Flora: Effortless Context Construction to Arbitrary Length and Scale

Tianxiang Chen[1,2], **Zhentao Tan**[4], **Xiaofan Bo**[3], **Yue Wu**[4], **Tao Gong**[1,2], **Qi Chu**[1,2*], **Jieping Ye**[4]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]Zhejiang Lab
[4]Independent Researcher
txchen@mail.ustc.edu.cn, {zhentaotan5, b1829544644}@gmail.com, {matthew.wy, yejieping.ye}@alibaba-inc.com, {tgong, qchu}@ustc.edu.cn

Abstract

Effectively handling long contexts is challenging for Large Language Models (LLMs) due to the rarity of long texts, high computational demands, and substantial forgetting of short-context abilities. Recent approaches have attempted to construct long contexts for instruction tuning, but these methods often require LLMs or human interventions, which are both costly and limited in length and diversity. Also, the drop in short-context performances of present long-context LLMs remains significant. In this paper, we introduce Flora, an effortless (human/LLM-free) long-context construction strategy. Flora can markedly enhance the long-context performance of LLMs by arbitrarily assembling short instructions based on categories and instructing LLMs to generate responses based on long-context meta-instructions. This enables Flora to produce contexts of arbitrary length and scale with rich diversity, while only slightly compromising short-context performance. Experiments on Llama3-8B-Instruct and QwQ-32B show that LLMs enhanced by Flora excel in three long-context benchmarks while maintaining strong performances in short-context tasks.

Code — https://github.com/txchen-USTC/Flora

Introduction

Large language models (LLMs) are widely used in many natural language processing tasks (Chen et al. 2024b). These tasks often require dealing with lengthy text inputs (Bai et al. 2023, 2024b), such as long conversation histories (Zhong et al. 2024) or long documents (Bai et al. 2024b). Thus, improving LLMs to handle long-contexts is critical.

There are two categories of approaches to expand LLM's context window. The first focuses on modifying the LLM structure, such as altering the positional encoding (Chen et al. 2023a; Ding et al. 2024) or the attention mechanism (Peng et al. 2023a; Munkhdalai, Faruqui, and Gopal 2024; Gu and Dao 2023), and these are referred to as model-level methods. Nonetheless, these methods typically aim to overcome limitations related to models, hardware, etc., in the pursuit of understanding long texts, such as contexts with over 1 million tokens. In addition, data quality and diversity also matter (Chen et al. 2023b) in improving the long-context modeling capabilities of LLMs. Therefore, the second category, data-level methods, focuses on enhancing LLM's long-context capability via long data-constructing aspects (Chen et al. 2024a; Tang et al. 2024; An et al. 2024; Zhang et al. 2024). However, as displayed in Table 1, most of the current data-level methods necessitate human or LLM involvement in data generation. This approach is expensive and constrained, typically involving datasets of less than 20k samples, each under 10k tokens long. Moreover, enhancing long-context performance in LLMs often markedly compromises their short-context capabilities, even though nearly half of their training data is short (Chen et al. 2023b).

To overcome the challenges, an intuitive way is to construct long-context data by stacking short-context data for the following reasons: 1) Compared to the insufficient long-context data, high-quality supervised fine-tuning (SFT)

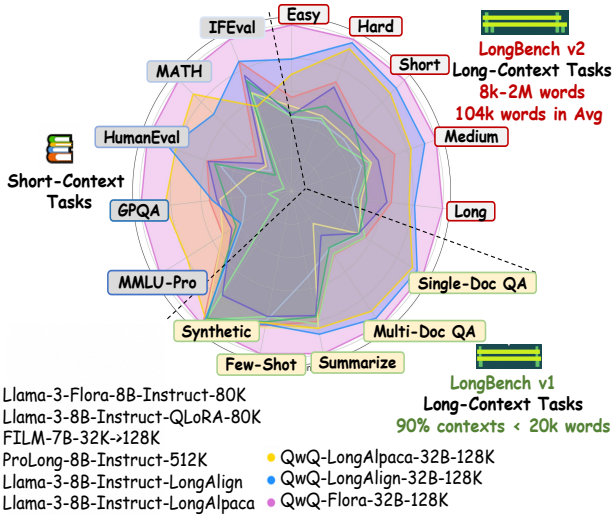

Figure 1: Average scores across long and short context tasks, normalized by the highest score on each task. The scores on LongBench v2 (Bai et al. 2024b) are evaluated in the zero-shot + CoT setting. Our Flora-enhanced models achieve state-of-the-art (SOTA) performances on all tasks, compared to other models of similar parameter scales.

*Qi Chu is the corresponding author.

Challenges	LC Construction Method	Concatenation Method	Flora (Ours)
LLM/Human-free	×	✓	✓
Low Demand for LC	×	✓	✓
Length & Diversity	×	✓	✓
Maintain SC Ability	×	✓	✓
Length Control	✓	×	✓
LC-specific	✓	×	✓

Table 1: Our Flora addresses the six major challenges of current long-context construction and concatenation strategies. Flora is LLM/human-free, does not require massive long contexts, offers diverse and infinitely lengthy data, preserves LLM's short context abilities, control input and output length and is tailored for long context tasks. "SC" stands for "short context," and "LC" stands for "long context."

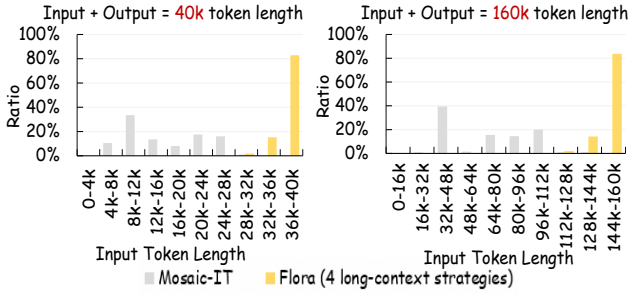

Figure 2: Comparison of output token length distributions between Flora-enhanced and Mosaic-IT enhanced data under fixed total token lengths. The x-axis shows the output token length, and the y-axis shows the ratio.

datasets of millions of samples are readily available (Lambert et al. 2024). 2) Existing pre-trained LLMs are not confused by the concatenated data due to the widespread use of data concatenation technologies in their training stage (Wolf et al. 2020; DeepSeek-AI et al. 2024). 3) These data naturally ensure the short-context capabilities of fine-tuned models. What's more, Mosaic-IT (Li et al. 2024a) has proven the potential of data concatenation in boosting the instruction-following capabilities of LLMs while simultaneously accelerating the training process. However, it cannot regulate the length of inputs and outputs, with outputs generally much longer than the inputs, as shown in Fig. 2. This approach produces suboptimal samples for long-context scenarios and lacks specific designs for enhancing key long-context abilities such as multi-document retrieval and QA.

To take a further step, we propose to concatenate short instructions and their responses as a "long context", and design corresponding instructions to enable the model to answer questions based on a full understanding of this "long context". Following this principle, we introduce Flora, an effortless data construction strategy that can generate instruction-tuning data of any length and scale. Specifically, after obtaining the concatenated long context, we design four instruction templates to simulate common long context understanding tasks such as multi-document question answering and summarization. Furthermore, the responses for the synthesized data are based on the original short context data, eliminating the need for human or LLM intervention. It allows for highly diverse and theoretically limitless-length contexts from existing short instruction tuning datasets with minimal impact on short-context performance and can also flexibly control the input and output length. We curated Flora-80k and Flora-128k (max 80k/128k tokens), and fine-tuned only 8.5% parameters of Llama-3-8B-Instruct and 5.5% of QwQ-32B to develop Flora-enhanced models. As displayed in Fig. 1, our Flora-enhanced models achieve SOTA on long-context benchmarks and excel in short-context tasks compared to similar-sized models.

Related Works

Model-level Long-Context LLMs

Model-level methods mainly adjust model structures. Some of these methods highlight improvements in position encoding design, which enhance the representation of positional information within models (Chen et al. 2023a; Liu et al. 2023b; Peng et al. 2023b; Ding et al. 2024). Additionally, (Bai et al. 2024a) optimizes batching strategies for efficient data processing. Parameter-efficient training focuses on reducing computational resources while maintaining performance (Chen et al. 2023b). There are also methods to innovate new model architectures by modifying attention mechanisms (Peng et al. 2023a; Dai 2019; Munkhdalai, Faruqui, and Gopal 2024; Gu and Dao 2023). However, model-level approaches usually train LLMs using target-length texts, but it's relatively rare to find extremely long-context training data. Therefore, how to construct high-quality long context data is important. Another development direction is the advancement of data-level long-context LLMs.

Data-level Long-Context LLMs

Data-level long-context LLMs rely on constructing long-context data. Current data construction methods are costly, often requiring LLMs or human labor. For instance, to tackle the "lost in the middle" (Liu et al. 2023a) issue, Prolong (Chen et al. 2024a) requires OPT-350m (Zhang et al. 2022) to create its large-scale long-context pre-training dataset. LOGO (Tang et al. 2024) requires prompting Qwen2-70B-Instruct (Yang et al. 2024) to generate questions for each data instance in its 0.3B token dataset. (Xiong et al. 2023) needs to prompt LLAMA 2 CHAT to generate synthetic self-instruct (Wang et al. 2022) long data. LongAlign (Bai et al. 2024a) is constructed via prompting Claude 2.1. FILM-7B (An et al. 2024) leverages a long-context QA dataset synthesized by GPT-4-Turbo (Achiam et al. 2023). Llama-3-8B-Instruct-QLoRA-80K (Zhang et al. 2024) prompts GPT-4 (Achiam et al. 2023) to synthesize its 3.5K long-context instruction tuning data. Prolong-8B-Instruct-512k (Gao et al. 2024) goes through continual pretraining on 40B token data from Llama3 to obtain long-context abilities. However, these resulting datasets are limited in length and diversity, and the resulting long-context LLMs generally suffer from severe drops in short-context abilities, despite many short contexts having been contained.

Unlike previous data-level methods, Flora is the first to eliminate human and LLM intervention in long-context

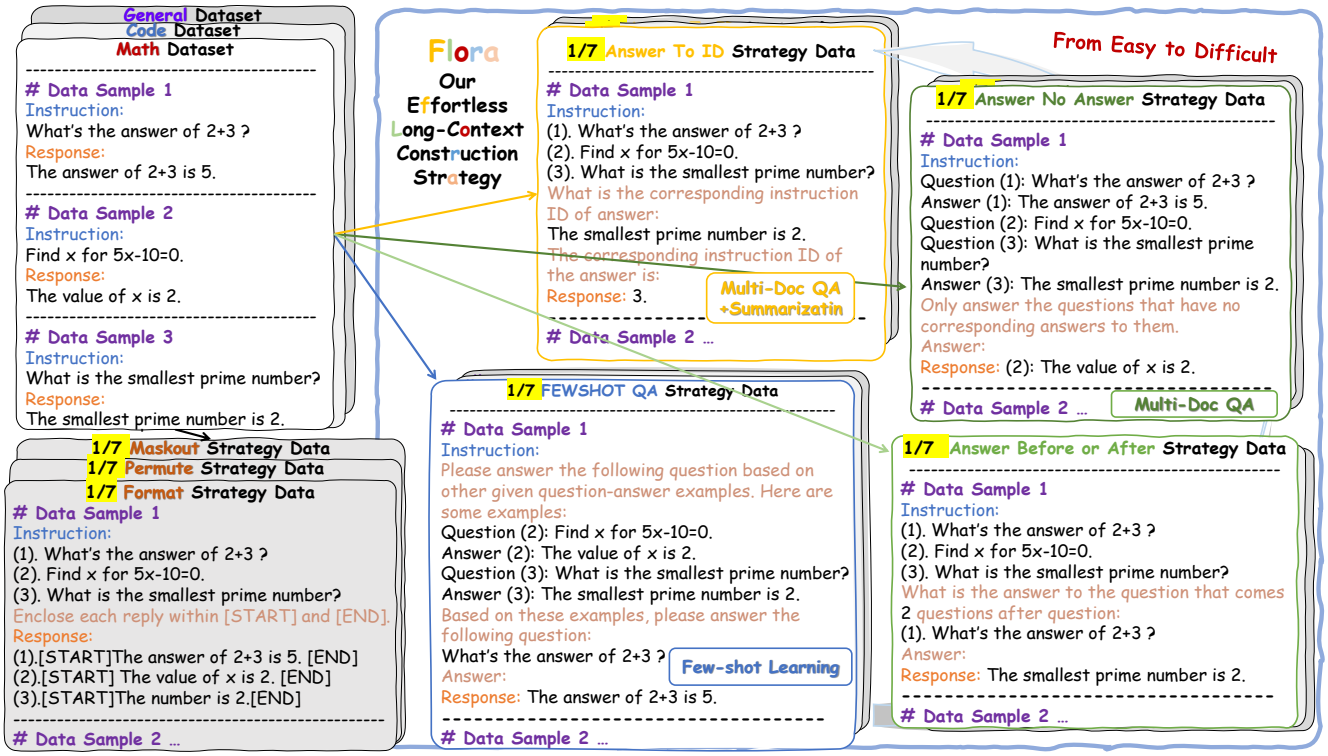

Figure 3: Illustration of Flora, our effortless long-context construction strategy that generates theoretically infinite long-context data without human or LLM intervention. It first categorizes short instruction-tuning datasets into three domains: math, coding, and general knowledge, and then applies targeted augmentations per domain to enhance LLMs' long-context capabilities.

data construction, enabling theoretically infinite-length contexts with rich diversity while preserving short-context abilities. These unique characteristics of Flora distinguish it from concatenation and data engineering methods, such as LifeLongICL (Xu, Ye, and Ren 2024) and Data Engineering (Fu et al. 2024). LifeLongICL evaluates LLMs' ability to retrieve relevant examples from concatenated few-shot demonstrations for answering new questions. This 'task haystack' approach focuses on assessing task retrieval skills of LLMs. Data Engineering holds that up-sampling long sequences while retaining the domain mixture of the pretraining corpora is crucial for context scaling during pretraining.

Methodology

Preliminaries

Concatenation Method: To illustrate data concatenation methods, we reference Mosaic Instruction Tuning (Mosaic-IT) (Li et al. 2024a), a technique designed to enhance LLMs' **instruction-following** capabilities. Mosaic-IT combines multiple short instruction inputs with a meta-instruction to form the input. The output is a meta-instruction-guided concatenation of selected short instruction outputs.

Training Objective: We first express the objective function of ordinary SFT to lay the groundwork for the introduction of our strategy. Given an SFT dataset D with n data samples, we can divide each sample into a triplet: $(\mathit{Instruction}, \mathit{Input}, \mathit{Response})$. For simplicity, we define $x = (\mathit{Instruction}, \mathit{Input})$ as the unified instruction, and y as the corresponding *Response*. Let $p_\theta(\cdot)$ denote the LLM with parameters θ that we aim to train. p_θ is fine-tuned by maximizing the objective function $\max_\theta \sum_{i=1}^{n} \sum_{j=1}^{l_i} \log p_\theta(y_{i,j}|x_i, y_{i,<j})$ over all N samples given as (x_i, y_i). Here, $y_{i,j}$ represents the j-th token of the response y_i, $y_{i,<j}$ denotes the sequence of tokens preceding $y_{i,j}$, and l_i indicates the token length of y_i.

Effortless Long-Context Construction

Our effortless long-context construction strategy, Flora, incorporates seven data augmentations along with a token length distribution rule, as illustrated in Fig. 3. Each of the seven augmentations contributes 1/7. Three of them are from Mosaic-IT to enhance the instruction following abilities, and the other four are proposed by us to enhance the three most critical long-context capabilities (Bai et al. 2023) of LLMs: multi-document retrieval, few-shot learning and summarization. Other long-context capabilities can be transferred from the three basic ones. The four long-context strategies, organized from easy to difficult in terms of learning difficulty, include: the Fewshot QA (FQA) strategy to boost few-shot learning skills; the Answer Before or After (ABA) and Answer No Answer (ANA) strategies to improve multi-document retrieval proficiency; and the Answer to ID (AID) strategy to enhance both multi-document retrieval and summarization capabilities. We further explore a token length

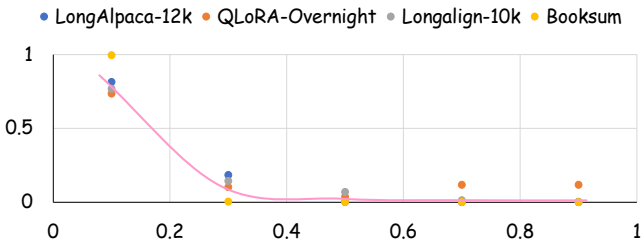

Figure 4: We show that the token length distribution of present public long context instruction tuning datasets can be fitted as a curve: $y = 2.411e^{-10.899x} + 0.017$, where x-axis measures the normalized token length ranges from 0 to 1, and y-axis measures the data sample proportion.

distribution rule of long-context datasets that can bring better long-context performance gains as a takeaway for constructing long-context datasets.

Fewshot QA (FQA) Strategy Few-shot learning ability is crucial for long-context learning since it boosts the LLM's comprehension and reasoning capabilities. Therefore, we design a Fewshot QA (FQA) strategy to enhance LLM's few-shot learning ability. In the FQA strategy, some arbitrary instructions are selected in the meta-instructions as few-shot examples and require LLMs to answer new instructions based on given examples. Thus the objective function can be expressed as: $\max_{\theta} \sum_{j=1}^{l} \log p_\theta([y_1^{'}, ..., y_\beta^{'}]_j | [(x_1, y_1)...(x_k, y_k), x_1^{'}, ..., x_\beta^{'}, I_f, I_{fqa}], [y_1^{'}, ..., y_\beta^{'}]_{<j})$, where k is the count of instructions as few-shot examples that have corresponding responses, $x_1^{'}, ..., x_\beta^{'}$ are the β instructions we instruct the LLM to answer, $y_1^{'}, ..., y_\beta^{'}$ are the corresponding responses, I_{fqa} is the meta-instruction of FQA strategy.

Answer Before or After (ABA) Strategy Designed to improve multi-document retrieval, this approach helps LLMs determine the relevance of information based on its position relative to other content, whether it appears before or after key context. Specifically, we instruct LLMs to answer questions that come n_i before or after question $x_i^{'}$, where there are β questions that need to be answered. Here, $n_i \in \{n_1, ..., n_\beta\}$, and $x_i^{'} \in \{x_i^{'}, ..., x_\beta^{'}\}$ is the arbitrarily selected question. For example, this could involve asking the LLMs to answer the question that comes 2 questions after the question "What's the answer to 2+3?". Thus the objective function can be formulated as: $\max_{\theta} \sum_{j=1}^{l} \log p_\theta([y_1^{'}, ..., y_{aba}^{'}]_j | [x_1, ..., x_k, x_i^{'}, ..., x_\beta^{'}, n_1, ..., n_\beta, I_f, I_{aba}], [y_1^{'}, ..., y_\beta^{'}]_{<j})$, where I_{aba} is the meta-instruction of ABA strategy.

Answer No Answer (ANA) Strategy Also focused on multi-document retrieval, this strategy will arbitrarily concatenate multiple questions and 4/5 of their answers into one instruction and instruct LLMs to only answer the 1/5 questions without corresponding answers. In this way, the LLM can learn to recognize whether the key information is presented in the documents. We define $x_1^{'}, ..., x_\beta^{'}$ as the β concatenated instructions without corresponding answers $y_1^{'}, ..., y_\beta^{'}$. The objective function of ANA can be formulated as: $\max_{\theta} \sum_{j=1}^{l} \log p_\theta([y_1^{'}, ..., y_\beta^{'}]_j | [(x_1, y_1)...(x_k, y_k), x_1^{'}, ..., x_\beta^{'}, I_f, I_{ana}], [y_1^{'}, ..., y_\beta^{'}]_{<j})$, where I_{ana} is the meta-instruction of ANA strategy, k is the number of instructions with concatenated answers.

Datasets	Domain	Sample Num	Max Token	Avg Token
LongAlpaca-12k	1	12k	70k	9.4k
LongAlign-10k	6	10k	86k	16.9k
QLoRA-Overnight	7	20k	80k	14k
Flora-80k (Ours)	15	93k	80k	9.5k
Flora-128k (Ours)	15	60k	128k	14.8k

Table 2: Comparison with other widely used long-context datasets in terms of domain (e.g., math, code, science, etc.), sample number, and token lengths. The token length is calculated by Llama-3 tokenizer.

Answer to ID (AID) Strategy This strategy aims to improve multi-document retrieval and summarization abilities. It involves randomly concatenating several questions into a single instruction, excluding their answers. The LLM is then tasked with identifying the question IDs given arbitrary answers to these questions. This approach requires the LLM to comprehend and summarize the answer, as well as retrieve the relevant question from among the multiple concatenated questions. We define k as the number of concatenated questions in an augmented instruction. $y_1, ..., y_\beta$ are the β answers given to the LLM to find the corresponding question IDs Let $y_1, ..., y_\beta$ represent the β answers provided to the LLM to determine the corresponding question IDs $y_1^{'}, ..., y_\beta^{'}$. Thus the objective function of AID can be formulated as: $\max_{\theta} \sum_{j=1}^{l} \log p_\theta([y_1^{'}, ..., y_\beta^{'}]_j | [x_1, ...x_k], [y_1, ...y_\beta, I_f, I_{aid}], [y_1^{'}, ..., y_\beta^{'}]_{<j})$, where I_{aid} is the meta-instruction of AID strategy.

Long-Context Dataset Construction We analyze some most widely used long-context SFT datasets (LongAlpaca-12k (Chen et al. 2023b), LongAlign-10k (Bai et al. 2024a), Booksum (Kryściński et al. 2021) and QLoRA-Overnight (Zhang et al. 2024)) and find that their ideal token length distribution follows $y = 2.411e^{-10.899x} + 0.017$, as shown in Fig. 4, where x is the normalized token length (0 to 1) and y is the data sample proportion. We follow this ideal distribution to construct our dataset.

We adopt the BUAA Infinity Instruct dataset, which has around 1.5M entries composed of extensive short-length SFT data samples collected by BAAI and enhance it with our Flora strategy to be the final long-context datasets. The token length distribution adheres to our functional rule. We categorize the Infinity Instruct dataset into math, code, and general knowledge. Each category is then enhanced with QAF, ABA, ANA, and AID augmentations, plus 3 Mosaic augmentations, before merging the augmented categories.

Models	Difficulty				Length (<32k; 32k-128k; >128k words)						LongBench v2 Avg.		SQA	MQA	Summ	FS	Syn	LongBench v1 Avg.
	Easy		Hard		Short		Medium		Long									
Model-Level Comparison with SOTA Long-Context LLMs																		
GLM-4-9B-Chat-128k	30.7	*34.4*	29.9	*28.6*	33.9	*35.0*	29.8	*30.2*	25.0	*25.0*	30.2	*30.4*	47.30	42.7	23.94	40.61	99	50.71
Llama-3.1-8B-Instruct-128k	30.7	*36.5*	29.6	*26.7*	35.0	*34.4*	27.9	*31.6*	25.9	*21.3*	30.0	*30.4*	47.71	37.46	23.75	42.68	98.5	50.02
Qwen2.5-7B-Instruct	29.2	*30.7*	25.7	*29.3*	36.1	*35.6*	23.7	*26.5*	18.5	*26.9*	27.0	*29.8*	41.00	37.74	22.54	43.2	84.84	45.86
Llama-3-8B-Instruct-262k	35.9	*32.8*	28.3	*26.4*	33.3	*35.6*	31.6	*25.6*	26.9	*24.1*	31.2	*28.8*	35.69	18.02	17.10	40.59	88.5	39.98
Claude-3.5-Sonnet-200k	46.9	*55.2*	37.3	*41.5*	46.1	*53.9*	38.6	*41.9*	37.0	*44.4*	41.0	*46.7*	-	-	-	-	-	-
Qwen3-235B-A22B-128k ♣	47.4	*56.4*	36.0	*46.2*	45.6	*58.3*	36.7	*44.1*	38.9	*48.6*	40.4	*50.1*	-	-	-	-	-	-
Data-Level Comparison with Long-Context Datasets																		
Mistral-FILM-7B	25.7	*27.7*	18.3	*21.5*	23.9	*24.6*	20.5	*25.0*	17.8	*20.4*	21.1	*23.9*	48.46	38.95	24	39.70	95	49.32
Llama-3.1-8B-Instruct-SEALONG	29.7	*37.0*	26.4	*27.3*	32.8	*36.7*	24.2	*28.8*	25.9	*25.9*	27.6	*31.0*	48.25	37.83	24.38	42.5	98.5	50.29
Prolong-8B-Instruct-512k (CPT)	30.9	*31.6*	25.2	*22.2*	32.4	*28.7*	27.5	*25.5*	18.6	*21.4*	27.3	*25.8*	44.44	20.03	24.76	43.24	95.75	45.64
Llama-3-8B-Instruct-QLoRA-80k	32.3	*27.6*	25.1	*20.6*	30.0	*27.2*	29.8	*22.3*	20.4	*18.5*	27.8	*23.3*	45.49	35.33	15.40	40.92	94.5	46.33
Llama-3-8B-Instruct-LifeLongICL	29.9	*25.9*	26.6	*29.4*	34.9	*36.7*	25.4	*23.6*	20.6	*22.1*	27.8	*28.1*	37.28	23.04	13.68	38.4	98	42.08
Llama-3-8B-Instruct-8k	0	0	0	0	0	0	0	0	0	0	0	0	40.59	28.34	14.05	32.67	79	38.93
+ LongAlpaca-12k	29.1	*29.2*	28.7	*25.1*	36.1	*32.8*	26.1	*23.3*	22.2	*23.1*	28.9	*26.6*	44.58	34.98	22.7	41.30	94.25	47.56
+ LongAlign-10k	29.5	*28.3*	29.1	*30.7*	32.7	*32.5*	26.7	*27.7*	28.7	*29.3*	29.3	*29.8*	44.38	27.36	22.5	39.65	77.13	42.24
+ Flora-80k (Ours)	33.3	*34.9*	33.4	*32.2*	45.0	*35.0*	30.7	*33.0*	19.4	*30.6*	33.4	*33.2*	48.68	39.36	23.6	42.51	99.5	50.73
QwQ-32B-128k ♣	-	*50.6*	-	*42.6*	-	*48.8*	-	*45.2*	-	*40.8*	-	*45.6*	83.11	77.54	26.78	44.95	98.75	66.23
+ LongAlpaca-12k ♣	-	*43.5*	-	*42.0*	-	*51.5*	-	*38.2*	-	*36.1*	-	*42.5*	80.05	74.26	24.82	41.66	95.5	63.26
+ LongAlign-10k ♣	-	*49.1*	-	*43.8*	-	*54.5*	-	*42.6*	-	*37.8*	-	*45.8*	83.21	78.56	25.87	42.15	98	65.56
+ Flora-128k (Ours) ♣	-	***61.8***	-	***44.8***	-	***58.8***	-	***46.2***	-	***46.4***	-	***50.5***	**87.12**	**83.06**	**29.38**	**51.39**	**99.5**	**70.07**

Table 3: Results (%) on LongBench v2 and v1: CoT prompting results in LongBench v2 are highlighted with italic. Bold numbers indicate the highest values per column. InternLM2-7B-LongWanjuan (Lv et al. 2024), marked with *, is closed-source, and only its official LongBench v1 results are reported. Llama-3-8B-Instruct-LifeLongICL denotes the model trained by Flora-enhanced LifeLongICL data. Reasoning models are indicated by ♣. QwQ-32B-128k results are self-evaluated and limited to its reasoning mode.

To better preserve short-context capabilities, we also incorporate original short data samples (those under 2k tokens) from the Infinity Instruct to replace the augmented samples under 2k tokens.

We compare our datasets with some other widely used public long-context datasets in Table 2, where our curated datasets cover more domains and is flexible in scale and length.We have omitted the thinking mode token during curating the dataset for QwQ, therefore training on QwQ can be regarded as an instruction tuning process. The model after SFT can still generate the thinking part.

Experimental Setup

SFT Details

We use QLoRA (Dettmers et al. 2024) to efficiently fine-tune Llama-3-8B-Instruct as our main model based on Llama-Factory. We apply LoRA on all Q, K, V, and O projections and additionally train the embedding layer. The LoRA rank is set to 256, with an alpha value of 128 and 4-bit quantization. The learning rate is set to 5e-5, with a linear decay schedule and no warm-up steps. The batch size is 8, and gradient checkpointing is enabled to optimize memory usage. The model is trained for one epoch with 8.5% trainable parameters for Llama3-8B and with 5.5% trainable parameters for QwQ-32B on $4 \times 8A100(80G)$ machines using DeepSpeed v2 offloading within a day. For Llama3-8B, we expand the RoPE base to 200M and increase max position embeddings to 81,920. For QwQ-32B, we keep the original RoPE base and increase max position embeddings to 131,072. The β is set to 1 for FQA, ABA and AID, and set to 20% of the concatenated instructions for ANA.

Evaluation Details

Compared Methods To prove the effectiveness of our strategy, we compare our datasets with other long-context datasets with similar training settings. To showcase the long-context capabilities of our final models, we evaluate them against other prevalent long-context LLMs with comparable parameters. Specifically, we fine-tune Llama-3-8B-Instruct and QwQ-32B (Team 2025) on long-text datasets: LongAlpaca-12k (Chen et al. 2023b), LongAlign-10k (Bai et al. 2024a) and our Flora dataset with maximum token lengths of 80k and 128k. This demonstrates the scalability of our strategy across model parameters and token lengths. Other compared data-level LLMs include Llama-3-8B-Instruct-QLoRA-80k (Zhang et al. 2024), Prolong-8B-Instruct-512k (Gao et al. 2024), InternLM2-7B-LongWanjuan (Lv et al. 2024), Mistral-FILM-7B (An et al. 2024), Llama-3.1-8B-Instruct-SEALONG (Li et al. 2024b) and our self-trained Llama-3-8B-Instruct on the data of LifeLongICL (Xu, Ye, and Ren 2024) enhanced by our Flora. For model comparison, we compare our model with other long-context LLMs trained to handle context windows $\geq$ 32k tokens (GLM-4-9B-Chat-128k (GLM et al. 2024), Qwen-2.5-7B-Instruct (Team 2024), Llama-3.1-8B-Instruct-128k (Dubey et al. 2024), ChatGLM3-6B-32k (GLM et al. 2024), Llama-3-8B-Instruct-262k, Claude-3.5-Sonnet, and Qwen3-235B-A22B-128k.

Long-context Benchmarks We adopt three benchmarks, LongBench v1 (Bai et al. 2023), LongBench v2 (Bai et al. 2024b) and Needle-In-A-HayStack (NIAH), to evaluate the long-context understanding ability of various LLMs. LongBench v1 is widely used to evaluate LLMs in handling inputs mostly below 20k words. We take 11 representative tasks from it to form 5 task types, including single-document question answering (SQA), multi-document question an-

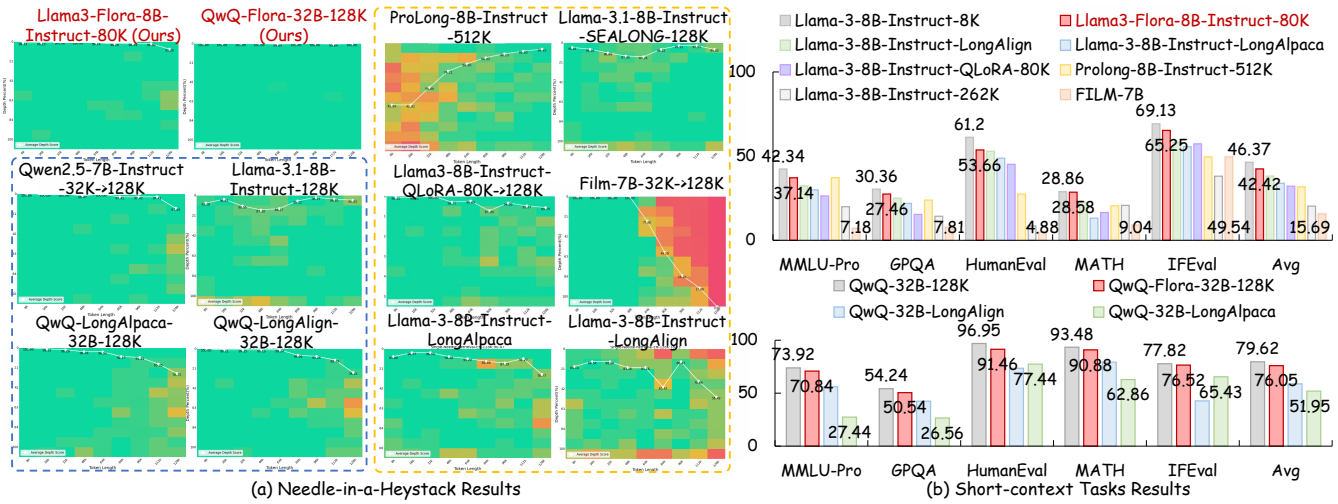

Figure 5: (a) Results of Single-Retrieval task in Needle-In-A-Haystack benchmark. The x-axis represents the context lengths, while the y-axis indicates the depth of the inserted needle. The green color signifies a score close to 1, and red denotes a score close to 0. (b) Results on five short-context tasks from the Open LLM Leaderboard 2.

swering (MQA), summarization (Summ), few-shot learning (FS) and synthetic tasks (Syn).

However, significant advancements in long-context LLMs have increased context window lengths significantly from 8k to 128k and even up to 1M tokens, making LongBench v1 inadequate to evaluate LLMs capable of handling more than 20k words. To address this, LongBench v2 and Needle-In-A-HayStack benchmarks are adopted. LongBench v2 includes 503 challenging multiple-choice questions across contexts ranging from 8k to 2 million words, with most 128k words. It focuses on deep understanding and reasoning across real-world multitasks and offers a more comprehensive and challenging assessment than v1. It can also serve as an OOD benchmark, as it is designed to assess (1) long-context reasoning and (2) the understanding of long structured data (including reasoning over lengthy tables and knowledge graphs)-task categories that are not emphasized by our concatenation strategy. Needle-In-A-HayStack challenges LLMs to recall irrelevant information inserted into a lengthy context and is assessed by GPT-3.5. We test the single needle retrieval tasks. For models with original maximum position embeddings below 128k, we extend their embeddings to 128k for these evaluations.

Short-context Benchmarks We select 5 tasks from Open LLM Leaderboard v2 to evaluate the short-context abilities of models, including **MMLU-Pro** (Wang et al. 2024), **GPQA** (Rein et al. 2023),**IFEval** (Zhou et al. 2023), **HumanEval** (Chen et al. 2021) and **MATH** (Hendrycks et al. 2021). The first three are employed to test the general knowledge abilities, while **MATH** (Hendrycks et al. 2021) and **HumanEval** (Chen et al. 2021) are used to test math and coding abilities, respectively. We employ the OpenCompass library to evaluate all these tasks.

Experiments and Analysis

Long Context Results

We evaluate our Flora-enhanced LLMs against a series of the latest long-context LLMs, including some prevalent LLMs and data-level LLMs to which our model belongs. The results on LongBench v2 and v1 are detailed in Table 3, where our model achieves SOTA performances on average on both benchmarks. Since the overall test length of LongBench v1 mainly covers the short length parts of LongBench v2, LongBench v2 can better reflect the long-context understanding and reasoning abilities of different LLMs. Notably, Llama-3-8B-Instruct-8k struggles with these demanding long-context challenges and outputs null results, resulting in zero scores on LongBench v2. Compared to other data-level models, Llama3-Flora-8B-Instruct-80k demonstrates significant superiority in handling hard questions and those below 128k words without CoT prompts. With the use of CoT, our model continues to markedly outperform others, even showing an enhanced ability to comprehend and reason through particularly long contexts (beyond 128k words).

The results on NIAH are visualized in Fig. 5 (a), where our models nearly achieves a 100% retrieval accuracy across all context lengths and the Llama3 model shows excellent generalization to new positions (80k-128k).

Short Context Results

We also test several long-context LLMs on five short-context tasks from the Open LLM Leaderboard V2 in Fig. 5 (b). All Llama3 and QwQ series models perform worse than baselines, suggesting that extending the context window compromises the model's ability to handle short contexts. This observation aligns with previous researches (Peng et al. 2023b; Zhang et al. 2024). Notably, the performance decline of other methods on short-context tasks is not due to a lack of short-context data. For instance, LongAlpaca-12k includes

Aug. (Norm.)	SQA			MQA			Summ			FS			Syn			v1	SC
	NQA	MQA-zh	Avg.	2Wiki	MSQ	Avg.	Gov	VCS	Avg.	SAM	LSHT	Avg.	PR-en	PR-zh	Avg.	Avg.	Avg.
Base (Mosaic-IT)	25.71	55.28	40.50	45.03	27.83	36.43	23.98	13.19	18.58	37.42	39.5	38.46	94.5	93	93.75	45.54	42.76
3/4 Base+1/4 FQA	27.09	55.16	41.13	45.81	31.05	38.43	25.88	12.22	19.05	**41.80**	42	**41.9**	96	97	96.5	47.40	42.89
3/4 Base+1/4 ABA	26.4	55.04	40.72	49.73	**33.56**	**41.65**	25.26	13.87	19.57	39.97	40	39.98	97.5	98	97.75	47.93	43.03
3/4 Base+1/4 ANA	27	55.23	41.12	49.48	32.88	41.18	26.04	14.38	20.21	38.14	41.5	39.82	98.5	97	97.75	47.97	42.87
3/4 Base+1/4 ATID	**28.41**	53.05	40.73	49.15	31.67	40.41	**27.85**	15.74	21.80	39.43	41.5	40.46	99	98	98.5	48.38	42.94
FQA+ABA+ANA+ATID	27.52	54.87	41.20	48.63	29.8	39.22	27.09	15.99	21.54	39.58	41	40.29	99	**98.5**	98.75	48.20	42.91
3/7 Base+4/7 All	28.10	55.88	41.99	49.51	32.06	40.79	28.01	16.18	22.10	40.66	42	41.33	99	98.5	98.75	48.99	43.02
3/7 Base+4/7 All †(Ours)	28.24	**55.93**	**42.08**	**49.76**	31.91	40.84	27.82	**16.37**	**22.10**	41.29	**42**	41.65	**99.5**	98	**98.75**	**49.08**	**43.41**

Table 4: Ablation studies on different long-context augmentations on LongBench v1 and short context (SC) tasks. All the augmented datasets contain 50k data samples, and the token lengths follow a normal distribution for fair comparisons. † means the augmented samples below 2k tokens are replaced by the original SFT samples below 2k tokens.

Model	Difficulty		Length			LongBench v2 Avg.	SQA	MQA	Summ	FS	Syn	LongBench v1 Avg.	Short Context Tasks Avg.
	Easy	Hard	Short	Medium	Long								
Baseline (Llama-3-8B-Instruct-8k)	0	0	0	0	0	0	40.59	28.34	14.05	32.67	79	38.93	46.37
Normal Distribution	**33.5**	28.7	36.6	27.7	25.8	30.5	44.77	38.01	21.14	39.86	98.25	48.40	43.56
$y = 0.2$ (Even)	30.7	29.9	**37.8**	29.8	18.5	30.2	44.54	38.80	21.01	39.94	97	48.26	43.83
$y = 2.375(x-0.5)^2 + 0.01$ (U-Shaped)	32.8	28.0	36.7	28.4	21.3	29.8	41.73	**40.33**	20.03	38.84	98.5	47.89	44.18
$y = 2.411e^{10.899(x-1)} + 0.017$ (Reverse)	31.8	26.9	33.3	28.2	22.2	28.7	43.90	38.30	20.24	40.13	97.25	47.96	40.65
$y = 2.411e^{-10.899x} + 0.017$ (Ours)	33.0	**31.8**	37.4	**30.2**	**27.8**	**32.3**	**45.05**	38.69	**21.69**	**41.95**	**98.75**	**49.23**	**44.82**

Table 5: Ablation studies on different token length distribution on LongBench v2, v1 and short context tasks. All experiments use 8k samples, the same number as the dataset of the reverse distribution, for fair comparison.

a significant amount of such data yet still underperforms. For LongAlign, its original training setting is different from ours: it first conducts continual pretraining over 10B tokens (with sequence lengths up to 64k), followed by instruction tuning on its LongAlign-10k with long and short context data. As our strategy focuses on the SFT stage, we ensure a fair comparison by evaluating directly against its SFT dataset LongAlign-10k.

However, our model boasts a significantly smaller performance decline compared to other similarly scaled long-context LLMs, with only a 3-4% average drop versus at least a 8% average drop for other models. It excels in retaining basic knowledge, coding, math, and instruction-following abilities. This is because our strategy involves splitting the original instruction tuning dataset into three categories before concatenating samples with diverse meta-instructions, while keeping samples under 2k tokens as original instruction tuning dataset samples. By concatenating samples of the same category, we preserve domain-specific knowledge. The arbitrary sample concatenation, along with the use of varied meta-instructions, helps maintain instruction-following skills. Retaining shorter samples as part of the original data also contributes to this preservation.

Ablation Studies

Table 4 presents an ablation study on various long-context augmentation strategies, showcasing the average results across different evaluation criteria on LongBench v1 and short-context tasks. Note that we only selected LongBench v1 for our experiments because it includes tasks designed to assess the corresponding effect of our augmentation strategies. We adopt Mosaic-IT as our baseline and concatenate the maximum data length to 80k tokens for fair comparison. The dataset sample length adheres to a normal distribution as Mosaic-IT and all the augmented datasets contain 50k data samples in Table 4. To evaluate the effectiveness of each augmentation strategy, we incorporate each strategy once, allowing its augmented data to constitute 1/4 of the total samples, while the remaining 3/4 consists of data augmented by Mosaic-IT. The results clearly demonstrate that each of our proposed strategies significantly enhances the corresponding long-context abilities of the original LLM. Implementing all four strategies together leverages their individual strengths, resulting in the best overall performance. Replacing the very short concatenated contexts with the original samples can better maintain short-context abilities.

We also investigate how different token length distributions affect LLM's understanding of long and short contexts. The findings, detailed in Table 5, are based on experiments using datasets of 8k samples, each with a maximum token length of 80k, ensuring a fair comparison. We analyze various distributions: normal, even, U-shaped, reverse, and our optimized rule. The results suggest that more short samples mitigates the decline in short-context understanding, while only a few extremely long samples are needed for excellent long-context capabilities. Our approach yields optimal proficiency in both long and short contexts.

Conclusion

Handling extremely long contexts remains a challenge for LLMs due to the scarcity of such data, high computational demands, and the issue of catastrophic forgetting of short context abilities. Existing methods often involve costly and limited human or model intervention. We introduce Flora, a new approach for constructing long contexts without human or model involvement. Flora assembles short SFT samples into theoretically infinite-length contexts, paired with high-level meta-instructions for training. This method enhances long-context capabilities with minimal impact on short-context abilities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62121002, 62472396), Anhui Provincial Natural Science Foundation (2508085QF212) and the advanced computing resources provided by the Supercomputing Center of the USTC.

References

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

An, S.; Ma, Z.; Lin, Z.; Zheng, N.; and Lou, J.-G. 2024. Make Your LLM Fully Utilize the Context. *arXiv preprint arXiv:2404.16811*.

Bai, Y.; Lv, X.; Zhang, J.; He, Y.; Qi, J.; Hou, L.; Tang, J.; Dong, Y.; and Li, J. 2024a. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Bai, Y.; Tu, S.; Zhang, J.; Peng, H.; Wang, X.; Lv, X.; Cao, S.; Xu, J.; Hou, L.; Dong, Y.; et al. 2024b. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks. *arXiv preprint arXiv:2412.15204*.

Chen, L.; Liu, Z.; He, W.; Li, Y.; Luo, R.; and Yang, M. 2024a. Long Context is Not Long at All: A Prospector of Long-Dependency Data for Large Language Models. *arXiv preprint arXiv:2405.17915*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023a. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Chen, T.; Tan, Z.; Gong, T.; Wu, Y.; Chu, Q.; Liu, B.; Ye, J.; and Yu, N. 2024b. Llama SLayer 8B: Shallow Layers Hold the Key to Knowledge Injection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 5991–6002.

Chen, Y.; Qian, S.; Tang, H.; Lai, X.; Liu, Z.; Han, S.; and Jia, J. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Dai, Z. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Guo, D.; Yang, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Zhang, H.; Ding, H.; Xin, H.; Gao, H.; Li, H.; Qu, H.; Cai, J. L.; Liang, J.; Guo, J.; Ni, J.; Li, J.; Wang, J.; Chen, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Song, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhao, L.; Wang, L.; Zhang, L.; Li, M.; Wang, M.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Tian, N.; Huang, P.; Wang, P.; Zhang, P.; Wang, Q.; Zhu, Q.; Chen, Q.; Du, Q.; Chen, R. J.; Jin, R. L.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Xu, R.; Zhang, R.; Chen, R.; Li, S. S.; Lu, S.; Zhou, S.; Chen, S.; Wu, S.; Ye, S.; Ye, S.; Ma, S.; Wang, S.; Zhou, S.; Yu, S.; Zhou, S.; Pan, S.; Wang, T.; Yun, T.; Pei, T.; Sun, T.; Xiao, W. L.; Zeng, W.; Zhao, W.; An, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Li, X. Q.; Jin, X.; Wang, X.; Bi, X.; Liu, X.; Wang, X.; Shen, X.; Chen, X.; Zhang, X.; Chen, X.; Nie, X.; Sun, X.; Wang, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yu, X.; Song, X.; Shan, X.; Zhou, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhu, Y. X.; Zhang, Y.; Xu, Y.; Xu, Y.; Huang, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Li, Y.; Wang, Y.; Yu, Y.; Zheng, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Tang, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Wu, Y.; Ou, Y.; Zhu, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Zha, Y.; Xiong, Y.; Ma, Y.; Yan, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Wu, Z. F.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Huang, Z.; Zhang, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Gou, Z.; Ma, Z.; Yan, Z.; Shao, Z.; Xu, Z.; Wu, Z.; Zhang, Z.; Li, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Gao, Z.; and Pan, Z. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Ding, Y.; Zhang, L. L.; Zhang, C.; Xu, Y.; Shang, N.; Xu, J.; Yang, F.; and Yang, M. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fu, Y.; Panda, R.; Niu, X.; Yue, X.; Hajishirzi, H.; Kim, Y.; and Peng, H. 2024. Data Engineering for Scaling Language Models to 128K Context. In *International Conference on Machine Learning*, 14125–14134. PMLR.

Gao, T.; Wettig, A.; Yen, H.; and Chen, D. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Kryściński, W.; Rajani, N.; Agarwal, D.; Xiong, C.; and Radev, D. 2021. Booksum: A collection of datasets

for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Lambert, N.; Morrison, J.; Pyatkin, V.; Huang, S.; Ivison, H.; Brahman, F.; Miranda, L. J. V.; Liu, A.; Dziri, N.; Lyu, S.; et al. 2024. T\" ULU 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124*.

Li, M.; Chen, P.; Wang, C.; Zhao, H.; Liang, Y.; Hou, Y.; Liu, F.; and Zhou, T. 2024a. Mosaic IT: Enhancing Instruction Tuning with Data Mosaics. *arXiv preprint arXiv:2405.13326*.

Li, S.; Yang, C.; Cheng, Z.; Liu, L.; Yu, M.; Yang, Y.; and Lam, W. 2024b. Large Language Models Can Self-Improve in Long-context Reasoning. *arXiv preprint arXiv:2411.08147*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Liu, X.; Yan, H.; Zhang, S.; An, C.; Qiu, X.; and Lin, D. 2023b. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*.

Lv, K.; Liu, X.; Guo, Q.; Yan, H.; He, C.; Qiu, X.; and Lin, D. 2024. Longwanjuan: Towards systematic measurement for long text quality. *arXiv preprint arXiv:2402.13583*.

Munkhdalai, T.; Faruqui, M.; and Gopal, S. 2024. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*.

Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Derczynski, L.; et al. 2023a. RWKV: Reinventing RNNs for the Transformer Era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14048–14077.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2023b. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Tang, Z.; Sun, Z.; Li, J.; Zhu, Q.; and Zhang, M. 2024. LOGO–Long cOntext aliGnment via efficient preference Optimization. *arXiv preprint arXiv:2410.18533*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models.

Team, Q. 2025. Qwq-32b: Embracing the power of reinforcement learning. *URL: https://qwenlm. github. io/blog/qwq-32b*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Xiong, W.; Liu, J.; Molybog, I.; Zhang, H.; Bhargava, P.; Hou, R.; Martin, L.; Rungta, R.; Sankararaman, K. A.; Oguz, B.; et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Xu, X.; Ye, Q.; and Ren, X. 2024. Stress-testing long-context language models with lifelong icl and task haystack. *arXiv preprint arXiv:2407.16695*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 Technical Report. *CoRR*.

Zhang, P.; Shao, N.; Liu, Z.; Xiao, S.; Qian, H.; Ye, Q.; and Dou, Z. 2024. Extending Llama-3's Context Ten-Fold Overnight. *arXiv preprint arXiv:2404.19553*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhong, W.; Guo, L.; Gao, Q.; Ye, H.; and Wang, Y. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19724–19731.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.