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

Dongqi Pu, Vera Demberg

Department of Computer Science
Department of Language Science and Technology
Saarland Informatics Campus, Saarland University, Germany
dongqi.me@gmail.com

UNIVERSITÄT DES SAARLANDES

European Research Council
Established by the European Commission

NAACL 2024

TL;DR

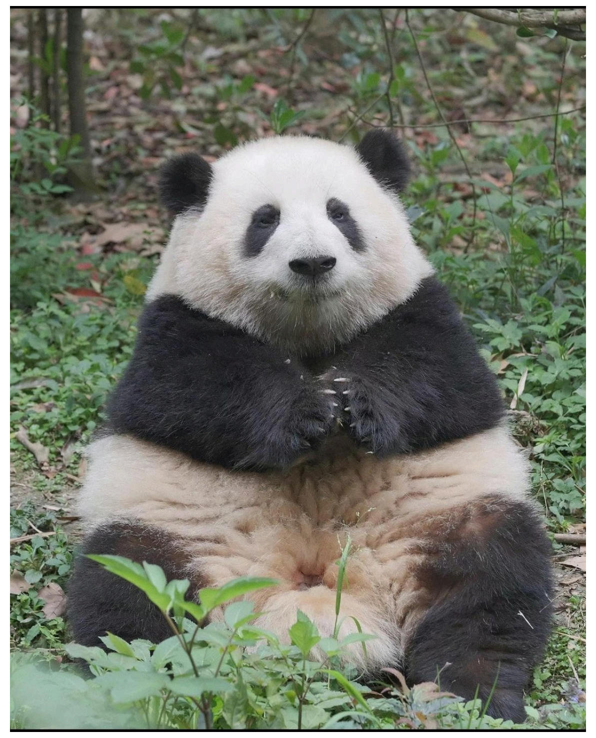
- **RST-LoRA** improves long document summarization by integrating **rhetorical structure theory** into the LoRA model, outperforming previous methods.

Motivation

- Why we need **low-rank** approximation?
- Why we need **discourse** knowledge?

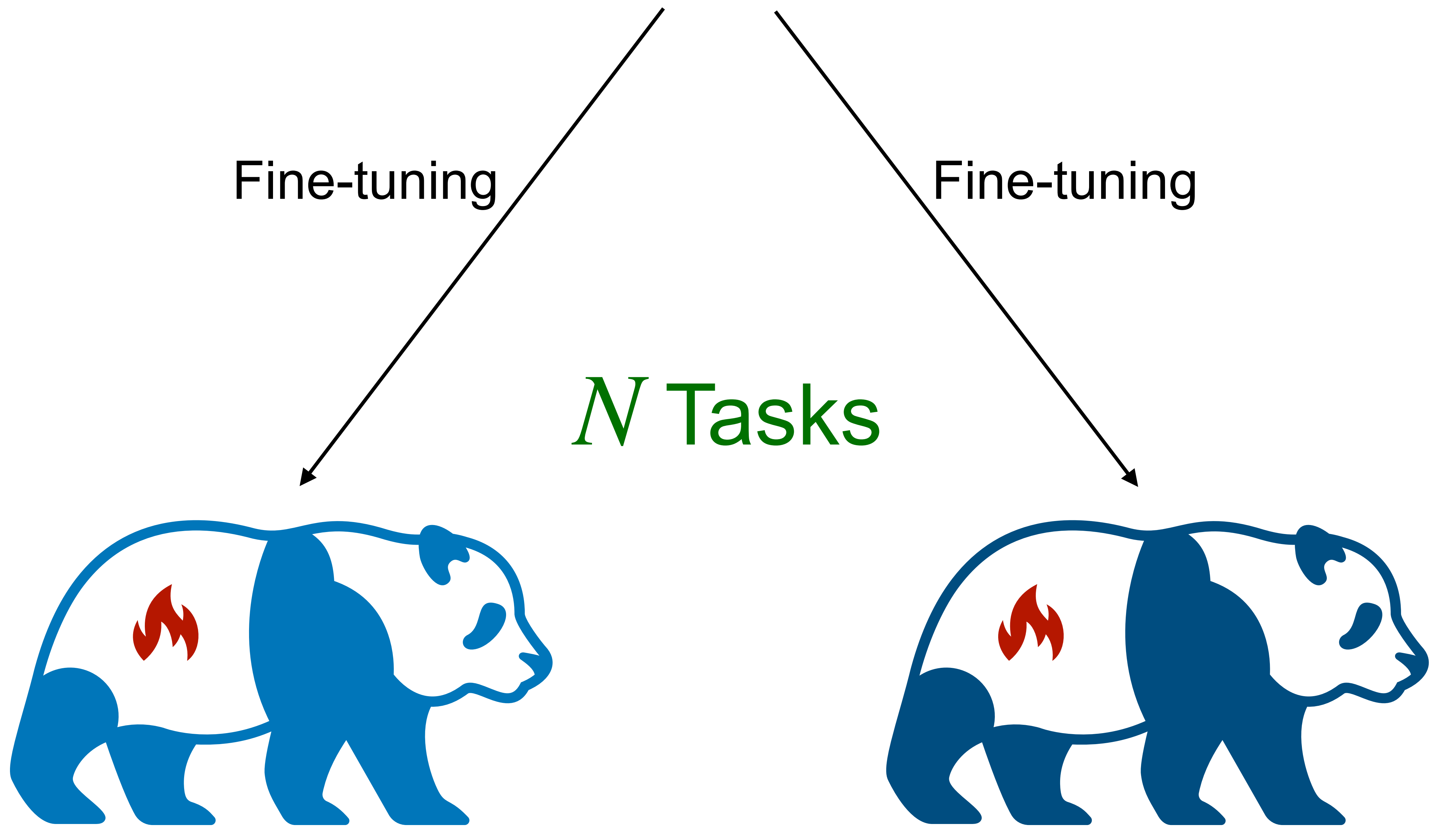
Motivation

- Why we need low-rank approximation?
  - Model size ↑ 👉 software and hardware ↑



M

Pre-trained model



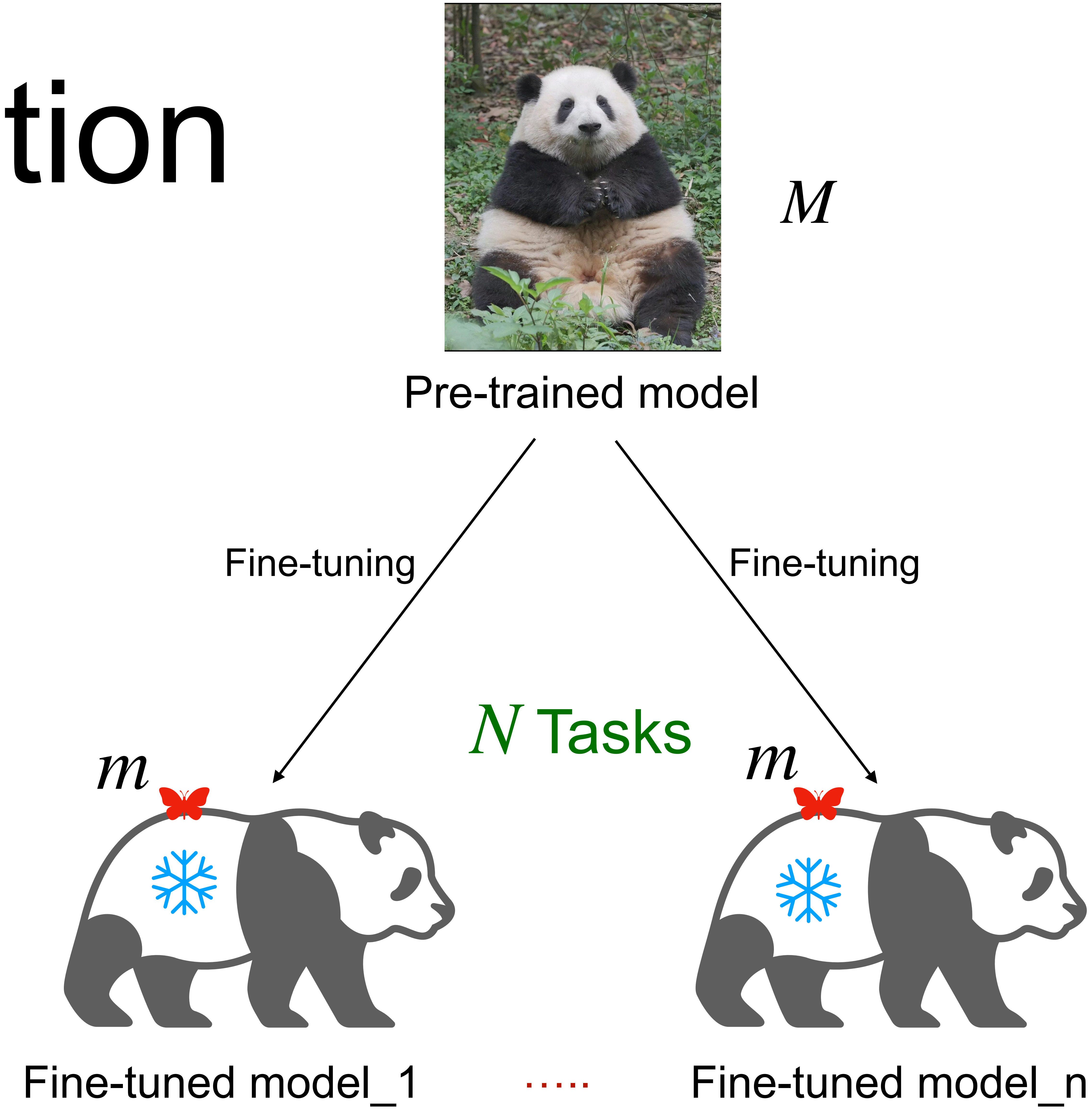
Fine-tuned model_1 ····· Fine-tuned model_n

Vanilla full-parameter fine-tuning (FFT)

$N \times M$

Motivation

- Why we need low-rank approximation?
  - Model size ↑ 👉 software and hardware ↑

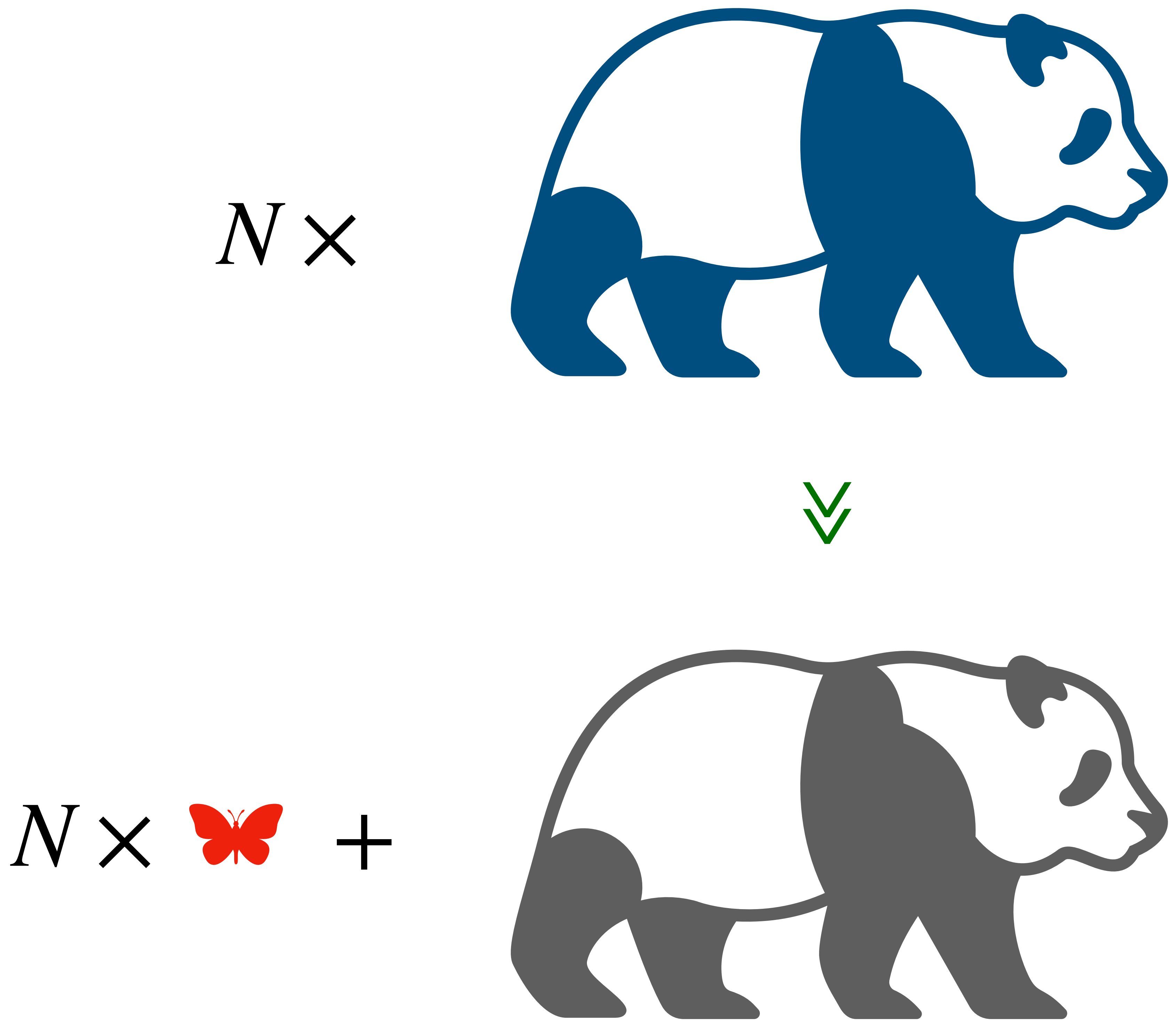


Parameter-efficient fine-tuning (PEFT)

$$N \times m + M$$

Motivation

- Why we need low-rank approximation?
  - Model size ↑ 👉 software and hardware ↑
  - Only 0.01–1% of the parameters, PEFTs ≈ FFT



FFT vs PEFT

Motivation

- Why we need discourse knowledge?
  - Challenges in PEFTs
    - Latent text relations
    - Importance level of different sentences

Ghazvininejad et al. (2022); Zhao et al. (2023)

Reason

EPFTs are not driven or guided by discourse knowledge during the training phase, as this is not explicitly present in the input data.

RST Prerequisite

- Rhetorical Structure Theory (RST) is helpful for determining:
  - Which sentences **should or should not** be included in the summary
    - Sentences relations
    - Discourse importance level

RST Prerequisite

- **EDU1** is the most pivotal component
- **EDU2** provides information for **EDU3**
  - It is not a problem to delete EDU2
  - It is still fine to delete both EDU2 and 3

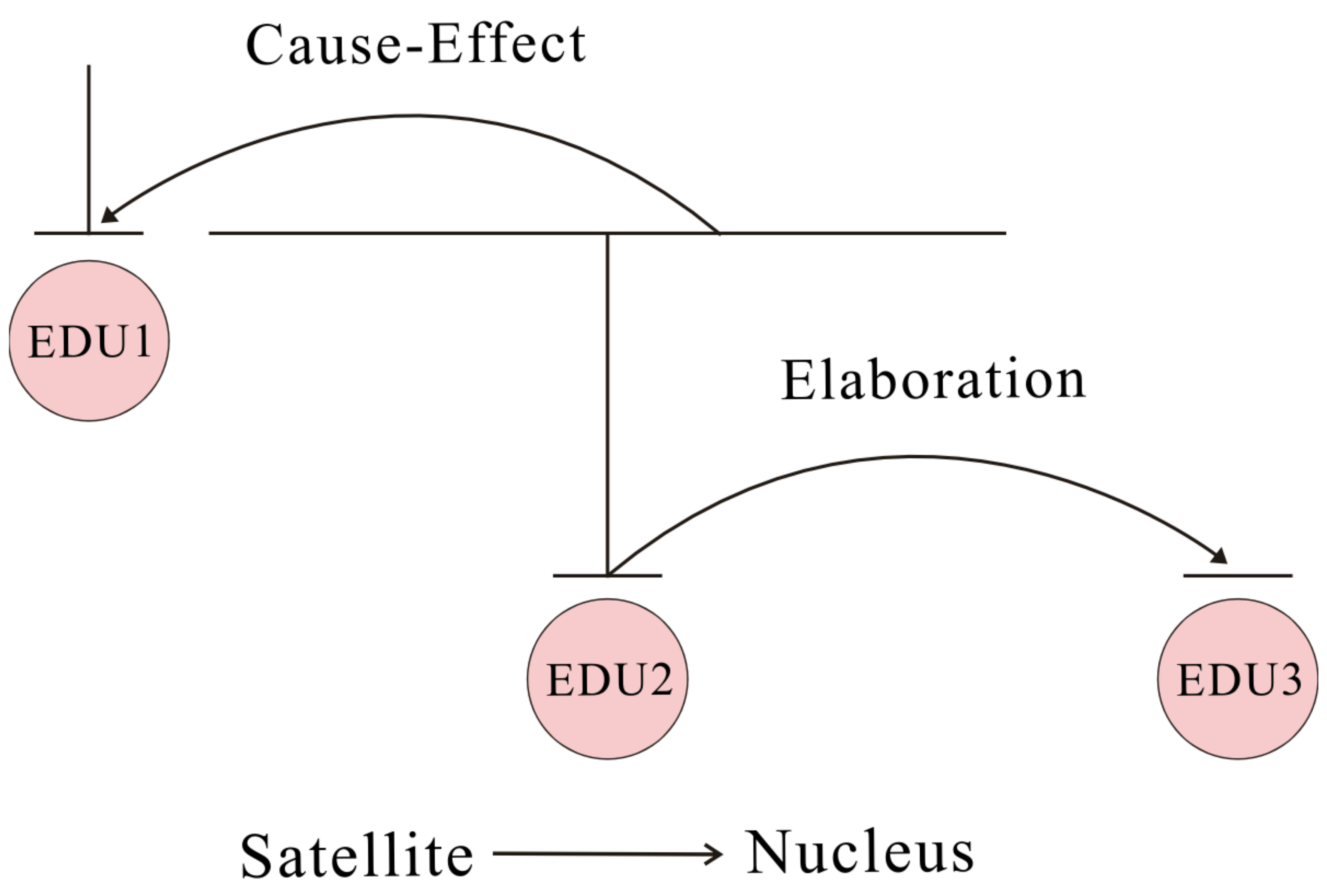


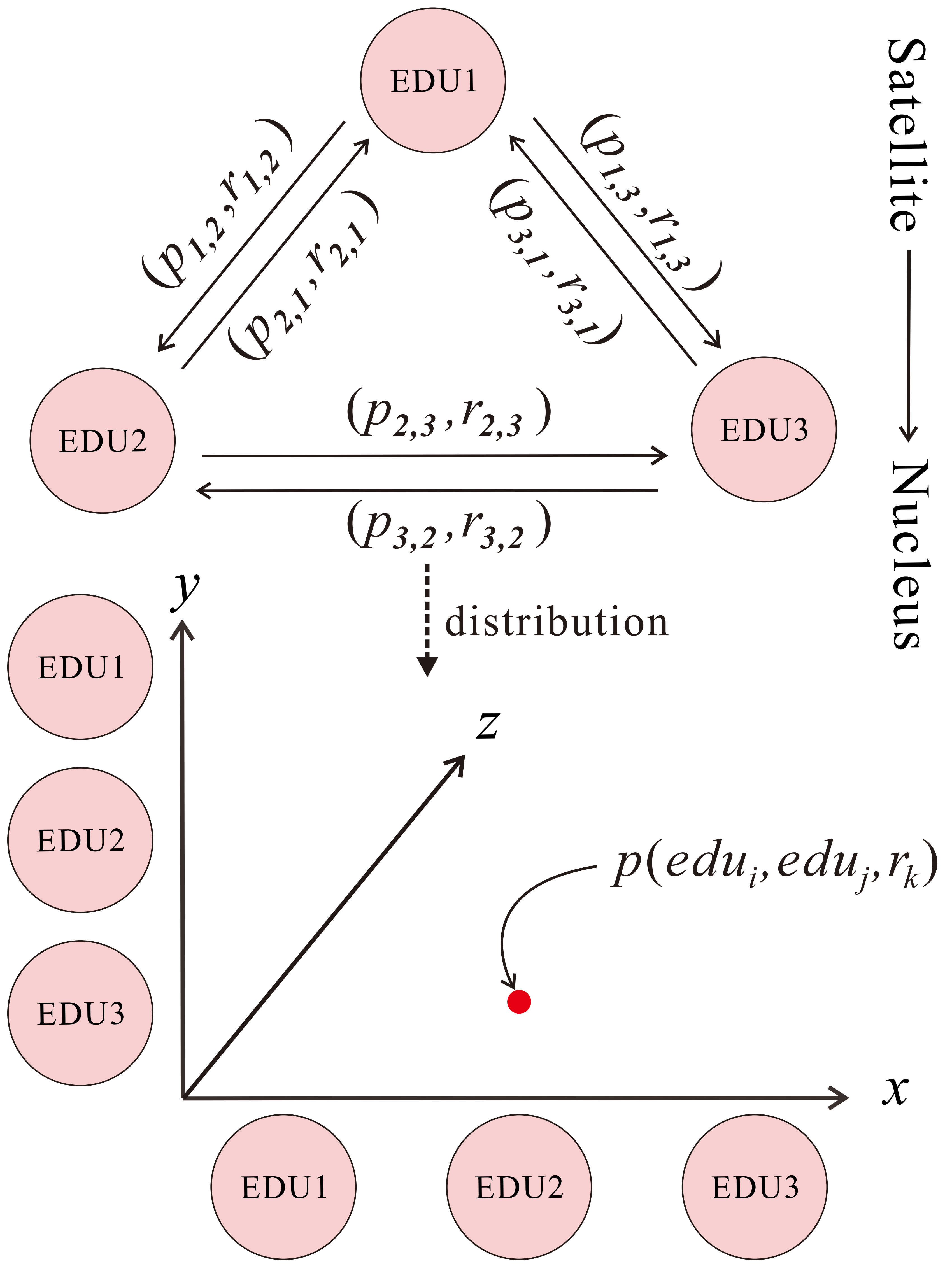
Figure 1: An example of RST tree: [*Utilizing discourse structure to enhance text summarization is beneficial.*]EDU1 [*This technique can be used to identify key ideas and capture often overlooked nuances.*]EDU2 [*Accurate capture of these complex structures facilitates the generation of good summaries.*]EDU3

Our Method

- RST Distribution
- RST-Aware Injection

Our Method

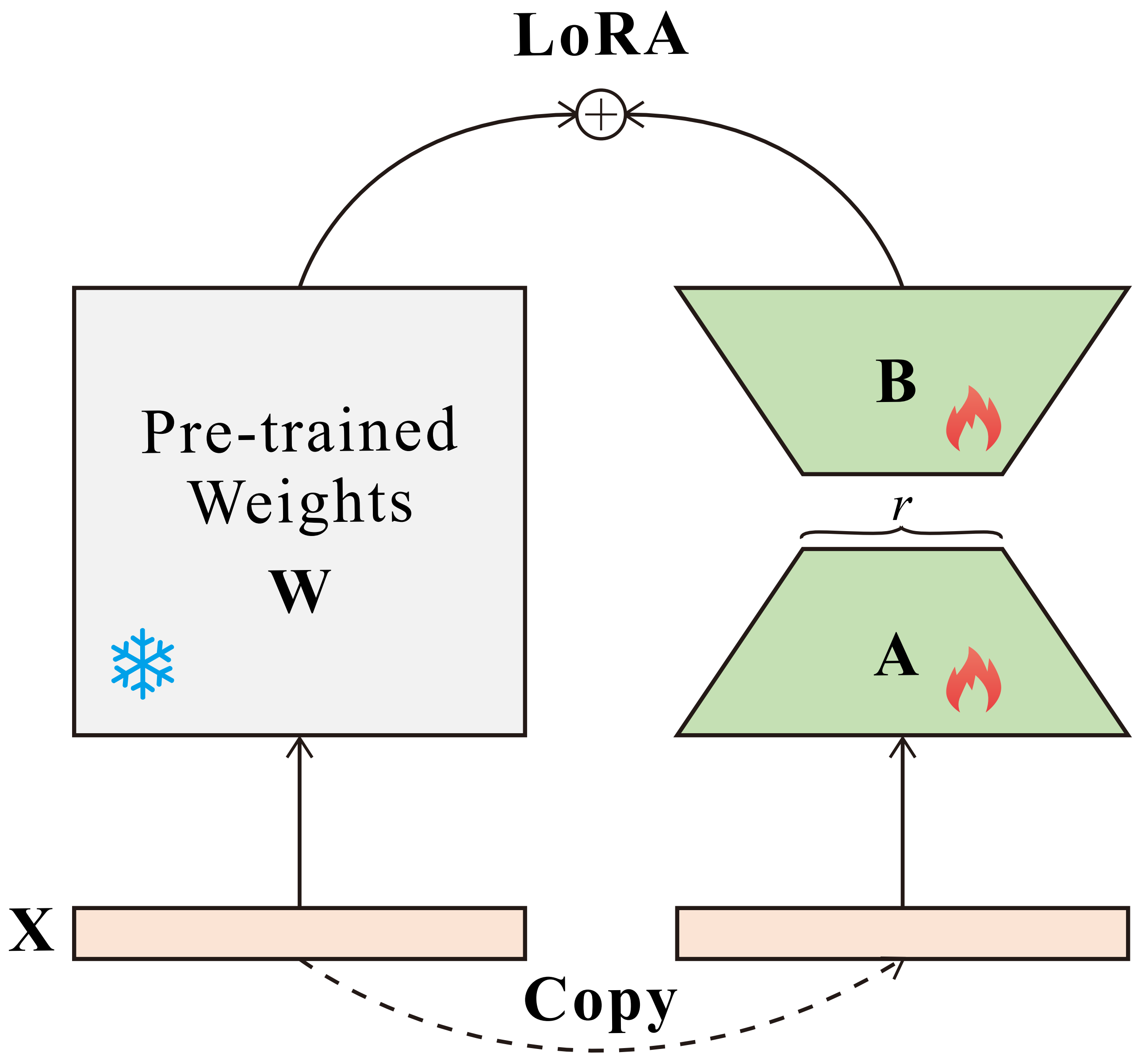
- RST Distribution
  - Each point indicates the probability value $p(edu_i, edu_j, r_k) \in [0,1] \subseteq \mathbb{R}$ that edu_i is the nucleus of edu_j with discourse relation r_k. (Pu et al., 2023)
  - We average and merge the y-axis of the matrix, and the merged value $c(edu_i, \overline{edu_j}, r_k)$ is called the importance index of edu_i with relation r_k.

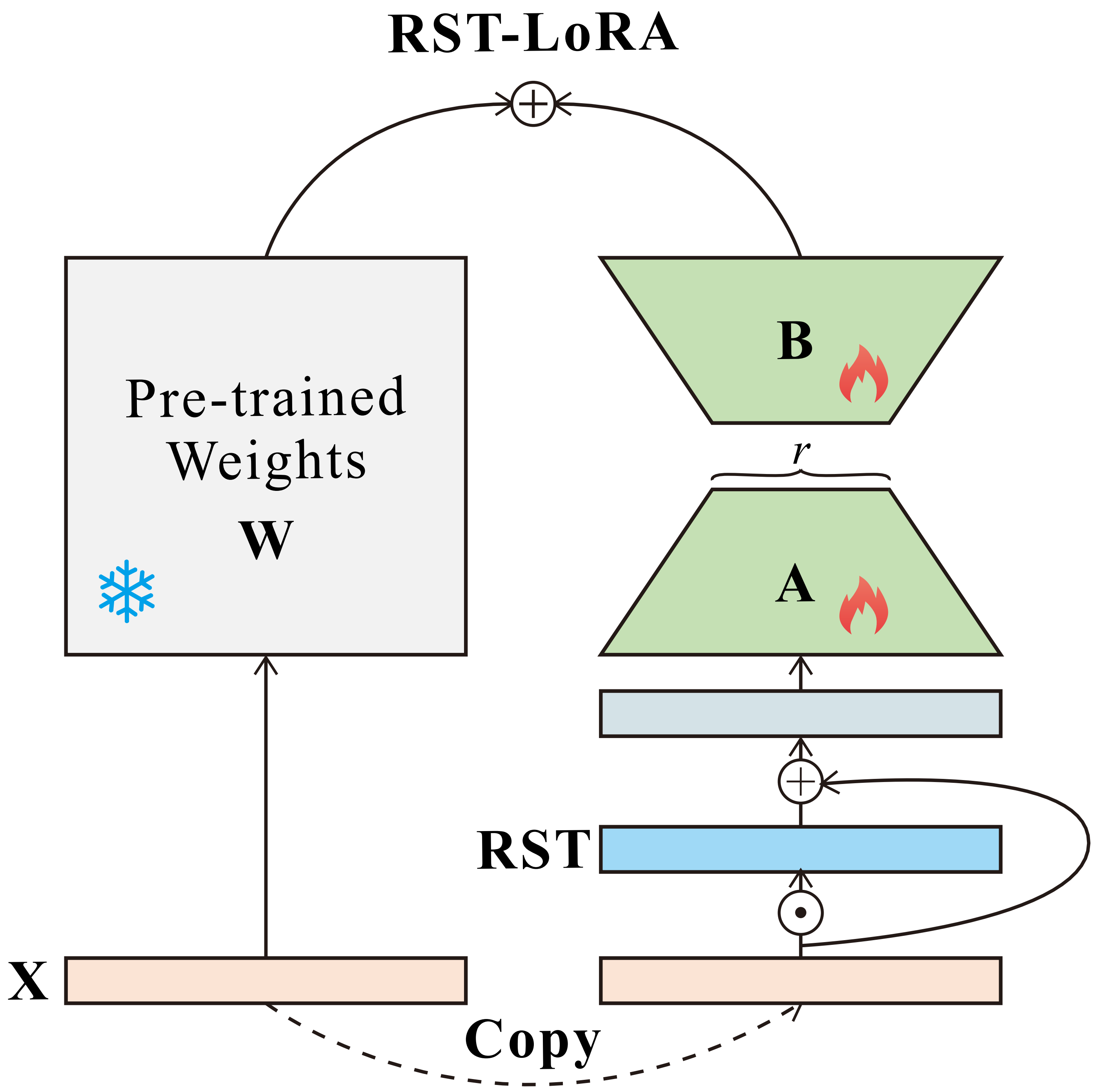


Our Method

- RST Distribution (4 variants)
  - RST^b_{wo}: Binary, label-agnostic representation (1 or 0)
  - RST^b_{w}: Binary distribution with relation labels
  - RST^p_{wo}: Label-omitted probabilistic representation
  - RST^p_{w}: Most fine-grained representation with relation types and probabilities

Our Method





- RST-Aware Injection

RST

- $h \leftarrow h + X(W^{down}_{A\times r} W^{up}_{r\times B})$ (vanilla LoRA)
- $h \leftarrow h + [(X \odot (1+\gamma))(W^{down}_{A\times r} W^{up}_{r\times B})$ (ours)

Experiments

- Experimental Settings
  - Datasets
  - Parser
  - Metrics
  - Training and Inference

Experiments

- **Datasets**
  - Multi-LexSum (ML, Shen et al., 2022)
  - eLife (Goldsack et al., 2022)
  - BookSum Chapter (BC, Kryscinski et al., 2022)

From legal documents, scientific papers, and books.

Experiments

- **Parser**
  - DMRST (Liu et al., 2020, 2021).
  - Extracting probabilities and type labels from final logits layer

Experiments

- **Metrics**
  - 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)
  - NIST (Lin and Hovy, 2003)

Experiments

- **Training and Inference**
  - Backbones
    - Longformer (Beltagy et al., 2020) 👉🏻 Seq2Seq
    - Vicuna13B-16k (Zheng et al., 2023) 👉🏻 GPT
  - Baselines
    - Backbones w/ FFT
    - Backbones w/ LoRA
    - GPT-4 (ZS & ICL)
  - Other SOTAs

RST variant performance

- **Label integration**
- **Uncertainty consideration**

Both complementarily enhance model performance

Data	Model	$R1_{f1}\uparrow$	$R2_{f1}\uparrow$	$RL_{f1}\uparrow$	$RLsum_{f1}\uparrow$
Multi-LexSum	Longformer$_{RST^b_{wo}-LoRA}$	45.82	21.32	23.81	43.40
	Longformer$_{RST^b_{w}-LoRA}$	46.02	21.34	23.87	43.39
	Longformer$_{RST^p_{wo}-LoRA}$	46.21	21.54	24.09	43.37
	Longformer$_{RST^p_{w}-LoRA}$	46.33	21.86	24.11	43.58
	Vicuna$_{RST^b_{wo}-LoRA}$	46.32	21.64	24.22	43.32
	Vicuna$_{RST^b_{w}-LoRA}$	47.33	22.70	24.25	43.31
	Vicuna$_{RST^p_{wo}-LoRA}$	47.39	22.79	24.35	43.33
	Vicuna$_{RST^p_{w}-LoRA}$	47.45	23.19	24.39	44.02
eLife	Longformer$_{RST^b_{wo}-LoRA}$	49.34	14.24	21.34	46.74
	Longformer$_{RST^b_{w}-LoRA}$	49.41	14.39	21.29	46.79
	Longformer$_{RST^p_{wo}-LoRA}$	49.87	14.49	21.83	47.15
	Longformer$_{RST^p_{w}-LoRA}$	49.89	14.68	22.11	47.64
	Vicuna$_{RST^b_{wo}-LoRA}$	48.73	14.68	21.89	47.11
	Vicuna$_{RST^b_{w}-LoRA}$	49.72	14.72	22.03	47.02
	Vicuna$_{RST^p_{wo}-LoRA}$	49.87	14.79	22.21	48.10
	Vicuna$_{RST^p_{w}-LoRA}$	49.92	14.92	22.41	48.21
BookSum Chapter	Longformer$_{RST^b_{wo}-LoRA}$	34.70	10.22	20.39	34.21
	Longformer$_{RST^b_{w}-LoRA}$	34.72	10.19	20.41	34.87
	Longformer$_{RST^p_{wo}-LoRA}$	35.29	11.38	21.62	35.11
	Longformer$_{RST^p_{w}-LoRA}$	35.40	11.76	21.88	35.27
	Vicuna$_{RST^b_{wo}-LoRA}$	37.28	12.35	22.13	38.33
	Vicuna$_{RST^b_{w}-LoRA}$	37.41	12.66	22.51	38.40
	Vicuna$_{RST^p_{wo}-LoRA}$	37.87	13.10	22.77	39.69
	Vicuna$_{RST^p_{w}-LoRA}$	37.92	13.24	22.93	40.31

Table 1: Performance of different RST variants

Main Results

- **LoRA vs. FFT:** Comparable, more efficient
- ***RST_w^p*-LoRA:** Best performance
- **GPT-4:** Poorest, lacks tuning

Dataset	Model	# Trainable Parameters	$R1_{f1}\uparrow$	$R2_{f1}\uparrow$	$RL_{f1}\uparrow$	$RLsum_{f1}\uparrow$	$BERTscore_{f1}\uparrow$	Meteor$\uparrow$	sacreBLEU$\uparrow$	NIST$\uparrow$
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
	$\text{Longformer}_{RST_w^p-LoRA}$	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
	$\text{Vicuna}_{RST_w^p-LoRA}$	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
	$\text{Longformer}_{RST_w^p-LoRA}$	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
	$\text{Vicuna}_{RST_w^p-LoRA}$	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
	$\text{Longformer}_{RST_w^p-LoRA}$	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
	$\text{Vicuna}_{RST_w^p-LoRA}$	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	-	-	-	-	-

Ablation Study

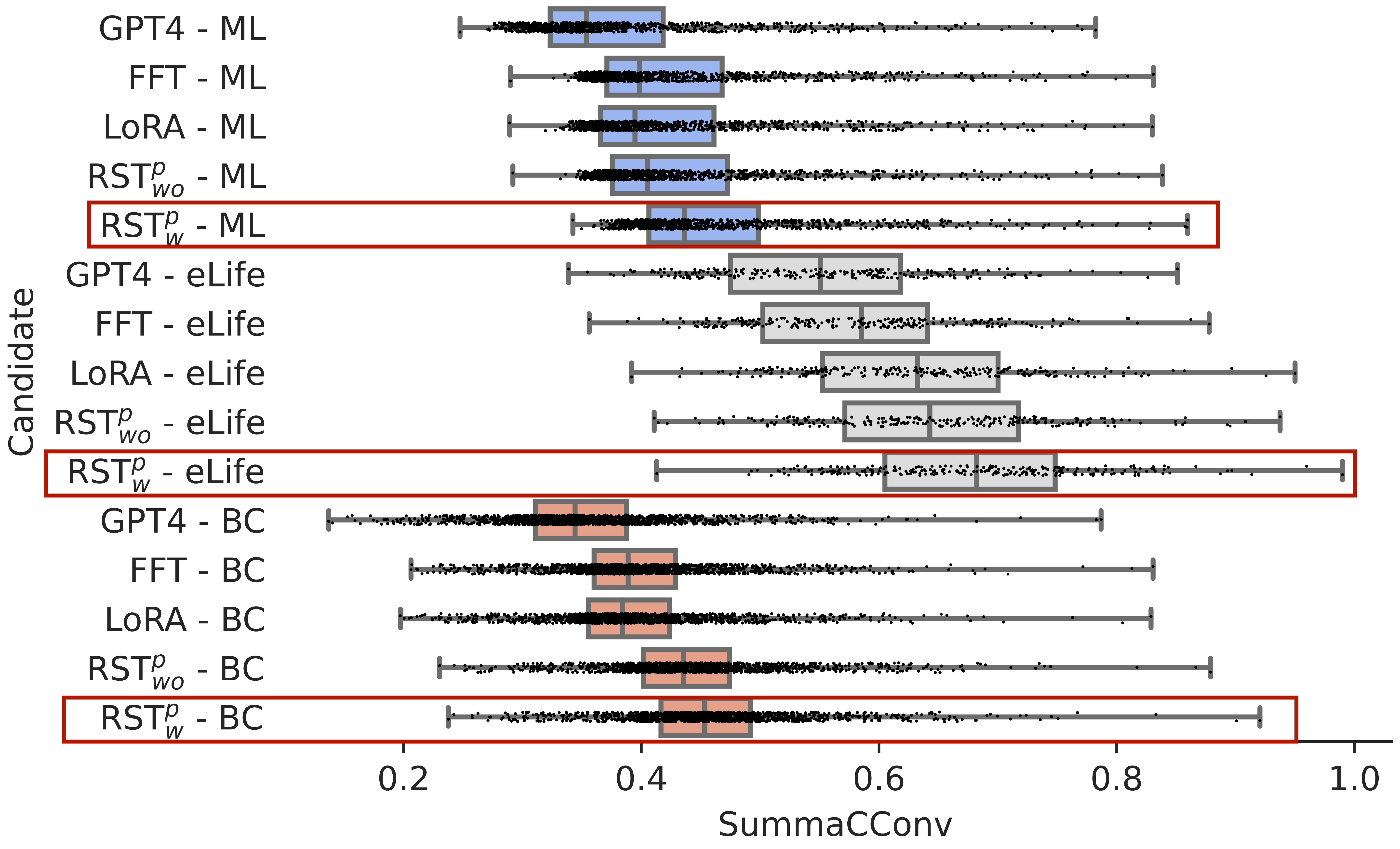
- **RST control conditions:** Even, Odd, Random
- **Vicuna** backbone testing
- Ablation shows **reduced performance**

Dataset	Model	$R1_{f1}\uparrow$	$R2_{f1}\uparrow$	$RL_{f1}\uparrow$	$RLsum_{f1}\uparrow$
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

Hallucination Checking

- **SummaC testing:** 0-1 score range
- **GPT-4:** Weakest consistency
- **RST enhances LoRA:** Reduces hallucinations



Impact of Different Rank r

$r = 8$ is a trade-off point between performance gain and computational cost

Impact of Parser Capability

- **Parser impact test:** 10%, 20%, 40%, 80% masking
- **Vicuna backbone:** Multi-LexSum dataset
- **Performance declines:** >40% noise

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
LoRA	**46.32**	**21.76**	**24.09**	**43.14**

Human Evaluation

- **Human evaluation:** BookSum, 10 instances
- **Evaluators:** CL/CS Graduate candidates, blind test
- **RST_w^p-LoRA:** Highest neural model performance

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%

Relevance (R), Informativeness (I), Conciseness (C), Faithfulness (F)

GPT-4 Evaluation

- **GPT-4 self-evaluation:** Lowest scores to own answers
- *RST_w^p*-**LoRA:** more closer to the quality of human-generated summaries

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%

Relevance (R), Informativeness (I), Conciseness (C), Faithfulness (F)

Conclusion

- A method for injecting **discourse knowledge** into the training of LoRA model.
- Discourse uncertainty and relation labels are **complementarily**.
- Our model **outperforms** current SOTA models in specific evaluation metrics.

More Info

- **Data & Code:** https://dongqi.me/projects/RST-LoRA
- **Questions:** dongqi.me@gmail.com

Thanks for listening

Q&A