K-CoT：基于关键词思维链提示的中文排比句生成研究

钟茂生，甘家其*，张鹤君，谢林康，李宏伟
江西师范大学，计算机信息工程学院，江西，南昌，330022
Email: {zhongmaosheng, jiaqi_gan, zhanghejun, xie.28, lihongwei}@jxnu.edu.cn

摘要

本文针对中文排比句研究面临的高质量语料匮乏和细粒度标注缺失两大挑战，构建了一个包含主题、情感基调、排比标志词和关键词多维标注的中文排比句语料库。基于此，本文提出了一种基于关键词引导的思维链排比句生成框架K-CoT，通过模拟人类修辞创作的认知过程，将排比句生成分解为“主题解构-特征映射-关键词生成-句式合成”的渐进式推理流程。在ChatGLM和Llama等主流模型上的实验表明，本文提出的K-CoT在排比句生成任务上取得了显著的性能提升。本文为排比句研究提供了一个新颖的数据集，也为生成模型的修辞能力优化提供了可解释的技术路径，其分阶段推理机制对提升语言模型的语义可控性具有普适意义。

关键词： 排比句生成；思维链提示；大语言模型

K-CoT：Chain-of-Thought Prompting with Keywords for Chinese Parallelism Generation

Maosheng Zhong, Jiaqi Gan*, Hejun Zhang, LinKang Xie, Hongwei Li
School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi, 330022
Email: {zhongmaosheng, jiaqi_gan, zhanghejun, xie.28, lihongwei}@jxnu.edu.cn

Abstract

For addressing two fundamental challenges in Chinese parallelism research—the scarcity of high-quality corpora and the absence of fine-grained annotations.This study constructs a multi-dimensionally annotated Chinese Parallelism Corpus encompassing topic, tone, parallel markers, and keywords. Building upon this foundation, we propose K-CoT (Keyword-guided Chain-of-Thought), a novel generation framework that simulates human rhetorical composition through a progressive reasoning pipeline:”topic deconstruction → feature mapping → keywords generation → syntactic synthesis”. Experimental results on mainstream models (ChatGLM, Llama) demonstrate that K-CoT achieves significant performance improvements in parallelism generation. Our work contributes a pioneering linguistic resource, an interpretable technical framework for enhancing generative models’ rhetorical capabilities, and a universally applicable staged-reasoning mechanism that advances semantic controllability in language models.

Keywords: Parallelism Generation , Chain-of-Thought , Large Language Model

*甘家其（通信作者）

第二十四届中国计算语言学大会论文集，第363页-第376页，济南，中国，2025年8月11日至14日。

1 引言

排比作为一种典型的修辞结构，其核心特征在于将三个或以上在语义相关、句法对称、韵律协调的短语或句子进行有序排列，形成具有特定修辞效果的句式结构(Harris and others, 1997)。从结构层次来看，排比句整体可以称为排比块，各分句可以称为排比单元或排比项(Dai et al., 2018)。已有研究表明(熊李艳et al., 2018)，这种修辞方式通过句式结构的规律性重复，能够在多个维度上增强表达效果：其一，强化语言的节奏感和韵律美；其二，突出核心语义的递进关系；其三，提升文本的说服力和感染力。

排比句生成任务是自然语言生成（Natural Language Generation，NLG）领域的一个重要分支。排比句的自动生成面临三个主要挑战：首先，需要保证各排比项在句法结构上的高度一致性；其次，要维持语义层面的逻辑连贯性；最后，还可能需兼顾修辞效果的艺术性。以排比句“科学是探索未知的火炬，照亮人类前行的道路；科学是破除蒙昧的利剑，斩断陈腐观念的枷锁；科学是文明进步的阶梯，托起时代发展的浪潮”为例，该句式不仅实现了比喻意象的递进式排列，更通过“火炬-利剑-阶梯”的隐喻系统，构建了完整的科学认知图景。

基于既有研究基础(Sun et al., 2025)，本研究进一步对排比句的类型学特征进行系统梳理。排比句的类型多种多样，根据不同的划分规则又可以细分为多种排比句。例如，按照分句间的逻辑关系进行划分可以分为并列排比、承接排比和递进排比；按照语言单位的结构层级分可以分为成分排比（排比项为词组或短语，共同充当句子的某一成分）和句子排比（排比项为完整的分句或独立句子）。相比于成分排比，句子排比以完整分句为排比项，每个分句包含主谓宾等完整成分，能独立表达完整命题。而成分排比依赖主句存在，表意片段化。认知心理学实验表明(Just et al., 1992)，人类工作记忆对完整命题的存储效率高于碎片化成分，因此句子排比的完整分句更符合认知组块规律。此外，句子排比能够实现多种修辞嵌套，如在排比句中再嵌入比喻、对偶等修辞手法，使句子更具修辞表现力。同时句子排比可以通过调整分句长度控制节奏，实现韵律的协调。基于上述分析，本文研究主要集中于句子排比生成任务，这一选择既符合人类认知处理的基本规律，又能充分发挥排比句的修辞表现力，为后续的自动生成研究提供更具理论合理性的切入点。

从语言学角度分析，排比句具有态度鲜明和强化主题的特性。通过重复相似的句法结构和韵律模式，排比句能够凸显说话者的主观立场，并引导听者或读者聚焦核心主题。此外，尽管当前大语言模型（如GPT-4、ChatGLM等）在通用文本生成任务中展现出卓越性能，但在排比句生成这一特定修辞任务上仍存在显著不足。我们的实验分析揭示了两类典型问题：结构不对齐和主题不一致问题，如附录A.1中表7所示。针对上述问题，思维链（Chain-of-Thought, CoT）提示提供了有效的解决方案，相较于传统微调范式，CoT参考了人类解决问题的过程，通过逐步思考的方式将复杂问题转化为简单问题逐一进行解决，这种方式也避免了大规模参数更新。此外，Brown (2020)的研究证明上下文学习（In-Context Learning，ICL）可有效替代特定任务的微调，展现了大语言模型强大的零样本或少样本学习能力。受此启发，本文旨在提升大语言模型在排比句生成任务中的表现，通过ICL增强提示学习，同时使用CoT提示来优化ICL，以增强大语言模型的泛化能力、推理能力和任务适应性。

综上所述，本文以人类创作排比句的思路作为出发点，提出了一种基于关键词引导的CoT提示框架K-CoT。该框架主要分为数据聚类、关键词推导、排比句生成三个阶段。具体而言，以给定主题（Topic）和基调（Tone）作为初始输入，在数据聚类阶段，通过聚类算法对数据进行分簇，从而选出形式多样、内容多样的上下文演示样例。在关键词推导阶段，构建了从概念到具体的认知推理链，该链条模拟了人类创作时“主题解构→特征映射→关键词生成”的思维过程。在排比句生成阶段，结合上两个阶段的结果生成完整的排比句，并在生成过程中融入大模型的自我反思过程以提升情绪支持效果，模仿了人类“先写初稿，再润色语言”的写作策略。本文通过人工构建的数据集分析了该框架在提高排比句生成质量方面的潜力和挑战，并探讨未来的研究方向。本文主要有以下三个贡献:

1. 构建了CPDAK（Chinese Parallelism Dataset with Annotated Keywords）数据集[-1]，这是一个具有关键词标注特征的中文排比句数据集，旨在通过结构化标注提升大语言模型生成排比句的质量，填补该领域高质量标注数据的空白。

[-1]https://anonymous.4open.science/r/Chinese-Parallelism-Dataset

2. 从认知语言学角度出发，通过模拟了人类构思排比句的思维过程，构建了一个具有认知合理性的生成框架。该框架将传统的黑箱式文本生成转化为可解释的渐进式推理流程，使生成过程具有更好的可解释性，为研究语言模型的修辞能力提供了新的理论视角和技术路径。

3. 提出了一种基于关键词CoT排比句生成框架K-CoT，在少样本场景下，实验结果表明，K-CoT能在多种大模型上实现了性能的提升，表明了K-CoT具有较强的泛化能力。

2 相关工作

2.1 中文排比句识别与生成

中文排比句生成是一项有趣的自然语言生成任务。自20世纪50年代起，自然语言生成技术初现于机器翻译领域，研究者通过人工规则和双语词典构建语言间的映射关系，将翻译视为符号的机械替换(Stahlberg, 2020)；至80至90年代，统计学习方法成为主流，基于马尔可夫模型(Gruber et al., 2007)和概率图模型(Teh et al., 2004)，系统能够从语料库中自动学习生成短文本，实现了初步的自动化语言生成；进入21世纪后，大模型的兴起推动自然语言生成进入新阶段，使模型不仅能够处理长文本和复杂语境，更开始展现对语义的深层理解与创造性表达能力(Wan et al., 2023)。自然语言生成的研究也逐渐向更多应用领域拓展，比如：在对话系统中实现多轮交互，在智能问答中完成精准推理，在辅助写作中生成连贯篇章。

排比句讲究结构对齐、逻辑连贯、内容关联(Mabona et al., 2019)，近年来与其相关的研究主要集中在排比句的识别方向。目前在自然语言处理领域，排比句的识别已经取得了一定的进展。早期主要是根据排比句的句法特征制定相应的语法规则进行识别(梁社会et al., 2013)。机器学习兴起时，Song(2016)提出了一种基于机器学习的学生作文句子排比识别方法，通过结合广义词汇对齐策略和词序列对齐度量，利用句法树核函数计算深层特征进行排比识别。随着深度学习技术的发展，利用其提取文本语义特征以自动识别排比句，已然成为主流的研究方法。穆婉清(2018)根据排比句的句法结构相似和内容相关的特点，设计了一种将卷积神经网络与结构相似度相结合的排比句识别方法；Dai(2018)提出了一种通过引入循环神经网络来捕捉句子的语义特征的排比句识别方法；朱晓亮(2021)则采用融合预处理算法与BERT模型来优化排比句的自动识别过程，这些工作共同推动了排比句识别技术的发展。

现有的方法想要提升排比句生成质量，往往需要大量训练数据和全参数微调，这在很多场景下是难以满足的。如钟茂生(2025)通过分析排比句的语法特征和句法结构，提出了一种基于词性对齐和依存关系的中文排比句生成方法，使模型的性能得到了较大的提升，同时该工作表明了排比句的语法特征和句法结构对排比句生成任务的重要性，但受限于小模型的局限性(Xu et al., 2023)，在处理较长的排比句生成时效果往往不佳。如表1所示，我们观察到，它缺乏对排比句的重要元素的标注，同时可以观察到平均每一句token数相对较短，且排比句风格集中于公文写作类型。本文认为排比句同样高频用于学生作文或文章等来抒情写景，为此我们构建了一个抒情写景类型的中文排比句数据集。本文还发现在较长的排比句写作中，可以与其他修辞手法相结合，使排比句更具感情色彩和修辞表现力。

	CPDAK	Parallelisms Dataset
# Sentence	10023	30452
# Annotated Parallelism	6529	/
# tokens	433K	777K
# tokens per sentence	68	26

表 1: 现有排比句数据集(钟茂生et al., 2025)和CPDAK数据集的统计信息

2.2 通过提示学习促进自然语言生成

随着预训练语言模型(Pre-trained Language Model，PLM)规模的不断扩大，传统微调方法面临高昂计算成本的挑战。在此背景下，提示学习因其高效性受到广泛关注，其中Wei(2022)提出的CoT提示通过显式引入中间推理步骤，显著提升了模型输出的合理性与质量(Yang et al.,

2023b)。该方法在NLG领域取得显著成果，包括文本风格迁移(Suzgun et al., 2022)、数学推理(Yue et al., 2023)和生成评估(Jiang et al., 2023; Chan et al., 2023; Chen et al., 2025)等任务。特别是在修辞生成方面，Shao(2024)将CoT应用于比喻句生成，以喻意为中间推理步骤，有效提升了生成质量。此外，还有如倪宣凡(2022)的工作等。然而，目前大模型在排比句生成方面的研究仍属空白，本文旨在系统探索CoT提示在中文排比句生成中的应用。

我们观察到，排比句的关键词可以作为思维链提示的一个中间步骤，能够有效地连接主题和基调，为模型生成高质量的排比句提供关键信息。在本研究中，通过利用人工构建的CPDAK语料库系统评估了大型语言模型在中文排比句生成任务中的表现。实验数据表明，当大模型结合本文提出的框架K-CoT后，模型生成的排比句在结构完整性、主题一致性和修辞丰富性等维度均取得显著改善。

3 中文排比数据集

3.1 定义

排比是一种重要的修辞格，广泛应用于各种语篇。从古典文学到现代演讲；从诗歌到电影对白；从哲学思辨到日常对话，排比都是加强信息传递和情感表达的有效手段。

为了系统刻画排比句的语言特征，本文建立了多维度标注体系。从概念维度出发我们标注了排比句的主题，它是排比句的核心论述对象，各排比项需要保持主题的一致性；从情感维度出发我们标注了排比句的基调，它代表了排比句的情感倾向或语义功能；从认知维度出发我们标注了排比句的关键词，它实现了主题具象化的语义节点。例如，“奋斗让我们在困境中成长；奋斗让我们在挑战中突破；奋斗让我们在坚持中成功。”这一排比句以“奋斗”为主题，以“激励”为基调，“困境、挑战、坚持”为关键词序列，旨在赞扬奋斗精神，呼吁我们不甘于平凡。

3.2 数据收集

为构建CPDAK语料库，本文从散文、小说、议论文及学生作文等富含修辞手法的中文文学资源中收集了大量潜在的排比句。表2展示了构建CPDAK语料库使用的主要文本数据统计信息。

Category	Books	Tokens	Sentences
Children’s Books	195	17M	0.58M
Chinese literature	336	64M	2.2M
Translated literature	854	121M	4.2M

表 2: 中文文本数据统计信息

初步筛选：首先，我们需要从大量的文本数据中筛选出潜在的排比句。根据排比句的定义，本文将排比句的识别问题转化为语义文本匹配(Jiang et al., 2019; Hu et al., 2019)问题，任意两个排比项间应该具有以下特点：

- 各排比项在相近位置具有相同的词语，又称为排比标；
- 各排比项具有结构相似性；
- 各排比项具有相同的情感极性(Ku and Chen, 2007)。

Guégan(2006)之前的研究表明排比具有传递性，即如果排比项1和排比项2属于排比句，排比项2和排比项3属于排比句，则排比项1和排比项3也属于排比句，由此可以得出一个完整的排比句{排比项1，排比项2，排比项3}。鉴于此，本文认为可以通过计算各排比项之间的语义匹配置信度。如果语义匹配置信度较高，则该句子更有可能为排比句而非字面表达。已有研究表明通过词向量可以获得词语之间的语义信息(Kusner et al., 2015)，如果任意两个排比分句分别表示为n维向量$\mathbf{w}$和$\mathbf{v}$，则它们的余弦相似度得分为：

$$\cos(\mathbf{w},\mathbf{v}) = \frac{\sum_{i=1}^{n} w_i v_i}{\sqrt{\sum_{i=1}^{n} w_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \quad (1)$$

第二十四届中国计算语言学大会论文集，第363页–第376页，济南，中国，2025年8月11日至14日。

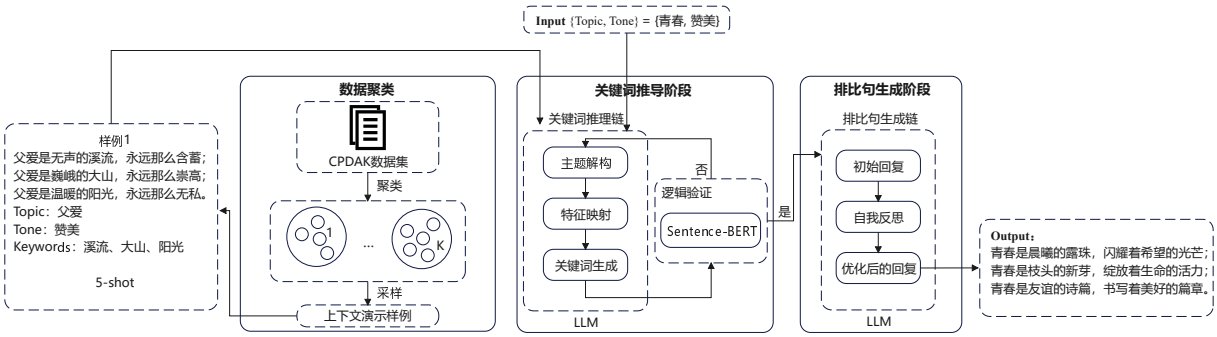

图 1: K-CoT框架

如果需要人工从大量的文本数据中筛选潜在的排比句，那将花费巨大的成本。因此，为减少人工筛选量，我们基于Text2vec[0]训练了一个中文排比句分类器，通过置信度阈值$\cos(\mathbf{w}, \mathbf{v}) \geq 0.673$自动预筛选潜在排比句。具体细节如附录A.2所示，该分类器能有效过滤非排比句，显著提升筛选效率。

3.3 数据标注

本文采用细粒度的四元组标注框架，对语料库中的每个排比句进行全面系统的标注。具体而言，每个排比实例被标注为（Topic, Flag, Tone, Keywords）核心特征的元组，具体示例如附录A.3中表8所示。

标注过程主要分为两个阶段：初步标注和精细标注。

- **初步标注：**在这一阶段，标注者团队主要由5名中国大学生构成，从初步筛选后的数据集中筛选出真实的排比句，并初步标注排比句的主题、基调、标志词和关键词。

- **精细标注：**在初步标注的基础上，第二批标注者重新完善标注。他们具有一定的中文文学基础，在标注过程中提供中文排比句标注指南作为参考，他们能够更加精确高效的标注排比句的各成分。最后统一由三名标注者审阅，以确保标注的准确性和一致性。

- **中文排比句标注指南：**我们在第二轮精细标注中使用了一种严谨且全面的标注方法，重点在于对中文排比句的各成分进行精细化标注。

  1）标注和质量检查规则：鉴于中文排比句结构的复杂性和多样性，必须统一标注所有相关要素，包括主题、排比词、基调和关键词。标注者需参考之前的标注结果，尤其注意主题和关键词的准确性，并关注基调判断的一致性，以避免可能的错误。在第二轮标注中，我们对格式要求进行了标准化，采用JSON格式进行标注。经人工校验后，可利用程序自动提取标注结果，用于构建高质量的标注排比句语料库。

  2）复核一致性：为了确保基调标注的准确性和一致性，本文提供了一份基调描述词列表。标注者需根据排比句表达的整体情感色彩，从列表中选择最合适的描述词。我们强调选择更精准的描述词，避免使用过于宽泛的词语。这份列表对于确保基调准确反映排比句的情感色彩或态度至关重要。关键词的提取是排比句标注的核心，标注的关键词需能与主题关联的核心词汇。应避免提取过于泛化的词语，尽量选择更具体的、更具代表性的词语。关键词的数量不宜过多，一般不超过5个，因为过长的排比句可能会引起读者疲劳。

4 K-CoT框架

本文提出的排比句生成框架如图1所示，分为三个阶段：数据聚类阶段，通过聚类算法对训练集数据分簇，形成高质量且多样的上下文提示示例；关键词推导阶段，在基调的基础上，生成与主题相关的关键词序列；排比句生成阶段，结合上两个阶段获得的上下文示例和关键词，引导大模型生成最终的排比句。

[0]https://github.com/shibing624/text2vec

4.1 任务定义

当在写作要使用排比句时，我们往往清楚排比句的主题以及情感色彩或态度。因此，该任务首先给定主题（Topic）和基调（Tone）作为初始条件，引导模型进行渐进式推理：模型需要先分析主题在特定情感基调下的语义特征，识别核心概念维度（如将“勇气”分解为“强度”和“担当”），然后将这些抽象特征映射为具象关键词（如“磐石”象征强度，“肩膀”代表担当），最终基于这些语义连贯的关键词序列生成结构规范的排比句。这种显式的推理链条能有效确保各分句在概念层面的内在关联性，从而解决排比句主题不一致问题。

4.2 数据聚类

目前已有相关研究表明，引入多样化的演示样本构建上下文提示，可以有效地提高模型地ICL能力(Ye et al., 2023)。受此启发，为了增加演示样本的多样性，本阶段的任务主要是将数据集中的训练数据进行聚类成簇。本文采用混合聚类策略来增强示例选择的多样性。通过结合BERT模型的[CLS] token实例级聚类（k=20）和关键词级的词汇嵌入聚类（k=20），我们构建了一个双通道聚类框架，该框架同时考虑了文本的整体语义特征和局部关键词分布模式。在实施过程中，我们通过轮廓系数（> 0.6 (Shutaywi and Kachouie, 2021)）优化聚类质量，确保每个簇内的样本既保持语义一致性又具备足够的表达多样性。这种分层聚类方法能够有效捕捉排比句在宏观语义结构和微观修辞特征上的变化，从而为后续的少样本提示学习提供更具代表性的示例集合。最终从20个语义簇中策略性采样构建提示示例库。

4.3 思维链提示

本文采用CoT提示策略来增强排比句生成的连贯性和逻辑性。具体而言，包含关键词推导、排比句生成两个阶段。在模型完成关键词推导生成后，我们将其输出结果作为后续提示的上下文信息，指导模型生成符合语义结构和情感要求的完整排比句。通过关键词构建递进式的生成流程，这种流程是分阶段、渐进式的提示设计，不仅保留了中间推理步骤的逻辑链条，还通过显式地引入前序生成结果作为约束条件，显著提升了最终排比句在主题一致性和修辞质量方面的表现。

关键词推导阶段：主题理解是排比句生成的基础，为实现这一目标，需要对主题语义进行多层次解析，以提升大语言模型对核心概念的捕捉能力，确保生成内容的概念一致性。为此，本文通过模拟人类构思排比句时的认知过程，构建“主题解构-特征映射-关键词生成”的三步推理链，其过程可以写成如下：

$$Chain_{\text{keywords}} : c \rightarrow a \rightarrow k \tag{2}$$

其中，c表示初始输入（主题和基调），a表示属性，k表示关键词。具体实施时，首先对输入主题（如“青春”）进行语义解构，识别其核心特征维度（如“短暂性”、“能量感”、“成长性”）；随后结合给定基调（如“活力”），将这些抽象特征映射为具象关键词（如“朝露-闪电-年轮”）。最后，根据推理链$Chain_{\text{keywords}}$构建的指令I、初始输入内容c和上下文演示示例E_c构建提示词，并将其作为大模型的输入，获得符合基调的关键词K_c:

$$\begin{aligned} prompt_{\text{keywords}} &\leftarrow c, \text{I}, \text{E}_c \\ K_c &= \text{LLM}(prompt_{\text{keywords}}) \end{aligned} \tag{3}$$

此外，为了确保每个关键词满足与主题的语义相关性，本文增加了逻辑验证步骤，即利用Sentence-BERT(Reimers and Gurevych, 2019)模型获得主题与各关键词的词嵌入表示，并计算主题与各关键词的余弦相似度，当相似度大于0.75(Song et al., 2016)则保留用于后续生成，否则需重新生成，该步骤能有效保证排比句的主题一致性。

排比句生成阶段：基于前阶段产生的关键词，再结合上下文演示样例提示大语言模型生成排比句。将生成过程分解，形成“生成-反思-再生成”的生成过程，通过“自我优化”使生成的排比句更加流畅工整:

$$Chain_{\text{parallelism}} : R_{\text{original}} \rightarrow (reflection) \rightarrow R_{\text{revised}} \tag{4}$$

具体而言，首先要求大语言模型根据给定的初始输入c、上下文的演示样例E_c与关键词K_c生成一个初始回复R_{original}，接着要求其根据初始回复，思考如何达到更好的情感表达效

果，以丰富句子情感表现力，最后根据思考内容优化初始回复，生成最终回复R_{revised}：

$$\begin{aligned} prompt_{R_{\text{original}}} &\leftarrow c, \mathrm{E}_c, K_c \\ R_{\text{original}} &= \mathrm{LLM}(prompt_{R_{\text{original}}}) \\ prompt_{R_{\text{revised}}} &\leftarrow prompt_{R_{\text{original}}}, \mathrm{I}, R_{\text{original}} \\ R_{\text{revised}} &= \mathrm{LLM}(prompt_{R_{\text{revised}}}) \end{aligned} \tag{5}$$

5 实验

5.1 实验设置

本文通过Hugging Face提供的模型参数进行本地化部署与管理，使用单块RTX-3090 GPU进行模型推理。为了验证本文提出的框架在排比句生成任务的有效性，同时为了避免大语言模型自身特性带来的性能影响。我们选择了几个具有代表性的大语言模型进行实验。此外，在超参数设置方面，除温度参数($T = 0.8$)外均保持默认设置。因为排比句生成是一个创造性任务，过小的超参数T会限制模型的创造性表达。所有模型均采用相同的提示模板和评估指标，确保实验结果的公平性和可比性。

- **ChatGLM3-6B**(Du et al., 2022)[1]是一个开源的、支持中英双语的对话语言模型，基于General Language Model架构，具有62亿参数。
- **Baichuan2-7B**(Yang et al., 2023a)[2]是百川公司推出的新一代开源大语言模型，在多个权威的中文、英文和多语言的通用、领域基准上取得同尺寸最佳的效果。
- **Qwen-2.5-7B**(Jin Xu, 2025)[3]是Qwen大语言模型的最新系列，在遵循指令、生成长文本和生成结构化输出方面有显著改进，因此能够处理复杂的排比句生成任务。
- **Yi-1.5-9B**(AI et al., 2024)[4]是Yi系列新一代从零开始训练的开源双语语言模型，在语言认知、尝试推理、阅读理解等方面表现优异。
- **Llama3-8B-Chinese-Chat**(Wang et al., 2024)[5]是一个基于Meta-Lama-3-8B-Instruct(Touvron et al., 2023)模型的针对中英文用户的指令调优语言模型，通过ORPO(Hong et al., 2024)专门为中英文用户微调的第一个模型，能够处理多种中文语言处理任务，包括排比句的生成。

5.2 基线方法

本文提出的框架不涉及模型参数的调整，因此从两个方面选取基线方法进行对比，一方面是传统的微调方法，另一方面是大模型领域的提示方法。由于排比句生成研究相对较少，对于传统的微调方法，本文选取了一个具有代表性的排比句生成模型作为基线：

- **CPGen-POS&DEP**(钟茂生et al., 2025)是一种基于词性对齐与依存关系的中文排比句生成模型，通过学习词性对齐特征与给定分句的依存关系生成高质量的排比句。

对于大模型领域的提示方法，所采用的基线方法如下：

- **Vanilla**：通过在提示信息中加入“Let’s think step by step”来激发大语言模型的推理能力，引导模型根据初始输入进行排比句生成。
- **CoT**：提示词中的演示样例是随机产生，其它过程与本文框架一致。

[1]https://huggingface.co/THUDM/chatglm3-6b-32k
[2]https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat
[3]https://huggingface.co/Qwen/Qwen2.5-7B
[4]https://huggingface.co/01-ai/Yi-1.5-9B-Chat
[5]https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat

Model	Method	BLEU↑		Distinct↑	
		BLEU-1	BLEU-4	MA-D-1	MI-D-1
CPGen-POS&DEP	Finetune	51.74	32.06	75.45	7.23
ChatGLM	Vanilla	52.29	36.15	67.62	12.91
	CoT	58.27	44.70	67.62	13.52
	K-CoT	61.53	46.62	74.82	17.18
Baichuan	Vanilla	59.46	43.09	66.53	11.68
	CoT	73.23	61.08	67.00	12.42
	K-CoT	75.81	63.03	75.34	13.58
Qwen	Vanilla	54.88	40.94	69.86	11.91
	CoT	74.16	62.84	70.00	11.56
	K-CoT	73.53	**66.37**	76.11	17.05
Yi	Vanilla	55.50	36.88	67.62	12.91
	CoT	74.42	62.52	67.78	12.94
	K-CoT	75.37	60.35	74.82	17.18
Llama	Vanilla	58.59	43.42	67.00	13.05
	CoT	74.52	63.00	67.51	12.86
	K-CoT	**76.85**	65.93	**79.95**	**19.26**

表 3: 自动评估结果

5.3 评估指标

鉴于排比句的复杂性和新颖性，仅凭自动评估指标可能无法充分反映生成质量。本文参考诗歌和对联生成等类似任务的评估方法，发现结合自动和人工评价是目前评估艺术性文本生成的主流做法。

自动评估指标：为了全面评估性能情况，本文选择了两个经典自然语言生成评估指标。我们使用BLEU值(Papineni et al., 2002)评估生成文本与参考文本的词汇重叠度，该指标能够有效反映排比句的语义准确性；同时采用Distinct指标(Li et al., 2016)计算n-gram的重复率，其中Macro-Distinct衡量单一样本的词汇丰富度，Micro-Distinct评估整体语料的多样性分布。

人工评估指标：为全面评估排比句生成质量，本文建立了五维人工评价体系。我们聘请了多位语言学相关专业评估员，从以下维度进行1-3分的评分，最后取平均分作为最终得分：

- 结构一致性（Structural）：评估排比项间的句法对齐程度，重点关注：各排比项的主题一致性；句子长度和韵律的协调性；标志性词语的重复模式。
- 流畅性（Fluency）：衡量排比句表达的流畅度，主要考察：词语表达是否符合语言习惯；逻辑递进关系的合理性；概念表达的完整性。
- 情感表现力（Emotional）：评价修辞效果的情感维度，具体包括：情感基调的一致性；情感强度的递进梯度；心理共鸣的激发效果。
- 创新性（Creativity）：评价语言表达的创新程度，着重分析：隐喻意象的原创性；概念组合的独特性；陈词滥调的规避程度。
- 语言生动性（Vividness）：考察修辞手法的运用效果，具体关注：比喻、夸张等修辞的恰当性；感官词汇的丰富度；意象构建的鲜明度。

5.4 实验结果与分析

实验结果表明（如表3所示），本文提出的K-CoT框架在排比句生成任务中显著优于Vanilla提示和标准CoT提示方法。具体而言，在Qwen模型上，K-CoT的BLEU-4指标达到66.37，较CPGen-POS&DEP基线提升107%，同时在其他指标上均有显著提升。这一优势

主要源于K-CoT对人类创作认知过程的建模：通过分阶段的关键词引导机制，实现了从抽象概念到具象表达的渐进式转换；基于思维链的推理过程确保了语义逻辑的连贯性；多维度约束使生成结果更符合语言表达规范。这些设计使得生成排比句在流畅性和人类偏好度方面接近人类创作水平。

人工评估结果如表4所示，K-CoT生成的排比句在五个维度上均优于基线模型。尤其在Llama模型上，K-CoT的最终得分达到2.35，较基线模型Vanilla提高了34%。这一结果表明，分阶段的关键词引导机制通过显式语义分解，确保了排比句的结构一致性和流畅性。思维链的渐进式推理模拟了人类修辞创作过程，使情感表达的形成合理递进，同时关键词间的逻辑约束有效避免了语义断裂。此外，与CoT相比，采用聚类的采样方式能取得更高的人类评估得分，进一步表明了对演示示例进行聚类采样能够提高模型的生成排比句质量。

Model	Method	Structural	Fluency	Emotional	Creativity	Vividness	Final_Score
ChatGLM	Vanilla	1.53	1.18	1.35	1.95	1.59	1.52
	CoT	2.27	2.17	1.79	1.75	2.04	2.00
	K-CoT	2.34	2.25	1.93	1.89	1.99	2.08
Baichuan	Vanilla	1.45	1.34	1.26	2.03	1.51	1.52
	CoT	2.34	2.24	1.81	1.75	1.86	2.00
	K-CoT	2.30	2.25	1.85	1.96	2.19	2.11
Qwen	Vanilla	2.23	2.08	1.77	2.01	1.91	2.00
	CoT	2.43	2.34	2.08	2.05	2.28	2.24
	K-CoT	**2.55**	**2.50**	2.17	**2.23**	**2.41**	**2.37**
Yi	Vanilla	2.26	2.31	1.89	2.13	2.12	2.14
	CoT	2.43	2.39	2.06	1.94	2.28	2.22
	K-CoT	2.50	2.49	2.18	2.05	2.34	2.31
Llama	Vanilla	1.76	1.70	1.68	1.97	1.64	1.75
	CoT	2.36	2.38	2.27	2.00	2.19	2.24
	K-CoT	2.52	2.46	**2.41**	2.12	2.22	2.35

表 4: 人工评估结果

5.4.1 不同LLMs的表现

根据表4和5，本文得出两个主要结论。首先，Yi、Qwen和Llama模型的整体表现优于ChatGLM和Baichuan，这种性能差异可能源于模型架构的迭代改进，较新发布的模型（如Qwen-2.5）采用了更高效的注意力机制和训练策略，使其在结构准确性和情感连贯性等关键指标上表现更优。此外，模型并非总能生成合理的中文排比句。这些结果表明，中文排比句生成仍是一个复杂且尚未被充分探索的任务。

5.4.2 人工评价指标与综合得分的相关性分析

根据表6的分析结果，各评估指标与最终得分均呈现出不同的相关性。其中流畅性和结构一致性与最终得分呈现强相关性，这表明评委在评估排比句时最重视语言表达的流畅程度和句式结构的规范性。情感表现力与创意性呈现中等程度相关，反映出在满足基础语言质量的前提下，修辞的情感渲染力和概念创新性对评分具有补充性影响。值得注意的是，语言生动性的相关性相对较低，可能源于两方面原因：其一，生动性评价本身具有一定主观性，不同评委对修辞手法的偏好存在差异；其二，当排比句在结构和流畅性方面表现突出时，评委可能降低对生动性的严格要求。这些发现为排比句生成模型的优化提供了明确方向，即应当优先保证语言流畅性和结构规范性，在此基础上再适当提升情感表达和创意性水平。

6 总结与展望

本文提出了一个标注完善的中文排比句数据集，包含约6500个句子，来源于广泛的中文文学形式，如散文、小说和文章等体裁。为了确保标注的准确性和一致性，本文制定了一套详尽的标注规范，用于指导标注者标注句子的主题、基调以及关键词等重要信息。此外，本文提出

Model	Method	ACC
ChatGLM	Vanilla	0.26
	CoT	0.31
	K-CoT	0.39
Baichuan	Vanilla	0.33
	CoT	0.46
	K-CoT	0.55
Qwen	Vanilla	0.44
	CoT	0.59
	K-CoT	**0.67**
Yi	Vanilla	0.41
	CoT	0.53
	K-CoT	0.62
Llama	Vanilla	0.51
	CoT	0.59
	K-CoT	0.65

表 5: 模型生成的句子是合理排比句的百分比

	Structural	Fluency	Emotional	Creativity	Vividness
Final_Score	0.75	**0.82**	0.55	0.56	0.48

表 6: 最终得分与人工评估评分之间的Pearson相关性

的K-CoT框架通过模拟人类创作排比句的认知过程，创新性地将关键词与思维链推理相结合，在排比句生成任务中取得了显著成效。实验结果表明，该框架不仅显著提升了生成质量，更重要的是建立了一个可解释的修辞生成范式，其分阶段的"主题解构-特征映射-关键词生成"推理链有效解决了排比句中的结构对齐不佳和主题不一致的问题。

在未来工作中，本文将K-CoT框架扩展至其他修辞生成任务，验证其泛化能力。另外，针对如何平衡规范性与创造性，未来可以探索引入基于人类反馈的强化学习机制，更好地平衡生成结果地规范性与创造性。最后，我们还计划扩展数据集，涵盖更多领域和语言，以更全面地评估方法的通用性和鲁棒性。我们相信，这项工作将为排比句生成的研究和应用提供有价值的参考和启示。

参考文献

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

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.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Wentong Chen, Yankai Lin, ZhenHao Zhou, HongYun Huang, YanTao Jia, Zhao Cao, and Ji-Rong Wen. 2025. ICLEval: Evaluating in-context learning ability of large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10398–10422, Abu Dhabi, UAE, January. Association for Computational Linguistics.

第二十四届中国计算语言学大会论文集，第363页-第376页，济南，中国，2025年8月11日至14日。

Yange Dai, Wei Song, Xianjun Liu, Lizhen Liu, and Xinlei Zhao. 2018. Recognition of parallelism sentence based on recurrent neural network. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 148–151. IEEE.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May. Association for Computational Linguistics.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial intelligence and statistics*, pages 163–170. PMLR.

Marie Guégan and Nicolas Hernandez. 2006. Recognizing textual parallelisms with edit distance and similarity degree. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 281–288.

Robert Harris et al. 1997. A handbook of rhetorical devices.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA, November. Association for Computational Linguistics.

Weiwei Hu, Anhong Dang, and Ying Tan. 2019. A survey of state-of-the-art short text matching algorithms. In *International conference on data mining and big data*, pages 211–219. Springer.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *The world wide web conference*, pages 795–806.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Just, A. M., Carpenter, and A. P. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1):122–122.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–1850.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 2284–2295, Hong Kong, China, November. Association for Computational Linguistics.

Xuanfan Ni and Piji Li. 2022. 融合提示学习的故事生成方法(a story generation method incorporating prompt learning). In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 166–177.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Wenhao Huang, Ge Zhang, and Jie Fu. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3357–3366, Torino, Italia, May. ELRA and ICCL.

Meshal Shutaywi and Nezamoddin N. Kachouie. 2021. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23(6).

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 794–803.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Maosong Sun, Jiye Liang, Xianpei Han, Zhiyuan Liu, Yulan He, Gaoqi Rao, Yubo Chen, and Zhiliang Tian, editors. 2025. *Chinese Computational Linguistics: 23rd China National Conference, CCL 2024, Taiyuan, China, July 25–28, 2024, Proceedings*, volume 14761 of *Lecture Notes in Computer Science*. Springer Nature Singapore, Singapore.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Yee Teh, Michael Jordan, Matthew Beal, and David Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024. Llama3-8b-chinese-chat (revision 6622a23).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023b. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

朱晓亮and 谯宇同. 2021. 基于bert 模型的排比句自动识别方法. 计算机应用与软件, 38:7.

梁社会, 陈小荷, and 刘浏. 2013. 先秦汉语排比句自动识别研究–以《孟子》《论语》中的排比句自动识别为例. 计算机工程与应用, 49(19):222–226.

熊李艳, 林晓乔, and 钟茂生. 2018. 面向自动写作的中文排比句抽取方法. 计算机应用研究, 35(06):1751–1755.

穆婉青, 廖健, and 王素格. 2018. 融合cnn 和结构相似度计算的排比句识别及应用. 中文信息学报, 32(2):139–146.

钟茂生, 刘蕾, 吴如萍, 甘家其, and 周新宇. 2025. 基于词性对齐与依存关系的中文排比句生成方法. 中文信息学报, 39(02):131–142.

附录

附录A.1 排比句生成初始分析

前期调研时，我们在多个大语言模型上进行排比句生成任务，发现模型输出往往得不到满意的结果，总结分析后，除了生成合理的排比句频率不高之外，还存在如表7中的两个问题，即结构不一致和主题不一致。第一个例子表明生成的句子各排比项结构不一致，不符合排比句结构对齐特性。第二个例子表明句子的核心主题不一致，从前两句排比项可知主题为勤奋，但第三句排比项的主题却为悠闲，尽管它的结构工整，但它仍然不是一个完美的排比句。我们认为这可能有两个原因：语言模型通常以token预测为目标，缺乏对修辞结构的显式建模；高质量排比句在预训练语料库中的占比有限。

1.勤奋是基石，为梦想筑牢根基； 勤奋是灯塔，照亮前行的每一条路； 勤奋是翅膀，助力我们飞向更高的天空。	**结构不对齐**
2.勤奋是阶梯，攀登成功的每一步； 勤奋是钥匙，开启智慧的每把锁； 悠闲是清风，拂过心灵的每一面。	**主题不一致**
3.勤奋是阶梯，攀登成功的每一步； 勤奋是钥匙，开启智慧的每把锁； 勤奋是航船，航行梦想的每一程。	**合格**

表 7: 大语言模型生成排比句存在问题与合格示例

附录A.2 排比句识别

针对现有排比句数据集仅包含正样本的局限性（(钟茂生et al., 2025)），本研究构建了一个包含正负样本的平衡数据集。负样本主要涵盖三类易混淆文本：

1. 对联文本：具有工整结构但语义对仗，如”校内桃李相映辉，园外柏松互比高”；
2. 反复句：通过重复相似句式强化语气，如”哪里是山，哪里是房屋，哪里是菜园”；
3. 普通陈述句：从学生作文中筛选的句子。

我们从多个网站获取284个反复句样本[6]，收集1000条典型对联[7]，并人工筛选2000个普通句（确保不含排比）。正样本包含2583个经人工验证的排比句[8]。最终构建的数据集共5195个样本（正：2583，负：2612），按8:1:1划分为训练集、验证集和测试集。基于这个新数据集，我们

[6]https://www.yuwenmi.com/zaoju/708260.html
[7]https://www.yuwenmi.com/duilian/
[8]https://www.chazidian.com/zaoju$_4$/

训练了一个基于Text2vec的$\text{RoBERTa}_{\text{Large}}$中文排比句分类器，在测试集上的排比句和非排比句的置信度得分分别为0.810和0.537，我们取平均值作为阈值，即当置信度得分大于0.673则视为潜在的排比句。为了训练这个模型，我们将学习率设置为5e-5，预热步长为200，批量大小设置为16，模型迭代次数为10次。

附录A.3 CPDAK数据集样例

我们在表8中展示CPDAK数据集的部分来源以及标注示例。

Source Type	Parallelism	Topic	Flag	Tone	Keywords
Prose	时间是生命的刻度，记录着我们成长的轨迹； 时间是历史的画卷，描绘着时代变迁的景象； 时间是自然的韵律，谱写着万物生长的乐章。	时间	时间是	比喻 哲理	刻度 画卷 韵律
Composition	青春是一场旅行，沿途风景书写回忆； 青春是一次飞翔，蓝天白云见证自由； 青春是一首诗篇，字里行间抒发激情。	青春	青春是	赞美 感激	一场旅行 一次飞翔 一首诗篇
Article	让阅读成为一种习惯，滋养心灵的沃土； 让思考成为一种习惯，点亮智慧的明灯； 让行动成为一种习惯，开辟成功的道路。	习惯	让...成为 一种习惯	成长 智慧	阅读 思考 行动

表 8: CPDAK数据集标注示例

附录A.4 实例分析

图2展示了不同模型在三个设置下的生成实例对比分析。在Vanilla中，模型生成的排比句普遍存在结构不对齐和主题不一致问题，如标红部分所示。引入随机示例和关键词思维链（CoT）后，有效缓解了结构不对齐和主题不一致问题，显著提升了生成质量，表现为：结构创新性增强，如Yi采用“如...般”的明喻结构，使排比句更具修辞表现力；语义深度拓展，如Llama的“露珠-太阳-雄鹰”形成微观到宏观的递进，相比于Vanilla的内容更加丰富；情感表达丰富，如Qwen的“旅行-飞翔-诗篇”多维度刻画青春，展现青春的多姿多彩。在K-CoT中，生成的排比句更能体现用户的基调倾向（如标蓝色部分），同时各模型展现出更强的修辞控制能力：ChatGLM通过“火焰-溪流-星空”构建多感官意象，Baichuan的“朝阳-花朵-江河”形成时间隐喻，Qwen则实现“清晨-午后-夜晚”的完整时间链。这验证了优质示例引导对模型修辞生成能力的关键作用，同时也揭示了关键词CoT能有效引导模型捕捉排比句的深层语义结构，显著提升排比句的流畅性和多样性。此外，我们还发现不同模型架构在语义理解和创造性表达方面的特性差异，即ChatGLM长于意象创新，Baichuan擅长抽象概念具象化，而Qwen则展现出更强的时间维度建模能力。

	Vanilla	**CoT**	**K-CoT**
ChatGLM	青春在奋斗中闪光； 生命在奉献中升华； 人生在拼搏中精彩。	青春是穿透云层的朝阳； 青春是雨过天晴的彩虹； 青春是永不熄灭的火焰。	青春如火焰般炽热奔放； 青春似溪流般清澈跃动； 青春若星空般浩瀚无垠。
Baichuan	青春充满无限可能； 青春洋溢蓬勃朝气； 青春绽放绚丽光彩。	青春是探索未知的勇气； 青春是蜕变成长的坚韧； 青春是超越自我的豪情。	青春是初升的朝阳，燃烧着激情与活力； 青春是盛放的花朵，绽放着梦想与希望； 青春是奔腾的江河，激荡着勇敢与拼搏。
Qwen	青春是初升的朝阳，照亮我们前行的道路； 青春是盛放的花朵，绽放我们无尽的活力； 青春是奔腾的江河，滋养我们不断前进的精神。	青春是一场旅行，充满探索与发现； 青春是一次飞翔，追求自由与梦想； 青春是一首诗篇，书写激情与活力。	青春是清晨的朝阳，温暖而充满希望； 青春是午后的清风，舒适而充满生机； 青春是夜晚的星空，璀璨而充满活力。
Yi	青春是一场旅行，沿途风景书写回忆； 青春是一次飞翔，翱翔天际追逐梦想； 青春是一首诗篇，吟咏生命的热烈篇章。	青春如盛夏的暴雨，激荡着热情与奔放； 青春如秋日的金黄，收获着梦想与收获； 青春如冬日的暖阳，温暖着心灵与希望。	青春如朝阳初升，充满希望的光芒； 青春如花朵绽放，洋溢活力的芬芳； 青春如风帆扬起，追逐梦想的方向。
Llama	青春是梦想的起点，点亮未来无限的可能； 青春是激情的火种，燃烧心中不屈的热情； 青春是勇气的风帆，航向人生的无限海洋。	青春如初升的朝阳，温暖每一个清晨； 青春如盛放的花朵，绚烂每一片草原； 青春如奔腾的江河，自由每一段旅途。	青春是清晨的露珠，闪耀着希望的光芒； 青春是初升的太阳，散发着无限的活力； 青春是翱翔的雄鹰，承载着梦想的力量。

图 2: 输入青春-希望与活力（主题-基调）对，不同模型生成的文本实例