大语言模型在中文文本纠错任务的评测

穆玲玲*, 王晓盈, 崔佳佳
郑州大学计算机与人工智能学院，河南省 郑州市，450001
iellmu@zzu.edu.cn; xywangnlp@qq.com; jjcui@gs.zzu.edu.cn

摘要

大语言模型（Large Language Models，LLMs）在信息抽取、机器翻译等自然语言处理任务上的能力已被广泛评估，但是在文本纠错方面还主要局限于评价GPT的英文语法纠错能力。中文文本纠错任务包括中文语法检测（Chinese Grammatical Error Detection，CGED）和中文语法纠错（Chinese Grammatical Error Correction，CGEC）两个子任务。本文使用提示的方法评估了国内外的主流大模型在中文语法检测和中文语法纠错任务上的能力。论文设计了不同的提示策略，对结果进行了整体和细粒度的分析。在NLPCC2018和CGED2018测试集上的实验结果表明，ERNIE-4和ChatGLM-4的中文文本纠错能力优于GPT-3.5-Turbo和LLaMa-2-7B-Chat，少样本思维链提示策略性能最优，对词序错误和拼写错误上纠正的准确率较高，说明大模型在低资源下具有较好的中文文本纠错能力。然而测试结果显示大模型的召回率比基线模型高至少14个百分点，说明大模型在中文文本纠错任务上存在过度校正的问题。

关键词： 大语言模型；中文文本纠错；提示；思维链

Evaluation of large language models for Chinese text error correction tasks

Lingling Mu*, Xiaoying Wang, Jiajia Cui
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou Henan, 450001
iellmu@zzu.edu.cn; xywangnlp@qq.com; jjcui@gs.zzu.edu.cn

Abstract

Large language models (LLMs) have been extensively evaluated for their capabilities in natural language processing such as information extraction and machine translation. However, their performance in text correction has primarily been limited to the evaluation of GPT's English grammar correction capabilities. Chinese text correction tasks encompass two sub-tasks: Chinese Grammatical Error Detection (CGED) and Chinese Grammatical Error Correction (CGEC). This paper assesses the capabilities of mainstream domestic and international LLMs on the CGED and CGEC tasks using prompting methods. The paper designs different prompting strategies and conducts holistic and granular analyses of the results. Experimental results on the NLPCC2018

*通讯作者

and CGED2018 test set indicate that ERNIE-4 and ChatGLM-4 outperform GPT-3.5-Turbo and LLaMa-2-7B-Chat in Chinese text correction. The few-shot chain of thought prompting strategy demonstrates optimal performance, with high accuracy in correcting word order and spelling errors, suggesting that LLMs possess good Chinese text correction capabilities under low-resource conditions. However, the test results show that LLMs have a recall rate at least 14 percentage points higher than the baseline models, indicating that LLMs have an issue of overcorrection in Chinese text correction tasks.

Keywords: CGEC , LLMs , Prompt , Chain of Thought

1 引言

中文文本纠错任务主要包括中文语法检测（Chinese Grammatical Error Detection，CGED）和中文语法纠错（Chinese Grammatical Error Correction，CGEC）两个子任务。中文语法检测任务的目标是自动检测中文文本中是否存在语法错误、判断错误类型并确定错误的具体位置(Xie et al., 2020)。中文语法纠错任务的目标是自动检测并修正文本中的包括标点、拼写、词汇、语序等方面的语法错误，以确保生成符合原意的正确句子(Chang et al., 2023)。

现阶段，文本纠错的主流方法是序列到序列（Sequence-to-sequence，seq2seq）和序列到编辑（Sequence-to-edit，seq2edit）。序列到序列方法将文本纠错任务视为翻译任务，输入为可能包含语法错误的句子，输出则为与输入对应的语法正确的句子(Zhao et al., 2019; Kiyono et al., 2019; Wang et al., 2021; Li et al., 2022a; Fang et al., 2023a)。序列到序列方法则将文本纠错任务视为序列标注任务，其输入与序列到序列相同，但输出是对输入文本进行修正的一组正确编辑操作(Malmi et al., 2019; Stahlberg and Kumar, 2020; Omelianchuk et al., 2020; Yuan et al., 2021)。尽管这两种方法在文本纠错任务上均取得了显著的进展，但它们的成功都极大地依赖于大规模的训练数据，当训练数据有限时，模型的性能将受到影响。

大语言模型（Large Language Models，LLMs）是指参数量达到数千亿级别（或更多）的Transformer架构模型，这些模型通过在庞大的文本数据上进行训练而得以构建(Shanahan, 2024)，例如ERNIE-4、ChatGLM-3、GPT-3.5和LLaMa-2等。大模型在NLP领域展现了显著的能力，主要体现在其强大的泛化能力，能够适应多种语言和不同类型的任务(Kwon et al., 2023; Loem et al., 2023; 罗文, 2024)。

大模型的主要使用方法是为特定任务设计恰当的提示策略，即通过一系列明确定义的指令任务对模型进行训练，这有助于提升模型对指令的响应精确度，并减少对大量样本的依赖性，即使在样本数量有限的场景下也能展现出卓越的性能。一种典型的提示策略是将任务描述和示范以自然语言文本的形式表达的上下文学习(Instruction with Context Learning, ICL)(Huang and Chang, 2023; Qiao et al., 2023; Dong et al., 2022)。另一种策略是思维链提示(chain-of-thought prompting, CoT)(Wei et al., 2022; Kojima et al., 2022; Fang et al., 2023b)，即在提示中嵌入一系列中间推理步骤以提高LLM的效果。基于提示的大模型已经在信息抽取、命名实体识别和序列标注等多个NLP任务中取得显著效果。

尽管基于提示的大模型已在信息抽取、命名实体识别、序列标注等多个自然语言处理任务中已经得到广泛的评测，但在中文文本纠错任务中的评测仍不够充分。目前，大模型在中文文本纠错任务上的评测研究主要集中在中文语法纠错和中文拼写检查两方面,采用的大模型主要是ChatGLM-3、LLaMa-2-7B-Chat和GPT3.5(Li et al., 2023b; Fang et al., 2023c; Fan et al., 2023; Song et al., 2023)，通常从整体效果进行分析，对提示模板的设计策略讨论较少，且缺少对中文语法检测任务的评测。

本文旨在使用不同的提示策略，从整体和细粒度错误类型两个方面评估更多国内外大模型在中文语法检测和中文语法纠错上的性能，并研究ICL和CoT等不同提示策略的影响。本文选取了国内外主流的大模型，包括LLaMa-2-7B-Chat、GPT-3.5-Turbo、ChatGLM-4、ERNIE-4、Baichuan2-7B-Chat和Qwen-7B-Chat，设计了无解释、专家提示、思维链和少样本思维链四种不同的提示策略，以评估这些大模型在中文文本纠错任务上的性能。此外，本文还利

用ChERRANT工具(Zhang et al., 2022a)提取了大模型在NLPCCC2018测试集和CGED2018测试集上纠正出的错误类型，对大模型在具体错误类型上的性能进行了评价。

本文的主要贡献包括以下几点：

- 本文针对中文语法检测和中文语法纠错两个子任务，探索了在低资源下大模型在中文文本纠错任务中的性能表现。
- 本文提出了针对中文文本纠错评价的少样本提示和上下文提示等策略，发现了大模型在使用少样本思维链提示模版下的性能最好。
- 本文对比了国内外不同大模型在中文文本纠错任务上的性能差异，发现ERNIE-4和ChatGLM-4的中文文本纠错能力优于GPT-3.5-Turbo和LLaMa-2-7B-Chat。
- 本文根据错误类型对大模型进行了细粒度纠错能力的评价，对纠错能力的评估更全面；

2 相关工作

2.1 用于中文语法纠错的序列到序列方法

自2016年神经机器翻译方法取得显著成果以来，语法纠错任务往往被视作文本生成任务，使用序列到序列的生成模型，直接生成正确的句子(Fang et al., 2023a; Li et al., 2022a; Liu et al., 2021; Wang et al., 2021)。在这种框架下，编码器将源输入X编码成向量，然后传递给解码器以生成校正结果Y。早期的研究主要采用基于CNN(Ren et al., 2018)或RNN(Zhou et al., 2018b)的模型来解决语法纠错任务。随着transformer(Vaswani et al., 2017)在机器翻译任务中取得巨大成功，它也被应用于语法纠错任务。HRG(Malmi et al., 2019)结合了拼写检查器、基于NMT的模型和序列编辑模型，但其拼写检查器基于语言模型，未能充分利用上下文信息。后续的研究主要基于预训练模型，如Kaneko (2019)提出将BERT融入到语法纠错的编码器-解码器模型，通过加载预训练模型对语法纠错模型进行微调以提升纠错性能。Zhang (2022a)提出了基于序列到序列架构的中文BART模型，适用于中文语法纠错任务。最新研究主要集中在以下三个方面：1）在推理速度方面，Sun (2021)提出使用并行解码和浅层解码器替代传统的平衡编码器-解码器的Transformer架构，从而提高推理速度。Yakovlev (2023) 提出使用非自回归解码来提升推理速度；2）通过添加额外的语法知识或利用特定语言知识来提高纠错性能(Mita and Yanaka, 2021; Fei et al., 2023; Kaneko et al., 2022; Zhang et al., 2022b)；3）在解码方法方面，Sun (2022)提出通过权衡精确度-召回率来提高纠错性能，Zhang (2023)提出通过重排序来选择正确率更高的纠错结果，从而提高纠错性能。

2.2 用于中文语法纠错的序列到编辑方法

序列到序列模型在输出纠错后的句子时，并不提供任何解释，例如错误的具体位置和类型。为了解决这个问题，研究人员提出了序列到序列模型，该模型将语法纠错任务视为序列标注任务，通过预测句子中每个token的编辑操作来纠正错误(Malmi et al., 2019; Awasthi et al., 2019; Stahlberg and Kumar, 2020)。LaserTagger(Malmi et al., 2019)提出将源输入转换为token级的编辑操作序列，包括保留、删除、添加和交换。PIE (Awasthi et al., 2019)将局部序列编辑问题转化为序列标注任务，并使用BERT对编辑后的输入token进行非自回归标注。GECToR (Omelianchuk et al., 2020)在传统编辑操作的基础上引入了自定义的g-transformations，如大小写转换、合并标记、更改单词后缀等。Parnow (2021)提出了一个类似GAN的序列标注模型来提升模型的性能，该模型由语法错误检测器作为鉴别器和语法错误标注器组成，Gumbel-Softmax采样作为生成器。Li (2022b)则提出一个序列到编辑模型，用于中文语法纠错任务，在每个字符前预测一个动作，动作由跳过、复制和生成三类构成，然后将这些动作与基本序列到序列框架融合，提供最终预测。Tan (2023)提出了一种基于纠错类型的校正方法，在每个训练实例中，额外构建多个涉及特定类型错误训练实例，然后使用这些额外构造的训练实例和原始的训练实例依次训练模型。序列到编辑模型的一个局限性在于其对编辑操作手动构建的依赖，这降低了模型的可迁移性并可能影响输出的流畅性(Li et al., 2023a)

2.3 大模型在语法纠错任务上的性能评估

近期的研究对多种大模型在语法纠错任务中的性能进行了广泛的评估，包括了闭源和开源的模型(Loem et al., 2023; Kwon et al., 2023; Li et al., 2023b; Fang et al., 2023c; Kaneko and Okazaki, 2023)。例如，Loem (2023)研究了基于GPT-3的提示方法在零样本和少样本语法纠错任务中的性能和可控性，证实了GPT-3在语法纠错任务上的有效性。Kwon (2023)对大模型在阿拉伯语法纠错上的表现进行了评估，发现提示方法和少样本学习在上下文中有效，GPT-4在专家提示下的F1值达到65.49，较基线提高了约5分。Li (2023b)评估了ChatGLM-3、Baichuan和GPT-3.5-Turbo等大模型在中文语法错误纠正和中文拼写检查任务上的表现。Fang (2023c)对ChatGPT在英语、德语和中文三种不同语言上的语法纠错任务表现进行了评估，强调了其在生成流畅句子方面的能力。Song (2023)评估了GPT-4在语法错误解释方面的能力，发现它使用一次性提示只能对60.2%的错误进行解释。Fan(2023)探讨了大型语言模型在中文语法纠错任务中的表现，主要在4个中文语法纠错数据集上使用了3种不同的大模型进行了实验。这些研究主要集中在评估ChatGLM-3、LLaMa-2-7B-Chat和GPT-3.5-Turbo等大模型在中文语法纠错和中文拼写检查任务中的表现。然而，目前的研究尚未探讨大模型在中文语法检测任务上的性能，也未曾进一步分析提示模板对大模型纠错性能的具体影响，以及模型在不同错误类型上的表现差异。本文旨在通过采用多种提示策略，对国内外的大模型在中文文本纠错任务中的性能进行评估，涵盖了中文语法检测和中文语法纠错两个子任务。

3 实验设置

3.1 数据集

为了与现有研究保持一致，本文使用CGED2018测试集(Rao et al., 2018)来评估大模型在CGED任务上的性能，并使用NLPCC-2018测试集(Zhao et al., 2018b)来衡量大模型在中文语法纠错任务中的表现。CGED2018数据集由北京语言大学提供，专为中文句法错误诊断技术评估而设计，训练集包括402条数据，测试集包括3549条数据。NLPCC2018的训练集是从lang8收集处理得到的，是汉语的母语使用者对汉语学习者作文的修改，处理后包含1092285条数据，测试集来源于北京大学中文系构建的北京大学汉语学习者语料库，该语料库包含外国大学生的作文，旨在促进国际教育和汉语作为中介语言的研究，包含2000条数据。数据集详细信息见表1。

数据集	句子数	错误占比（100%）
NLPCC2018测试集	2000	99.2
CGED2018测试集	3549	85.71

表 1: 数据集信息

3.2 评价指标

为了与现有研究保持一致，本文采用了两种评估指标，分别针对中文语法检测和中文语法纠错两个不同的任务。在中文语法检测任务中，本文使用准确率(Precision, P)、召回率(Recall, R)和F1作为评价指标；而在中文语法纠错任务中，本文采用准确率(Precision, P)、召回率(Recall, R)和F0.5作为评价指标。

CGED2018评测任务的评价涵盖了检测、识别、位置三个子任务(Zhao et al., 2018a)。具体而言，检测任务旨在判断句子是否含有错误，若存在错误则判定为不正确，可视为一个二分类问题。识别任务旨在确定错误的类别，这可视为一个多分类问题。位置任务旨在确定错误的精确位置。这三个子任务的难度逐渐增加，且均使用准确率(Precision, P)、召回率(Recall, R)、F1这三个评价指标*。

NLPCC2018评测任务的目标是改正句子中的语法错误，评价依据模型对错误语句的编辑与标准编辑集合的匹配程度。评价指标为准确率(Precision, P)、召回率(Recall, R)以及F0.5，本文使用MaxMatch（M2）†工具来计算上述指标。

http://lang-8.com/
*https://github.com/blcuicall/CCL2022-CLTC
†https://github.com/nusnlp/m2scorer

3.3 大模型选择

本文选用了国内外主流的大模型，对它们在中文语法检测和中文语法纠错两个任务上的性能进行了评估。在选择模型时，我们特别考虑了国内大模型，这是因为国内大模型通常利用了大量具有中国特色的数据集进行训练，而国外大模型则可能依赖于全球范围内的数据资源。具体来说，我们选取了LLaMa-2和GPT-3.5两个常用的国外大模型，以及ChatGLM-3、ERNIE-4、BaiChuan-2和Qwen等国内大模型。

LLaMa-2 [‡] 是Meta AI发布的一个大语言模型，以其较小的参数规模实现了卓越的性能。LLaMA遵循大型语言模型的工作原理，即通过预测下一个单词来递归生成文本。

GPT-3.5 [§] 是一类基于Transformer架构的预训练语言模型，由OpenAI开发，通过在海量文本数据上进行无监督预训练，实现了对自然语言文本的理解和生成。

ChatGLM-3 [¶] 是一个基于千亿参数规模的中英双语模型，它集成了问答、多轮对话和代码生成功能，通过代码预训练和有监督微调等技术提升了其功能性能。

ERNIE-4 [‖] 是百度提出的一种基于知识增强的持续学习语义理解框架，该框架结合大数据预训练与多源知识，通过持续学习技术不断吸收文本数据中的词汇、结构和语义知识，实现模型的持续进化。

BaiChuan-2 [**] 是百川智能推出的一款新一代开源大语言模型，其训练语料为2.6万亿Tokens的高质量数据集，在多个权威的中文、英文和多语言通用及领域基准测试中取得了最佳效果。

Qwen [††] 是阿里云研发的720亿参数规模的模型，基于Transformer架构，支持最高32K的上下文长度和15万的词表，在3万亿Tokens的超大规模预训练数据上进行训练，涵盖了多样化的预训练数据类型。

3.4 提示模板

本文主要设计四种提示模板：无解释（None）、专家提示（Expert）、思维链（CoT）和少样本思维链（Few-shot CoT）(Fang et al., 2023b; Xu et al., 2023)。这四种提示模板以递进方式构建，思维链在专家提示的基础上加入了中文文本纠错的思维推理过程，而少样本思维链则在此基础上增加了纠错样本示例。以下是对这四种提示模板的详细描述。

None 该提示模板仅要求模型完成语法纠错任务，不提供额外的提示或解释。具体指令为："请对输入句子进行**语法纠错**。若输入句子存在语法错误，仅输出纠正后的正确句子，无需任何额外解释；若句子语法无误，则直接输出原始输入。"

Explain Xu (2023)引入了一种新的策略，即利用大模型的类专家能力。该方法将专家角色赋予模型，并提供具体说明，以提升结果的相关性和质量。受Xu (2023)的启发，本文在Explain提示模板中将模型描述为专业的中文语法纠错工具，并阐明语法纠错任务的含义，即在None提示模板的基础上对语法纠错任务进行了全面且详细的阐释。具体指令为："你是一个**专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力**。若输入句子存在语法错误，仅输出纠正后的正确句子，无需任何额外解释；若句子语法无误，则直接输出原始输入。"

CoT Kojima (2022)和Fang (2023b) 提出了CoT提示策略，即在提示中明确任务的推理路径。受Kojima (2022)和Fang (2023b)的启发，本文在CoT提示模板中指出了语法纠错的实现过程，引导大模型按照给定的推理路径完成任务。具体指令为："你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。**在处理输入文本时，你会深入理解和分析句子结构，精准定位语法错误的位置和类型，然后迅速修正为符合语法的正确句子**。若输入句子存在语

[‡]https://github.com/meta-llama/llama
[§]https://openai.com/chatgpt
[¶]https://github.com/THUDM/ChatGLM3
[‖]https://wenxin.baidu.com/
[**]https://github.com/baichuan-inc/Baichuan2
[††]https://github.com/QwenLM/Qwen

法错误，仅输出纠正后的正确句子，不会提供任何多余的解释；若句子语法无误，则原样返回原始输入。”

Few-shot CoT Few-shot CoT借鉴了ICL(Huang and Chang, 2023; Qiao et al., 2023; Dong et al., 2022)提示策略，在CoT模板的基础上添加了少量语法纠错示例，以帮助大模型学习蕴含在示例中的语法纠错知识，从而更有效地完成中文文本纠错任务。示例的选择主要基于人工从验证集中挑选的典型样例，以确保数据集的一致性。具体指令为：“你是一款专业的中文语法纠错工具，具备卓越的语法错误识别和纠正功能。在处理输入的文本时，你会深入剖析句子结构，精准定位语法错误的类型和位置，然后迅速给出符合语法的修正方案。一旦发现输入句子存在语法错误，你只会简洁地输出修正后的正确句子，不会添加任何额外的解释；而当句子语法无误时，你则原样返回原始输入。**下面是一个示例：（1）输入：通过六个月的学习，我的汉语和同声传译水平有了很大提高。输出：通过六个月的学习，我的汉语和同声传译水平有了很大的提升。**示例结束，接下来将是你要纠正的句子。”

4 实验结果

在本节中，本文将在CGED2018和NLPCC2018两个测试集上分别评估大模型在中文语法检测和中文语法纠错任务上的性能。本文在本地加载并测试了Qwen-7B-Chat、Baichuan-2-7B-Chat、ChatGLM-3-6B和LLaMa-2-7B-Chat的性能，接口调用ERNIE-4、GLM-4和GPT-3.5-Turbo。

4.1 中文语法检测实验结果

模型	语法错误识别			错误类型识别			错误类型定位l		
	P	R	F1	P	R	F1	P	R	F1
序列到编辑									
Li and Qi (2018)	58.7	63.34	61.31	46.22	41.42	37.77	17.53	11.38	12.47
Zhang (2018)	72.66	74.08	73.36	58.31	49.55	53.57	38.39	29.66	33.46
Liang(2020)	–	–	78.0	–	–	64.0	–	–	39.9
Luo (2020)	**85.5**	78.6	**81.9**	68.1	62.1	**65**	48	**41.3**	**44.4**
序列到序列									
Fu (2018b)	82.76	**99.95**	75.63	**71.07**	**97.52**	55.03	**53.41**	38.86	36.12
Zhao (2018a)	63.77	75.84	69.29	40.84	41.61	41.22	6.3	6.09	6.2
Han (2020)	–	–	77.55	–	–	61.38	–	–	44.51
国外大模型									
LLaMa-2-7B-Chat	56.43	81.55	66.7	32.23	36.54	34.25	4.22	4.34	4.28
GPT-3.5-Turbo	61.10	78.62	68.76	41.76	50.01	45.52	14.54	18.90	16.44
国内大模型									
Qwen-7B-Chat	52.62	65.17	58.23	33.83	44.6	38.48	9.27	14.3	11.25
BaiChuan-2-7B-Chat	57.71	67.13	62.06	37.80	44.84	41.02	11.15	15.79	13.07
ChatGLM-3-6B	53.28	77.16	63.04	41.11	44.15	42.58	12.13	13.32	12.70
ChatGLM-4	**62.96**	73.08	67.64	34.61	56.97	43.06	10.80	22.27	14.54
ERNIE-4	58.58	**91.17**	**71.33**	**42.01**	**67.88**	**51.9**	**16.87**	**31.39**	**21.95**

表 2: 大模型在CGED2018测试集上结果

如表2所示，大模型在中文语法检测任务上的表现相较于当前的有监督模型仍然存在差距。然而，在低资源下，大模型已经展现出了较好的性能。此外，在中文语法检测任务上，国外大模型的表现不及国内大模型，而小参数（6-7B）的大模型的性能也低于大参数的大模型。值得注意的是，在大参数国内大模型中，其在语法错误类型识别子任务上的表现优于其他两个子任务。

为了研究大模型参数数量对性能的影响，本文对比了6B参数的ChatGLM-3与更大参数规模的ChatGLM-4接口在中文语法检测任务上的表现。如表2所示，6B的ChatGLM-3与大参数ChatGLM-4接口在语法错误识别子任务上的F1值分别为63.04和67.64，在错误类型识别子任务上的F1值分别为42.06和43.06，在错误类型定位子任务上的F1值分别为12.7和14.54。这些结

果表明，随着参数数量的增加，大模型能够更全面地学习训练数据，从而在中文语法检测任务上展现出更好的性能。

在语法错误识别子任务上，国外大模型与国内大模型的性能相近，但在其他两个子任务上与国内大模型存在较大差距。这可能是因为中文和英文在语言表达上的差异，错误类型定义的不同，以及国外大模型未使用大量中文数据进行训练，导致模型在错误识别方面表现良好，但在错误类型识别和错误类型定位方面存在局限性。

如表2所示，国内大模型（包括ChatGLM-4、ERNIE-4）在语法错误识别子任务上的F1值分别为67.64和71.33，在错误类型识别子任务上的F1值分别为43.06和51.9，在错误类型定位子任务上的F1值分别为14.54和21.96。这些结果表明，在低资源下，国内大模型在中文语法检测任务上的表现已经接近部分有监督模型，可以有效的完成中文语法检测任务。然而，这些模型普遍存在准确率相对较低而召回率较高的情况，因此可以在模型输出结果基础上做进一步改进。

模型	P	R	F0.5
序列到编辑			
HRG (Hinson et al., 2020)	36.79	27.82	34.56
SG-GEC (Wu and Wu, 2022)	50.56	25.24	42.11
序列到序列			
AliGM(Zhou et al., 2018a)	41	13.75	29.36
YouDao(Fu et al., 2018a)	35.24	18.64	29.91
BLCU(Li et al., 2019)	47.63	12.56	30.57
Qiu and Qu (2019)	36.88	18.94	31.01
MaskGEC(Zhao and Wang, 2020)	44.36	22.18	36.97
TEA(王辰成et al., 2020)	39.43	22.8	34.41
WCDA(Tang et al., 2021)	47.29	23.89	39.49
Copy(Zhao et al., 2019)	51.25	32.55	**45.97**
SynGEC(Zhang et al., 2022b)	49.96	**33.04**	45.32
TemplateGEC (Li et al., 2023a)	**54.50**	27.40	45.50
国外大模型			
LLaMa-2-7B-Chat	11.79	11.46	11.72
GPT-3.5-Turbo	26.64	22.75	25.76
国内大模型			
Qwen-7B-Chat	19.95	22.53	20.42
BaiChuan-7B-Chat	20.87	23.28	21.31
ChatGLM-3-6B	19.52	21.09	19.82
ChatGLM-4	24.27	32.34	25.72
ERNIE-4	**32.78**	**38.15**	**33.72**

表 3: 大模型在NLPCC2018测试集上结果

4.2 中文语法纠错实验结果

如表3所示，大模型在中文语法纠错任务上的表现与在中文语法检测任务相似，均不如现有的有监督模型，但在低资源下，仍展现出可观的性能。此外，在中文语法纠错任务上，国外大模型的表现不及国内大模型，且参数规模较小的大模型（6-7B）的性能也低于大型参数模型。

6-7B参数量级的国内大模型（包括ChatGLM-3-6B、Qwen-7B-Chat和BaiChuan-2-7B-Chat）在中文语法纠错任务上的F0.5值分别为19.82，20.42和21.31，表现较弱。这种性能不足可能由两个主要因素导致：首先，中文语法纠错任务相较于中文语法检测任务更为复杂，它不仅要求模型能够识别错误，还要求模型能够进行错误的修正，这涉及到词序、标点使用和词汇选择等多个错误类型，增加了任务的难度。其次，这些模型可能未经过针对中文语法纠错数据集的专门训练，且模型参数规模相对较小，导致模型在学习过程中可能未能充分捕捉到任务所需的特征和规律。

国外大模型（如GPT-3.5-Turbo和LLaMa-2-7B-Chat）在中文语法纠错上的F0.5值分别为11.72和25.76，表现不佳，这可能是因为它们在训练时未接触大量中文数据，无法充分适应中文的语言特性。例如，源输入“中国，悠久的历史，灿烂的文化，真是在历史上最难忘的国家。”，参考答案为“中国，拥有悠久的历史，灿烂的文化，真是在历史上让人最难忘的国

家。”，GPT-3.5-Turbo模型的输出结果为“中国，悠久的历史，灿烂的文化，真是历史上最难忘的国家。”。这表明模型的纠错结果更倾向于英文的表达方式。

国内大模型在中文语法纠错任务上的表现较为出色，ChatGLM-4和ERNIE-4的F0.5值分别为25.72和33.72。虽然与现有的有监督模型相比存在一定差距，但在低资源下，已经展现出卓越的性能。如表3所示，国内大模型普遍存在召回率高而准确率低的情况。这一现象可能是由于模型主要接受了大量正确句子的训练，而在面对错误句子时，模型倾向于根据其训练数据对其进行修正。由于同一语意的句子可以有多种表达方式，这种修正方法可能不适合正式或非标准的语言环境，有时甚至可能导致过度校正。此外，模型在语义理解方面也存在不足，具体例子如表4所示。表4的标注旨在对比模型输入、模型输出与正确答案之间的差异，与传统的纠错标注方法有所区别。

类型	输入	模型输出	正确答案
标点符号使用不当	因为读书准备就业的时间越少了。	因为读书准备就业的时间**更少**了。	因为读书，准备就业的时间**更少**了。
拼写错误	**我看来**北京人很爽快。	**在我看来**，北京人很爽快。	**在我看来**北京人很爽快。
	人生就是过山车，**又高又低**，**又**光明又黑暗。	人生就是过山车，**又高有低**，**有光明**又有黑暗。	人生就是过山车，**有高有低**，**有光明**又有黑暗。
搭配不当	不**带**口罩的有些朋友们特别后悔**了**。	有些不**带**口罩的朋友们特别后悔。	有些不**戴**口罩的朋友们特别后悔。
	现代社会的就业**难太厉害**。	现代社会的就业**太难厉害**。	现代社会的就业**太难**。
量词使用不当	但是这种想法太**近视眼的**，而有很大的错误。	但是这种想法太**近视**，而且有很大的错误。	但是这种想法太**短浅**，而且有很大的错误。
	本人在贵公司实习了**二个**多月，**此时间**经验了许多新鲜的事情。	本人在贵公司实习了**二个**多月，**此间**经历了许多新鲜的事情。	本人在贵公司实习了**两个**多月，**此间**经历了许多新鲜的事情。
名词使用不当	十二月是瑞典一年中黑夜最长白天短的**一天**。	十二月是瑞典一年中黑夜最长白天最短的**一个月**。	十二月是瑞典一年中黑夜最长白天最短的**一月**。
	勤奋刻苦，不折不挠的精神是从黑暗走向光明的人的必要条件。	勤奋刻苦，不折不挠的精神是从黑暗走向光明的必要条件。	勤奋刻苦，不折不挠的精神是**人**从黑暗走向光明的必要条件。
过度校正	四个人在**同桌**吃一顿饭是很难得。	四个人在**桌**吃一顿饭是很难得的。	四个人在**同桌**吃一顿饭是很难得的。
	这样，你就会尝到泰国人**死爱**的味道。	你将会尝到泰国人**喜爱**的味道	这样，你就会尝到泰国人**爱死**的味道。
	抽烟对身体非常不好，对身体**非常危害**。	抽烟对身体非常不好，对身体**危害极大**。	抽烟对身体非常不好，对身体**危害非常大**。

表 4: ERNIE-4在NLPCC2018测试集中个别样例

5 大模型提示策略影响分析

本节将从三个方面探讨影响ERNIE-4模型在中文语法纠错任务中性能的因素：提示模板中的关键字选择、提示策略的设计以及少样本思维链提示策略中样本的数量。

5.1 提示模板中的关键字对大模型性能的影响

鉴于大模型提示模板中的关键词对其性能具有显著影响，本文在 3.4提出的None模板基础上，设计了多种不同的关键词提示，并通过调用ERNIE-4接口进行了测试。这些关键词分为两大类：“直接指示”和“角色扮演”。在“直接指示”类别中，模板明确指示模型执行语法纠错的指令，例如：“查找并校对输入句子中的任何语法错误。”、“查找并修复这个句子中的任何错误。”以及“在不改变原语义和原语言的情况下，润色输入句子。”。这些模板旨在引导模型专注于纠正语法错误，同时保持句子的原始意义和语言风格。在“角色扮演”类别中，模板通过指定模型扮演与语法纠错相关的特定角色来进行提示，例如：“作为专业的中文语法纠错工具”、“担任资深的编辑”或“作为中文教师”。这些模板旨在引导模型从特定专家的角度出发，执行语法纠错任务。具体的提示模板示例和详细信息见附录 8。

如附录 8所示，测试的几种提示模板均能够实现语法纠错的任务。然而，部分模板倾向于在源句的基础上进行不必要的扩展，或是改变原有的句式结构，从而导致与正确答案相比改动

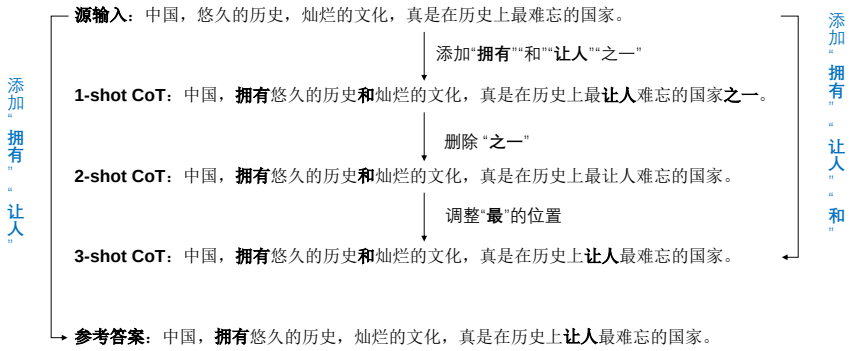

图 1: ERNIE-4使用Few sot-CoT提示策略的输出样例

较大，出现过度校正的现象。因此，本文选取了修改正确且改动最小的关键字作为基础模板的关键字，即“你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。”。

5.2 提示策略对大模型性能的影响

ERNIE-4	NLPCC2018测试集		
	P	R	F0.5
None	30.37	39.59	31.86
Explain	30.77	**40.19**	32.28
CoT	32.47	37.36	32.63
Few-shot CoT	**32.78**	38.15	**33.72**

表 5: 不同提示策略对中文语法纠错任务性能的影响

如表5所示，不同提示模板对ERNIE-4模型在中文语法纠错任务中的性能表现产生了显著影响。性能表现从低到高依次为：无额外解释的提示模板（None）、加入专家解释的提示模板（Explain）、加入思维链（CoT）的提示模板以及加入纠错示例的提示模板（Few-shot CoT）。

1）当大模型仅接收到执行语法纠错任务的指令而缺乏额外解释时，模型必须依赖其先前学习的知识来推断语法纠错的本质，这种理解可能与语法纠错的真实意图存在偏差，导致性能最差。2）当模型被告知语法纠错任务的具体定义时，能够依据这一明确的指示完成任务，从而减少了理解上的偏差，模型性能因此得到提升。3）如果在提示模板中明确指出语法纠错的步骤，即构建思维链，这将引导大模型模型按照标准化的推理过程来执行任务，从而更有效地提高模型性能。4）在思维链的基础上，提供额外的纠错示例能够使模型学习到示例中蕴含的丰富信息，从而更深入地理解语法纠错任务，并实现更卓越的性能表现。

5.3 少样本思维链提示策略中样本数量对大模型性能的影响

ERNIE-4	NLPCC2018测试集		
	P	R	F0.5
1-shot CoT	31.05	37.36	32.14
2-shot CoT	32.20	**38.94**	33.35
3-shot CoT	**32.78**	38.15	**33.72**

表 6: 大模型提示模板中样本数量对中文语法纠错性能的影响

为研究少量示例提示模板对大模型纠错能力的影响，本文通过调整提示模板中示例的数量进行了实验。如表6所示，随着提示模板中纠错示例数量的增加，大模型的纠错准确率显著提高，尤其在处理复杂和微妙的语言错误时效果显著。如图 1所示。这可能是因为当大模型接收到包含示例的提示模板时，它们能够更快地识别文本中的不一致性，并从中归纳出纠错的规则。这种学习机制不仅使大模型能够机械地纠正错误，而且能够更深入地理解文本纠错的本

质，即如何在保持原文意蕴不变的前提下，提高语言的准确性和流畅性。实验结果验证了在提示模板中添加少量示例可以提升大模型纠错能力方面的有效性，通过模仿和学习大量示例，大模型能够增强执行语法纠错任务的能力。

	W	S	M	R	无
gold	**202**	1961	963	698	17
None	162(25%)	2108(22%)	1117(13%)	802(21%)	206
Explain	177(26%)	2092(22%)	1079(14%)	873(21%)	206
CoT	170(25%)	2205(21%)	1119(14%)	895(20%)	210
1-shot CoT	157(28%)	2398(17%)	1141(12%)	907(18%)	**277**
2-shot CoT	169(24%)	2322(20%)	**1165(14%)**	**921(19%)**	200
3-shot CoT	168(24%)	**2401(17%)**	1140(11%)	919(19%)	263

表 7: 基础错误类型分析（数量/准确率）

6 针对不同错误类型的纠错结果分析

为了深入研究大模型在中文语法纠错任务中的性能表现，本文对ERNIE-4在不同错误类型上的纠错效果进行了细致分析。本文采用了ChERRANT工具(Zhang et al., 2022a)来提取错误类型，这些错误类型被划分为基础错误和详细错误两大类。具体来说，本文通过对比大模型生成结果与标准答案之间的差异，对大模型在不同错误类型上的表现进行了评估。基础错误和详细错误的分析结果分别展示在表7和8中，其中，gold代表标准答案，而None、Explain、CoT、1-shot CoT、2-shot CoT和3-shot CoT则代表了大模型所采用的不同提示模板。

	W	PUNCT	SPELL	NUM	PREP	OTHER	ADV
gold	**202**	323	**390**	34	**230**	639	226
None	162(25%)	491(11%)	314(54%)	39(17%)	194(24%)	631(8%)	242(16%)
Explain	177(26%)	505(11%)	319(52%)	46(21%)	206(24%)	648(8%)	257(13%)
CoT	170(25%)	515(12%)	317(55%)	47(17%)	213(24%)	722(6%)	260(14%)
1-shot CoT	157(28%)	511(10%)	296(49%)	**51(19%)**	221(22%)	758(5%)	293(11%)
2-shot CoT	169(24%)	**520(11%)**	314(52%)	47(19%)	219(25%)	755(6%)	283(13%)
3-shot CoT	168(24%)	478(10%)	287(52%)	45(24%)	206(25%)	**792(6%)**	**297(11%)**
	AUX	NOUN	VERB	CONJ	PRON	ADJ	QUAN
gold	286	369	735	118	141	107	24
None	467(17%)	391(20%)	787(23%)	163(11%)	185(12%)	97(18%)	26(38%)
Explain	448(19%)	395(19%)	765(24%)	151(12%)	173(15%)	101(16%)	30(23%)
CoT	466(18%)	422(19%)	790(22%)	154(14%)	185(14%)	99(18%)	29(27%)
1-shot CoT	481(16%)	449(15%)	**864(18%)**	196(8%)	189(12%)	106(16%)	31(22%)
2-shot CoT	**498(17%)**	432(20%)	860(21%)	177(10%)	179(15%)	96(13%)	28(25%)
3-shot CoT	479(16%)	**450(15%)**	885(17%)	**190(9%)**	**198(12%)**	**120(17%)**	**33(21%)**

表 8: 详细错误类型分析（数量/准确率）

如表7所示，W代表词序错误，S代表替换错误，M代表缺失错误，R代表冗余错误。在基础错误类型上，从准确率的角度来看，大模型在处理缺失错误方面的表现较弱，而在词序错误方面则表现较为出色。从识别出的错误数量上来分析，大模型在纠错过程中不倾向于修改词序错误，而更倾向于修正替换错误。总体来看，虽然生成的替换错误数量较多，但正确率并不高，这表明大模型倾向于将正确的词语修改为模型先前学习到的词语，从而导致过度校正的问题。相比之下，生成的词序错误数量较少，但正确率较高，这可能是因为大模型的训练数据偏向于口语化表达，而测试集中的数据偏向于学术化表达，导致大模型在面对测试集中口语化句子时，倾向于认为其是正确的。

如表8所示，其中SPELL代表拼写错误，PREP代表介词错误，NUM代表数字错误，ADJ代表形容词错误，QUAN代表量词错误等。在详细错误类型方面，大模型在识别SPELm56778L和PREP错误类型上的数量低于gold，而在其他错误类型中，大模型识别出的错误数量均超过了gold。从准确率的角度来看，大模型在处理PUNCT、CONJ和PRON错误类型时表现较差，而在SPELL和QUAN错误类型上表现较好。这说明大模型在纠错过程中不倾

向于修改拼写错误，但其正确率较高，这可能是因为大模型在大量数据训练中更多地学习了正确的token，而未能学习到测试集中的拼写错误token。大模型在量词错误上的良好表现可能是因为中文语法中量词错误相对容易修正。另一方面，大模型在介词错误上的识别数量少于标准答案，这可能因为中文语法中介词用法的复杂性，导致大模型未能充分学习中文介词的使用规则。这些分析揭示了大模型在中文语法纠错任务中的优势和局限性，为未来针对特定错误类型优化大模型的性能提供了一些参考依据。

7 总结

本文采用不同提示策略，评估了国内外主流大模型在中文语法检测和中文语法纠错任务中的性能表现，并对大模型的纠错能力进行了基于错误类型的细粒度评价。实验结果表明，国内大模型在中文文本纠错任务上的表现优于国外大模型，尤其是采用少样本思维链提示策略时，性能表现最为显著，对词序错误和拼写错误的纠正准确率较高。尽管大模型在低资源下能够有效完成中文文本纠错任务，但与有监督模型相比，仍存在一定差距，并存在过度校正的问题。

为了进一步优化大模型的性能，未来的研究可以从以下三个方面着手：1）对大模型的纠错结果进行再训练，以减少过度校正的问题；2）将大模型与有监督的小模型相结合，通过知识蒸馏的方式，在大模型中提取有针对性的知识，以提高小模型的纠错性能；3）通过强化或改进少样本思维链提示策略，进一步提升大模型在中文文本纠错任务中的表现。

致谢

感谢审稿人帮助改善论文的建议。本文受国家自然科学基金青年项目（62306284），中国博士后科学基金面上项目（2023M743189），河南省自然科学基金青年项目（232300421386）资助。

参考文献

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4260–4270, Hong Kong, China, November. Association for Computational Linguistics.

Hongxiang Chang, Yang Liu, Meng Xu, Yingying Wang, Cunliang Kong, Liner Yang, Yang Erhong, Maosong Sun, Gaoqi Rao, Renfen Hu, Zhenghao Liu, 鸿翔常, 洋刘, 萌徐, 莹莹王, 存良孔, 麟儿杨, 尔弘杨, 茂松孙, 高琦饶, 韧奋胡, and 正皓刘. 2023. CCL23-eval 任务7总结报告: 汉语学习者文本纠错(overview of CCL23-eval task: Chinese learner text correction). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 239–249, Harbin, China, August. Chinese Information Processing Society of China.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms fornative chinese grammatical error correction withsupervised fine-tuning. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 69–80, Cham. Springer Nature Switzerland.

Tao Fang, Jinpeng Hu, Derek F. Wong, Xiang Wan, Lidia S. Chao, and Tsung-Hui Chang. 2023a. Improving grammatical error correction with multimodal feature integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344, Toronto, Canada, July. Association for Computational Linguistics.

Tao Fang, Yang Shu, Kaixin Lan, Derek Wong, Jinpeng Hu, Lidia Chao, and Yue Zhang. 2023b. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation, 04.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023c. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *ArXiv*, abs/2304.01746.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada, July. Association for Computational Linguistics.

Kai Fu, Jin Huang, and Yitao Duan. 2018a. Youdao's winning solution to the nlpcc-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *Natural Language Processing and Chinese Computing*.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018b. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 52–59, Melbourne, Australia, July. Association for Computational Linguistics.

Yingjie Han, Yingjie Yan, Yangchao Han, Rui Chao, and Hongying Zan. 2020. Chinese grammatical error diagnosis based on RoBERTa-BiLSTM-CRF model. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 97–101, Suzhou, China, December. Association for Computational Linguistics.

Charles Hinson, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Heterogeneous recycle generation for Chinese grammatical error correction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2191–2201, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July. Association for Computational Linguistics.

Masahiro Kaneko and Naoaki Okazaki. 2023. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *ArXiv*, abs/2309.11439.

Masahiro Kaneko, Kengo Hotate, Satoru Katsumata, and Mamoru Komachi. 2019. TMU transformer system using BERT for re-ranking at BEA 2019 grammatical error correction on restricted track. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, and Torsten Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–212, Florence, Italy, August. Association for Computational Linguistics.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7176–7187, Dublin, Ireland, May. Association for Computational Linguistics.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China, November. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid), December. Association for Computational Linguistics.

Changliang Li and Ji Qi. 2018. Chinese grammatical error diagnosis based on policy gradient LSTM model. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 77–82, Melbourne, Australia, July. Association for Computational Linguistics.

Si Li, Jianbo Zhao, Guirong Shi, Yuanpeng Tan, Huifang Xu, Guang Chen, Haibo Lan, and Zhiqing Lin. 2019. Chinese grammatical error correction based on convolutional sequence to sequence model. volume 7, pages 72905–72913.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022a. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8335–8351, Dublin, Ireland, May. Association for Computational Linguistics.

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. 2022b. Sequence-to-action: Grammatical error correction with action guided sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10974–10982, Jun.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. 2023a. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892, Toronto, Canada, July. Association for Computational Linguistics.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Y. Li, F. Zhou, Haitao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *ArXiv*, abs/2307.09007.

Deng Liang, Chen Zheng, Lei Guo, Xin Cui, Xiuzhang Xiong, Hengqiao Rong, and Jinpeng Dong. 2020. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 57–66, Suzhou, China, December. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. 2021. Understanding and improving encoder layer fusion in sequence-to-sequence learning.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada, July. Association for Computational Linguistics.

Yikang Luo, Zuyi Bao, Chen Li, and Rui Wang. 2020. Chinese grammatical error diagnosis with graph convolution network and multi-task learning. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 44–48, Suzhou, China, December. Association for Computational Linguistics.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China, November. Association for Computational Linguistics.

Masato Mita and Hitomi Yanaka. 2021. Do grammatical error correction models realize grammatical generalization? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4554–4561, Online, August. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Kevin Parnow, Zuchao Li, and Hai Zhao. 2021. Grammatical error correction as GAN-like sequence labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3284–3290, Online, August. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada, July. Association for Computational Linguistics.

Zhaoquan Qiu and Youli Qu. 2019. A two-stage model for chinese grammatical error correction. *IEEE Access*, 7:146772–146777.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Hongkai Ren, Liner Yang, and Endong Xun. 2018. A sequence to sequence learning for chinese grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 401–410, Cham. Springer International Publishing.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79, jan.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *ArXiv*, abs/2311.09517.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online, November. Association for Computational Linguistics.

Xin Sun and Houfeng Wang. 2022. Adjusting the precision-recall trade-off with align-and-predict decoding for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Dublin, Ireland, May. Association for Computational Linguistics.

Xin Sun, Tao Ge, Furu Wei, and Houfeng Wang. 2021. Instantaneous grammatical error correction with shallow aggressive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5937–5947, Online, August. Association for Computational Linguistics.

Minghuan Tan, Min Yang, and Ruifeng Xu. 2023. Focal training and tagger decouple for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5978–5985, Toronto, Canada, July. Association for Computational Linguistics.

Zecheng Tang, Yixin Ji, Yibo Zhao, and Junhui Li. 2021. 基于字词粒度噪声数据增强的中文语法纠错(Chinese grammatical error correction enhanced by data augmentation from word and character levels). In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 813–824, Huhhot, China, August. Chinese Information Processing Society of China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5), dec.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Xiuyu Wu and Yunfang Wu. 2022. From spelling to grammar: A new framework for Chinese grammatical error correction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 889–902, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Haihua Xie, Zhiyou Chen, Jing Cheng, Xiaoqing Lyu, and Zhi Tang. 2020. 基于数据增强和多任务特征学习的中文语法错误检测方法(Chinese grammar error detection based on data enhancement and multi-task feature learning). In Maosong Sun, Sujian Li, Yue Zhang, and Yang Liu, editors, *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 761–770, Haikou, China, October. Chinese Information Processing Society of China.

Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *ArXiv*, abs/2305.14688.

Konstantin Yakovlev, Alexander Podolskiy, Andrey Bout, Sergey Nikolenko, and Irina Piontkovskaya. 2023. GEC-DePenD: Non-autoregressive grammatical error correction with decoupled permutation and decoding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1546–1558, Toronto, Canada, July. Association for Computational Linguistics.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yongwei Zhang, Qinan Hu, Fang Liu, and Yueguo Gu. 2018. CMMC-BDRC solution to the NLP-TEA-2018 Chinese grammatical error diagnosis task. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 180–187, Melbourne, Australia, July. Association for Computational Linguistics.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022b. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2023. Bidirectional transformer reranker for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3801–3825, Toronto, Canada, July. Association for Computational Linguistics.

Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *AAAI Conference on Artificial Intelligence*.

Jianbo Zhao, Si Li, and Zhiqing Lin. 2018a. Contextualized character representation for Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 172–179, Melbourne, Australia, July. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018b. Overview of the nlpcc 2018 shared task: Grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 439–445, Cham. Springer International Publishing.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018a. Chinese grammatical error correction using statistical and neural models. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 117–128, Cham. Springer International Publishing.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li, 2018b. *Chinese Grammatical Error Correction Using Statistical and Neural Models: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, pages 117–128. 08.

王辰成, 杨麟儿, 王莹莹, 杜永萍, and 杨尔弘. 2020. 基于transformer增强架构的中文语法纠错方法. 中文信息学报, 34(6):106–114.

王厚峰罗文. 2024. 大语言模型评测综述. 中文信息学报, 38(1):1–23.

8 附录A.测试提示模板样例

模板
- 模板1：你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。
- 模板2：你是一位资深的编辑，请帮助校对以下内容，完成纠错和润色工作。
- 模板3：你是中文教师，教授外国人学习中文。
- 模板4：在不改变原语义和原语言的情况下，润色输入句子。
- 模板5：查找并校对输入句子中的所有语法错误。
- 模板6：查找并修复这个句子中的所有错误。

输入：任何事情都是各有利弊，众所周知越建立工业越对经济方面有所发展。

正确答案：任何事情都是各有利弊，众所周知**越发展工业，经济方面就越**有所发展。

模板1：任何事物都是各有利弊，众所周知，**越发展工业，对经济发展越**有所**助益**。

模板2：任何事物都**各有利弊**，众所周知，**工业的建立和发展对经济有着显著的推动作用**。

模板3：任何事物都**有其利与弊**，众所周知，**工业的发展对经济有着显著的推动作用**。

模板4：任何事物都**存在其正面和负面效应**，众所周知，**工业的兴起和发展对经济的增长具有显著推动作用**。

模板5：任何事物都**各有利弊**，众所周知，**工业越发展，对经济的推动作用越显著**。

模板6：任何事情都**有利有弊**，众所周知，**工业发展得越充分，对经济发展越有利**。

输入：不官是从口、眼、鼻子进去这样会伤害身体的建康。

正确答案：不**管**是从口、眼**还是**鼻子进去，**都**会伤害身体的**健康**。

模板1：**无论**是从口、眼、鼻子进**入，都**会伤害身体**健康**。

模板2：**无论**是**通过**口**腔**、眼**睛还是**鼻子进**入，这些都可能对身体健康造成伤害。**

模板3：**无论**是**通过**口、眼**还是鼻进入，这些都可能对身体健康造成伤害**。

模板4：**无论**是经口、眼**还是**鼻进**入，这些都可能损害**身体**健康**。

模板5：不**管**是**通过**口、眼、鼻进**入，这些都可能对身体造成伤害**。

模板6：不**管**是**通过**口、眼、鼻进**入，这都**会伤害身体的**健康**。

输入：对一个生名来说空气污染是很危害的问题，对身体不好。

正确答案：对一个生**命**来说空气污染是**有**很**大**危害的问题，对身体不好。

模板1：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体不好。

模板2：对**于生命而言，**空气污染是**一个非常严重**的问题，**它**对身体**健康极为不利**。

模板3：对**于生命**来说，空气污染是**一个非常严重**的问题，对身体**健康有害**。

模板4：对**于任何生物而言，**空气污染**都**是**一个极具**危害的问题，对身体**健康极为不利**。

模板5：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体**有害**。

模板6：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体**有害**。

图 2: 提示模板测试样例