释现实（这叫\合理化叙事"）。但现在社会变化太快，这种说法开始被挑战，人们越来越觉得社会阶层应该可以变化（这叫\社会流动性"），原来的那套解释也开始不管用了。"}

<输入>
{sentence}
</输入>
"""