题目：**基于大语言模型数据增强的文本评测研究**

论文提纲

1. **引言**

**（**本章节将详细阐述自然语言处理领域中文本评分任务的重要性及其面临的挑战。特别指出标注数据不足对模型性能的影响，并介绍数据增强技术在解决这一问题中的潜在价值。

* 1. **研究背景及意义**

在自然语言处理（NLP）领域，尤其是在文本评分任务中，数据集的质量和规模对模型性能有着决定性的影响。然而，高质量的标注数据往往难以获得，因为它们需要大量的人力和时间来创建和验证。此外，对于特定领域，如罕见语言或专业主题，高质量的标注数据可能微乎及微。而在其中，数据增强技术可以通过生成额外的训练样本来扩大和多样化数据集，从而提高模型的泛化能力和鲁棒性。在文本评分领域，这意味着能够更准确地评估文本的质量，无论是在教育、出版还是内容审核等领域。本研究旨在探讨如何通过数据增强技术，特别是利用大语言模型和文本蒸馏方法，来提高文本评分模型的性能。本研究提出了一种基于大语言模型的数据增强稳步评测研究，探究在数据紧缺的情况下，借由大语言模型进行数据增强和数据蒸馏方法，是否能够得出用于训练的较好质量的数据集，以提升文本分类模型的性能。

* 1. **国内外研究现状分析**

在自然语言处理领域，数据增强技术已经成为提升模型性能的重要手段，特别是对于文本分类任务。当前对于文本数据不足想出来的数据增强方法有很多，如Wei等人[1]发明的EDA方法通过同义词替换，随机插入，随机删除，随机交换，回译，文本重组的方式对数据进行增强，从而实现数据的多样化。Qian等人[2]发明的AEDA则通过更简单地插入标点符号的方式来进行数据增强，更加简化了对文本的处理，实现处理的轻量化。还有研究者通过相似度的计算来进行未标注文本的分类，变向实现从未标注数据到标注数据的转化，从而实现数据增强(如Su等人[5])。

还有一些研究者，为了适应不同领域作文所带来的作文分布差异，通过优化模型的方式构建了一些鲁棒性较好的跨领域作文自动评分模型。

随着近些年transfomer的发展，大语言模型如GPT等被证实在数据增强上具有较高的可靠性，因此也出现了其他数据增强的方法，比如运用大语言模型进行数据增强。

然而，在现有的方法中，传统的方法会使得模型学到的特征比较单一，使得训练得出的模型灵活性较低。跨领域的作文评分模型能够通过模型较好的鲁棒性实现较为不错的评分效果，但是建模却较为复杂，且难以适应不同评分标准的作文数据集。而现代运用大语言模型生成数据的方法优化大多在于修改大语言模型的提示词（即运用提示词工程）以得到更高质量的文本（如Chen等人[4]的研究）。但由于大语言模型生成的随机性，即使是通过缜密的提示词工程得到的文本依然无法保证所有文本的高质量性。

总的来说，现有方法面临着三重挑战：传统文本增强易导致特征单一化学习；跨领域模型构建存在复杂度与适配性矛盾；大语言模型生成数据的质量不可控性。这些瓶颈制约了自动评分系统的实用化进程，亟待建立有效的生成数据质量控制机制。

* 1. **本文研究内容**

本研究提出基于大语言模型及数据蒸馏的双阶段增强框架LLM-DA，通过生成-蒸馏协同机制提升文本评分模型性能。研究内容包含三个核心模块，分别为数据增强模块，数据蒸馏（过滤）模块以及数据改写（再生成）模块。通过此架构可以得出较优质的增强数据集，为模型构建提供良好训练集。

总体来说，本研究的贡献在于：1. 提出了LLM-DA双阶段增强框架，解决了大语言模型因为生成的随机性导致数据质量参差不齐的情况。2. 将数据增强运用于AES模型构建当中，并通过蒸馏得到的高质量数据集得出了评分能力良好的模型。3. 通过将模型进行对比分析，证明了此方法优于传统数据增强，及未经过数据蒸馏训练得出的模型。

* 1. **论文结构**

相关研究（本章节将概述一些面对数据不足的当今研究方法。包括当前数据增强技术的发展状况以及本研究旨在解决的具体问题：如何通过数据增强技术提升文本评分模型在数据稀缺情况下的性能。）

任务定义（本章节将介绍aes评分任务）

研究方法介绍（本章节将详细介绍本研究采用的数据增强技术，包括大语言模型的文本生成方法和文本蒸馏技术的原理及其工作流程）

模型构建与实验设计（本章节将描述基准模型的构建过程，包括多种不同条件下的模型：未经过数据增强的模型、经过大语言模型数据增强但没有经过蒸馏的模型、经过大语言模型数据增强且数据经过蒸馏的模型，运用其他数据增强方法（如eda）得出的模型，以及通过其他非此主题文章训练得出的模型。详细说明实验的设计，包括数据集的选择、大模型提示词的更改，模型的选择和验证过程，以及性能评估的标准。）

1. **相关技术介绍**
   1. **机器学习技术**

机器学习是一种让计算机通过数据来学习和改进的技术，它使计算机能够在不进行明确编程的情况下执行任务。机器学习的核心思想是让计算机从大量的历史数据中自动学习规律和模式，然后利用这些规律和模式对新的数据进行预测或决策。

机器学习的主要方法包括监督学习、无监督学习和强化学习。监督学习是指利用带有标签的训练数据来训练模型，模型通过学习输入数据与输出标签之间的映射关系来预测新的输入数据的输出标签。无监督学习则是利用没有标签的数据来发现数据中的内在结构和规律，例如聚类分析和降维等。强化学习则是通过让智能体在环境中进行试错学习，根据环境反馈的奖励信号来调整行为策略，以达到最大化累积奖励的目标。

在文本评测领域，机器学习技术可以用于构建文本评分模型，通过对大量已标注的文本数据进行学习，模型能够掌握文本质量的评估标准，从而对新的文本进行自动评分。

* 1. **深度学习技术**

深度学习是机器学习的一个重要分支，它通过构建具有多个层次的神经网络模型来学习数据中的复杂模式和特征。深度学习模型通常由输入层、多个隐藏层和输出层组成，每一层都对输入数据进行非线性变换，从而能够自动提取数据中的高级特征。

常见的深度学习模型包括卷积神经网络（CNN）和循环神经网络（RNN）。CNN主要用于处理图像数据，通过卷积层和池化层的操作来提取图像的局部特征和空间层次结构。RNN则适用于处理序列数据，如文本、语音等，它具有循环结构，能够对序列中的历史信息进行建模和利用。此外，近年来出现的Transformer模型在自然语言处理领域取得了巨大的成功，它通过自注意力机制能够有效地捕捉文本中的长距离依赖关系。

深度学习技术在文本评测中可以用于构建更强大的文本评分模型，通过学习文本的深层次语义特征，提高模型的评分准确性和泛化能力。

* 1. **大语言模型技术**

大语言模型是一种基于深度学习技术的自然语言处理模型，它通过在大规模的文本数据上进行无监督学习来获取语言知识。大语言模型通常具有数十亿甚至上百亿的参数，能够对自然语言进行理解和生成。

大语言模型的训练过程主要包括预训练和微调两个阶段。预训练阶段，模型在大规模的无标注文本数据上进行训练，学习语言的统计规律和语义知识。微调阶段，则是在特定的自然语言处理任务上对预训练模型进行进一步的调整和优化，使其更好地适应具体的应用场景。

在文本评分任务当中，大语言模型可以用于生成高质量的文本数据，通过数据增强的方法来扩充训练集，从而提高文本评分模型的性能。此外，大语言模型还可以作为评分模型的一部分，利用其对语言的理解能力来辅助评估文本的质量。

1. **基于大语言模型的数据增强**
   1. **研究动机**

在文本评分任务中，标注数据的规模和质量对模型性能起着决定性作用。然而，获取大量高质量标注数据往往成本高昂且耗时费力，尤其在专业领域或罕见语言中，标注数据稀缺。数据增强技术成为解决这一问题的关键手段，能够通过生成额外训练样本，提升数据集的规模和多样性，增强模型的泛化能力和鲁棒性。

传统数据增强方法，如 EDA 及其变体 AEDA，通过同义词替换、插入标点等操作实现数据多样化，但可能使模型学到的特征单一，限制了模型灵活性。跨领域作文评分模型虽具备一定鲁棒性，但构建复杂，难以适应不同评分标准的数据集。而基于大语言模型的数据增强方法，虽能生成高质量文本，但其生成的随机性导致文本质量参差不齐，需要有效的质量控制机制。

本研究的动机正是针对现有方法的不足，探索利用大语言模型和文本蒸馏技术，提出一种创新的数据增强框架，以克服传统方法的局限性，提高文本评分模型在数据稀缺情况下的性能，推动自动评分系统的实用化进程。

* 1. **数据集的选择**

本研究采用Kaggle平台发布的ELLIPSE Corpus[[1]](#footnote-1)数据集作为基准数据源，其选取依据主要因为：(1) 其包含多维度的语言特征标注，如作文分数涵盖语法，主题一致性等不同角度的细分；(2)其拥有完整的评估体系，提供详尽的1-5分量表评分标准；(3)其涵盖超过2000条经过专业评分的英语学习者写作样本。为了构建符合研究目标的实验数据集，本研究基于主题一致性原则，选定了"Distance Learning"主题下297篇文本作为基础语料库。其次，本研究采用分层随机抽样方法，按0.5分间隔将1.0-5.0评分区间分为9个层级，并从每个层级随机抽取一篇代表性样本，构建包含9篇文本的初始文本数据集，该设计有效模拟真实场景中数据严重缺失的情况，即每一个评分只有一篇文章。最后，将剩余的288篇文本构成验证集，用于模型评分能力评估。

针对原始评分体系，本研究决定对数据执行离散化处理，将1.5-5.0的连续评分向下取整至最邻近整数，形成1-5分的五分类评估体系。该处理方法既能符合ELLIPSE数据集官方发布的整数分评分规则，又能降低模型复杂度。

* 1. **LLM-DA框架实现数据增强**

接下来，将重点介绍我们实现数据增强以及蒸馏的LLM-DA框架，其具体框架图如下图x所示，其主要包含3个核心模块：

大语言模型数据增强模块：通过当前给定的作文评分标准，构建提示词模板，运用大语言模型生成大量对应评分的数据。

数据蒸馏模块：设计基于词汇复杂及多样性指数评估，文章句型运用与主题连贯性评估，以及基于规则语法检测的三通道过滤机制，对进入的数据集进行蒸馏过滤。将数据集进行分类，分离出符合要求的高质量数据集和不符合要求的低质量数据集。

数据自适应循环改写模块：记录下不符合要求的低质量数据集，根据其未达标的指标构建不同的大语言提示词模版，重新利用大语言模型进行修改，直至符合数据过滤模块的标准，得到质量良好的数据集。

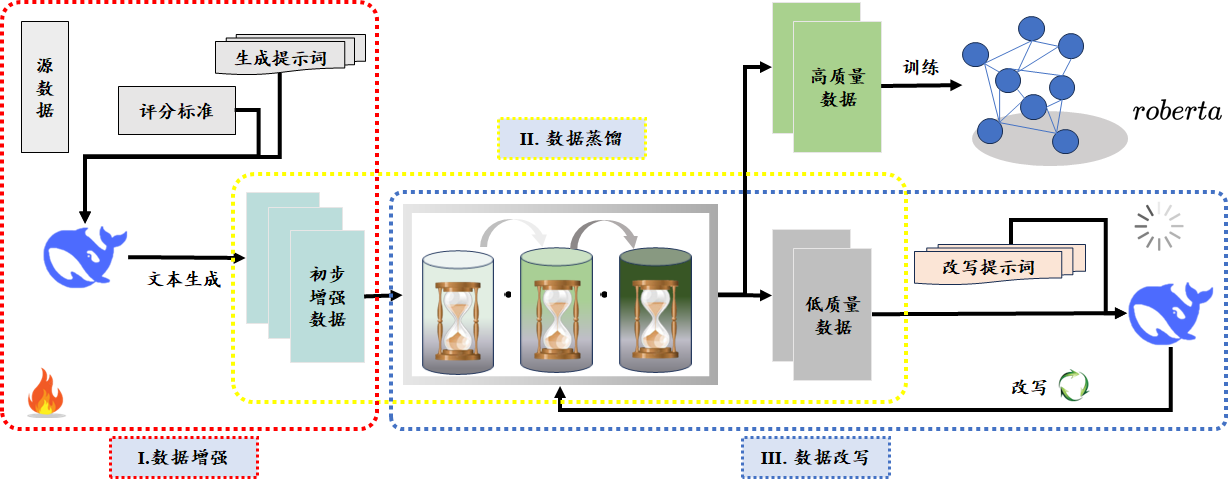


图 x LLM-DA框架

* + 1. **数据生成模块**

在本模块当中，大语言模型将根据提示词生成对应的文本，实现初步的数据增强。在提示词的构建当中，本研究结合了相关的官方评分标准（如下表I所示）与对应分数来进行提示词模版的构建，具体的构建思路及模版如下图1所示，其中代表作文分数，代表对应作文分数的评价指标，代表作文主题，本研究中，代表’Distance Learning’。

本研究用作文分数来代替写作者的水平，分数越低意味着此篇作文的作者水平越低，同时结合了表I的评分标准，构建了图1精密的提示词模版。在大语言模型的选用中，我们采用了7B的deepseek蒸馏模型，有三点考量：其一，deepseek:7B拥有深度思考的优势，使得其能够基本思考并读懂提示词的要求，生成较高质量的文本。其二，此模型参数量较低，方便本地部署和运行，对硬件的需求不高，且免费开源无需有API收费。

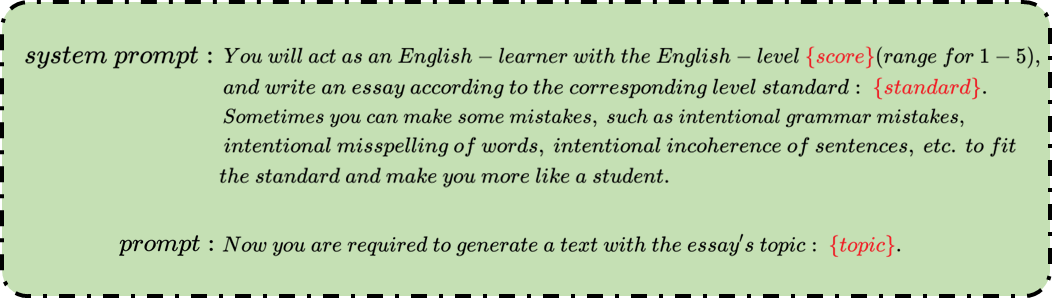


图1 生成模块的提示词构建模版

表I 评分标准

|  |  |
| --- | --- |
| 分数 | 标准 |
| 5 | Native-like facility in the use of language with syntactic variety, Appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication. |
| 4 | Facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication. |
| 3 | Facility limited to the use of common structures and generic vocabulary;  organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases. |
| 2 | Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies. |
| 1 | A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies. |
|  |  |

* + 1. **数据蒸馏**

本研究采用了三通道过滤机制，即基于词汇复杂及多样性指数评估，文章句型运用与连贯性评估，以及基于规则语法检测的过滤机制，从词汇，文章连贯性，语法运用等角度过滤数据，筛选出符合标准的数据。

1. 基于词汇复杂及多样性指数评估：本通道的过滤算法会根据初始的9份作文数据集，计算其词汇丰富度，并建立与此作文词汇得分的非线性回归函数，由此得出词汇复杂及多样性的指数评估。具体而言，本研究先会将文章进行分词，去除停用词等操作，然后记录下不重复出现词语占全词语的比例，以此得到文章的词汇指数。特别地，为凸显词汇复杂度对整体语言表达能力的影响，本研究对长度超过八个字符的词汇赋予更高的权重，即每个此类词汇相当于1.5个普通词汇的权重。本次研究词汇复杂度指数计算运用二次函数进行拟合，其具体的公式如下公式一，二所示，所拟合的曲线函数如下图2所示：

公式（一） 公式（二）

公式（三）

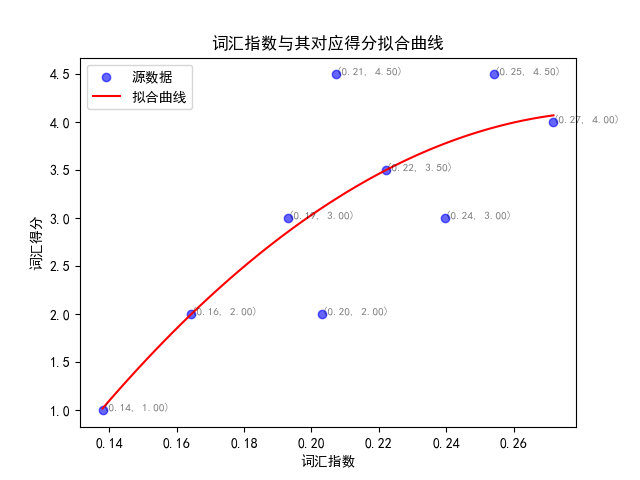


图2 词汇指数计算的拟合曲线

上述公式中，是经过预处理（去除停用词、标点符号和数字，并统一为小写）后的单词列表。是中去重后的单词集合。是单词的权重，具体定义为上述公式三所示。为预测的词汇得分，为计算词汇复杂度与丰富度指数的函数。，，为拟合好的参数值。

2. 基于文章句型运用与连贯性评估：与词汇指数计算相类似，本算法也采用了拟合的相关算法，对文章进行连贯性即句型运用的指数评估。具体而言，本研究会细分求出文章的连接词，连接短语以及高级连接句型的使用频率，其具体的构建分类如下图3所示：

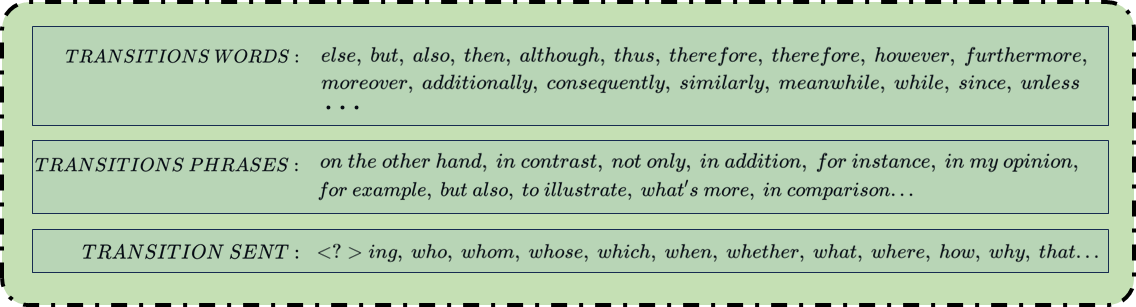


图3 文章连贯词，短语句子的匹配指标

由于一个句子可能出现多个连接词或连接短语，但出现的连接句如从句连接句的比例较少。两者的评判标准差异过大，无法简单的如同词汇指数计算那样获得权重标准。因此，本研究将其连接句和短语分成一类，连接句分为一类进行平面的拟合，算出对应的权重关系，定义拟合函数为：

公式（四）

其中和分别为连接词，连接短语及连接句使用的频率，具体计算方式为下面公式五，六所示。为拟合函数，也被用作此通道的过滤函数。其中与分别是小写去停用词之后的词语和句子集合，，，分别是连接词，连接短语，连接句的集合。

公式（五）

公式（六）

对应的拟合平面如下图6所示：

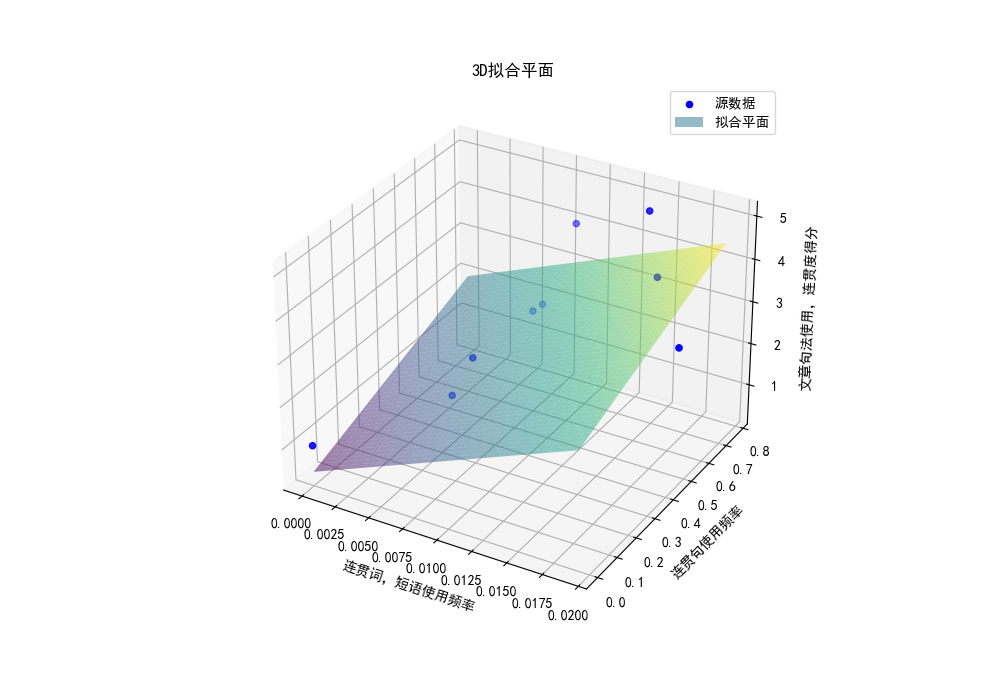


图3 文章连贯度拟合平面

3. 基于规则的语法检测：在本研究中，语法检测工作通过language-tool-python工具予以实现。选择该工具的原因主要基于以下三点考量：其一，该工具具有开源属性，凭借一套可灵活定制的规则体系，能够精准检测语法及拼写层面的错误，且无需额外下载或运行复杂模型，使用起来便捷高效；其二，从检测效能来看，它具备出色的识别能力，涵盖诸多常见语法错误类型，诸如主谓一致、时态运用不当、冠词误用等问题，同时对拼写错误也能精准捕捉；其三，该工具配备了实用的纠错功能，可在检测出错误后，为用户提供合理且有效的修改建议，助力文本质量提升。

在过滤算法方面，本研究引入相关拟合函数作为过滤函数。在分析样例数据时发现，不同英语水平学习者撰写的论文在语法质量上存在显著差异。例如，在所统计的样例数据中，评分仅为1分的文章之一经检测后发现存在超过100条语法错误，而评分高达5分的文章语法错误数量仅为3个。面对如此巨大的语法错误数量差异，为了使过滤算法能够精准适配不同水平文本的处理需求，本研究经过严谨论证，最终选定指数函数进行拟合，并依据该函数构建过滤算法，以实现对文本的高效、精准处理。具体的计算公式和拟合函数图如下公式七和图4所示所示

公式（七）

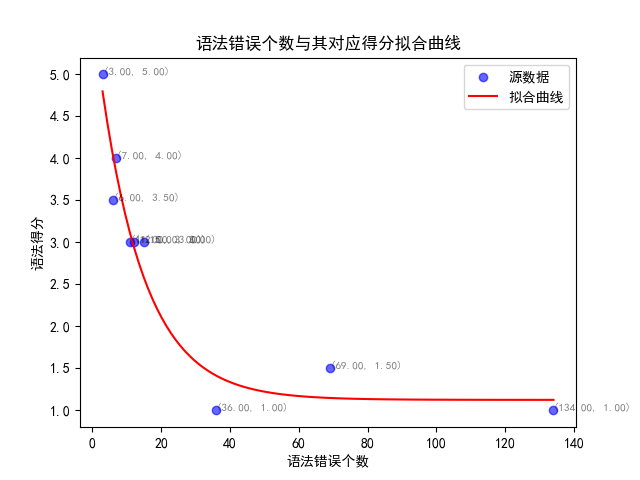


图4语法得分的拟合曲线

数据在经过三重过滤的蒸馏操作之后，筛选其与实际要求分数正负1分之内，且三个分数的平均分在实际要求分数正负0.5分之内的数据，作为高质量数据输出。而剩下的不符合要求的数据作为低质量数据输入到数据改写模块，对生成的内容进行进一步的改写。

* + 1. **数据改写**

本研究中，我们针对第二轮数据蒸馏得到的低质量数据，采用了改写而非重新生成的策略。这一选择基于以下考虑：直接使用大型语言模型进行生成，其结果往往具有不稳定性。相比之下，让模型对已有的文本进行修改，能够更高效地获得高质量的样本。因此，我们选择了对第一轮生成的文本进行改写，而非完全重新生成。在改写阶段，为了让大语言模型更加理解提示词的含义，我们选择了更大参数量的deepseek:14B模型对数据生成阶段deepseek:7B模型生成的数据进行更改，使其更能适应更改任务。

对于更改提示词模型的设置，我们利用数据蒸馏得出的三个结果值（词汇得分，连贯得分，语法得分）进行统计，并根据得分差异构建对应的更改提示词模版，对应构建模版如下图5所示。

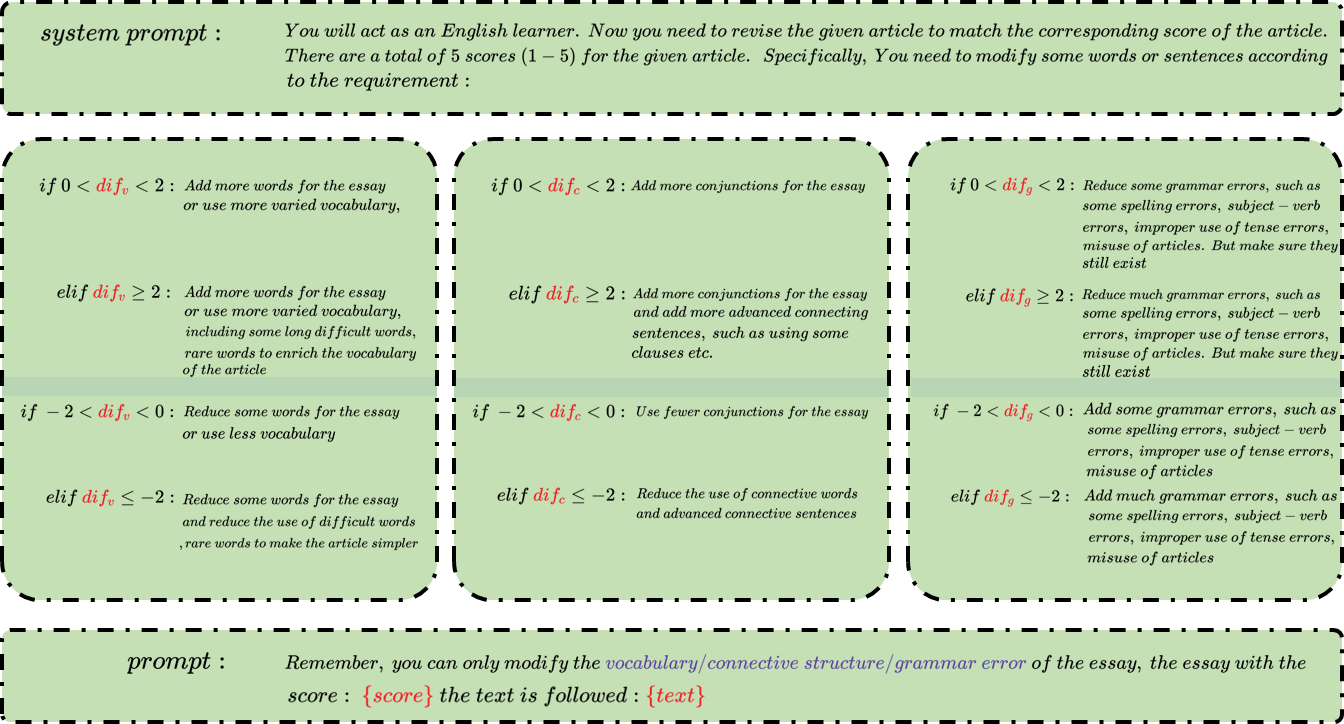


图5改写提示词构建模版

在上图当中，，,分别代表了需要的实际分数与蒸馏过滤算法计算的词汇得分，连贯得分，语法得分相减的值，例如，若需要一篇4分的文章，但实际生成经过蒸馏过滤算法得到了3分的词汇得分、5分的连贯得分和4分的语法得分，此时的，,分别为1，-1，0）。其计算公式如下公式八所示。

公式（八）

上述公式中，代表需要的实际分数，表示由蒸馏过滤算法得出的评估得分。值越高，说明生成的文章质量越低于实际要求，需要加强；反之则说明质量越高，可能需要适当降低。

在上述模版中，代表了实际的需求得分，代表了原始文章。我们根据值的大小，分别使用不同的提示词模板对文章进行处理，使大语言模型在改写文章时更加灵活。对于过滤和改写的顺序，本研究选择：先改写词汇，再改写连贯性，最后改写语法。这一顺序是基于预实验的结果，发现先进行语法改写可能会在后续的词汇或连贯性改写时再次改变句子的语法，因此我们决定先对文章进行影响较小的改写，再进行影响较大的改写。

1. **实验结果与分析**
   1. **数据集**

本次试验所采用Kaggle平台发布的ELLIPSE Corpus[[2]](#footnote-2)的Distance Learning主题的剩余288篇数据作为测试所用的数据集。模型同意采用预训练的roberta分类器进行训练，这是由于

* 1. **基准模型**
  2. **评价指标**
  3. **结果分析与讨论**

**5 总结**

结果分析与讨论（本章节将展示实验结果，对比不同模型的性能，并分析数据增强与文本蒸馏技术对模型性能的影响。通过对比分析，展示本数据增强技术在经过数据蒸馏之后如何提升模型在数据稀缺情况下的性能）

局限性探讨（本章节将讨论模型可能会存在的潜在问题，以及其对应的可能解决方法，表面后续研究努力的方向）

1. https://www.kaggle.com/datasets/matthewjansen/ellipse-corpus/data [↑](#footnote-ref-1)
2. https://www.kaggle.com/datasets/matthewjansen/ellipse-corpus/data [↑](#footnote-ref-2)