Rephrasing Invokes Better Generations for Large Language Models

Haoran Yang , Hongyuan Lu , Wai Lam
The Chinese University of Hong Kong
{hryang, hylu, wlam}@se.cuhk.edu.hk

Abstract

In the realm of emerging multitasking abilities of Large language models (LLMs), methodologies like prompt tuning enable low-cost adaptation to downstream tasks without retraining the model. However, automatic input pre-processing when LLMs are unavailable is currently under-studied. This paper proposes RELLM (**Re**phrasing for **LLM**s), a method that automatically paraphrases input content for better output generations. RELLM replaces low-frequency lexical items with their high-frequency counterparts. This substitution is particularly beneficial for low-resource language tasks that lack sufficient training data and resources. RELLM is user-friendly and requires no additional LLM training. Experimental results in cross-lingual summarization, and natural language inference demonstrate the effectiveness of RELLM.

1 Introduction

Large language models (LLMs, Ouyang et al. 2022; Yang et al. 2023) such as ChatGPT [1] and LLaMA-2 (Touvron et al., 2023) have exhibited their power in tackling various tasks by providing corresponding prompts. A prompt typically consists of two parts (Wang et al., 2023b): instruction that describes the nature of the task, and input that describes the specific context of the task. One of the keys to the success of eliciting the desired information from LLM is prompt tuning (Lester et al., 2021; Liu et al., 2022; Yang et al., 2022) which often involves experimenting with different prompt structures, wording, or formatting. Yet, this stream of works usually focus on the refinement of the instruction part (Wei et al., 2022; Zhou et al., 2023) and overlook the value of modifying input contents.

To this end, we propose a novel method called RELLM, which rephrases the input content to the same meaning while written in different expressions to improve the generation quality. Specifically, RELLM substitutes the low-frequency words in input with their high-frequency counterparts that represent the same meaning. Such a methodology is inspired by the fact that replacing low-frequency words in the pre-training procedure can improve language models (Bai et al., 2022; Wang et al., 2023a). By employing RELLM, we can derive benefits, especially for low-resource languages where access to ample training data is limited. However, this raises another question – how do we define low-frequency? Since the pre-training data for LLMs are usually not publicly released – like ChatGPT, it could be unusual to define the frequency. We surprisingly found that using word frequency statistics that are online available can empirically impressively work well for RELLM.

We conduct experiments on two different tasks, cross-lingual summarization (Narayan et al., 2018) and natural language inference (Bowman et al., 2015; Williams et al., 2018). We found that RELLM can invoke better generations on these tasks compared to unmodified inputs. For example, the gain for summarization is up to 2x BLEU-4 points (summarize texts written in English to Latgalian). Our contributions are three-fold :

- We propose RELLM as a novel method that replaces the low-frequency words with their high-frequency paraphrases for the input content into LLMs for better generation.
- We surprisingly found that using online available word statistics brings good improvements. We adopt this, as the training data for LLMs are frequently not open-resourced.
- We conduct experiments on cross-lingual summarization and natural language inference. The results illustrate the effectiveness of RELLM that invokes better generation for low-resource languages.

[1] https://openai.com/blog/chatgpt

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 8–15

2 Method

2.1 Intuition

LLMs have emerged as remarkable tools for tackling various tasks. Prompt tuning plays a crucial role in harnessing the full potential of these models. By fine-tuning the prompts, users can adjust instructions to tailor the model's output. This process significantly impacts the quality of the generations. Most of the prompt tuning methods including continuous prompt tuning (Li and Liang, 2021; Lester et al., 2021) and discrete prompt tuning (Deng et al., 2022; Wen et al., 2023) are heavily relying on the access to the weights of models to calculate gradients to optimize. However, since the weights of a lot of prevailing LLMs such as ChatGPT, Bard [2], are not available, tuning the prompts automatically is almost impossible.

To this end, we propose Rephrasing for LLMs (ReLLM). ReLLM rephrases the inputs without changing their original meaning. Specifically, ReLLM enhances the performance of language models by replacing low-frequency words with high-frequency words in prompts, more detailly, the input part of the prompts. The intuitions behind our proposed method are twofold. Firstly, in monolingual tasks, low-frequency words are less commonly encountered in training data and may pose challenges for LLMs to accurately generate coherent and relevant responses. Secondly, in many multilingual tasks such as cross-lingual summarization, aligning words between different languages is crucial. Low-frequency words in the source language might lack direct translation equivalents in the target language, making alignment challenging. By strategically substituting such words with high-frequency alternatives, we aim to provide the model with more robust and representative input, leading to improved performance. Moreover, this substituing process is totally performed by the LLM itself (e.g., ChatGPT). Therefore, the weights of the LLMs are not required to be accessible since no tuning stage is performed.

2.2 ReLLM

Formally, given the prompt $\mathbf{p} = (t, x)$ where t is the instruction related to the task (for example, for translation task, t may be "translate the following sentence from English to German: ") and x is the input of the task. Most of prompt tuning methods focus on adjusting t while ReLLM focuses on refining x. Instead of directly feeding $\mathbf{p}$ to the LLMs, we firstly rephrase x to $\hat{x}$ and then input $\hat{\mathbf{p}} = (t, \hat{x})$ to the LLMs. Since many tasks are sensitive to changes in sentence structure, rephrasing the whole sentence may have a negative effect on the performance. Therefore, we only replace the low-frequency words with their high-frequency counterparts.

However, one difficulty is that we are unable to count the frequency of words in the situation that the training corpus of LLMs is not publicly available. To solve this issue, we turn to exploit online available word frequency statistics to help replace low-frequency words in x. Specifically, we use the google-10000-english[3], containing 10000 English words ordered by frequency from high to low based on Google's Trillion Word Corpus, as the high-frequency word dictionary D_H. If a word x_i does not belong to D_H, we think this word is a low-frequency word that should be replaced by its high-frequency counterpart. Moreover, to avoid mistakenly replacing some special words, for example, names, locations, or numbers, we introduce another word dictionary D_L[4] which contains a large number of normal words. We only substitute the words that are not in D_H but in D_L.

After spotting the low-frequency words, we next need to determine their high-frequency counterpart. The challenge lies in keeping the meaning of the sentence unchanged after replacing the words. We utilize ChatGPT to accomplish this challenging task. Specifically, given an input x and a low-frequency word $x_i \in x$, we use the below prompt to obtain the desired output:

> Given a word x_i and a paragraph: x, find the word's synonym that has a higher frequency and does not change the meaning of the paragraph. The output format is a dictionary where the key is the word and the value is its synonym.

We only post-process the output with the format $\{x_i : \hat{x}_i\}$ by replacing x_i in x with $\hat{x}_i$.

3 Experiments

We choose ChatGPT as the LLM to complete the word substitution task due to its impressive performance across various tasks and domains. Specifically, we use **gpt-3.5-turbo**. This is a ChatGPT

[2] https://bard.google.com/?hl=en

[3] https://github.com/first20hours/google-10000-english

[4] https://github.com/dwyl/english-words/tree/master

	gpt-3.5-turbo				text-davinci-003			
	baseline		ReLLM		baseline		ReLLM	
Language	**ROUGE-L**	**BLEU-4**	**ROUGE-L**	**BLEU-4**	**ROUGE-L**	**BLEU-4**	**ROUGE-L**	**BLEU-4**
aeb_Arab	9.6	1.16	**9.7**	**1.37**	8.0	0.50	**8.2**	**0.77**
ast_Latn	17.8	**3.37**	**18.0**	3.24	16.7	2.84	**17.0**	**3.08**
ayr_Latn	6.2	0.60	**6.5**	**0.62**	4.8	**0.46**	**4.8**	0.45
ban_Latn	11.0	**1.37**	**11.0**	1.20	9.1	0.84	**9.1**	**1.02**
szl_Latn	9.3	1.08	**9.5**	**1.20**	8.0	0.61	**8.0**	**0.85**
bho_Deva	13.3	**1.30**	**13.3**	1.26	**9.2**	0.63	8.8	**0.65**
smo_Latn	19.8	**2.48**	**20.4**	2.43	18.8	1.92	**18.9**	**2.06**
lus_Latn	16.4	**2.37**	**16.4**	2.18	15.6	2.03	**15.9**	**2.09**
lij_Latn	10.3	0.93	**10.6**	**0.93**	11.4	0.97	**11.6**	**1.13**
lim_Latn	15.1	1.65	**15.4**	**1.90**	13.9	1.30	**14.0**	**1.38**
ltg_Latn	5.2	0.52	**5.4**	**1.05**	**4.4**	0.32	4.3	**0.42**
gla_Latn	17.1	2.03	**17.1**	**2.09**	14.8	**1.30**	**14.8**	1.16
fur_Latn	17.3	**2.58**	**17.3**	2.23	**17.2**	2.48	17.1	**2.53**

Table 1: Automatic Evaluation Results on Cross-Lingual Summarization.

	SNLI			MultiNLI		
Language	**baseline**	**ReLLM(v1)**	**ReLLM(v2)**	**baseline**	**ReLLM(v1)**	**ReLLM(v2)**
eng_Latn	**0.448**	0.422	0.420	**0.450**	0.414	0.436
aeb_Arab	0.282	0.288	**0.308**	**0.376**	0.362	0.368
bho_Deva	**0.314**	0.292	0.312	0.350	0.346	**0.370**
lij_Latn	0.284	0.292	**0.294**	0.394	**0.408**	0.388
lim_Latn	**0.302**	0.278	0.296	0.390	0.378	**0.410**
ltg_Latn	0.298	0.302	**0.304**	0.328	**0.336**	0.324
gla_Latn	0.282	0.292	**0.304**	0.330	**0.348**	0.338
fur_Latn	0.308	0.310	**0.324**	0.360	**0.408**	0.386
ace_Arab	0.298	0.294	**0.312**	0.330	0.322	**0.340**
ace_Latn	0.290	0.296	**0.304**	0.304	**0.320**	0.314
ydd_Hebr	0.298	**0.312**	0.302	0.314	0.312	**0.318**
bem_Latn	0.302	**0.312**	0.300	0.310	0.306	**0.328**
san_Deva	0.304	0.298	**0.316**	**0.368**	0.354	0.366
fur_Latn	0.308	0.310	**0.324**	0.360	**0.408**	0.386
pol_Latn	0.290	0.298	**0.312**	0.430	0.422	**0.436**

Table 2: Accuracy on SNLI and MultiNLI.

model accessed via the official API through Python. We conducted all word-replacing experiments during April and May.

We evaluate ReLLM on two different tasks: cross-lingual summarization and natural language inference. The first task is multilingual and we intend to demonstrate that it is easier for LLMs to align high-frequency words to the words in other languages. On the other hand, the natural language inference task focuses on examining the impact of low-frequency words within the same language.

3.1 Cross-Lingual Summarization

Setup We conduct experiments on XSum (Narayan et al., 2018), in which each document is summarized into one sentence, both written in English. To investigate whether high-frequency words are better aligned in other low-resource languages, we convert the monolingual summarization to cross-lingual summarization. Specifically, we preserve the original input text while translating the ground-truth targets into low-resource languages using NLLB [5]. To investigate the translation quality, we translate targets back into English using NLLB and calculate the similarity between the original targets and those translated back. The results are shown in Table 3 in Appx A. We found that the translation quality is generally good. We adopt gpt-3.5-turbo and text-davinci-003 to perform this task. The prompts without rephrasing are regarded as the baseline. We use BLEU-4 [6] and ROUGE-L [7] as the automatic evaluation metrics.

Prompt We use the following prompt to obtain the output from LLMs:

[5] https://huggingface.co/facebook/nllb-200-3.3B

[6] https://github.com/mjpost/sacrebleu

[7] https://github.com/pltrdy/rouge

```
The task is to summarize an article with only one sentence. Here are two examples: [example_1], [example_2]. Given the following article: [text], output its summary and translate it to [language].
```

Results The results on 13 low-resource languages[8] are reported in Table 1. It is evident that ReLLM generally outperforms the baseline in terms of ROUGE and BELU metrics for both gpt-3.5-turbo and text-davinci-003 models. Remarkable improvements are observed, such as a doubling of the score in Latgalian (ltg_Latn), where the BLEU score increases from 0.52 to 1.05. Additionally, it is worth noting that gpt-3.5-turbo exhibits superior performance compared to text-davinci-003 in the cross-lingual summarization task. We provide some cases in Appx. B.

3.2 Natural Language Inference

In contrast to previous experiments focusing on multilingual tasks, this particular section evaluates the performance of ReLLM specifically on monolingual natural language inference. The primary objective of this experiment is to demonstrate the effectiveness of ReLLM in languages that have not undergone adequate training in LLMs due to limited training data.

Setup We conduct experiments on two canonical natural language inference tasks: SNLI (Bowman et al., 2015) and its upgraded version MultiNLI (Williams et al., 2018). They serve as benchmarks for assessing a model's ability to understand the logical relationships between sentences, such as entailment, contradiction, and neutrality. Most language models perform well on English language tasks because they have been extensively trained on large-scale English corpora. Under this scenario, the utility of ReLLM in the English language domain may be limited. Therefore, we first rephrase the English data and then translate them to other languages with the help of NLLB-200-3.3b. We use accuracy as the evaluation metric.

ReLLM in NLI Different from summarization, which relies on word alignment between source and target sentence, NLI focuses on sentence understanding. Therefore, except the original replacement operation, that word x_i in sentence x is replaced by word $\hat{x}_i$ (ReLLM(v1)), we propose another strategy in which we provide the substitution $\hat{x}_i$ as well as keeping the original word x_i. For this strategy, we just replace x_i to $x_i(\hat{x}_i)$ (ReLLM(v2)). Some examples are provided in Appx. C.

Prompt We use the same prompt that is adopted by Zhong et al. (2023):

```
Given the sentence [text_1] written in [language], determine if the following statement is entailed or contradicted or neutral: [text_2]. Only output the label.
```

Results We present the results in Table 2. It is notable that when sentences are provided in English (eng_Latn), the baseline approach (prompt without rephrasing) achieves the highest performance in both SNLI and MultiNLI tasks. This outcome can be attributed to the fact that ChatGPT has already been fully trained on the English corpora and therefore has a strong understanding of low-frequency words. The potential imperfect modifications to the input are detrimental to performance.

On other low-resource languages, ReLLM(v2) demonstrates superior accuracy in 11 out of 14 non-English languages for the SNLI task. As for the MultiNLI task, the highest scores are distributed in a ratio of 6:6:2 among ReLLM(v2), ReLLM(v1), and the baseline. These results highlight the positive impact of providing high-frequency words in enhancing LLMs' understanding of sentences. When comparing ReLLM(v1) and ReLLM(v2), it can be observed that ReLLM(v2) performs on par with ReLLM(v1) for MultiNLI and surpasses ReLLM(v1) for SNLI. This suggests that for tasks that do not necessitate alignment between the source and target, retaining both high-frequency words and their low-frequency counterparts is more effective than substitution alone.

4 Conclusion

In this paper, we introduce ReLLM, a method designed to rephrase the input part of a prompt. Our approach involves the substitution of low-frequency words in the input with their high-frequency counterparts. We experimentally demonstrate that the rephrased prompt yields improved results in eliciting the desired information from LLMs compared to the original prompt. Importantly, the entire rephrasing process can be executed without accessing the weights and training data of LLMs. This capability proves particularly valuable in scenarios where only APIs are available.

[8] https://github.com/facebookresearch/flores/tree/main/flores200

Limitations

RELLM has only been evaluated on a limited set of tasks, and its usefulness in various other generation and classification tasks remains unconfirmed. Additionally, the number of languages in which RELLM has been tested is also restricted. Moreover, caution should be exercised when applying RELLM to high-resource languages, as it may potentially have a negative impact. Further research and experimentation are necessary to assess the broader applicability and potential limitations of RELLM.

Ethics Statement

There is no ethical issue known to us in this work. Our methods and conducted experiments are based on the well-known and widely used LLM.

References

He Bai, Tong Wang, Alessandro Sordoni, and Peng Shi. 2022. Better language model with hypernym class prediction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yile Wang, Yue Zhang, Peng Li, and Yang Liu. 2023a. Language model pre-training with linguistically motivated curriculum learning.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le,

and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Haoran Yang, Piji Li, and Wai Lam. 2022. Parameter-Efficient Tuning by Manipulating Hidden States of Pretrained Language Models For Classification Tasks. *arXiv e-prints*, page arXiv:2204.04596.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Translation Quality

As for the translation quality, although it is difficult to directly assess the translation quality in these low-resource languages, we adopt an indirect method: we translate the translated text back into English and compute the similarity between the original English corpus and the corpus which are translated back from low-resource language. We report results on 8 low-resource languages of the XSum dataset.

Language	BLEU-1	BLEU-2	ROUGE-L
aeb_Arab	53.9	30.7	50.6
ast_Latn	67.5	41.4	65.1
ayr_Latn	35.6	16.5	38.3
ban_Latn	69.7	47.7	64.9
bho_Deva	74.1	52.1	66.6
lij_Latn	71.6	51.3	68.9
lim_Latn	70.7	52.2	70.2
lus_Latn	57.1	32.0	48.8

Table 3: Similarity between original targets and targets translated back.

B Summarization Cases

At present, we have not conducted human evaluation due to the limited number of individuals proficient in low-resource languages and the associated high costs involved. We would like to provide some cases that are translated back from low-resource languages for your reference. The cases in Table 4 are from XSum dataset which aims to summarize a given article.

C NLI Cases

We present some rephrased cases of the natural language inference task in Table 5.

Gold Reference:	A leading human rights organisation has urged Nato to investigate fully the deaths of civilians in air strikes in Libya last year.
Ours:	Human Rights Watch is demanding that NaTO take responsibility in places where it claimed at least 72 deaths of child fighters caused by airstrikes during its campaign in Libya last year.
Baseline:	Human Rights Watch is demanding credible investigations into NATO airstrikes, which the organization believes killed 72 civilians last year in Libya, with NATO insisting that it cannot take responsibility for its lack of presence on the ground to confirm the deaths, something in which Amnesty International also agreed, calling it "deeply decent".
Gold Reference:	The Italian parliament has approved a long-debated and extensive electoral reform that aims to give the country more political stability.
Ours:	The lower house of the Italian parliament approved electoral reforms aimed at ending shaky alliances by guaranteeing a majority of seats to the political party that wins the most votes in the election.
Baseline:	The lower house of the Italian parliament has approved an electoral reform package that will guarantee the party that wins the most votes a majority of seats, but critics argue that giving parties too much power at the expense of the voter.

Table 4: Some cases from XSum dataset. We omit the content of articles since they are too long.

	ReLLM(v1)	ReLLM(v2)
text-1: He feels perturbed.	He feels uncomfortable.	He feels perturbed (uncomfortable).
text-2: He wants to sleep.	He wants to sleep.	He wants to sleep.
text-1: Five people are sitting on horses at a rodeo.	Five people are sitting on horses at a cowboy show.	Five people are sitting on horses at a rodeo (cowboy show).
text-2: Bandits are sitting on horses as they prepare for a robbery.	Bandits are sitting on horses as they prepare for a theft.	Bandits are sitting on horses as they prepare for a robbery (theft).
text-1: A woman is looking into a mirror, brushing her hair.	A woman is looking into a mirror, combing her hair.	A woman is looking into a mirror, brushing (combing) her hair.
text-2: The woman is taking a shower.	The woman is taking a shower.	The woman is taking a shower.

Table 5: Some rephrased cases of natural language inference task.