C3LRSO: A Chinese Corpus for Complex Logical Reasoning in Sentence Ordering

Xiaotao Guo, Jiang Li, Xiangdong Su[*], Fujun Zhang

College of Computer Science, Inner Mongolia University, Hohhot, China
National & Local Joint Engineering Research Center of Intelligent Information
Processing Technology for Mongolian, Hohhot, China
Inner Mongolia Key Laboratory of Multilingual
Artificial Intelligence Technology, Hohhot, China
32209060@mail.imu.edu.cn, cssxd@imu.edu.cn

Abstract

Sentence ordering is the task of rearranging a set of unordered sentences into a coherent and logically consistent sequence. Recent work has primarily used pre-trained language models, achieving significant success in the task. However, existing sentence ordering corpora are predominantly in English, and comprehensive benchmark datasets for non-English languages are unavailable. Meanwhile, current datasets often insert specific markers into paragraphs, inadvertently making the logical sequence between sentences more apparent and reducing the models' ability to handle genuinely unordered sentences in real applications. To address these limitations, we develop C3LRSO, a high-quality Chinese sentence ordering dataset that overcomes the aforementioned shortcomings by providing genuinely unordered sentences without artificial segmentation cues. Furthermore, given the outstanding performance of large language models on NLP tasks, we evaluate these models on our dataset for this task. Additionally, we propose a simple yet effective parameter-free approach that outperforms existing methods on this task. Experiments demonstrate the challenging nature of the dataset and the strong performance of our proposed method. These findings highlight the potential for further research in sentence ordering and the development of more robust language models. Our dataset is freely available at https://github.com/JasonGuo1/C3LRSO.

1 Introduction

Semantic coherence is necessary for text readability, accurate semantic expression, and effective information transmission. The sentence ordering task (Barzilay and Lapata, 2008) aims to reconstruct a coherent paragraph from a set of unordered sentences and has been shown to improve coherence in various NLP tasks, including multi-document summarization (Barzilay and Elhadad, 2002; Nallapati et al., 2017), conversational analysis (Zeng et al., 2018) and text generation (Konstas and Lapata, 2013; Holtzman et al., 2018). This task is crucial for assessing the machine's understanding of causal and temporal relationships. To achieve this goal, it is essential to ensure that the semantic and logical relationships between sentences are accurate, as their correctness significantly impacts the readability of the ordering results.

Early research on sentence ordering primarily employed probabilistic transition models and rule-based models, such as Hidden Markov Models (HMM), entity and content models. These methods often relied on manually crafted features (Lapata, 2003; Barzilay and Lee, 2004; Barzilay and Lapata, 2008). Despite their complexity, these designs required significant manual effort and time, necessitated expert knowledge, and were challenging to generalize to other contexts. In recent years, inspired by the advent of pre-trained language models in deep learning, various approaches based on pre-trained language models have been proposed, achieving remarkable success in the sentence ordering task(Cui et al., 2020; Chowdhury et al., 2021; Jia et al., 2023).

Despite progress in this task, the lack of appropriate language resources has hindered further research momentum. Currently, existing sentence ordering datasets are primarily in English (Chen et al., 2016; Yin et al., 2019; Wang and Wan, 2019), and there is no comprehensive benchmark dataset for Chinese. Additionally, current sentence ordering datasets are often created by inserting specific markers into existing paragraphs to segment them, making the logical order between sentences too apparent. This approach may cause models to learn only superficial logical relationships, which could weaken their ability to handle genuinely unordered sentences in real-world applications.

To address these challenges, we develop

[*] Corresponding Author

Proceedings of the 31st International Conference on Computational Linguistics, pages 4085–4095
January 19–24, 2025. ©2025 Association for Computational Linguistics

C3LRSO, a high-quality dataset specifically designed for Chinese sentence ordering tasks. The corpus is derived from real-world civil service exams, where candidates are required to reorder a set of unordered sentences into coherent and logical paragraphs by analyzing their logical relationships and content coherence. The dataset covers a wide variety of topics, such as culture, science, society, and the economy. These contents are carefully crafted to reflect the complexity and depth of the examination, ensuring that they effectively test the model's abilities in logical reasoning and language expression. Unlike previous sentence ordering datasets, our proposed corpus provides a higher level of challenge as it simulates real logical reasoning and language processing scenarios.

Furthermore, we conduct benchmark evaluation experiments on C3LRSO to assess the sentence logical ordering capabilities of several large pre-trained language models, validating both the dataset's quality and the models' performance. In addition to traditional benchmarks, our research delves deeper into how well large language models (LLMs) perform on the sentence ordering task. Recent studies have highlighted LLMs' remarkable proficiency in various reasoning tasks (OpenAI, 2023; Wei et al., 2022), particularly their ability to process long contexts and handle complex reasoning.

However, it remains uncertain how these models would fare in sentence ordering tasks without any supervised training. To explore this, we evaluate their performance on the challenging C3LRSO dataset, which was specifically designed to test logical coherence and sentence reordering. Additionally, we incorporate retrieval-augmented generation techniques, combining Chain-of-Thought (CoT) prompting (Wei et al., 2022) with Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), to enhance their reasoning performance.

To summarize, our contributions include:

- We introduce C3LRSO, the first comprehensive Chinese dataset for sentence ordering. We conduct experiments on this dataset using traditional methods, including BERSON(Cui et al., 2020), RE-BART(Basu Roy Chowdhury et al., 2021).
- To the best of our knowledge, we for the first time evaluate the performance of sentence ordering task using LLMs in zero-shot settings.
- We propose a novel approach that combines Retrieval-Augmented Generation (RAG) with the Chain-of-Thought (CoT) methodology for the task, aiming to mitigate the limitations of large language models (LLMs) during inference. The result outperforms those of direct prompting LLMs and traditional models.
- The experimental results and analysis underscore the challenging nature of our corpus, demonstrate the potential of large language models (LLMs) for sentence ordering, and reveal the limitations of current methods.

2 Related Work

2.1 Sentence Ordering

The sentence ordering task involves taking a set of potentially disordered sentences $S = \{s_1, s_2, \ldots, s_n\}$ as input. The goal is to determine the optimal order $O^* = \{o_1, o_2, \ldots, o_n\}$ that maximizes the global coherence of the reordered sentence sequence $S_{O^*} = \{s_{o_1}, s_{o_2}, \ldots, s_{o_n}\}$.

Early sentence ordering research attempted to use probabilistic transition methods based on linguistic features (Lapata, 2003), content models (Barzilay and Lee, 2004), and entity-based methods (Barzilay and Lapata, 2008; Prabhumoye et al., 2020). In recent years, neural network models have been applied to the sentence ordering task, adopting generative or ranking structures (Chen et al., 2016; Gong et al., 2016; Kumar et al., 2020; Chowdhury et al., 2021; Zhu et al., 2021a,b). Generative models treat sentence ordering as a sequence prediction problem, exploring the relationship between sentences and their positions to generate coherent sentence sequences from a set of input sentences. In the early stages, some encoder-decoder-based approaches (Logeswaran et al., 2018) treated unordered input sentences as permutation sequences and encoded them sequentially. This sequential modeling approach encodes erroneous sentence order and sentence-level semantic logic, potentially leading to the generation of incoherent paragraphs by the decoder. Specifically, when using sequential modeling, different permutations of the same paragraph may produce different paragraph representations, resulting in different output sentence orders, which is unreasonable and counter intuitive to human perception. To address this issue, Cui et al. (2018) first proposed a deep attention-based sentence ordering network that combines self-attention mechanisms to learn reliable and consistent paragraph representations. Wang and Wan

Datasets	Lg.	Content Type	Content Domain	Length Statistics		Dataset Split		
				Avg	Max	Train	Val	Test
NIPS Abstract	En	abstracts from conference papers	single	6	15	2427	408	377
AAN Abstract	En	abstracts from research community	single	5	20	8569	962	2626
NSF Abstract	En	abstracts from research awards	single	8.9	40	96070	10185	21580
arXiv Abstract	En	abstracts of arxiv website	single	5.38	35	884912	110614	110615
SIND	En	visual narrative	multiple	5	5	40155	4990	5055
ROCStory	En	short stories	multiple	5	5	78529	9816	9817
Accidents	En	narratives from accident database	multiple	11.5	19	100	-	100
Earthquakes	En	earthquakes articles	single	10.4	32	100	-	99
C3LRSO (Ours)	Zh	Chinese civil servant examination questions	multiple	19.53	82	976	100	100

Table 1: Comparison between our dataset and other datasets. Lg. denotes language: En for English and Zh for Chinese.

(2019) proposed a novel hierarchical attention network that captures word clues and dependencies between sentences.

Recent works are mostly based on pre-trained language models. Cui et al. (2020) proposed BERSON, which combines a BERT enhanced hierarchical relational sentence encoder and a self-attention based paragraph encoder. This architecture helps capture global dependencies between sentences while also encoding each individual sentence. Yin et al. (2019) used a sentence-entity graph to represent paragraphs. Also, a graph recurrent network is employed to recursively perform semantic transformations between connected nodes, enabling the model to effectively handle long-distance dependencies and aggregate semantic information at the paragraph level. Prabhumoye et al. (2020) used topological sort to find the correct order of the sentences in a document. It utilizes a classifier to predict relative order constraints between sentences and enhances this process with a BERT model to improve document coherence. Basu Roy Chowdhury et al. (2021) utilized BART to formalize the task as a conditional text-to-marker generation problem. Lai et al. (2021) proposed an iterative pair-wise ranking prediction framework, which enhances graph representation by iteratively predicting the order between pairs of sentences, thereby facilitating the generation of well-ordered sentences. In addition, Bin et al. (2023) proposed NAON, which employs a non-autoregressive decoder, thereby addressing the limitations of traditional autoregressive methods by leveraging bilateral dependencies and enabling parallel sentence prediction. Jia et al. (2023) enhanced pair-wise ranking-based and sequence generation-based methods by plugging a coherence verifier. This approach does not alter the parameters of the baseline model but instead verifies the coherence of the candidate rankings generated by the baseline model and re-ranks by beam search.

2.2 Sentence Ordering Datasets

We have meticulously reviewed the existing sentence ordering datasets, as presented in Table 1. Most previous sentence ordering corpora can be grouped into two categories by content domain. Prior studies (Yin et al.; Chen et al., 2016; Logeswaran et al., 2018) are primarily centered on abstracts limited to a single content domain (scientific papers). Other works based on narratives and stories (Huang et al., 2016; Wang and Wan, 2019) encompass multiple domains.

We observed that most of them are predominantly in English and relevant non-English datasets are scarce. Chinese sentence ordering datasets are unavailable, particularly multi-domain ones. Besides, existing sentence ordering datasets are dominantly written language instead of expert-designed exam questions. Thus, our proposed C3LRSO from real examination could be a valuable supplement for sentence ordering tasks and language reasoning tasks.

3 C3LRSO Dataset

3.1 Data Source

Having reviewed existing corpora, we determine to construct a new dataset that adapts to the requirements. To this end, we introduce C3LRSO (**C**hinese **C**orpus for **C**omplex **L**ogical **R**easoning in **S**entence **O**rdering), a corpus that features complex sentence ordering in Chinese. The dataset is sourced from real-world Chinese civil servant examination questions for several reasons: (1) These questions cover a wide range of domains, includ-

Example
①这其中一脉相承地贯穿了中国传统山水文化的精神和理念，体现了天人合一的历史文化的延续性。 This consistently carries the spirit and philosophy of traditional Chinese landscape culture, reflecting the continuity of the historical and cultural concept of harmony between human and nature.
②主要原因在于历史上的杭州人将传统山水文化的理念和西湖的治理融合在一起，并将这种融合延续下来。两者缺一不可。 The main reason lies in the historical fact that the people of Hangzhou integrated the philosophy of traditional landscape culture with the governance of West Lake and have continued this integration. Both elements are indispensable.
③中国拥有湖泊的城市很多，但为什么城市发展与景观和谐并存的鲜而有之，而尤以杭州与西湖这一例凸显了出来? China has many cities with lakes, but why is it so rare to see urban development coexisting harmoniously with the landscape, with Hangzhou and West Lake standing out as a prime example?
④这种天人合一的延续性，是中国其他城市普遍缺失的。 This continuity of harmony between human and nature is something that is generally lacking in other Chinese cities.
⑤我们翻阅西湖的历史，那就是一部保护与治理的历史，就是城市建设与景观建设相辅相成的历史。 When we look back at the history of West Lake, it is a history of protection and governance, a history where urban development and landscape construction have complemented each other.
The correct order of the five sentences is: Answer： ⑤①④③②

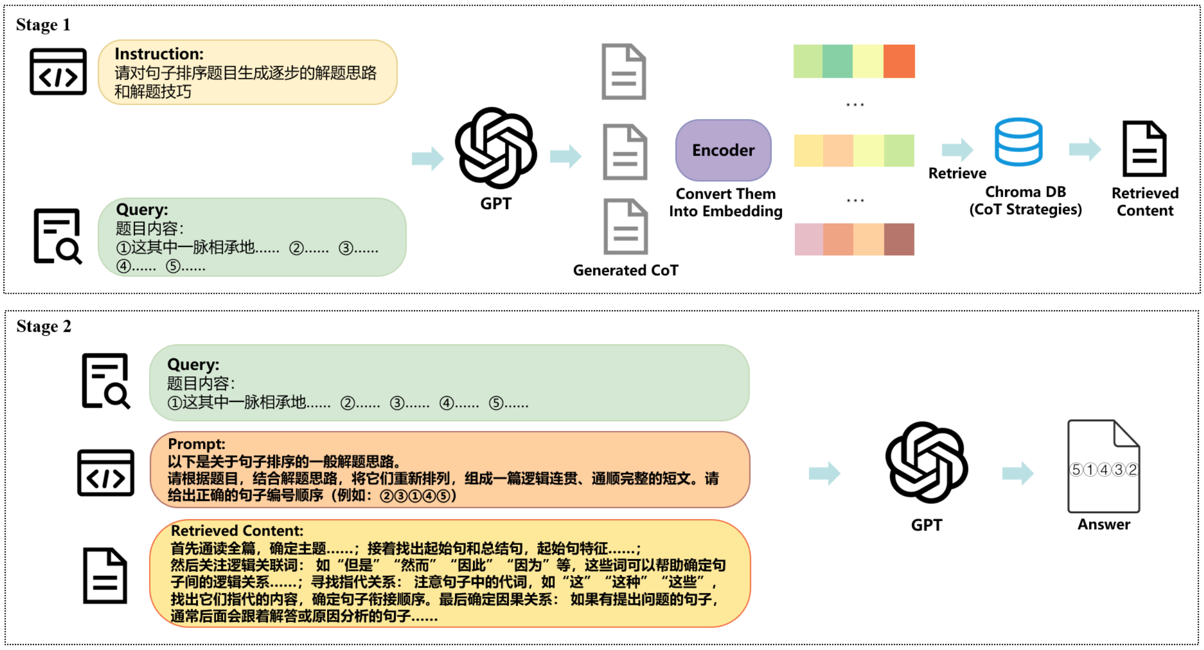
Table 2: A simplified example of C3LRSO. In the first five pairs of sentences, the former is the original Chinese question text, and the latter is its English machine translation output. The last line shows the correct order of these five sentences.

ing but not limited to economics, law, and culture, making them ideal for evaluating the model's ability to order sentences across different fields; (2) Civil service exam questions are known for their intricate text structures and strict logic, which are essential for testing the model's capability in producing coherent and logically structured text; (3) The sentences in these questions require not only an understanding of individual sentence meanings but also recognizing the interrelationships between sentences and the overall coherence of the text. Such attributes ensure that this dataset provides a high standard for assessing the model's performance in text generation and logical reasoning. Ultimately, using this challenging corpus offers a more precise assessment of a model's strengths and weaknesses in managing complex text.

The construction of the dataset follows a three-stage process: data collection, format verification, and quality control. Initially, we gathered sentence ordering questions from publicly available Chinese civil service exam resources. Next, Python scripts were used to clean the data, eliminating any issues with incorrect Chinese characters, incomplete content, or wrong answers. This was followed by manual corrections. Lastly, the dataset was converted into a structured JSON format, ensuring the quality of questions.

3.2 Data Collection

A pilot screening on publicly accessible data reveals that numerous websites share exam questions, so we alternatively utilize material from authoritative websites as our primary sources. We developed Python scripts to automate the extraction of the questions and their corresponding correct answers from the raw web pages. These scripts also performed preliminary cleaning tasks, such as removing any HTML tags, normalizing the encoding to UTF-8, and converting non-standard numbering formats to standardized circled numbers (e.g., converting '1.', '(1)' to '①'). The data collection process was conducted over a one-month period, employing web scraping tools to automate the extraction of questions from selected sources. We removed a few questions that contained sensitive content and corrected the erroneous answers to two questions after comparing them with multiple authoritative sources. We collected a total of 1,203 questions, all in Mandarin Chinese. After a careful filtering process (see section 3.3), we finally obtained 1,176 unordered paragraphs (questions) across 8 content domains: Economy, Entertainment, Culture, Law, Health, Science, Education,



Instruction:
请对句子排序题目生成逐步的解题思路和解题技巧

Query:
题目内容:
①这其中一脉相承地...... ②...... ③......
④...... ⑤......

Query:
题目内容:
①这其中一脉相承地...... ②...... ③...... ④...... ⑤......

Prompt:
以下是关于句子排序的一般解题思路。
请根据题目，结合解题思路，将它们重新排列，组成一篇逻辑连贯、通顺完整的短文。请给出正确的句子编号顺序（例如：②③①④⑤）

Retrieved Content:
首先通读全篇，确定主题......；接着找出起始句和总结句，起始句特征......；然后关注逻辑关联词：如“但是”“然而”“因此”“因为”等，这些词可以帮助确定句子间的逻辑关系......；寻找指代关系：注意句子中的代词，如“这”“这种”“这些”，找出它们指代的内容，确定句子衔接顺序。最后确定因果关系：如果有提出问题的句子，通常后面会跟着解答或原因分析的句子......

Figure 1: Illustration of our pipeline.

and Philosophy. Each question consists of a few sentences in random order (ranging from 4 to 7 typically). Each sentence is assigned a sequential number for ordering purposes. It is worth mentioning that since these questions have been designed by experts and tested by candidates, they are already very scientific, comprehensive, and challenging, requiring no extra automated or manual annotation. Nevertheless, we manually inspected each question to ensure its quality, confirming that the content of the questions complies with standards, contains no typographical errors, and has correct answers.

3.3 Quality Control

To ensure the quality of the dataset, we filtered out questions according to the following criteria: (1) We drop questions that lack complete sentence content; (2) Any question without sequence number is discarded because the sentences need to be distinguished by numbers; For non-standard numbers, we standardize them to corresponding circled numbers, e.g. ①,②; (3) Content that is unreasonable or not based on actual questions is also neglected. This process also includes removing duplicate questions, excluding incomplete sentence sets. Moreover, questions containing politically sensitive topics, personal information, or culturally insensitive material were excluded to maintain ethical standards and ensure the dataset's suitability for diverse applications.

A team of native Chinese teachers with expertise in language processing conducted manual reviews. They meticulously checked each question for grammatical accuracy, logical coherence, and relevance to ensure high-quality standards. Subsequently, we standardized the dataset format, utilizing JSON file format to facilitate readability and usability. Eventually, we obtained a fully standardized dataset, as illustrated in Table 2.

3.4 Dataset Statistics

The comparison of multiple corpora is presented in Table 1. As for the data split, there are 976 questions designated for training, 100 for validation, and 100 for testing. Length statistics represent the average and maximum number of words per clause. The statistics are based on English words for all corpora in the table except for our work. Since our dataset consists of Chinese text, we need to count the number of Chinese words. Hence, we used the PKUSEG package (Luo et al., 2019) to tokenize the clauses and calculate the mean and max word counts. It is worth noting that the clauses in our dataset contain a wider variety of Chinese words and cover a wide range of topics, which is more conducive to evaluating the model's performance in the sentence ordering task.

4 The Proposed Method

With the collected dataset, we propose an LLM-based method to perform sentence ordering on C3LRSO. In this setting, we design this method to rely solely on LLMs to find the answer, and it performs better than direct prompting. Our approach is based on Retrieval-Augmented Generation (RAG) (Gao et al., 2022), which can conduct precise dense retrieval. We first apply it to sentence ordering. More specifically, we first apply the Chain-of-Thought (CoT) (Wei et al., 2022) to break down common solutions solving the sentence ordering problem, followed by the application of our Retrieval-Augmented Generation (RAG) pipeline.

The complete process is illustrated in Figure 1. First, we input general problem-solving techniques for sentence ordering into a large language model (LLM) to generate a Chain-of-Thought (CoT) explanation, which is then converted into vectors and stored in a Chroma vector database. Next, for each problem, we use the HyDE method (Gao et al., 2022) to prompt the LLM to generate a problem-solving approach. This approach is utilized to retrieve CoT content from the vector database. Finally, the model combines the problem with this Chain-of-Thought as input, performs reasoning, and generates the final numbered answer.

This method enhances the model's reasoning ability because only by producing more reasonable reasoning sentences can the LLM retrieve more useful explanatory information. It alleviates the hallucination problem often encountered when directly asking the model to output answers. The additional CoT explanations provide clearer assistance for the model's reasoning, leading to more logical ordering and ensuring that the model's thinking is more aligned with the original problem. This approach also tests the summarization and inductive capabilities of the LLM used for generating explanations.

The utilization of CoT explanations helps structure the reasoning process, making it easier for the model to follow logical steps and generate coherent answers. The method ensures that the model's reasoning process stays closely aligned with the original question, enhancing the relevance and correctness of the answers.

As previously mentioned, our pipeline consists of two stages. We have designed corresponding prompt templates for each stage to achieve the experimental objectives. Specifically, illustrated in Table 3 and Table 6. Additionally, we also provide a template for direct prompting method with Chinese prompts and their English translations, as presented in Table 4.

5 Experiments

5.1 Experimental Settings

Models In our experiment, we first examine the performance of BERSON (Cui et al., 2020) and RE-BART (Basu Roy Chowdhury et al., 2021) on our dataset. Second, we evaluate the effectiveness of LLMs in direct prompting on the sentence ordering task without any fine-tuning. To accomplish this, we evaluate ChatGPT (GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o) from OpenAI (OpenAI, 2023), ERNIE-3.5-8K from the ERNIE model family, GLM4 from the GLM model family, as well as iFLYTEK Spark3.5 from the Spark model family.

Overall, experiments are conducted using traditional methods, direct prompting, and RAG-based prompting. Following that, we apply the aforementioned method to enhance the performance of LLMs.

Metrics To evaluate the sentence ordering task, we use PMR (Perfect Match Ratio), ACC (Accuracy), and Kendall's τ as the metrics. PMR is used to calculate the accuracy of exact matches at the paragraph level. It measures whether the entire predicted sequence is completely consistent with the true (gold standard) sequence. It is one of the strictest evaluation metrics for calculating the accuracy of exact matches at the paragraph level. ACC calculates the accuracy of absolute position prediction at the sentence level, focusing on whether each sentence is placed in its correct position. Kendall's τ measures the relative order of all sentence pairs in the predicted paragraph, considering the relative positional relationship of all sentence pairs in the predicted sequence.

5.2 Results

Table 5 shows the performance of traditional models, 6 LLMs, and LLMs using our method on our proposed dataset. The labels in our dataset are correctly ordered numerical sequences. Overall, both BERSON and RE-BART perform decently in the standard setting, surpassing direct prompting LLMs on most metrics. In our experiments, the BERSON method utilizes the bert-base-chinese pre-trained model. We set the batch size to 4 and the learning rate to 5e-5, training the model for a total of 20 epochs. For the RE-BART method, we

RAG Prompts Template Stage 1 (中文)	**RAG Prompts Template Stage 1 (English)**
请对句子排序题目生成通用的逐步的解题思路和解题技巧：	Please create a step-by-step common reasoning approach and techniques for solving the sentence ordering question:
题目内容：（具体题目）	Question Content: (Specific question)
①这其中一脉相承地......	①Among these, there is a continuous lineage that...
②主要原因在于历史上的杭州人......	②The main reason lies in the historical residents of Hangzhou...
③中国拥有湖泊的城市很多......	③There are many cities in China that have lakes...
④这种天人合一的延续性......	④This continuity of the harmony between nature and humanity...
⑤我们翻阅西湖的历史......	⑤When we leaf through the history of West Lake...
本题解题思路：首先......	Solution approach for this question: First, ...

Table 3: An example of the Stage 1 RAG prompts template, featuring Chinese prompts and their English translations, where the content inside parentheses is intended to be filled in.

Direct Prompts Template (中文)
请根据以下无序的句子，将它们重新排列，组成一篇逻辑连贯、通顺完整的短文。请给出正确的句子编号顺序（例如：②③①④）。 题目：（题目内容） 回答：
Direct Prompts Template (English)
Please rearrange the following unordered sentences to form a logically coherent and smoothly flowing short paragraph. Please provide the correct sentence number sequence (e.g., ②③①④). Question: (Content) Answer:

Table 4: Direct prompts template example with Chinese prompts and their English translation. The content inside parentheses is intended to be filled in.

Method	**ACC ↑**	**PMR ↑**	τ ↑
Traditional Methods			
BERSON	0.494	0.28	0.567
RE-BART	0.343	0.16	0.237
Direct Prompting			
GLM4	0.220	0.02	0.130
GPT-3.5 Turbo	0.255	0.03	0.031
GPT-4 Turbo	0.386	0.12	0.187
GPT-4o	0.235	0.03	0.159
iFLYTEK Spark3.5	0.368	0.10	0.111
ERNIE-3.5-8K	0.412	0.13	0.238
w/ RAG			
GLM4+	0.524	0.32	0.367
GPT-3.5 Turbo+	0.363	0.10	0.170
GPT-4 Turbo+	**0.570**	**0.36**	**0.571**
GPT-4o+	0.426	0.25	0.408
iFLYTEK Spark3.5+	0.553	0.32	0.329
ERNIE-3.5-8K+	0.408	0.14	0.242

Table 5: Experimental Results. The best metrics are highlighted in bold. The reported results are based on the versions of the APIs and the C3LRSO corpus as of May 15, 2024.

employ the bart-large-chinese pre-trained model, maintaining the same batch size of 4 and learning rate of 5e-5, with training conducted over 20 epochs. All experiments were performed using a single Tesla V100 GPU. All APIs are called using official links. Finally, BERSON outperformed all direct prompting LLMs on all three metrics. RE-BART outperformed the majority of direct prompting methods in PMR and τ metrics and surpassed GLM4, GPT-3.5-turbo, and GPT-4 on ACC. Based on the results, we found that traditional models still have a significant advantage over direct prompting large language models in the sentence ordering task.

Comparing RAG models and direct prompting LLMs, we noticed that our method significantly improved the performance of the corresponding models. Among them, GPT-4-Turbo performed the best, achieving the state-of-the-art performance in sentence ordering on our dataset. This indicates that our method significantly enhances the reasoning ability and performance in sentence ordering tasks for large language models.

6 Conclusion

This paper addresses two key, under-explored challenges in sentence ordering datasets. First, existing datasets often insert markers to segment paragraphs, making sentence relationships overly explicit and limiting models' ability to handle genuinely unordered sentences in real-world scenarios. Second, there is a notable lack of non-English sentence ordering datasets, particularly for Chinese. To tackle these issues, we created C3LRSO, a novel Chinese sentence ordering dataset, and proposed a method based on Retrieval-Augmented Generation and Chain-of-Thought reasoning. In addition, we conducted a comprehensive evaluation of both large language models and traditional models, providing detailed analyses. For future work, we aim

RAG Prompts Template Stage 2 (中文)	RAG Prompts Template Stage 2 (English)
以下是一些关于句子排序的一般解题思路，可供参考：	Below are some general strategies for sentence ordering that you may refer to:
（检索到的思维链内容）	(Retrieved chain-of-thought content)
请根据题目，结合上述解题思路，	Please, based on the question and the above strategies,
将它们重新排列，组成一篇逻辑连贯、通顺完整的短文。	rearrange them to form a logically coherent and smoothly flowing short paragraph.
请给出正确的句子编号顺序（例如：②③①④）	Please provide the correct sentence number sequence (e.g., ②③①④)
题目：（题目内容）	Question: (Content)
回答：	Answer:

Table 6: An example of the Stage 2 RAG prompts template, featuring Chinese prompts and their English translations, where the content inside parentheses is intended to be filled in.

to further augment the dataset and explore architectures better suited for the sentence ordering task.

Limitations

We plan to expand the dataset in future. Additionally, in all zero-shot and few-shot experiments, we utilized identical prompts across all models. However, prompt selection is of significant importance for large language models. We plan to utilize a greater variety of prompts and prompt engineering techniques in future experiments on carefully curated datasets.

Ethics Considerations

All data used in this study are publicly available and were obtained from open source platforms. We have adhered to all relevant policies and guidelines to ensure that our use of these data does not infringe upon any copyright or intellectual property rights. Our corpus developed in this study is intended solely for academic research purposes. The questions in the dataset inherently have correct answers. The manual review and human evaluation were carried out by members of our research group in collaboration.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), National Education Science Planning Project (Grant No. BIX230343), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2022YFHH0077), The Central Government Fund for Promoting Local Scientific and Technological Development (Grant No. 2022ZY0198), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Inner Mongolia Autonomous Region Science and Technology Planning Project (Grant No. 2023YFSH0017), Joint Fund of Scientific Research for the Public Hospitals of Inner Mongolia Academy of Medical Sciences (Grant No.2023GLLH0035).

References

Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi Bin, Wenhao Shi, Bin Ji, Jipeng Zhang, Yujuan Ding, and Yang Yang. 2023. Non-autoregressive sentence ordering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4198–4214, Singapore. Association for Computational Linguistics.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order?

a simple way to order sentences. *arXiv preprint arXiv:2104.07064*.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349.

Baiyun Cui, Yingming Li, and Zhongfei Zhang. 2020. Bert-enhanced relational sentence ordering network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649, Melbourne, Australia. Association for Computational Linguistics.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Sainan Jia, Wei Song, Jiefu Gong, Shijin Wang, and Ting Liu. 2023. Sentence ordering with a coherence verifier. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9301–9314.

Ioannis Konstas and Mirella Lapata. 2013. A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, 48:305–346.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8115–8122.

Shaopeng Lai, Ante Wang, Fandong Meng, Jie Zhou, Yubin Ge, Jiali Zeng, Junfeng Yao, Degen Huang, and Jinsong Su. 2021. Improving graph-based sentence ordering with iteratively predicted pairwise orderings. *arXiv preprint arXiv:2110.06446*.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *arXiv preprint arXiv:1906.11455*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

OpenAI. 2023. Introducing chatgpt. Accessed: 2024-05-30.

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020. Topological sort for sentence ordering. *arXiv preprint arXiv:2005.00432*.

Tianming Wang and Xiaojun Wan. 2019. Hierarchical attention networks for sentence ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7184–7191.

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.

Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. Graph-based neural sentence ordering.

Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. 2019. Graph-based neural sentence ordering. *arXiv preprint arXiv:1912.07225*.

Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018. Microblog conversation recommendation via joint modeling of topics and discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 375–385.

Yutao Zhu, Jian-Yun Nie, Kun Zhou, Shengchao Liu, Yabo Ling, and Pan Du. 2021a. Bert4so: Neural sentence ordering by fine-tuning bert. *arXiv preprint arXiv:2103.13584*.

Yutao Zhu, Kun Zhou, Jian-Yun Nie, Shengchao Liu, and Zhicheng Dou. 2021b. Neural sentence ordering based on constraint graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14656–14664.

A Dataset Construction Details

We selected real-world Chinese civil service examination questions as our primary source for the following reasons: (1) Covering various fields such as economics, law, and culture, these questions allow for testing the model's ability to order sentences across different domains; (2) Civil service exam questions typically feature more complex text structures and rigorous logic, making them ideal for testing whether models better understand and generate structured and logical texts; (3) Such sentences require the model to not only understand the meaning of individual sentences but also to grasp the relationships between them and the overall coherence of the text. This is crucial for enhancing the model's capabilities in text understanding and generation. By using such challenging and high-standard corpus to test and evaluate NLP models, it is possible to more accurately identify the models' strengths and weaknesses in handling complex texts, thereby facilitating progress in this field.

The construction of our dataset consists of three major stages: data collection, format verification, and quality control. We first obtained sentence ordering questions and answers from authoritative open-source websites sharing Chinese civil service exam questions. Next, we employed Python scripts to check the raw material to ensure there were no incorrect Chinese characters, incomplete content, or incorrect answers, followed by manual correction. Finally, we extracted and converted them into JSON format. We also implemented several quality control measures during the process to ensure the dataset's quality and reliability.

To construct a high-quality dataset suitable for complex sentence ordering tasks in Chinese, we initiated a comprehensive data collection process. Initially, we conducted a pilot study to survey publicly accessible data sources. We found that numerous websites provide collections of Chinese civil service examination questions, particularly focusing on the language and comprehension sections, which often include sentence ordering tasks.

To ensure the reliability and authority of our dataset, we decided to source materials from reputable and authoritative websites, such as official educational platforms and well-recognized online education providers. We compiled a preliminary collection of 1,203 Mandarin Chinese sentence ordering questions.

Each question in our dataset comprises a set of sentences presented in a random order. The number of sentences per question varies, typically ranging from four to seven. Each sentence is prefixed with a sequentially increasing number (e.g., ①, ②, ③), which serves as an identifier for ordering purposes.

These examination questions cover a diverse range of domains, including Economics, Entertainment, Culture, Law, Health, Science, Education, Philosophy, and more. This diversity ensures that our dataset can evaluate models' abilities across various topics and contexts.

Given that these questions are originally designed by experts in the field and have been utilized in actual civil service examinations, they inherently possess a high level of complexity and logical rigor. Therefore, they are well-suited for testing models' capabilities in understanding and generating coherent and logically structured texts without the need for additional manual annotation or modification.

Following the initial data collection, we proceeded to preprocess the data. We developed Python scripts to automate the extraction of the questions and their corresponding correct answers from the raw web pages. These scripts also performed preliminary cleaning tasks, such as removing any HTML tags, normalizing the encoding to UTF-8, and converting non-standard numbering formats to standardized circled numbers (e.g., converting ’1.’, ’(1)’, or ’一、’ to ’①’).

To ensure the high quality and usability of the dataset, we implemented a multi-stage quality control process involving both automated and manual verification steps.

Firstly, we applied automated scripts to check for common data issues. The scripts scanned the dataset to identify questions with missing sentences, incomplete content, or incorrect Chinese characters resulting from encoding errors. They also verified that each sentence within a question was properly numbered in sequential order.

Secondly, we performed manual reviews to catch any issues that automated scripts might have missed. A team of native Chinese speakers with expertise in language processing carefully examined the dataset. They checked for grammatical correctness, logical coherence, and overall quality of the sentences.

We applied the following specific criteria for filtering: Completeness: Questions lacking complete sentence content were removed. This ensured that each question provided sufficient information for the sentence ordering task.

Proper Numbering: Any question without sequence numbers or with incorrect numbering was discarded. For questions with non-standard numbering, we standardized them to the corresponding circled numbers (e.g., ①, ②, ③).

Content Reasonableness: Content that was unreasonable, nonsensical, or not based on actual examination questions was excluded. This maintained the authenticity and relevance of the dataset.

Duplication Removal: Duplicate questions or sentences were identified and removed to prevent redundancy in the dataset.

Error Correction: For two questions, which included typographical and grammatical errors, manual corrections were made to ensure accuracy and maintain the original meaning.

After the quality control process, we refined the dataset to 1,176 high-quality unordered paragraphs suitable for training and evaluating models on complex sentence ordering tasks.

Additionally, we converted the finalized dataset into a standardized JSON format to facilitate easy integration with various machine learning frameworks. Each JSON entry includes the unordered sentences, their correct order, and metadata such as the domain category.

Our dataset covers a broad range of topics, which enhances its utility in evaluating models' performance in understanding and ordering sentences across different domains. The diverse content and complex structures of the sentences make the C3LRSO dataset a valuable resource for advancing research in natural language processing tasks related to text coherence and logical reasoning.