Evaluation of ChatGPT Models on Sentence Simplification

Xuanxin WU[1] Yuki Arase[1]
[1]Graduate School of Information Science and Technology, Osaka University
{xuanxin.wu,arase}@ist.osaka-u.ac.jp

Abstract

Sentence simplification, which rewrites a sentence to be easier to read and understand, is promising technique to help people with various reading difficulties. Recent investigations have shown that instruction-tuned large language models, namely ChatGPT-3.5, perform strongly on sentence simplification via prompting. However, it has yet to be clear how well the most advanced ChatGPT-4 addresses the problem. Also, it is unclear how effective fine-tuning of ChatGPT-3.5 is on sentence simplification. This study evaluates the capabilities of these two models, by comparing their performance with the current state-of-the-art supervised model of Control-T5. The results show that prompting ChatGPT-4 generally outperforms fine-tuned ChatGPT-3.5 and Control-T5, while lexical paraphrasing remains as a challenge.

1 Introduction

Sentence simplification aims to make sentences easier to read and understand by modifying their wordings and structure, without changing the meaning. It helps people with reading difficulties, such as non-native speakers [1, 2], individuals with aphasia [3], dyslexia [4, 5], or autism [6].

In previous years, data-driven approaches have been predominant in the field of sentence simplification, relying on large corpora of aligned complex-simple sentences [7]. These approaches often employed a sequence-to-sequence model as the core framework, then has been enhanced by integrating various sub-modules into it. Among them, fine-tuning pre-trained language models incorporating control tokens representing both lexical and syntactic complexity has achieved state-of-the-art performance. Specifically, MUSS [8] employed BART [9] while Control-T5 [10] employed T5 [11]. Despite advancements, existing simplification models do not meet the level of direct usefulness to end users [7].

Recent developments have seen the rise of large language models (LLMs). Scaling pre-trained language models, such as increasing model or data size, enhances their capacity for downstream tasks. These LLMs with tens or hundreds of billions of parameters possess unique features like in-context learning, which smaller models lack [12]. Furthermore, instruction-tuning has enabled LLMs to follow users' instructions given as prompts. Unlike earlier models that required fine-tuning, these LLMs can be effectively prompted with zero- or few-shot examples for task-solving. Notably, the ChatGPT families relesed by OpenAI demonstrate exceptional general and task-specific abilities [13, 14, 15]. In terms of sentence simplification, some studies compared LLMs' performance with the state-of-the-art supervised models. For instance, Feng et al. [16] evaluated the performance of prompting ChatGPT and GPT-3.5, later Kew et al. [17] compared 44 LLMs varying in size, architecture, pre-training methods, and with or without instruction tuning. Their findings indicate that OpenAI's LLMs generally surpass the previous state-of-the-art sentence simplification models.

However, these previous investigations have limitations in three points. First, there is a lack of comprehensive exploration regarding the most advanced ChatGPT model to date, ChatGPT-4.[1)]. Second, the efficacy of fine-tuning ChatGPT-3.5 for sentence simplification has yet to be investigated. Third, they lack sufficient exploration on prompts; they adopted a uniform prompts with few-shot examples, which applies identical instructions and offering the same quantity of examples across datasets irrespective to their varying features in simplification strategies. ChatGPT models are known to be sensitive to prompts, the previous studies overlooked the unique needs of each dataset, possibly underutilizing their potential on simplification.

We evaluate the performance of ChatGPT-4 and

1) We denote as ChatGPT-3.5/4 to make it distinctive with LLMs without instruction-tuning.

ChatGPT-3.5[2] in English sentence simplification, using prompt engineering and fine-tuning techniques on three representative datasets on sentence simplification: Turk [18], ASSET [19], and Newsela [20]. We compare these models to Control-T5, which has demonstrated the best performance in Turk and ASSET. Our findings are summarized as follows:

- ChatGPT-4 indicates generally higher performance than fine-tuned ChatGPT-3.5 and Control-T5, while lexical paraphrasing is still a challenge.
- Fine-tuning ChatGPT-3.5 yielded inferior performance, and increasing the number of training samples did not lead to improvements.

2 Evaluation Datasets and Metrics

In this study, we employed the standard datasets and metrics commonly used for English sentence simplification as detailed in the following.

2.1 Datasets

We used validation and test sets from the three standard corpora on English sentence simplification.[3] The numbers of complex-simple sentence pairs in these sets are summarized in Table 1. Remarkably, these datasets have distinctive features as summarized below.

- **Turk [18]:** This dataset comprises 2,359 sentences from English Wikipedia, each paired with 8 simplified references written by crowd-workers. It is created primarily focusing on **lexical paraphrasing**.
- **ASSET [19]:** This dataset uses the same 2,359 source sentences as the Turk dataset. It differs from Turk by aiming at rewriting sentences with more **diverse transformations**, i.e., paraphrasing, deleting phrases, and splitting a sentence, and provides 10 simplified references written by crowd-workers.
- **Newsela [20, 21]:** This dataset originates from a collection of news articles accompanied by simplified versions written by professional editors. Subsequently, it was aligned from article-level to sentence-level, resulting in approximately $94k$ sentence-level simplifications. After careful observation, we found that **deletions of words, phrases, and clauses** predominantly characterize the Newsela dataset.

Table 1: Number of complex-simple sentence pairs in the validation and test sets of each dataset

Dataset	Validation	Test
Turk	2,000	359
ASSET	2,000	359
Newsela	1,129	1,077

2.2 Metrics

Our primary focus is on **SARI** [18], a widely recognized metric for evaluating sentence simplification. **SARI** evaluates a simplification by comparing it against reference(s) and the source sentence, focusing on the words that are added, kept, and deleted. For calculating **SARI** at the corpus level, we use the EASSE package [22], which allows us to gauge the overall quality of the model's simplifications.

Additionally, we report on **LENS** [23], a recent sentence-level simplification evaluation metric. **LENS** leverages RoBERTa to perform automated evaluations by training it to predict human judgment scores, considering both the semantic similarity and the edits performed by the model compared to the source sentence and references. Following the implementation from the original LENS study [23], we computed the average LENS scores of all test samples as the final score. Note that as LENS is a model-based evaluation metric, it does not produce any clues to understand the grounds of a specific score.

3 Tuning LLMs for Simplification

This section describes the processes of our prompt engineering for ChatGPT-4 and fine-tuning of ChatGPT-3.5, respectively.

3.1 Prompt Engineering

Aiming to optimize ChatGPT-4's sentence simplification capabilities, we conducted prompt engineering based on three principal components:

- **Dataset-Specific Instructions:** We tailored instructions to each dataset's unique features and objectives, as detailed in Section 2.1. For the Turk and ASSET datasets, we created instructions referring to the

2) We used the 'gpt-4-0613' and 'gpt-3.5-turbo' models, respectively, and called them via OpenAI's APIs. The latter is the latest fine-tuning capable model to date.

3) These validation sets were used for prompt engineering and fine-tuning.

Table 2: Prompt engineering impact on SARI scores

Valid Set	SARI Diff.	Best Prompts
Turk	8.3	Turk style + Few-shot + Single ref
ASSET	4.5	ASSET style + Few-shot + Single ref
Newsela	3.6	Newsela style + Few-shot + Multi refs

guidelines provided to the crowd-workers who composed the references. In the case of Newsela, where such guidelines are unavailable, we created instructions following the styles used for Turk and ASSET, with an emphasis on deletion.

- **Varied Number of Examples:** We varied the number of examples to attach to the instructions: 0, 1, and 3.
- **Varied Number of References:** We experimented with single or multiple (namely, three) simplification references used in the examples. For Turk and ASSET, we manually selected a high-quality reference from their multiple references. For Newsela, which is basically a single-reference dataset, we extracted references targeting different levels of simplicity of the same source sentence as multiple references.

We integrated these components into prompts, resulting in the creation of 15 variations (Figure 1). These prompts were then applied to each validation set, excluding selected examples. This resulted in SARI score variations: 8.3 for Turk, 4.5 for ASSET, and 3.6 for Newsela. Prompts that achieved the highest SARI scores were designated as 'Best Prompts', which are summarized in Table 2. For more detailed information, refer to the Appendix A.

Results reveal a direct alignment between the best prompt's instructional style and its respective dataset. Additionally, these top-performing prompts all use a few-shot examples of 3. The optimal number of simplification references varies; Turk and ASSET show strong results with a single reference, whereas Newsela benefits from multiple references, likely due to the intricacies involved in ensuring meaning is preserved amidst deletions. Overall, prompt engineering notably enhances ChatGPT-4's sentence simplification output, as evidenced by the significant increase in SARI scores. Following this, we used the best prompts to produce simplifications from the respective test sets.

3.2 Fine-Tuning

We fine-tuned ChatGPT-3.5 by calling the API. We used the WikiLarge [21] training set, following the fine-tuning settings of Control-T5 to ensure consistency. ChatGPT-3.5 requires only a small amount of training dataset[4], we made efforts to sample as clean sentence pairs as possible. Our filtering based on manually designed heuristics left 58k complex-simple sentence pairs out of the original 296k pairs. We randomly sampled 50 and 1,000 pairs for fine-tuning ChatGPT-3.5. We employed prompts that yielded the highest SARI scores on ChatGPT-4.

Table 3: Performance comparison of different models

	Model	SARI	*add*	*keep*	*del*	LENS
Turk	Control-T5	**43.7**	11.2	70.2	49.7	**66.7**
	ChatGPT-4	42.9	10.7	68.1	49.9	50.9
	ChatGPT-3.5_50	37.1∗	2.8	70.0	38.4	43.1
	ChatGPT-3.5_1k	36.8∗	2.8	69.5	38.0	41.8
ASSET	Control-T5	44.9	12.3	63.0	59.4	**68.1**
	ChatGPT-4	**47.3**∗	13.2	58.2	70.4	58.9
	ChatGPT-3.5_50	42.6∗	7.9	58.2	61.7	54.9
	ChatGPT-3.5_1k	40.8∗	8.0	61.3	53.1	52.1
Newsela	Control-T5	38.6	4.6	38.4	72.7	63.7
	ChatGPT-4	41.4∗	5.6	37.4	81.2	**64.4**
	ChatGPT-3.5_50	**41.5**∗	4.6	37.4	82.5	41.7
	ChatGPT-3.5_1k	36.3∗	4.8	40.1	64.1	38.7

4 Results

As the previous state-of-the-art, we replicated the Control-T5 model [10], which fine-tuned the T5-base using the training set of WikiLarge for Turk and ASSET and that of Newsela, respectively. The results are presented in Table 3 where the best scores are emphasized in bold.

4.1 Analysis of SARI Scores

We report SARI scores alongside individual scores for each edit operation: Add (*add*), Keep (*keep*), and Delete (*del*). To assess the statistical significance of the variations in SARI scores, we employed a randomization test against Control-T5. Table 3 marks SARI scores with an asterisk (∗) where statistically significant differences were confirmed.

The results reveal that ChatGPT-4 outperforms Control-T5 in sentence simplification on the ASSET and Newsela, as indicated by its higher SARI scores. Specifically, on the ASSET test set, ChatGPT-4 achieved a SARI score of 47.3, while Control-T5 scored 44.9. Additionally, ChatGPT-4

4) OpenAI suggests starting from 50 samples: https://platform.openai.com/docs/guides/fine-tuning/preparing-your-dataset

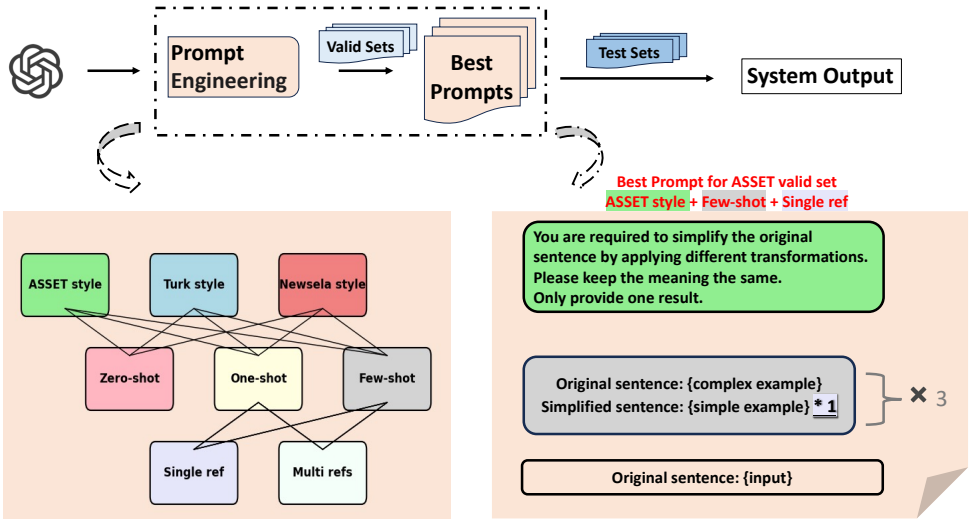

Figure 1: ChatGPT-4 prompt engineering; the identified best prompt on the ASSET validation set is illustrated here.

showed a tendency towards more extensive editing, with lower *keep* scores and higher *add* and *del* scores compared to other models. On the Newsela test set, ChatGPT-4 was scored 41.4 against Control-T5's 38.6, with a notable increase in *del* scores. In contrast, ChatGPT-4's performance on Turk is on par with Control-T5, which implies that lexical paraphrasing remains challenging for ChatGPT-4.

The fine-tuned ChatGPT-3.5 models generally received lower SARI scores across all datasets, suggesting less efficacy in sentence simplification than ChatGPT-4 and Control-T5. One exception was Newsela, where the `ChatGPT-3.5_50` model showed similar performance to ChatGPT-4 with a SARI score of 41.5 and comparable addition, keep, and deletion scores. Nevertheless, upon closer examination, we discovered that fine-tuned ChatGPT-3.5 models exhibit errors such as copies of the original sentences and fragmented sentences, which were not observed in ChatGPT-4's simplifications. It is notable that increasing training samples from 50 to 1,000 led to decreased performance in all test sets.

4.2 Analysis of LENS scores

In addition to the corpus-level evaluation using SARI, we report LENS, which is based on sentence-level evaluation. The results from LENS present a more nuanced picture: Control-T5 outrperforms ChatGPT-4 on the Turk dataset, but fails behind ChatGPT-4 on the Newsela dataset. This is consistent with the SARI results, but the score differences between the two models are large. Additionally, on the ASSET dataset, the LENS rates Control-T5 more favorably, in contrast to the SARI which shows a preference for ChatGPT-4. We found that LENS scores significantly vary among sentences; the standard deviations on ChatGPT-4 were 11.7 (Newsela) to 15.6 (Turk) and on Control-T5 were 14.3 (ASSET) to 17.0 (Newsela). A more sophisticated method to consolidate sentence-level LENS scores to corpus-level score may be necessary.

Similar to SARI results, fine-tuned ChatGPT-3.5 models were consistently scored lower than Control-T5 in LENS across all datasets. Again, increasing training samples from 50 to 1,000 did not improve LENS performance. These outcomes highlight the need for continued research into the fine-tuning of ChatGPT models for simplification.

5 Conclusion

In this study, we performed an extensive evaluation of ChatGPT-4 and ChatGPT-3.5 in sentence simplification by comparing with the state-of-the-art supervised baseline. Our findings suggest that prompting advanced ChatGPT-4 model may outperform the supervised baseline in this field. However, fine-tuning ChatGPT-3.5 was consistently inferior to the baseline, underscoring the complexities and limitations of fine-tuning ChatGPT models for sentence simplification. We also identified inconsistencies in some results between SARI and LENS, indicating areas that require further exploration. We are currently conducing human assessments of errors on outputs of these models to thoroughly understand their abilities on simplification and practical impacts from human perspectives.

References

[1] Gustavo Henrique Paetzold. Lexical simplification for non-native english speakers, September 2016. Publisher: University of Sheffield.

[2] Advaith Siddharthan. An architecture for a text simplification system. In **Language Engineering Conference, 2002. Proceedings**, pp. 64–71, 2002.

[3] John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. Simplifying text for language-impaired readers. In Henry S. Thompson and Alex Lascarides, editors, **Ninth Conference of the European Chapter of the Association for Computational Linguistics**, pp. 269–270, Bergen, Norway, June 1999. Association for Computational Linguistics.

[4] Luz Rello, Clara Bayarri, Azuki Gòrriz, Ricardo Baeza-Yates, Saurabh Gupta, Gaurang Kanvinde, Horacio Saggion, Stefan Bott, Roberto Carlini, and Vasile Topac. Dyswebxia 2.0! more accessible text for people with dyslexia. In **Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility**, W4A '13, New York, NY, USA, 2013. Association for Computing Machinery.

[5] Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. Simplify or help? text simplification strategies for people with dyslexia. In **Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility**, W4A '13, New York, NY, USA, 2013. Association for Computing Machinery.

[6] Eduard Barbu, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L. Alfonso Ureña-López. Language technologies applied to document simplification for helping autistic people. **Expert Systems with Applications**, Vol. 42, No. 12, pp. 5076–5086, 2015.

[7] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-driven sentence simplification: Survey and benchmark. **Computational Linguistics**, Vol. 46, No. 1, pp. 135–187, 2020.

[8] Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 1651–1664, Marseille, France, June 2022. European Language Resources Association.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

[10] Kim Cheng Sheang and Horacio Saggion. Controllable sentence simplification with a unified text-to-text transfer transformer. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 341–352, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.

[11] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[12] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.

[13] OpenAI. Gpt-4 technical report, 2023.

[14] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 16646–16661, Singapore, December 2023. Association for Computational Linguistics.

[15] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[16] Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. Sentence simplification via large language models, 2023.

[17] Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. BLESS: Benchmarking large language models on sentence simplification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13291–13309, Singapore, December 2023.

[18] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 401–415, 2016.

[19] Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4668–4679, Online, July 2020.

[20] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. **Transactions of the Association for Computational Linguistics**, Vol. 3, pp. 283–297, 2015.

[21] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 584–594, Copenhagen, Denmark, September 2017.

[22] Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. EASSE: Easier automatic sentence simplification evaluation. In Sebastian Padó and Ruihong Huang, 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): System Demonstrations**, pp. 49–54, Hong Kong, China, November 2019.

[23] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. LENS: A learnable evaluation metric for text simplification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 16383–16408, Toronto, Canada, July 2023.

A Best Prompts

You are required to simplify the original sentence by using simpler concepts, words, or phrases. Please keep the meaning the same. Only provide one result.

Original sentence: San Francisco Bay is located in the U.S. state of California, surrounded by a contiguous region known as the San Francisco Bay Area, dominated by the large cities San Francisco, Oakland and San Jose.
Simplified sentence: San Francisco Bay is located in the U.S. state of California, surrounded by a contiguous region known as the San Francisco Bay Area, influenced by the large cities, San Francisco, Oakland and San Jose.

Original sentence: The book chronicles events which take place in the fictional space colony of Windhaven.
Simplified sentence: The book chronicles events which take place in the space colony of Windhaven.

Original sentence: Some academic journals do refer to Wikipedia articles, but are not elevating it to the same level as traditional references.
Simplified sentence: Some academic journals do refer to Wikipedia articles, but are not using it to the same level as common references.

Original sentence: {input}

Figure 2: Turk style + Few-shot + Single ref

You are required to simplify the original sentence by applying different transformations. Please keep the meaning the same. Only provide one result.

Original sentence: Rollins retired in 1962 and opted to become a coach.
Simplified sentence: Rollins retired in 1962. He then chose to become a coach.

Original sentence: Tourism is concentrated in the mountains, particularly around the towns of Davos / Arosa, Laax and St. Moritz / Pontresina.
Simplified sentence: Tourism takes place in the mountains around the towns of Davos / Arosa, Laax and St. Moritz / Pontresina.

Original sentence: First Fleet is the name given to the 11 ships which sailed from Great Britain on 13 May 1787 with about 1,487 people to establish the first European colony in New South Wales.
Simplified sentence: 11 ships sailed from Great Britain on 13 May 1787 carrying about 1,487 people. These ships aimed to establish the first European colony in New South Wales. These 11 ships were named First Fleet.

Original sentence: {input}

Figure 3: ASSET style + Few-shot + Single ref

You are required to simplify the original sentence. You can delete information that makes the sentence difficult to understand. Only provide one result.

Original sentence: Becker was trailing an underwater camera that will help him and the other scientists figure out how to wrench out an extensive network of oyster racks held up by some 4,700 wooden posts sunk into the Estero 's sandy bottom.
Simplified sentences:
The camera will help scientists figure out how to remove the oyster racks.
The posts are sunk into the Estero 's sandy bottom.
The racks are held up by about 4,700 wooden posts.

Original sentence: He also announced a 15 percent increase in the minimum wage, effective next month, and an increase in scholarships for high school and college students.
Simplified sentences:
He said the minimum wage for workers will go up.
President Maduro said he would fix some things.
The minimum wage is the least amount of money someone can get paid to work.

Original sentence: The monitoring site, more than 5,000 feet above sea level on a pine-studded overlook above the lowest layer of the atmosphere, gives Faloona access to undisturbed air from across the Pacific before it is fouled by U.S. pollution sources.
Simplified sentences:
The spot is more than 5,000 feet above sea level.
His measuring instruments are located on Chews Ridge in the Santa Lucia Mountains.
There he can test the air blowing in from across the Pacific.

Original sentence: {input}

Figure 4: Newsela style + Few-shot + Multi refs