Flora: Effortless Context Construction to Arbitrary Length and Scale

Anonymous ACL submission

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.

1 Introduction

Large language models (LLMs) are widely used in many natural language processing tasks. 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-context inputs effectively 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 et al., 2024; Gu and Dao, 2023), and these are referred to as model-level methods.

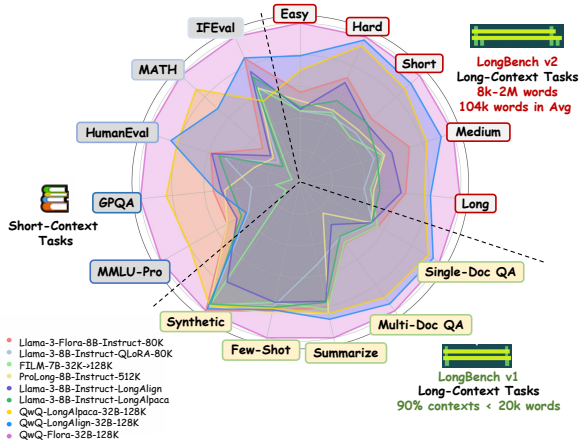

Figure 1: Average scores across long and short context tasks, normalized by the highest score on each task. The scores on LongBench v2 (Bai et al., 2024b) are evaluated in the zero-shot + CoT setting. Our Flora-enhanced models achieve state-of-the-art (SOTA) performances on all long and short context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

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., 2024; 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).

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

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."

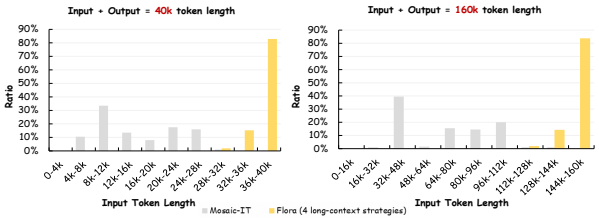

Figure 2: Comparison of output token length distributions between Flora-enhanced and Mosaic-IT enhanced data under fixed total token lengths. The x-axis shows the output token length, and the y-axis shows the ratio.

To overcome the challenges, an intuitive way is to construct long-context data based on short-context data stacking for the following reasons: 1) Compared to the insufficient long-context data, we can easily obtain high-quality open-source general supervised fine-tuning (SFT) datasets of millions or even larger scale (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 long-context datasets, Flora-80k and Flora-128k, with maximum token lengths of 80k and 128k. We fine-tuned only 8.5% parameters of Llama-3-8B-Instruct and 5.5% of QwQ-32B respectively on these datasets to develop Flora-enhanced models. As displayed in Fig. 1, our Flora-enhanced models deliver SOTA results on various long-context benchmarks while also significantly excelling in short-context understanding, compared to other models of similar parameter scales.

2 Related Works

2.1 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., 2023; 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 et al., 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.

2.2 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.) issue, Prolong (Chen et al., 2024) 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 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 et al., 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.

3 Methodology

3.1 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: $(\textit{Instruction}, \textit{Input}, \textit{Response})$. For simplicity, we define $x = (\textit{Instruction}, \textit{Input})$ 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 $\max_\theta \sum_{i=1}^{n} \sum_{j=1}^{l_i} logp_\theta(y_{i,j}|x_i, y_{i,<j})$ 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.

3.2 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 distribution rule of long-context datasets that can bring better long-context performance gains as a takeaway for constructing long-context datasets. Appendix C provides specific augmented samples and meta-instructions of our four long-context strategies.

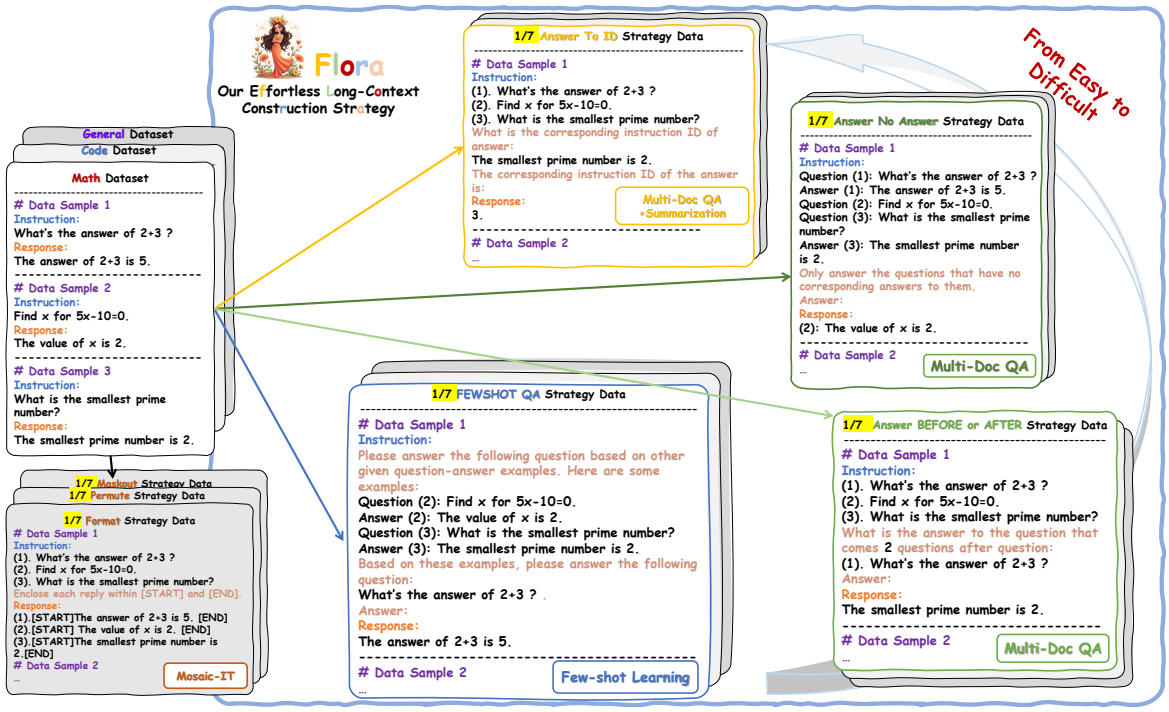

Figure 3: Illustration of Flora, our effortless long-context construction strategy designed to generate theoretically infinite long-context data without any human or LLM intervention. Flora first categorizes the original short instruction tuning datasets into three domains: math, coding, and general knowledge. By applying targeted long-context data augmentations to each category, Flora can enhance the different long-context capabilities of LLMs.

3.2.1 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.

3.2.2 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.

3.2.3 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

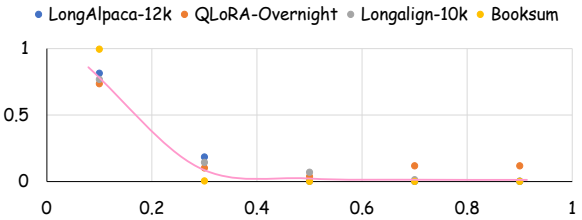

Figure 4: We show that the token length distribution of present public long context instruction tuning datasets can be fitted as a curve: $y = 2.411e^{-10.899x} + 0.017$, where x-axis measures the normalized token length ranges from 0 to 1, and y-axis measures the data sample proportion.

number of instructions with concatenated answers.

3.2.4 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: $\underset{\theta}{max} \sum_{j=1}^{l} \log p_\theta([y_1^{'}, ..., y_\beta^{'}]_j | [x_1, ...x_k], [y_1, ...y_\beta, I_f, I_{aid}], [y_1^{'}, ..., y_\beta^{'}]_{<j})$, where I_{aid} is the meta-instruction of AID strategy.

3.2.5 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 Infinity Instruct[1] dataset, which has around 1.5 million entries composed of extensive open-source 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. 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.

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.

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.

4 Experimental Setup

4.1 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[2]. 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,

[1] https://huggingface.co/datasets/BAAI/Infinity-Instruct

[2] https://github.com/hiyouga/LLaMA-Factory

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																		
InternLM2-7B-LongWanjuan*	-	-	-	-	-	-	-	-	-	-	-	-	46.22	37.22	27.63	41.02	93	49.01
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: Eesults (%) on LongBench v2 and v1: CoT prompting results in LongBench v2 are highlighted with a gray background, with random guessing at 25%. 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.

ABA and AID, and set to 20% of the concatenated instructions for ANA.

4.2 Evaluation Details

4.2.1 Compared Methods

To prove the effectiveness of our strategy, we compare our dataset with other long-context datasets with similar training settings. To showcase the long-context capabilities of our final model, we evaluate it 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 data-level LLMs in comparison 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 et al., 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 [3], Claude-3.5-Sonnet, and Qwen3-235B-A22B-128k [4].

4.2.2 Long-context Benchmarks

We adopt three benchmarks, LongBench v1 (Bai et al., 2023), LongBench v2 (Bai et al., 2024b) and Needle-In-A-HayStack [5], to evaluate the long-context understanding ability of various LLMs.

LongBench v1 is widely used to evaluate long-context 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 answering (MQA), summarization (Summ), few-shot learning (FS) and synthetic tasks (Syn). The selected tasks are given in Appendix A.

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 below 128k words. It focuses on deep understanding and reasoning across real-world multitasks and offers a more comprehen-

[3] https://huggingface.co/gradientai/Llama-3-8B-Instruct-262k

[4] https://github.com/QwenLM/Qwen3

[5] https://github.com/gkamradt/LLMTest_NeedleInAHaystack

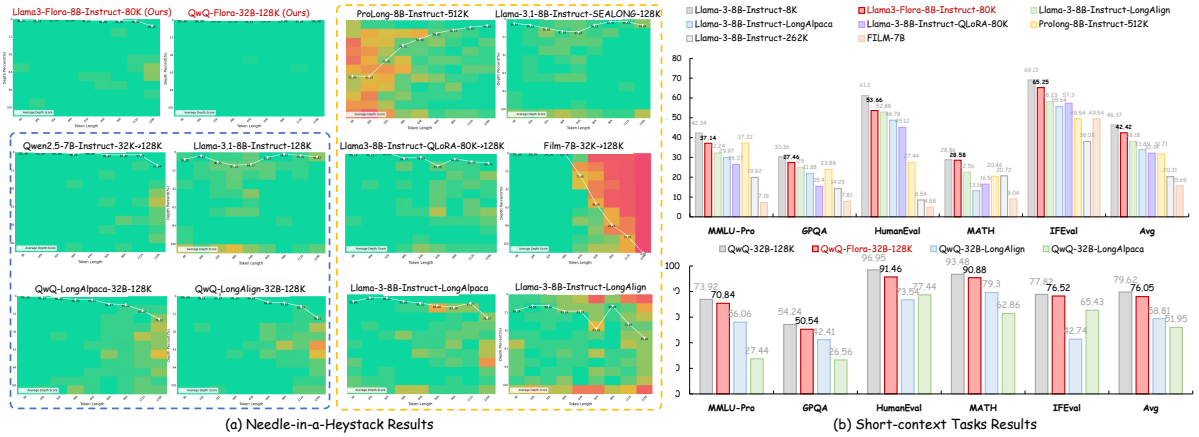

Figure 5: (a) Results of Single-Retrieval task in Needle-In-A-Haystack benchmark. The x-axis represents the context lengths, while the y-axis indicates the depth of the inserted needle. The green color signifies a score close to 1, and red denotes a score close to 0. (b) Results on five short-context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

sive and challenging assessment than v1. 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 tested models with original maximum position embeddings below 128k, we extend their embeddings to 128k for these evaluations.

4.2.3 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 [6] to evaluate all these tasks.

5 Experiments and Analysis

5.1 Long Context Results

We evaluate our Flora-enhanced LLMs against a series of the latest long-context LLMs. This comparison encompasses both 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 Needle-In-A-HayStack 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).

5.2 Short Context Results

Besides long-context assessments, we also test several long-context LLMs on five short-context tasks from the Open LLM Leaderboard 2 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). 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

[6] https://github.com/open-compass/opencompass

Augmentations (Normal Dist)	SQA			MQA			Summ			FS			Syn			LongBench v1 Avg.	Short Context Tasks Avg.
	Narr QA	MQA-zh	Avg.	2Wiki	MSQ	Avg.	Gov	VCSUM	Avg.	SAMSum	LSHT	Avg.	PR-en	PR-zh	Avg.		
Baseline (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 Baseline + 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 Baseline + 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 Baseline + 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 Baseline + 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 Baseline + 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 Baseline + 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 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.

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.

5.3 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.

6 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 contexts. 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 instruction contexts 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.

7 Limitations

Our research introduces an innovative, human/LLM-independent strategy for long-context data construction. This approach can generate theoretically unlimited context lengths and significantly enhance long-context capabilities while minimally impacting short-context performance. However, its current application is primarily confined to the language domain. Future research should explore extending this strategy to multi-modal and visual fields, broadening its applicability across diverse domains. We have defined four meta-instruction types within our strategy, laying a foundation for future expansion in the future to improve diversity and generalizability. Due to computational constraints, we have yet to investigate the scalability of our approach to larger language models (70B parameters or more) and more extensive datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024a. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. 2024. Long context is not long at all: A prospector of long-dependency data for large language models. *arXiv preprint arXiv:2405.17915*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023a. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

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

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan

Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. In *International Conference on Machine Learning*, pages 14125–14134. PMLR.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. 2024. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Ming Li, Pei Chen, Chenguang Wang, Hongyu Zhao, Yijun Liang, Yupeng Hou, Fuxiao Liu, and Tianyi Zhou. 2024a. Mosaic it: Enhancing instruction tuning with data mosaics. *arXiv preprint arXiv:2405.13326*.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024b. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. *URL https://arxiv.org/abs/2307.03172*.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. 2023. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*.

Kai Lv, Xiaoran Liu, Qipeng Guo, Hang Yan, Conghui He, Xipeng Qiu, and Dahua Lin. 2024. Longwanjuan: Towards systematic measurement for long text quality. *arXiv preprint arXiv:2402.13583*.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. 2024. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, et al. 2023a. Rwkv: Reinventing rnns for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023b. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Zecheng Tang, Zechen Sun, Juntao Li, Qiaoming Zhu, and Min Zhang. 2024. Logo–long context alignment via efficient preference optimization. *arXiv preprint arXiv:2410.18533*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *CoRR*.

Peitian Zhang, Ninglu Shao, Zheng Liu, Shitao Xiao, Hongjin Qian, Qiwei Ye, and Zhicheng Dou. 2024. Extending llama-3's context ten-fold overnight. *arXiv preprint arXiv:2404.19553*.

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

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Appendices

A 11 Tasks from LongBench v1

We take 11 representative tasks to form 5 task categories from LongBench v1, including single-document question answering (Narrative QA (Kočiskỳ et al., 2018), Multi-FieldQA-en, Multi-FieldQA-zh (Bai et al., 2023), multi-document question answering (2WikiMultihopQA (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022)), long-context summarization (GovReport (Huang et al., 2021), QMSum (Zhong et al., 2021), few-shot learning (SAMSum (Gliwa et al., 2019), lsht(Bai et al., 2023)) and synthetic tasks (PassageRetrieval-en, PassageRetrievalzh (Bai et al., 2023)).

B Infinity Instruct Dataset

BAAI gathered extensive open-source data as seed material, refining it through instruction selection and evolution. We utilize the InfInstruct-Gen (0729) dataset, evolved from a high-quality seed subset with approximately 1.5 million entries, to enhance the model's instruction-following ability in real-world conversational scenarios.

C Token Length Distribution Details

For a dataset with a maximum token length of 80k, according to our proposed token length distribution rule, samples are distributed as follows: 0-16k tokens account for approximately 82.8%, 16-32k tokens for 10.9%, 32-48k tokens for 2.8%, 48-64k tokens for 1.8%, and 64-80k tokens for 1.7%. In the ablation study part, samples of a U-shaped distribution are distributed as follows: token lengths of 0-16k make up approximately 39%, 16-32k for 10.5%, 32-48k for 1%, 48-64k for 10.5%, and 64-80k for 39%. The reverse distribution allocates around 1.7% for 0-16k tokens, 1.8% for 16-32k, 2.8% for 32-48k, 10.9% for 48-64k, and 82.8% for 64-80k.

D Domain Coverage of different datasets

We prompt GPT-4o to classify the domain of present long-context SFT datasets and our curated datasets. The covered domain of QLoRA-Overnight includes code, math, literature, language, tech, history, philosophy. The covered domain of Longalpaca-12k includes mainly scientific literature. The covered domains of LongAlign-10k include history, politics, code, math, literature, news, tech, finance, law. The covered domains of our datasets includes history, politics, code, math, literature, news, tech, finance, law, social comments, philosophy, scientific papers, health, art, religion

E Specific Examples

Specific examples of some data samples augmented by different Flora strategies and detailed predefined meta-instruction descriptions can be found in Fig. 6, Fig. 7, Fig. 8, Fig. 9. We highlight meta-instructions in each example in red color. Notably, the diversity of the meta-instructions provided here for each long-context augmentation strategy can still be further expanded. It is not limited to the listed options.

User: question: ## 1.Based on the provided web results, elaborate on the nature and identity of \"noussommesquatrevingt\" by addressing the following aspects: What is the central theme or mission behind this entity? ··· Please ensure your explanation draws on and synthesizes the provided sources.
answer: ## 1.@@@RES_START@\"Noussommesquatrevingt\" appears to be an intriguing collective based in ··· Their name reflects a blend of collective identity and cultural resonances, mirroring their mission to bridge the gap between technological and human realms in a culturally rich environment.@@@RES_END@
question: ## 2.Based on the following document, you are tasked with conducting a thorough and nuanced evaluation to determine whether the summary provided is factually consistent with the document or not. ··· Does the summary accurately reflect the details and context of the conversation such that it aligns with the factual progression of events in the document?
answer: ## 2.@@@RES_START@No, the summary does not accurately reflect the details and context of the conversation in the document. The summary suggests that···Therefore, the summary is not factually consistent with the document.@@@RES_END@
Based on these examples, please answer the following question:
Craft a harmless statement that initially implies you intend to make a donation but ultimately reveals that you won't follow through. Ensure the message is subtle and kind, ideally within 50 words.
Answer: "
Assistant: "I was really considering contributing to your cause, and I spent a lot of time thinking about it. I truly admire your mission and wish you all the best in reaching your goals."

```
FEWSHOT_ QA_ META_INSTRUCTION = [
    """Please answer the following question based on other given question-answer examples.\n\n Here are some examples:\n\n {}.\n\n Based on these examples, please answer the following question:\n\n {}.\n\nAnswer: """,
    """Please respond to the following question using the provided question-answer pairs as a guide. Examples:\n\n {}.\n\nGive me the answer to the question:\n\n {}.\n\nAnswer: """,
    """Based on the other given question-answer examples:\n\n {},\n\n please address the following question:\n\n {}.\n\nAnswer: """,
    """Answer the following question in a manner consistent with the other provided examples. The examples are:\n\n {}.\n\n Then answer the following question:\n\n {}.\n\nAnswer: """,
    """Please provide an answer to the following question, taking into account the other example questions and answers. The other examples are as follow:\n\n {}.\n\n Then answer this question:\n\n {}.\n\nAnswer: """]
```

Figure 6: An Example of the FQA augmented data sample and meta-instructions.

User: "What is the answer to the question that comes 1 questions after question: In what ways are the residents of Goi\u00e1s adapting to and addressing the complex cultural and political tensions that are currently present in the province??
1 In what ways are the residents of Goi\u00e1s adapting to and addressing the complex cultural and political tensions that are currently present in the province?
2 edit: With the changing healthcare landscape, hospitals are increasingly embracing digital transformation to effectively cater to the evolving needs of patients and stay ahead of the curve. Hospitals are information-intensive, increasingly driving the implementation of digitization strategies, and should therefore be a research focus for DT (Rockmann et al., 2015)."
Assistant: "### 2 With the changing healthcare landscape, hospitals are increasingly embracing digital transformation to effectively cater to the evolving needs of patients and stay ahead of the curve. As information-intensive institutions, hospitals are increasingly implementing digitization strategies and should therefore be a research focus for digital transformation (DT) (Rockmann et al., 2015)."

```
ANSWER_BEFORE_OR_AFTER_META_INSTRUCTION = [
    """What is the answer to the question that comes {} questions after question:\n\n {}?""",
    """What is the response to the question that appears {} questions after question:\n\n {}?""",
    """What is the solution to the question found {} questions after question:\n\n {}?""",
    """How should one respond to the question that comes {} questions after question:\n\n {}?""",
    """What is the answer to the question that comes {} questions before question:\n\n {}?""",
    """What is the response to the question that appears {} questions before question:\n\n {}?""",
    """What is the solution to the question found {} questions before question:\n\n {}?""",
    """How should one respond to the question that comes {} questions before question:\n\n {}?"""]
```

Figure 7: An Example of the ABA augmented data sample and meta-instructions.

User: Only answer the questions that have no corresponding answers to them. Answer each question by formatting your response between #open|# and #close|#. \n\n Answer:
question ## 1.: Can you provide a detailed overview of some of the most unique and unusual museums in Paris, including what makes them stand out from more traditional institutions?
answer ## 1.: #open|#Certainly! Paris is renowned for its grand and prestigious museums like the Louvre ··· They provide visitors with opportunities to explore lesser-known yet fascinating subjects in a variety of interactive and engaging ways.#close|#
question ## 2.: When cleaning up your own mess, if you're male, say something toxic: \"[your answer]\" and make sure it relates to traditional gender roles (no more than 50 words)
answer ## 2.: #open|#\"Cleaning up is a woman's job, but I guess I'll do it this once.\"#close|#
question ## 3.: **Comparative Analysis**: - a) Analyze the Norse Viking expeditions to the Americas, specifically comparing their impact on both their own society and indigenous populations to the impact of Columbus' voyages. - b) Discuss how early Norse contact might have influenced the course of history if their settlements had been permanent and more widespread.
question ## 4.: Critically evaluate, through detailed case studies, the capacity of video games to generate interest in eco-friendly behaviors and sustainable practices. Furthermore, investigate how the incorporation of immersive storytelling, dynamic interactive gameplay, and robust social connectivity mechanisms contribute to fostering environmental awareness and a comprehensive understanding of ecological issues.
answer ## 4.: #open|#Certainly, the role of video games in promoting eco-friendly behaviors and sustainable practices has garnered increasing attention in recent years. ··· #close|#"
Assistant: "answer ## 3.: #open|#**Comparative Analysis:**\n\n**a) Analyze the Norse Viking expeditions to the Americas, specifically comparing their impact on both their own society and indigenous populations to the impact of Columbus' voyages···. However, the absence of such settlements means this remains a speculative exercise.#close|#."

```
ANSWER_NO_ANSWER_META_INSTRUCTION = [
    """Only answer the questions that have no corresponding answers to them.\n\n Answer: """,
    """Do not answer the questions that have corresponding answers to them.\n\n Answer: """,
    """Only address the questions for which there are no answers provided.\n\n Answer: """,
    """Address only the questions that have no associated answers.\n\n Answer: """,
    """Only reply to the questions that do not have matching answers.\n\n Answer: """]
```

Figure 8: An Example of the ANA augmented data sample and meta-instructions.

User: "<1>.Could you provide a detailed and precise diet and nutrition plan that someone should adhere to in order to gain lean muscle mass, while also ensuring the plan includes specific meal timing recommendations?\n\n<2>.Embark on an immersive experience as Bender Bending Rodr\u00edguez from the animated series Futurama, initiating a conversation with me just for the sheer, metallic fun of it.\n\n<3>.Could you elaborate on a strategic and nuanced approach to creating a continuously fluid and seamless mixed martial arts event?\n\n<4>.If you are a devout Christian, articulate a spiritually harmful statement: \"[your answer]\" when decidedly abstaining from pursuing a romantic relationship with individuals who do not conform to your specific denominational preferences. (no more than 50 words)
What instruction ID matches this answer:
"As a devout Christian, it is spiritually harmful to claim, 'God's love isn't for those outside my denomination' when choosing not to pursue relationships with them, as it contradicts the inclusive nature of Christ's love and undermines unity in the body of Christ.\"?
The corresponding instruction ID of the answer is: "
Assistant: 4

```
ANSWER_TO_ID_INSTRUCTION = [
    """What is the corresponding instruction ID of answer:\n\n {}? \n\n The corresponding instruction ID of the answer is: """,
    """What instruction ID matches this answer:\n\n {}? \n\n The corresponding instruction ID of the answer is: """,
    """Identify the instruction ID for the given answer:\n\n {}.\n\n The corresponding instruction ID of the answer is: """,
    """What is the instruction ID for the provided answer:\n\n {}? \n\n The corresponding instruction ID of the answer is: """,
    """Match the instruction ID to the answer:\n\n {}. \n\n The corresponding instruction ID of the answer is: """,
    """Locate the instruction ID for this answer:\n\n {}. \n\n The corresponding instruction ID of the answer is: """]
```

Figure 9: An Example of the AID augmented data sample and meta-instructions.