基于大语言模型（LLM）的文本智能批改系统

**中文学术类文本纠错标注规范**

目录

[1 前言 3](#_Toc143546356)

[2 标注原则与形式 3](#_Toc143546357)

[2.1 标注原则 3](#_Toc143546358)

[2.2 标注形式 4](#_Toc143546359)

[3 中文文本错误分类 4](#_Toc143546360)

[3.1 标点错误 4](#_Toc143546361)

[3.1.1 标点冗余 5](#_Toc143546362)

[3.1.2 标点缺失 5](#_Toc143546363)

[3.1.3 标点误用 6](#_Toc143546364)

[3.2 拼写错误 6](#_Toc143546365)

[3.2.1 中文术语拼写错误 6](#_Toc143546366)

[3.2.2 英文术语拼写错误 7](#_Toc143546367)

[3.2.3 其它拼写错误 7](#_Toc143546368)

[3.3 语法错误 7](#_Toc143546369)

[3.3.1 成分冗余 7](#_Toc143546370)

[3.3.2 成分残缺 8](#_Toc143546371)

[3.3.3 搭配不当 9](#_Toc143546372)

[3.3.4 语序不当 10](#_Toc143546373)

[3.3.5 不合逻辑 10](#_Toc143546374)

[3.3.6 句式杂糅 10](#_Toc143546375)

[4 历史版本 11](#_Toc143546376)

[4.1 第一版 11](#_Toc143546377)

[参考文献 12](#_Toc143546378)

# 1 前言

中文文本纠错是利用自然语言处理技术自动识别并纠正中文语料中包含的各类错误，包括标点错误、拼写错误、语法错误等多方面的内容。

当前中文语病语料库存在的主要问题有二：总体样本数量较小、数据集精细化程度不高。目前公开的中文文本语病数据集有MucGEC、NacGEC等，它们大多聚焦于通用文本类型，没有针对学术类文本进行准确标注和深入研究。而对于文本纠错研究，数据集的质量好坏将在很大程度上决定研究的效果。

为了解决这一短板，本项目旨在生成一份高质量的学术论文数据集，研发并完善一种主要基于大语言模型（LLM）的智能文本纠错系统，针对学术论文领域中常见的拼写错误和语法错误问题提出解决方案。

为了帮助标注人员更高效准确地标注中文学术类文本纠错数据，我们编写了下面的标注规范。本规范将定义常见的中文学术类文本错误分类体系，并给出相应的例句；介绍各类文本错误的纠正方法，并给出纠错示例；介绍文本纠错语料的标注准则和技巧。

值得注意的是，尽管本研究的最终目的是在中文学术类文本上表现出出色的纠错性能，但在目前的标注任务中主要针对的是文本中出现的明显错误，而尽量不修改文本的表述风格。在取得高质量数据集后，我们将首先确保实现基本的检错和改错功能。此后，在条件允许的情况下，则会进一步根据学术类文本的特点尝试对文本的叙述视角、语气等要素进行优化。

# 2 标注原则与形式

## 2.1 标注原则

由于本次标注的语料全部为学术论文，与以往的通用文本内容有所不同，我们在这里简单地规定汉语学术类文本纠错数据标注的准则。

（1）标注者标注的数据绝大多数为纯中文数据，但如果遇到了中英文混搭的句子，那么对于中文需要纠正所有类型的错误，而对于英文只纠正拼写错误。

（2）对于标点符号出错，需要将其视为一个语病并加以改正。

（3）对于句中出现的重复现象，如果不是写作者有意为之，则需要进行纠正。

（4）若句子的含义较清晰，但表述混乱，则需要进行纠正。

（5）若单个句子中出现多种类型的语言错误（尽管这种情况发生的可能性不高），请选取最合适的一种方式进行纠正。

标注过程中，对于拼写类型和词语误用类型的错误，要善于利用已有的词典资源，先充分理解再标注。请尽量不要贸然标注，影响质量。可以利用的资源如：百度汉语（hanyu.baidu.com），在线新华字典（<https://zidian.aies.cn>）等。若遇到人名、地名、机构名等命名实体类型的拼写错误，可以借助互联网搜索引擎判别是否正确。

## 2.2 标注形式

在Word文档中进行标注时，请将原文中出现错误的句子的头尾两处用字符串“[[[”“]]]”进行框选。请注意包含一整个句子，避免丢失句首、句末的标点符号。然后，在原始错误句子的“]]]”字符串后，用字符串“[[[”“]]]”框选改正后的句子的头部和尾部。示例如下。

原文：本文综述了近年来在计算机科学领域的一些重要研究进展。首先，我们回顾了机器学习领域的发展，包括深度学习、强化学习和迁移学习等方面的研究。其次，我们探讨了计算机视觉领域的最新进展，如目标检测、图像分割和人脸识别等技术。然而，我们还讨论了自然语言处理领域的研究现状，包括机器翻译、情感分析和问答系统……等方面的研究。最后，我们总结了当前计算机科学领域的挑战和未来的研究方向，包括解释性AI、可解释性和隐私保护等问题。

标注后的文档：本文综述了近年来在计算机科学领域的一些重要研究进展。首先，我们回顾了机器学习领域的发展，包括深度学习、强化学习和迁移学习等方面的研究。其次，我们探讨了计算机视觉领域的最新进展，如目标检测、图像分割和人脸识别等技术。[[[然而，我们还讨论了自然语言处理领域的研究现状，包括机器翻译、情感分析和问答系统……等方面的研究。]]][[[此外，我们还讨论了自然语言处理领域的研究现状，包括机器翻译、情感分析和问答系统等方面的研究。]]]最后，我们总结了当前计算机科学领域的挑战和未来的研究方向，包括解释性AI、可解释性和隐私保护等问题。

# 3 中文文本错误分类

为了便于理解，下面将依次列举学术类文本中常见的中文文本错误，并给出一定的例子。请标注者在进行标注之前首先阅读下面的示例。

## 3.1 标点错误

标点错误主要是在语句中多用、少用或误用了标点符号，根据实际情况比较容易判断。

### 3.1.1 标点冗余

标点冗余是指在句子中不必要的地方插入了标点符号。

【例1】

错误句子：根据调查结果，我们发现：大多数受访者对这个问题持有积极的态度。

正确句子：根据调查结果，我们发现大多数受访者对这个问题持有积极的态度。

【例2】

错误句子：这个理论在实践中有着丰富的应用，和广阔的前景。

正确句子：这个理论在实践中有着丰富的应用和广阔的前景。

【例3】

错误句子：在这个研究中，我们使用了多种方法，包括实地观察、问卷调查、深度访谈……等等。

正确句子：在这个研究中，我们使用了多种方法，包括实地观察、问卷调查、深度访谈等等。

### 3.1.2 标点缺失

标点丢失主要是指在句中、句末漏写了本应存在的标点。

【例1】

错误句子：我们使用了多种统计方法，例如t检验方差分析回归分析等。

正确句子：我们使用了多种统计方法，例如t检验、方差分析、回归分析等。

说明：这里在多条列举时，没有使用顿号进行分隔。

【例2】

错误句子：研究结果表明这个理论有广泛的应用领域比如人工智能数据挖掘等。

正确句子：研究结果表明，这个理论有广泛的应用领域，比如人工智能、数据挖掘等。

说明：这里整个长句未使用标点符号。

【例3】

错误句子：根据计算机视觉算法与应用一书中的方法，我们进行了图像分割的研究。

正确句子：根据《计算机视觉算法与应用》一书中的方法，我们进行了图像分割的研究。

说明：这里引用了书籍的名称，但未使用书名号。

### 3.1.3 标点误用

标点误用主要是指混淆了某些标点符号的用法，并在句中错误使用。

【例1】

错误句子：我们得出了一个重要的结论:这个新方法非常有效,可以提高产品质量和生产效率.

正确句子：我们得出了一个重要的结论：这个新方法非常有效，可以提高产品质量和生产效率。

说明：这里将中文标点符号误用为英文标点符号。

【例2】

错误句子：我们的研究目标是解决一个关键问题，如何提高系统的性能，稳定性，实用性和便捷性。

正确句子：我们的研究目标是解决一个关键问题，如何提高系统的性能、稳定性、实用性和便捷性。

说明：这里将顿号误用为逗号，导致“一逗到底”。

【例3】

错误句子：根据《大数据分析方法》的理论，我们进行了《数据挖掘》的实验研究。

正确句子：根据《大数据分析方法》的理论，我们进行了“数据挖掘”的实验研究。

说明：书名号和引号发生混淆。

## 3.2 拼写错误

主要指由于作者疏忽导致句子中出现明显不符合原意的词语，即错别字。学术类文本中，要特别关注专业术语是否出现此类错误。

### 3.2.1 中文术语拼写错误

【例1】

错误句子：统计学中的“方程”分析方法可以用于数据处理。

正确句子：统计学中的“方差”分析方法可以用于数据处理。

【例2】

错误句子：在人工智能领域，我们研究了机械学习和深入学习的算法。

正确句子：在人工智能领域，我们研究了机器学习和深度学习的算法。

### 3.2.2 英文术语拼写错误

【例】

错误句子：在人工智能领域，我们研究了natural network（神经网络）的训练方法。

正确句子：在人工智能领域，我们研究了neural network（神经网络）的训练方法。

### 3.2.3 其它拼写错误

【例1】

错误句子：在生物学研究中，我们使用了“基因饭组”进行分析。

正确句子：在生物学研究中，我们使用了“基因组”进行分析。

【例2】

错误句子：在本文中，我们引用了分子构造和反应机理理论。

正确句子：在本文中，我们应用了分子构造和反应机理理论。

## 3.3 语法错误

### 3.3.1 成分冗余

主要指相同或相近意思的词语在句子中同时出现，往往是两个词相邻出现，导致语义重复。

【例1】

错误句子：城镇化攸关到亿万人民的生活质量，它不是简单的城镇人口比例增加和城市面积扩张，而是在人居环境、社会保障、生活方式等方面实现由“乡”到“城”的转变。

正确句子：城镇化攸关亿万人民的生活质量，它不是简单的城镇人口比例增加和城市面积扩张，而是在人居环境、社会保障、生活方式等方面实现由“乡”到“城”的转变。

说明：句中“攸关”本身包含“到”的含义。

【例2】

错误句子：根据我们的初步估计，这项研究的样本处理时间大约在2个月左右。

正确句子：根据我们的初步估计，这项研究的样本处理时间大约在2个月。

说明：“大约”和“左右”同义。

【例3】

错误句子：通过对文本的深入研究和批评分析，我们逐渐培养起从经典作品中理解和感知美学价值的能力，产生出来了解读文学作品的本领。

正确句子：通过对文本的深入研究和批评分析，我们逐渐培养起从经典作品中理解和感知美学价值的能力，产生了解读文学作品的本领。

【例4】

错误句子：这款新一代芯片具有多达到64个核心，能够同时处理复杂的计算任务。

正确句子：这款新一代芯片具有多达64个核心，能够同时处理复杂的计算任务。

### 3.3.2 成分残缺

主要指句中必要的词语成分缺失，导致表意不明或歧义。比较常见的情况有主语缺失、宾语缺失、谓语缺失等。

【例1】

错误句子：在过去几十年的研究中，发现了许多与遗传相关的疾病。

正确句子：在过去几十年的研究中，研究者发现了许多与遗传相关的疾病。

说明：句中缺少主语。

【例2】

错误句子：自2000年起就掀起了“互联网+”的新浪潮。

正确句子：自2000年起全社会就掀起了“互联网+”的新浪潮。

说明：句中缺少主语。

【例3】

错误句子：只有按照实验设计，才能确保结果的有效性。

正确句子：只有按照实验设计操作，才能确保结果的有效性。

说明：句中缺少谓语。

【例4】

错误句子：在治疗方案后，研究团队对患者的病情进行了跟踪观察。

正确句子：在实施治疗方案后，研究团队对患者的病情进行了跟踪观察。

说明：句中缺少谓语。

【例5】

错误句子：本文介绍了LK-99超导体的制备过程与制备方法，提出了该材料成为跨时代的介质。

正确句子：本文介绍了LK-99超导体的制备过程与制备方法，提出了该材料成为跨时代的介质的结论。

说明：句中缺少宾语。

【例6】

错误句子：针对“AI机器人会抢美国工人饭碗”，新版路线图指出，过去五六年里美国新增约60万个制造业岗位。

正确句子：针对“AI机器人会抢美国工人饭碗”的观点，新版路线图指出，过去五六年里美国新增约60万个制造业岗位。

说明：句中缺少宾语。

【例7】

错误句子：对于第一组实验数据，第二组实验数据，表现出相当的稳定性与鲁棒性。

正确句子：对于第一组实验数据，以及第二组实验数据，都表现出相当的稳定性与鲁棒性。

说明：句中缺少关联词。

【例7】

错误句子：当技术进步提出了人性的考验，就可能引发伦理困境。

正确句子：当技术进步提出了对人性的考验，就可能引发伦理困境。

说明：句子缺少介词，导致表意不明。

### 3.3.3 搭配不当

主要指对某个词语的词义、词性运用不当。

【例1】

错误句子：研究人员通过实地考察和文献综述，给出了一个诡异的观点。

正确句子：研究人员通过实地考察和文献综述，给出了一个奇异的观点。

【例2】

错误句子：在这个研究领域，有很多前人的经验可供我们借鉴和抄袭。

正确句子：在这个研究领域，有很多前人的经验可供我们借鉴和参考。

【例3】

错误句子：我们通过对样本的深入分析，得出了一个不可能的发现和结论。

正确句子：我们通过对样本的深入分析，得出了一个此前未出现过的发现和结论。

### 3.3.4 语序不当

主要指。

【例1】

错误句子：

正确句子：

【例2】

错误句子：

正确句子：

【例3】

错误句子：

正确句子：

### 3.3.5 不合逻辑

主要指。

【例1】

错误句子：

正确句子：

【例2】

错误句子：

正确句子：

【例3】

错误句子：

正确句子：

### 3.3.6 句式杂糅

主要指。

【例1】

错误句子：

正确句子：

【例2】

错误句子：

正确句子：

【例3】

错误句子：

正确句子：

# 4 历史版本

## 4.1 第一版

完成了标注手册总体框架的搭建，界定了常见的12种中文文本错误类型，并从学术类文本中找出了一些例子。

# 参考文献

[1] Ma S, Li Y, Sun R, et al. Linguistic Rules-Based Corpus Generation for Native Chinese Grammatical Error Correction[J]. arXiv preprint arXiv:2210.10442, 2022.