Distilling Structured Rationale from Large Language Models to Small Language Models for Abstractive Summarization

Linyong Wang[1,2], Lianwei Wu[1,2*], Shaoqi Song[1], Yaxiong Wang[3], Cuiyun Gao[4], Kang Wang[1]
[1]ASGO, School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.
[4]Harbin Institute of Technology, Shenzhen, China
linyongwang@mail.nwpu.edu.cn, wlw@nwpu.edu.cn, songshaoqi@mail.nwpu.edu.cn, wangyx15@stu.xjtu.edu.cn,
gaocuiyun@hit.edu.cn, wk0_0@mail.nwpu.edu.cn

Abstract

Large Language Models (LLMs) have permeated various Natural Language Processing (NLP) tasks. For the summarization tasks, LLMs can generate well-structured rationales, which consist of Essential Aspects (EA), Associated Sentences (AS) and Triple Entity Relations (TER). These rationales guide smaller models (≤1B) to produce better summaries. However, their high deployment costs (≥70B), such as substantial storage space and high computing requirements, limit their utilization in resource-constrained environments. Furthermore, effectively distilling these structured rationales from LLMs into Small Language Models (SLMs) models remains a challenge. To address this, we propose the LLM-based Structured Rationale-guided Multi-view Weak-gated Fusion framework (LSR-MWF). The framework initially employs LLMs to dig structural rationales from a document, considering multiple viewpoints such as EA, AS, and TER. Then, it develop a multi-step summary generation evaluation strategy to select high-quality structured rationales. Subsequently, it aligns with these rationales using additional modules organized in a hierarchical structure. Finally, the framework integrates the features output by these modules with original abstractive model through a weak-gated mechanism. Experimental results on two publicly available CNN/DailyMail and XSum datasets show that our method improves the performance of the abstractive model, outperforming baselines by 11.2% and 5.8%, respectively. In addition, our method improves the interpretability of summary generation from the viewpoints of EA, AS and TER.

Code — https://github.com/Wangdoudou8/LSR-MWF

Introduction

Large language models (LLMs), such as GPT-series (Ouyang et al. 2022; Achiam et al. 2023), Llama (Touvron et al. 2023), PaLM (Chowdhery et al. 2023), and Chinchilla (Hoffmann et al. 2022), have permeated every aspect of natural language processing (NLP), encompassing tasks like question-answering (QA) systems (Longpre et al. 2023), etc. For summarization tasks, LLMs can generate high-quality structured rationales from a document, which consist of Essential Aspects (**EA**), Associated Sentences (**AS**), and Triple Entity Relations (**TER**), as shown in Figure 1. These rationales function similarly to Chain-of-Thought (CoT) (Wang, Zhang, and Wang 2023; Wen et al. 2023b,a) in guiding Small Language Models (SLMs) to produce better summaries. However, deploying LLMs (≥70B) requires substantial computational resources, limiting their utility in resource-constrained environments (≤1B) (Strubell, Ganesh, and McCallum 2020). Furthermore, effectively distilling these structured rationales from LLMs to SLMs remains a significant challenge due to the lack of careful rationale extraction, rationale selection, or training strategies (Pham et al. 2023; Jiang et al. 2024; Liu et al. 2024a).

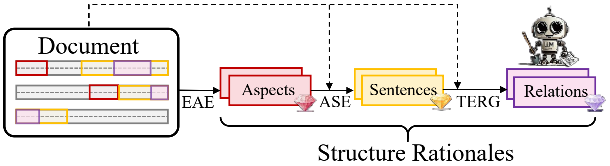


Figure 1: A simple demonstration of structured rationales distillation of documents via LLMs, including Essential Aspects Extraction (EAE), Associated Sentences Extraction (ASE) and Triples Entity Relations Generation (TERG).

Abstractive summarization is a pivotal task that condenses lengthy texts into concise, informative summaries (Radev, Hovy, and McKeown 2002). BART (Lewis et al. 2019) and PEGASUS (Zhang et al. 2020) are two prominent pre-trained Seq2Seq language models widely used in academic research. GSum (Dou et al. 2020), built upon BART, further enhances its performance by incorporating guidance from an extractive summarizer. Recently, SeqCo (Xu et al. 2022) emerged as a method that harnesses contrastive learning to bolster the effectiveness of BART in abstractive summarization. However, existing methods for summarization largely focus on the overall content of the source document, neglecting to explore its hierarchical structured information. This often results in coarse-grained, poor-quality summaries (Gekhman et al. 2023; Liu et al. 2023). Recognizing the hierarchical nature of source documents, recent research has indicated the potential of LLMs in generating structured rationales by extracting core themes from documents (Min et al. 2023;

[*]Corresponding author.

Jiang et al. 2024).

To impart the structured abstractive summarization capabilities of LLMs to SLMs, with the aim of enhancing both performance and interpretability, we introduce a novel framework, **L**LM-based **S**tructured **R**ationale-guided **M**ulti-view **W**eak-gated **F**usion framework (LSR-MWF), which not only makes the process of generating summaries clearer but also provides deeper insights into the content. Our work encompasses three main components: 1) Exploring structured rationales of documents by LLMs, 2) Selecting the best rationales, and 3) Training the local small model. First, we design three sub-tasks for LLMs to "dig" EA, AS and TER (three types of gems) of a document based on the LLM-based Structured Rationale-guided sub-framework (LSR). Simultaneously, these "gems" compose the structured rationales, offering insights into the document from three distinct viewpoints. Secondly, To ensure the overall quality of structured rationales, we adopt a multi-step summary generation evaluation strategy to select high-quality structured rationales for subsequent local small model training. Finally, we train our local small model using the Multi-view Weak-gated Fusion sub-framework (MWF). Specifically, we add three hierarchically structured modules to gradually align features of EA, AS, and TER. These features are then controlled through the weak-gated mechanism to fuse the outputs of the abstractive model.

The main contributions of this paper are as follows:

- **Effective Generation of Structured Rationales in LLMs:** We propose an innovative method, LLM-based Structured Rationale-guided sub-framework (LSR), which harnesses the extraction and generation capabilities of LLMs to progressively extract or generate EA, AS, and TER of documents.
- **Effective Knowledge Distillation of Structured Rationales to SLMs:** We ensure that SLMs can fully absorb and utilize these structured rationales. This is achieved by adopting the Multi-view Weak-gated Fusion sub-framework (MWF), significantly enhancing the performance of SLMs in summarization tasks.
- **Experimental Validation of Method Superiority and Interpretability:** Experimental results demonstrate that our proposed method significantly outperforms baselines on two public standard datasets, CNN/DailyMail and XSum. Finally, the entire framework, LSR-MWF, is innovative as it bridges the gap between LLMs and SLMs, enabling SLMs to inherit the structured abstractive summarization capabilities of LLMs while maintaining high performance and interpretability by showing specific cases and visualizations.

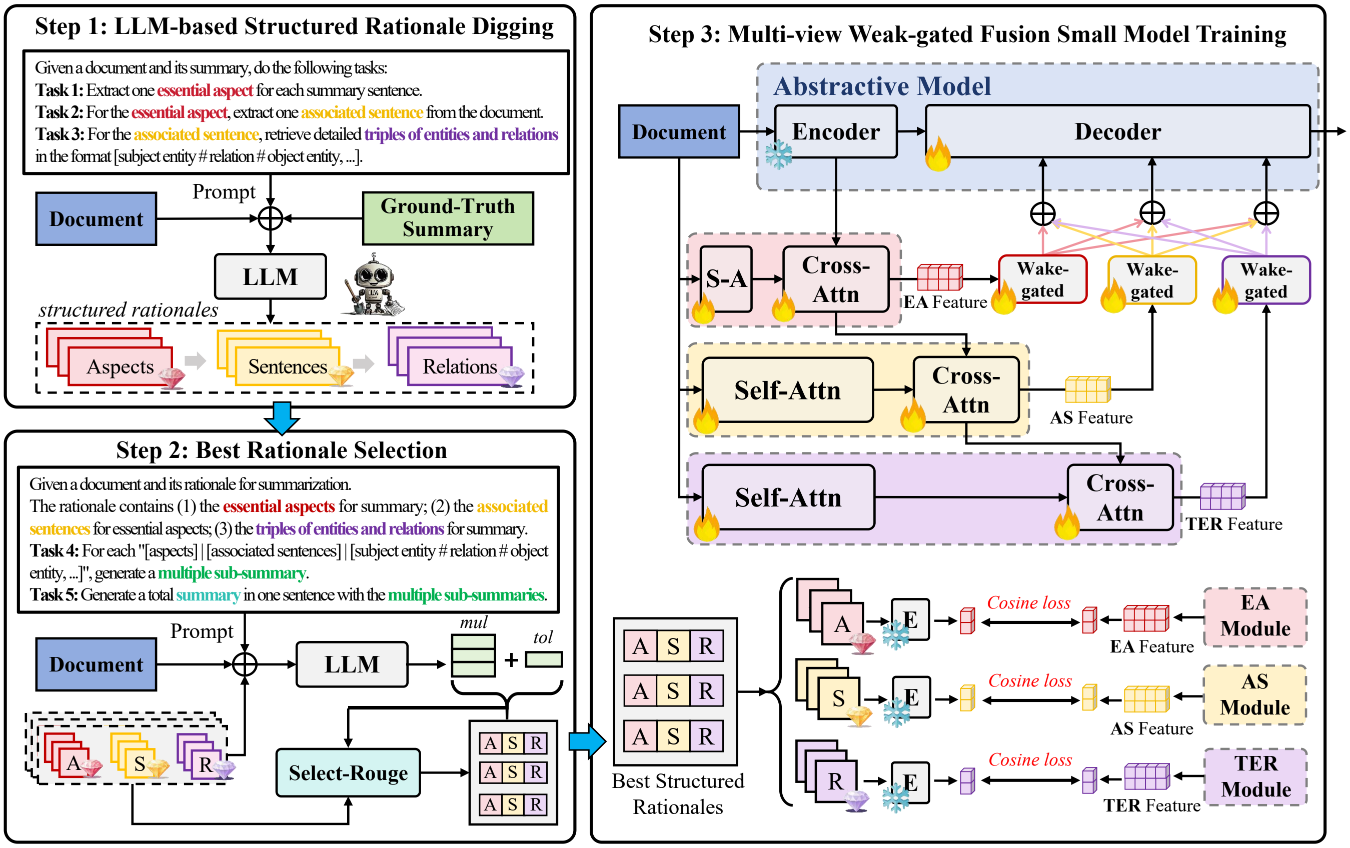
Related Work

LLMs for Abstractive Summarization

With technological advancements, numerous LLMs capable of performing summarization tasks have emerged, such as ChatGPT, GPT-4 (Achiam et al. 2023), and PaLM (Anil et al. 2023). These models, trained on vast amounts of text corpus with billions of parameters, exhibit exceptional performance in abstractive summarization tasks. Notably, their performance can be further enhanced when guided through step-by-step reasoning (Wei et al. 2022; Liu et al. 2024b). However, despite their impressive capabilities, the substantial resource requirements of these LLMs pose a challenge to their widespread adoption. Additionally, when utilizing LLM-as-a-service APIs, data privacy concerns cannot be overlooked, particularly when handling sensitive information. This underscores the importance of running SLMs locally. To leverage the powerful reasoning capabilities of LLMs in abstractive summarization, Wang et al. (2021) ingeniously utilized these models to enhance the quality of tags generated for headlines, while Jiang et al. (2024) harnessed aspect-triple rationales generated by LLMs to improve the summary quality of SLMs. Despite these advancements, existing methods still fail to fully transfer the comprehensive extraction and generation capabilities of LLMs to SLMs.

Knowledge Distillation and Interpretability in Abstractive Summarization

Knowledge distillation techniques (Hinton, Vinyals, and Dean 2015; Kim and Rush 2016; Guo et al. 2023) aim to extract specialized knowledge from larger models to tailor smaller models for specific tasks. This technology has found wide application in various domains (Shleifer and Rush 2020; Avram et al. 2021; Zhou, Xu, and McAuley 2021; Jiao et al. 2019; Jia et al. 2024). For abstractive summarization tasks, Jia et al. (2020) and Liu, Yang, and Chen (2024) focused on extractive and abstractive summarization, respectively, both leveraging knowledge distillation techniques to enhance summary generation quality. However, their approaches lack effective visualization, leaving interpretability as an area for further improvement. As the complexity of deep neural networks increases, the interpretability of models becomes increasingly crucial. To enhance model interpretability, researchers have begun exploring rationales generation techniques (Ho, Schmid, and Yun 2022; Hsieh et al. 2023; Li and Chaturvedi 2024; Wu et al. 2023b,a; Jiang et al. 2024). For abstractive summarization, the creation of rationales not only enhances model interpretability but also provides insights for keypoint generation, contributing to the production of higher-quality and more structured summaries. Recent work has leveraged the structured rationales generated by LLMs to enhance the performance and transparency of smaller summarization models. For instance, Jiang et al. (2024) utilized rationales and summaries generated by LLMs to train a smaller model through a multi-round curriculum learning approach. Similarly, other researchers, such as Ho, Schmid, and Yun (2022) employed the reasoning samples produced by LLMs to fine-tune smaller models, achieving substantial improvements in reasoning capabilities and transparency across various tasks. Even with these advancements, the comprehensive extraction and generation capabilities of LLMs in abstractive summarization remain underexplored. To further explore the limits of LLMs' reasoning abilities, this paper refines the rationale generation method and introduces the LSR-MWF to more comprehensively absorb and utilize these structured information.



Given a document and its summary, do the following tasks:
Task 1: Extract one **essential aspect** for each summary sentence.
Task 2: For the **essential aspect**, extract one **associated sentence** from the document.
Task 3: For the **associated sentence**, retrieve detailed **triples of entities and relations** in the format [subject entity # relation # object entity, ...].

Given a document and its rationale for summarization.
The rationale contains (1) the **essential aspects** for summary; (2) the **associated sentences** for essential aspects; (3) the **triples of entities and relations** for summary.
Task 4: For each "[aspects] | [associated sentences] | [subject entity # relation # object entity, ...]", generate a **multiple sub-summary**.
Task 5: Generate a total **summary** in one sentence with the **multiple sub-summaries**.

Figure 2: The complete process of distilling the structured rationales of LLM into local SLM. In the figure, "E" represents the encoder, "S-A" stands for Self-Attention for illustrative purposes, and "A", "S", and "R" denote essential aspects, associated sentences, and triple entity relations, respectively. Modules depicted with fire symbols indicate parameter updates, while snowflake symbols represent parameter freezing.

Methodology

Overview

We introduce the LSR-MWF. The overall architecture consists of three main components, as shown in Figure 2. It outlines the entire process, including LLM-based Structured Rationale Digging, Best Rationale Selection, and Multi-view Weak-gated Fusion Small Model Training. This method transfers document summarization capabilities from LLMs ($\geq$70B) to local SLMs ($\leq$1B). We elaborate on each process of the overall method as follows.

Step 1: LLM-based Structured Rationale Digging

Based on LLM-based Structured Rationale-guided subframework (LSR), we leverage the powerful extraction and generation capabilities of LLMs, along with tailored prompt templates, to conduct an extensive exploration of the latent reasoning process, termed "structured rationale", between the source document and its corresponding ground-truth summary. This exploration progresses through the hierarchical digging of three distinct viewpoints: Essential Aspects (EA), Associated Sentences (AS), and Triple Entity Relations (TER), which are guided by the requirements of the subsequent three tasks. To facilitate further detailed discussion, we outline some important concepts below.

Gem 1: Essential Aspect (EA) Defined as the topic words $a_{1\sim n}$ extracted from the source document that correspond to each summary sentence.

Gem 2: Associated Sentence (AS) Defined as the most relevant sentences $s_{1\sim n}$ in the document corresponding to the EA.

Gem 3: Triples Entity Relation (TER) Defined as the structured triple entity relations $r^*_{1\sim m} = \langle s|r|o \rangle_{1\sim m}, (m \geq n)$ extracted from the AS, including the subjects $s_{1\sim m}$, relations $r_{1\sim m}$ and objects $o_{1\sim m}$.

Task 1: EA Extraction This task is defined as extracting the EA in the document D and its corresponding ground-truth summary S^* , to ultimately obtain a set $A = \{a_1, a_2, \ldots, a_n\}$. Assuming $|S^*|$ to represent the number of sentences in S^*, the formula is given by:

$$a_i \sim P\left(A|D, S^*\right) = \prod_{i=1}^{|S^*|} p\left(a_i|D, s_i^*, a^{<i}\right) \quad (1)$$

Task 2: AS Extraction Given the extracted EA, this task is defined as extracting the AS from the document D to form a set $S = \{s_1, s_2, \cdots, s_n\}$, where:

$$s_i \sim P\left(S|A, D, S^*\right) = \prod_{i=1}^{|S^*|} p\left(s_i|a_i, D, s_i^*, s^{<i}\right) \quad (2)$$

Task 3: TER Extraction Based on the extracted AS, this task is defined as extracting the corresponding TER from each s_i to obtain $R = \{r_1^*, r_2^*, \cdots, r_m^*\}$, where:

$$\begin{aligned} r_i^* &= \{\langle s_j|r_j|o_j\rangle, \cdots\} \sim P\left(R|S, D, S^*\right), \\ P\left(R|S, D, S^*\right) &= \prod_{i=1}^{|S|} p\left(r_i|s_i, D, s_i^*, r^{*,<i}\right) \end{aligned} \quad (3)$$

Step 2: Best Rationale Selection

For each training sample, we set the temperature parameter τ of LLMs to 0 to ensure the uniqueness of the extracted structured rationales and generate multiple sub-summaries S^{mul} and an total summary S^{tol} using a multi-step summary generation evaluation strategy. This strategy aligns with the requirements of the following two tasks.

Task 4: Multiple Sub-summaries Generation Given a document D, we utilize LLMs again to generate a summary for each structured rationale, referred to as a sub-summary. Ultimately, we obtain a set of multiple sub-summaries S^{mul}. The structured rationales $R^* = \{(A_i, S_i, R_i)\}_{i=1}^n$, the corresponding formula is:

$$s_i^{\text{mul}} \sim P\left(S^{\text{mul}}|D, R^*\right) = \prod_{i=1}^{|R^*|} p\left(s_i^{\text{mul}}|D, r_i^*, s^{\text{mul},<i}\right) \quad (4)$$

Task 5: Total Summary Generation Based on the already generated multiple sub-summaries S^{mul} , we further compress them using LLMs to obtain a shorter total summary S^{tol} . The formula is as follows:

$$s_i^{\text{tol}} \sim P\left(S^{\text{tol}}|S^{\text{mul}}, D\right) = \prod_{i=1}^{|R^*|} p\left(s_i^{\text{tol}}|s_i^{\text{mul}}, D, s^{\text{tol},<i}\right) \quad (5)$$

Then, we calculate the ROUGE$_{\text{N}}$ scores for S^{mul} and S^{tol}. Following the evaluation method adopted by Liu et al. (2022); Zhang et al. (2013) during abstractive model validation, we use Eq. (7) and Eq. (8) respectively for CNNDM (Hermann et al. 2015; Nallapati et al. 2016) and XSum (Narayan, Cohen, and Lapata 2018) datasets to calculate the quality scores of the structured rationales. We discard the training examples with low scores and ultimately obtain two new datasets, as shown in Table 1.

$$\text{ROUGE}_{\text{N}} = \frac{\sum_{S \in Ref} \sum_{gram_s \in S} Count_{match}(gram_s)}{\sum_{S \in Ref} \sum_{gram_s \in S} Count(gram_s)} \quad (6)$$

$$Score_1 = 1 - \frac{\text{ROUGE}_1 \times \text{ROUGE}_2}{\text{ROUGE}_1 + \text{ROUGE}_2} \quad (7)$$

$$Score_2 = 1 - \frac{\text{ROUGE}_1 + \text{ROUGE}_2 + \text{ROUGE}_3}{3} \quad (8)$$

Step 3: Multi-view Weak-gated Fusion Small Model Training

In this subsection, we comprehensively train our local SLM using the Multi-view weak-gated Fusion sub-framework (MWF). This framework aligns with structured rationales from LLMs by utilizing additional modules organized in a hierarchical structure. Subsequently, it integrates the features output by these modules with the original abstractive model through a weak-gated mechanism. For simplicity, we use $\langle A\rangle$, $\langle S\rangle$ and $\langle R\rangle$ to represent the features of EA, AS, and TER, respectively. The ultimate goals of this framework are twofold:

Multi-view Hierarchical Aligning of Structured Rationales. We construct three hierarchically structured modules for three viewpoints of structured rationales: the essential aspects module, the associated sentences module, and the triple entity relations module. All modules are based on the Transformer architecture (Vaswani et al. 2017). The input to all modules is the same source document D. After passing through a shared embedding layer (omitted in the Figure 2 for simplicity), each module processes the input D through its own self-attention layer to enrich the semantic content. To ensure that the semantic features A_{out} , S_{out} and R_{out} output by these modules align closely with $\langle A\rangle$, $\langle S\rangle$ and $\langle R\rangle$, we previously encode $\langle A\rangle$, $\langle S\rangle$ and $\langle R\rangle$ using the encoder of the abstractive model. Next, we apply average pooling to the encoded $\langle A\rangle$, $\langle S\rangle$ and $\langle R\rangle$. Finally, we compute the cosine similarity between the likewise average-pooled semantic features "A_{out} , S_{out} , R_{out}" and "$\langle A\rangle$, $\langle S\rangle$, $\langle R\rangle$" using the formula $Cosine\, loss = sim\langle x, y\rangle = (xy) / (||x||||y||)$, such as $\mathcal{L}_{\text{EA}} = sim\langle A_{\text{out}}, \langle A\rangle\rangle$.

Features Fusion through Weak-gated Mechanism. To dynamically adjust the fusion degree of various features extracted from different viewpoints at each decoding layer, based on the current context. After A_{out}, S_{out} and R_{out} enter their respective weak-gated networks, each is copied L times, where L represents the number of layers in the decoder of the abstractive model. Once the output X^{en} from the encoder in the abstractive model is obtained, during the decoding phase, the following operations are performed at each layer of the decoder:

$$X_{i+1}^{de} = \text{MulHead}\left(W_i^Q X_i^{de}, W_i^K X_{\text{new},i}^{en}, W_i^V X_{\text{new},i}^{en}\right) \quad (9)$$

Here, $X_{\text{new},i}^{en}$ is the result of hierarchically features fusion between the output of the abstractive model's encoder and structured rationales, and where $i \in L$:

$$X_{\text{new},i}^{en} = X_i^{en} + g_i^A \cdot A_{\text{out},i} + g_i^S \cdot S_{\text{out},i} + g_i^R \cdot R_{\text{out},i} \quad (10)$$

Here, g_i^A represents the weak-gated unit at the i-th layer specifically designed to incorporate semantic features related to $\langle A\rangle$. Its value range is $[0, 1]$. It is a continuous value that can be adaptively updated during training. The responsibility of g_i^S and g_i^R is similar to that of g_i^A . Different from previous work, which uses a fixed ReLU activation function as a sturdy gate (Yao et al. 2020; Sun, Ren, and Xie 2024), we treat the

	# Examples			# Avg Words	
Datasets	Train	Valid	Test	Doc.	Sum.
CNNDM	287K	13K	11K	791.6	55.6
CNNDM*	203K	13K	11K	773.2	57.8
XSum	203K	11K	11K	429.2	23.3
XSum*	126K	11K	11K	457.6	25.5

Table 1: Datasets Statistics. "*" represents the datasets processed through steps 1 and 2.

gate as a kind of adaptively learned weight network parameter, thus called weak-gated unit. By observing the values of weak-gated units at different layers, we can gain insight into the model's dependence on distinct features at different decoding stages, which aids in deeper understanding of the abstractive model's decision-making process and working mechanism.

Training Objective of Loss Function To preserve and enhance the generative capabilities of the abstractive model, we adopt a combined loss function that integrates the sequence-level cosine loss ($\mathcal{L}_{\text{EA}} + \mathcal{L}_{\text{AS}} + \mathcal{L}_{\text{TER}}$) with the token-leve cross-entropy loss $\mathcal{L}_{cross-entropy}$. Our composite loss function is formulated as follows:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{cross-entropy} + \gamma_2 (\mathcal{L}_{\text{EA}} + \mathcal{L}_{\text{AS}} + \mathcal{L}_{\text{TER}}) \qquad (11)$$

Here, γ_1 and γ_2 are two hyper-parameters. Notably, the sequence-level cosine loss effectively complements the token-level cross-entropy loss. This is because the cosine loss captures the overall structural similarity, while the cross-entropy loss acts as a normalization mechanism, ensuring that the model can assign a balanced probability distribution across the entire sequence.

Experiments

Datasets and Metrics

We conduct experiments on two widely-used abstractive summarization datasets: CNN/DailyMail[1] (CNNDM) (Hermann et al. 2015; Nallapati et al. 2016) and XSum[2] (Narayan, Cohen, and Lapata 2018). These datasets differ in text length and level of abstraction, allowing us to demonstrate the generalization ability of our method. For original datasets, we first perform preprocessing, which includes removing empty items from documents or summaries. The document-summary pairs are then filtered through two processing steps, where the threshold for $Score_1$ is set to 85 and the threshold for $Score_2$ is set to 65. The dataset sizes before and after processing for CNNDM and XSum are shown in Table 1. We use ROUGE (Lin 2004) to measure the quality of abstracts in our results, specifically reporting F1 scores of ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) between ground-truth summaries and the generated abstracts. Additionally, we use BERTScore (BS) to measure semantic similarity between the generated summary and the reference summary.

Baselines

We choose a variety of strong-performing baseline models for comparison, including BERTSumAbs (Liu and Lapata 2019), T5 (Raffel et al. 2020), BART (Lewis et al. 2019), PEGASUS (Zhang et al. 2020), GSum (Dou et al. 2020), BigBird (Zaheer et al. 2020), SimCLS (Liu and Liu 2021), SeqCo (Xu et al. 2022), GLM (Du et al. 2021), BRIO (Liu et al. 2022), GPT-3.5 (Ouyang et al. 2022), and TriSum (Jiang et al. 2024).

Setup

We used Llama3-70B[3] as our LLM and BART-large from Hugging Face (Wolf et al. 2020) as our origin abstractive model. The overall parameter number of LSR-MWF is 439M($\leq$1B). All experiments are conducted on 2 NVIDIA RTX A6000 GPUs. We employ the Adam optimizer (Kingma and Ba 2014) with learning rate scheduling, where the learning rate lr is calculated as $2 \times 10^{-3} \min\left(\text{step}^{-0.5}, \text{step} \cdot \text{warmup}^{-1.5}\right)$, where step representing the number of update steps, and warmup set to 10000. For CNNDM, the initial weak-gated units are all set to 0.02, $\gamma_1 = 0.6$ and $\gamma_2 = 0.4$. For XSum, the initial weak-gated units are all set to 0.01, $\gamma_1 = 0.7$ and $\gamma_2 = 0.3$.

Results

The results are shown in Table 2. When utilizing structured rationales, Llama3-70B shows excellent summarization ability, outperforming all baselines, which indicates that structured rationales contribute to generating higher quality summaries. Furthermore, LSR-MWF outperforms many models across both datasets, highlighting its strength and adaptability. More specifically, LSR-MWF outperforms the original BART$_{\text{large}}$ by 11.2% and 5.8% on the two datasets, respectively. Therefore, LSR-MWF not only retains the ability of the abstractive model but also improves the quality of summary generation. It is worth noting that LSR-MWF performs better than the model TriSum, which also utilizes structured rationales, and this illustrates the effectiveness of our refined structured rationales and model design.

Ablation Study Table 3 examines the impact of removing different "gems" of structured rationales and their corresponding module on model performance. The results reveal that for both CNNDM and XSum datasets, removing TER (relation-level information) has a significant effect on model performance. Specifically, the model's performance degrades most when TER is absent, indicating its crucial role in summary generation. Additionally, we observe that relying solely on TER doesn't yield as good results as using AS (sentence-level information) alone. We assume that this is because TER is extracted from sentences, and without the support of AS, the model struggles to effectively utilize TER for CNNDM. Conversely, for XSum, TER alone demonstrates greater benefit when compared to AS alone. This may be attributed to the higher abstract level of summaries in XSum. To sum up, this multi-view constraint and hierarchically structured modules

[1] https://cs.nyu.edu/~kcho/DMQA/

[2] https://github.com/EdinburghNLP/XSum

[3] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

	CNNDM				XSum			
Model	**R-1**	**R-2**	**R-L**	**BS**	**R-1**	**R-2**	**R-L**	**BS**
BERTSumAbs (Liu and Lapata 2019)	41.18	18.73	37.22	0.8576	38.81	16.48	31.00	0.8723
T5$_{\text{Large}}$ (Raffel et al. 2020)	42.42	20.78	39.93	0.8722	40.12	17.23	32.34	0.9073
BART$_{\text{Large}}$ (Lewis et al. 2019)	44.01	21.12	40.58	0.8798	45.42	22.31	37.28	0.9162
PEGASUS (Zhang et al. 2020)	44.23	21.57	41.30	0.8737	46.71	24.38	38.89	0.9190
GSum (Dou et al. 2020)	45.52	22.32	42.13	0.8783	45.12	21.53	36.55	0.9123
BigBird$_{\text{Large}}$ (Zaheer et al. 2020)	43.83	21.12	40.74	0.8803	47.13	24.06	38.77	0.9197
SimCLS (Liu and Liu 2021)	45.57	21.91	41.02	0.8828	46.59	24.20	39.11	0.9078
SeqCo (Xu et al. 2022)	45.02	21.79	41.81	0.8747	45.59	22.38	37.02	0.9135
GLM$_{\text{RoBERTa}}$ (Du et al. 2021)	43.82	20.97	40.45	0.8733	45.51	23.48	37.33	0.8855
BRIO-Mul (Liu et al. 2022)	47.63	23.53	44.49	0.8874	47.10	24.52	39.15	0.9238
GPT-3.5$_{\text{zero-shot}}$ (Ouyang et al. 2022)	37.42	13.78	29.10	0.8770	26.63	06.71	18.78	0.8767
TriSum (Jiang et al. 2024)	45.72	22.70	41.93	0.8850	47.33	**24.39**	39.01	0.9217
GPT-3.5$_{\text{TriSum}}$ (Jiang et al. 2024)	46.68	23.48	40.73	0.8920	34.44	12.61	28.43	0.8925
Llama-3-70B$_{\text{zero-shot}}$ (Touvron et al. 2023)	38.56	14.69	30.78	0.8795	33.42	11.66	26.73	0.8916
Llama-3-70B w/ structured rationale	**50.85**	**25.27**	**46.42**	**0.9106**	**49.74**	**27.41**	**40.54**	**0.9325**
LSR-MWF($\leq$1B)	**48.54**	**23.91**	**45.10**	**0.8977**	**47.47**	24.32	**39.28**	**0.9257**

Table 2: Performance comparison of ROUGE and BERTScore scores on CNN/DailyMail and XSum datasets. We highlight the top-2 results in bold font. Our backbone model BART$_{\text{Large}}$ is colored gray for reference.

	CNNDM			XSum		
Model	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
LSR-MWF	48.54	23.91	45.10	47.47	24.32	39.28
w/o EA	48.08	23.13	44.29	47.34	24.01	38.97
w/o AS	47.55	22.87	43.78	46.93	23.58	38.66
w/o TER	46.37	22.73	42.22	45.22	22.95	37.39
w/o AS&TER	44.58	21.37	41.06	45.09	22.37	36.89
w/o EA&TER	45.22	21.89	41.39	44.73	22.25	36.78
w/o EA&AS	44.76	21.58	41.15	46.69	23.46	38.43

Table 3: Ablation study of LSR-MWF. “w/o” means without. “EA” means essential aspects, “AS” means associated sentences, and “TER” means triple entity relations.

Gate Init.	**R-1**	**R-2**	**R-L**	**BS**
0.001	47.21	22.76	43.92	0.8825
0.015	47.74	23.32	44.38	0.8851
0.020	48.54	23.91	45.10	0.8977
0.020*	46.60	20.33	42.83	0.8785
0.025	48.02	23.53	44.76	0.8943
0.100	15.02	5.44	14.72	0.8513

Table 4: Analytical experiments on initialization weights of weak-gatd units on CNNDM. “*” represents that the weights of weak-gated unit are fixed.

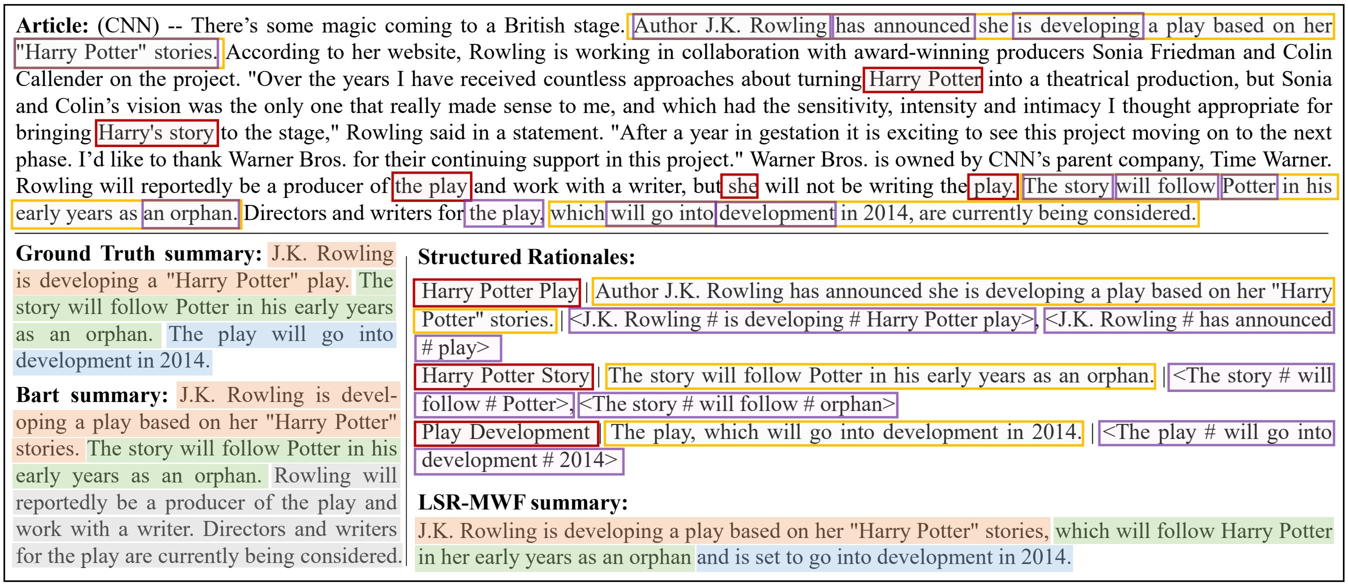
aid the abstractive model in making more accurate inferences and generations in complex contexts.

Analysis

Superiority of Weak-gated Mechanism To observe from Table 4, the initialization weights of weak-gated units should not be excessively large or overly small. Furthermore, we found that when weights of weak-gated units are set to a fixed value and not updated during training, the performance of the model is significantly reduced. This suggests that in the process of integrating structured rationales into the abstractive model, the weights of weak-gated units adaptively and dynamically change to effectively fuse with the features of different decoder layers to improve model’s performance.

Visualization of Weak-gated Mechanism According to Figure 4, we further illustrate the dynamic changes of the weights of weak-gated units during training for LSR-MWF with three types of “gems” inputs. For the CNNDM dataset, it can be observed that the weight fluctuation of the weak-gated units corresponding to the first six decoder layers is relatively small, maintaining a value close to 0.02. However, the fluctuations in the last 6 layers are much larger. Additionally, we note that the weak-gated weights of TER change more rapidly. This may be due to the critical nature of TER, which is highly required by different decoder layers, necessitating quick adaptation of its weights. For XSum, it is evident that only the weak-gated units in the last four layers are active. Notably, the weak-gated units in the final layer exhibit rapid fluctuations across the three layers of EA, AS, and TER. This suggests that the information from the last decoder layer of the abstractive model may be important, requiring the weights to rise rapidly to adapt.

Study of Sequence-level and Token-level Loss Figure 5 indicates that the optimal ratio of γ_1:γ_2 for the CNNDM dataset is 6:4, while the best ratio for the XSum dataset is 7:3. This suggests that CNNDM prefers sequence-level cosine loss, whereas



Article: (CNN) -- There's some magic coming to a British stage. Author J.K. Rowling has announced she is developing a play based on her "Harry Potter" stories. According to her website, Rowling is working in collaboration with award-winning producers Sonia Friedman and Colin Callender on the project. "Over the years I have received countless approaches about turning Harry Potter into a theatrical production, but Sonia and Colin's vision was the only one that really made sense to me, and which had the sensitivity, intensity and intimacy I thought appropriate for bringing Harry's story to the stage," Rowling said in a statement. "After a year in gestation it is exciting to see this project moving on to the next phase. I'd like to thank Warner Bros. for their continuing support in this project." Warner Bros. is owned by CNN's parent company, Time Warner. Rowling will reportedly be a producer of the play and work with a writer, but she will not be writing the play. The story will follow Potter in his early years as an orphan. Directors and writers for the play, which will go into development in 2014, are currently being considered.

Ground Truth summary: J.K. Rowling is developing a "Harry Potter" play. The story will follow Potter in his early years as an orphan. The play will go into development in 2014.

Bart summary: J.K. Rowling is developing a play based on her "Harry Potter" stories. The story will follow Potter in his early years as an orphan. Rowling will reportedly be a producer of the play and work with a writer. Directors and writers for the play are currently being considered.

Structured Rationales:

Harry Potter Play | Author J.K. Rowling has announced she is developing a play based on her "Harry Potter" stories. | <J.K. Rowling # is developing # Harry Potter play>, <J.K. Rowling # has announced # play>

Harry Potter Story | The story will follow Potter in his early years as an orphan. | <The story # will follow # Potter>, <The story # will follow # orphan>

Play Development | The play, which will go into development in 2014. | <The play # will go into development # 2014>

LSR-MWF summary: J.K. Rowling is developing a play based on her "Harry Potter" stories, which will follow Harry Potter in her early years as an orphan and is set to go into development in 2014.

Figure 3: An example of CNNDM. EA, AS and TER are respectively wrapped in red, yellow and purple boxes, like the document of Figure 1. For the three summaries in the figure, the semantically identical parts are colored with the same color.

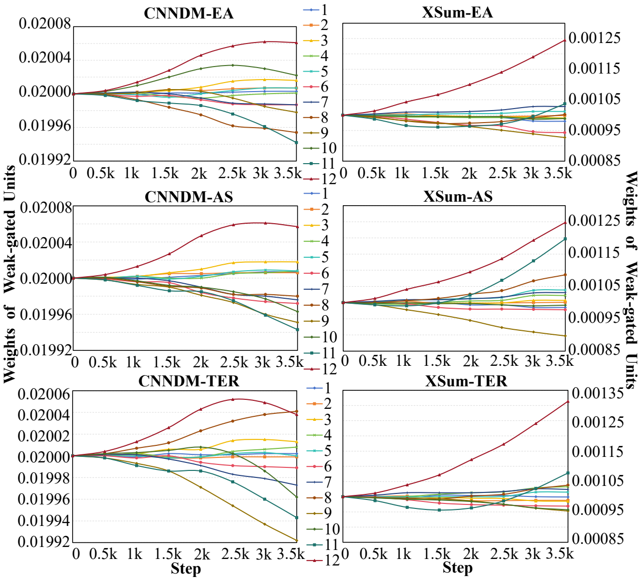


Figure 4: Visualization of weak-gated parameter value.

XSum prefers token-level loss. This difference may be attributed to the higher level of abstraction in the summaries in XSum.

Case Study Figure 3 compares a summary of a CNNDM article discussing the upcoming production of J.K. Rowling's Harry Potter script. On the one hand, BART's summary specifies the characters and storyline in detail, it omits some key information such as the release event and time. On the other hand, the structured rationales of our method are hierarchically progressive, starting with EA, progressing to AS, and finally to TER. The EA presents a high-level overview of the event and its consequences. The AS consists of the most relevant and direct statements from the documents. The TER further refines these statements to clarify the relationship between the entities involved in AS. This technique ensures the completeness of the summary and improves clarity, allowing the reader to follow the content of the summary back to its main aspects and detailed triples for a deeper understanding of the summarization process.

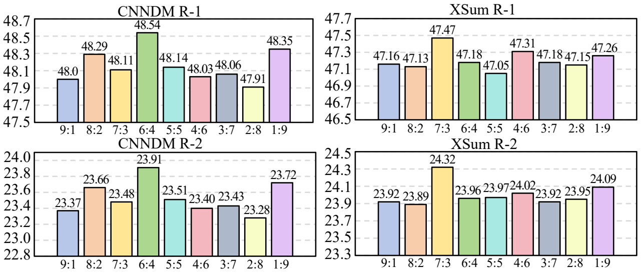


Figure 5: Model performance for different γ_1 and γ_2 coefficients is weighted by loss Eq. (11), each ratio on the abscissa represents $\gamma_1 : \gamma_2$. where $\gamma_1 + \gamma_2 = 1$.

Conclusion

In this work, we propose a distillation method, LSR-MWF, which leverages LLMs and employs specific strategies to obtain high-quality structured rationales. These rationales are then used to hierarchically guide SLMs. Our method bridges the gap between LLMs and SLMs, enabling SLMs to inherit the structured abstractive summarization capabilities of LLMs while maintaining high performance and interpretability. We believe that the model's performance can be further enhanced by refining and extending the structured rationales, such as oppositional or anaphora relations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62202381 and U22B2036, in part by Shenzhen Science and Technology Program and Guangdong Basic and Applied Basic Research Foundation (2024A1515010087), General Program of Chongqing Natural Science Foundation (No. CSTB2022NSCQ-MSX1284), Young Talent Fund of Association for Science and Technology in Shaanxi, China, the National Postdoctoral Innovative Talents Support Program for L. Wu. We would like to thank the anonymous reviewers for their constructive comments.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Avram, A.-M.; Catrina, D.; Cercel, D.-C.; Dascălu, M.; Rebedea, T.; Păiş, V.; and Tufiş, D. 2021. Distilling the knowledge of Romanian BERTs using multiple teachers. *arXiv preprint arXiv:2112.12650*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Dou, Z.-Y.; Liu, P.; Hayashi, H.; Jiang, Z.; and Neubig, G. 2020. GSum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Gekhman, Z.; Herzig, J.; Aharoni, R.; Elkind, C.; and Szpektor, I. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*.

Guo, G.; Han, L.; Wang, L.; Zhang, D.; and Han, J. 2023. Semantic-aware knowledge distillation with parameter-free feature uniformization. *Visual Intelligence*, 1(1): 6.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D. d. L.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-K.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Jia, R.; Cao, Y.; Shi, H.; Fang, F.; Liu, Y.; and Tan, J. 2020. Distilsum: Distilling the knowledge for extractive summarization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2069–2072.

Jia, Z.; Sun, S.; Liu, G.; and Liu, B. 2024. MSSD: multi-scale self-distillation for object detection. *Visual Intelligence*, 2(1): 8.

Jiang, P.; Xiao, C.; Wang, Z.; Bhatia, P.; Sun, J.; and Han, J. 2024. TriSum: Learning Summarization Ability from Large Language Models with Structured Rationale. *arXiv preprint arXiv:2403.10351*.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Kim, Y.; and Rush, A. M. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, H.; and Chaturvedi, S. 2024. Rationale-based Opinion Summarization. *arXiv preprint arXiv:2404.00217*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, P.; Wu, L.; Wang, L.; Guo, S.; and Liu, Y. 2024b. Step-by-Step: Controlling Arbitrary Style in Text with Large Language Models. In *LREC-COLING 2024*, 15285–15295.

Liu, Y.; and Lapata, M. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Liu, Y.; and Liu, P. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.

Liu, Y.; Liu, P.; Radev, D.; and Neubig, G. 2022. BRIO: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Liu, Y.; Shi, K.; He, K. S.; Ye, L.; Fabbri, A. R.; Liu, P.; Radev, D.; and Cohan, A. 2023. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Liu, Y.; Yang, Y.; and Chen, X. 2024. Improving Long Text Understanding with Knowledge Distilled from Summarization Model. In *ICASSP*, 11776–11780. IEEE.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*, 22631–22648. PMLR.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pham, M. Q.; Indurthi, S. R.; Chollampatt, S.; and Turchi, M. 2023. Select, prompt, filter: Distilling large language models for summarizing conversations. In *EMNLP*, 12257–12265.

Radev, D.; Hovy, E.; and McKeown, K. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4): 399–408.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Shleifer, S.; and Rush, A. M. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13693–13696.

Sun, X.; Ren, X.; and Xie, X. 2024. A Novel Multimodal Sentiment Analysis Model Based on Gated Fusion and Multi-Task Learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8336–8340. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want to reduce labeling cost? GPT-3 can help. *arXiv preprint arXiv:2108.13487*.

Wang, Y.; Zhang, Z.; and Wang, R. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wen, Z.; Tian, Z.; Huang, Z.; Yang, Y.; Jian, Z.; Wang, C.; and Li, D. 2023a. GRACE: Gradient-guided Controllable Retrieval for Augmenting Attribute-based Text Generation. In *Findings of ACL 2023*, 8377–8398.

Wen, Z.; Tian, Z.; Wu, W.; Yang, Y.; Shi, Y.; Huang, Z.; and Li, D. 2023b. GROVE: A Retrieval-augmented Complex Story Generation Framework with A Forest of Evidence. In *Findings of EMNLP 2023*, 3980–3998.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

Wu, L.; Liu, P.; Yuan, Y.; Liu, S.; and Zhang, Y. 2023a. Context-aware style learning and content recovery networks for neural style transfer. *Information Processing & Management*, 60(3): 103265.

Wu, L.; Liu, P.; Zhao, Y.; Wang, P.; and Zhang, Y. 2023b. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 211–225.

Xu, S.; Zhang, X.; Wu, Y.; and Wei, F. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 11556–11565.

Yao, F.; Sun, X.; Yu, H.; Yang, Y.; Zhang, W.; and Fu, K. 2020. Gated hierarchical multi-task learning network for judicial decision prediction. *Neurocomputing*, 411: 313–326.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. *NeurIPS*, 33: 17283–17297.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, 11328–11339. PMLR.

Zhang, Y.; Zhang, H.; Nasrabadi, N. M.; and Huang, T. S. 2013. Multi-metric learning for multi-sensor fusion based classification. *Information Fusion*, 14(4): 431–440.

Zhou, W.; Xu, C.; and McAuley, J. 2021. BERT learns to teach: Knowledge distillation with meta learning. *arXiv preprint arXiv:2106.04570*.