

RST-LoRA: A Discourse-Aware Low-Rank Adaptation for Long Document Abstractive Summarization

Dongqi Liu Vera Demberg

Department of Computer Science

Department of Language Science and Technology

Saarland Informatics Campus, Saarland University, Germany

{dongqi, vera}@lst.uni-saarland.de

Abstract

For long document summarization, discourse structure is important to discern the key content of the text and the differences in importance level between sentences. Unfortunately, the integration of rhetorical structure theory (RST) into parameter-efficient fine-tuning strategies for long document summarization remains unexplored. Therefore, this paper introduces **RST-LoRA** and proposes four RST-aware variants to explicitly incorporate RST into the LoRA model. Our empirical evaluation demonstrates that incorporating the type and uncertainty of rhetorical relations can complementarily enhance the performance of LoRA in summarization tasks. Furthermore, the best-performing variant we introduced outperforms the vanilla LoRA and full-parameter fine-tuning models, as confirmed by multiple automatic and human evaluations, and even surpasses previous state-of-the-art methods¹.

1 Introduction

The advent of pre-trained large language models (LLMs), such as LLaMA-2 (Touvron et al., 2023), Vicuna (Zheng et al., 2023), and GPT-related models from OpenAI (OpenAI, 2023), has greatly accelerated research progress of Natural Language Processing (NLP). With the continual growth in the scale of LLMs, the requirements for both software and hardware in order to fully fine-tune LLMs to adapt to downstream tasks, especially in processing long sequence data, will become increasingly demanding (Gu et al., 2022; Pu et al., 2024).

Parameter-Efficient Fine-Tuning (PEFT) strategies are noteworthy in mitigating the aforementioned problem by reducing the number of parameters that need to be adjusted (Chen et al., 2022; AkbarTajari et al., 2022; Mao et al., 2022; Gheini et al., 2023; Badola et al., 2023; Zhang et al., 2023b;

Lawton et al., 2023). Some studies have highlighted that by updating only 0.01–1% of the (additional) parameters and freezing all other parameters of LLMs, PEFT methods can match or even exceed the performance of vanilla full-parameter fine-tuning (Li and Liang, 2021; Hu et al., 2022; Asai et al., 2022; Yang et al., 2022; Gu et al., 2023; Liao et al., 2023; Zhang et al., 2023b; Li et al., 2023; Lei et al., 2023; Zhang et al., 2023a; Chen et al., 2023; Lawton et al., 2023). Among these methods, LoRA algorithm (Low-Rank Adaptation, Hu et al., 2022) has achieved state-of-the-art (SOTA) performance due to its ability to circumvent the latency associated with adapter tuning (Houlsby et al., 2019) as well as the input length constraints of prefix/prompt tuning (Li and Liang, 2021; Lester et al., 2021) during model training and inference (He et al., 2022; Ghazvininejad et al., 2022; Dettmers et al., 2023; Zhang et al., 2023a; Whitehouse et al., 2023; Ding et al., 2023).

Recent investigations (Üstün and Cooper Stickland, 2022; Ponti et al., 2023; Zhao et al., 2023; Zeng et al., 2023; Zhang et al., 2023b; Wan et al., 2023; Liu et al., 2023a) have revealed that PEFT strategies face challenges in distinguishing latent text relations and determining the importance level of different sentences during fine-tuning. This issue arises because such distinctions are not a primary focus in PEFT’s learning process and are not explicitly represented in the input data. However, this is essential for the task of long document summarization since generating a good summary often requires natural language generation (NLG) models to have the ability to discern salient information within the text and comprehend the intricate interrelations among different text components.

Our approach proposed here takes inspiration from Ishigaki et al. (2019); Xiao et al. (2020); Dong et al. (2021); Cao and Wang (2022); Guo et al. (2022); Pu et al. (2023), who have advised that explicitly integrating document structure and/or

¹The project information is available at <https://dongqi.me/projects/RST-LoRA>.

discourse knowledge can enhance the performance of neural summarization models when fully fine-tuning the NLG models. This motivates us to investigate the following research questions: *Can the Rhetorical Structure Theory (RST, Mann and Thompson, 1987) improve the performance of LoRA strategy in summarizing lengthy documents?* Specifically, we want to explore and verify whether infusing RST knowledge into LoRA can improve the performance of long document summarization. To answer this question, this paper will propose, introduce, and integrate four RST structure variants to guide the training of LoRA. These variants include (i) binary (ii) probability RST distribution, both with (w) and without(w/o) relation labels.

In summary, our contributions are as follows:

- We introduce a method for injecting discourse knowledge into the training of the LoRA model. Our approach is compatible with both Seq2Seq and GPT transformer-based architectures, allowing for its easy adoption across different LLMs.
- Our empirical findings demonstrate that discourse uncertainty and relation labels are complementary, and both can contribute to the improvement of final performance. Notably, our model also outperforms current SOTA full-parameter fine-tuning (FFT) models in specific evaluation metrics.
- We offer quantitative and qualitative analyses showing that our model surpasses baseline models in factual consistency checking. Moreover, the results of human evaluation and GPT-4 examination reveal that our model produces summaries that are closer in quality to those generated by humans.

2 RST Prerequisite Knowledge

Rhetorical Structure Theory (RST) is a discourse framework that is helpful for determining which sentences in a document should or should not be included in a summary (Marcu, 1997, 1999, 2000; Kikuchi et al., 2014; Goyal and Eisenstein, 2016; Pu et al., 2023). To be specific, RST delineates a set of coherence relations between text segments, known as Elementary Discourse Units (EDUs), at the document level (e.g., one segment might provide clarification for another, or conversely, two segments could present contrasting viewpoints). Moreover, RST categorizes EDUs based on their discourse importance, labeling central EDUs as ‘nuclei’, and less central EDUs as

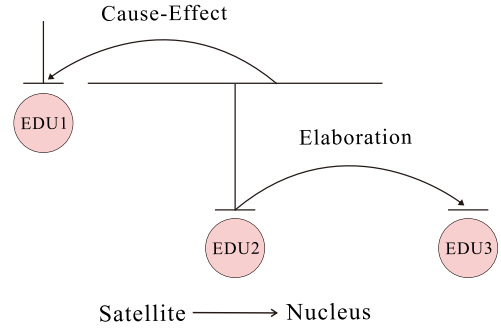


Figure 1: An example of RST tree: [Utilizing discourse structure to enhance text summarization is beneficial.]^{EDU1} [This technique can be used to identify key ideas and capture often overlooked nuances.]^{EDU2} [Accurate capture of these complex structures facilitates the generation of good summaries.]^{EDU3}

‘satellites’ (Marcu, 1999; Bosselut et al., 2018; Isonuma et al., 2019; Xu et al., 2020).

For example, consider an RST tree as depicted in Figure 1. In this instance, EDU1 serves as the most pivotal component within the entire example, thus constituting the nucleus for both EDU2 and EDU3. EDU2 is tasked with elucidating and providing supplementary information to EDU3, positioning it as a satellite unit in relation to EDU3. Given the relatively diminished importance of EDU2, merging EDU1 and EDU3 while pruning EDU2, the semantic essence of the example would remain intact. In a more extreme scenario, retaining only EDU1 as the summary sentence and omitting both EDU2 and EDU3, the primary information conveyed by the entire example would still be preserved. As has also been argued by, e.g., Marcu (1997); Louis et al. (2010); Cohan et al. (2018); Liu et al. (2019); Li et al. (2020); Xu et al. (2020); Dong et al. (2021); Chen and Yang (2021) that ‘satellite’ EDUs play a subordinate role in summarization, with most summary sentences deriving from ‘nuclei’.

3 Related Work

3.1 Document Summarization with RST

RST is a linguistic discourse framework that provides a way to organize text into a hierarchical tree structure, which helps to better understand the overall organization and inter-part relations of text. Early research by Marcu (1997) and Louis et al. (2010) uncovered that human-written summaries often align with the nucleus EDUs in RST trees. This correlation underscores the validity of RST as a theoretical motivation for summarization tasks. Building on this insight, subsequent studies have

demonstrated the value of explicitly incorporating RST trees into neural summarization models. For example, [Kikuchi et al. \(2014\)](#) boosted the summarization performance of the RNN-based model by constructing RST trees, where satellite EDUs were pruned to retain only the nucleus EDUs, thus focusing on the document’s key content. Pre-trained language models have a noted tendency to capture some superficial aspects of discourse relations without explicit training ([Miaschi et al., 2020](#); [Qian et al., 2021](#); [Schuster and Linzen, 2022](#)), but the latent discourse information is often not captured correctly. To alleviate this challenge, [Xu et al. \(2020\)](#) and [Dong et al. \(2021\)](#) enhanced summarization models by incorporating discourse structure within transformer-based and graph neural network models, respectively.

More recently, [Pu et al. \(2023\)](#) proposed an approach that incorporates the uncertainty of RST structures into the attention mechanisms of summarization models and achieved SOTA results on multiple datasets. However, all the above approaches require full fine-tuning of NLG models, which is very expensive. As the model parameters increase, this issue will be further amplified. Incorporating RST into PEFT might potentially lower the barrier to fine-tuning by structuring the learning process around the inherent rhetorical patterns in the data.

3.2 Document Summarization with LoRA

LoRA, presented by [Hu et al. \(2022\)](#), is a low-rank approximation strategy that reduces the number of trainable parameters by freezing the pre-trained model weights and injecting trainable rank decomposition matrices into each layer of the transformer architecture. The initial research also demonstrated through summarization tasks that applying LoRA on the GPT-3 model ([Brown et al., 2020](#)) with less than 1% of the parameters could even outperform FFT. Expanding on this, studies by [Dettmers et al. \(2023\)](#), [Xu et al. \(2024\)](#), and [Li et al. \(2024\)](#) enhanced generalization ability in downstream summarization tasks by quantifying LoRA matrices and adopting mixed-precision techniques. Furthermore, [Zhu et al. \(2023\)](#) combined LoRA with layer pruning, achieving notable improvements in specialized applications like medical report summarization. Recently, [Liao et al. \(2023\)](#) validated the feasibility of using task-neutral sparse masks to improve the performance in text summarization with LoRA.

In a similar work, [Ghazvininejad et al. \(2022\)](#) in-

tegrated hierarchical document structure (i.e. blocking structure) into prefix-tuning to simulate the high-level discourse relation and achieved improvements in the task of text generation. However, there is still an unexplored potential in explicitly integrating fine-grained RST structures into the summarization process with PEFT methods, since comprehending the coherence of discourse elements could positively impact the quality of generated summaries ([Li et al., 2016](#); [Liu and Chen, 2019](#); [Huang and Kurohashi, 2021](#)), particularly in the context of summarizing long documents ([Cohan et al., 2018](#); [Xu et al., 2020](#); [Li et al., 2020](#); [Gabriel et al., 2021](#); [Dong et al., 2021](#); [Balachandran et al., 2022](#); [Pu et al., 2023](#)).

4 Proposed Approach

A limitation observed with LoRA and other PEFT methods during the fine-tuning phase is that, while their primary function is to act as a low-rank approximation of the weight matrices in LLMs, they do not adequately capture textual context relations ([He et al., 2022](#); [Ghazvininejad et al., 2022](#); [Wan et al., 2023](#); [Zhang et al., 2023b](#)). One of the reasons is that LoRA is not driven or guided by discourse knowledge during the training phase, because this part of knowledge is not explicitly present in the input data ([Üstün and Cooper Stickland, 2022](#); [Zhao et al., 2023](#)). In addition, the matrices obtained by low-rank approximation strategies may have more difficulty in capturing complex textual discourse relations due to the smaller semantic space that can be expressed when compared to LLMs’ weight matrices ([Wan et al., 2023](#); [Zhang et al., 2023b](#); [Tomanek et al., 2021](#)). Hence, we propose a method that directly and explicitly incorporates discourse architecture into LoRA. This approach allows LoRA’s adaptation mechanism for downstream tasks to discern intricate text relations through soft guidance, which can leverage contextual discourse connections and steer the learning trajectory toward a discourse-informed summarization process.

4.1 RST Distribution

Our approach builds upon prior practice ([Chen et al., 2017](#); [Xiao et al., 2020](#); [Bugliarello and Okazaki, 2020](#); [Pu and Sima’an, 2022](#); [Pu et al., 2023](#); [Zhao et al., 2023](#)) of integrating linguistic structures (such as syntactic structure, discourse structure, etc.) into neural NLG models. To in-

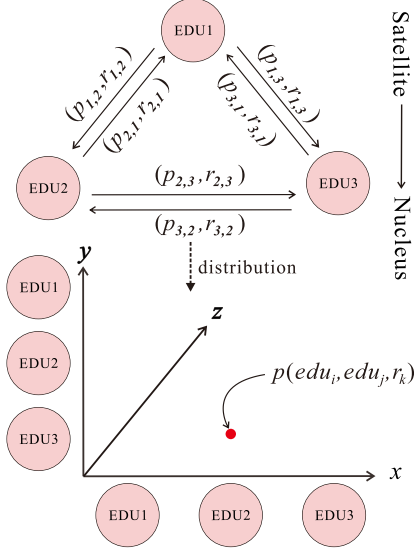


Figure 2: RST distribution

fuse discourse structure into LoRA, we begin by converting RST structures, generated by an RST parser² into a compact matrix representation. Figure 2 exemplifies how to transmute the full potential RST structures (n-best RST forests) into a three-dimensional discourse matrix (Pu et al., 2023). In this matrix, the x and y axes correspond to the elementary discourse units (EDUs) within the source document, while the z -axis denotes the discourse relation label³. Each point of the matrix indicates the probability value $p(edu_i, edu_j, r_k) \in [0, 1] \subseteq \mathbb{R}$ that edu_i is the nucleus of edu_j with discourse relation r_k . It should be noted that $\forall i = j, p(edu_i, edu_j, r_k) = 0$, since no unit is self-dependent. Next, we average and merge the y -axis of the matrix, and the merged value $c(edu_i, \overline{edu_j}, r_k)$ is called the importance index of edu_i with relation r_k . The *RST* distribution is then obtained by combining all $c(edu_i, \overline{edu_j}, r_k)$. Based on this, we propose four fine-grained RST matrix distributions:

- RST_{wo}^b : A binary, label-agnostic representation collapsing probabilities into a simple 1-or-0 regarding discourse connections.
- RST_w^b : An extension of the binary distribution that includes relation labels, enriching the binary decisions with relational types.
- RST_{wo}^p : A probabilistic representation that omits labels, focusing instead on the probabilities to

express uncertainty in discourse connections.

- RST_w^p : The most granular representation, retaining both types of discourse relations and their probabilistic weights for a full-fledged representation of discourse nuances.

The inclusion of the relation label is contingent on whether we perform average-and-merge along the relation dimension (z -axis). Whether the approach is binary or based on uncertainty hinges on whether we replace the probability value with 1 or 0. In the binary cases, probabilities equal to or above 0.5 are replaced with 1, else with 0. Previous researchers (such as Xu et al. (2020) and Dong et al. (2021)) considered the 1-best tree, representing binary relations outputted from parsers into summarization models (also the case of our first two variants). The latter two variants utilize the parser’s output probabilities as confidence indicators for discourse connections (Pu et al., 2023). This probabilistic approach softens the impact of potential errors or ambiguities in the parser’s output, blending its uncertainty into the model.

4.2 RST-Aware Injection

In the process of vanilla LoRA fine-tuning, let $W_{A \times B}^{fine-tuned}$ denote the fine-tuned LLM’s parameters, and $W_{A \times B}^{pre-trained}$ represent the parameters before fine-tuning. The change in parameters is represented by $\Delta W_{A \times B}$, where A and B correspond to the dimensions of the parameter matrix:

$$W_{A \times B}^{fine-tuned} = W_{A \times B}^{pre-trained} + \Delta W_{A \times B}$$

In other words, the parameters after fine-tuning can be obtained by adding a matrix representing the variation to the parameters of the original, pre-fine-tuned model.

$$\Delta W_{A \times B} \simeq \Phi[(W_{A \times r}^{down} W_{r \times B}^{up})]$$

$$r \ll \min(A, B)$$

The objective of the LoRA strategy aims to learn the mapping method Φ that can provide an approximation of the matrix representing parameter variations (Hu et al., 2022). Typically, the rank value r is considerably smaller than both A and B , so that the total number of parameters of $W_{A \times r}^{down}$ and $W_{r \times B}^{up}$ is significantly smaller than $W_{A \times B}$. For a given input document X to the linear projection in the model’s hidden layer, LoRA modifies the projection output (hidden representation) h as follows:

²https://github.com/seq-to-mind/DMRST_Parser

³Appendix A details the grouping of discourse relations.

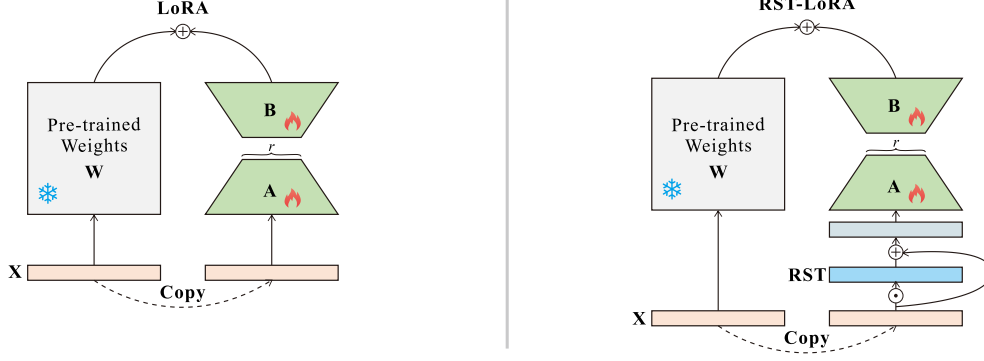


Figure 3: Model architecture: The diagram illustrates the integration of the RST matrix into the LoRA model. The left side is the original LoRA, while the right side depicts our proposed method RST-LoRA.

$$h \leftarrow h + X(W_{A \times r}^{down} W_{r \times B}^{up})$$

In its current form, LoRA treats both satellite and nucleus EDUs in documents equally and only recognizes their difference during the back-propagation process. This issue is also noted in the analyses by Ghazvininejad et al. (2022); Zhao et al. (2023), who also discovered that PEFT faces challenges in understanding the complex relations between sentences and the differences in importance level between text segments during its learning process. Therefore, we soft-guide the learning process by injecting the RST structure (i.e., the matrix presentation mentioned above) into the text embedding matrix of LoRA, as shown in Figure 3. Specifically:

$$h \leftarrow h + [(X \odot (1 + \gamma))(W_{A \times r}^{down} W_{r \times B}^{up})]$$

Here, γ denotes the weight coefficient matrix, or more precisely, the RST distribution matrix. The operation \odot signifies element-wise multiplication, and the motivation behind employing element-wise multiplication is that it can significantly amplify the impact of probability values on the input X matrix, creating an RST-injected matrix with greater distributional variance, in contrast, element-wise addition would exert a lesser impact on X . It should be noted that RST parser operates at the EDU level, meaning that sub-word units within the same EDU share the same multiplication factor, embedding the same probability value across the entire EDU into X . The estimates of learned parameters $W_{A \times r}^{down}$ and $W_{r \times B}^{up}$ are adjusted to match the utility of discourse knowledge for the ultimate summarization purpose. Each element of γ is constrained to be non-negative. The operation of $1 + \gamma$ functions as a

residual connection, allowing discourse knowledge to exert a subtle influence on the adjustment of the low-rank weight matrix. If we set all elements of γ to a uniform value δ , including zero, the adjustment to the low-rank matrices would revert to the conventional LoRA approach.

5 Experiments and Analysis

5.1 Experimental Settings

Datasets Our experiments are conducted on three recent long document summarization datasets: Multi-LexSum (ML, Shen et al., 2022), eLife (Goldsack et al., 2022) and BookSum Chapter (BC, Kryscinski et al., 2022). These datasets are sourced from the fields of legal documents, scientific papers, and books, respectively. We select these datasets because they exhibit a high degree of heterogeneity, and we want to test whether our proposed approach could maintain sufficient generalization performance across different data domains. Statistics of these datasets are provided in Appendix B.

Parser For automatic parsing of source documents, we employ DMRST parser (Liu et al., 2020, 2021) which enables us to extract probabilities or uncertainties of discourse relations and type labels from its final logits layer.

Automatic Metrics Aligning with previous work for evaluating summarization systems (Narayan et al., 2018; Liu et al., 2023c; Blinova et al., 2023), we use F1 scores of Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum) (Lin, 2004), BERTScore (Zhang et al., 2020), METEOR (Banerjee and Lavie, 2005), sacreBLEU (Post, 2018) and NIST (Lin and Hovy, 2003) for model’s performance evaluation. A description of these metrics can be found in Appendix C.

Training & Inference We operate Longformer (Beltagy et al., 2020) and Vicuna13B-16k (Zheng et al., 2023) as our baseline backbone models. Longformer is a state-of-the-art, open-source model optimized for handling long documents under Seq2Seq architecture. Meanwhile, Vicuna is another SOTA model based on GPT architecture. Our objective in using these models is to demonstrate the generalizability of our strategy across different architectural frameworks. We also include GPT-4 (OpenAI, 2023) as one of our comparative models. It should be noted that for GPT-4, we use both zero-shot learning (ZS) and in-context learning (ICL) with demonstrations from two randomly selected samples from the training datasets⁴. Besides, we compare our results with both the original full parameter fine-tuning (FFT) and the vanilla LoRA fine-tuning. All open-source models, including the baseline, proposed, and ablation models, adhere to identical hyperparameter settings. These settings are elaborated in Appendix D.

5.2 Experimental Results

General Results The differences in performance of different RST variants are shown in Table 1. Among our proposed RST-injected variants, models integrating discourse relation labels generally outperformed those without this integration. Similarly, models considering the uncertainty in discourse relations fare better than those disregarding it. This suggests that integrating parser uncertainty and coherence labels into the model improves the robustness of the model against potential misinformation to a certain extent when compared to the parser’s 1-best binary decisions.

Table 2 shows the performance differences between our final strategy (the best RST variant) and other comparative models. Specifically, GPT-4 exhibits the poorest overall performance, attributable to a lack of parameter tuning. The performance of the models based on Vicuna as backbone is overall better than the models based on Longformer due to the larger number of parameters. Regarding parameter-efficient settings, vanilla LoRA’s performance is marginally lower than FFT across most datasets, except eLife. However, LoRA achieves comparable results to FFT while only requiring adjustments of 0.25% of parameters for Longformer and 0.05% for Vicuna, highlighting LoRA’s efficiency.

⁴Prompts can be found in Appendix E.1 and E.2.

Data	Model	R1 _{f1} ↑	R2 _{f1} ↑	RL _{f1} ↑	RLsum _{f1} ↑
Multi-LexSum	Longformer _{RST_{wo}^b-LoRA}	45.82	21.32	23.81	43.40
	Longformer _{RST_w^b-LoRA}	46.02	21.34	23.87	43.39
	Longformer _{RST_{wo}^p-LoRA}	46.21	21.54	24.09	43.37
	Longformer _{RST_w^p-LoRA}	46.33	21.86	24.11	43.58
	Vicuna _{RST_{wo}^b-LoRA}	46.32	21.64	24.22	43.32
	Vicuna _{RST_w^b-LoRA}	47.33	22.70	24.25	43.31
	Vicuna _{RST_{wo}^p-LoRA}	47.39	22.79	24.35	43.33
	Vicuna _{RST_w^p-LoRA}	47.45	23.19	24.39	44.02
eLife	Longformer _{RST_{wo}^b-LoRA}	49.34	14.24	21.34	46.74
	Longformer _{RST_w^b-LoRA}	49.41	14.39	21.29	46.79
	Longformer _{RST_{wo}^p-LoRA}	49.87	14.49	21.83	47.15
	Longformer _{RST_w^p-LoRA}	49.89	14.68	22.11	47.64
	Vicuna _{RST_{wo}^b-LoRA}	48.73	14.68	21.89	47.11
	Vicuna _{RST_w^b-LoRA}	49.72	14.72	22.03	47.02
	Vicuna _{RST_{wo}^p-LoRA}	49.87	14.79	22.21	48.10
	Vicuna _{RST_w^p-LoRA}	49.92	14.92	22.41	48.21
BookSum Chapter	Longformer _{RST_{wo}^b-LoRA}	34.70	10.22	20.39	34.21
	Longformer _{RST_w^b-LoRA}	34.72	10.19	20.41	34.87
	Longformer _{RST_{wo}^p-LoRA}	35.29	11.38	21.62	35.11
	Longformer _{RST_w^p-LoRA}	35.40	11.76	21.88	35.27
	Vicuna _{RST_{wo}^b-LoRA}	37.28	12.35	22.13	38.33
	Vicuna _{RST_w^b-LoRA}	37.41	12.66	22.51	38.40
	Vicuna _{RST_{wo}^p-LoRA}	37.87	13.10	22.77	39.69
	Vicuna _{RST_w^p-LoRA}	37.92	13.24	22.93	40.31

Table 1: Performance of different RST variants

We also observe consistent performance improvements in LoRA when integrating RST structure into its training process without increasing the number of fine-tunable parameters, and in most cases even exceeds the FFT model. Our final model RST_w^p-LoRA, integrates both discourse relation types and uncertainty into LoRA’s training, achieving the best experimental outcomes. It also defeats SOTA models (fully fine-tuned with complicated strategies) on some metrics, including the current most advanced model (Pu et al., 2023) that incorporates RST structure to improve summarization performance.

Ablation Results To further assess the impact of the RST matrix on model performance, we specify three additional control conditions:

- **RST_{Even}**: In the RST matrix, we set values to 1 at even positions and 0 at odd positions.
- **RST_{Odd}**: We assign values of 1 at odd positions and 0 at even positions in the RST matrix.
- **RST_{Random}**: We assign random values $\in [0, 1] \subseteq \mathbb{R}$ to the RST matrix without considering the probability of discourse relations.

In ablation experiments, we use Vicuna as backbone for testing. The motivation behind setting these three ablation conditions is to simulate the extreme scenario where the RST parser completely fails to deliver valuable discourse information. Table 3 indicates that different ablation integration

Dataset	Model	# Trainable Parameters	R1 _{f1} ↑	R2 _{f1} ↑	RL _{f1} ↑	RLsum _{f1} ↑	BERTscore _{f1} ↑	Meteor↑	sacreBLEU↑	NIST↑
Multi-LexSum	Longformer _{FFT}	0.44B	45.81	21.32	23.71	43.25	87.21	33.30	12.06	2.23
	Longformer _{LoRA}	1.13M	45.78	21.30	23.65	43.12	87.31	33.31	12.00	2.28
	Longformer _{RST_w^p-LoRA}	1.13M	46.33 ^{†‡}	21.86 ^{†‡}	24.11 ^{†‡}	43.58 ^{†‡}	92.01 ^{†‡}	34.55 ^{†‡}	13.11 ^{†‡}	3.21 ^{†‡}
	Vicuna _{FFT}	13B	46.40	21.88	24.15	43.28	90.02	33.19	13.56	3.32
	Vicuna _{LoRA}	6M	46.32	21.76	24.09	43.14	89.45	33.22	13.44	3.31
	Vicuna _{RST_w^p-LoRA}	6M	47.45 [†]	23.19 ^{†‡}	24.39^{†‡}	44.02^{†‡}	93.89^{†‡}	35.31^{†‡}	14.02^{†‡}	4.11^{†‡}
	GPT-4 _{ZS}	-	38.74	13.39	18.26	37.67	60.91	24.24	7.43	1.55
	GPT-4 _{ICL}	-	42.14	15.27	20.37	40.12	71.32	28.14	10.22	1.90
	Pu et al. (2023)	-	46.42	22.89	-	43.98	86.70	33.94	-	-
	Shen et al. (2022)	-	53.73	27.32	-	30.89	42.01	-	-	-
eLife	Longformer _{FFT}	0.44B	47.59	13.58	20.75	45.25	85.50	28.21	6.86	2.90
	Longformer _{LoRA}	1.13M	48.31	13.69	21.10	45.80	85.63	28.18	7.05	3.12
	Longformer _{RST_w^p-LoRA}	1.13M	49.89 ^{†‡}	14.68 ^{†‡}	22.11 ^{†‡}	47.64 ^{†‡}	87.64 ^{†‡}	31.23 ^{†‡}	7.78 ^{†‡}	3.79^{†‡}
	Vicuna _{FFT}	13B	48.32	14.06	21.31	45.57	85.71	30.28	7.00	2.91
	Vicuna _{LoRA}	6M	48.41	14.32	21.40	46.01	86.06	31.00	6.62	2.88
	Vicuna _{RST_w^p-LoRA}	6M	49.92^{†‡}	14.92^{†‡}	22.41^{†‡}	48.21^{†‡}	87.81^{†‡}	33.22^{†‡}	8.15^{†‡}	3.42 ^{†‡}
	GPT-4 _{ZS}	-	42.73	9.05	17.93	40.15	61.21	25.13	3.47	2.32
	GPT-4 _{ICL}	-	44.62	11.35	20.03	44.09	73.23	27.36	5.66	2.45
	Tang et al. (2023)	-	35.22	9.73	-	32.33	-	-	-	-
	Pu et al. (2023)	-	48.70	14.84	-	46.13	84.70	29.53	-	-
BookSum Chapter	Longformer _{FFT}	0.44B	34.68	10.02	20.35	33.71	81.02	27.30	3.32	1.62
	Longformer _{LoRA}	1.13M	34.63	9.96	20.22	33.79	81.33	27.32	3.55	1.86
	Longformer _{RST_w^p-LoRA}	1.13M	35.40 ^{†‡}	11.76 ^{†‡}	21.88 ^{†‡}	35.27 ^{†‡}	83.99 ^{†‡}	29.03 ^{†‡}	5.94^{†‡}	2.02 ^{†‡}
	Vicuna _{FFT}	13B	37.21	12.38	22.07	38.21	82.31	28.01	3.45	1.70
	Vicuna _{LoRA}	6M	37.30	12.26	21.84	38.23	82.23	27.83	3.34	1.68
	Vicuna _{RST_w^p-LoRA}	6M	37.92 ^{†‡}	13.24^{†‡}	22.93^{†‡}	40.31^{†‡}	84.12 ^{†‡}	29.22^{†‡}	5.48 ^{†‡}	2.32^{†‡}
	GPT-4 _{ZS}	-	35.25	7.46	17.52	34.23	58.56	26.50	3.36	1.54
	GPT-4 _{ICL}	-	37.42	10.06	19.49	36.11	79.56	27.56	3.52	1.72
	Pu et al. (2023)	-	34.02	10.28	-	32.87	85.30	27.47	-	-
	Cao and Wang (2023)	-	41.11	10.63	-	40.20	-	-	-	-
	Scirè et al. (2023)	-	42.13	10.53	16.75	-	-	-	-	-

Table 2: Model performance. The bold numbers represent the best results concerning the given test set. [†] and [‡] indicate statistical significance ($p < 0.05$) of our final model (RST_w^p-LoRA) against the FFT and LoRA model via paired t-test based on the same backbone respectively. *FFT* for full fine-tuning, *ZS* for zero-shot learning and *ICL* for in-context learning. Each result of the SOTA models is directly replicated from their original papers.

Dataset	Model	R1 _{f1} ↑	R2 _{f1} ↑	RL _{f1} ↑	RLsum _{f1} ↑
ML	RST _{Even}	46.21	21.39	23.66	42.55
	RST _{Odd}	46.26	21.37	23.82	42.90
	RST _{Random}	46.30	21.73	24.07	43.10
eLife	RST _{Even}	47.10	14.28	20.86	45.33
	RST _{Odd}	47.04	14.20	20.98	45.31
	RST _{Random}	47.32	14.29	21.36	45.71
BC	RST _{Even}	37.09	12.20	21.75	38.06
	RST _{Odd}	37.01	12.18	21.72	38.10
	RST _{Random}	37.27	12.23	21.80	38.19

Table 3: F1 scores for ablation study

strategies not only fail to enhance the model’s performance but even detract from it. Experiments by introducing random noise exhibit that these arbitrary values reduce the model’s performance to a level marginally lower than the original LoRA. Furthermore, this also implies that when the RST parser fails to provide meaningful knowledge (as in the case of random noise), the impact of noise on the performance of the model is limited.

5.3 Analysis

Hallucination Checking We delve deeper into the level of factual consistency of the generated

summaries, which we test using the SummaC method (Laban et al., 2022). The score of SummaC ranges from 0 to 1, and the higher the score, the better the consistency. The results of the assessment using Vicuna as backbone are depicted in Figure 4. We observe that GPT-4 exhibits the weakest factual consistency, while the original LoRA also shows a comparatively lower level of factual accuracy than FFT. However, explicitly incorporating RST structure into LoRA mitigates the issue of hallucinations/inaccuracies in generated summaries, achieving better results than FFT model.

Impact of Different Rank r Figure 5 and Figures 6, 7 in Appendix F illustrate the impact of different ranks on model performance (Vicuna backbone). Across different datasets, the RST-aware model consistently outperforms the original LoRA at various ranks and achieves similar performance as the FFT model at lower ranks. Furthermore, a larger rank r will help to improve the performance of the model, which is also aligned with the findings of He et al. (2022); Zhang et al. (2023a). However, a higher rank correlates with an increased

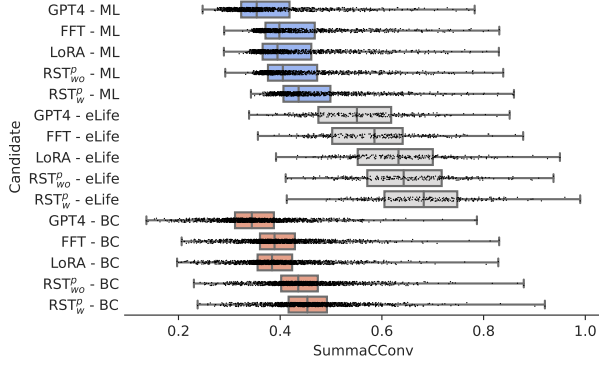


Figure 4: Factual consistency analysis

number of parameters requiring adjustment. Importantly, $r = 8$ is a trade-off point between performance gain and computational cost, when r continues to increase, the gain rate of performance improvement begins to slow down.

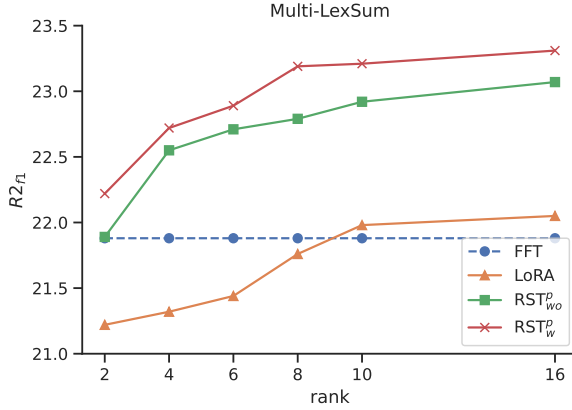


Figure 5: Impact of different r on ML dataset

Impact of Parser Capability To rigorously evaluate the parser’s impact on our method, we conduct an experiment that involves intentionally altering the RST parser’s output. This is designed to simulate varying levels of parser performance instability, thereby allowing us to observe its influence on our model’s efficacy. Specifically, we introduce random masking to the parser’s output at incremental thresholds of 10%, 20%, 40%, and 80%, assigning random values within the range of 0 to 1 to portions of the RST matrix. Table 4 presents the findings from this experiment, with Vicuna serving as backbone for RST_W^p -LoRA model on the Multi-LexSum dataset.

These results illustrate the direct correlation between the RST parser’s performance and the performance of our model’s output. Notably, even under conditions of compromised parser performance (with up to 20% of the information being

Model	$R1_{f1}\uparrow$	$R2_{f1}\uparrow$	$RL_{f1}\uparrow$	$RLsum_{f1}\uparrow$
RST_10%	47.33	23.01	24.33	43.45
RST_20%	47.09	22.78	24.23	43.37
RST_40%	46.52	21.76	24.13	43.20
RST_80%	46.32	21.75	24.06	43.15

Table 4: Impact of random masking on the parser

randomly masked), our model still demonstrates a good capacity to enhance summary generation quality by leveraging the learned discourse structure knowledge. However, it is observed that when the level of induced noise surpassed 40%, the negative impact became pronounced, relegating the model’s performance to levels akin to that of the original vanilla LoRA.

Human Evaluation To better analyze the quality of the summaries generated by the models, we randomly select 10 instances from the BookSum dataset and conduct a human evaluation. The evaluators we have recruited are graduate and doctoral candidates with specializations in Computer Science or Computational Linguistics, each possessing advanced proficiency in English. They receive compensation at the University’s established hourly rate. Evaluators are asked to read the corresponding original document, as well as five candidate summaries (from FFT, LoRA, and RST_W^p -LoRA with Vicuna backbone, GPT-4 and human). The human evaluators are blind to the condition, i.e. they do not know which summary comes from which system (or human author). Each sample is independently evaluated by three distinct human raters (thus 150 evaluation samples in total). Evaluators should rate the candidate summaries on a scale of 1 to 5 for relevance (R), informativeness (I), conciseness (C), and faithfulness (F), with a higher score indicating better quality. They also need to give an overall ranking of the five summaries. The detailed guidelines for human evaluation are available in Appendix F. The results, presented in Table 5, show the average values for each metric, as well as the proportions of times each model’s output is considered the best or worst among the candidates. The scores of Fleiss’ Kappa coefficient for R, I, C, and F are 0.812, 0.705, 0.683, and 0.688, respectively, with an average score of 0.722, indicating substantial agreement.

From Table 5, it is evident that human-generated summaries surpass all neural summarization models in terms of quality. Among the four neural

Candidate	R	I	C	F	Best	Worst
Human	4.70	4.83	4.53	4.67	83.3%	0.0%
GPT-4 _{ICL}	3.76	2.27	3.25	2.33	0.0%	56.7%
Vicuna _{LoRA}	4.03	2.37	3.20	2.50	0.0%	20.0%
Vicuna _{FFT}	4.27	2.57	3.67	2.77	6.67%	13.3%
Vicuna _{RST_w^P-LoRA}	4.53	3.90	4.03	3.17	13.3%	10.0%

Table 5: Human evaluation results

models, GPT-4 shows the least performance, with LoRA coming in second, having a 20% probability of being rated as the worst. The FFT model fares slightly better than the LoRA model. The RST_w^P-LoRA model outperforms other neural summarization systems across all metrics, and its average scores on some indicators approach the level of human performance. Moreover, compared to other neural summarization systems, the RST_w^P-LoRA model is more likely to be recognized for producing the highest quality summaries and less likely to be considered as generating the poorest quality summaries.

GPT-4 Evaluation Inspired by Liu et al. (2023b), we engage GPT-4 to assess our candidate models using the same guidelines as our human evaluators. To ensure experimental consistency, all experiments use the identical hyper-parameters settings detailed in Appendix D. To avoid potential biases from previous interactions, we reset the conversation history prior to each query and abstain from making any further modifications. In our initial investigation, we aim to explore the extent to which GPT-4 evaluations⁵ generally concur with human assessments in terms of both relative ranking and average scores within the same subset of 10 samples delineated in human evaluations. We then extend the evaluation to include all samples from the test sets⁶.

The outcomes for these tests are shown in Table 6, as well as in Table 9, 10 in Appendix H. We find that in GPT-4 evaluation, GPT-4 tends to assign the lowest scores to its own answers compared to those generated by other fine-tuned models. Summaries written by humans receive the highest scores and are generally regarded as the highest quality. In line with human evaluation findings, GPT-4 also recognizes LoRA as yielding inferior outcomes. In addition, the RST_w^P-LoRA model scored higher

⁵Utilizing the same iteration of the GPT-4 model as employed in prior summary generation tasks.

⁶Prompt can be found in Appendix E.3.

Candidate	R	I	C	F	Best	Worst
Human	4.89	4.76	4.67	4.72	96.8%	0.0%
GPT-4 _{ICL}	4.02	3.81	4.47	3.12	0.0%	35.3%
Vicuna _{LoRA}	4.20	3.82	4.43	3.37	0.0%	29.5%
Vicuna _{FFT}	4.31	4.04	4.49	3.55	0.0%	25.5%
Vicuna _{RST_w^P-LoRA}	4.46	4.44	4.60	4.12	3.2%	9.7%

Table 6: GPT-4 evaluation results on BC dataset

than both LoRA and FFT. We further discuss the error analysis (case study) in Appendix I.

6 Conclusion

We present RST-LoRA, a novel discourse-aware LoRA model tailored for long document summarization. Our approach primarily incorporates rhetorical knowledge into the LoRA training process by transforming RST structures into RST distributions. We develop four RST-LoRA variants, examining the impact of uncertainty in RST relational connections and discourse labels on overall performance. Empirical evidence from our studies demonstrates a consistent improvement in the performance of the standard LoRA model. By fine-tuning less than 0.5% of the LLMs parameters, our best RST-LoRA variant not only surpasses the performance of LoRA and FFT but also exceeds previous state-of-the-art methods. Furthermore, our analysis underscores the efficacy of our approach in leveraging discourse knowledge, which strengthens LoRA’s capabilities in producing more factually consistent and better-quality summaries.

7 Ethics Considerations

The datasets employed in our research are accessible to the public. Throughout the stages of data processing, experimental analysis, and model training/evaluation, our approach detects no violations of privacy. Regarding human evaluation, all participants engage voluntarily and are appropriately compensated. Additionally, we guarantee a safe and supportive setting during the evaluation period, following the ACM Code of Ethics in our experimentation and analysis.

8 Limitations

Data All long document summarization datasets we use are open-source and peer-reviewed datasets. While these data sources are of high quality, inherent bias may exist within them. Exploring bias falls outside the scope of our study. In addition,

the datasets we selected are from different fields (books, scientific publications, legal documents), and the heterogeneity between datasets is relatively high, but it is important to note that these data only represent a small fraction of real-world data and do not cover all possible long document summarization fields. Furthermore, although the RST parser we chose is multilingual, the exclusivity of English in our dataset can be seen as a limitation because it does not contain data for other languages, and we have not discussed RST-LoRA for non-English languages.

Model In our experiments, we employ two state-of-the-art long document pre-trained LLMs: Longformer and Vicuna. These models may carry biases from their pre-training phase. However, evaluating the extent of bias in these models has not been conducted, as it also lies outside the scope of this study. Additionally, the substantial cost of requesting GPT-4 (non-open source) API for generating summaries poses considerable financial barriers. We acknowledge this as a limitation and designate it as a potential space to explore in our future research. Furthermore, in our experiments, we compare the performance differences of different scales of LLMs (Longformer and Vicuna) and demonstrate that RST-LoRA can improve the summarization performance of LLMs at different scales. However, the number of potential LLMs is infinite, we have not tested on other models, and we leave the exploration of how the performance of RST-LoRA changes with different sizes of LLMs to future research.

Automated Evaluation Although we apply a range of widely used automated evaluation metrics in our experiments to systematically assess candidate models from multiple perspectives on the test set. While these metrics provide a multifaceted view of model performance, we are aware of their inherent limitations and the possibility that they may not fully encapsulate the models’ comprehensive performance.

Parser In line with prior research integrating discourse structures into summarization models, our work also needs an RST parser. In addition, manually annotated RST trees are extremely costly, we are unable to compare the summarization difference between RST parser output against human-annotated RST trees. Furthermore, incorporating RST into the LoRA training process does not signif-

icantly increase the amount of calculation or time complexity (just one more element-wise multiplication operation), but using a parser to generate the discourse structure still requires corresponding calculations. Our paper argues that discourse structure aids in summarization, which is orthogonal to the use of a specific discourse parser. We recognize that similar or greater improvements would be observed when employing a better parser.

Generalizability Since the main research scope of this paper is long document summarization, we have not delved deeply into applying our proposed method to other NLP tasks, such as machine translation, question answering, or text simplification. Although our method could potentially be adapted to other NLP tasks involving LLMs without considerable modification, this aspect remains unexplored and is earmarked for subsequent research studies.

Human Assessment The sample size for human evaluation in our study is constrained due to the nature of the extensive length of the original documents, which often extends over multiple pages. Scaling up the evaluation process through methods such as crowd-sourcing becomes challenging. Therefore, similar to many preceding studies (Atri et al., 2023; Tang et al., 2023; Phang et al., 2023), we evaluate only a set of 10 documents, which may not provide a fully representative view of the entire dataset. All human evaluators we recruit are Master’s or Ph.D. students, not all are experts in the field of text summarization domain, nor are they necessarily proficient in reading across diverse domains. As such, their assessments, while valuable, should not be the sole basis for judgment.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). We are grateful to the anonymous reviewers and area chairs for their exceptionally detailed and helpful feedback.



European Research Council
Established by the European Commission

References

- Mohammad AkbarTajari, Sara Rajaei, and Mohammad Taher Pilehvar. 2022. [An empirical study on the transferability of transformer modules in parameter-efficient fine-tuning](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10617–10625, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Akari Asai, Mohammadreza Salehi, Matthew Peters, and Hannaneh Hajishirzi. 2022. [ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6655–6672, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yash Atri, Arun Iyer, Tanmoy Chakraborty, and Vikram Goyal. 2023. [Promoting topic coherence and inter-document consorts in multi-document summarization via simplicial complex and sheaf graph](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2154–2166, Singapore. Association for Computational Linguistics.
- Kartikeya Badola, Shachi Dave, and Partha Talukdar. 2023. [Parameter-efficient finetuning for robust continual multilingual learning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9763–9780, Toronto, Canada. Association for Computational Linguistics.
- Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. [Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. [SIMSUM: Document-level text simplification via simultaneous summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.
- Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. [Discourse-aware neural rewards for coherent text generation](#). 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 173–184, New Orleans, Louisiana. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Emanuele Bugliarello and Naoaki Okazaki. 2020. [Enhancing machine translation with dependency-aware self-attention](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1618–1627, Online. Association for Computational Linguistics.
- Shuyang Cao and Lu Wang. 2022. [HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.
- Shuyang Cao and Lu Wang. 2023. [Awesome: Gpu memory-constrained long document summarization using memory mechanism and global salient content](#). *arXiv preprint arXiv:2305.14806*.
- Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. 2022. [Revisiting parameter-efficient tuning: Are we really there yet?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2612–2626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Huadong Chen, Shujian Huang, David Chiang, and Jijun Chen. 2017. [Improved neural machine translation with a syntax-aware encoder and decoder](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1936–1945, Vancouver, Canada. Association for Computational Linguistics.
- Jiaao Chen and Diyi Yang. 2021. [Structure-aware abstractive conversation summarization via discourse](#)

- and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.
- Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. [Parameter-efficient fine-tuning design spaces](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. [A discourse-aware attention model for abstractive summarization of long documents](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. [Sparse low-rank adaptation of pre-trained language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4133–4145, Singapore. Association for Computational Linguistics.
- Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. [Discourse-aware unsupervised summarization for long scientific documents](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.
- Saadia Gabriel, Antoine Bosselut, Jeff Da, Ari Holtzman, Jan Buys, Kyle Lo, Asli Celikyilmaz, and Yejin Choi. 2021. [Discourse understanding and factual consistency in abstractive summarization](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 435–447, Online. Association for Computational Linguistics.
- Marjan Ghazvininejad, Vladimir Karpukhin, Vera Gor, and Asli Celikyilmaz. 2022. [Discourse-aware soft prompting for text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4570–4589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mozhdeh Gheini, Xuezhe Ma, and Jonathan May. 2023. [Know where you’re going: Meta-learning for parameter-efficient fine-tuning](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11602–11612, Toronto, Canada. Association for Computational Linguistics.
- Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. [Making science simple: Corpora for the lay summarisation of scientific literature](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Naman Goyal and Jacob Eisenstein. 2016. [A joint model of rhetorical discourse structure and summarization](#). In *Proceedings of the Workshop on Structured Prediction for NLP*, pages 25–34, Austin, TX. Association for Computational Linguistics.
- Naibin Gu, Peng Fu, Xiyu Liu, Zhengxiao Liu, Zheng Lin, and Weiping Wang. 2023. [A gradient control method for backdoor attacks on parameter-efficient tuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3508–3520, Toronto, Canada. Association for Computational Linguistics.
- Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. [PPT: Pre-trained prompt tuning for few-shot learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.
- Juncai Guo, Jin Liu, Yao Wan, Li Li, and Pingyi Zhou. 2022. [Modeling hierarchical syntax structure with triplet position for source code summarization](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–500, Dublin, Ireland. Association for Computational Linguistics.
- Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. [Towards a unified view of parameter-efficient transfer learning](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-efficient transfer learning for NLP](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.
- 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*.

- Yin Jou Huang and Sadao Kurohashi. 2021. [Extractive summarization considering discourse and coreference relations based on heterogeneous graph](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.
- 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.
- Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. [Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.
- Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. [Single document summarization based on nested tree structure](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. [BOOKSUM: A collection of datasets for long-form narrative summarization](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [SummaC: Re-visiting NLI-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Neal Lawton, Anoop Kumar, Govind Thattai, Aram Galstyan, and Greg Ver Steeg. 2023. [Neural architecture search for parameter-efficient fine-tuning of large pre-trained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8506–8515, Toronto, Canada. Association for Computational Linguistics.
- Tao Lei, Junwen Bai, Siddhartha Brahma, Joshua Ainslie, Kenton Lee, Yanqi Zhou, Nan Du, Vincent Y Zhao, Yuxin Wu, Bo Li, Yu Zhang, and Ming-Wei Chang. 2023. [Conditional adapters: Parameter-efficient transfer learning with fast inference](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The power of scale for parameter-efficient prompt tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jonathan Li, Will Aitken, Rohan Bhambhoria, and Xiaodan Zhu. 2023. [Prefix propagation: Parameter-efficient tuning for long sequences](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1408–1419, Toronto, Canada. Association for Computational Linguistics.
- Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. [The role of discourse units in near-extractive summarization](#). In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatzakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2024. [Loftq: LoRA-fine-tuning-aware quantization for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Zhenwen Li, Wenhao Wu, and Sujian Li. 2020. [Composing elementary discourse units in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Online. Association for Computational Linguistics.
- Baohao Liao, Yan Meng, and Christof Monz. 2023. [Parameter-efficient fine-tuning without introducing new latency](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4260, Toronto, Canada. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chin-Yew Lin and Eduard Hovy. 2003. [Automatic evaluation of summaries using n-gram co-occurrence statistics](#). In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

- Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023a. [RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yang Liu, Ivan Titov, and Mirella Lapata. 2019. [Single document summarization as tree induction](#). 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 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.
- Zechun Liu, Barlas Oguz, Aasish Pappu, Yangyang Shi, and Raghuraman Krishnamoorthi. 2023c. [Binary and ternary natural language generation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 65–77, Toronto, Canada. Association for Computational Linguistics.
- Zhengyuan Liu and Nancy Chen. 2019. [Exploiting discourse-level segmentation for extractive summarization](#). In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121, Hong Kong, China. Association for Computational Linguistics.
- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. [Multilingual neural RST discourse parsing](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. [DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing](#). In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.
- Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. [Discourse indicators for content selection in summarization](#). In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156, Tokyo, Japan. Association for Computational Linguistics.
- William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.
- Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabza. 2022. [UniPELT: A unified framework for parameter-efficient language model tuning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264, Dublin, Ireland. Association for Computational Linguistics.
- Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.
- Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.
- Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.
- Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venturi. 2020. [Linguistic profiling of a neural language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Jason Phang, Yao Zhao, and Peter Liu. 2023. [Investigating efficiently extending transformers for long input summarization](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961, Singapore. Association for Computational Linguistics.
- Edoardo Maria Ponti, Alessandro Sordani, Yoshua Bengio, and Siva Reddy. 2023. [Combining parameter-efficient modules for task-level generalisation](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 687–702, Dubrovnik, Croatia. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Dongqi Pu and Khalil Sima’an. 2022. [Passing parser uncertainty to the transformer: Labeled dependency distributions for neural machine translation](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 41–50, Ghent, Belgium. European Association for Machine Translation.

- Dongqi Pu, Yifan Wang, and Vera Demberg. 2023. [Incorporating distributions of discourse structure for long document abstractive summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.
- Dongqi Pu, Yifan Wang, Jia Loy, and Vera Demberg. 2024. [Scinews: From scholarly complexities to public narratives – a dataset for scientific news report generation](#). In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.
- Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernández Astudillo. 2021. [Structural guidance for transformer language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online. Association for Computational Linguistics.
- Sebastian Schuster and Tal Linzen. 2022. [When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.
- Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. [Echoes from alexandria: A large resource for multilingual book summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 853–867, Toronto, Canada. Association for Computational Linguistics.
- Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F. Hurtado, and José González. 2022. [DACSA: A large-scale dataset for automatic summarization of Catalan and Spanish newspaper articles](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931–5943, Seattle, United States. Association for Computational Linguistics.
- Noam Shazeer and Mitchell Stern. 2018. [Adafactor: Adaptive learning rates with sublinear memory cost](#). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.
- Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. [Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Chen Tang, Shun Wang, Tomas Goldsack, and Chenchua Lin. 2023. [Improving biomedical abstractive summarisation with knowledge aggregation from citation papers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 606–618, Singapore. Association for Computational Linguistics.
- Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vailancourt, and Fadi Biadsy. 2021. [Residual adapters for parameter-efficient ASR adaptation to atypical and accented speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6751–6760, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Ahmet Üstün and Asa Cooper Stickland. 2022. [When does parameter-efficient transfer learning work for machine translation?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7919–7933, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yixin Wan, Kuan-Hao Huang, and Kai-Wei Chang. 2023. [PIP: Parse-instructed prefix for syntactically controlled paraphrase generation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10372–10380, Toronto, Canada. Association for Computational Linguistics.
- Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2023. Parameter-efficient multilingual summarisation: An empirical study. *arXiv preprint arXiv:2311.08572*.
- Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. [Do we really need that many parameters in transformer for extractive summarization? discourse can help !](#) In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 124–134, 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.
- Yuhui Xu, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhengsu Chen, XI-AOPENG ZHANG, and Qi Tian. 2024. [QA-loRA: Quantization-aware low-rank adaptation of large language models](#). In *The Twelfth International Conference on Learning Representations*.

Zhuoyi Yang, Ming Ding, Yanhui Guo, Qingsong Lv, and Jie Tang. 2022. [Parameter-efficient tuning makes a good classification head](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7576–7586, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guangtao Zeng, Peiyuan Zhang, and Wei Lu. 2023. [One network, many masks: Towards more parameter-efficient transfer learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7564–7580, Toronto, Canada. Association for Computational Linguistics.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023a. [Adaptive budget allocation for parameter-efficient fine-tuning](#). In *The Eleventh International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Zhen-Ru Zhang, Chuanqi Tan, Haiyang Xu, Chengyu Wang, Jun Huang, and Songfang Huang. 2023b. [Towards adaptive prefix tuning for parameter-efficient language model fine-tuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1239–1248, Toronto, Canada. Association for Computational Linguistics.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. [Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. 2023. Parameter-efficient fine-tuning with layer pruning on medical report summarization and medical dialogue generation. *arXiv preprint arXiv:2305.08285*.

RST type	RST label
<i>Temporal</i>	Asynchronous, Synchronous
<i>Contingency</i>	Cause, Condition
<i>Comparison</i>	Contrast, Concession
<i>Expansion</i>	Explanation, Elaboration, Conjunction

Table 7: RST relation category

A RST Relation Category

B Datasets Statistics

Table 8 shows the statistics of the datasets. Within the table, *Coverage* measures how much of the summary directly utilizes tokens from the source material. A higher coverage score suggests a greater proportion of the summary’s tokens originate from the source document. *Density* calculates the average length of source text segments each summary token is associated with. A higher density score could imply the inclusion of longer continuous text segments in the summary (Segarra Soriano et al., 2022). The *compression ratio* represents the relationship between the source document’s length and that of the summary. A lower compression ratio indicates a summary that is comparatively more concise.

C Automatic Evaluation Metrics

- ROUGE (Lin, 2004) evaluates the overlap of n-grams between the machine-generated summaries and human-crafted references. Our analysis includes the F1 scores for Rouge-1 (R1), Rouge-2 (R2), Rouge-L (RL), and Rouge-Lsum (RLsum).
- BERTScore (Zhang et al., 2020) uses BERT embeddings to analyze semantic similarity.
- METEOR (Banerjee and Lavie, 2005) computes the harmonic mean of uni-gram precision and recall, placing additional emphasis on recall for a balanced assessment.
- sacreBLEU (Post, 2018) measures linguistic alignment and the fluidity between generated and reference summaries.
- NIST (Lin and Hovy, 2003) appraises the n-grams’ informativeness by assigning weights based on the novelty of information they contain, as determined by their frequency in the corpus.

D Hyper-parameters Settings

All experiments are optimized using the Adam (Kingma and Ba, 2015) optimizer (with $\beta_1 = 0.9$,

Dataset	Training	Validation	Test	Avg. Doc Tokens	Avg. Summary Tokens	Coverage	Density	Compression Ratio
Multi-LexSum	3177	454	908	75543.21	646.14	0.93	3.39	96.4
eLife	4346	241	241	10132.12	382.58	0.82	1.77	27.65
BookSum Chapter	9600	1431	1484	5339.62	505.42	0.78	1.69	15.97

Table 8: Datasets statistics

$\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and weight decay = 0.1) and Adafactor (Shazeer and Stern, 2018), with a warm-up ratio of 0.2. The initial learning rate is set to $5e-5$, with a cosine learning rate schedule.

Additionally, within the LoRA strategy, we set a constant rank r to 8, the scaling α to 32, and the dropout rate to 0.1. During training, we save checkpoints that achieve the highest Rouge-2 F1 score on the validation set as the final model. All experiments are run for 50 epochs with a batch size of 16, and early stopping is implemented to prevent over-fitting (all models converged before 50 epochs). For model inference, we employ a beam search of size 4 with a length penalty of 3.0 and set a no-repeat n-gram size of 3.

For GPT-4⁷, we employ GPT-4 Turbo version (*gpt-4-1106-preview*), which is, at the time of experimentation (between 10 October 2023 and 15 December 2023), the best-performing publicly accessible version provided by OpenAI. For the hyper-parameters setting, we set temperature=1, top_p=1, frequency penalty=0.2, and presence penalty=0.2. The remaining hyper-parameters are set to their default values as recommended by OpenAI.

E GPT-4 Prompts

E.1 Prompt for Zero-shot Summaries Generation

Document: {Document}
Summary:

E.2 Prompt for In-context Summaries Generation

Document: {Document}
Summary: {Summary}

Document: {Document}
Summary: {Summary}

Document: {Document}
Summary:

E.3 Prompt for Summaries Evaluation

Source Document: {Document}
Summary of Candidate1: {Candidate1}
Summary of Candidate2: {Candidate2}
Summary of Candidate3: {Candidate3}
Summary of Candidate4: {Candidate4}
Summary of Candidate5: {Candidate5}

Note: The summaries are presented in order, with their respective candidate numbers from 1 to 5.

Please review the following evaluation guidelines to assess the quality of the above five candidate summaries, and rank them from best to worst:

Evaluation Guidelines: {Guidelines}

Please use the following format for your output (scores ONLY):

Relevance of Candidate1:
Informativeness of Candidate1:
Conciseness of Candidate1:
Faithfulness of Candidate1:
Relevance of Candidate2:
Informativeness of Candidate2:
Conciseness of Candidate2:
Faithfulness of Candidate2:
Relevance of Candidate3:
Informativeness of Candidate3:
Conciseness of Candidate3:
Faithfulness of Candidate3:
Relevance of Candidate4:
Informativeness of Candidate4:
Conciseness of Candidate4:
Faithfulness of Candidate4:
Relevance of Candidate5:
Informativeness of Candidate5:
Conciseness of Candidate5:
Faithfulness of Candidate5:
Ranking (using candidate number):

⁷<https://platform.openai.com/docs/models/>

F Impact of Different r

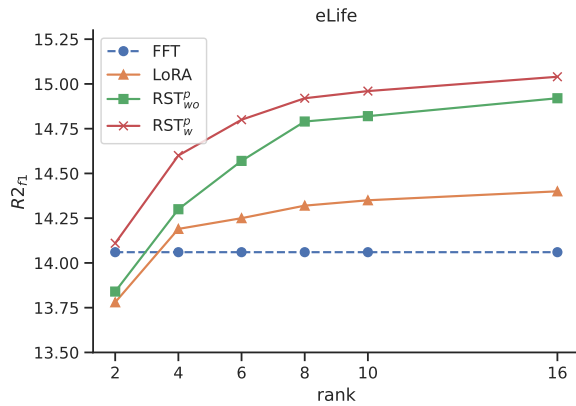


Figure 6: Impact of different r on eLife dataset

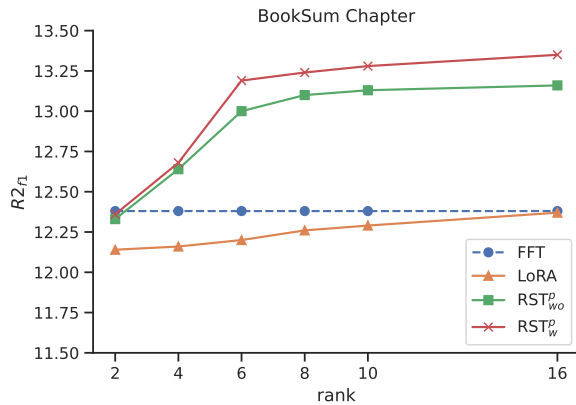


Figure 7: Impact of different r on BC dataset

G Human Evaluation Guideline

Here, we offer a more detailed explanation of the metrics and evaluation criteria used for our human evaluation process.

Prerequisites: Eligibility for this evaluation requires simultaneous fulfillment of two conditions: (1) being a master’s or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrating greater than or equal to C2 English proficiency⁸. If you do not meet both criteria, we respectfully ask you to refrain from participating in this task. Those who qualify are encouraged to proceed and follow the instructions below.

We invite you to carefully review the following long document along with five candidate summaries. After a thorough examination of each summary, please rate them based on the following four criteria, using a Likert scale from 1 (worst) to 5 (best), where a higher score denotes better quality:

- **Relevance:** This metric assesses the extent to which the summary content accurately reflects the source text. A relevant summary should encompass topics pertinent to the source document.
- **Informativeness:** This metric assesses the extent to which the summary provides a comprehensive understanding of the key points and essential details from the source text. An informative summary should encapsulate the core ideas, facilitating a clear and precise comprehension of the main arguments and findings of the source document.
- **Conciseness:** This metric assesses the extent to which the summary excludes less important information from the source text. A concise summary should effectively eliminate non-essential content from the source document during the generation process.
- **Faithfulness:** This metric assesses the extent to which the candidate is incorrect in that it contradicts the information from the source document. A faithful summary adheres strictly to the information provided in the source document, avoiding the inclusion of unverified facts.

Next, you are also expected to rank the candidates from best to worst based on overall quality.

⁸https://en.wikipedia.org/wiki/C2_Proficiency

H GPT-4 Evaluation Results

Candidate	R	I	C	F	Best	Worst
Human	4.67	4.70	4.52	4.83	94.2%	0.0%
GPT-4 _{ICL}	4.43	3.88	3.62	3.19	0.0%	43.3%
Vicuna _{LoRA}	4.52	4.03	4.20	3.40	0.0%	28.4%
Vicuna _{FFT}	4.52	4.06	4.28	3.58	0.0%	21.6%
Vicuna _{RST_w^p-LoRA}	4.57	4.33	4.31	4.22	5.8%	6.7%

Table 9: GPT-4 evaluation results on ML dataset

Candidate	R	I	C	F	Best	Worst
Human	4.80	4.81	4.72	4.78	96.3%	0.0%
GPT-4 _{ICL}	4.22	3.91	4.35	3.45	0.0%	45.2%
Vicuna _{LoRA}	4.47	4.12	4.41	3.58	0.0%	30.1%
Vicuna _{FFT}	4.59	4.23	4.47	3.82	0.2%	16.3%
Vicuna _{RST_w^p-LoRA}	4.62	4.49	4.63	4.39	3.5%	8.4%

Table 10: GPT-4 evaluation results on eLife dataset

I Error Analysis

The passage discusses a comparison between RST-LoRA and GPT-4 in generating summaries using two samples from the eLife test set, as shown in Tables 11 and 12. It is observed that the summaries generated by GPT-4 tend to be overly general, lacking relevance and informativeness to the source document, and often include the model’s own judgments or inferences. For instance, in the example provided, GPT-4 deviates from the main discussion topic and fails to align the generated summary with factual consistency from the source document. This comparison also aligns with the findings of [Atri et al. \(2023\)](#) suggesting that general-purpose LLMs like GPT-4, without task-specific fine-tuning, still tend to focus more on language fluency but struggle to maintain fidelity to factual information and consistency with the source document.

Model	Output
Target-1	<p>Imagine a gymnastics competition in which participants take turns to cartwheel and somersault across the floor. The routines on display comprise sequences of precisely timed movements learned through practice. This is also true for many of the actions we perform every day, such as reaching for a cup of coffee. A region of the brain called the cerebellum helps us learn sequences of movements. But how does it do this? To find out, Khilkevich et al. came up with a new version of an old experiment. Rabbits were first trained to blink their eye in response to a specific external cue. This type of learning, called associative learning, has been shown before in the cerebellum. But Khilkevich et al. wondered whether the cerebellum could also use internal feedback signals from the eye-blink as a cue to learn the next movement. If so, this might explain how the cerebellum can chain movements together in a sequence. As predicted, Khilkevich et al. found that rabbits could learn to blink their eye in response to an initial signal, and then blink again in response to the first blink. Control experiments confirmed that the second eye-blink was coupled to the first, and not to the original cue. Moreover, on many trials, the rabbits showed a third and even fourth eye-blink. This is because feedback signals from the first, second, or third blink were the same. Thus, the feedback signals from the first blink triggered the second blink, feedback from the second triggered the third, and so forth. Rabbits could also learn to use a blink of the left eye as a cue for a blink of the right eye. Similar patterns of neuronal activity accompanied each blink, suggesting that the same mechanism generated them all. The cerebellum can thus use feedback from one movement as a cue to learn the proper timing of the next movement in a sequence. A key question is whether this mechanism of sequence learning extends beyond movement. The cerebellum has extensive connections to the brain's outer layer, the cortex, including many areas involved in cognition. Future experiments should test whether the cerebellum might help guide sequences of cortical activity during cognitive tasks.</p>
GPT-4 _{ICL}	<p>In a recent study, researchers explored the idea of using rabbits to understand human movement, focusing on how the cerebellum, a brain part, might be involved. They looked at how rabbits respond to various light and sound cues, seemingly to understand if these animals could learn complex dance-like moves. The study, which seemed to have a loose structure, suggested that the rabbits' performance in these tasks could somehow be linked to human learning of sequences, like playing a musical instrument or participating in sports. However, the connection between the rabbits' activities and human motor skills was not established, leaving the real impact of the study on human neuroscience quite ambiguous. Additionally, the research didn't delve deeply into the differences between rabbit and human brains, making its applicability to human subjects questionable.</p>
Vicuna _{RST_w^p-LoRA}	<p>The cerebellum's role in movement and learning was explored in a study by Khilkevich et al., where rabbits were trained to blink in response to cues. This research investigated the cerebellum's capacity for associative learning and its ability to use internal feedback from one movement as a cue for the next. Findings revealed that rabbits could learn sequential blinking, with each blink triggering the subsequent one. This suggests that the cerebellum not only coordinates simple movements but can also chain complex sequences of movements together. The study provides insights into the cerebellum's broader role in learning and cognition, particularly in understanding how it processes and coordinates the timing of sequential movements.</p>

Table 11: Comparison of target summary with the summary generated by RST-LoRA and GPT-4

Model	Output
Target-2	To grow and multiply, a living cell must take a variety of factors into account, such as its own energy levels and the availability of nutrients. A protein called mTOR sits at the core of a signaling pathway that integrates these and other sources of information. Problems with the mTOR pathway contribute to several diseases including diabetes and cancer. The mTOR protein occurs in two distinct protein complexes, called mTORC1 and mTORC2. These complexes contain a mix of other proteins – known as accessory proteins. They also sense different cues and act upon distinct targets in the cell. Recent research reported the structure of mTORC1, which provided clues about how this complex works. Yet, much less was known about the mTORC2 complex. Stutfeld, Aylett et al. have now used a technique called cryo-electron microscopy to reveal the three-dimensional architecture of the human version of mTORC2. Comparing the new mTORC2 structure to the existing one for mTORC1 showed that they have many features in common but important differences too. The overall shape of both complexes is similar and each complex contains two copies of mTOR arranged similarly. Also, the main accessory proteins in each complex interact with almost the same parts of mTOR, but the accessory proteins in mTORC2 are organized differently from those of mTORC1. The different accessory proteins also have distinct shapes. These differences could help to explain why the complexes respond to different cues and recognize different targets. These new findings provide an entry point for further studies on how mTORC2 works in cells. The next step is to get a higher-resolution image of the structure of this complex to see the finer details of all the components. This may in the future help scientists to develop drugs that inhibit mTORC2 to treat cancer and other diseases.
GPT-4 _{ICL}	In this paper, researchers delve into the intriguing world of cellular biology, focusing on a protein known as the mammalian target of rapamycin (mTOR). This protein, often likened to a cellular conductor, orchestrates a symphony of biological processes, guiding cells through complex growth and metabolic pathways. The study particularly highlights the dual roles of mTOR in different cellular complexes, drawing attention to its versatile nature. Interestingly, the paper touches upon the enigmatic behavior of mTOR in various diseases, hinting at its Jekyll and Hyde personality in cellular functioning. This study provides a glimpse into the intricate ballet of molecular interactions, painting a picture of mTOR as a pivotal figure in the dance of life.
Vicuna _{RST_w^p–LoRA}	The cells of animals, yeast, and other eukaryotes all contain a wide range of proteins that carry out different roles. A protein called mTOR sits at the core of this protein, where one letter of the protein sits at one end and the other at the other. The number of mRNAs is so small that it is thought that mTOR is only recently shown to be present in living cells. However, mTOR has only been studied in the laboratory in the form of single-celled structures called human cells. This has made it impractical for researchers to dissect the structure of the mTOR protein and its corresponding protein. Now, Wassagano et al. have used a technique called cryo-electron microscopy to reveal the three-dimensional architecture of mTOR proteins. The experiments reveal that this complex contains a small region called the ‘active’ site that is responsible for producing mTOR. This complex is also involved in the release of energy from blood sugar in cells and is also responsible for many of the processes that occur in the pancreas and immune responses in cells. In particular, the active state of the innermost mTOR complex is activated when the cell is starved for nutrients. Further experiments showed that the active site is involved in both the production and the activity of a protein called sIN1. Wassagani et al al.’s findings show that both mTOR and SIN1 have a similar active site to each other in human cancers. These findings shed new light on the role of how mTOR interacts with other proteins in human cells, which may help to identify new targets for developing treatments against these kinds of cancer.

Table 12: Comparison of target summary with the summary generated by RST-LoRA and GPT-4