Hierarchical Attention Adapter for Abstractive Dialogue Summarization

Raymond Li[†] **, Chuyuan Li**[†] **, Gabriel Murray**[‡] **, Giuseppe Carenini**[†]
[†] University of British Columbia, Vancouver, BC, Canada
[‡] University of Fraser Valley, Abbotsford, BC, Canada
{raymondl, carenini}@cs.ubc.ca, chuyuan.li@ubc.ca
gabriel.murray@ufv.ca

Abstract

Dialogue summarization is still a very challenging task even for large language models (LLMs). On the one hand, some previous approaches have pre-trained language models specifically for dialogue understanding and summarization, but they have been limited to relatively small language models such as BART and T5. On the other hand, other works have tried to directly exploit the dialogue semantics and discourse structures in their modeling effort, but by construction, they require access to those structures, which is in itself a largely unsolved problem. In this paper, we synergistically combine these two ideas in an approach that can be seamlessly integrated into the decoder-only architecture adopted by the most state-of-the-art LLMs. In particular, our novel solution leverages the parameter-efficient fine-tuning (PEFT) paradigm to model the hierarchical structure of dialogues, where input sequences are naturally segmented into dialogue turns, and then fine-tune the model for abstractive summarization. From experiments on two datasets, we find that Hierarchical Attention Adapter outperforms all baseline adapter methods on SummScreen, where our approach can also be combined with LoRA to achieve the best performance on SamSum.

1 Introduction

The explosion in real-time messaging, consultation forums, and online meetings has resulted in a vast amount of conversational data, necessitating more efficient methods for understanding and extracting key information. Dialogue summarization, which aims to automatically distill salient information from dialogues, has been widely applied across various scenarios in different domains. These include task-oriented dialogues such as customer service (Feigenblat et al., 2021), law (Duan et al., 2019), medical care (Joshi et al., 2020), and open-ended dialogues like chit-chat (Chen et al., 2021b), screen plays (Chen et al., 2021a), and forum discussions (Chowdhury and Chakraborty, 2019).

Conventional dialogue summarization models typically approach the task as a sequence-to-sequence problem and fine-tune encoder-decoder models such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020). Although these methods show promising results (Liu et al., 2021a; Wang et al., 2023; Zhong et al., 2022; Cho et al., 2024), they are limited to relatively small Transformer architecture models and cannot be easily employed on large decoder-only models. Another line of research examines dialogue specific features such as speaker marker, discourse structure, topic changes, and coreference information (Chen and Yang, 2021a; Liu et al., 2021b; Cho et al., 2024) and incorporates dialogue semantics and structural information into model pre-training, oftentimes involving an intermediate stage for dialogue-specific information extraction. However, these methods require additional effort and the results constructed from such information may not be accurate.

On the other hand, recent advances in Large Language Models (LLMs) have revolutionized the field of NLP and have become an essential building block in various intelligent user-facing applications (Bang et al., 2023; Bubeck et al., 2023), resulting in a shift in focus from relatively small encoder-decoder models to large-scale decoder-only models. The remarkable achievements of these LLMs can largely be attributed to research on model scaling (Brown et al., 2020; Chowdhery et al., 2023; Workshop et al., 2022), where increasing the number of model parameters and the volume of pre-training data can lead to significant enhancements of their capabilities to understand and generate human language. However, despite the success of sophisticated prompting (Wei et al., 2022; Zhou et al., 2023) and demonstration selection (Lewis et al., 2020b; Rubin et al., 2022) strategies, there remains a noticeable performance gap compared to

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 17–30
November 8, 2025 ©2025 Association for Computational Linguistics

fine-tuning (Liu et al., 2022; Mosbach et al., 2023), especially for tasks such as dialogue summarization where the input sequences can be long and possess hierarchical structures.

In this paper, we synergistically combine the ideas of using modern LLMs with dialogue structural information in dialogue summarization task, without explicitly injecting rigid linguistic structures. To this end, we select GPT-style decoder-only architecture as our LLM backbone. This simplified architecture allows for more efficient pretraining through the language modeling objective, where the model can quickly process and generate tokens without first transforming the input sequence into an abstract representation by the encoder. However, an innate drawback of such architecture is the token-level unidirectional flow, which limits the ability to model the full context of the dialogue, especially the nuances and dependencies that emerge from future turn back to past ones. By contrast, encoder-decoder architecture encodes the input sequence using bidirectional conditioning and results in inherently stronger representations compared to the causal conditioning representation. In order to fine-tune decoder-only models more effectively, we propose a novel parameter-efficient fine-tuning (PEFT) architecture that encodes the input with bidirectional contextualization.

Another great challenge in dialogue summarization is the length of text, which can sometimes exceed the model's input limit. While recent efforts have managed to increase the context window of the model by scaling the positional embeddings (Press et al., 2022; Su et al., 2024) and reducing the complexity of attention through sparsity (Child et al., 2019; Jaszczur et al., 2021), the theoretical support for long context are often measured with the language modeling loss and synthetic tasks, which do not comprehensively demonstrate their effectiveness in practical applications. Inspired by previous studies (Nguyen et al., 2020; Madaan et al., 2023; Du et al., 2023) demonstrating that hierarchical structures of input can significantly enhance downstream performance, especially for tasks with naturally segmented input sequences, we propose encoding dialogues at both the speech turn level and the dialogue level. This is achieved through two attention layers in our adapter module. Our proposed adapter, called **Hierarchical Attention Adapter**, can incorporate the interactions of speech turns in dialogues naturally without the need for external structural integration.

To summarize, our contributions are threefold: (1) We propose a novel PEFT architecture: Hierarchical Attention Adapter, that incorporates bidirectional contextualization to model the hierarchical structure of the dialogue sequence; (2) Our experiments on two dialogue summarization datasets demonstrate the effectiveness of our approach, where we achieve the best overall performance over the baseline either using our method directly or combining with LoRA; (3) We analyze the importance of each layer of adapters and ways to represent dialogue rounds, and provide useful insights for future PEFT approaches.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization, aimed at distilling the salient information of dialogue into a concise summary, has received more attention as virtual conversations have become increasingly prevalent (Jia et al., 2023). While the standard approach for neural abstractive summarization follows the sequence-to-sequence generation paradigm (Sutskever et al., 2014), where an autoregressive model generates the summary conditioned on the input text (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Lewis et al., 2020a; Zhang et al., 2020), studies on dialogue summarization often exploited the semantic and discourse graph structures of the input by explicitly incorporating those structures to model the high-level interactions between dialogue segments (Hardy and Vlachos, 2018; Chen and Yang, 2021b; Hua et al., 2023). However, these methods typically require additional steps to obtain the dialogue graphs making them impractical for most application scenarios. Other studies have proposed pre-training language models specifically for dialogue understanding and summarization (Gu et al., 2021; Zhong et al., 2022), their scales are order-of-magnitude smaller compared to the general-purpose Large Language Models (Touvron et al., 2023b; Jiang et al., 2023) resulting in often inferior performance in poor generalizability. In contrast, our approach can be seamlessly integrated into the decoder-only architecture adopted by the most state-of-the-art LLMs.

2.2 Modeling Structures

While Graph Neural Networks (GNNs) have traditionally been the de-facto standard for graph representation, the Transformer model has garnered a

lot of popularity due to its superior performance in modeling graph structures (Hussain et al., 2022; Wu et al., 2022, 2023). Similar to GNNs, the self-attention mechanism in the Transformer encoder aggregates the node embeddings to update the representation of each node in a fully connected graph. In the domain of NLP, one popular area of research is to model explicitly defined linguistic structures within the input sequence. While a majority of studies have focused on sentence-level dependency or constituency trees (Wu et al., 2018; Hao et al., 2019; Strubell et al., 2018; Wang et al., 2019a,b), multi-sentential discourse structures have also been found to be beneficial for more practical downstream tasks such as summarization (Xiao et al., 2020; Xu et al., 2020; Feng et al., 2020; Dong et al., 2021). For instance, Chen and Yang (2021b) incorporated discourse relations between utterances and action for dialogue summarization, while Du et al. (2023) utilized discourse structures to propagate hidden representation for question answering. Meanwhile, another line of work aims to implicitly learn the structure of language through architectural design, allowing the model to have more flexibility in learning representations beneficial for the task (Nguyen et al., 2020; Madaan et al., 2023). Given the substantial proficiency of LLMs in representing and comprehending natural language, it is questionable whether the integration of rigid linguistic structures, which often require multi-stage pipelines to acquire, can further enhance the model's capabilities in downstream tasks. Therefore, our proposed technique aims to allow the model to learn multi-sentential structures based on the internal representations of the LLM.

2.3 Large Language Models

While earlier work on masked language models are often designed to encode a contextualized representation of the input sequence (Peters et al., 2018; Devlin et al., 2019), autoregressive pre-training was found to be much effective for language generation tasks (Lewis et al., 2020a; Raffel et al., 2020). The current lineage of large language models can trace their origins to the GPT-family (Radford et al., 2018, 2020; Brown et al., 2020), where they find that scaling the decoder-only architecture can lead to an improved model capacity on downstream tasks (following scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022)) while displaying emergent abilities to solve complex tasks through few-shot demonstrations (Brown et al., 2020) and chain-of-thought prompting (Wei et al., 2022). Following the success of proprietary LLMs (e.g., ChatGPT, Claude, etc.) as general-purpose task solvers through instruction-tuning (Ouyang et al., 2022), the open-source community has also released a large number of publicly available models for researchers to work on (Zhang et al., 2022; Bai et al., 2023; Almazrouei et al., 2023). For example, Llama (Touvron et al., 2023a) is one of the first collections of open-sourced models pre-trained exclusively using publicly available datasets, Llama-2 (Touvron et al., 2023b) improved upon the Llama by training on more data and doubling the context length from 2K to 4K tokens, XGen (Nijkamp et al., 2023) further increased to context length to 8K through pre-training in stages with increasing sequence length, while Mistral (Jiang et al., 2023) used sliding window attention (Child et al., 2019) to support a theoretical attention span of approximately 131K tokens. While other techniques have also been proposed to improve the long-context capabilities of LLMs through position encoding refinement and continual pre-training (Xu et al., 2023; Xiong et al., 2023), these techniques require tuning the full set of parameters which is far too expensive for practical applications.

3 Method

3.1 Hierarchical Attention Adapter

An overview of the architecture is presented in Figure 1. In our hierarchical attention adapter, we first project the LLM hidden states to dimension d using a single linear layer. At each output time step j, we decompose the projected hidden states based on input and out tokens where $H = [h_i; h_o]$, with $h_i \in \mathbb{R}^{n \times d}$ and $h_o \in \mathbb{R}^{j \times d}$ representing the hidden states of input and output tokens respectively. We model the hierarchical structure of the input dialogue sequence using a hierarchical self-attention module. To compute the dialogue turn embeddings $H_t \in \mathbb{R}^{t \times d}$, where t is the number of dialogue turns, we first apply an attention encoder to the hidden states of each dialogue turn. Finally, to model the coarse interactions between dialogue turns, we use another layer of attention layer on the turn embeddings to obtain the hierarchical representation H_d of the dialogue sequence. This module allows us to construct representations of the input sequence with bidirectional contextualization while incorporating the hierarchical structure of the dialogue. The hierarchical representation is ex-

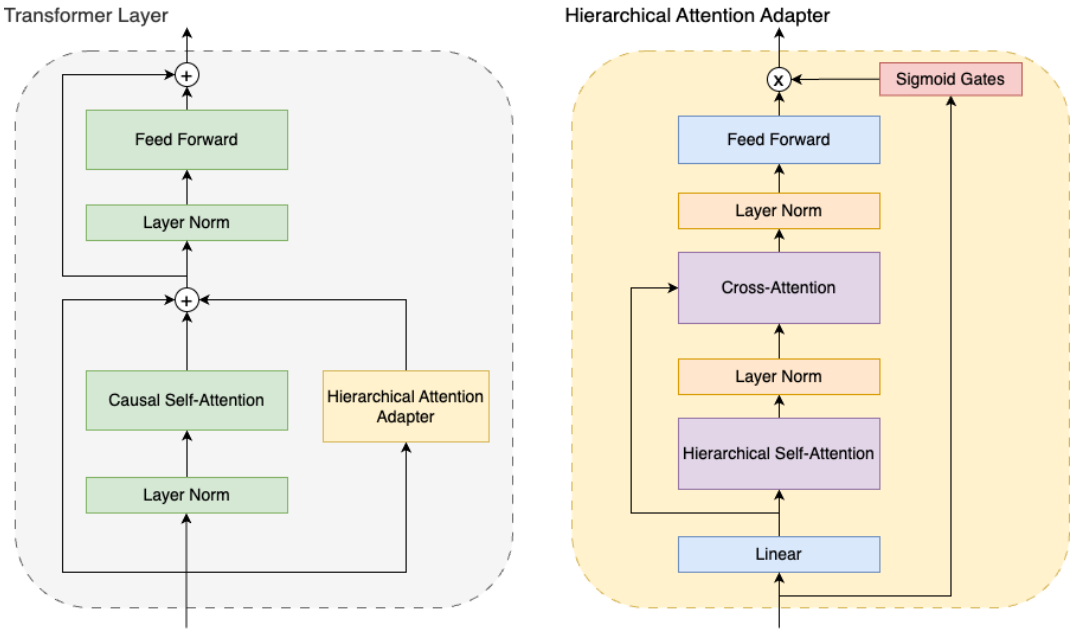

Figure 1: Overview of the Hierarchical Attention Adapter inserted in-parallel to the causal self-attention layer of the decoder transformer model. Specifically, the Hierarchical Self-Attention contains two attention layers: one encodes the hidden state of individual speech turn and one encodes the interactions of different speech turns.

pressed in Equation 1, where M_i denotes the mask for selecting the i^{th} dialogue turn.

$$H_t = [\text{Attn}(M_i \odot h^{(i)}); \ldots; \text{Attn}(M_t \odot h^{(i)})] \quad (1)$$

$$H_d = \text{Attn}(H_t) \quad (2)$$

To enable the decoder to use the hierarchical representation of the input sequence, we follow the intuition of the cross-attention mechanism in the original transformer architecture (Vaswani et al., 2017). Specifically, we compute the weighted sum of the projected hidden states using a cross-attention layer with the hierarchical representation H_d as query and H as key and value. Finally, we use a two-layer feed-forward module to project the adapter representation to the dimension of LLM hidden states. The final adapter representation H_a can be expressed in Equation 3.

$$H_a = \text{FF}\big(\text{Attn}(Q = H^{(d)}, K = V = H)\big) \quad (3)$$

3.2 Gated Addition

Following the work by Mao et al. (2022), we use a gated additive method to inject the adapter output into the hidden states of the LLM. To do so, our learnable gate consists of a two-layer feed-forward module that takes the previous LLM hidden layer as input and uses a sigmoid activation function to map the final output to a value between 0 and 1. The gate acts as a learnable scaling factor that estimates the importance of our module based on the input hidden states. We insert our adapter in parallel with the LLM self-attention layer, where the gated adapter output is combined with the attention output and residual connection through element-wise addition.

4 Experiments

4.1 Model

We adopt Mistral-7B (Jiang et al., 2023) as the base model for all our experiments. This model has 32 hidden layers (32 heads per layer) and uses the sliding window attention (SWA) (Child et al., 2019) with a window size of 4096 to support sequences of up to 4096×32 tokens. The main reason for choosing this model is due to the memory efficiency of SWA, as the standard self-attention mechanism has a quadratic complexity w.r.t. the sequence length.

4.2 Datasets and Metrics

We choose two widely used dialogue datasets: SamSum (chit-chat) (Gliwa et al., 2019) and SummScreen (screen plays) (Chen et al., 2022). From the statistics in Table 1, we can see that SamSum is an easier dataset with both short dialogues and reference summaries, while SummScreen is a much harder dataset with both extremely long dialogues and summaries requiring the model to learn the

Dataset	Domain	Dialogue	Summary	Train Size	Validation Size	Test Size
SamSum	Chat	83.9	20.3	14,731	818	819
SummScreen	Screenplay	6,612.5	337.4	18,915	1,795	1,793

Table 1: Statistics of the three datasets used in our experiments. The number of sentences in original dialogues and reference summaries, the number of documents in train, validation, and test sets are reported.

long-term dependencies between turns to generate a coherent summary. For SummScreen, we remove ultra-long examples (>16,000 tokens) to avoid out-of-memory during training.

For evaluation metrics, we use the popular ROUGE (Lin, 2004), which measures n-gram overlaps with the reference summary, as well as GPT3Score (Fu et al., 2023), which employs generative pre-trained models to evaluate text quality by calculating the length-normalized conditional log probability of the evaluated text (reference given candidate and candidate given reference) given task-specific prompts and aspect definitions. For ROUGE metric, we compute ROUGE-1, ROUGE-2, and ROUGE-Lsum (all are F1 scores) using the `rouge-score`[1] package, which respectively measures the overlap of unigram, bigram, and the longest common sub-sequence for each sentence. Following Grusky (2023), we compute the ROUGE scores without stemming and stopword removal, which is consistent with the original `ROUGE-1.5.5` implementation by Lin (2004). For GPTScore, we use `text-davinci-002` (Brown et al., 2020) since it is currently the most powerful text completion model accessing through the OpenAI API[2] that supports token probabilities and has shown to be highly correlated with human judgment (Fu et al., 2023). We compute the harmonic mean of the conditional probability for the candidate summary predicted by the reference and vice versa. The conditional probabilities are computed based on three aspects, namely, Informativeness (I), Naturalness (N), and overall Quality (Q).

4.3 Baselines

To evaluate the performance of our proposal, we compare against three baseline PEFT methods, namely, Low-Rank Adaptation (LoRA) (Hu et al., 2022), Bottleneck Adapter (Houlsby et al., 2019), and the standard Attention Adapter. In particular, LoRA injects trainable rank decomposition matrices to approximate the gradient updates during fine-tuning, Bottleneck Adapter injects two-layer MLPs sequential to the self-attention and feed-forward modules of the LLM, while Attention Adapter is a standard decoder attention layer with casual self-attention followed by a feed-forward layer. From results presented in prior studies (Yu et al., 2023), we expect both LoRA and the Bottlenet Adapter to be competitive baselines.

4.4 Hyperparameter Settings

We apply our Hierarchical Attention Adapter in parallel to the self-attention sub-layer of the decoder LLMs. We first project the LLM hidden states to our adapter dimension of $d_{\text{adapter}} = 128$, before applying our hierarchical and cross-attention modules with 4 attention heads each. Finally, we apply the two-layer feed-forward module with the same SiLU activation (Elfwing et al., 2018) as Mistral. Following the settings by Mao et al. (2022), we train our model with AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $1e^{-4}$ with linear decay to 0 and a batch size of 64. For the baseline methods, we apply rank 32 LoRA vectors to the query and key vectors, following a hyper-parameter search for rank $\in \{8, 16, 32, 64\}$. For both the Bottleneck and Attention Adapter, we use the same hidden size of 128, where the Attention Adapter has the same number of attention heads (4) and learnable gate dimension as our Hierarchical Attention Adapter. All our experiments are conducted on machines with 4×A100 SXM4 GPU (40GB of memory).

To reduce the number of parameters and improve task performance, we also experiment with combining LoRA with our Hierarchical Attention Adapter. Following the findings by Li et al. (2023), where they find that only the top layers of the pre-trained model can effectively utilize injected adapters, we inject our adapters to the top-4 (excluding the last) layers of the LLM when combined with LoRA.

4.5 Results

From the results presented in Table 2, we see that combining LoRA with Hierarchical Attention

[1] https://pypi.org/project/rouge-score
[2] https://platform.openai.com/docs/models

	SamSum						SummScreen					
	ROUGE			GPTScore			ROUGE			GPTScore		
Model	R1	R2	RLs	I	N	Q	R1	R2	RLs	I	N	Q
LoRA	.543	.303	.504	-.913	-.901	-.891	.305	.085	.295	-.666	-.667	-.666
Bottleneck	.540	.300	.501	-.909	-.895	-.887	.278	.068	.265	-.780	-.781	-.779
Attention	.519	.275	.480	-.930	-.915	-.907	.263	.078	.249	-.653	-.652	-.652
Ours	.545	.303	.507	-.911	-.897	-.888	**.334**	**.089**	**.321**	**-.632**	**-.633**	**-.632**
Ours + LoRA	**.546**	**.306**	**.508**	**-.902**	**-.889**	**-.881**	.326	.089	.318	-.702	-.703	-.704

Table 2: Evaluation comparisons with baseline models and our methods across three datasets using ROUGE and GPTScore metrics. R1, R2, RLs denote resp. ROUGE-1, ROUGE-2, and ROUGE-Lsum; I, N, and Q denote resp. Informativeness, Naturalness, and Quality.

Model	R1	R2	RL	RLs
ConDigSum (2021a)	.543	.293	.452	-
GPT3-finetuned (2022)	.534	.298	.459	-
ChatGPT (2023)	.408	.137	.315	-
InstructDS (2023)	**.553**	**.313**	**.467**	-
Ours	.545*	.303*	.466*	**.507**

Table 3: Comparison with state-of-the-art benchmarks on SAMSum dataset. We report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum metrics. * not significantly different with the best score (in bold).

Model	R1	R2	RL	RLs
DialogLM-sparse (2022)	.358	.083	.187	-
SPECTRUM (2024)	**.358**	**.095**	**.212**	-
Ours	.334	.089	.169	.321

Table 4: Comparison with state-of-the-art benchmarks on SummScreen. We report ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-LSum metrics.

Adapter (*LoRA + Ours*) achieved the best results on SamSum and using only the Hierarchical Attention Adapter achieved the best results on SummScreen. We also see that our method (with and without LoRA) achieves the highest improvement over the baselines on the SummScreen dataset. This is mainly due to the fact that the examples in SummScreen have the longest average input and output length, making the hierarchical representation more beneficial even without LoRA. This finding also confirms our hypothesis that decoder-only architectures struggle with significantly long contexts. On the contrary, we see that our model (Ours) does not outperform the baselines in the SamSum datasets, where we only see improvements when combined with LoRA. We hypothesize that due to the short context of SamSum, the casual self-attention of the decoder can sufficiently encode the input dialogue for summarization. This could be due to the fact that hierarchical attention helps model the dialogue context while tuning the decoder self-attention (via LoRA) improves the context selection for generation. This leads to long summaries generated by our model, and partially explains the poor performance (lower precision) of our method on SamSum, where the short output length (20 tokens) requires the model to precisely generate the summary based on the context.

In addition, for reference, we present the performances of state-of-the-art models fine-tuned specifically for dialogue summarization task without using the PEFT paradigm. The benchmarks on SAMSum are (1) topic-aware BART model trained with contrastive learning ConDigSum (Liu et al., 2021a), (2) GPT3 fine-tuned with LoRA (Hu et al., 2022), (3) ChatGPT with instruction tuning (Wang et al., 2023), and (4) Flan-T5-XL with instruction tuning (Wang et al., 2023). As shown in Table 3, our method is comparable with the SOTA InstructDS model on all ROUGE metrics, demonstrating the effectiveness of our hierarchical adapter on decoder-only model. On SummScreen, the best performing models are given in Table 4: (1) DialogLM (Zhong et al., 2022) – an encoder-decoder model pre-trained using dialogue-tailored noise; (2) SPECTRUM (Cho et al., 2024), a speaker-enhanced model pre-trained on PEGASUS. These models are specially trained with dialogue data, while our approach can applied to any general-purpose LLMs. With the scaling effect, we expect larger decoder-only models to lead to greater improvements.

	R-1	R-2	R-L	GPTScore (Ave)
Last Token	0.534	0.295	0.496	-0.902
Mean-Pooling	0.541	0.303	0.502	-0.893
Attention	**0.546**	**0.306**	**0.508**	**-0.891**

Table 5: Results on SamSum for different turn embedding methods. R-1/2/L refer to ROUGE-1/2/Lsum metrics.

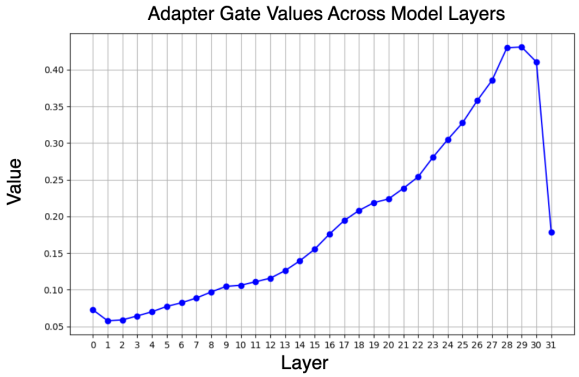

Figure 2: Gate value for our adapter at each layer of the Mistral-7B model. Values are averaged on the development set of SamSum.

5 Analysis

5.1 Dialogue Turn Representations

We first study the techniques for constructing the representations for dialogue turns. Using the SamSum dataset, we use three different turn representations while keeping the rest of the architecture identical. From the results in Table 5, we see that using an additional layer of attention outperforms mean-pooling and using the last token of each turn. It is worth noting that not only is using the last token representation outperforms mean-pooling. We believe that it is mainly due to the autoregressive nature of the decoder-only architecture, where the representation of each token encodes information from all prior tokens.

5.2 Summary Length

We find that adapter-based models tend to generate longer summaries than LoRA. Conversely, overly brief summaries tend to omit salient information; we provide concrete examples in subsection 5.4. In particular, on SamSum, the average number of summary tokens of our Hierarchical Attention Adapter is 23, while LoRA produces an average of 18 tokens. However, when combining LoRA with our method, the predicted summary has an average of 20 tokens. On the SummScreen dataset, the average summery length for LoRA is 460 while the length for Hierarchical Attention Adapter and Hierarchical Attention Adapter + LoRA are 526 and 549, respectively.

5.3 Gate Values

To assess the relative importance of our adapter at each layer of the model, we analyze the gate output by computing the average absolute values over the development set of SamSum. For the results presented in Figure 2, we see that the average gate value gradually increases before drastically dropping off at the last layer. This is in contrast to the findings by previous work on encoder models (Rücklé et al., 2021; Li et al., 2023), where they reported performance benefits from inserting adapter modules in the final layer of the pre-trained model. We hypothesize the main reason is that since the last layer of the pre-trained decoder is directly used to predict the next word distribution over the vocab, the model cannot effectively learn to use the modified representation (from adapters) to encode the fine-grained contextual nuances required for word prediction. We believe these insights will benefit future studies on developing PEFT methods for pre-trained decoder models.

5.4 Qualitative Comparison

We perform a qualitative comparison between the summary generated by LoRA and our Hierarchical Attention Adapter. From the output snippets presented in Table 6, we see that on SamSum, our model tends to be more verbose during prediction, resulting in longer summaries that contains richer information. For the first example of SummScreen, we see that while both generated summaries do not capture the same content as the reference summary, the summary generated by LoRA begins to repeat towards the end. In contrast, the summary generated by our method captures more aspects of the conversation as it introduced six characters (rather than the three for LoRA). For the second example, we see that while both summaries contain information about *Kendall's trial*, the summary generated by LoRA also becomes repetitive towards the end. We hypothesize that this could be due to the intrinsic capacity of LoRA, as it only uses a small set of parameters to update the self-attention weights.

6 Conclusions and Future Work

In this work, we present Hierarchical Attention Adapter, a novel PEFT architecture that uses bidi-

Summary Type	Generated Summary
SamSum	
Reference	Kate broke her arm and she's going to the hospital. She'd like to know whether her medical insurance covers hospital costs. Greg suggests her to call Linda or ask someone at the reception about it.
LoRA	Kate broke her arm and is going to the hospital.
Ours	Kate broke her arm and is being sent to the hospital. She wants to know if their medical insurance covers hospital costs.
Reference	Casey got a new nail polish and did her nails herself. It took her nearly 4 hours, so she won't do her friends' nails, as it takes too long.
LoRA	Casey did her nails herself. It took her 4 hours.
Ours	Casey did her nails herself and they took her 4 hours. Amelia and Kristen want her to do their nails too but she refuses.
SummScreen	
Reference	Adam bails Krystal out of jail and makes her an "offer" in exchange for her telling him where their mutual grandson is. She denies that she knows he's alive and refuses his offer. But JR gets Bianca to admit that he is alive...
LoRA	Jonathan and Greenlee continue to argue about Ryan's actions. Jonathan wants to know if Greenlee is going to stand by Ryan. Greenlee tells Jonathan that she will stand by Ryan. Jonathan tells Greenlee that he wants to know if she is going to stand by him...
Ours	Greenlee and Jonathan argue about Ryan and Kendall. Maggie comes in and tells them that she just saw Bianca. Greenlee and Jonathan argue about Kendall. Maggie comes in and tells them that she just saw Bianca...
Reference	Kendall's trial begins. Bianca urges Kendall to listen to what she learned in her dream about something she needs to find in order to help Kendall get acquitted. Jack proposed to Erica and she accepts. Edmund suddenly regains consciousness...
LoRA	Kendall is nervous about her trial. Bianca tells her that she had a dream that she was at Michael's condo the night he was killed. Kendall tells her that it was impossible. Bianca tells her that she is worried about her. Kendall tells her that she is worried about her...
Ours	Kendall is in court for her trial. Bianca tells Kendall that she had a dream that she was at Michael's condo the night that he died. Kendall tells Bianca that she cannot believe that she slept the whole day...

Table 6: Comparison of generated summary between LoRA and our Hierarchical Attention Adapter for two examples in the SamSum (top 2) and SummScreen (bottom 2) dataset. LoRA tends to produce overly brief summaries on SAMSum and repetitive content on ScreenSumm, indicating a limitation in generating longer outputs.

rectional contextualization to model the hierarchical structure of the input sequence for dialogue summarization task. Experiments on two datasets show that our proposed method outperforms other baselines for summaries with long context and achieves the best overall performance when combined with LoRA. We perform analysis on the average gate value to assess the relative importance of our adapter at each layer of the model and find that while the adapters of upper layers have higher importance, the model learns to not use the final layer since it is used for computing the next-token probability.

For future work, we wish to perform additional analysis to study the usefulness of the different components in our proposed architecture and perform further experiments on additional datasets. We will also perform further evaluation to compare the faithfulness and factuality of summaries generated by different models. Lastly, while our

current proposal requires segment annotation of the input sequence (i.e., dialogue turns), we intend to extend our approach to implicitly learn the segment boundaries during training and generalize to other summarization tasks such as scientific paper summarization.

Limitations

Automatically assessing the quality of dialog summaries is a huge challenge. We recognize the importance of manual annotators for results comparison. However, human evaluation is costly and inefficient. While most of the dialogue summarization work relies heavily on the ROUGE score, we also report on the GPTScore with instruction prompts, which is an automatic metric that gives multi-faceted evaluation and is closely related to human judgment.

Other types of dialogue summarization tasks such as meeting summarization (e.g., AMI (Kraaij et al., 2005) ICSI meeting corpus (Shriberg et al., 2004)) often do not have enough examples for sufficiently fine-tuning an LLM. Real-world meeting dialogues often span multiple topics and include disfluencies, interruptions, and other artifacts. These characteristics make summarization more realistic but also more challenging. We plan to address these practical issues in future work.

Lastly, while PEFT methods such as LoRA are well optimized in existing libraries such as Ollama[3], adapter-based methods often lack in inference speed due to the computation of additional modules. We hope our work can motivate future studies to efficiently integrate adapter modules into LLMs.

Ethical Considerations

We have taken proactive steps to address ethical concerns related to our research. Our testing datasets were carefully selected to minimize potential issues with biased or hateful content. If this method is applied to new datasets that involve recording multi-party dialogues, informed consent should be obtained from all participants. Because dialogue data may contain sensitive personal information, we urge caution in such applications, especially in summarization.

[3]https://github.com/ollama/ollama

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback and suggestions. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jiaao Chen and Diyi Yang. 2021a. 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.

Jiaao Chen and Diyi Yang. 2021b. 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. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021a. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021b. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Sangwoo Cho, Kaiqiang Song, Chao Zhao, Xiaoyang Wang, and Dong Yu. 2024. Spectrum: Speaker-enhanced pre-training for long dialogue summarization. *arXiv preprint arXiv:2401.17597*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tanya Chowdhury and Tanmoy Chakraborty. 2019. Cqasumm: Building references for community question answering summarization corpora. In *Proceedings of the ACM india joint international conference on data science and management of data*, pages 18–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. 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. Association for Computational Linguistics.

Haowei Du, Yansong Feng, Chen Li, Yang Li, Yunshi Lan, and Dongyan Zhao. 2023. Structure-discourse hierarchical graph for conditional question answering on long documents. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6282–6293. Association for Computational Linguistics.

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1361–1370.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm-a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2020. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *arXiv preprint arXiv:2012.03502*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New*

Frontiers in Summarization, pages 70–79. Association for Computational Linguistics.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. 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 3682–3692. Association for Computational Linguistics.

Jie Hao, Xing Wang, Shuming Shi, Jinfeng Zhang, and Zhaopeng Tu. 2019. Multi-granularity self-attention for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 887–897. Association for Computational Linguistics.

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

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*.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883. Association for Computational Linguistics.

Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. 2022. Global self-attention as a replacement for graph convolution. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 655–665, New York, NY, USA. Association for Computing Machinery.

Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. 2021. Sparse is enough in scaling transformers. In *Advances in Neural Information Processing Systems*.

Qi Jia, Yizhu Liu, Siyu Ren, and Kenny Q. Zhu. 2023. Taxonomy of abstractive dialogue summarization: Scenarios, approaches, and future directions. *ACM Comput. Surv.*, 56(3).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Raymond Li, Gabriel Murray, and Giuseppe Carenini. 2023. Mixture-of-linguistic-experts adapters for improving and interpreting pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9456–9469. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Junpeng Liu, Yanyan Zou, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Caixia Yuan, and Xiaojie Wang. 2021a. Topic-aware contrastive learning for abstractive dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1229–1243.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021b. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lovish Madaan, Srinadh Bhojanapalli, Himanshu Jain, and Prateek Jain. 2023. Treeformer: Dense gradient trees for efficient attention computation. In *The Eleventh International Conference on Learning Representations*.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 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. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. 2020. Tree-structured attention with hierarchical accumulation. In *International Conference on Learning Representations*.

Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. 2023. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2020. Language models are unsupervised multitask learners. *OpenAI*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Xing Wang, Zhaopeng Tu, Longyue Wang, and Shuming Shi. 2019a. Self-attention with structural position representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1403–1409. Association for Computational Linguistics.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019b. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Qitian Wu, Wentao Zhao, Zenan Li, David Wipf, and Junchi Yan. 2022. Nodeformer: A scalable graph structure learning transformer for node classification. In *Advances in Neural Information Processing Systems*.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. 2023. Simplifying and empowering transformers for large-graph representations. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wei Wu, Houfeng Wang, Tianyu Liu, and Shuming Ma. 2018. Phrase-level self-attention networks for universal sentence encoding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3729–3738. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

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. Association for Computational Linguistics.

Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830*.

Bruce Yu, Jianlong Chang, Lingbo Liu, Qi Tian, and Chang Wen Chen. 2023. Towards a unified view on visual parameter-efficient transfer learning.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.