# 基于大语言模型数据增强的文本评测研究

# 摘　要： 本文研究了基于大语言模型的数据增强技术在文本评分任务中的应用。数据增强技术通过生成额外的训练样本，能够有效扩大数据集规模并提高模型的泛化能力。而本研究提出了一种基于大语言模型和数据蒸馏的双阶段增强框架（LLM-DA），通过生成-蒸馏协同机制提升增强的数据质量，以提升文本评分模型的性能。该框架包含数据增强、数据蒸馏和数据改写三个核心模块，能够在数据稀缺的情况下生成高质量的训练数据集。实验结果表明，该方法优于传统数据增强方法及未经过数据蒸馏的模型，为文本评分任务提供了一种有效的解决方案。 ****关键词****：数据增强；大语言模型；文本评分；数据蒸馏；自然语言处理

Abstract：This paper investigates the application of data augmentation techniques based on large language models (LLMs) in text scoring tasks. Data augmentation can effectively expand the scale of the dataset and improve the generalization ability of models by generating additional training samples. This study proposes a dual-stage data augmentation framework based on large language models and data distillation (LLM-DA), which enhances the quality of augmented data through a generation-distillation collaborative mechanism, thereby improving the performance of text scoring models. The framework comprises three core modules: data augmentation, data distillation, and data rewriting, and is capable of generating high-quality training datasets under data-scarce conditions. Experimental results demonstrate that this method outperforms traditional data augmentation methods and models without data distillation, providing an effective solution for text scoring tasks.

**Keywords**：Data Augmentation; Large Language Models; Text Scoring; Data Distillation; Natural Language Processing

目 录

[摘　要 1](#_Toc193813753)

[Abstract： 1](#_Toc193813754)

[1. 引言 3](#_Toc193813755)

[1.1 研究背景及意义 3](#_Toc193813756)

[1.2 国内外研究现状分析 3](#_Toc193813757)

[1.3 本文研究内容 4](#_Toc193813758)

[1.4 论文结构 5](#_Toc193813759)

[2 相关技术介绍 5](#_Toc193813760)

[2.1 机器学习技术 5](#_Toc193813761)

[2.2 深度学习技术 6](#_Toc193813762)

[2.3 大语言模型技术 6](#_Toc193813763)

[3 基于大语言模型的数据增强 7](#_Toc193813764)

[3.1 研究动机 7](#_Toc193813765)

[3.2 数据集的选择 8](#_Toc193813766)

[3.3 LLM-DA框架实现数据增强 9](#_Toc193813767)

[4 实验结果与分析 17](#_Toc193813768)

[4.1 数据集 17](#_Toc193813769)

[4.2 基准模型 17](#_Toc193813770)

[4.3 评价指标 18](#_Toc193813771)

[4.4 结果分析与讨论 19](#_Toc193813772)

[5. 总结 21](#_Toc193813773)

[参考文献 21](#_Toc193813774)

# 引言

## 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其微。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。

## 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AIGC技术的发展，大语言模型如GPT[[1]](#footnote-1)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。

## 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。

## 论文结构

1.引言以及研究方法介绍：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。此外，将概述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。同时还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。

2.相关技术介绍：本章节将介绍文本设计到的相关技术，概述讲述相关技术的发展内容

3.模型构建与实验设计：本章节将描述基准模型的构建过程，包括四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，以及初始数据量丰富的模型。详细说明实验的设计，包括数据集的选择、模型的训练和验证过程，以及性能评估的标准。

4.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，展示数据增强技术如何提升模型在数据稀缺情况下的性能，并讨论其在不同领域和任务中的适用性。

# 相关技术介绍

## 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。

## 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[19]和循环神经网络RNN以及LSTM[20]。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和LSTM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。随着近些年技术的发展，一些更复杂的深度学习模型被提出，比如将要介绍了大预言模型，以及其衍生的T5，BERT模型等等。

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。

## 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。大语言模型主要为Transformer结构，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。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。Transformer是由2017年的论文《Attention is All You Need》中被Vaswani等人在 [16]中首次提出，奠定了后续大语言模型的发展方向。现如今如OpenAI的GPT系列和Google的BERT[17]，或是T5[21]等模型都是基于Transformer架构的变体，这些模型在文本生成、问答系统、文本分类等任务中取得了显著的成果。包括本研究训练所用到的RoBERTa[18]预训练模型也是基于Transformer架构提出的，针对BERT改进的模型。

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。

# 基于大语言模型的数据增强

## 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，但其本质并没有对作文内容进行实质性的更改，着会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，且难以适应不同评分标准的数据集。而目前的一些基于大语言模型的数据增强方法，虽能生成高质量文本，但由于其生成的随机性，会导致生成的文本质量参差不齐，因此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LLM-DA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。

## 数据集的选择

本研究采用Kaggle平台发布的ELLIPSE Corpus[[2]](#footnote-2)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篇文本构成测试集，用于模型评分能力评估。初始数据集的样例格式大致如下表I所示：

表I 样例数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 主题 | 文章 | 总得分 | 词汇得分 | 连贯得分 | 语法得分 | 名字性别等 |
| Distance  Learning | Text1 | 2 | 2.5 | 1.5 | 2 | ... |
| Career Commitment | Text2 | 3 | 2.5 | 3 | 3.5 | ... |
| Success And Failure | Text3 | 4.5 | 4 | 4.5 | 4.5 | ... |
| Distance  Learning | Text4 | 1 | 1 | 1 | 1 | ... |
| ... | ... | ... | ... | ... | ... | ... |
| Lunch Menus | TextN | 5 | 4.5 | 4.5 | 5 | ... |

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。

## LLM-DA框架实现数据增强

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1所示，其主要包含3个核心模块：

1. 大语言模型数据增强模块（数据生成）：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。

2. 数据三通道蒸馏过滤模块（数据蒸馏）：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。

3. 数据自适应循环改写模块（数据改写）：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。

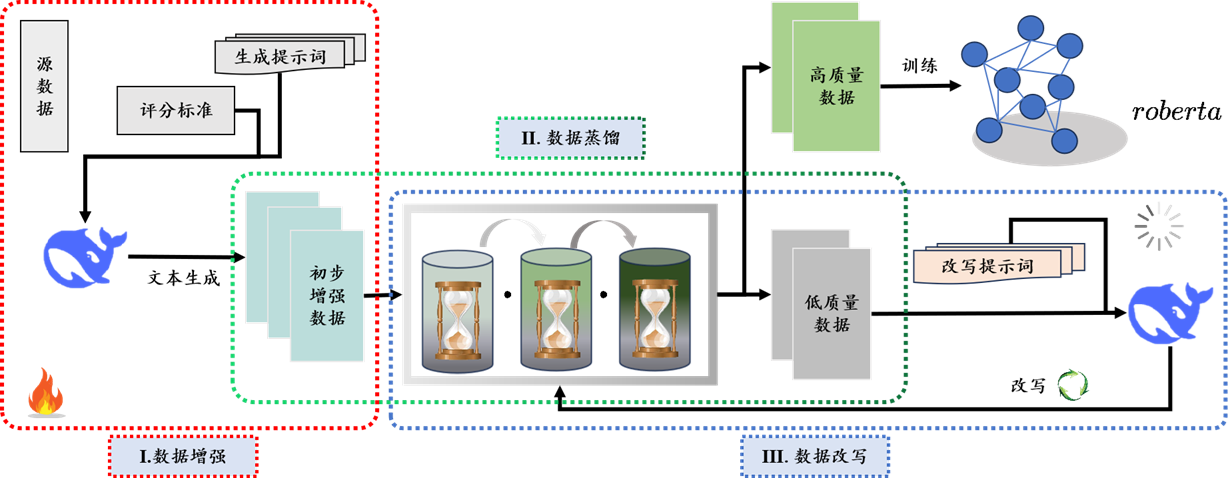


图 1 LLM-DA框架

### 数据生成模块

在本模块当中，大语言模型将结合评分标准，构建对应的生成提示词，给机器生成对应的文本，实现初步的数据增强。在提示词的构建当中，本研究结合了ELLIPSE Corpus[[3]](#footnote-3)官方评分标准（如下表II所示）进行构建，具体的构建思路及模版如下图2所示，在图中，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，在本研究中既代表’Distance Learning’。

本研究将以作文分数代替写对应的语言学习者的写作水平，分数越低意味着此篇作文的作者写作水平越低，在进行提示词构建时结合了表I的评分标准，构建了图2精密的提示词模版。在大语言模型的选用中，本研究采用了7B的DeepSeek蒸馏模型，主要有三点考量：其一，DeepSeek:7B拥有长CoT[[4]](#footnote-4)深度思考的优势，使得其能够基本读懂相应提示词的要求，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。

通过本模块，数据的数量将会得到提升，得到初始经过增强后的数据集，为后续的实验提供了原始的数据材料。

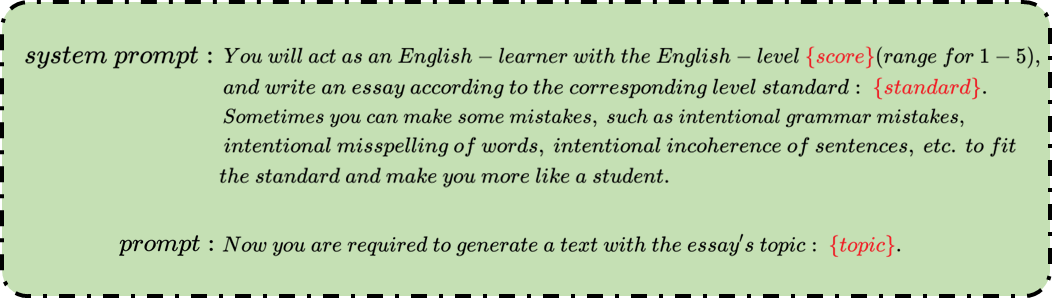


图2 生成模块的提示词构建模版

表II 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. |
|  |  |

### 数据蒸馏

本模块采用了三通道蒸馏过滤机制，这是一种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，及基于规则语法检测的过滤机制。此模块会从词汇，文章连贯性，语法运用等不同的角度蒸馏并过滤数据，筛选出符合标准的数据。

概括而言，初步增强后的文章数据在此三个通道中会被分别计算词汇指数，连贯指数和语法指数。之后，通道会筛选其与实际要求分数正负1分之内，且三个指数的分数平均分在实际要求分数正负0.5分之内的数据，以此作为高质量数据输出。而剩下的不符合要求的数据则作为低质量数据输入到下一部分(数据改写模块)进行进一步的改写。

三个蒸馏过滤通道的具体设置机制如下：

### （1）基于词汇复杂及多样性指数评估：

在本通道中，我们会根据初始的9份作文数据集，计算其词汇丰富度，并建立与此实际作文词汇得分两者的非线性回归函数，由此得出文章词汇复杂及多样性和对应词汇得分的关系曲线。具体而言，在此通道中，文章数据会被进行分词，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作为文章的词汇复杂度和丰富度的指标（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，本研究对长度大于或等于8个字符的词汇赋予更高的权重比例，具体而言，我们令每个长度大于或等于8个字符的词汇相当于1.5个普通词汇的权重。在此通道中，词汇指数计算采用了二次函数进行拟合，其具体的公式如下公式1，2所示，根据9篇现有的样例数据所拟合的曲线函数如下图3所示：

公式（1）

公式（2）

公式（3）

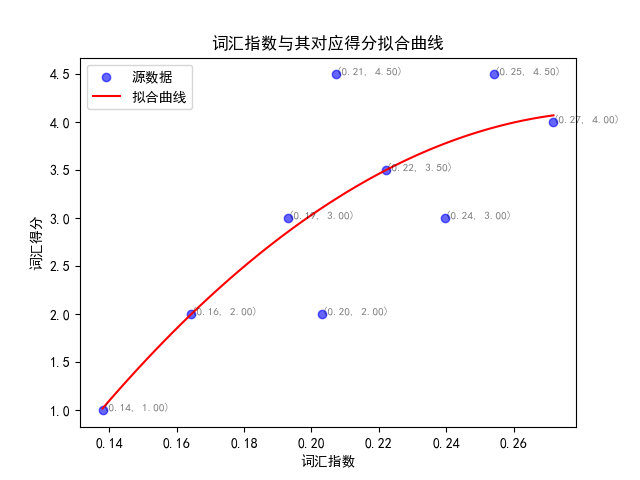


图3 词汇指数计算的拟合曲线

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。

此通道会通过关系函数预测输入文章的词汇得分，并过滤出其得分与实际要求分数正负1分的数据。

### （2）基于文章连贯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，对文章连贯性以及连接句型的运用进行评估（连贯指数）。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：

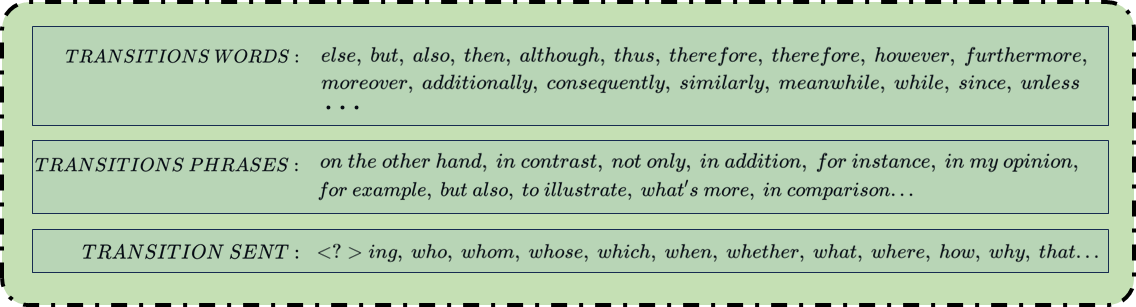


图3 文章连贯词，短语句子的匹配指标

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，二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两者的权重标准。因此，本研究将直接把连接句和短语分成一类，连接句分为一类来进行参数拟合，从而计算出对应的权重关系，定义拟合函数为：

公式（4）

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5，6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4的和既是我们所要找到的权重比例。

公式（5）

公式（6）

根据初始9份数据集所得到的拟合平面如下图4所示，此通道会通过关系函数预测输入文章的连贯得分，并过滤出其得分与实际要求分数正负1分的数据。

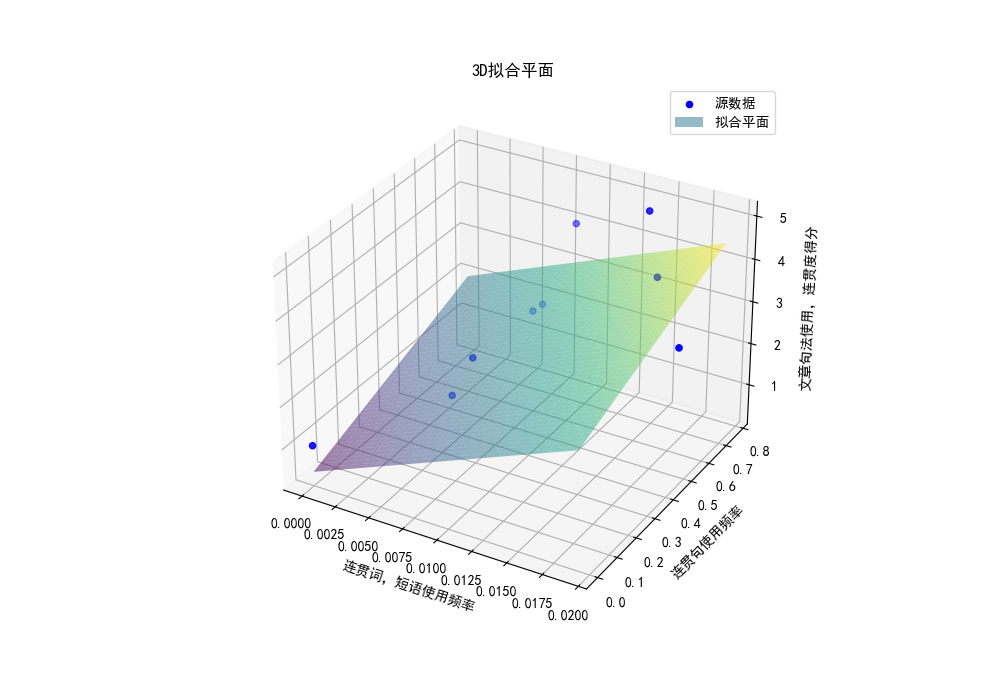


图4 文章连贯指数拟合平面

### （3）基于规则的语法检测：

在本通道当中，语法检测工作运用language-tool-python工具实现。选择该工具的原因主要基于以下三点考量：其一，该工具开源，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，本此通道依然引入相关拟合函数作为过滤函数。我们在初步进行样例数据分析时发现，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，且为了使过滤算法能够精准适配不同水平文本的处理需求，本研究经过严谨考虑，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7所示。

公式（7）

最终经初始9份数据集所得到的拟合关系曲线如下图5所示，此通道会通过关系函数预测输入文章的语法得分，并过滤出其得分与实际要求分数正负1分的数据。

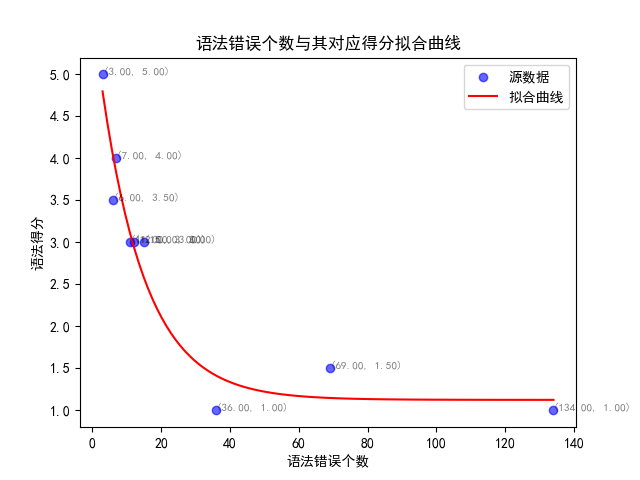


图5语法得分的拟合曲线

### 数据改写

此模块会针对第二轮数据蒸馏过滤得到的低质量数据，采用了改写而非重新生成的策略。这一选择主要是因为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，本研究选择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，我们加入了更大参数量的DeepSeek:14B模型对数据生成阶段中DeepSeek:7B模型生成的数据进行更改，这是由于此模块中的提示词设置更为复杂，为了使模型能够更好理解提示词的要求，此模块中提高了大预言模型的参数量级。

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，具体的构建模版如下图6所示。

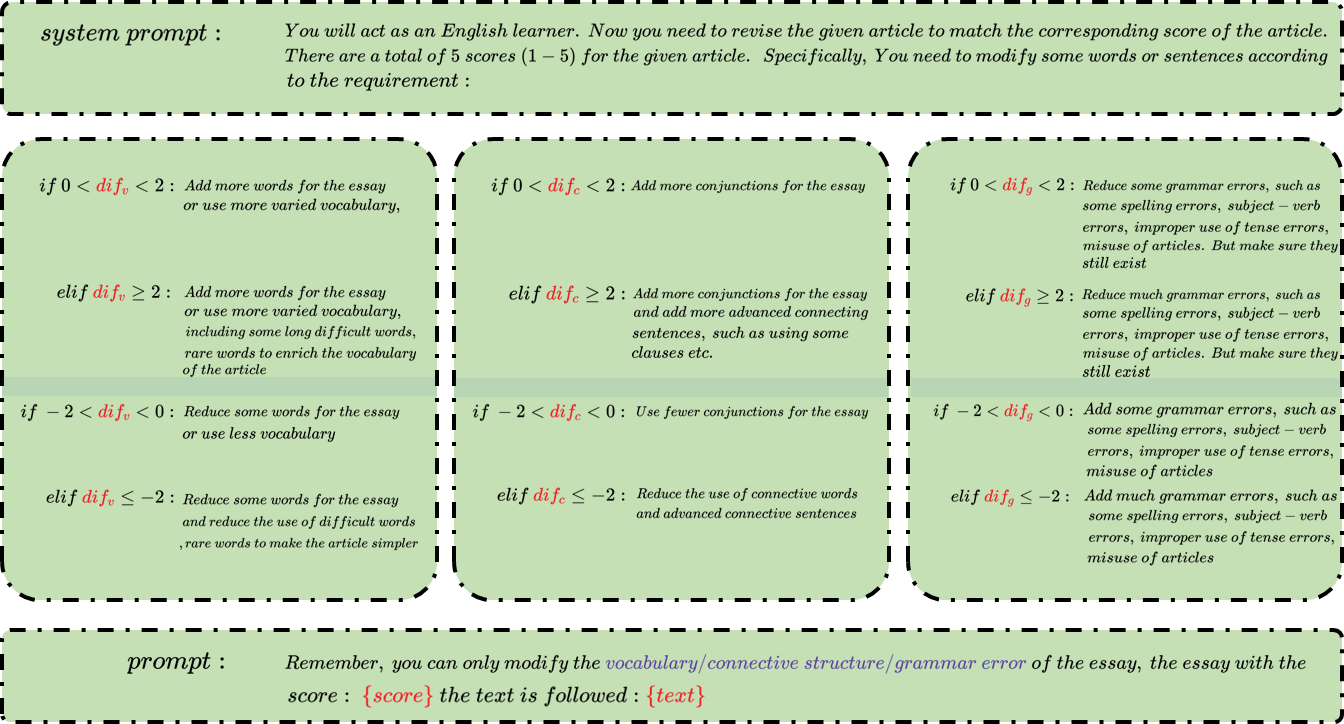


图6改写提示词构建模版

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，但实际生成的文章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8所示。

公式（8）

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，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会对实际结果造成影响，本研究选择：先改写词汇，再改写连贯性，最后改写语法的顺序。这一顺序是基于预实验的结果，实验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。

# 实验结果与分析

## 数据集

本次试验所采用Kaggle平台发布的ELLIPSE Corpus[[5]](#footnote-5)的Distance Learning主题的剩余288篇数据作为测试所用的数据集，对模型进行性能测试。

## 基准模型

本研究主要采用了预训练的RoBERTa分类器作为评测模型，这是由于RoBERTa基于Transformer架构，能够有效捕捉文本中的长距离依赖关系和语义信息。在处理作文评分任务时，它能够理解作文中的复杂句式结构、段落间的逻辑关系以及整体的连贯性，从而更准确地评估作文的质量。与传统的单向语言模型不同，RoBERTa的双向Transformer编码器可以同时利用上下文信息来理解每个词的含义。在作文评分中，这意味着模型能够更全面地理解词汇的语义及其在上下文中的使用是否恰当，判断词汇的搭配是否合理并理解是否存在语义重复或矛盾等问题。相比于BERT，RoBERTa的初始训练数据量远远大于BERT，这使得RoBERTa能够学习到更丰富、更全面的语言特征和语义信息，以及可以更好地理解各种不同的写作风格、主题和语言表达方式，从而提高评分的准确性和泛化能力。

我们一共构建了6种通过不同数据集进行训练的RoBERTa模型，这些数据集分别是1.仅用大语言模型进行数据增强的数据集。2. 用大语言模型进行数据增强，同时进行词汇过滤的数据集。3. 用大语言模型进行数据增强，同时进行词汇，连贯过滤数据集。4. 用大语言模型进行数据增强，同时进行词汇，连贯，语法过滤数据集。5. Elipse-corpus的其他跨主题数据集。6. 经过传统EDA增强的数据集。本研究采用交叉熵损失作为模型训练的损失函数，为了防止模型的过拟合，当平均损失小于0.75就会停止模型的训练，其具体的损失函数公式如下公式9所示：

公式（9）

其中，是样本数量，是类别数量，是样本的真是标签，如果样本属于类别，则，否则为0。是模型属于的概率。交叉熵损失是处理多分类问题的标准选择，因为她基于预测概率分布与真实标签分布之间差异进行计算，能够很好的衡量模型预测的准确性。且其对模型的输出变化敏感，能够提供有效的梯度信号，帮助模型更快地收敛。

## 评价指标

本研究采用的评价指标为QWK（Quadratic Weighted Kappa）值。QWK是一种用于衡量分类任务中模型预测与真实标签之间一致性的指标，常被用于作文自动评分，它考虑了分类的顺序性，对相邻类别之间的错误给予较小的惩罚，而对相隔较远的类别错误给予较大的惩罚。QWK的取值范围通常在-1到1之间，其中1表示完全一致，0表示一致性与随机猜测相同，负值表示一致性低于随机猜测。

在作文自动评分任务中，QWK能够很好地反映模型评分与人工评分之间的一致性程度。通过计算模型预测分数与人工标注分数之间的QWK值，可以评估模型的评分准确性。QWK的计算公式如下公式10所示：

公式（10）

其中，是分类的总数。是一个权重矩阵，用于衡量分类和分类之间的差异。对于二次加权Kappa，权重通常定义为2。是观测到的分类和分类之间的一致性频数。是期望的分类和分类之间的一致性频数（这里假设了分类的随机性）。分式的分子是观测到的加权不一致性总和。分母是期望的加权不一致性总和。

经调研，其对应的QWK值以及评分能力相关标准如下大概可以如下表III所示：

表III 评分能力与QWK值关系

|  |  |
| --- | --- |
| QWK | 评分能力 |
| < 0.2 | 非常低 |
| 0.2~0.4 | 低 |
| 0.4~0.6 | 中等 |
| 0.6~0.8 | 高 |
| 0.8~1.0 | 基本完美 |

## 结果分析与讨论

经过训练，不同训练集所训练出来的RoBERTa模型QWK评分值如下表IV所示：

表IV 不同数据集训练的RoBERTa模型评分能力评估

|  |  |  |
| --- | --- | --- |
| 数据集 | 模型QWK得分 | 评分能力 |
| Elipse-corpus的其他跨主题数据集 | 0.5403 | 中等 |
| 仅用EDA进行数据增强 | 0.3278 | 低 |
| 仅用大语言模型进行数据增强 | 0.1367 | 非常低 |
| 用大语言模型进行数据增强，同时进行词汇过滤 | 0.2190 | 低 |
| 用大语言模型进行数据增强，同时进行词汇和连贯过滤 | 0.3529 | 低 |
| 用大语言模型进行数据增强，同时进行词汇，连贯，语法过滤 | 0.5880 | 中等 |

从QWK值的评分结果可以看出，仅经过大预言模型进行数据增强得到的数据在给模型训练之后，在测试集的QWK得分仅为0.1367，评分能力非常低，甚至只有传统的数据增强方法（EDA）训练出来模型约1/2的评分能力。由此可知，由大预言模型生成出来的初步数据污染性较高，并不能够作为优质的训练数据供模型训练得到良好的评分效果。

随着数据的蒸馏过滤，模型的评分能力逐渐提高，QWK得分逐渐上升，经过三通道过滤蒸馏的数据集，QWK得分接近0.6，评分能力接近高水平的标准。说明了数据在经过蒸馏之后更加接加接近了人为撰写的数据，数据的质量变高，模型的预测能力增强。从QWK得分的增长幅度可以看出，对文章影响越大的改动（如改动语法与更改词汇相比），其最后得出的数据集质量越高，QWK得分净增长越高。

在表IV的统计当中，我们跨领域的训练集在RoBERTa上也表现出了较好的水平，评分能力也达到了中等的水平，但由于跨领域的数据分布差异，模型依然不如经过大预言模型进行数据增强，再进行三通道蒸馏得到数据集所训练出来的模型。

# 5. 总结

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-蒸馏协同机制提升数据集的质量，适应当今数据集缺失领域的情况，通过数据增强提升了文本评分模型性能。

研究提出的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。同时，本研究将数据增强运用于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。并通过将模型进行对比分析，证明了该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，也超过了同等条件下跨领域的模型性能。

参考文献

[1] Wei, X., & Zou, K. (2020). Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv preprint arXiv:2004.06495.

[2] Qian, T., & Zhou, B. (2021). AEDA: An Easier Data Augmentation Technique for Text Classification Tasks. arXiv preprint arXiv:2106.09138.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL.

[4] [PromptDA: Label-guided Data Augmentation for Prompt-based Few Shot Learners](https://aclanthology.org/2023.eacl-main.41/) (Chen & Shu, EACL 2023)

[5] Sun, W., Wang, S., & Li, Y. (2022). mlpas\_encoder only essay scoring. arXiv preprint arXiv:2201.12345.

[6] L. Edjindjeba, E. Omar, O. Barakat, T. Desmettre, and T. Marx, “Leverage chatgpt as tool for data augmentation for supervised machine learning tasks in healthcare natural language processing,” 2023.

[7] K. Rosenblatt, “Chatgpt passes mba exam given by a wharton professor,” Retrieved Jan., vol. 26, p. 2023, 2023.

[8] D. A. Balota and J. I. Chumbley, “Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage,” Journal of Experimental Psychology: Human Perception and Performance, vol. 10, no. 3, p. 340, 1984.

[9] K. Cain and H. M. Nash, “The influence of connectives on young readers’ processing and comprehension of text,” Journal of Educational Psychology, vol. 103, no. 2, p. 429, 2011.

[10] S. Crossley and D. McNamara, “Interlanguage talk: What can breadth of knowledge features tell us about input and output differences?,” in Twenty-Third International FLAIRS Conference, 2010.

[11] Y. Cui, J. Zhu, L. Yang, X. Fang, X. Chen, Y. Wang, and E. Yang, “Clap for chinese: a linguistic complexity feature automatic calculation platform,” in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 5525–5538, 2022.

[12] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi, “Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text,” arXiv preprint arXiv:2107.01294, 2021.

[13] S. Mitrovic, D. Andreoletti, and O. Ayoub, “Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text,” arXiv preprint arXiv:2301.13852, 2023.

[14] M. Nasseri and P. Thompson, “Lexical density and diversity in dissertation abstracts: Revisiting english l1 vs. l2 text differences,” Assessing Writing, vol. 47, p. 100511, 2021.

[15] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., “Training language models to follow instructions with human feedback,” Advances in neural information processing systems, vol. 35, pp. 27730–27744, 2022.

[16] Vaswani, A., et al. (2017). Attention Is All You Need. In I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems (pp. 5998-6008). Curran Associates, Inc.

[17]Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

[18]Liu, Yinhan, et al. "RoBERTa: A robustly optimized BERT pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[19]LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[21]Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 21(140), 1-67.

[22] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “Detectgpt: Zero-shot machine-generated text detection using probability curvature,” in \*International Conference on Machine Learning\*, pp. 24950–24962, PMLR, 2023.

[23] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, “A watermark for large language models,” in \*International Conference on Machine Learning\*, pp. 17061–17084, PMLR, 2023.

[24] Y. Luchen, F. Yuan, W. Xin, and R. Wenbo, “Research on content detection algorithm and bypass mechanism of large language model,” \*Information Security Research\*, vol. 9, no. 6, pp. 524–, 2023.

[25] W. H. Gomaa, A. A. Fahmy, et al., “A survey of text similarity approaches,” \*international journal of Computer Applications\*, vol. 68, no. 13, pp. 13–18, 2013.

[26] K. B. Dempsey, P. M. McCarthy, and D. S. McNamara, “Using phrasal verbs as an index to distinguish text genres.,” in \*FLAIRS\*, pp. 217–222, 2007.

[27] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, “How close is chatgpt to human experts? comparison corpus, evaluation, and detection,” \*arXiv preprint arXiv:2301.07597\*, 2023.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 24824–24837.

致　　谢

行文至此，我的毕业论文已经画上了句号，而此刻内心满溢的感恩之情，却如同即将起笔的篇章，有着说不尽的话语。在这段充满挑战与成长的学术旅程中，我衷心向您表示感谢，李霞老师。

回首过往，您总是不断的帮助我，并激励我前行。在大二时，我正式加入实验室，您总是倾囊相授，为我打开知识的宝库。大三时，您让我加入了研究生的组会，在其中，我增长了很多见识，也结交了很多师友，科研思维有长足进步。

直至现在我依然记得实验室的氛围！热闹时，大家讨论着前沿的知识，分享着自己的作品与经历；安静时，大家埋头阅读文献，奋笔撰写论文。即使是初来乍到，这种友好的气氛也让我迅速适应下来。

那是我第一次与“科研”走的如此接近，感谢您将我引入前沿的海洋，与大家一同感受自然语言的魅力，在这里，我找到了比图书馆更广阔的天地。

感谢您对我的倾情付出，总是愿意“帮一帮我”。您总是特别可靠，解决我的多数担忧，您愿意在参加计设时接受和指导我招入没有经验的队友，提出合理的意见；在我论文因为硬件设备原因拖慢进度时，您愿意借出您刚装好的高性能电脑，让我顺利的推进实验。在我因为实习而困扰时，也愿意对外向我引荐。您了解我的近况，关心我的成长，然后对我倾囊相助。我这段大学的成长之路身边，处处都有您的身影。

感谢您对我的特别关照，总是愿意“等一等我”。您总是特别耐心，井井有条。每一次项目，每一次论文讲解，每一次的比赛，您从各个角度提出专业的意见，让我们意识到自己的不足，又总是在进行教导之后继续笑着肯定我们的能力，然后鼓励我们做得更好。记得第一次的论文撰写，您帮我改到了凌晨四点。那时的我并没有任何经验，如同笨拙的幼鸟扑棱翅膀。但你依然用心的教导着我，您总是先向我示范前路应该如何前行，然后在前方耐心地等待我跟上来。其实我知道您也根本就没有多少时间，您有很多重要的课程要上，有很多优秀的学生要带，有很多篇顶会的论文想发表，还有很多有意义的国家项目想完成。但您却依然尽力将不多的时间分于我，我于此感激不尽。

感谢您对我的无私教导，总是愿意“拍一拍我”。您总是特别温柔，教导我不同的生活道理。您愿意在我失意的时候依然鼓励我，跟我说，要往前看，要向前走，要大胆做。您教导我——“如果你因为太阳落山而流泪，那你也将错过群星。” 您也教导我——“如果因为害怕凋零而拒绝开花，春天将失去意义。”。您教我学会思考，在我前路一片迷茫的时候，轻轻拨开迷雾，让我看清楚脚下的路，告诉我现有的选择，然后再把决定权交由于我。支持我走向未知的道路，告知我可能遇到的风险。感谢您总是时不时地拍一拍我，向我加油打气，肯定我的才能。即使我并非千里马，但我依然感激遇到了伯乐。

最后，我想衷心地对您说一声：老师，感谢您一路以来的陪伴与支持！纵使我们终将分别，我也定不辜负您的期望，我会带着您的教导，继续砥砺前行！

1. https://openai.com/chatgpt [↑](#footnote-ref-1)
2. https://www.kaggle.com/datasets/matthewjansen/ellipse-corpus/data [↑](#footnote-ref-2)
3. https://www.kaggle.com/datasets/matthewjansen/ellipse-corpus/data [↑](#footnote-ref-3)
4. 意为思维链，最早由Google团队的Wei等人[28]提出，是一种加强大语言模型内容理解的方法 [↑](#footnote-ref-4)
5. https://www.kaggle.com/datasets/matthewjansen/ellipse-corpus/data [↑](#footnote-ref-5)