

InstructCMP: Length Control in Sentence Compression through Instruction-based Large Language Models

Juseon-Do¹, *Jingun Kwon¹, Hidetaka Kamigaito², and Manabu Okumura³

¹Chungnam National University, ²Nara Institute of Science and Technology (NAIST)

³Tokyo Institute of Technology

doju00@o.cnu.ac.kr

jingun.kwon@cnu.ac.kr

kamigaito.h@is.naist.jp

oku@pi.titech.ac.jp

Abstract

Extractive summarization can produce faithful summaries but often requires additional constraints such as a desired summary length. Traditional sentence compression models do not typically consider the constraints because of their restricted model abilities, which require model modifications for coping with them. To bridge this gap, we propose Instruction-based Compression (InstructCMP), an approach to the sentence compression task that can consider the length constraint through instructions by leveraging the zero-shot task-solving abilities of Large Language Models (LLMs). For this purpose, we created new evaluation datasets by transforming traditional sentence compression datasets into an instruction format. By using the datasets, we first reveal that the current LLMs still face challenges in accurately controlling the length for a compressed text. To address this issue, we propose an approach named “length priming,” that incorporates additional length information into the instructions without external resources. While the length priming effectively works in a zero-shot setting, a training dataset with the instructions would further improve the ability of length control. Thus, we additionally created a training dataset in an instruction format to fine-tune the model on it. Experimental results and analysis show that applying the length priming significantly improves performances of InstructCMP in both zero-shot and fine-tuning settings without the need of any model modifications.

1 Introduction

Sentence compression is a task of creating a concise summary from an original sentence while conveying its key information, by deleting words in the sentence. Generally, sentence compression in extractive summarization provides more faithful summaries than abstractive summarization (Cao et al., 2018).

* corresponding author

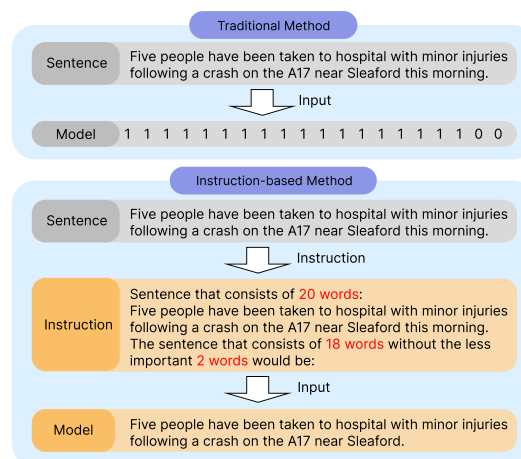


Figure 1: Process of transforming a traditional labeled dataset into an instruction-based format. The binary output of “1” or “0” from the traditional methods corresponds to keeping or dropping words, respectively. Length constraints in “length priming” are highlighted in red in the instruction.

While traditional sentence compression methods used tree trimming, the approaches can be affected by parsing errors (Jing, 2000; Knight and Marcu, 2000; Berg-Kirkpatrick et al., 2011; Filippova and Altun, 2013). The introduction of LSTM-based Seq2Seq approaches aims to address this issue although their performance tends to degrade in handling longer sentences (Filippova et al., 2015). To solve this problem, Kamigaito and Okumura (2020) incorporated syntactic dependency trees into the Seq2Seq attention mechanism (Kamigaito et al., 2018) by jointly learning the dependency trees and sentence compression models. However, the state-of-the-art model required a considerable amount of ground-truth data for training (Filippova and Altun, 2013; Hasegawa et al., 2017).

Recently, unsupervised sentence compression has gained attention by exploiting BERT-based encoder models (Devlin et al., 2019). These models incorporated various scoring functions that target improving fluency and faithfulness in compress-

sion without relying on ground-truth data (Niu et al., 2019; Zhou and Rush, 2019; Schumann et al., 2020; Ghalandari et al., 2022). However, these approaches are inefficient because they require extensive model modifications, such as including classifiers or modifying beam search for objective-specific fine-tuning.

In general, summarization requires additional constraints such as a summary length (Takase and Okazaki, 2019; Dou et al., 2021; Kwon et al., 2023a). The traditional task setting for sentence compression often did not consider this factor because of the restricted model abilities, which require model modifications to handle such constraints (Schumann et al., 2020; Ghalandari et al., 2022).

Recently, LLMs have gained considerable attention for their remarkable zero-shot task-solving abilities, especially under instruction-based settings (Ouyang et al., 2022; Wei et al., 2022a). Inspired by these latest advancements, we present Instruction-based Compression (InstructCMP), a novel approach to sentence compression that accommodates a length constraint through explicit instructions, without necessitating model modifications. To the best of our knowledge, this approach represents the first implementation of sentence compression in an instruction-based framework. For this purpose, we transformed traditional sentence compression datasets into an instruction-based format for evaluation.

However, recent LLMs do not consistently generate an output of the precise length, even when specific instructions to include such constraints are provided in a zero-shot manner (Zhou et al., 2023; Qin et al., 2023). Furthermore, as we validate it later, even when testing with the latest powerful models, such as ChatGPT (GPT-4) and ChatGPT (GPT-4-1106-preview),¹ accurately adhering to length constraints remains a substantial challenge.

To address this problem, we propose an instruction approach for better length control, which is named “length priming.” We incorporate additional length information (Misra et al., 2020) into the instruction. In addition to specifying the number of deleted words for the desired length, we include the length to be retained and the number of words in the source sentence in the instruction, without any external resources. To further improve length controllability, we additionally created a training

¹<https://chat.openai.com/>

Work	Length Const.	Mod.
Filippova et al. (2015)*	✗	✗
Zhao et al. (2018)*	✗	✗
Kamigaito and Okumura (2020)*	✗	✗
Schumann et al. (2020)	✓	✗
Ghalandari et al. (2022)	✓	✗
Ours (InstructCMP)	✓	✓

Table 1: Comparison of various sentence compression models with InstructCMP. * indicates that the model was learned in a supervised manner, while others were learned in an unsupervised manner. Mod. indicates a requirement of model modifications for constraints.

dataset with the instructions to fine-tune the model using the dataset. Figure 1 shows the transformation process for an instruction format.

We conducted experiments on four benchmark datasets and performed an in-depth analysis to evaluate the effectiveness of LLMs in compressing sentences under the length constraint. The analysis considers the following factors: the model type and the number of parameters for the model size. Experimental results show that InstructCMP with length priming compresses sentences in a zero-shot setting while successfully keeping the desired length without model modifications. The performance can be further improved by fine-tuning it with the created instruction-based training dataset. The “length priming” method proves effective in both zero-shot and fine-tuning settings, as shown by significant improvements in the ROUGE metrics and adherence to the length constraint, even when using ChatGPT (GPT-4) and ChatGPT (GPT4-1106-preview). Our in-depth analysis also showed that InstructCMP can compress sentences while maintaining faithfulness. Our experiments show that instruction-based models like ChatGPT can effectively control the length when provided with more specific length-related information.²

2 Problem Statement

The traditional approach to sentence compression is considered as a sequential labeling task (Filippova et al., 2015; Wang et al., 2017; Zhao et al., 2018; Kamigaito and Okumura, 2020; Schumann et al., 2020; Ghalandari et al., 2022). Each source token in a sequence, represented as $\mathbf{x} = \{x_0, x_1, \dots, x_n\}$, is processed using a sentence compression model to predict a corresponding label sequence, which is

²Our code and datasets are available at: <https://github.com/JuseonDo>.

$\mathbf{y} = \{y_0, y_1, \dots, y_n\}$, where $y_i \in \{1, 0\}$.

While the method is straightforward, it has limitations in incorporating additional constraints such as a desired length. Addressing these requirements in the traditional approach typically involves modifications to the model, which is inefficient (Schumann et al., 2020; Ghalandari et al., 2022).

To overcome these limitations, we utilize the recent powerful instruction-based LLMs for the sentence compression task (Touvron et al., 2023; Chung et al., 2022). Table 1 shows a comparison between previous work on traditional sentence compression and InstructCMP. Unlike the previous work, InstructCMP incorporates a length constraint directly into the instruction format, allowing models to process and learn the constraint as a part of their input. This allows an efficient and flexible solution for practical sentence compression, without extensive model modifications.

3 Instruction-based Compression

In this section, we describe InstructCMP. We consider “length priming” for a length constraint in it. We created new evaluation datasets by transforming traditional sentence compression datasets into an instruction format. To further improve performances of InstructCMP, we also created a new training dataset in an instruction-based template.

3.1 Instruction Template

Table 2 shows instructions that include a length constraint. The first instruction permits InstructCMP to compress text by deleting words without any constraints. However, in general, summarization requires a desired length for compressed texts (Makino et al., 2019; Dou et al., 2021; He et al., 2022; Kwon et al., 2023a).

Length Priming. To apply the length constraint, we first construct an instruction that deletes words to meet a desired length (Constraint 2). It is easy to calculate the number of words to be deleted for any desired length.

However, LLMs do not consistently follow instructions, particularly when processing length constraints (Zhou et al., 2023; Qin et al., 2023). To address this issue, we propose the “length priming” method for the length constraint instruction for enhanced length comprehension. Constraint 3 considers the total length of the source text and the number of words that should be kept and deleted together. Considering such additional length infor-

mation can enable InstructCMP to recognize the length constraint more effectively. The number of words that should be kept is automatically calculated from the target desired length.

Constraint 3-1 applies the “length priming” only to the source text based on its length, whereas Constraint 3-2 applies it solely to the target text based on the number of words that should be kept and deleted together.

3.2 Dataset Creation

We consider four benchmark datasets. The Google dataset (**Google**) was automatically created by considering the syntactic dependency tree structure from news headlines (Filippova and Altun, 2013). The training, validation, and test datasets consist of 200,000, 1,000, and 1,000 pairs, respectively. For the test dataset used in the evaluation, the gold compression ratio is 0.45. The Broadcast (**Broad**) and BNC (**BNC**) datasets (Clarke and Lapata, 2008) comprise manually compressed sentences. Each of these datasets contains 1,370 and 1,629 evaluation pairs, respectively. The gold compression ratios of these datasets, which are 0.76 and 0.72 respectively, are longer than those of other evaluation datasets. DUC2004 (TASK1) (**DUC**) comprises 500 pairs with a gold compression ratio of 0.39. Unlike other evaluation datasets, this dataset includes abstract summaries as its ground truth.

We created new datasets by transforming traditional sentence compression datasets into an instruction format. For length constraint instructions, we inject lengths of ground-truth summaries.

3.3 Instruction-based Fine-tuning

To improve performances by leveraging LLM’s generalizability (Wang et al., 2022; Wei et al., 2022a; Chung et al., 2022), we also created a training dataset for instruction-based fine-tuning by sampling 5% of the training dataset from **Google**. Through this fine-tuning, we aim to enhance a model for better learning and improving abilities to handle length constraints in compressing sentences without any model modifications.

4 Experiments

4.1 Experimental Settings

Evaluation Metrics. F_1 scores of ROUGE-1 (R-1), -2 (R-2), and -L (R-L), the F_1 -score for kept tokens (F_1), and the BERT score (BS) (Zhang* et al., 2020) were used to evaluate compression

#	Constraint	Instruction
1	\times	Sentence:\n{src}\n\nThe sentence without the less important words would be:\n
2	Length w/o priming	Sentence:\n{src}\n\nThe sentence without the less important {del} words would be:\n
3	Length	Sentence that consists of {src len} words:\n{src}\n\nThe sentence that consists of {keep} words without the less important {del} words would be:\n
3-1	Length w/o tgt priming	Sentence that consists of {src len} words:\n{src}\n\nThe sentence without the less important {del} words would be:\n
3-2	Length w/o src priming	Sentence:\n{src}\n\nThe sentence that consists of {keep} words without the less important {del} words would be:\n

Table 2: Instruction formats for length constraints, created by transforming a traditional dataset. “src” indicates the placeholder for a source sentence. “del” denotes the placeholder for the number of deleted words. “keep” and “src len” denote additional length information.

quality. The ROUGE scores were calculated using the implementation provided by Google Research.³

To evaluate performances related to a length constraint, we calculated ΔCR , the difference between the model-generated compression ratio and the gold compression ratio. ΔCR evaluates how close the compression ratio of model-generated outputs is to the gold compressed summary (Kamigaito et al., 2018; Kamigaito and Okumura, 2020). Because InstructCMP can produce novel words, we counted the number of the novel words in the model-generated compressed summaries. Thus, *novel* represents the ratio of novel words that do not appear in the source text.

Implementation Details. We employed the instruction-based open-source Llama2-13B-chat model (Touvron et al., 2023)⁴ as our backbone model. We tested various instructions on the validation dataset from Google and made selections based on their performance. To explore various parameter numbers, we experimented with 4-bit and 8-bit quantizations, as well as without quantization (Jacob et al., 2018) using PyTorch.⁵ We also evaluated the performance across various model sizes, including 7B and 70B, and compared various model types, specifically the encoder-decoder based models of FLAN-T5-XXL (11B) (Chung et al., 2022)⁶ and FLAN-UL2 (20B) (Tay et al., 2023).⁷

For instruction-based fine-tuning, we considered QLoRA, which can preserve the full 16-bit fine-tuning performance (Detmeters et al., 2023).

³<https://github.com/google-research/google-research/tree/master/rouge>

⁴<https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

⁵<https://github.com/pytorch/pytorch>

⁶<https://huggingface.co/google/flan-t5-xxl>

⁷<https://huggingface.co/google/flan-ul2>

QLoRA is an extended version of Low-Rank Adapters (LoRA) (Hu et al., 2022), an improved Parameter-Efficient Fine-Tuning (PEFT) (Mantroukar et al., 2022) method for LLMs. This method combines low-rank and trainable matrices with the frozen weights in each layer of Transformer, building upon the foundational approach of LoRA. We incorporated low-rank matrices into the query and value weights using a LoRA attention dimension of 8. During training, we used 8-bit quantization for QLoRA, and during inference, we employed 4-bit quantization.

4.2 Main Results

Table 3 shows the performances of InstructCMP based on the Llama2-13B-chat model in a zero-shot setting, used directly without additional training, and in the QLoRA instruction-tuning setting, which involves fine-tuning of InstructCMP. Because prompting techniques for LLMs, such as few-shot (Min et al., 2022), directional stimulus (Li et al., 2023), and generated knowledge (Liu et al., 2022) methods, require external resources, we compared “length priming” to prompting techniques of chain-of-thought (Wei et al., 2022b) and tree-of-thought in a single prompt (Yao et al., 2023; Hulbert, 2023) by adding them at the beginning of length constraint instructions (#2 in Table 2).

Performance in Instruction-based Zero-shot.⁸

Even in a zero-shot setting, InstructCMP without a length constraint (#1 in Table 2) successfully compresses sentences, while it cannot necessarily meet the length. In applying length constraints with “length priming”, consistently improved performances are observed in both ROUGE and ΔCR . In addition, our “length priming” significantly outper-

⁸Experiments considering various instructions on the validation dataset from Google are detailed in Appendix A.

Dataset	Setting	Instruction	Prompting	R-1	R-2	R-L	F ₁	BS	ΔCR	novel
Google	Zero-shot	#1	\times	65.88	55.48	65.42	0.66	0.66	+30.22	0.28
		#2	Chain-of-Thought	65.74	<u>56.12</u>	<u>65.56</u>	0.66	0.66	+32.46	0.11
		#2	Tree-of-Thought	65.56	55.34	65.19	0.66	0.66	+30.99	0.17
		#3	Priming	74.59[†]	62.45[†]	73.69[†]	0.74[†]	0.73	+10.13 [†]	0.57
	QLoRA fine-tuning	#1	\times	82.85	75.15	82.58	0.84	0.82	-1.28	0.17
		#2	Chain-of-Thought	<u>84.88</u>	<u>77.20</u>	<u>84.56</u>	<u>0.86</u>	0.83	<u>-0.90</u>	0.18
		#2	Tree-of-Thought	84.69	76.89	84.26	0.85	0.83	-1.90	0.17
		#3	Priming	86.88[†]	79.55[†]	86.26[†]	0.87[†]	0.84	-0.16 [†]	0.17
Broad	Zero-shot	#1	\times	<u>79.30</u>	65.54	<u>78.27</u>	<u>0.79</u>	0.76	+4.21	0.32
		#2	Chain-of-Thought	78.94	<u>65.76</u>	78.21	0.79	0.75	+3.99	0.19
		#2	Tree-of-Thought	78.02	63.90	77.32	0.78	0.74	+4.17	0.33
		#3	Priming	80.27[†]	66.62[†]	79.30[†]	0.80[†]	0.76	-0.01 [†]	0.33
	QLoRA fine-tuning	#1	\times	70.14	58.15	69.70	0.68	0.68	-15.88	0.34
		#2	Chain-of-Thought	<u>78.24</u>	<u>65.61</u>	<u>77.78</u>	<u>0.77</u>	0.72	-3.96	0.36
		#2	Tree-of-Thought	77.68	64.94	77.06	0.76	0.71	-7.46	0.32
		#3	Priming	82.63[†]	69.76[†]	81.16[†]	0.81[†]	0.75	-1.38 [†]	0.35
BNC	Zero-shot	#1	\times	<u>74.81</u>	61.21	73.64	0.75	0.70	+10.38	0.37
		#2	Chain-of-Thought	74.46	61.03	<u>73.66</u>	<u>0.75</u>	0.69	+3.57	0.11
		#2	Tree-of-Thought	73.81	60.11	72.82	0.74	0.68	+7.01	0.26
		#3	Priming	75.78[†]	61.76	74.52[†]	0.76[†]	0.70	+0.16 [†]	0.25
	QLoRA fine-tuning	#1	\times	61.28	49.61	60.51	0.60	0.59	-24.21	0.27
		#2	Chain-of-Thought	<u>75.58</u>	<u>62.55</u>	<u>74.76</u>	<u>0.74</u>	0.68	-4.35	0.27
		#2	Tree-of-Thought	73.37	60.22	72.30	0.72	0.66	-10.81	0.25
		#3	Priming	77.54[†]	64.38[†]	76.00[†]	0.76[†]	0.70	-4.13	0.26
DUC	Zero-shot	#1	\times	<u>27.09</u>	<u>8.72</u>	<u>22.65</u>	<u>0.23</u>	0.33	+37.97	0.25
		#2	Chain-of-Thought	26.28	8.35	21.86	0.23	0.32	+40.53	0.10
		#2	Tree-of-Thought	26.13	8.20	21.75	0.23	0.32	+40.31	0.19
		#3	Priming	28.19[†]	9.66[†]	24.56[†]	0.24[†]	0.34	+15.08 [†]	0.81
	QLoRA fine-tuning	#1	\times	27.31	9.21	24.34	0.24	0.35	+0.28	0.18
		#2	Chain-of-Thought	26.29	8.62	23.40	0.23	0.34	-3.10	0.19
		#2	Tree-of-Thought	26.28	8.38	23.58	0.23	0.34	-2.29	0.20
		#3	Priming	26.83	8.57	23.96	0.23	0.33	+0.78	0.21

Table 3: Experimental results of InstructCMP using Llama2-13B-chat on Google, Broad, BNC, and DUC. Checkmark indicates not applying a length constraint. [†] indicates the improvement is significant ($p < 0.05$) compared with the underlined (generally, the best baseline score) on each dataset.

forms other prompting methods, chain-of-thought and tree-of-thought, in both length controllability and ROUGE metrics.

However, controlling the length of outputs for Google and DUC proved to be more challenging than Broad and BNC, specifically, in a zero-shot setting. We think this challenge arises from the nature of datasets, whose compression ratio is lower. Table 4 shows the results based on a target compression ratio of 0.2 and a target word count of 5 words, respectively. We observed that when the compression ratio is lower, the LLMs have difficulties maintaining both informativeness and length controllability.

Performance in Instruction-based Fine-tuning.⁹

Following instruction-based QLoRA fine-tuning, the created training dataset further improves per-

formances of InstructCMP. As shown in ΔCR for Broad and BNC, the model without the length constraint was trained to compress sentences more closely aligned with the gold compression ratio of Google. However, the performance degradation was observed on DUC when fine-tuning was applied using Google, due to the different natures of their abstractive and extractive ground-truth summaries.

Length Priming. The ablation results for “length priming” in instructions are presented in Table 6. We first compare performances of “length priming” in an unsupervised zero-shot setting. It significantly improved performances on all datasets in terms of ΔCR compared to w/o priming. Even in a supervised instruction-based fine-tuning, “length priming” largely improved performances in both ROUGE metrics and length controllability. The exception is on DUC because of its nature of the abstractive gold summary.

⁹Experiments considering 0.5% and 1% randomly sampled training datasets from Google are detailed in Appendix B.

Data	Boundary	cnt	R-1	R-2	R-L	F ₁	ΔCR	src len	tgt len	gen len
Google	0.8~1.0	32	86.05	74.22	85.18	0.85	-1.02	-	-	-
	0.6~0.8	180	81.09	69.64	79.96	0.80	8.24	-	-	-
	0.4~0.6	343	77.86	66.87	77.15	0.78	10.96	-	-	-
	0.2~0.4	403	70.15	56.94	69.15	0.69	11.05	-	-	-
	0.0~0.2	42	53.78	39.52	53.36	0.51	11.18	-	-	-
	20~	13	80.43	69.01	79.67	0.79	-	38.08	20.85	25.31
	15~20	127	78.22	66.98	76.86	0.77	-	29.46	16.31	18.58
	10~15	518	75.97	64.40	75.03	0.76	-	26.74	11.68	14.75
	5~10	338	71.11	57.75	70.45	0.70	-	25.90	7.69	10.16
	0~5	4	55.42	43.45	55.52	0.58	-	27.25	4.00	7.75
DUC	0.8~1.0	8	11.23	3.43	10.26	0.15	-9.44	-	-	-
	0.6~0.8	20	18.18	5.27	15.36	0.16	15.43	-	-	-
	0.4~0.6	118	30.51	10.48	26.12	0.27	14.60	-	-	-
	0.2~0.4	326	29.56	10.24	25.93	0.24	17.64	-	-	-
	0.0~0.2	18	20.61	5.91	17.88	0.18	11.86	-	-	-
	15~20	26	22.97	5.96	18.45	0.24	-	32.15	15.65	19.96
	10~15	363	29.95	10.23	26.07	0.26	-	33.55	11.62	17.06
	5~10	101	25.66	9.37	22.81	0.19	-	33.06	8.38	14.11

Table 4: Effect of compression ratio and word count. *cnt* indicates the number of instances in each boundary.

Data	Setting	Output	Gram.	Faith.	Info.
Google	QLoRA	13B	4.14 [†]	4.09	4.06 [†]
	Zero-shot	13B	<u>4.06</u>	4.09	<u>4.00</u>
	Gold	-	4.03	4.11	4.05
Broad	Zero-shot	13B	3.92	3.88	3.86
		70B	3.90	3.87	3.90 [†]
	Gold	-	3.92	3.88	<u>3.85</u>
BNC	Zero-shot	13B	3.98	3.93	3.93
		70B	3.96	3.91	3.96
	Gold	-	3.96	3.94	3.92

Table 5: Human evaluation results. The notations are the same as those in Table 3.

We also compare the effectiveness of “length priming,” using larger models, such as Llama-2-70B-chat-hf, ChatGPT (GPT-4), and ChatGPT (GPT-4-1106-preview). Figure 2 shows the results. We confirm that “length priming” is essential for length constraints, even in the most recent and powerful LLMs.¹⁰

5 Analysis

5.1 Parameter Sizes

The left graph of Figure 3 shows the F₁ score for kept tokens and the model-generated compression ratio (CR), compared to the gold compression ratio, based on zero-shot InstructCMP without a length

¹⁰When we additionally tested the chain-of-thought and tree-of-thought prompting methods on these larger models, their length controllability was similar to each other, which is similar to the results in Table 3.

Data	Method	Instruction	R-1	R-2	R-L	F ₁	BS	ΔCR	
Google	Zero-shot	#2	63.73	54.04	63.54	0.64	0.64	+38.44	
		#3	74.59 [†]	62.45 [†]	73.69 [†]	0.74 [†]	0.73	+10.13 [†]	
		#3-1	67.32	57.61	67.01	0.68	0.67	+30.63	
		#3-2	73.72	60.66	72.94	0.72	0.72	+9.58	
	QLoRA fine-tuning	#2	84.99	77.43	84.69	0.86	0.83	+1.45	
		#3	86.88 [†]	79.55 [†]	86.26 [†]	0.87 [†]	0.84	-0.16 [†]	
		#3-1	85.20	77.46	84.72	0.86	0.83	+0.76	
		#3-2	86.80	79.58	86.29	0.87	0.84	+0.12	
	Broad	Zero-shot	#2	81.08	67.79	80.55	0.81	0.77	+8.78
			#3	80.27	66.62	79.30	0.80	0.76	-0.01 [†]
			#3-1	81.13	68.14	80.55	0.81	0.77	+6.91
		#3-2	78.64	64.58	77.63	0.78	0.74	-1.42	
QLoRA fine-tuning		#2	80.34	67.77	79.81	0.78	0.75	-1.02	
		#3	82.63 [†]	69.76 [†]	81.16 [†]	0.81 [†]	0.75	-1.38	
	#3-1	82.80	70.39	82.05	0.81	0.77	+0.90		
#3-2	82.66	69.81	81.16	0.81	0.75	-1.08			
BNC	Zero-shot	#2	77.36	63.64	76.59	0.78	0.72	+10.46	
		#3	75.78	61.76	74.52	0.76	0.70	+0.16 [†]	
		#3-1	77.24	63.52	76.50	0.77	0.72	+8.53	
		#3-2	73.16	59.17	71.82	0.73	0.68	-4.05	
	QLoRA fine-tuning	#2	73.74	61.52	72.92	0.72	0.68	-5.50	
		#3	77.54 [†]	64.38 [†]	76.00 [†]	0.76 [†]	0.70	-4.13 [†]	
		#3-1	77.62	64.58	76.45	0.77	0.70	-1.49	
		#3-2	77.40	64.20	75.81	0.76	0.68	-4.03	
	Zero-shot	#2	26.23	8.38	21.70	0.23	0.31	+46.37	
		#3	28.19 [†]	9.66 [†]	24.56 [†]	0.24 [†]	0.34	+15.08 [†]	
		#3-1	26.53	8.59	22.33	0.23	0.32	+41.51	
		#3-2	28.41	9.85	24.66	0.24	0.34	+16.45	
QLoRA fine-tuning	#2	27.20	8.98	24.27	0.24	0.35	+0.47		
	#3	26.83	8.57	23.96	0.23	0.33	+0.78		
	#3-1	26.25	8.27	23.49	0.23	0.34	-1.22		
	#3-2	26.46	8.31	23.62	0.23	0.33	+1.32		

Table 6: Ablation study for “length priming.” The notations are the same as those in Table 3.

constraint on the Llama2-chat model with 7B, 13B, and 70B parameters. On Google and DUC, the F₁ scores increased with enlarging the model size, achieving compression closer to the gold compression ratio. However, on Broadcast and BNC, which have high gold compression ratios, InstructCMP with the 70B model compresses sentences more concisely, resulting in a compression ratio that significantly deviates from the gold compression ratio,

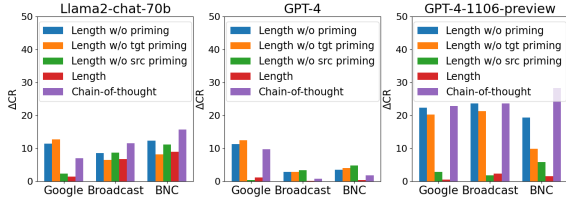


Figure 2: Absolute ΔCR for “length priming” types.

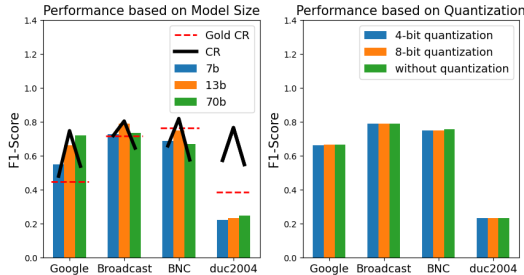


Figure 3: Performances for different model sizes and quantizations.

consequently decreasing F_1 scores compared to the 13B model.

To further investigate this, we conducted human evaluations. We sampled 100 instances each from Google, Broad, and BNC. By using Amazon Mechanical Turk, we assigned in total 120 evaluators who obtained both US high school and US bachelor’s degrees for grading the results with scores from 1 to 5 (5 is the best) in terms of grammatical correctness (Gram), factual consistency (Faith), and a balance of redundancy and informativeness (Info). Table 5 shows the results. Because of the automatically constructed nature of Google, QLoRA and zero-shot settings can yield higher grammaticality scores than the gold summary. These results also indicate gold summaries of Broad and BNC are actually redundant (Ghalandari et al., 2022), and our instruction-based approach can generate faithful, informative, and grammatical summaries.

The right graph of Figure 3 shows the results of zero-shot InstructCMP without a length constraint on Llama2-13B-chat. Interestingly, there are no significant differences in performance among the 4-bit, 8-bit, and nonquantized versions.

5.2 Model Types

It is also of interest to draw comparisons with other instruction-based models, such as FLAN-T5-XXL and FLAN-UL2, both of which employ the encoder-decoder architecture. However, they did not effectively compress sentences using instruction templates in Table 2. We think this is due

Data	Model	Instruction	R-1	R-2	R-L	F_1	ΔCR
Google	T5-XXL	#1	60.06	50.52	59.81	0.60	+47.84
		#2	<u>62.41</u>	51.18	<u>61.90</u>	<u>0.61</u>	<u>+35.72</u>
		#3	66.22[†]	51.68	65.43[†]	0.62[†]	+19.51[†]
	UL2	#1	63.53	<u>45.79</u>	62.35	0.57	+11.92
		#2	<u>64.72</u>	44.38	<u>63.87</u>	<u>0.57</u>	+1.11
		#3	66.06[†]	47.24[†]	65.39[†]	0.59[†]	+6.34
Broad	T5-XXL	#1	82.45	69.33	81.93	0.81	+12.72
		#2	74.42	59.57	72.96	0.72	+2.18
		#3	77.68	63.47	76.58	0.76	+4.78
	UL2	#1	73.82	<u>56.45</u>	<u>70.90</u>	<u>0.70</u>	-7.77
		#2	68.79	52.27	66.70	0.66	-9.84
		#3	74.31	59.12[†]	72.79[†]	0.71[†]	-4.04 [†]
BNC	T5-XXL	#1	75.35	61.44	74.33	0.74	+11.30
		#2	63.99	48.06	61.90	0.61	-5.48
		#3	65.43	49.78	63.55	0.62	-3.43 [†]
	UL2	#1	67.42	49.90	63.88	0.62	-10.64
		#2	60.40	43.54	57.45	0.56	-13.62
		#3	64.88	49.03	62.81	0.61	-8.17 [†]

Table 7: Experimental results from zero-shot instruction-based FLAN models using encoder-decoder architectures. The notations are the same as those in Table 3.

to the nature of their pre-training, which causes potential gaps between the pre-training steps and the instruction templates for extractive summarization settings (Kwon et al., 2023a). Thus, we used slightly modified instruction templates.¹¹ Table 7 shows the results. Our “length priming” can improve length controllability by keeping ROUGE metrics compared to w/o priming.

5.3 Case Study

Table 8 shows the outputs of zero-shot InstructCMP based on the Llama-13B-chat model. The first example shows the controllability of the length constraint instruction. Even when instructed to delete zero words, InstructCMP follows the instruction correctly. The second example shows the flawless grammatical capabilities of LLMs (Mitrović et al., 2023). When deleting a single word can cause a grammatical error, InstructCMP can correct the error by paraphrasing, represented as *novel* in Table 3. The third example shows the output of InstructCMP in response to the length constraint. “Length priming” assists InstructCMP to compress a source text to meet a desired length, performing better than the length constraint without priming.

5.4 Comparison with the Baselines

We compare InstructCMP with traditional state-of-the-art (SOTA) baselines, specifically SCRL,¹² which employs reinforcement learning optimized

¹¹Experimental results using instruction templates in Table 2 and modified instruction templates are in Appendix C.

¹²<https://github.com/complementizer/r1-sentence-compression>

Source. Eni has won a license for exploration block SM-857 offshore Brazil.
Instruction. Sentence that consists of 11 words:\n{source}\nThe sentence that consists of 11 words without the less important 0 words would be\n:
InstructCMP. Eni has won a license for exploration block SM-857 offshore Brazil.
Source. Rick Riordan has revealed the cover for his latest crossover short story, "Staff of Serapis", which features Annabeth Chase and Sadie Kane.
InstructCMP. Rick Riordan has revealed the cover for his latest crossover short story, featuring Annabeth Chase and Sadie Kane.
Source. Chinese shares closed lower Wednesday dragged down by the bio-pharmaceutical sector and small enterprises with growth potential.
Length const. w/o priming. Chinese shares closed lower Wednesday dragged down by the bio-pharmaceutical sector.
Length const. Chinese shares closed lower Wednesday.
Gold: Chinese shares closed lower Wednesday.

Table 8: Outputs of InstructCMP on Google.

Data	Model	R-1	R-2	R-L	F ₁	BS	len
Unsupervised							
Google	SCRL*	70.22	53.03	69.84	0.71		10.8
	SCRL	70.53	53.30	70.07	0.71	0.65	10.3
	InstructCMP	74.92 [†]	62.53 [†]	73.83 [†]	0.75 [†]	0.75	10.8 [†]
Broad	SCRL	83.04	66.64	82.64	0.82	0.74	81%
	InstructCMP	77.93	63.33	76.85	0.78	0.74	77% [†]
BNC	SCRL	79.55	62.24	78.69	0.79	0.69	79%
	InstructCMP	75.11	60.56	74.03	0.75	0.70	74% [†]
DUC	SCRL	26.78	8.14	23.30	0.22	0.25	10.0
	InstructCMP	28.14 [†]	9.43 [†]	24.82 [†]	0.23 [†]	0.32	10.6 [†]
Supervised							
Google	SLAHAN*				0.86		
	SLAHAN	82.98	74.35	82.75	0.83	0.78	9.3
	InstructCMP	82.85	75.15 [†]	82.58	0.84 [†]	0.82	9.5

Table 9: Comparison with traditional state-of-the-art baselines. * indicates the reported score in the original paper. *len* indicates the generated summary length. The notations are the same as those in Table 3.

in unsupervised settings, and SLAHAN,¹³ which recursively tracks parent and child words and leverages BERT embeddings optimized in supervised settings, trained on Google (Kamigaito and Okumura, 2020).

Following SCRL, we set a desired length of 11 for Google and DUC. In line with the previous work, we truncated model-generated outputs to 75 characters and used ROUGE recall scores for DUC (Schumann et al., 2020; Ghalandari et al., 2022). For Broadcast and BNC, the desired length was set to 75% of the length of the source sentence. Table 9 shows the results. Because zero-shot InstructCMP faces challenges in compressing sentences with length constraints when the gold compression ratio is low, we increased the model capability by using Llama2-70B-chat for Google and DUC instead of

¹³<https://github.com/kamigaito/SLAHAN>

Data	Size	R-1	R-2	R-L	F ₁	ΔCR
Google		87.45	80.47	87.00	0.88	0.69
Broad	10%	79.21	66.31	77.51	0.78	-1.44
BNC		83.38	70.29	81.84	0.81	0.42
DUC		27.02	8.34	23.85	0.23	2.09
Google		89.01	82.24	88.56	0.89	0.39
Broad	15%	79.72	66.47	78.27	0.79	0.02
BNC		82.92	69.65	81.90	0.82	0.57
DUC		26.30	7.92	23.53	0.23	2.14

Table 10: LoRA fine tuned model: training dataset size 10% and 15% from randomly sampled from Google dataset with #3 instruction based on the 13B model

Llama2-13B-chat. We observed comparable performances of InstructCMP to SCRL.

We also compare InstructCMP, based on Llama-13B-chat, with SLAHAN. Following the previous work, we fine-tuned InstructCMP without a length constraint and achieved significant improvement, even after using 5% of the training dataset.

5.5 Increasing Training Dataset Size

We provide additional experimental results using larger datasets for QLoRA fine-tuning with 10% and 15% google training datasets. Table 10 shows the results. Three different benchmark results on Google, Broad, and BNC support that length priming is necessary, except for DUC due to its abstract summary nature, and indicate the generalization of the length priming instruction.

6 Related Work

Sentence Compression. Early studies on sentence compression in both supervised and unsupervised learning frameworks have used linguistic constraints, such as tree trimming methods (Jing, 2000; Knight and Marcu, 2000; Hori and Furui, 2004; Clarke and Lapata, 2006; Berg-Kirkpatrick et al., 2011; Filippova and Altun, 2013). To avoid potential parsing errors in the tree trimming, LSTM-based models have been introduced for deletion-based compression (Filippova et al., 2015) by jointly considering eye-tracking data (Klerke et al., 2016) and incorporating a score function of an ILP-based tree trimming method (Wang et al., 2017). Zhao et al. (2018) explored reinforcement learning for a syntax-based language model, that does not use explicit parsed trees. Kamigaito et al. (2018); Kamigaito and Okumura (2020) proposed Seq2Seq approaches that jointly learn sentence compression and dependency trees within their attention networks inspired by supervised head attention

(Kamigaito et al., 2017), an extensible approach to document-level summarization (Ishigaki et al., 2019) similar to the case of graph neural networks (Xu et al., 2020; Kwon et al., 2021). Alternatively, some recent work has utilized LLMs, such as BERT, for sentence compression to optimize fluency in unsupervised frameworks (Zhou and Rush, 2019; Niu et al., 2019; Schumann et al., 2020). Because a high-quality compressed sentence can infer from the original sentence, encoder-decoder-based autoencoder approaches have been also explored (Miao and Blunsom, 2016; Févry and Phang, 2018; Malireddy et al., 2020). For better optimization, reinforcement learning has been used (Wang et al., 2018; Ghalandari et al., 2022).

Length Control. Despite the success of previous studies, practical summarization requires additional constraints such as a summary length for compressing sentences (Liu et al., 2018; Takase and Okazaki, 2019; Li et al., 2020; He et al., 2022). The approach for controlling the output to a desired length required modifying model parameters (Kikuchi et al., 2016), applying direct constraints (Takase and Okazaki, 2019; Makino et al., 2019; Kwon et al., 2023a), or splitting the training dataset into specific length ranges (He et al., 2022) due to the limited model abilities. Traditionally, sentence compression heavily relies on the model modifications for constraints such as lengths (Schumann et al., 2020; Ghalandari et al., 2022).

Instruction-based LLMs. LLMs can perform various tasks in a zero-shot setting, using instruction-formatted inputs (Brown et al., 2020; Radford et al., 2019). The emergence of instruction-based LLMs, such as ChatGPT and GEMINI,¹⁴ has demonstrated a significant improvement in performance, particularly in their zero-shot problem-solving abilities (Feng et al., 2023; Fang et al., 2023). Because performance varies greatly with various instructions, previous studies focused on finding better instructions (Zhu et al., 2023; Wang et al., 2023; Yao et al., 2023). Various prompting methods have been investigated, such as few-shot, directional stimulus, generated knowledge, chain-of-thought, and tree-of-thought (Min et al., 2022; Li et al., 2023; Liu et al., 2022; Wei et al., 2022b; Yao et al., 2023). These new types of LLMs mark the beginning of a new era in the field of natural language processing.

While the capabilities of these LLMs continue to grow with an increasing number of parameters,

challenges are introduced for these models in training and testing steps to provide robust and generalized outputs (Rae et al., 2022; Smith et al., 2022; Chowdhery et al., 2022; Chung et al., 2022; Brown et al., 2020; Tay et al., 2023). To address this issue, PEFT methods such as LoRA have been introduced. These methods combine low-rank and trainable matrices with frozen weights in each layer of Transformer and even consider quantization (Hu et al., 2022; Dettmers et al., 2023).

As a related approach to priming, label embedding (Xiong et al., 2021; Zhang et al., 2021) can also incorporate label-related information into the input to enhance generation, as mentioned by Kwon et al. (2023b). However, in contrast to priming, label embedding cannot precisely control the generation itself and requires additional training.

To conduct the sentence compression task with instructions, we focus on priming that incorporates additional constraint-specific information to enhance performance, particularly for the length constraint, rather than just paraphrasing instructions to direct the task.

7 Conclusion

We proposed InstructCMP to conduct sentence compression by incorporating length constraints without model modifications. For this new approach, we constructed new evaluation datasets by transforming traditional sentence compression datasets into an instruction format, while we also created new training datasets. Additionally, we introduced “length priming” into the instructions and demonstrated its effectiveness in zero-shot and instruction-based fine-tuning settings on four benchmark datasets. We also conducted an in-depth analysis, including the model size and type.

Limitations

Although our length priming successfully compresses sentences, it might be challenging to consider it in document summarization, which requires considering multiple sentences. Therefore, it remains a topic for future studies. In the future, we will consider sentence relationships for prompting to summarize documents. Furthermore, there can be cases where keyword constraints are required for controllable summarization to take into account the content of summaries, which also remains a potential area for future investigation.

¹⁴<https://gemini.google.com/>

Acknowledgements

We would like to gratefully acknowledge the anonymous reviewers for their helpful comments and feedbacks.

References

- Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. [Jointly learning to extract and compress](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact-aware neural abstractive summarization. *AAAI’18/IAAI’18/EAAI’18*. AAAI Press.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pilla, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- James Clarke and Mirella Lapata. 2006. [Models for sentence compression: A comparison across domains, training requirements and evaluation measures](#). In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 377–384, Sydney, Australia. Association for Computational Linguistics.
- James Clarke and Mirella Lapata. 2008. Global inference for sentence compression : an integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. [GSum: A general framework for guided neural abstractive summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.
- Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. [Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation](#).
- Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. [Sentence simplification via large language models](#).
- Thibault Févry and Jason Phang. 2018. [Unsupervised sentence compression using denoising auto-encoders](#). In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

- Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. [Sentence compression by deletion with LSTMs](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.
- Katja Filippova and Yasemin Altun. 2013. [Overcoming the lack of parallel data in sentence compression](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.
- Demian Ghalandari, Chris Hokamp, and Georgiana Ifrim. 2022. [Efficient unsupervised sentence compression by fine-tuning transformers with reinforcement learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1280, Dublin, Ireland. Association for Computational Linguistics.
- Shun Hasegawa, Yuta Kikuchi, Hiroya Takamura, and Manabu Okumura. 2017. [Japanese sentence compression with a large training dataset](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Vancouver, Canada. Association for Computational Linguistics.
- Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. [CTRLsum: Towards generic controllable text summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chiori Hori and Sadaoki Furui. 2004. Speech summarization: An approach through word extraction and a method for evaluation. *IEICE Transactions*, 87-D:15–25.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Dave Hulbert. 2023. [Using tree-of-thought prompting to boost chatgpt’s reasoning](#). <https://github.com/dave1010/tree-of-thought-prompting>.
- Tatsuya Ishigaki, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2019. [Discourse-aware hierarchical attention network for extractive single-document summarization](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 497–506, Varna, Bulgaria. INCOMA Ltd.
- Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. [Quantization and training of neural networks for efficient integer-arithmetic-only inference](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hongyan Jing. 2000. [Sentence reduction for automatic text summarization](#). In *Sixth Applied Natural Language Processing Conference*, pages 310–315, Seattle, Washington, USA. Association for Computational Linguistics.
- Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2018. [Higher-order syntactic attention network for longer sentence compression](#). 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 1716–1726, New Orleans, Louisiana. Association for Computational Linguistics.
- Hidetaka Kamigaito, Katsuhiko Hayashi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2017. [Supervised attention for sequence-to-sequence constituency parsing](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 7–12, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Hidetaka Kamigaito and Manabu Okumura. 2020. [Syntactically look-ahead attention network for sentence compression](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8050–8057.
- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. [Controlling output length in neural encoder-decoders](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.
- Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. [Improving sentence compression by learning to predict gaze](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.
- Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization - step one: Sentence compression. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, page 703–710. AAAI Press.
- Jingun Kwon, Hidetaka Kamigaito, and Manabu Okumura. 2023a. [Abstractive document summarization with summary-length prediction](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 618–624, Dubrovnik, Croatia. Association for Computational Linguistics.

- Jingun Kwon, Hidetaka Kamigaito, Young-In Song, and Manabu Okumura. 2023b. [Hierarchical label generation for text classification](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 625–632, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jingun Kwon, Naoki Kobayashi, Hidetaka Kamigaito, and Manabu Okumura. 2021. [Considering nested tree structure in sentence extractive summarization with pre-trained transformer](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4039–4044, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. [Keywords-guided abstractive sentence summarization](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8196–8203.
- Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding large language models via directional stimulus prompting. *arXiv preprint arXiv:2302.11520*.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. [Generated knowledge prompting for commonsense reasoning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.
- Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. [Controlling length in abstractive summarization using a convolutional neural network](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119, Brussels, Belgium. Association for Computational Linguistics.
- Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. [Global optimization under length constraint for neural text summarization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048, Florence, Italy. Association for Computational Linguistics.
- Chanakya Malireddy, Tirth Maniar, and Manish Shrivastava. 2020. [SCAR: Sentence compression using autoencoders for reconstruction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 88–94, Online. Association for Computational Linguistics.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>.
- Yishu Miao and Phil Blunsom. 2016. [Language as a latent variable: Discrete generative models for sentence compression](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 319–328, Austin, Texas. Association for Computational Linguistics.
- Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. [Exploring BERT’s sensitivity to lexical cues using tests from semantic priming](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.
- Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. [Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text](#).
- Tong Niu, Caiming Xiong, and Richard Socher. 2019. [Deleter: Leveraging BERT to perform unsupervised successive text compression](#). *arXiv preprint, arXiv:1909.03223*.
- 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](#). *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is ChatGPT a general-purpose natural language processing task solver?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. page 9. OpenAI.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Mari-beth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen,

- Siddhant Jayakumar, Elena Buchatskaya, David Buden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsim-poukelli, Nikolai Grigorev, Doug Fritz, Thibault Sot-tiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. [Scaling language models: Methods, analysis insights from training gopher](#).
- Raphael Schumann, Lili Mou, Yao Lu, Olga Vechto-mova, and Katja Markert. 2020. [Discrete optimization for unsupervised sentence summarization with word-level extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042, Online. Association for Computational Linguistics.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. [Using deepspeed and megatron to train megatron-turing nlG 530b, a large-scale generative language model](#).
- Sho Takase and Naoaki Okazaki. 2019. [Positional encoding to control output sequence length](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. [U12: Unifying language learning paradigms](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-bert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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, Di-ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-stein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subrama-nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-lor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Ro-driguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Liangguo Wang, Jing Jiang, Hai Leong Chieu, Chen Hui Ong, Dandan Song, and Lejian Liao. 2017. [Can syntax help? improving an LSTM-based sentence compression model for new domains](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1385–1393, Vancouver, Canada. Association for Computational Linguistics.
- Liangguo Wang, Jing Jiang, and Lejian Liao. 2018. Sen-tence compression with reinforcement learning. In *Knowledge Science, Engineering and Management*, pages 3–15, Cham. Springer International Publish-ing.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large language models are not fair evaluators](#).
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormo-labashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Puro-hit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. [Super-NaturalInstructions: Generaliza-tion via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. As-sociation for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. [Finetuned language models are zero-shot learners](#). In *International Con-ference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022b. [Chain of thought prompting elicits reasoning in large language models](#). *CoRR*, abs/2201.11903.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. [Fusing label embedding into BERT: An efficient improvement for text classification](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [Discourse-aware neural extractive text summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#).

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2021. [A language model-based generative classifier for sentence-level discourse parsing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2432–2446, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Zhao, Zhiyuan Luo, and Akiko Aizawa. 2018. [A language model based evaluator for sentence compression](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 170–175, Melbourne, Australia. Association for Computational Linguistics.

Jiawei Zhou and Alexander Rush. 2019. [Simple unsupervised summarization by contextual matching](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy. Association for Computational Linguistics.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. [Controlled text generation with natural language instructions](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. [Multilingual machine translation with large language models: Empirical results and analysis](#).

A Performance in Instruction Selection

To determine task-specific instructions, we manually composed several candidates and evaluated their performances on the validation dataset from

Google. Table 11 shows the results. Based on their performances, we selected the 5th instruction as our base instruction for the setting without a length constraint.

B Performance in Varying Training Dataset Size

To investigate the impact of the training dataset size on performance, we also prepared 0.5% and 1% training datasets randomly sampled from the Google dataset. Table 12 shows the results of QLoRa fine-tuning for constraints in supervised settings. As observed, increasing the dataset size correlates with improved performance. Table 13 shows the results of an ablation study on “length priming”. Similarly, our “length priming” proves to be essential for performance improvements even in small datasets.

C Performance of the FLAN Models and Modified Instruction Templates

Table 14 shows the results for the FLAN models using instruction templates in Table 2. They did not effectively compress sentences, as denoted by ΔCR .

Table 15 shows the modified templates for the FLAN models.

#	Instruction	R-1	R-2	R-L	F ₁	Δ CR	novel
1	Sentence:\n{input}\nThe sentence without the non-essential words would be:\n	64.73	54.59	64.23	0.65	+35.76	0.36
2	Sentence:\n{input}\nThe compressed version of the original sentence without generating new words:\n	60.53	48.10	59.43	0.61	+38.09	1.12
3	Sentence:\n{input}\nCompress the sentence by removing the non-essential words:\n	62.37	50.95	61.61	0.62	+36.84	0.80
4	Sentence:\n{input}\nDelete the non-essential words by keeping the original meaning:\n	61.50	51.87	61.18	0.62	+43.23	0.35
5	Sentence:\n{input}\nThe sentence without the less important words would be:\n	66.99	56.79	66.54	0.68	+30.21	0.25
6	Original Sentence:\n{input}\nMake a new sentence without the non-essential words. New sentence would be:\n	64.23	52.26	62.79	0.65	+31.62	0.92
7	Sentence:\n{input}\nThe sentence without the unnecessary words would be:\n	63.85	53.33	63.29	0.64	+36.48	0.51

Table 11: Performances of different instructions using zero-shot InstructCMP based on the Llama2-chat-13B model on the validation dataset of Google.

Data	Size	Instruction	R-1	R-2	R-L	F ₁	Δ CR
Google	0.5%	#1	80.50	72.22	80.22	0.81	+1.49
		#3	83.56	75.33	82.97	0.84	-0.21
Broad	0.5%	#1	71.46	59.30	70.88	0.70	-14.37
		#3	80.62	68.34	79.31	0.79	-5.68
BNC	0.5%	#1	64.28	52.43	63.42	0.63	-19.59
		#3	73.49	60.88	72.07	0.72	-10.89
DUC	0.5%	#1	26.91	8.61	23.59	0.23	+3.06
		#3	26.15	8.07	23.25	0.22	+0.9
Google	1%	#1	81.68	73.55	81.40	0.83	+2.05
		#3	85.45	77.55	84.83	0.86	+0.46
Broad	1%	#1	72.54	60.57	72.04	0.71	-13.04
		#3	82.25	69.63	80.62	0.80	-3.38
BNC	1%	#1	64.63	52.80	63.74	0.64	-18.93
		#3	76.49	63.46	74.73	0.75	-6.64
DUC	1%	#1	27.69	8.95	24.24	0.24	+3.77
		#3	26.63	8.57	23.93	0.23	+1.73
Google	5%	#1	82.85	75.15	82.58	0.84	-1.28
		#3	86.88	79.55	86.26	0.88	-0.16
Broad	5%	#1	70.14	58.15	69.70	0.68	-15.88
		#3	82.63	69.76	81.16	0.81	-1.38
BNC	5%	#1	61.28	49.61	60.51	0.60	-24.21
		#3	77.54	64.38	76.00	0.76	-4.13
DUC	5%	#1	27.31	9.21	24.34	0.24	+0.28
		#3	26.83	8.57	23.96	0.23	+0.78

Table 12: Experimental results of InstructCMP using Llama2-13B-chat for different training dataset sizes on Google, Broad, BNC, and DUC.

Data	Size	Instruction	R-1	R-2	R-L	F ₁	Δ CR
Google	0.5%	#2	80.35	72.20	80.08	0.81	+1.79
		#3	83.56	75.33	82.97	0.84	-0.21
		#3-1	81.38	72.68	81.00	0.82	0.00
		#3-2	83.23	74.67	82.73	0.84	-0.64
Broad	0.5%	#2	72.09	59.95	71.59	0.71	-12.55
		#3	80.62	68.34	79.31	0.79	-5.68
		#3-1	76.87	64.49	76.32	0.76	-7.51
		#3-2	80.31	68.12	79.51	0.79	-4.98
BNC	0.5%	#2	64.89	52.96	63.95	0.64	-10.53
		#3	73.49	60.88	72.07	0.72	-10.89
		#3-1	70.81	58.33	69.81	0.70	-12.57
		#3-2	72.76	60.32	71.48	0.72	-0.85
DUC	0.5%	#2	27.16	9.02	23.95	0.23	+3.04
		#3	26.15	8.07	23.25	0.22	+0.90
		#3-1	25.51	8.16	22.68	0.22	-0.43
		#3-2	26.33	8.29	23.65	0.23	+0.15
Google	1%	#2	81.93	73.86	81.61	0.83	+0.07
		#3	85.45	77.55	84.83	0.86	+0.46
		#3-1	83.91	75.56	83.51	0.85	+0.82
		#3-2	85.18	77.07	84.55	0.86	-0.65
Broad	1%	#2	72.37	60.34	71.94	0.71	-12.90
		#3	82.25	69.63	80.62	0.80	-3.38
		#3-1	81.85	69.38	81.25	0.80	-0.67
		#3-2	81.17	68.81	79.91	0.79	-5.12
BNC	1%	#2	64.15	52.32	63.34	0.63	-19.47
		#3	76.49	63.46	74.73	0.75	-6.64
		#3-1	77.50	64.64	76.65	0.77	-1.77
		#3-2	74.27	61.72	72.73	0.73	-8.34
DUC	1%	#2	27.34	9.05	24.36	0.24	-0.22
		#3	26.63	8.57	23.93	0.23	+1.73
		#3-1	25.84	8.54	23.14	0.22	-1.32
		#3-2	26.14	8.16	23.38	0.23	-0.48
Google	5%	#2	84.99	77.43	84.69	0.86	+1.45
		#3	86.88	79.55	86.26	0.87	-0.16
		#3-1	85.20	77.46	84.72	0.86	+0.76
		#3-2	86.80	79.58	86.29	0.87	+0.12
Broad	5%	#2	80.34	67.77	79.81	0.78	-1.02
		#3	82.63	69.76	81.16	0.81	-1.38
		#3-1	82.80	70.39	82.05	0.81	+0.90
		#3-2	82.66	69.81	81.16	0.81	-1.08
BNC	5%	#2	73.74	61.52	72.92	0.72	-5.50
		#3	77.54	64.38	76.00	0.76	-4.13
		#3-1	77.62	64.58	76.45	0.77	-1.49
		#3-2	77.40	64.20	75.81	0.76	-4.03
DUC	5%	#2	27.20	8.98	24.27	0.24	+0.47
		#3	26.83	8.57	23.96	0.23	+0.78
		#3-1	26.25	8.27	23.49	0.23	-1.22
		#3-2	26.46	8.31	23.62	0.23	+1.32

Table 13: Ablation study of “length priming” for different training dataset sizes on Google, Broad, BNC, and DUC.

Data	Model	Instruction	R-1	R-2	R-L	F ₁	ΔCR
Google	T5-XXL	#1	58.43	49.61	58.42	0.59	+55.16
		#2	58.33	49.50	58.31	0.59	+54.70
		#3	58.51	49.69	58.49	0.59	+54.67
	UL2	#1	59.17	50.33	59.03	0.60	+51.05
		#2	58.59	49.68	58.52	0.59	+53.35
		#3	60.59	51.66	60.53	0.61	+49.27
Broad	T5-XXL	#1	85.11	72.65	85.11	0.85	+23.35
		#2	85.08	72.56	85.08	0.85	+22.85
		#3	85.22	72.77	85.21	0.85	+22.94
	UL2	#1	84.27	70.97	83.56	0.84	+17.59
		#2	83.84	70.78	83.76	0.84	+21.16
		#3	83.99	70.92	83.84	0.84	+19.95
BNC	T5-XXL	#1	81.43	68.46	81.43	0.82	+27.06
		#2	81.39	68.49	81.39	0.82	+27.18
		#3	81.42	68.51	81.40	0.81	+26.97
	UL2	#1	80.20	67.07	79.00	0.80	+19.77
		#2	80.29	67.01	80.03	0.80	+24.12
		#3	79.81	66.61	79.32	0.80	+21.92

Table 14: Experimental results for the zero-shot instruction-based FLAN models using instruction templates in Table 2.

#	Constraint	Instruction
1	χ	Sentence:\n{src }\nSummarize without the less important words would be:\n
2	Length w/o priming	Sentence:\n{src }\nSummarize without the less important {del} words would be:\n
3	Length	Sentence with {src len} words:\n{src }\nSummarize in {keep} words without the less important {del} words would be:\n

Table 15: Modified instruction templates for the FLAN models.