

# LONGLoRA: EFFICIENT FINE-TUNING OF LONG-CONTEXT LARGE LANGUAGE MODELS

Yukang Chen<sup>1</sup> Shengju Qian<sup>1</sup> Haotian Tang<sup>2</sup> Xin Lai<sup>1</sup>  
 Zhijian Liu<sup>2</sup> Song Han<sup>2</sup> Jiaya Jia<sup>1</sup>  
<sup>1</sup>CUHK <sup>2</sup>MIT

## ABSTRACT

We present LongLoRA, an efficient fine-tuning approach that extends the context sizes of pre-trained large language models (LLMs), with limited computation cost. Typically, training LLMs with long context sizes is computationally expensive, requiring extensive training hours and GPU resources. For example, training on the context length of 8192 needs  $16\times$  computational costs in self-attention layers as that of 2048. In this paper, we **speed up the context extension** of LLMs in two aspects. On the one hand, although *dense global* attention is needed during inference, *fine-tuning the model can be effectively and efficiently done by sparse local attention*. The proposed **shift short attention** (S<sup>2</sup>-Attn) effectively enables context extension, leading to non-trivial computation saving with similar performance to fine-tuning with vanilla attention. Particularly, it can be implemented with only *two lines of code* in training, while being optional in inference. On the other hand, we **revisit the parameter-efficient fine-tuning regime for context expansion**. Notably, we find that LoRA for context extension works well under the premise of trainable embedding and normalization. LongLoRA demonstrates strong empirical results on various tasks on LLaMA2 models from 7B/13B to 70B. LongLoRA adopts LLaMA2 7B from 4k context to 100k, or LLaMA2 70B to 32k on a single  $8\times$  A100 machine. LongLoRA extends models' context while retaining their original architectures, and is compatible with most existing techniques, like FlashAttention-2. In addition, to make LongLoRA practical, we collect a dataset, LongQA, for supervised fine-tuning. It contains more than 3k long context question-answer pairs. All our code, models, dataset, and demo are available at [github.com/dvlab-research/LongLoRA](https://github.com/dvlab-research/LongLoRA).

利用稀疏局部注意，可以有效地进行模型的微调。

shift short attention  
扩展了上下文，节约了计算资源

训练实现时只需要两行代码

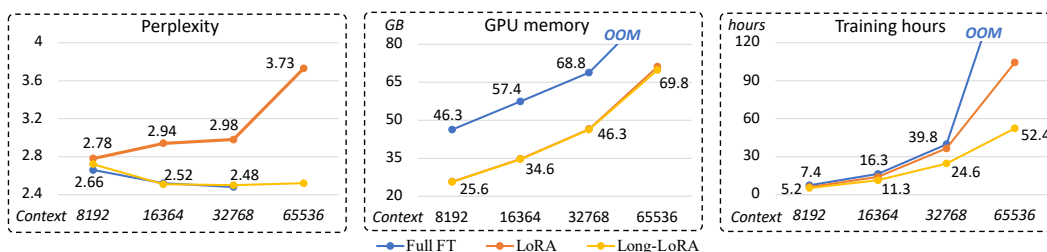


Figure 1: Performance and efficiency comparison between full fine-tuning, plain LoRA, and our LongLoRA. We fine-tune LLaMA2 7B on various context lengths, with FlashAttention-2 (Dao, 2023) and DeepSpeed (Rasley et al., 2020) stage 2. Perplexity is evaluated on the Proof-pile (Azerbaiyev et al., 2022) test set. Plain LoRA baseline spends limited GPU memory cost, but its perplexity gets worse as the context length increases. LongLoRA achieves comparable performance to full fine-tuning while the computational cost is much less.

## 1 INTRODUCTION

Large language models (LLMs) are typically trained with a **pre-defined context size**, such as 2048 tokens for LLaMA (Touvron et al., 2023a) and 4096 tokens for LLaMA2 (Touvron et al., 2023b).

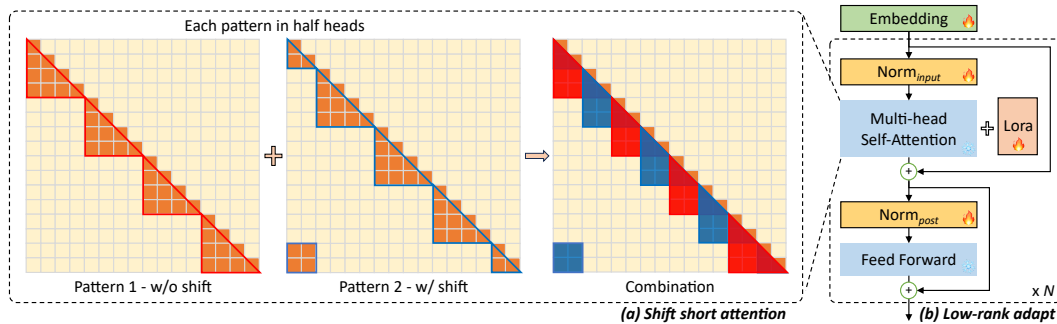


Figure 2: Overview of LongLoRA designs. LongLoRA introduces **shift short attention** during fine-tuning. The trained model can retain its original standard self-attention during inference. In addition to plain LoRA weights, LongLoRA additionally makes embedding and normalization layers trainable, which is essential to long context learning, but takes up only a small proportion of parameters.

训练后的模型在推理过程中可以保持原来的标准自注意

LongLoRA还使嵌入层和归一化层可训练，这对于长上下文学习是必不可少的，但只占用了一小部分参数。

However, the pre-defined size limits LLMs in many applications, like summarizing long documents or answering long questions. To resolve this limitation, some recent works (Chen et al., 2023; Twokowski et al., 2023; Mohtashami & Jaggi, 2023) train or fine-tune LLM to longer context. However, training an LLM from scratch with long sequences poses computational challenges, and fine-tuning an existing pre-trained LLM is also considerably expensive. For instance, Position Interpolation (Chen et al., 2023) spent 32 A100 GPUs to extend LLaMA models from 2k to 8k context, and 128 A100 GPUs for longer context fine-tuning. FOT (Twokowski et al., 2023) used 32 TPUs for standard transformer training and 128 TPUs for LongLLaMA. These computation resources are typically unaffordable for common researchers, which naturally leads us to question: **Can we extend the context window of LLMs efficiently?**

One straightforward approach is to fine-tune a pre-trained LLM via low-rank adaptation (LoRA) (Hu et al., 2022). LoRA modifies the linear projection layers in self-attention blocks by utilizing low-rank matrices, which are generally efficient and reduce the number of trainable parameters. However, our empirical findings indicate that training long context models in this manner is neither sufficiently effective nor efficient. In terms of effectiveness, plain low-rank adaptation results in a high perplexity in long context extension, as in Table 3. Increasing the rank to a higher value, e.g., rank = 256, does not alleviate this issue. In terms of efficiency, regardless of whether LoRA is employed or not, computational cost increases dramatically as the context size expands, primarily due to the standard self-attention mechanism (Vaswani et al., 2017). As shown in Figure 1, even with LoRA, the training hours for the standard LLaMA2 model increase substantially when the context window expands.

In this work, we introduce LongLoRA, an efficient fine-tuning approach that extends the context windows of pre-trained LLMs, e.g., LLaMA2 (Touvron et al., 2023b). LoRA (Hu et al., 2022) uses low-rank weight updates to approximate full fine-tuning. Similarly, we find that short attention is also able to approximate long context during training. We present shift short attention ( $S^2$ -Attn) as an efficient substitute for standard self-attention. As shown in Figure 2, we split context length into several groups and conduct attention in each group individually. In half attention heads, we shift the tokens by half group size, which ensures the information flow between neighbouring groups. For example, we use  $S^2$ -Attn with group size 2048 to approximate the total 8192 context length training. This shares a high-level spirit with Swin Transformer (Liu et al., 2021).

Models fine-tuned via  $S^2$ -Attn retain the original attention architecture during inference. This facilitates most existing optimization and infrastructure. Techniques for common LLMs can also be applied to ours. For example, FlashAttention-2 (Dao et al., 2022; Dao, 2023) is compatible with our method in both training and inference time. The reason behind this is that short attention resembles the attention scheme in the pre-training stage of LLMs. Other efficient attentions, e.g., dilated or sparse attention, have a large gap to the standard style in the pre-training stage, as shown in Table 2.

We empirically show that learnable embedding and normalization layers are the key to unlocking long context LoRA fine-tuning, in Table 3. Embedding and normalization layers take up a small

我们的经验表明，可学习的嵌入层和规范化层是解锁长上下文LoRA微调的关键

能否有效地扩展llm的上下文窗口吗？

随着上下文大小的扩展，计算成本急剧增加，这主要是由于标准的自我注意机制

同样，我们发现在训练过程中，短注意力也能够近似于长情境。我们提出了转移短注意力（ $S^2$ -Attn）作为标准的自我注意的有效替代品

通过 $S^2$ -Attn微调的模型在推理期间保留了原始的注意力架构。

这背后的原因是，短注意力类似于llm训练前阶段的注意方案

如图2所示，我们将上下文长度分成几个组，并在每个组中分别进行注意。在一半的注意力头中，我们移动一半的标记，这确保了相邻组之间的信息流动。

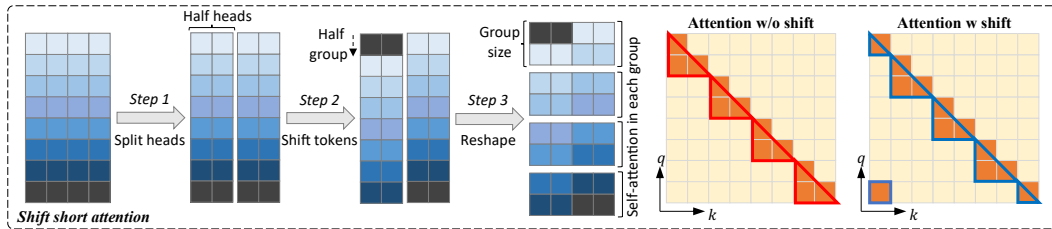


Figure 3: Illustration of shift short attention. **Shift short attention** involves three steps. First, it splits features along the head dimension into two chunks. Second, tokens in one of the chunks are shifted by half of the group size. Third, we split tokens into groups and reshape them into batch dimensions. Attention only computes in each group in ours while standard self-attention computes among all tokens. The information flows between groups via shifting.

proportion of parameters in the entire LLM. For example, embedding has ( $< 2\%$ ) parameters, and normalization has ( $\leq 0.004\%$ ) parameters in LLaMA2 7B. This ratio decreases for even larger LLMs.

In experiments, we show that LongLoRA is effective and efficient. We present experimental results of extending the context window for LLaMA2 7B, 13B, and 70B. Following the experimental settings of Position Interpolation (Chen et al., 2023), we fine-tune models with proper position embeddings. The trained models achieve comparable performance to the full-attention and fully fine-tuned results, while the computational cost is much less as shown in Figure 1. LongLoRA can fine-tune LLaMA2 7B up to 100k context, or a 70B model up to 32k, on a single  $8 \times$  A100 machine.

In addition, we present a dataset, LongQA, for supervised fine-tuning (SFT). LongQA contains more than 3k long questions and the corresponding answers. We design various types of questions for technical paper, science fiction, and other books. SFT is important for improving the chat ability of LLMs. We present some examples of our trained models in the appendix.

## 2 RELATED WORK

**Long-context Transformers.** A large body of research has been developed to increase the context length of transformers. Some of these approaches are retrieval-based (Karpukhin et al., 2020; Izacard et al., 2022; Guu et al., 2020), which augment language models via fetching related documents and including the retrieved results into contexts. Our work is complementary to these works, as our attention mechanism is unmodified during inference. Many works modify multi-head attention to be approximated ones (Wang et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020; Kitaev et al., 2020; Bulatov et al., 2022; Ding et al., 2023; Qiu et al., 2020). They alleviate the quadratic complexity of the self-attention computation. For example, Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020) use sparse attention to handle long sequences. Other works (Wu et al., 2022; Bulatov et al., 2022) utilize memory mechanisms as a compression on past inputs, to look up relevant tokens. One limitation of these works is that these compression has a large gap to full attention, making it infeasible to fine-tune pre-trained LLMs. Although our work also involves an approximation of attention mechanism, it has a similar shape and a small gap to standard attention. This enables ours to fine-tune pre-trained LLMs and maintain full attention during inference.

为了增加Transformer的上下文长度，已经进行了大量的研究

现在已经有预训练好的LLM，提出token方法要和之前的注意力类似，否则无法直接用在训练好的LLM

**Long-context LLMs.** LLMs are typically pre-trained with a pre-defined context length, such as 2048 for LLaMA (Touvron et al., 2023a) and 4096 for LLaMA2 (Touvron et al., 2023b). Training LLMs with long context from scratch is prohibitively expensive for most researchers. Recently, a number of works have tried to extend the context length of LLMs via fine-tuning. Position Interpolation (Chen et al., 2023) introduces a modification upon rotary position encoding (Su et al., 2021) and extends the context length of LLaMA to 32768. Focused Transformer (Tworkowski et al., 2023) utilizes contrastive learning to train LongLLaMA. Both of them rely on full fine-tuning, which is computationally expensive (128 A100 GPUs / 128 TPUv3 for training). Landmark attention (Mohtashami & Jaggi, 2023) is an efficient approach, but somewhat lossy. It compresses long context inputs into retrieved tokens. Our method saves substantial fine-tuning costs, while

我们的方法节省了大量的微调成本，并且保留了原始attention的质量

Table 1: Ablations on different training patterns and target context length. ‘Short’ means 1/4 of the target context length. ‘Long’ equals to the target context length. Models are fully fine-tuned upon an LLaMA2 (Touvron et al., 2023b) model in 7B size, on RedPajama (Computer, 2023) dataset. Results are tested in perplexity on PG19 (Rae et al., 2020) validation split.

Setting	Position Embedding	Training		Target Context Length		
		Attention	Shift	8192	16384	32768
Train-free	PI (Chen et al., 2023)	<u>w/o fine-tuning</u>		15.82	94.57	236.99
	NTK-Aware (ntk, 2023)			10.89	88.44	932.85
Full Attn	PI (Chen et al., 2023)	Long	-	8.02	8.05	8.04
Short Attn		Short	✗	8.29	8.83	9.47
S <sup>2</sup> -Attn		Short	✓	8.04	8.03	8.08

preserving the quality of the original attention. **Ours maintain full access to the entire input via unmodified attention during inference.**

我们在推理过程中通过未经修改的注意力来保持对整个输入的完整访问。

Some literature focuses on the position embedding modification of LLMs for long context extension, including Position Interpolation (Chen et al., 2023), NTK-aware (ntk, 2023), Yarn (Peng et al., 2023), positional Skipping (Zhu et al., 2023), and the out-of-distribution related method (Han et al., 2023). **Our method focuses on efficient fine-tuning and retaining the original architecture during inference, which is orthogonal to these position embedding methods.** Our models apply the Position Interpolation (Chen et al., 2023) in experiments.

**Efficient Fine-tuning.** This work is based on LoRA (Hu et al., 2022), a classical efficient fine-tuning approach. In addition to LoRA (Hu et al., 2022), there are many other parameter-efficient fine-tuning methods, including prompt tuning (Lester et al., 2021), prefix tuning (Li & Liang, 2021), hidden state tuning (Liu et al., 2022), bias tuning (Zaken et al., 2022), and masked weight learning (Sung et al., 2021). Input-tuning (An et al., 2022) introduces an adapter to tune input embedding. Although the input embedding layers are also trainable in ours, this is not enough for long context extension. We make a comprehensive analysis on layer types in experiments, in Table 3.

### 3 LONGLoRA

#### 3.1 BACKGROUND

**Transformer.** LLMs are typically built with transformers. Taking LLaMA2 (Touvron et al., 2023b) for example, as shown in Figure 2, an LLM model consists of an embedding input layer and a number of decoder layers. Each decoder layer comprises a self-attention module. It maps input features into a set of queries, keys, and values  $\{q, k, v\}$ , via linear projection layers with weight matrices  $\{W_q, W_k, W_v\}$ . Given  $\{q, k, v\}$ , it computes the outputs  $o$  as

$$o = \text{softmax}(qk^T)v \quad (1)$$

The outputs are then projected by a linear layer with a weight matrix  $W_o$ . And MLP layers are followed. Before and after self-attention modules, layer normalization (Ba et al., 2016) is applied. A final normalization is conducted after all decoder layers.

For long sequences, self-attention struggles with computation cost, which is quadratic to the sequence length. This dramatically slows down the training procedure and increases GPU memory costs.

**Low-rank Adaptation.** LoRA (Hu et al., 2022) hypothesizes that the weight updates in pre-trained models have a low intrinsic rank during adaptation. For a pre-trained weight matrix  $W \in \mathbb{R}^{d \times k}$ , it is updated with a low-rank decomposition  $W + \Delta W = W + BA$ , where  $B \in \mathbb{R}^{d \times r}$  and  $A \in \mathbb{R}^{r \times k}$ . The rank  $r \ll \min(d, k)$ . During training,  $W$  is frozen with no gradient updates, while  $A$  and  $B$  are trainable. This is the reason why LoRA training is much more efficient than full fine-tuning.

In the Transformer structure, LoRA only adapts the attention weights ( $W_q, W_k, W_v, W_o$ ) and freezes all other layers, including MLP and normalization layers. This manner is simple and parameter-efficient. However, **we empirically show that only low-rank adaptation in attention weights does not work for long context extension.**

我们的经验表明，只有注意权重的低秩adaptati on不适用于长上下文扩展。

Table 2: Ablation on attention patterns during fine-tuning. We fine-tune an LLaMA2 7B model to 32768 context length via various attention patterns, with the improved LoRA setting. We include four typical efficient attention designs, *e.g.*, shift, dilate, stride sparse for comparison. ‘*cro. heads / layers*’ means to swap different attention settings across attention heads or sequential layers. Taking  $S^2$ -Attn as an example, ‘*cro. layers*’ is to swap between w/ and w/o shift in sequential self-attention layers. ‘*only P1/P2*’ means all attention heads use pattern 1 (all no shift) or Pattern 2 (all shift) in Figure 2. Dilated attention (Ding et al., 2023) varies dilated rate from 1 to 4 in attention heads. Stride sparse attention is introduced in (Child et al., 2019), which is also swapped between local and stride attention in attention heads.

Test w/ Full-Attn	$S^2$ -Attn				Dilate <i>cro. heads</i>	Stride sparse <i>cro. heads</i>
	<i>cro. heads</i>	<i>cro. layers</i>	<i>only P1.</i>	<i>only P2.</i>		
✗	8.64	8.63	9.17	9.64	8.75	31.46
✓	<b>8.12</b>	9.70	<u>8.39</u>	9.81	11.78	>1000

#### Algorithm 1: Pseudocode of Shift Short Attention in PyTorch-like style.

```
# B: batch size; S: sequence length or number of tokens; G: group size;
# H: number of attention heads; D: dimension of each attention head

# qkv in shape (B, N, 3, H, D), projected queries, keys, and values
# key line 1: split qkv on H into 2 chunks, and shift G/2 on N
qkv = cat((qkv.chunk(2, 3)[0], qkv.chunk(2, 3)[1].roll(-G/2, 1)), 3).view(B*N/G, G, 3, H, D)

# standard self-attention function      half heads      shift操作, 将分区分
out = self_attn(qkv)                    向后移动半个分组大小

# out in shape (B, N, H, D)
# key line 2: split out on H into 2 chunks, and then roll back G/2 on N
out = cat((out.chunk(2, 2)[0], out.chunk(2, 2)[1].roll(G/2, 1)), 2) 还原shift操作

cat: concatenation; chunk: split into the specified number of chunks; roll: roll the tensor along the given dimension.
```

### 3.2 SHIFT SHORT ATTENTION

Standard self-attention pattern cost  $O(n^2)$  computations, making LLMs on long sequences high memory cost and slow. To avoid this issue during training, we propose **shift short attention ( $S^2$ -Attn)**, as shown in Figure 2. In the following, we explain our designs step by step.

**Pilot Study.** In Table 1, we first validate the importance of fine-tuning. Without fine-tuning, models perform worse as the context length grows up, even with proper position embeddings (Chen et al., 2023; ntk, 2023) equipped. **We build up a standard baseline that is trained and tested with full attention and full fine-tuning, which presents consistently good quality in various context lengths.**

The first trial is to train with **short attention, only pattern 1** in Figure 2. As we know for a long context, the high cost mainly comes from self-attention modules. **Thus, in this trial, since the input is long, we split into several groups in self-attention.** For example, the model takes 8192 tokens as input in both the training and testing stages, but self-attention is conducted in each group with a 2048 size. **The group number is 4**, as ablated in Table 11. This pattern is efficient but still does not work in a very long context, as shown in Table 1. **The perplexity becomes larger as the context length increases. The reason behind this is that there is no information exchange between different groups.**

To introduce communication between groups, we include a **shifted pattern**, as shown in Figure 2. We **shift the group partition by half group size in half attention heads.** Taking the overall 8192 context length for example, in **pattern 1**, the first group conducts self-attention from 1<sup>st</sup> to 2048<sup>th</sup> tokens. **In Pattern 2**, the group partition is shifted by 1024. The first attention group begins from 1025<sup>th</sup> and ends at 3072<sup>th</sup> tokens, while the first and the last 1024 tokens belong to the same group. **We use patterns 1 and 2 in each half self-attention heads respectively.** This manner does not increase additional computation cost but enables the information flow between different groups. We show that it gets close to the standard attention baseline in Table 1.

**Consistency to Full Attention.** Existing efficient attention designs can also improve the efficiency of long-context LLMs. In Table 2, we compare the proposed  $S^2$ -Attn with several typical efficient

建模8192tokens, 分组计算注意力, 每一组都进行自注意力, 大小为2048人。

随着上下文长度的增加, 困惑度也会变得更大。这背后的原因是, 在不同的组之间没有信息交换

在一半的自注意头中将分区分区向后移动半个分组大小

在模式2中, 分区分被移动了1024。第一个注意组从第1025代标记开始, 到第3072代标记结束, 而第一个和最后1024个标记属于同一组。

我们在每个半自注意集中分别使用模式1和模式2

这种方式不会增加额外的计算成本, 但可以实现不同组之间的信息流。



Table 3: Ablation on fine-tuning and ablations in various settings. Models are trained based on LLaMA2 (Touvron et al., 2023b) model in 7B size, with the proposed Shift Short Attention. The target context length is 32768. ‘+ Normal / Embed’ means including normalization or embedding layers as trainable. We use RedPajama (Computer, 2023) dataset for training. Results are tested in perplexity on PG19 (Rae et al., 2020) validation set. For long context adaptation, standard LoRA (Hu et al., 2022) has a large gap to the full fine-tuning result. Without trainable normalization or embeddings, larger ranks in LoRA have no effects.

Method	Full FT	LoRA (rank)						LoRA (rank = 8)	
		8	16	32	64	128	256	+ Norm	+ Norm + Embed
PPL	8.08	11.44	11.82	11.92	11.96	11.97	11.98	10.49	8.12

attention, including short attention, dilated attention (Ding et al., 2023), and stride sparse attention (Child et al., 2019). We show that  $S^2$ -Attn not only enables efficient fine-tuning but also supports full attention testing.

Some efficient attention designs are infeasible for long-context fine-tuning. The transformers (Qiu et al., 2020; Child et al., 2019), developed for training from scratch, have gaps to the standard full attention, which is used in pre-training. Thus, these attentions are not suitable for long context fine-tuning.  $S^2$ -Attn supports full attention testing, although the model is fine-tuned with shift short attention, as shown in Table 2. Although other attentions, like dilated attention (Ding et al., 2023) and stride sparse attention (Child et al., 2019), can also be used in long context fine-tuning, models must be tested with the attention used during fine-tuning. Shifting prevents models from being over-fitted to specific attention patterns. In  $S^2$ -Attn, pattern 1 or 2 only does not work as in Table 2.

**Easy Implementation.** Shift short attention is easy to implement. It involves only two steps: (1) shifting tokens in half attention heads, and (2) transposing features from token dimension to batch dimension. Two lines of code are enough. We provide a PyTorch-style code in Algorithm 1. In the following, we make a pilot study and clarify the reasons for our design step by step.

### 3.3 IMPROVED LORA FOR LONG CONTEXT

LoRA (Hu et al., 2022) is an efficient and popular manner for adapting LLMs to other datasets. It saves much trainable parameters and memory cost, compared to full fine-tuning. However, adapting LLMs from short context length to long is not easy. We empirically observe an obvious gap between LoRA and full fine-tuning. As shown in Table 3, the gap between LoRA and full fine-tuning grows as the target context length becomes larger. And LoRA with larger ranks cannot reduce the gap.

To bridge this gap, we open embedding and normalization layers for training. As shown in Table 3, they occupy limited parameters but make effects for long context adaptation. Especially for normalization layers, the parameters are only 0.004% in the whole LLaMA2 7B. We denote this improved version of LoRA as LoRA<sup>+</sup> in experiments.

## 4 EXPERIMENT

### 4.1 EXPERIMENTAL SETTINGS

**Models.** We extend the pre-trained 7B, 13B, and 70B LLaMA2 (Touvron et al., 2023b) models. The maximum extended context window sizes are up to 100k for 7B models, 65536 for 13B models, and 32768 for 70B models. The position indices for these models are re-scaled with Position Interpolation (Chen et al., 2023).

**Training Procedure.** We follow most training hyper-parameters in Position Interpolation (Chen et al., 2023), except that our batch size is smaller as we use a single 8× A100 GPUs machine in some cases. All models are fine-tuned via the next token prediction objective. We use AdamW (Loshchilov & Hutter, 2019) with  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ . The learning rate is set to  $2 \times 10^{-5}$  for 7B and 13B models, and  $10^{-5}$  for 70B models. We also use a linear learning rate warmup. The weight decay is

Table 4: Evaluation perplexity on proof-pile dataset (Rae et al., 2020) test split.  $S^2$ -Attn: Shift Short Attention. LoRA<sup>+</sup>: improved LoRA with embedding and normalization layers trainable. We fine-tune LLaMA2 (Touvron et al., 2023b) in 7B and 13B model sizes. We use RedPajama (Computer, 2023) dataset for training. Models fine-tuned with LongLoRA show progressively lower perplexity with longer evaluation context length. We use the same training setting as the model evaluated on PG19 (Rae et al., 2020), which is introduced in Table 9 in the appendix.

Size	Training Context Length	LongLoRA		Evaluation Context Length				
		$S^2$ -Attn	LoRA <sup>+</sup>	2048	4096	8192	16384	32768
7B	8192	✓		3.14	2.85	2.66	-	-
		✓	✓	3.15	2.86	2.68	-	-
		✓	✓	3.20	2.91	2.72	-	-
	16384	✓		3.17	2.87	2.68	2.55	-
		✓	✓	3.17	2.87	2.66	2.51	-
	32768	✓		3.20	2.90	2.69	2.54	2.49
		✓	✓	3.35	3.01	2.78	2.61	2.50
13B	8192	✓		2.96	2.69	2.53	-	-
		✓	✓	3.01	2.74	2.57	-	-
		✓	✓	3.04	2.77	2.60	-	-
	16384	✓		2.99	2.72	2.53	2.40	-
		✓	✓	3.03	2.74	2.55	2.41	-
	32768	✓		3.04	2.75	2.56	2.42	2.33
		✓	✓	3.05	2.76	2.57	2.42	2.32

Table 5: Maximum context length that we can fine-tune for various model sizes on a single 8 × A100 machine. We fine-tune LLaMA2 (Touvron et al., 2023b) in 7B, 13B, and 70B model sizes, using RedPajama (Computer, 2023) dataset, and evaluate the perplexity on Proof-pile dataset (Rae et al., 2020) test split. We use FlashAttention-2 (Dao, 2023) and DeepSpeed (Rasley et al., 2020) in Stage 3 during fine-tuning. With LongLoRA, the maximum context length for 7B, 13B, and 70B models are 100k, 64k, and 32k respectively. Evaluation on PG19 (Rae et al., 2020) is Table 10 in the appendix.

Size	Training Context Length	Evaluation Context Length						
		2048	4096	8192	16384	32768	65536	100,000
7B	100,000	3.36	3.01	2.78	2.60	2.58	2.57	2.52
13B	65536	3.20	2.88	2.66	2.50	2.39	2.38	-
70B	32768	2.84	2.57	2.39	2.26	2.17	-	-

zero. We set the per-device batch size as 1 and gradient accumulation steps as 8, which means that the global batch size equals 64, using 8 GPUs. We train our models for 1000 steps.

**Datasets.** We use the Redpajama (Computer, 2023) dataset for training. We evaluate the long-sequence language modeling performance of our fine-tuned models on the book corpus dataset PG19 (Rae et al., 2020) and the cleaned Arxiv Math proof-pile dataset (Azerbayev et al., 2022). We use the test split of PG19 (Rae et al., 2020), consisting of 100 documents. For the proof-pile dataset, we also use the test split of it for evaluation. We follow Position Interpolation (Chen et al., 2023) for Proof-pile data processing. We evaluate perplexity by using a sliding window approach with  $S = 256$ , following (Press et al., 2022).

In addition, we build a long context QA dataset, LongQA, for supervised fine-tuning. Although the models fine-tuned with Redpajama (Computer, 2023) present good perplexities, their chat ability is limited. We collect more than 3k question-answer pairs, relating to the materials like technical paper, science fiction, and other books. The questions we designed include summarization, relationships, characters, and other details related to the material. For more details, please refer to the appendix.

Table 6: Evaluation on topic retrieval using LongChat (Li et al., 2023). We compare our model to other open LLMs with long contexts. This task involves retrieving target topics from a very long conversation with lengths around 3k, 6k, 10k, 13k, and 16k. As some questions in the evaluation set are longer than 16k, our model is fine-tuned via 18k context length upon LLaMA2 13B. It achieves comparable performance to LongChat-13B (Li et al., 2023), the state-of-the-art model in this task, while ours is from an efficient fine-tuning manner.

Evaluation Context	3k	6k	10k	13k	16k
ChatGLM2-6B (Du et al., 2022)	0.88	0.46	0.02	0.02	0.02
MPT-30B-chat (Team, 2023a)	0.96	<b>1.0</b>	0.76	-	-
MPT-7B-storywriter (Team, 2023b)	0.46	0.46	0.28	0.34	0.36
LongChat-13B (Li et al., 2023)	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>0.98</b>	0.9
Ours	<b>1.0</b>	0.98	0.98	<b>0.98</b>	<b>0.94</b>

Table 7: Efficiency profile in terms of FLOPs on various context lengths. We break down the LLaMA2 7B model into FFN (feed-forward layers), Proj (projection layers for queries, keys, values, and attention outputs), Attn (self-attention kernel), and Others (e.g., embedding, normalization, LLM head). The ratio of attention in the overall model increases as the context length increases.  $S^2$ -Attn reduces the FLOPs by a large margin, especially when the context length is large.

Context Length	$S^2$ -Attn	FLOPs (T)				
		Attn	Proj	FFN	Others	Total
8192	$\times$	35.2	35.2	70.9	2.2	143.5
	$\checkmark$	8.8				117.1
16384	$\times$	140.7	70.4	141.8	4.3	357.2
	$\checkmark$	35.2				251.7
32768	$\times$	562.9	140.7	283.7	8.7	996.0
	$\checkmark$	140.7				573.8
65536	$\times$	2251.8	281.5	567.4	17.3	3118.0
	$\checkmark$	562.9				1429.1

## 4.2 MAIN RESULTS

**Long-sequence Language Modeling.** In Table 4 and Table 9, we report the perplexity for our models and baseline on Proof-pile (Azerbayev et al., 2022) and PG19 datasets. **Under certain training context lengths, our models achieve better perplexity with longer context sizes.** This indicates the effectiveness of our efficient fine-tuning method. In Table 4, for the same training and evaluation context length cases, the perplexity decreases as the context size increases. By increasing the context window size from 8192 to 32768, for LLaMA2 7B model, we observe that the perplexity gets better from 2.72 to 2.50 by -0.22. For LLaMA2 13B model, we observe that the perplexity reduces from 2.60 to 2.32 by -0.28.

In Table 5, we further examine the maximum context length that we can fine-tune on a single  $8 \times A100$  machine. We extend LLaMA2 7B, 13B, and 70B to 100k, 65536, 32768 context length respectively. LongLoRA achieves promising results on these extremely large settings. In addition, **we find some perplexity degradation on small context sizes for the extended models. This is a known limitation of Position Interpolation** (Chen et al., 2023).

**Retrieval-based Evaluation.** In addition to long-sequence language modeling, we also conduct experiments on retrieval in long contexts. In Table 6, we compare our model with other open LLMs on the topic retrieval task introduced in LongChat (Li et al., 2023). This task is to retrieve the target topic from a very long conversation, with lengths varying from 3k, 6k, 10k, 13k, to 16k. As some questions in LongChat (Li et al., 2023) are longer than 16k, we fine-tuned LLaMA2 13B with a context length of 18k. The training cost is similar to that for 16k. Our model achieves comparable performance to LongChat-13B (Li et al., 2023), the state-of-the-art model in this task. Unlike LongChat-13B (Li et al., 2023), which is fully fine-tuned on self-collected long context conversation text, our model is efficiently adapted on the open RedPajama (Computer, 2023) via next-token generation. Our model even slightly outperforms LongChat-13B in the 16k evaluation.



Table 8: Ablation on fine-tuning steps in both full fine-tuning and low-rank training (with trainable normalization and embedding). We fine-tune LLaMA2 (Touvron et al., 2023b) 7B with the proposed Shift Short Attention. The target context length is 8192. We use RedPajama (Computer, 2023) for training and PG19 (Rae et al., 2020) validation set for perplexity testing. Full fine-tuning has a faster convergence than the low-rank at the beginning, while the final gap is not large.

Training	Number of fine-tuning steps										
	0	100	200	300	400	500	600	700	800	900	1000
Full FT	15.82	8.17	8.10	8.07	8.06	8.03	7.99	7.99	7.96	7.95	7.94
LoRA <sup>+</sup>	15.82	8.63	8.16	8.15	8.14	8.12	8.11	8.10	8.08	8.04	8.02

### 4.3 ABLATION STUDY

**Efficiency Profile.** In Table 7, we breakdown LLaMA2 7B (Touvron et al., 2023b) into various types of layers, including FFN - feed-forward layers, Proj - projection for queries, values, keys, and attention outputs, Attn - self-attention computation, Others - other layers like embedding, normalization, LLM head. We analyze FLOPs. For full attention, the proportion of Attn sharply increases as the context length increases. For example, Attn has 24.5% of the total FLOPs at the 8192 context length while it increases to 72.2% at the 65536 context length. It decreases to 39.4% when  $S^2$ -Attn is used.

**Ablation on Fine-tuning Steps.** We report the relationship between perplexity and fine-tuning steps for an LLaMA2 7B model extending to the 8192 context length on the PG19 validation set, in Table 8. We see that without fine-tuning, at step 0, the model has a limited long context capability, *e.g.*, 15.82 perplexity. We show that the perplexity drops quickly. Full fine-tuning converges faster than low-rank training. They come closer after 200 steps, without a large gap at the end.

**Attention Patterns.** In Table 2, we show the effects of different attention patterns during fine-tuning. We fine-tune an LLaMA2 7B (Touvron et al., 2023b) model to 32768 context length on Redpajama (Computer, 2023) datasets and evaluate the perplexity on PG19 (Rae et al., 2020) validation set. We first examine the manner of swapping among various settings. For the shift operation we used in LongLoRA, there are three choices: disabling it, shifting between sequential layers, and shifting among attention heads. We show that shifting between layers is acceptable but not the best. In addition, setting all attention heads as pattern 1 or pattern 2 does not work.

We then test other types of efficient attention designs, including dilated attention (Ding et al., 2023), and stride sparse attention (Child et al., 2019). For dilated attention (Ding et al., 2023), we vary the dilate rate from 1 to 4 evenly among attention heads. Stride sparse attention (Child et al., 2019) contains both local and stride patterns. These attention patterns are invented in training-from-scratch transformers. This experiment is to examine their capability of fine-tuning on pre-trained LLMs (Touvron et al., 2023b), toward long context adaptation. Dilated attention performs well in full fine-tuning but is not well with low-rank adaptation. Fine-tuning with stride sparse attention is harmful. They have a large gap to full attention, which is applied in the pre-training stage.

## 5 CONCLUSION

In this work, we propose LongLoRA that can efficiently extend the context length of LLMs to be significantly larger. LongLoRA has less GPU memory cost and training time than standard full fine-tuning, with minimal accuracy compromise. At the architecture level, we propose shift short attention to approximate the standard self-attention pattern during training. Shift short attention is easy to implement, requiring only two lines of code. Moreover, models trained via shift short attention retain the original standard attention architecture during inference, making most pre-existing infrastructure and optimization reusable. At the training level, we bridge the gap between LoRA and full fine-tuning with trainable normalization and embedding. Our method can extend LLaMA2 7B to 100k context length and 70B model to 32k context length, on a single  $8 \times$  A100 machine. We believe that LongLoRA is a general method that could be compatible with more types of LLMs and position encodings, which we plan to investigate in the future.

**Acknowledgement** We would like to thank Xiuyu Li and Bohao Peng for the helpful discussions.

---

## REFERENCES

- Ntk-aware scaled rope, 2023. URL [https://www.reddit.com/r/LocalLLaMA/comments/141z7j5/ntkaware\\_scaled\\_rope\\_allows\\_llama\\_models\\_to\\_have/](https://www.reddit.com/r/LocalLLaMA/comments/141z7j5/ntkaware_scaled_rope_allows_llama_models_to_have/).
- Shengnan An, Yifei Li, Zeqi Lin, Qian Liu, Bei Chen, Qiang Fu, Weizhu Chen, Nanning Zheng, and Jian-Guang Lou. Input-tuning: Adapting unfamiliar inputs to frozen pretrained models. CoRR, abs/2203.03131, 2022.
- Zhangir Azerbayev, Edward Ayers, and Bartosz Piotrowski. Proof-pile, 2022. URL <https://github.com/zhangir-azerbayev/proof-pile>.
- Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. CoRR, abs/1607.06450, 2016.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. CoRR, abs/2004.05150, 2020.
- Aydar Bulatov, Yuri Kuratov, and Mikhail S. Burtsev. Recurrent memory transformer. In NeurIPS, 2022.
- Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. CoRR, abs/2306.15595, 2023.
- Yukang Chen, Gaofeng Meng, Qian Zhang, Shiming Xiang, Chang Huang, Lisen Mu, and Xinggang Wang. RENAS: reinforced evolutionary neural architecture search. In CVPR, pp. 4787–4796, 2019.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. CoRR, abs/1904.10509, 2019.
- Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023. URL <https://github.com/togethercomputer/RedPajama-Data>.
- Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. CoRR, abs/2307.08691, 2023.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In NeurIPS, 2022.
- Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1, 000, 000, 000 tokens. CoRR, abs/2307.02486, 2023.
- Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 320–335, 2022.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: retrieval-augmented language model pre-training. CoRR, abs/2002.08909, 2020.
- Chi Han, Qifan Wang, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Simple on-the-fly length generalization for large language models. CoRR, abs/2308.16137, 2023.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In ICLR, 2022.
- Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. CoRR, abs/2208.03299, 2022.

- 
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, pp. 6769–6781, 2020.
- Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.
- Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *EMNLP*, pp. 3045–3059, 2021.
- Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. How long can open-source llms truly promise on context length?, June 2023. URL <https://lmsys.org/blog/2023-06-29-longchat>.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *ACL*, pp. 4582–4597, 2021.
- Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *NeurIPS*, 2022.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 9992–10002, 2021.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- Amirkeivan Mohtashami and Martin Jaggi. Landmark attention: Random-access infinite context length for transformers. *CoRR*, abs/2305.16300, 2023.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *CoRR*, abs/2309.00071, 2023.
- Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *ICLR*, 2022.
- Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for RGBD semantic segmentation. In *ICCV*, pp. 5209–5218, 2017.
- Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip H. S. Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2):969–984, 2022.
- Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. In *EMNLP*, volume EMNLP 2020 of *Findings of ACL*, pp. 2555–2565, 2020.
- Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. Compressive transformers for long-range sequence modelling. In *ICLR*, 2020.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, pp. 3505–3506. ACM, 2020.

- 
- Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021.
- Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks. In *NeurIPS*, pp. 24193–24205, 2021.
- MosaicML NLP Team. Introducing mpt-30b: Raising the bar for open-source foundation models, 2023a. URL [www.mosaicml.com/blog/mpt-30b](http://www.mosaicml.com/blog/mpt-30b). Accessed: 2023-06-22.
- MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023b. URL [www.mosaicml.com/blog/mpt-7b](http://www.mosaicml.com/blog/mpt-7b).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b.
- Szymon Tworkowski, Konrad Staniszewski, Mikolaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Milos. Focused transformer: Contrastive training for context scaling. *CoRR*, abs/2307.03170, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.
- Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768, 2020.
- Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *ICLR*, 2022.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020.
- Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *ACL*, pp. 1–9, 2022.
- Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Pose: Efficient context window extension of llms via positional skip-wise training, 2023.

Table 9: Evaluation perplexity on PG19 dataset (Rae et al., 2020) test split. S<sup>2</sup>-Attn: Shift Short Attention. LoRA<sup>+</sup>: improved LoRA with embedding and normalization layers trainable. We fine-tune LLaMA2 (Touvron et al., 2023b) in 7B and 13B sizes along 8192, 16384, and 32768 context length.

Size	Training Context Length	LongLoRA		Evaluation Context Length				
		S <sup>2</sup> -Attn	LoRA <sup>+</sup>	2048	4096	8192	16384	32768
7B	8192	✓		7.55	7.21	6.98	-	-
		✓	✓	7.53	7.20	7.01	-	-
				7.70	7.35	7.14	-	-
	16384	✓		7.56	7.21	6.97	6.80	-
		✓	✓	7.65	7.28	7.02	6.86	-
	32768	✓		7.76	7.36	7.09	7.04	7.03
		✓	✓	8.29	7.83	7.54	7.35	7.22
13B	8192	✓		6.95	6.60	6.43	-	-
		✓	✓	6.94	6.63	6.45	-	-
				7.03	6.73	6.58	-	-
	16384	✓		6.90	6.58	6.37	6.22	-
		✓	✓	7.05	6.70	6.47	6.31	-
	32768	✓		7.14	6.76	6.52	6.39	6.36
		✓	✓	7.14	6.78	6.55	6.38	6.29

Table 10: Evaluation perplexity on PG19 dataset (Rae et al., 2020), with the maximum context length that we can fine-tune on a single 8× A100 machine. We fine-tune LLaMA2 (Touvron et al., 2023b) models, using RedPajama (Computer, 2023) dataset. We use the same setting to that in Table 5.

Size	Training Context Length	Evaluation Context Length						
		2048	4096	8192	16384	32768	65536	100,000
7B	100,000	8.38	7.90	7.57	7.33	7.16	7.06	7.04
13B	65536	7.63	7.21	6.94	6.75	6.62	6.57	-
70B	32768	5.93	5.63	5.44	5.32	5.27	-	-

## APPENDIX

**Environments.** All our experiments are conducted on an 8× A100 machine. We train all models using PyTorch (Paszke et al., 2019) with the DeepSpeed (Rasley et al., 2020) and FlashAttention-2 (Dao, 2023). Gradient checkpoint is used by default, which is a common technique in the Pefit codebase Mangrulkar et al. (2022). Note that sometimes, like fine-tuning 7B models to 8192 context size, 3090 Ti GPUs are acceptable.

**Evaluation Perplexity on PG19 Test Split.** In Table 9 and Table 10, we present the evaluation results on the PG19 test split. We use the same training settings as the models in Table 4 and Table 5. Similarly, for a model trained on a certain context length, as the evaluation context length increases, our models achieve better perplexity. Note that the perplexity in Table 9 and Table 10 is higher than that in the Proof-pile dataset, as PG19 (Rae et al., 2020) has very different writing styles.

**Ablation on Group Sizes.** In Table 11, we provide an ablation study on the group size of the shift short attention. We experimented on fine-tuning LLaMA2 7B to 8192 context length via LongLoRA. The group size varies from {1/2, 1/4, 1/6, 1/8} of the target context length. For example, the group size is 1024 for 1/8 of the context length 8192. We find that the 1/2 and 1/4 settings have minor gap to full attention fine-tuning. Group sizes less than 1/4 would be not good enough. We set the group size as 1/4 of the context length in experiments by default.

**LongQA for Supervised Fine-tuning.** To improve the chat ability of our models, we build up a long context QA dataset, LongQA, for supervised fine-tuning (SFT). It contains more than 3k question-answer pairs. We build the prompt format as the following line:

---

Table 11: Ablation on group size. We conduct experiments upon an LLaMA2 7B model and fine-tune it to 8192 context length via LongLoRA on PG19. We vary the group size of Shift Short Attention from  $\{1/2, 1/4, 1/6, 1/8\}$  of the target context length. ‘Full’ means the standard full attention.

Group	Full	1/2	1/4	1/6	1/8
PPL	8.02	8.04	8.04	8.10	8.16

Below is {material\_type}. Memorize the content and answer my question after the paper. {material\_content}  $n$  Now the material ends. {question}

{material\_type} can be "book", "paper", and others. {material\_content} is the long-context content in the document. {question} is the questions we design. We list some example questions as the following:

- Please tell me what high-level idea the author wants to indicate in this book.
- Please describe the relationship among the roles in the book.
- What are the main contributions and novelties of this paper?
- What are some limitations of the proposed method?
- Why doesn't Professor Snape seem to like Harry?
- ...

These questions can be some commonly used ones, like summarization and limitation. Or they can be specific to the material, like the question that is related to some roles in the book. The context length of each question is no longer than 32k tokens.

For SFT on LongQA, we use the models that have already been fine-tuned on Redpajama (Computer, 2023) for context extension in this step. We use the same learning rate, weight decay, and batch sizes as the context extension step. We train the models for 3 epochs. In the following, we provide some example questions and the answers from our model, in Figure 4, Figure 5, and Figure 6.



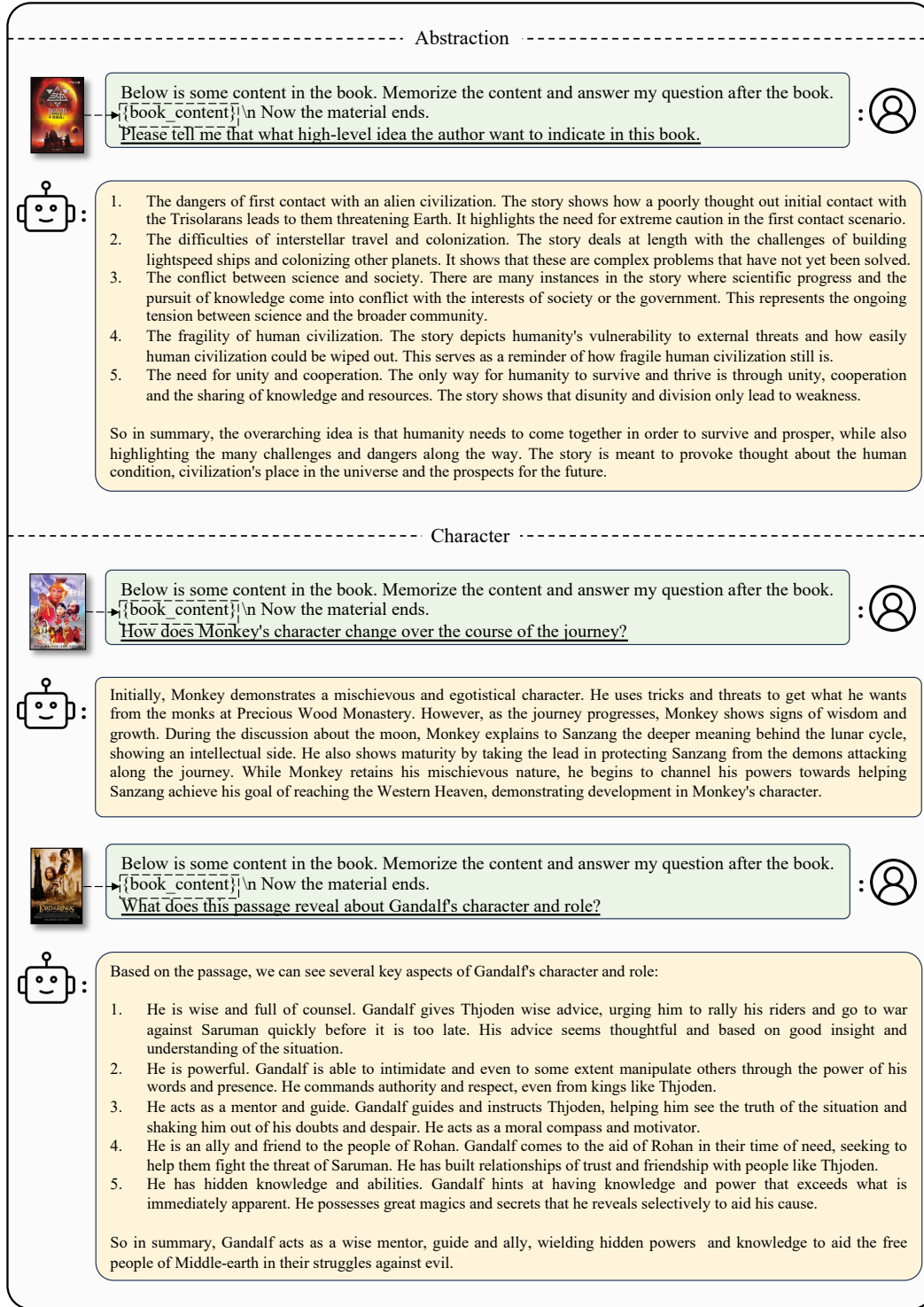



Figure 4: Examples on book sections<sup>123</sup> and questions related to abstraction and character. For all these examples, we select the related section in each book with less than 32k tokens.

<sup>1</sup>[https://en.wikipedia.org/wiki/Death%27s\\_End](https://en.wikipedia.org/wiki/Death%27s_End)

<sup>2</sup>[https://en.wikipedia.org/wiki/Journey\\_to\\_the\\_West](https://en.wikipedia.org/wiki/Journey_to_the_West)

<sup>3</sup>[https://en.wikipedia.org/wiki/The\\_Two\\_Towers](https://en.wikipedia.org/wiki/The_Two_Towers)


Relationship



Below is some content in the book. Memorize the content and answer my question after the book.

`{{book_content}}` Now the material ends.

Please describe the relationship among the roles in the book.




The book depicts a complex set of relationships among the various characters. Some of the key relationships are:

1. Harry Potter and Ron Weasley: They are close friends. Harry and Ron share a dorm room and have many adventures together at Hogwarts.
2. Harry and Hermione Granger: They are also good friends. Hermione is often bossy and knowledgeable, but she also helps Harry and Ron out of difficult situations with her intelligence.
3. Harry and Draco Malfoy: They are enemies. Draco is part of the Slytherin house and is always trying to cause trouble for Harry and his friends.
4. Harry and Professor Dumbledore: Dumbledore is the headmaster of Hogwarts and a mentor figure to Harry. He trusts and believes in Harry even when others doubt him.
5. Harry and Professor Snape: They have an antagonistic relationship. Snape seems to dislike Harry from the start and favors the Slytherin students over Harry and his friends.
6. Harry and Professor Lockhart: Lockhart enjoys the fame and attention that comes from being connected to Harry. However, Harry finds Lockhart vain and insincere.

Overall, the book depicts complex and nuanced relationships that develop and change over time, between students, teachers, and other characters at Hogwarts. The relationships help shape Harry's story and journey at the school of witchcraft and wizardry.

Details



Below is some content in the book. Memorize the content and answer my question after the book.

`{{book_content}}` Now the material ends.

Why doesn't Professor Snape seem to like Harry?



While Professor Snape says he doesn't "dislike" Harry but hates him, there are a few reasons he seems to treat Harry poorly:

1. Snape was a friend of Harry's parents, especially Lily Potter. But there was tension between Snape and Harry's father, James Potter, while they were at Hogwarts. So Snape's negative feelings toward James may transfer to Harry.
2. Snape wanted the Defense Against the Dark Arts teaching position but never got it. Professor Quirrell holds the position when Harry arrives at Hogwarts. Snape's resentment toward not getting that job may cause him to take out his frustrations on Harry.
3. As Harry gains fame for surviving Voldemort's attack as a baby, Snape likely resents Harry for achieving fame and attention that Snape never received, despite Snape's talents. This jealousy may fuel Snape's hostility toward Harry.

In summary, a combination of past history between Snape and Harry's father, professional jealousy, and resentment all seem to contribute to Snape's unfair treatment of Harry during his first Potions lesson.


Summarization



Below is some content in the book. Memorize the content and answer my question after the book.

`{{book_content}}` Now the material ends.

Please summarize the book in one paragraph.



War and Peace tells the story of five aristocratic Russian families set against the backdrop of the Napoleonic Wars. The novel explores themes of free will and determinism, fate and chance, and faith and reason. The main characters struggle to find meaning and purpose in life amidst the tumultuous events of the time. The book traces the growth and development of the characters over time, highlighting how they are shaped by war, love, family relationships, and the larger social and political upheavals during a turbulent and transformative period in Russian history. Despite all of the hardships and uncertainties, the characters find moments of joy, happiness, and connection with others that give meaning to their lives. The novel concludes with a sense of hope while also acknowledging the inevitability of death and the fleeting nature of human existence.

Figure 5: Examples on book sections<sup>456</sup> and questions of relationship, details, and summarization.

<sup>4</sup>[https://en.wikipedia.org/wiki/Harry\\_Potter\\_and\\_the\\_Philosopher%27s\\_Stone](https://en.wikipedia.org/wiki/Harry_Potter_and_the_Philosopher%27s_Stone)

<sup>5</sup>[https://en.wikipedia.org/wiki/Harry\\_Potter\\_and\\_the\\_Chamber\\_of\\_Secrets](https://en.wikipedia.org/wiki/Harry_Potter_and_the_Chamber_of_Secrets)

<sup>6</sup>[https://en.wikipedia.org/wiki/War\\_and\\_Peace](https://en.wikipedia.org/wiki/War_and_Peace)

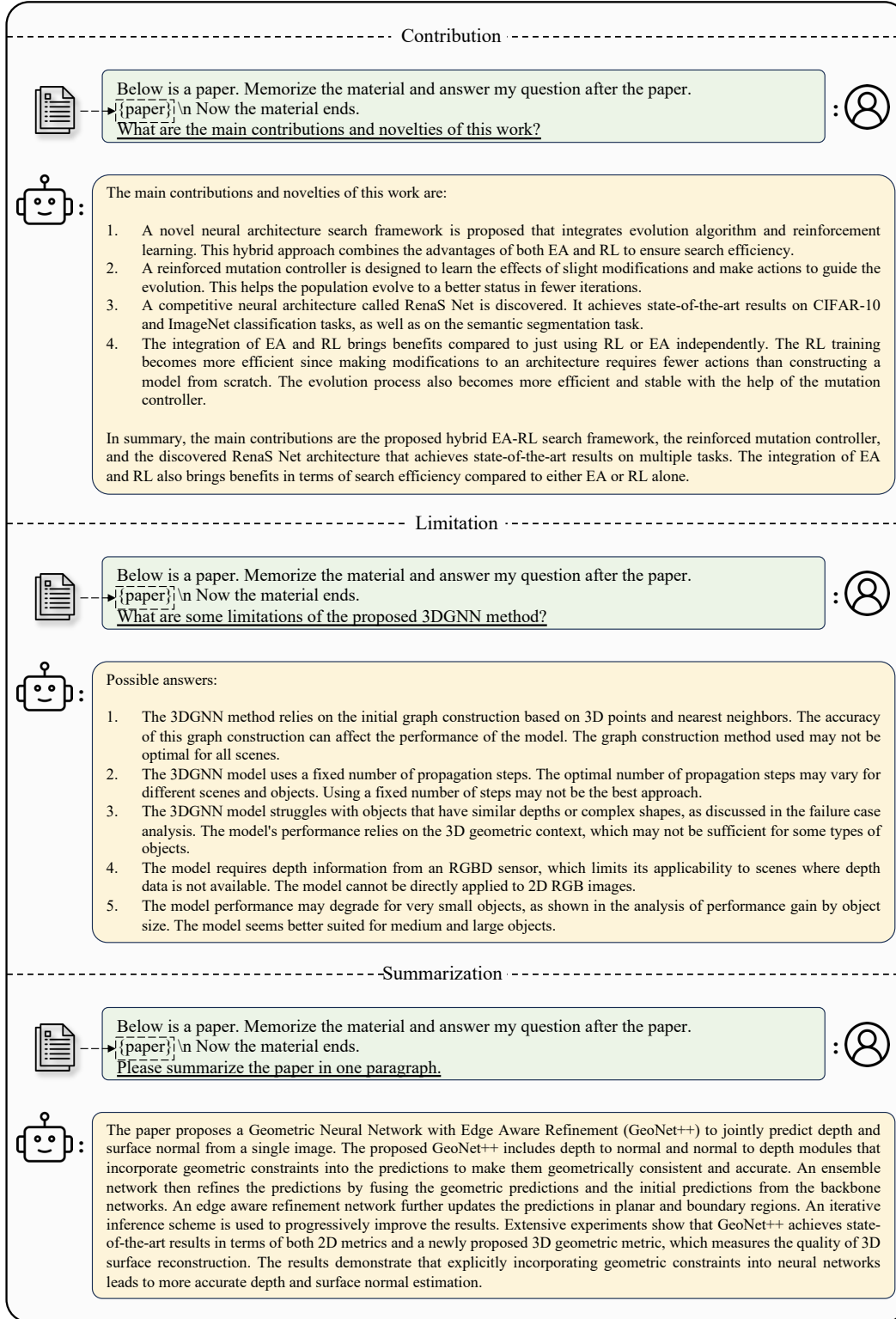


Figure 6: Examples on paper (Chen et al., 2019; Qi et al., 2017; 2022) and questions related to contribution, limitation, and summarization.