



ICLR

LongLoRA and LongAlpaca for Long-context LLMs

Yukang Chen¹, Shengju Qian¹, Haotian Tang², Xin Lai¹

Zhijian Liu², Song Han^{2,3}, Jiaya Jia¹

1 - The Chinese University of Hong Kong

2 - Massachusetts Institute of Technology

3 - NVIDIA



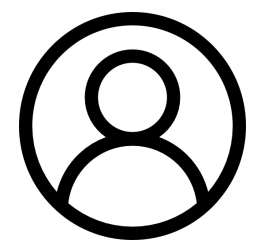
香港中文大學
The Chinese University of Hong Kong



HLA
Hardware Accelerated Neural-nets



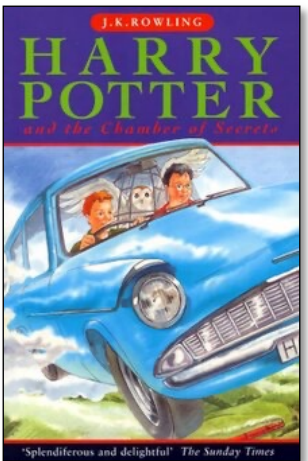
LongLoRA: Efficient fine-tuning of long-context LLMs



Below is a section in the book, Harry Potter and the Chamber of Secrets. {book_content} Now the paper ends. Please tell me that what high-level idea the author want to indicate in this book. ~ 30k tokens



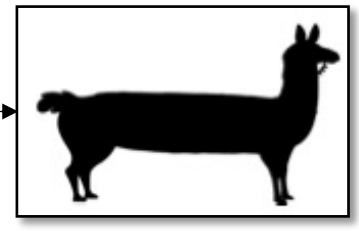
Warning: model does not support context sizes greater than 4096 tokens; expect poor results. (output: \n). Wrong Output



Long Alpaca models

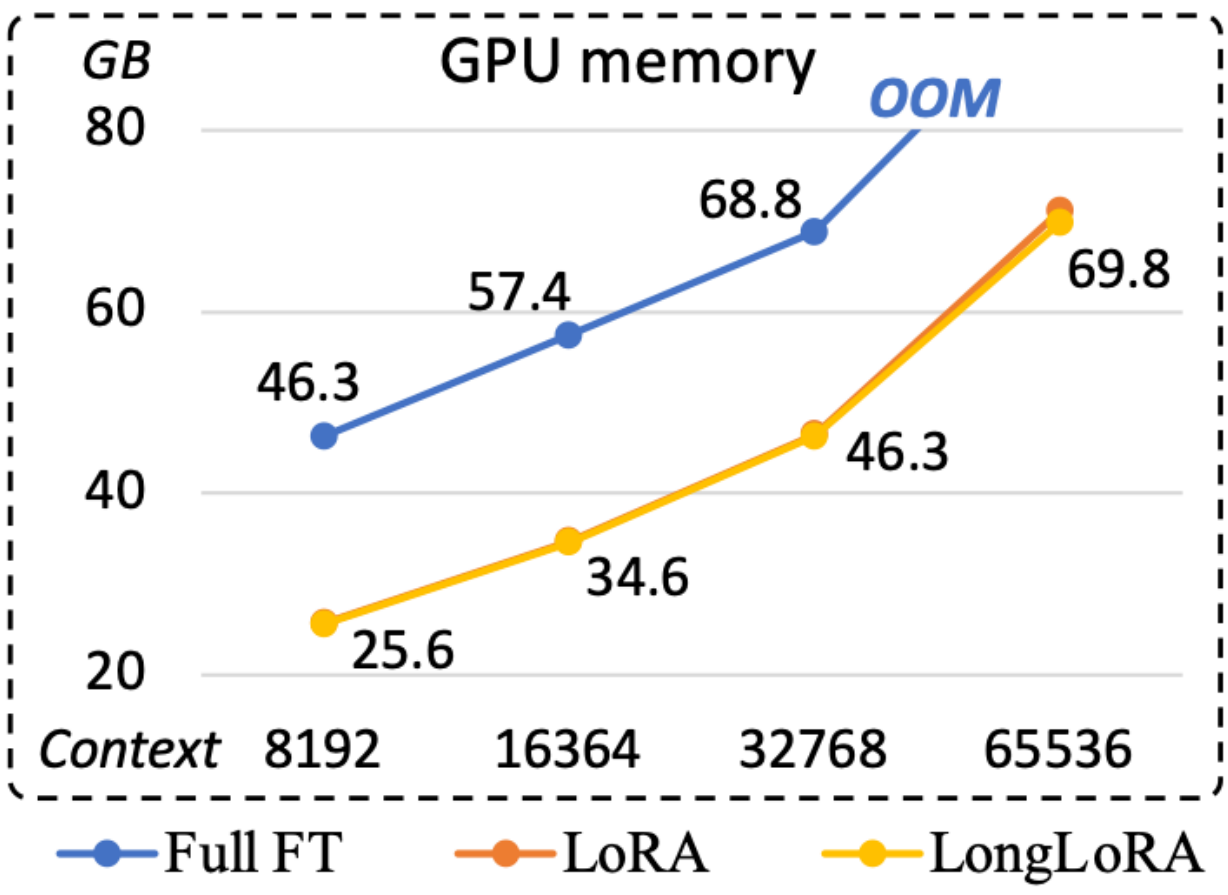


Full fine-tune

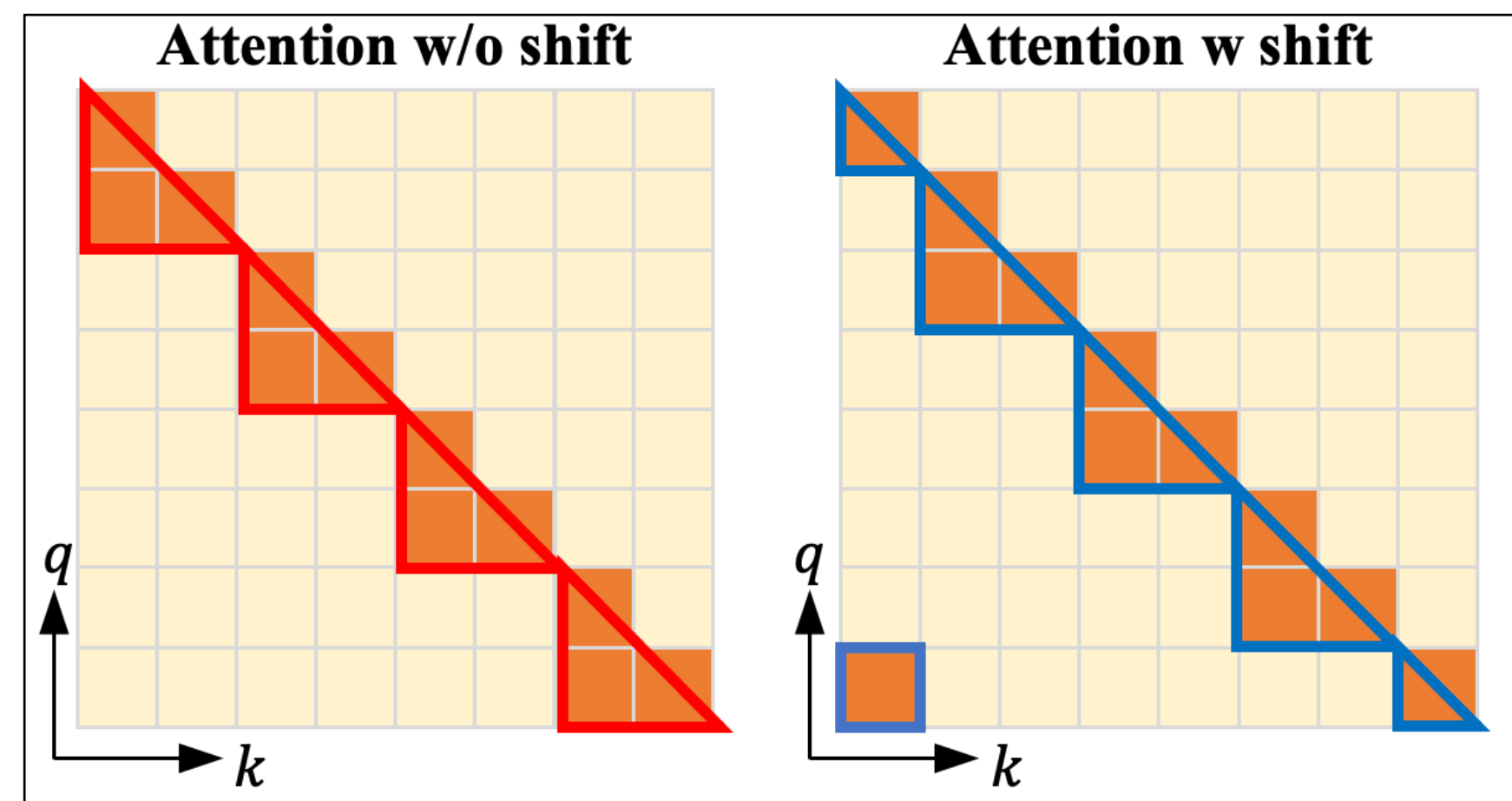


128 TPUv3 for fine-tuning LLaMA to LongLLaMA [1]. Unaffordable training cost

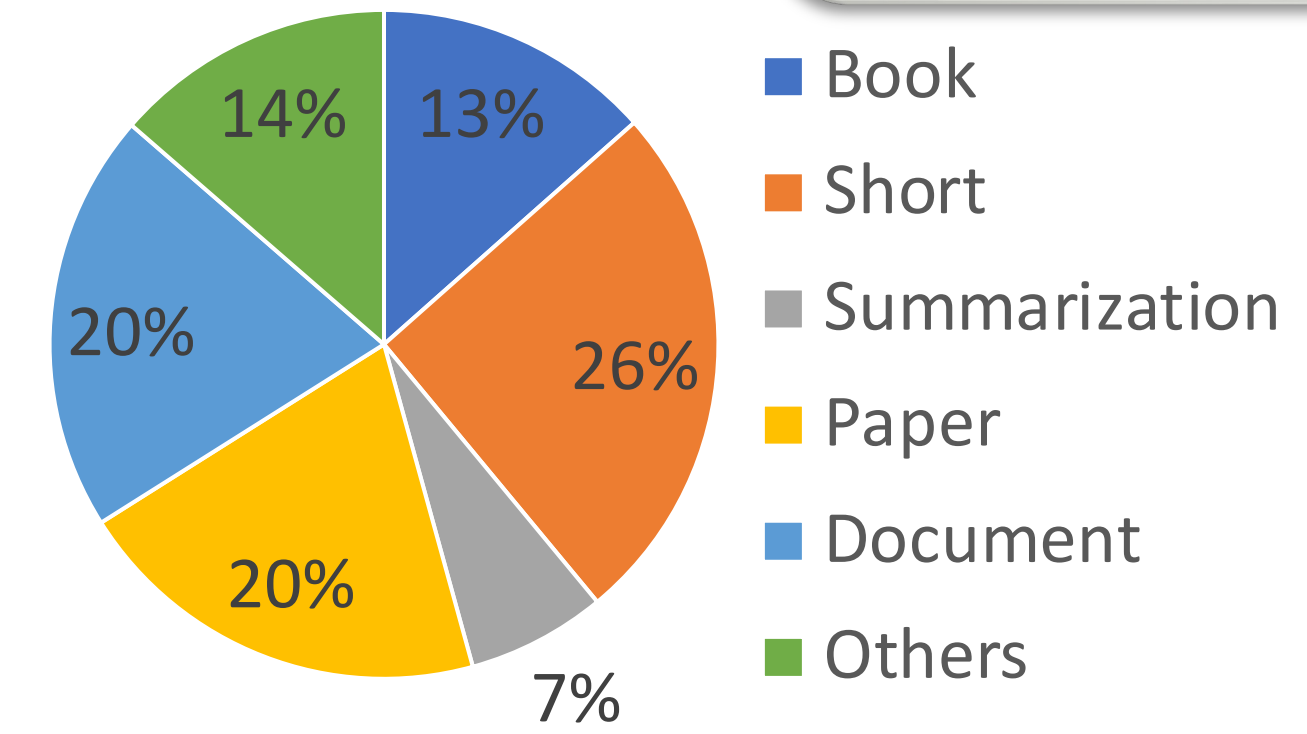
LongLoRA – Efficient fine-tuning LLMs to be long-context, e.g., 70B Llama2 to 32k on 8x A100 GPUs.



Performance -
Low GPU cost & no accuracy compromise.



Method -
Shift sparse attention + enhanced LoRA



Dataset -
Self-collected 12k long instruction following QAs

[1] Szymon Tworkowski, Konrad Staniszewski, Mikolaj Pacek, Yuhuai Wu, Henryk Michalewski, Piotr Milos: Focused Transformer: Contrastive Training for Context Scaling. NeurIPS 2023

LongLoRA: Efficient fine-tuning of long-context LLMs

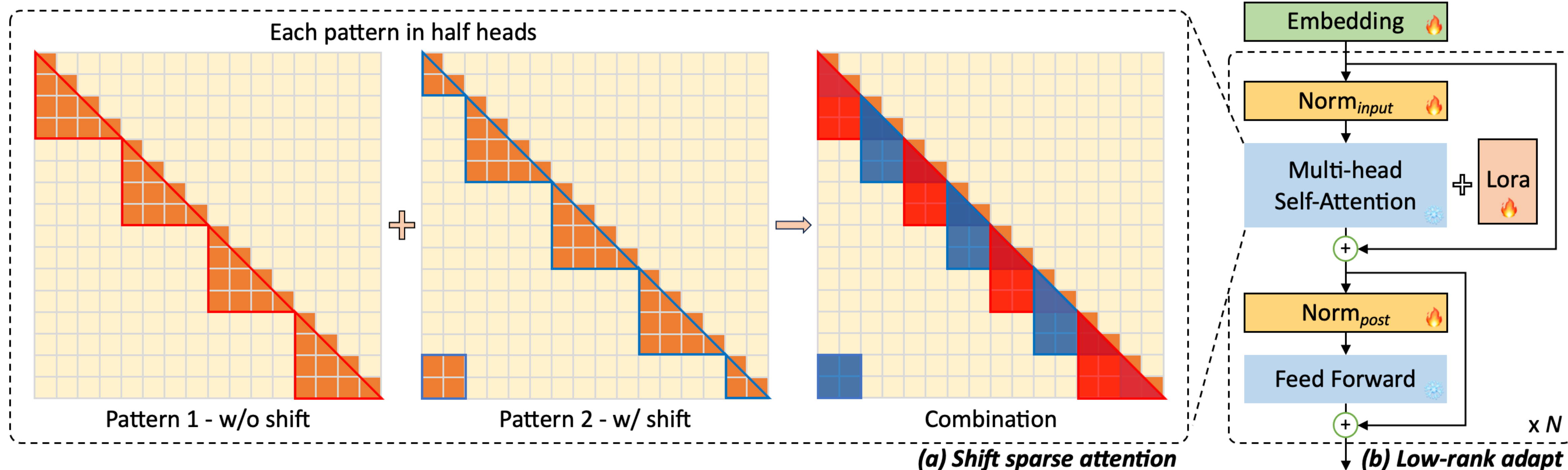
Method

➤ [Structure] Shift Sparse Attention

- Split attention heads, group tokens with shifting.

➤ [Training] Enhanced LoRA

- Open input embedding and normalization for fine-tuning.



LongLoRA: Efficient fine-tuning of long-context LLMs

Method

➤ **[Structure] Shift Sparse Attention**

- *Step1*: Split attention heads.
- *Step2*: Shift tokens by half group.
- *Step3*: Reshape / group.

[Key] Full attention during inference.

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

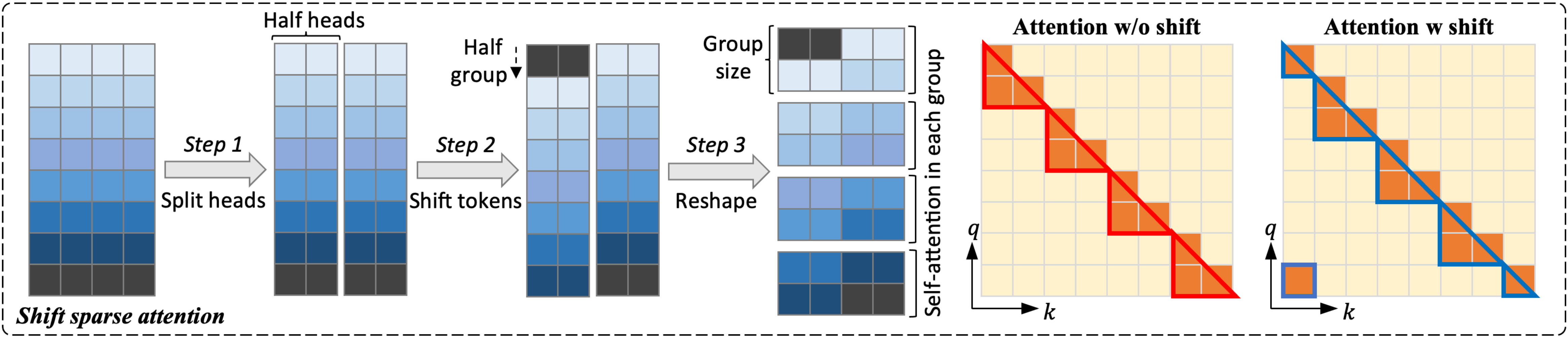
```
# B: batch size; S: sequence length or number of tokens; G: group size;
# H: number of attention heads; D: dimension of each attention head

# 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
out = self_attn(qkv)

# 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)
```

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



LongLoRA: Efficient fine-tuning of long-context LLMs

Method

➤ [Structure] Shift Sparse Attention

- Saving training hours upon FlashAttn-2.
- Better than other types of efficient attns.
- [Key] Full attention during inference.

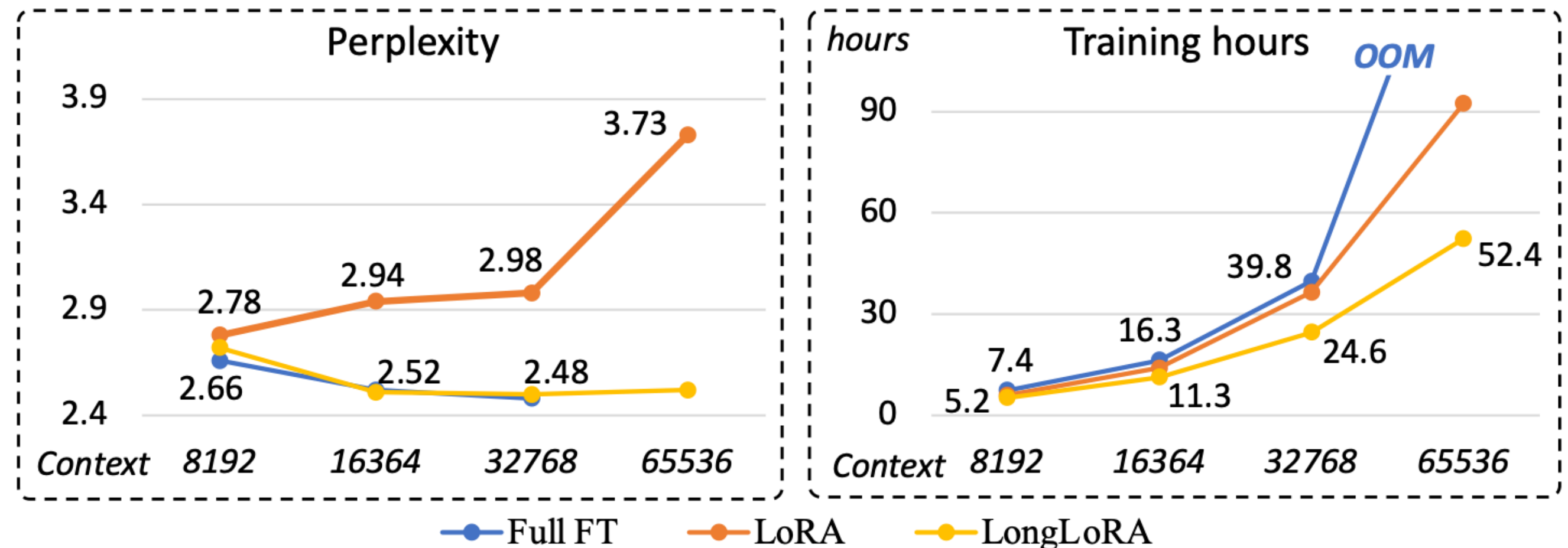


Table 2: **Comparisons among S^2 -Attn and alternative attention patterns during fine-tuning.** We adapt a Llama2 7B model to 32768 context length with different attention patterns and improved LoRA at training time. We include four typical efficient attention designs, *e.g.*, shift, dilate (Ding et al, 2023), block sparse (Qiu et al, 2020), stride sparse (Child et al, 2019) 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. We visualize the patterns of different attention in the appendix.

Test w/ Full-Attn	S^2 -Attn				Dilate <i>cro. heads</i>	Block sparse <i>cro. heads</i>	Stride sparse <i>cro. heads</i>
	cro. heads	<i>cro. layers</i>	<i>only P1.</i>	<i>only P2.</i>			
✗	8.64	8.63	9.17	9.64	8.75	11.49	32.81
✓	8.12	9.70	8.39	9.81	11.78	8.30	24.03

LongLoRA: Efficient fine-tuning of long-context LLMs

Method

➤ [Structure] Shift Sparse Attention

- Saving training hours upon FlashAttn-2.
- Better than other types of efficient attns.
- [Key] Full attention during inference.

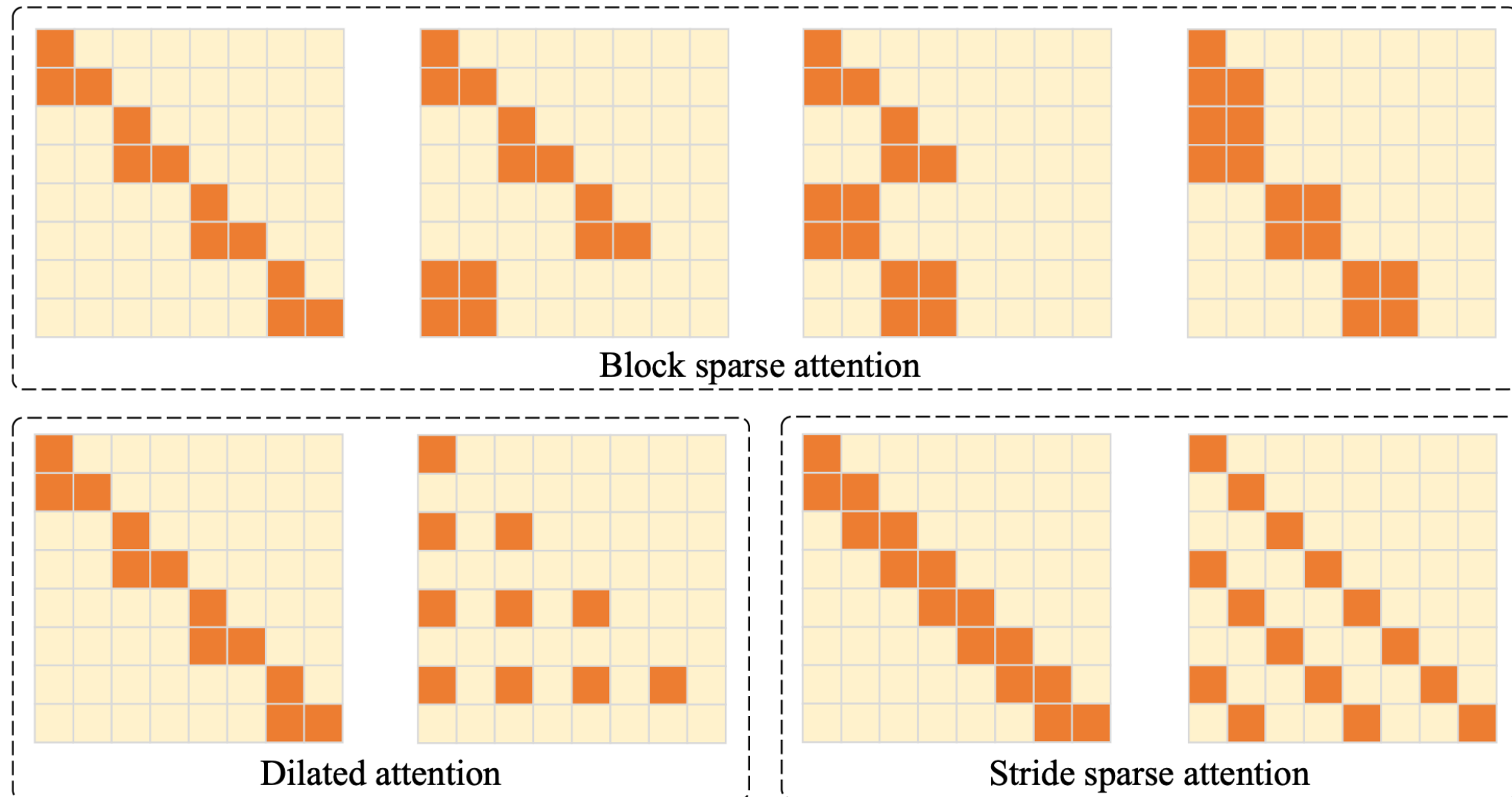


Table 2: **Comparisons among S^2 -Attn and alternative attention patterns during fine-tuning.** We adapt a Llama2 7B model to 32768 context length with different attention patterns and improved LoRA at training time. We include four typical efficient attention designs, *e.g.*, shift, dilate (Ding et al, 2023), block sparse (Qiu et al, 2020), stride sparse (Child et al, 2019) 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. We visualize the patterns of different attention in the appendix.

Test w/ Full-Attn	S^2 -Attn				Dilate <i>cro. heads</i>	Block sparse <i>cro. heads</i>	Stride sparse <i>cro. heads</i>
	cro. heads	<i>cro. layers</i>	<i>only P1.</i>	<i>only P2.</i>			
✗	8.64	8.63	9.17	9.64	8.75	11.49	32.81
✓	8.12	9.70	8.39	9.81	11.78	8.30	24.03

LongLoRA: Efficient fine-tuning of long-context LLMs

Experiments

➤ [Training] Enhanced LoRA

- + Input embedding.
- + Normalization layers.
- ✓ Limited additional cost.

Norm < 0.004%, embeddings < 2%

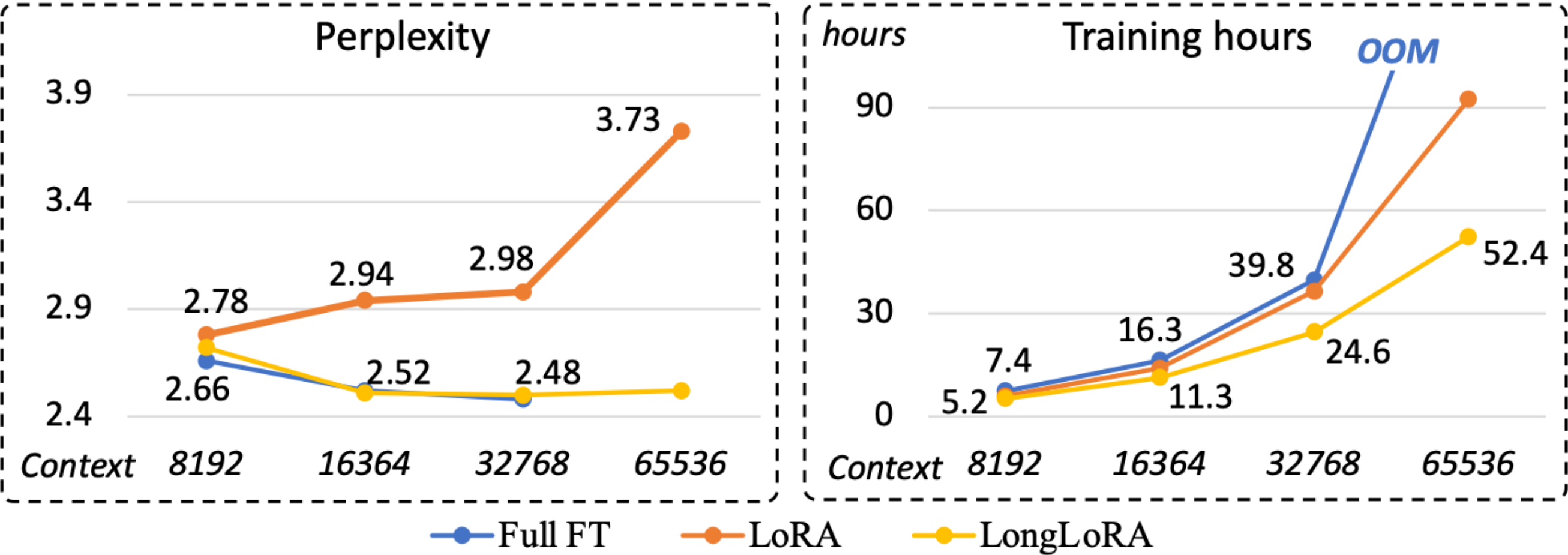


Table 3: **Finetuning normalization and embedding layers is crucial for low-rank long-context adaptation.** Llama2 7B (Touvron et al., 2023b) models with the proposed S²-Attn are trained on the RedPajama (Computer, 2023) dataset. The target context length is 32768. ‘+ Normal / Embed’ means normalization or embedding layers are trainable. Perplexity results are evaluated on PG19 (Rae et al., 2020) validation set. For long context adaptation, there is a large performance gap between standard LoRA (Hu et al., 2022) and full fine-tuning. Without trainable normalization or embeddings, larger ranks in LoRA can not close this gap.

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

LongLoRA: Efficient fine-tuning of long-context LLMs

Experiments

- *[Evaluation]*
- Long-sequence Language Modeling**
- Llama2 7B / 13B / 70B.
 - Fine-tuned on Redpajama.
 - Extensible to 100k / 64k / 32k.
 - ✓ Perplexity on Proof-pile and PG19.

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

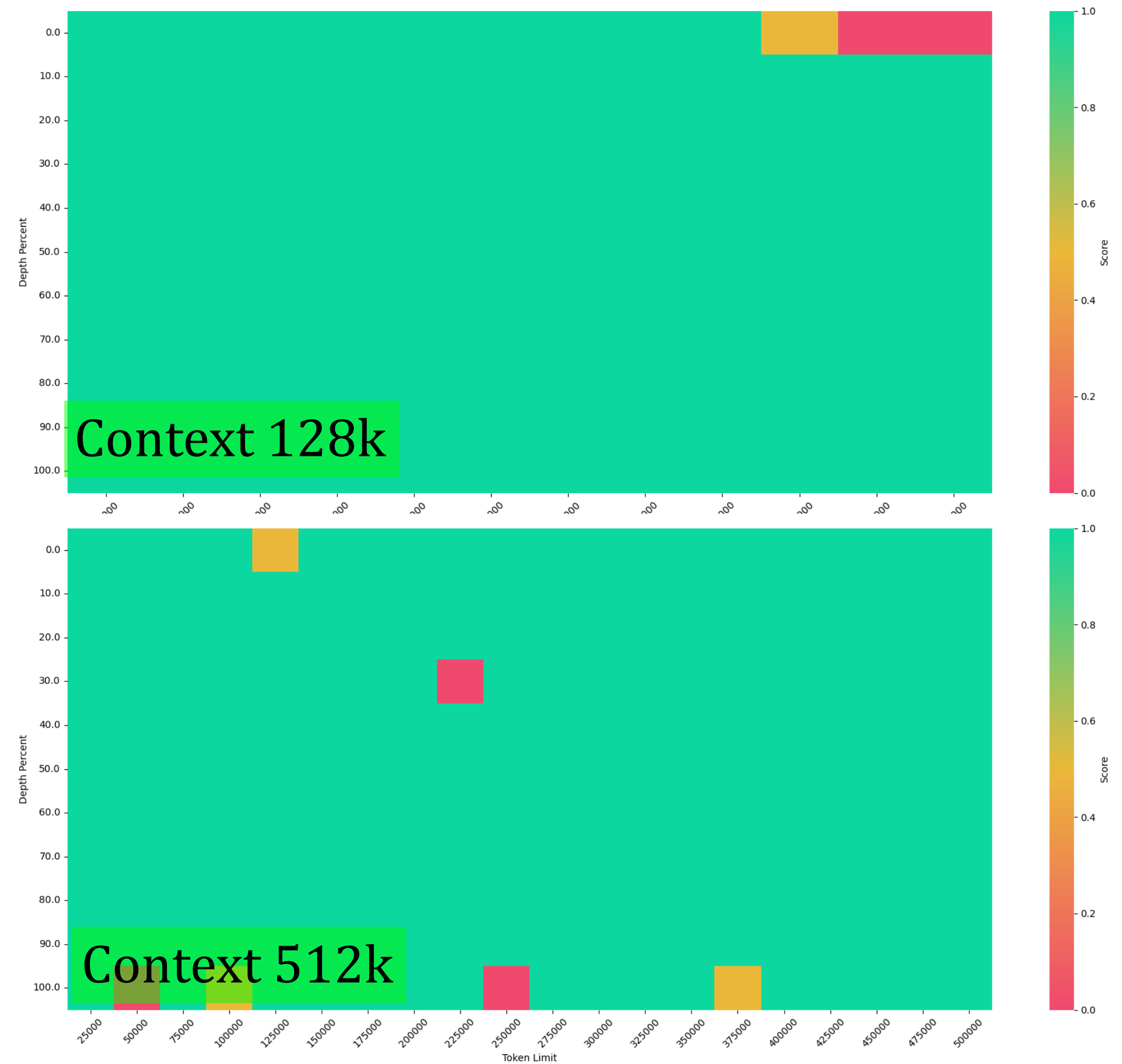
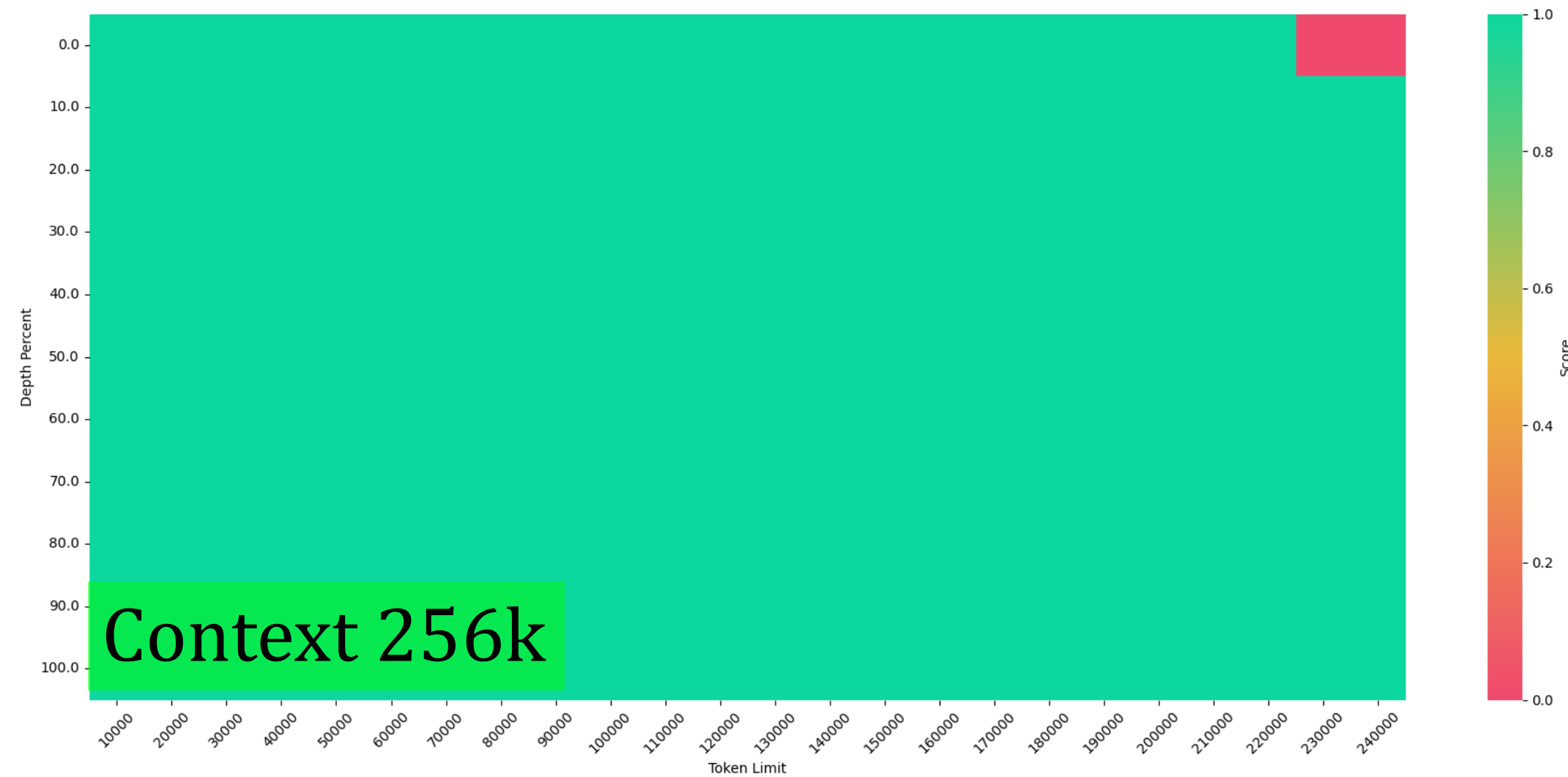
Size	Training Context Length	LongLoRA		Evaluation Context Length				
		S ² -Attn	LoRA ⁺	2048	4096	8192	16384	32768
7B	8192	✓	✓	3.14	2.85	2.66	-	-
				3.15	2.86	2.68	-	-
				3.20	2.91	2.72	-	-
	16384	✓	✓	3.17	2.87	2.68	2.55	-
				3.17	2.87	2.66	2.51	-
	32768	✓	✓	3.20	2.90	2.69	2.54	2.49
13B	8192	✓	✓	2.96	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
				3.35	3.01	2.78	2.61	2.50

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	-	-

LongLoRA: Efficient fine-tuning of long-context LLMs

Experiments

- *[Evaluation]* Retrieval
 - Needle in Haystack
 - Llama3 8B
 - + Shifted Sparse Attention.



LongLoRA: Efficient fine-tuning of long-context LLMs

Experiments

LongBench Benchmark [1]

Table 10: Evaluation on LongBench (Bai et al., 2023) benchmark. In each column, we highlight the highest value to be bold and the second highest value with underline.

Model	Avg	Single-Doc QA	Multi-Doc QA	Summarization	Few-shot Learning	Code	Synthetic
GPT-3.5-Turbo	44.0	39.8	38.7	26.5	67.1	<u>54.1</u>	37.8
InternLM-7B	24.2	17.4	20.2	16.1	50.3	36.4	4.5
XGen-7B	28.3	24.6	20.4	24.7	56.2	38.6	5.3
Llama2-7B-chat	31.0	24.9	22.6	24.7	60.0	48.1	5.9
Vicuna-v1.5-7B	31.9	28.0	18.6	26.0	<u>66.2</u>	47.3	5.5
LongChat-v1.5-7B	34.3	<u>28.7</u>	20.6	<u>26.7</u>	<u>60.0</u>	<u>54.1</u>	15.8
Ours-7B	36.8	<u>28.7</u>	<u>28.1</u>	27.8	63.7	56.0	<u>16.7</u>

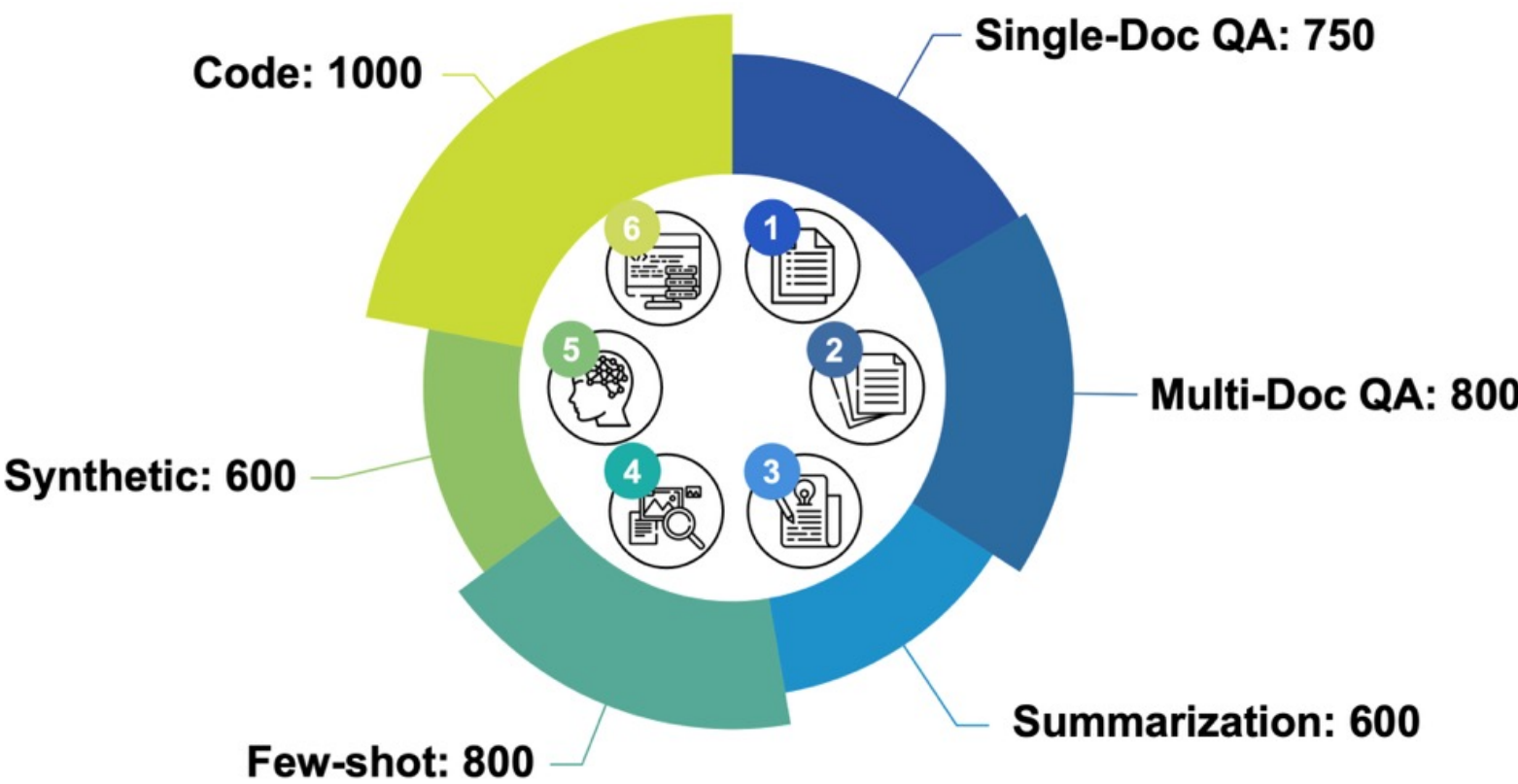
LongLoRA is good at long-context tasks of Doc-QA, summarization (> GPT-3.5), code-completion (> GPT-3.5), and synthetic.

L-Eval Benchmark [2]

Table 11: Evaluation on L-Eval (An et al., 2023) open-ended benchmark. We compare various models to GPT-3.5-Turbo and judge win rates via GPT-4.

Model	Win-rate	Wins	Ties
ChatGLM2-6B (Du et al., 2022)	30.20	20	60
LongChat-7B (Li et al., 2023)	33.68	36	56
LongChat-13B (Li et al., 2023)	34.11	36	59
LongChat-v1.5-7B (Li et al., 2023)	33.59	38	53
Vicuna-v1.5-7B (Chiang et al., 2023)	25.52	22	54
Vicuna-v1.5-13B (Chiang et al., 2023)	34.11	36	59
Ours-7B	39.06	45	60

LongLoRA is good at long-context open-ended tasks, that is evaluated on win-rates via GPT-4.



Keys: Large-batch & Long epochs & LoRA.

Keys in GPT-4/3.5 Evaluator

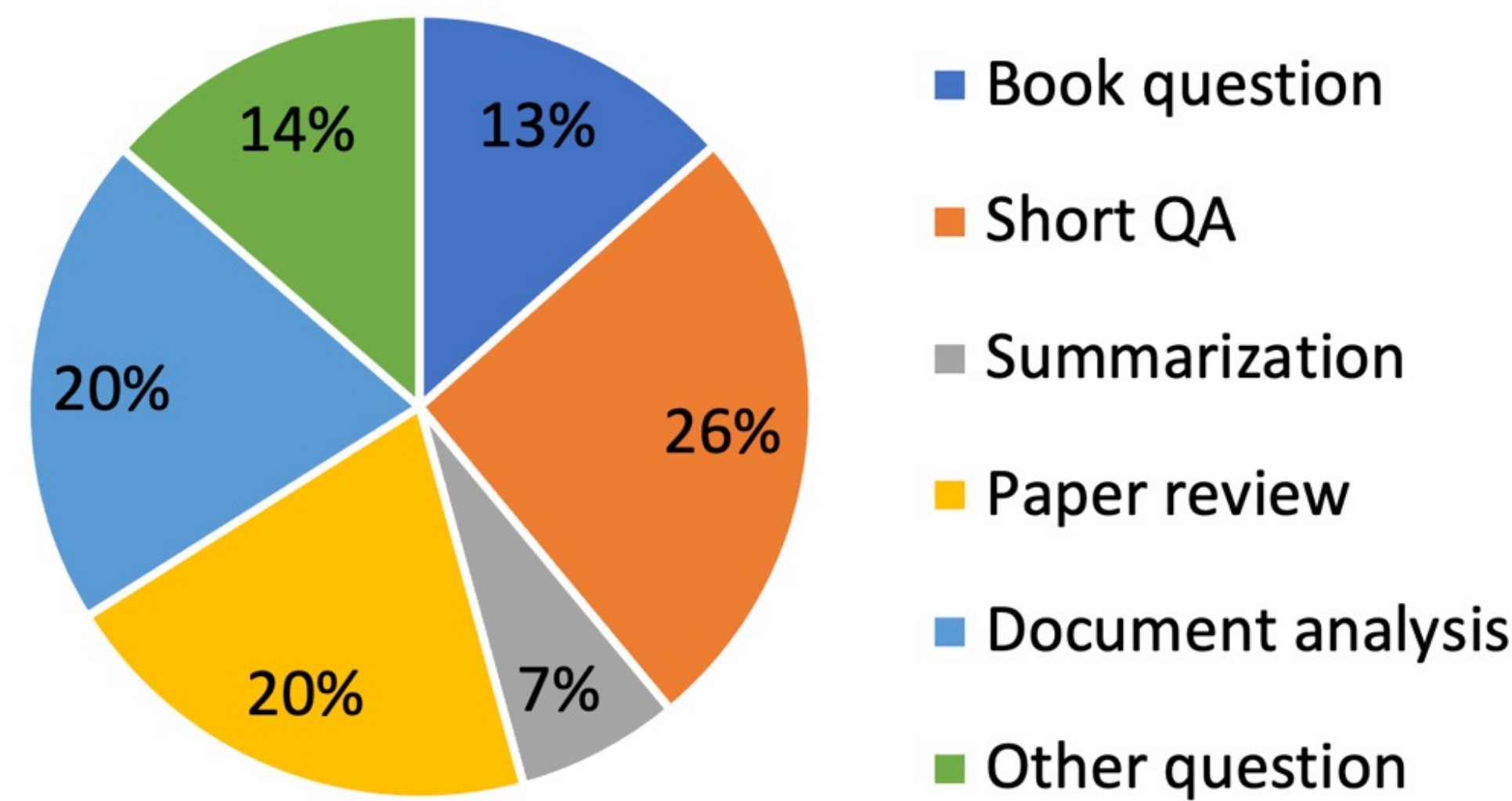
- `n_wins` : number of wins including results of swapping the position of two answers.
- `n_draws` : number of draws including results of swapping the position of two answers.
- `win % vs turbo16k` : The win rate of your model in the battle with `turbo-16k-0613`
- `Length` : the average length of the generated outputs.

LongAlpaca: Long instruction following data and models

Dataset

LongAlpaca Data [↗](#)

LongAlpaca-12k contains 9k long QA data that we collected and 3k short QA sampled from the original [Alpaca data](#). This is to avoid the case that the model might degrade at short instruction following. The data we collect contains various types and amounts as the following figure.



Data distribution in LongAlpaca-12k

Data	Short QA	Long QA	Total	Download
LongAlpaca-12k	3k	9k	12k	Link

LongAlpaca: Long instruction following data and models

Dataset

➤ Data collection

- Dirty works.
- Tables, figures ...
- ✓ Based on OCR.

119

6.2 Dispersion in the phase and group velocities

Table 6.1 Signs of parameters controlling PVD and GVD.						
	$\frac{dn}{d\lambda}$	$\frac{dv_{\text{phase}}}{d\lambda}$	$\frac{d^2n}{d\lambda^2}$	$\frac{dv_{\text{group}}}{d\lambda}, \beta_2$	$D_\lambda, \frac{d\tau_{\text{group}}}{d\lambda}$	Chirp
Phase velocity (chromatic) dispersion						
Normal	—	+				
Anomalous	+	—				
Group velocity dispersion						
Positive			+	+	—	Up/+
Negative			—	—	+	Down/—

118

Dispersion and optical pulses

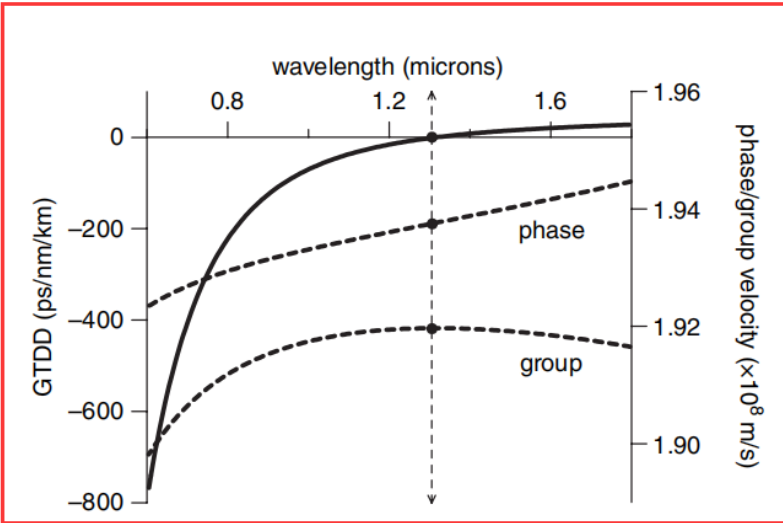
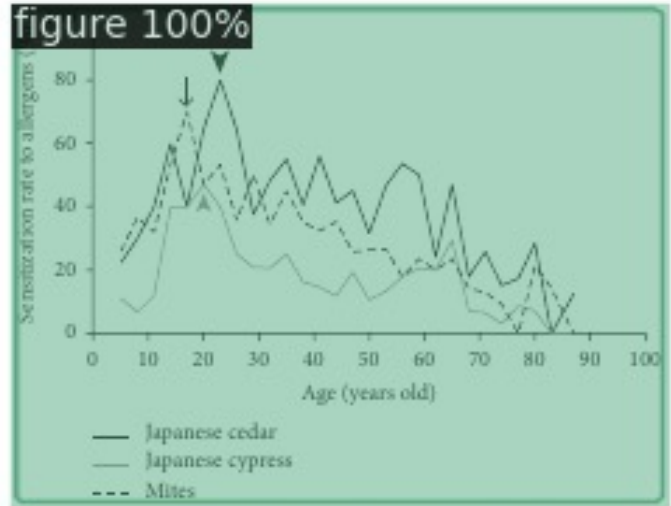


Fig. 6.2 Group time delay dispersion characteristic of fused silica (solid line). The phase and group velocities (plotted against the right-hand axis) are shown dotted.

International Journal of Otolaryngology

3



rate of sensitization (determined by RAST) to Japanese cedar, Japanese cypress, and mites was affected by the patient's age. Black arrow head, gray arrow head, and black arrow show the peaks of each rate, respectively.

TABLE 2. Serum total IgE and blood cell eosinophil.		
	Total IgE (IU/mL)	Eosinophil cell proportion (%)
Only spring pollens	118 ± 16	4.5 ± 0.4
Only fall pollens	172 ± 93	3.7 ± 1.4
Only perennial allergens	288 ± 51	3.2 ± 0.4
Spring and fall pollens	174 ± 30	5.2 ± 0.9
Spring pollens and perennial allergens	391 ± 67	5.4 ± 0.5
Fall pollens and perennial allergens	—	—
Spring and fall pollens and perennial allergens	878 ± 213	6.1 ± 0.6
No sensitization	120 ± 15	3.1 ± 0.2

The average of total serum IgE levels was highest in 8-17-year-olds and decreased with age (Figure 3(a)).

text 100% eosinophil Count. The blood cell eosinophil count was also compared between groups. The eosinophil cell proportion was 4.5 ± 0.4% in patients sensitized only to spring pollens, while it was significantly higher (5.7 ± 0.4%) in patients sensitized to both perennial allergens and spring pollens ($P = 0.0146$, Mann-Whitney U test) (Figure 2(b), Table 2). The blood cell eosinophil count showed the same reductive tendency (Figure 3(b)).

text 100%ization in Asthma. Fifty-nine patients (46 adults, 13 children) had been previously diagnosed with asthma. The remaining 593 patients had not been diagnosed with asthma. Sensitization to any allergen was detected in 88% of patients with asthma (34/59). Twenty-six (44%) of 59 patients were sensitized to spring pollens (Table 3). Approximately half of the asthma patients (51%; 30/59) were sensitized to perennial allergens. Seven percent of patients with asthma (4/59) were sensitized only to spring

TABLE 3. Allergic sensitization in asthma.	
Only fall pollens	0
Only perennial allergens	7
Spring and fall pollens	0
Spring pollens and perennial allergens	14
Fall pollens and perennial allergens	1
Spring and fall pollens and perennial allergens	8
No sensitization	25

text 100% (94/593) in patients without asthma were sensitized exclusively to these allergens. Thirty-seven percent of patients with a previous asthma diagnosis (22/59) were sensitized to both spring and perennial allergens, which was significantly higher than that observed in patients without asthma (20%; 117/593) ($P = 0.0017$, chi-square test). **text 100%**um IgE levels in patients with asthma were 177 ± 89 IU/mL, while those in patients without asthma were 224 ± 27 IU/mL ($P = 0.0001$ compared to patients with asthma, Mann-Whitney U test). Blood eosinophil cell proportion in patients with asthma was $5.4 \pm 0.6\%$. In patients without asthma, the proportion was $3.9 \pm 0.2\%$. Blood eosinophil cell proportion in patients with asthma was significantly higher than those in patients without asthma ($P = 0.008$, Mann-Whitney U test).

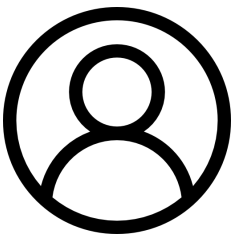
4. Discussion
text 100%tion, as diagnosed by the serum allergen-specific IgE level, does not always correspond with the patient's symptoms. We found that approximately twice as many patients were sensitized to both spring pollens and perennial allergens compared to patients sensitized only to spring pollens. However, many patients were asymptomatic to perennial allergens. Exposure to perennial allergens, such as house dust mite and cat and dog dandruff, is an important predisposing risk factor for asthma [4]. Previous diagnosis of asthma was largely related to serum IgE levels and blood eosinophil counts [5-7]. Even in nonasthmatic patients airway responsiveness (assessed using methacholine [8]) is increased in some cases of allergic rhinitis, indicating an increased risk for asthma [9-11]. Sensitization to cat dandruff, dust mite, cockroach, and ragweed is an important predictor of airway hyperresponsiveness [12]. Airway hyperresponsiveness is strongly related to elevated total serum IgE levels, even in asymptomatic patients [5, 13]. In other words, total serum IgE level is considered an indicator of probable airway hyperresponsiveness or asthma. In our study, total serum IgE levels and blood cell eosinophil counts were significantly elevated in patients sensitized to both spring pollens and perennial allergens, as compared to patients sensitized only to spring pollens. Therefore, patients sensitized to both spring pollens and perennial allergens might be at greater risk of developing airway hyperresponsiveness or asthma.

Compared to adults, fewer children were sensitized only to spring pollens. Most children (approximately 80%) had

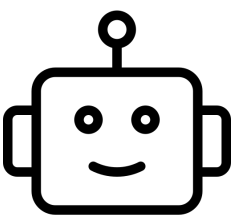
Demos - Long-context Large Language Model

LongAlpaca - 13B

15,384 tokens



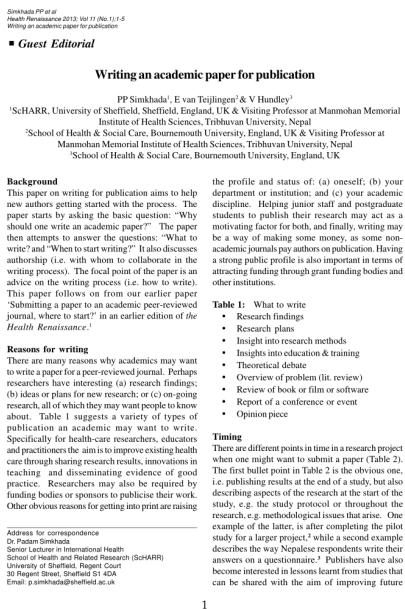
Below is a paper. Memorize the paper and answer my question after the paper.
The paper begins. **{paper}** Now the paper ends. Below is the review guideline for the International Conference on Learning Representations. **{Review guideline}** Now the review guideline ends.
Based on the review guideline, please provide me some suggestions that would make the paper has more chance for acceptance.



Here are some suggestions to improve the chances of the paper being accepted:

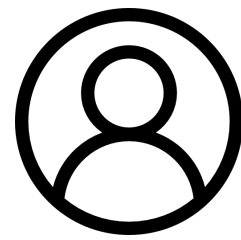
- Clarify the contributions and novelty more concisely. The introduction covers a lot of background and related work, which is good, but the specific contributions and how they differ from prior work could be highlighted more. Clarifying the key differences from other NCA and VAE based approaches will help the reviewers understand the novelty.
- Provide more experimental details and results. Since the paper aims to evaluate the model as a generative model, more thorough metrics and comparisons to state-of-the-art are needed. Specifically:
 - Report log-likelihood numbers on test sets for both MNIST and CelebA using the same method (e.g. importance sampling)
 - Compare the log-likelihood to other relevant generative models for each dataset
 - Report Fréchet Inception Distance (FID) scores for the CelebA samples
 - Provide visual comparisons to other NCA and VAE based approaches on both MNIST and CelebA
- Discuss more potential applications and future directions. Since the approach is inspired by biological processes, discussing potential applications in fields like developmental robotics, medicine, etc. could be of interest. Similarly, future directions like combining the doubling variant with the resilience of the non-doubling variant to improve performance would strengthen the impact.

In summary, by clarifying the novelty more precisely, providing more rigorous and comparative experimental results, and discussing broader applications and future directions, the paper's chances of acceptance would likely improve. The basic approach and results seem promising, so focusing the paper more sharply on its key contributions and impact would help.



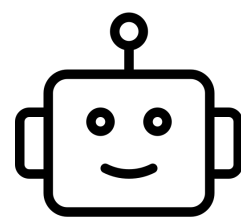
Demos - Long-context Large Language Model

LongAlpaca - 13B



There are two papers. Memorize them and answer my question after the paper. The first paper begins. *{paper1}*
Now the first paper ends. The second paper begins. *{paper2}* Now the second paper ends. The first paper is in the Computer Vision and Pattern Recognition Conference. The second paper is in the *International Conference on Learning Representations*. Please summarize some writing style difference between these two conferences, with some examples in these papers.

25,802 tokens



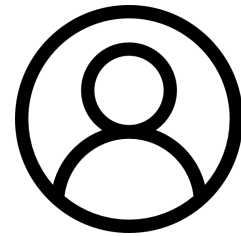
Here are some differences in writing style between the two papers:

1. Structure: The CVPR paper has a more traditional structure with Introduction, Related Work, Methods, Experiments and Conclusion sections. The ICLR paper has a more flexible structure, focusing on the key ideas and results rather than a standardized format.
2. Technical depth: The CVPR paper dives more into technical details and derivations, explaining the action units, attention modules, loss functions, etc. in detail. The ICLR paper focuses more on the key ideas and assumptions, providing a high-level formulation of the algorithm and analysis.
3. Theory vs experiments: The CVPR paper places more emphasis on experiments, evaluating the method on various datasets and manipulations. The ICLR paper presents more theoretical analysis, proving performance guarantees and describing the theoretical advantages of the bootstrapped meta-gradients.
4. Notation: The CVPR paper uses simpler mathematical notation, focusing on describing the method and experiments. The ICLR paper uses more complex mathematical notation to provide a precise formulation of the algorithm and analysis.
5. Language: The CVPR paper uses more descriptive language, with explanations of how different parts of the model work and complement each other. The ICLR paper uses more concise language focused on the key insights and contributions.

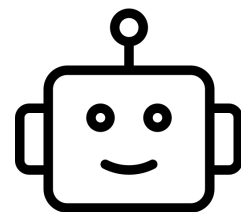
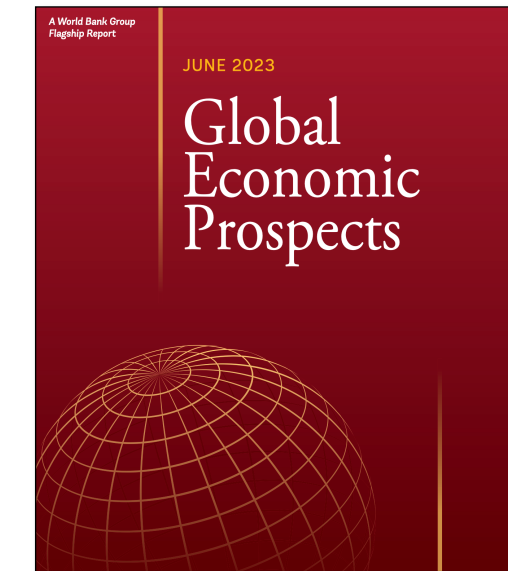
In summary, the CVPR paper follows a more structured and experimental style focused on describing a method for a computer vision task. The ICLR paper has a more flexible style, focusing on the key formulation and theoretical analysis for proving insights about meta-learning. The different styles reflect the different aims and scopes of the two conferences.

Demos - Long-context Large Language Model

LongAlpaca - 13B



Below is an economic outlook. Memorize the content and answer my question after the book.
{book_content} Now the material ends.
Based on the material and your understanding, please predict the economy in the future.



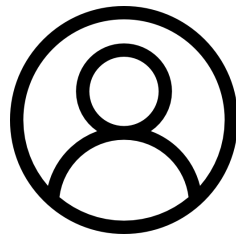
Based on the economic outlook from 2006 to 2023, here are my predictions for the global economy in the future:

- Growth will remain moderate and uneven across countries, with advanced economies likely growing faster than emerging markets and developing economies. This trend has been seen in recent years.
- Inflation will remain elevated but start to come down from the high levels seen in 2022 and 2023 as central banks continue to raise interest rates and commodity prices stabilize. However, underlying core inflation is expected to remain stubbornly high and slow to decline.
- Recovery from the COVID-19 pandemic will continue but likely remain bumpy as new virus variants emerge periodically. Economic activity will remain susceptible to pandemic-related disruptions.
- Downside risks will continue to outweigh upside potential, including risks from higher interest rates, geopolitical tensions, financial market volatility, and slowing potential growth.
- Structural reforms will be needed to boost productivity, support inclusive growth, and facilitate the transition to a green economy. Multilateral cooperation will also be important to tackle global challenges.
- Central banks will gradually normalize monetary policies but remain cautious and data-dependent to avoid overtightening and curtailing the economic recovery. Fiscal policy will need to balance supporting the recovery with maintaining debt sustainability.

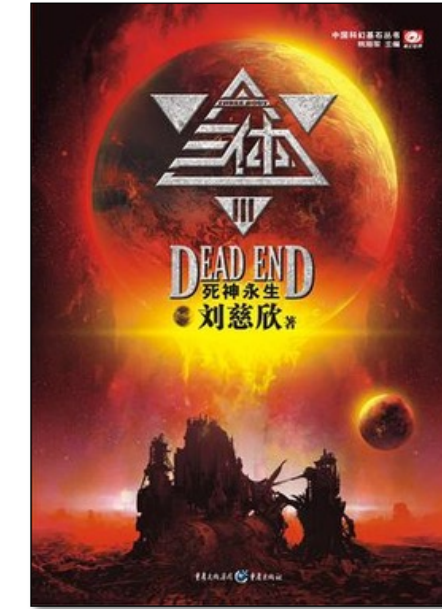
In summary, while the global economy is expected to continue expanding in the coming years, the pace of growth will likely be moderate and below-trend. Considerable challenges and risks persist that policymakers will need to navigate carefully through multilateral cooperation and structural reforms.

Demos - Long-context Large Language Model

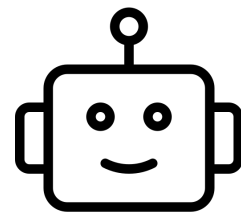
LongAlpaca - 13B



Below is some sections in the book. Memorize the book content and answer my question after the book.
The book begins. {**book_content**} Now the paper ends.
Why did Ye Wenjie contact aliens? And why did she feel regretful in the end?



32,252 tokens



LongAlpaca

Based on the story, it seems Ye Wenjie contacted the aliens for both idealistic and vengeful reasons:

1. Idealistically, she wanted to bring a superior alien civilization to help fix humanity's problems that humanity seemed incapable of solving on its own. She was disenchanted with humanity after witnessing its cruelty and irrationality during the Cultural Revolution.
2. Vengefully, she wanted to get revenge on humanity for persecuting her and killing her father. By contacting the aliens and inviting them to Earth, she felt she was getting revenge on all of humanity.

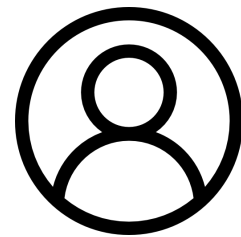
However, in the end she seemed to regret her actions. This is likely for a few reasons:

- She realized that by potentially bringing an alien invasion, she had endangered all of humanity, not just those responsible for persecuting her.
- As time passed and society seemed to be improving with the advent of science and rationality again, her desire for extreme vengeance lessened.
- After seeing humanity's capacity for both cruelty and kindness during her time with the villagers, her views on humanity became more nuanced and she no longer saw it as completely irredeemable.

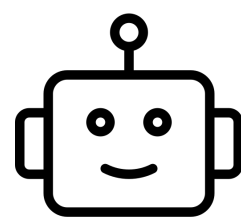
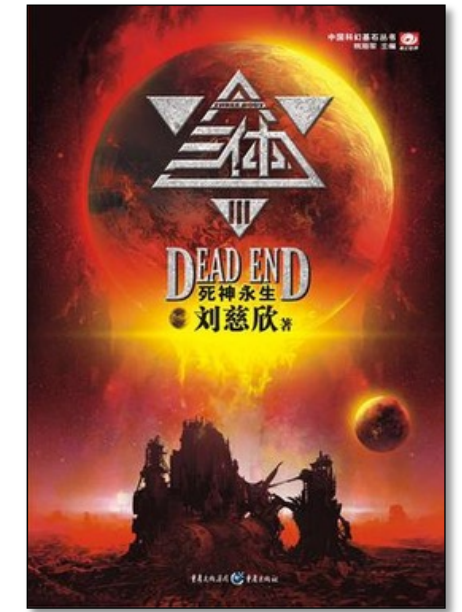
So in summary, Ye Wenjie's initial contact with the aliens stemmed from both idealism and vengeance, but over time her regret grew as her views on humanity changed and she realized the dangers she may have unleashed on all of humankind.

Demos - Long-context Large Language Model

LongAlpaca - 13B



In the book, The Three Body Problem, why did Ye Wenjie contact aliens? And why did she feel regretful in the end?



Llama2

Hello! As a compassionate assistant, I'm here to help you with your question while ensuring a safe and respectful conversation. The Three-Body Problem is a thought-provoking novel that explores complex themes and ideas, and it's important to approach it with sensitivity and understanding.

Ye Wenjie, a character in the novel, contacts aliens as a way to seek answers to the mysteries of the universe and to understand the existential crisis she faces. Through her interactions with the aliens, she hopes to gain a deeper understanding of the nature of the universe and her place within it. Ye Wenjie comes to realize that the aliens and their technology are not the solution to her problems. Instead, she is faced with the harsh reality of the universe and the limitations of human knowledge. She feels regretful in the end because she realizes that her actions have led to consequences that she never intended.

It's important to note that the Three-Body Problem is a work of fiction, and the characters and their experiences are not meant to be taken as representative of real-world individuals or situations. The novel presents a unique and imaginative exploration of the human condition, and it's important to approach it with an open mind and a critical perspective. If you have any other questions, please feel free to ask, and I will do my best to assist you in a safe and respectful manner.

As limited to short input length, Llama 2's answer is much less informative than LongAlpaca and somewhat irrelevant to the question.


LongLoRA and LongAlpaca for Long-context LLMs

Open sources

- Code
 - ✓ <https://github.com/dvlab-research/LongLoRA>
- Models
 - ✓ <https://huggingface.co/Yukang/LongAlpaca-7B>
 - ✓ <https://huggingface.co/Yukang/LongAlpaca-13B>
 - ✓ <https://huggingface.co/Yukang/LongAlpaca-70B>
- Data
 - ✓ <https://huggingface.co/datasets/Yukang/LongAlpaca-12k>

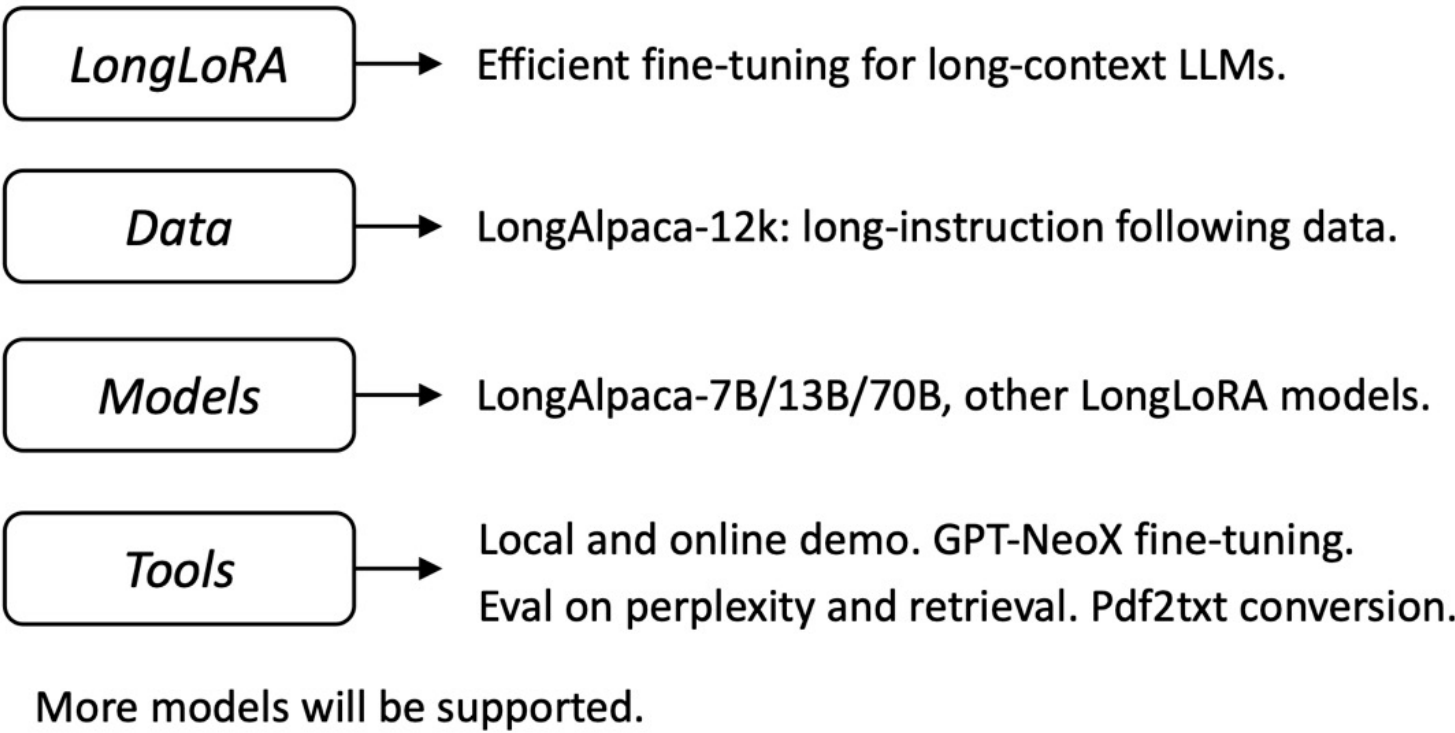
 Yukang/LongAlpaca-13B
📄 Text Generation • Updated 2 minutes ago • ⬇️ 19 • ❤️ 1

 Yukang/LongAlpaca-7B
📄 Text Generation • Updated 2 minutes ago • ❤️ 2

 Yukang/LongAlpaca-70B
📄 Text Generation • Updated 2 minutes ago • ❤️ 1



Long Alpaca



LongLoRA and LongAlpaca for Long-context LLMs [↗](#)

[Gradio](#) [Online Demo](#) [Models](#) [Huggingface Models](#) [Data](#) [LongAlpaca 12k](#) [Paper](#) [Arxiv](#)

[Code License](#) [Apache 2.0](#) [Data License](#) [CC By NC 4.0](#) [Weight License](#) [CC By NC 4.0](#)

TABLE OF CONTENTS [↗](#)

1. [News](#)
2. [Highlights](#)
3. [How to contribute](#)
4. [Requirements](#)
5. [Installation and quick guide](#)
6. [LongAlpaca Data](#)
7. [Models](#)
8. [Training](#)
9. [Evaluation](#)
10. [Demo](#)
11. [Data Generation via Pdf2Text](#)
12. [Examples](#)



Authors



Yukang Chen
CUHK



Shengju Qian
CUHK



Haotian Tang
MIT



Xin Lai
CUHK



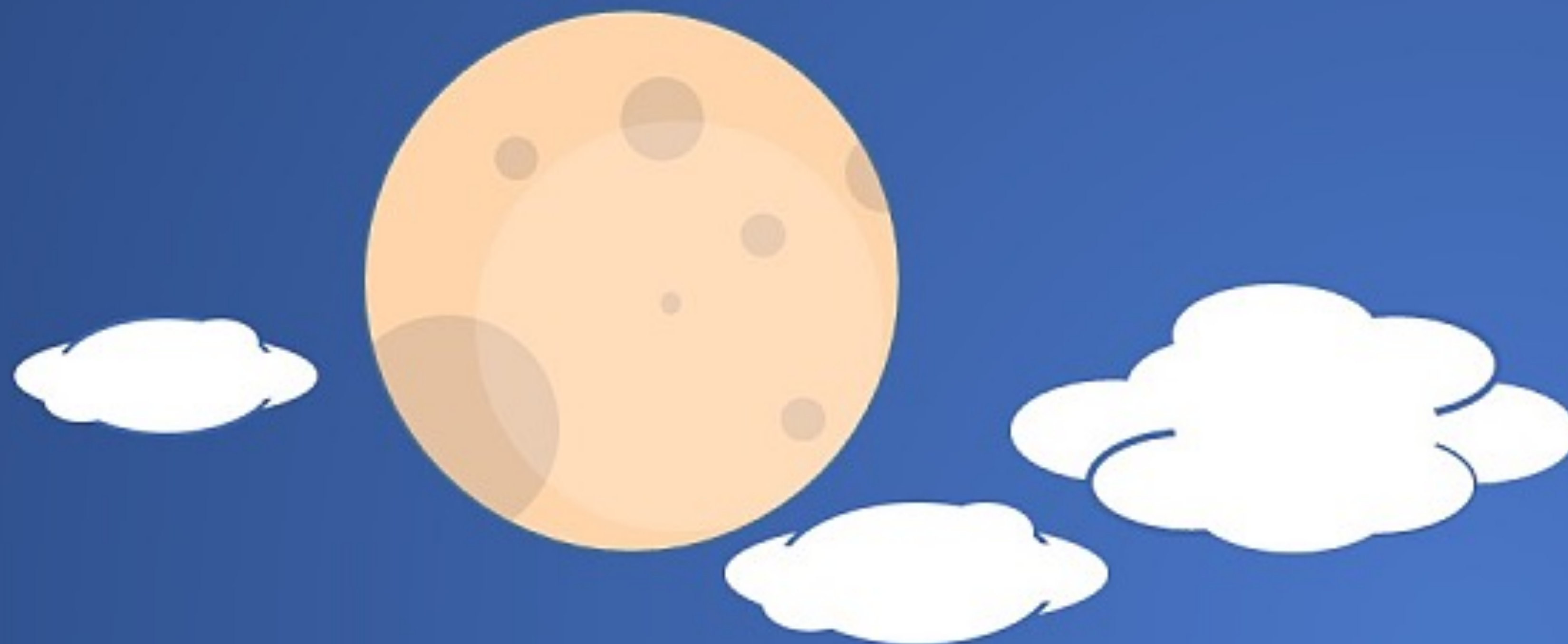
Zhijian Liu
MIT



Song Han
MIT & NVIDIA



Jiaya Jia
CUHK



Thanks