# Never Lost in the Middle: Mastering Long-Context Question Answering with Position-Agnostic Decompositional Training

**Junqing He, Kunhao Pan, Xiaoqun Dong,**
**Zhuoyang Song, Yibo Liu, Qianguo Sun,**
**Yuxin Liang, Hao Wang, Enming Zhang, Jiaxing Zhang**
International Digital Economy Academy, Shenzhen, China
hejunqing@idea.edu.cn

## Abstract

While large language models (LLMs) are equipped with longer text input capabilities than before, they struggle to seek correct information in long contexts. The "lost in the middle" problem challenges most LLMs, referring to the dramatic decline in accuracy when correct information is located in the middle. To overcome this crucial issue, this paper proposes to enhance the information searching and reflection ability of LLMs in long contexts via specially designed tasks called **P**osition-**A**gnostic **M**ulti-step QA (PAM QA). Trained with this task, our model excels in focusing more precisely on the desired information. Experimental results show substantial improvement in Multi-doc QA and other benchmarks, surpassing state-of-the-art models by a 13.7% absolute gain in shuffled settings and by 21.5% in the passage retrieval task. We release our model and code to promote related research in the community.[1]

## 1 Introduction

Large Language Models (LLMs), renowned for their exceptional generative and zero-shot learning abilities across diverse natural language processing (NLP) fields, have found extensive downstream applications (OpenAI, 2023; Boiko et al., 2023; Cheng et al., 2023; Waisberg et al., 2023; Hu et al., 2023). However, LLMs suffer from severe hallucinations, significantly compromising their performance in knowledge-oriented QA, dialogue, and writing (Roberts et al., 2020; Agrawal et al., 2023). Retrieval Augmented Generation (RAG) is an effective solution to hallucinations, and remarkable improvements have been achieved by incorporating supporting knowledge into the input of LLMs (Lewis et al., 2020b; Shuster et al., 2021; Thoppilan et al., 2022; Shi et al., 2023a). The most fundamental challenge to address in RAG is long context and Multi-document question answering (Multi-doc QA).

Some research works around the problem with a complicated pipeline or system (Chen et al., 2023a; Lee et al., 2024), but we aim to improve foundation models as they are a core component of those methods. Thorough research has been conducted to deal with long context inputs, categorized into three mainstreams: The first is to expand the context window using a sliding window (Dai et al., 2019; Xiao et al., 2023). Other researchers proposed to enhance the extrapolation ability by improving the Relative Positional Encoding in Transformers, the backbone of most LLMs (Su et al., 2021; Press et al., 2021; Luo et al., 2022; Vaswani et al., 2017). These two kinds of modifications both show substantial improvement in language modelling (LM). The third category of studies focuses on the recurrent compression of memory for long-range sequence learning (Rae et al., 2019; Peng et al., 2023). This methodology effectively learns the comprehensive representation of context, demonstrating notable proficiency in rapid computation and cost-effectiveness during inference. Though the methods above show strong performance in specific tasks and support LLMs with extra-long context windows, i.e. GPT3.5-Turbo-16K, Claude-v1.3-100K and Longchat (Dacheng et al., 2023), LLMs fail to produce correct answers if related documents are located in the middle of the context, called ***"lost in the middle"*** (Liu et al.). It is fatal for Multi-doc QA. However, whether a similar deterioration exists in Chinese LLMs has been unexplored and solutions to this problem have rarely been researched.

We hypothesise that the scale of attention scores of the beginning context grows large after pre-training and instruction tuning while that of the middle context, whose position is less trained, remains small for a long distance to the current token.

[1] It is publicly available at `https://huggingface.co/IDEA-CCNL/Ziya-Reader-13B-v1.0` and `https://github.com/hejunqing/never-lost-in-the-middle`



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642
August 11-16, 2024 ©2024 Association for Computational Linguistics

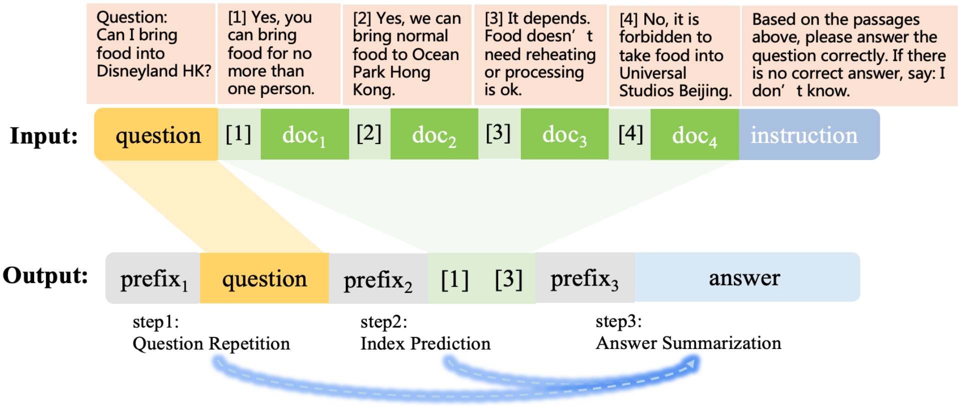



Figure 1: The workflow of PAM QA. The blue dashed lines indicate information flows. The desired output of a sample is composed of three parts, corresponding to three steps: Question repetition, index prediction, and answer summarization. [i] refers to the index of the $i$-th document. An input sample is displayed on the top.

This limits the contribution of related information to the answer and results in lower QA accuracy.

To overcome the pitfall, we proposed position-agnostic decompositional training to even up the attention scores over input context. Concretely, we designed a tailored Multi-doc QA task in which positive documents are located at arbitrary positions in contexts among noisy documents. The task presents a significant challenge, compelling the models to extract and summarize information despite the interference of useless ones (Ye et al., 2022). As human beings routinely solve complex tasks by decomposition to obtain higher quality outcomes (Cheng et al., 2015; Correa et al., 2023), we modified the Multi-doc QA task as a multi-step reasoning task, called **P**osition-**A**gnostic **M**ulti-step QA (**PAM** QA), combining the Chain-of-Thought (COT, Wei et al.) and position-agnostic Multi-doc QA. Trained with explicit extraction of the question and the index of supporting documents before generating answers, models learn to distinguish correct information from noisy ones and attend to them. It also forces attention to the question and supporting indexes stronger although the attention scale decays with increasing distance (Su et al., 2021).Empirical results on Multi-doc QA and other benchmarks show that, with only 1/2 or 1/4 context window size, our model improves upon state-of-the-art (SOTA) models by 7.0% in the top-ranked setting and by 13.7% in the shuffled setting. Competitive results are shown in other attention-dependent tasks including passage retrieval and summarization.

The contribution of this paper is threefold:

- This paper proposed a novel task named PAM QA to tackle the "lost in the middle" issue, which is fatal for knowledge-intensive scenarios. To our knowledge, it is the first attempt to solve the problem by training models on special tasks.
- We investigate the model's behaviour in-depth, revealing that failing to focus on target information may be the cause of "lost in the middle".
- Comprehensive experiments have shown that the proposed PAM QA is effective in solving the "lost in the middle" problem. Our model surpasses SOTA in Multi-doc QA and other related tasks on renowned Chinese benchmarks. It is non-trivial that the general QA ability of the model is also strong and satisfying. The model is open-sourced to boost future research in the community.

## 2 Position-Agnostic Multi-step QA

Multi-doc QA refers to a type of QA task where a model is presented with multiple documents and asked to answer questions correctly. It is difficult for models and humans alike, requiring accurate retrieval, information aggregation and comprehension from noisy candidates while struggling with fading memory.

In this situation, task decomposition, identifying subproblems and reasoning about them, becomes essential (Correa et al., 2023). We decomposed the difficult Multi-doc QA to PAM QA. This innovative task comprises three steps, as depicted in Figure 1.

The entire process of PAM QA unfolds as follows: when receiving a question, a set of candidate documents, and a specific instruction, the model initiates by generating $prefix_1$. It then proceeds to restate the question, predicting the indexes of related evidence after incorporating a connecting phrase, denoted as $prefix_2$. Finally, it formulates an answer by aggregating previous information, following an answer indicator, $prefix_3$.

### 2.1 Question repetition

The first step is question repetition (QR). The questions are placed at the front as a contextual-aware representation (Liu et al.). The subtask is started with $prefix_1$, "As for the question:" (or expressions with identical meaning) to prompt the model.

### 2.2 Index Prediction

Supporting evidence not only helps LLMs verify themselves but also aids users in evaluating responses (Menick et al., 2022). Remarkable results have been shown in generating quotes and citations (Thoppilan et al., 2022; Menick et al., 2022). We hypothesize that the indicator helps to encode and navigate the attention to corresponding documents.

Accordingly, the second step is index prediction (IP), namely to predict the indexes of the supported documents for the question as an MRC task, beginning with $prefix_2$: "Based on the information numbered". Unlike previous works that predict a verbatim quote extracted from a longer source retrieved, the indexes of corresponding evidence are the targets. For the case in Figure 1, the label of this step is "Based on the information numbered [1],[3]". Considering the indexes in the second step only count for very few tokens and are hard to emphasize in the sequential cross-entropy loss during training, an MRC task that only asks to predict indexes of correct documents is added as a supplement.

### 2.3 Answer Summarization

The third step is to generate the final answer after information aggregation. Thanks to the steps above, it can be simplified as an answer summarization (AS) task. The step starts with an indicator like "my answer is" as $prefix_3$.

In line with the proverb *"the palest ink is better than the best memory,"* we teach the model to take notes, turning these annotations into a highway to the relevant knowledge. It can reduce the distraction of extraneous information and make the attention to the question and supporting index stronger because the attention scale decays with increasing distance.

## 3 Training Data Construction

We equipped our model with distinguishing ability through instruction tuning. The training procedure is composed of two stages. We expand the LLM's context window to 8K in the first stage. In the second stage, the model was trained with PAM QA data to solve the attention (or memory) failure called "lost in the middle".

### 3.1 Context Window Expansion

We used about 300k selected data for general supervised finetuning (SFT). The data cover various categories of tasks including QA, MRC, role-playing, writing, coding, translation, brainstorming, math, Language Modeling (LM), and other natural language understanding (NLU) tasks like text classification. The data are packed to 8K window size in a multi-turn conversation style except for the LM task, which calculates the cross-entropy loss on the whole sequence.

### 3.2 PAM QA

Data are constructed by formatting inputs and concatenating target outputs of the steps in PAM QA. We first generated Multi-doc QA data and adapted it to PAM QA data.

First, we filtered out 30K samples of the Fact category with a single answer from DuReader2.0 dataset (He et al., 2018) and 20K samples from WebCPM (Qin et al., 2023). DuReader2.0 is the largest Chinese MRC dataset collected from Web documents and community QA, containing 200K questions, 420K answers and 1M documents. To ensure the quality of data, we creatively utilize a reward model to score the samples and select the high-quality part of them with thresholds, inspired by Li et al.. The reward model is trained with 69K human-ranked samples for alignment in general tasks, following Köpf et al. and Ouyang et al.. As both datasets only contain positive samples, negative samples are ingeniously generated subsequently.

As collaborative learning is beneficial to RAG (Izacard et al., 2022), we built a search engine with all the documents in the corresponding dataset. For each sample, documents in the whole collection except the positive ones are regarded as negative samples. We retrieved documents from the search engine as negative candidates for a partition comprising 70% of the data, while we randomly sampled from the original negative candidates for the remaining portion of data. The retrieved negative samples are more relevant to questions and harder to distinguish from the positive ones than random samples. Next, documents are shuffled within each sample in 50% of the data to prevent positive ones from consistently being at the beginning of contexts. Next, 25K samples were sampled from retrieval benchmarks, T2Rank (Xie et al., 2023) as the relevance MRC, a supplement for task 2. The negative samples are randomly sampled from the hard negative collections and shuffled with positive candidates. The indexes of positive documents are recorded.

The max length of each sample is sampled from 1K to 8K under the uniform distribution. This ensures our model can deal with samples of various input lengths with correct documents located at any position.

To enable the model to recognize situations where the correct document is absent, we generate "Synthetic Unknown" samples, where all documents are negative. The answer for these samples is a constant term indicating "I don't know." This category of data accounts for a proportion of 5%.

Finally, We sampled some general SFT data, taking a 20% ratio in this stage to alleviate the catastrophic forgetting (McCloskey and Cohen, 1989; Rebuffi et al., 2017). The total training samples in this stage summed up to 90K.

### 3.3 Training

We trained our model based on a pre-trained LLM that adapted from LLaMA2, called Ziya2-13B-Base (Touvron et al., 2023; Gan et al., 2023; Zhang et al., 2023). We trained for 2 epochs on 16 A100 GPUs in both stages with constructed data. The learning rate began with 1e-5 then decayed to 1e-6 with a warmup for the first 0.05% steps in the first stage. The max learning rate for the second stage was 5e-6. Flash Attention (Dao et al., 2022) was utilized to accelerate the training procedure. Sampling is turned on for all models during testing in the benchmarks. The hyperparams for testing are listed in Appendix A.

## 4 Experiments

In this section, we evaluate the long-context QA abilities of our model and existing representative LLMs. By inspecting the performance, we can verify whether our model overcomes the so-called "lost in the middle" problem (Liu et al.).

| Datasets | Avg length | Source | Metrics |
|---|---|---|---|
| Multi-doc. | 15,768 | DuReader | Rouge-L |
| Synt. | 6,745 | C4 Chinese | Accuracy |
| Summ. | 15,380 | VCSUM | Rouge-L |
| Single-doc | 6,701 | Multifield QA | F1 |
| RGB NR. | 1,105.7 | Self Generated | EM |

Table 1: The statistics of input lengths of the testing datasets. Multi-doc. is short for Multi-doc QA. Synt. and Summ. represent Synthetic Tasks and Summarization respectively while RGB NR is the abbreviation of RGB noise robustness task.

### 4.1 Benchmarks

We conducted experiments on a long context benchmark, LongBench (Bai et al., 2023) and Retrieval-Augmented Generation Benchmark (RGB, Chen et al.). LongBench measures various abilities of the testee given long input contexts. Specifically, we tested models on four related tasks in LongBench: Chinese Multi-doc QA, Synthetic tasks, summarization and single-doc QA. We also used the noise robustness testbed in RGB to test the QA ability in short texts, which examines the information extraction ability given a certain ratio of noise documents.

The synthetic task is a document retrieval task, where given a summary, the goal is to find the corresponding document from a large number of candidates. This task evaluates the information retrieval ability of LLMs in long contexts. The summarization task gives extremely long meeting records from multiple speakers and asks for a summary. It assesses the model's memory and summarization capabilities. Single-doc QA is a long-context QA task that is less similar to multi-doc QA. We conduct experiments on this task to test the robustness of the model. The context lengths and other statistics of the datasets are listed in Table 1. The evaluation scripts were provided by the LongBench official website[2] and RGB official repository.

We also re-constructed the synthetic task to examine whether the models are "lost in the middle". The correct passages are relocated at the 1st, 5th, 10th, 15th and 20th with passages located beyond the 20th removed. The results are in Figure 2.

Considering that the documents in the samples of Multi-doc QA tasks are basically sorted by relevance, we shuffled the first 10 candidate documents in each sample to make the real performance exposed, called Multi-doc QA shuffled.

[2] https://github.com/THUDM/LongBench

| Model | Multi-doc QA | Synthetic Tasks | Summarization | Single-doc QA |
|---|---|---|---|---|
| (Baichuan2-Turbo-192K) | 36.8 | 90.0 | 18.4 | 44.7 |
| Longchat-v1.5-7B-32K | 19.5 | 7.6 | 9.9 | 29.1 |
| ChatGLM2-6B-32K | 37.6 | 64.5 | **16.1** | 32.8 |
| (ChatGLM3-6B-32K) | 44.8 | 94.0 | 17.8 | 62.3 |
| GPT3.5-Turbo-16K | 28.7 | 77.5 | 16.0 | **61.2** |
| Vicuna-v1.5-7B-16K | 19.3 | 5.0 | 15.1 | 43.0 |
| Xgen-7B-8K | 11.0 | 3.5 | 2.2 | 14.8 |
| InternLM-7B-8K | 16.3 | 0.9 | 12.4 | 33.6 |
| Qwen-14B-Chat | 18.7 | 40.0 | 13.9 | 31.4 |
| Our model | **44.6** | **98.5** | 15.6 | 34.4 |

Table 2: The results are Rouge-L percentage for Multi-doc QA and Summarization while Synthetic Tasks compute the accuracy (EM scores). Models are separated in lines by context window sizes. ChatGLM3-6B-32K and Baichuan2-Turbo-192K are new models after our work.

In addition, we conducted a comprehensive human evaluation of model capabilities to see if training on PAM QA harms the general abilities of LLM. The test set contains 200 questions from a wide range of categories.

### 4.2 Baselines

We compared the performance of the most popular LLMs with a long context window. These strong baselines include: GPT3.5T-Turbo-16K extends the context window to 16K tokens, while both Longchat-v1.5-7B-32K (Dacheng et al., 2023) and ChatGLM2(3)-6B-32K (Du et al., 2022) further push the boundary to 32K tokens. Vicuna-v1.5-7B-16K (Zheng et al., 2023) and Xgen-7B-8K (Nijkamp et al., 2023) offer fine-tuned models on user-shared conversations and 8K sequences respectively. Baichuan2-13B-Chat (Yang et al., 2023) stands out in few-shot learning with a 4K token window, alongside a larger closed-source variant. Lastly, Qwen-14B-Chat introduces a 14B parameter model with dynamic NTK (dyn, 2023), trained on a window size of up to 8K tokens. We refer to retrieval-augmented models as those trained with retrieval-augmented data or paradigms. Baichuan2-13B-Chat and Baichuan2-Turbo-192k are both retrieval-augmented models (Yang et al., 2023).

## 5 Results and Discussion

In this section, we analyze the experimental results of the LLMs and discuss the reason for the findings. An ablation study is also conducted for in-depth attribution. Other details are in the Appendix.

### 5.1 Longer window size does not guarantee better performance

As shown in Table 2, our model has a Rouge-L of 44.6% in the Multi-doc QA task, 7.0% higher than ChatGLM2-6B-32K, which was the SOTA model. With only 1/4 window size, our model can outperform ChatGLM2-6B-32K at this task. It reveals the strong attention ability of our model since it is an open-book QA task. This Chinese Multi-doc QA dataset does not need to consider all of the contexts, as the correct documents are located at the beginning of contexts.

In the Synthetic Task, namely an abstract retrieval task, our model achieves the highest result with an accuracy of 98.5%, among models with longer context capabilities. This indicates that the "lost in the middle" issue is almost solved by the proposed method in this paper, as long as the average length is covered.

As for summarization, ChatGLM2-6B-32K and GPT3.5-Turbo-16K have similar performance with different context window sizes, showing that longer context window sizes do not guarantee better performance. The Rouge-L of our model is only 0.5% lower than SOTA, without any summarization data in the PAM QA training. As the average length of the task is much longer than 8K tokens, our model with a longer context length will have a promising improvement.

We observe a moderate result in Single-doc QA from our model and find it competitive among 8K models. GPT3.5-Turbo-16K achieves the highest result of 61.2% F1 score (before ChatGLM3-6B-32K), surpassing the longest model, Baichuan2-Turbo-192K.

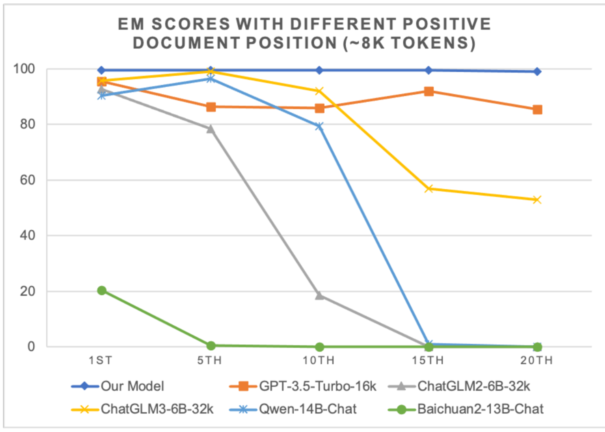



Figure 2: The EM score on Synthetic (passage retrieval) Task from LongBench with correct document inserted to certain position ranging from 1st to 20th.

### 5.2 PAM QA alleviates lost in the middle (and tail) problem

Experiments on the re-constructed Synthetic (passage retrieval) benchmark display the performance of models on different positive document positions. Concretely, the correct passage for each sample is inserted into the 1st, 5th, 10th, 15th, and 20th locations respectively among the other documents in each experiment. Theoretically, we should see a U-curve described in (Liu et al.), called "lost in the middle". Results are displayed in Figure 2.

However, the figure suggests that most open-source LLMs are lost not only in the middle but also in the tail. A significant decrease is observed when positive documents are placed at the 10th position. Despite the employment of techniques such as Alibi (Press et al., 2022) or NTK to expand the context window (i.e. Baichuan2-13B-Chat and Qwen-14B-Chat), models still demonstrate low results. In contrast, our model can survive in different settings of positions, holding a record of 99%. It reveals the effects of PAM QA training.

### 5.3 Models defeated by shuffled dataset, attention failure being the culprit

Figure 4 demonstrates the models' performance on Multi-doc QA before and after shuffling. We can see a sharp decline in all three models except ours. The largest gap reaches 17.3%, from ChatGLM2-6B-32K. Meanwhile, Baichuan2-13B-Chat also has a 7% reduction although the shuffled documents are within its context length. Therefore, LLMs without extra long context windows also have difficulty dealing with the challenge. Our model is the most robust model with a 3.7% decrease.

| Noise Ratio | 0 | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|---|
| GPT3.5-Turbo | 95.67 | **94.67** | **91.00** | **87.67** | **70.67** |
| ChatGLM2-6B | 86.67 | 82.33 | 76.67 | 72.33 | 54.00 |
| (ChatGLM3-6B) | 91.67 | 90.00 | 89.00 | 84.67 | 66.33 |
| Baichuan2-13B-Chat | 93.00 | 90.33 | 89.00 | 82.33 | 63.33 |
| Qwen-14B-Chat | 94.67 | 92.00 | 88.00 | 85.30 | 69.67 |
| Our model | **96.00** | 90.67 | 90.00 | 85.50 | 67.33 |

Table 3: Performance in RGB noise robustness testbed. EM scores are in percentage. ChatGLM3-6B is a new model after our work.

To unearth the cause of the decline and examine the attention capabilities of models, we visualize the attention scores of the last layer for the identical input. We repeat a sentence including the correct answer 20 times as the context to find if all of them will be highlighted in the self-attention procedure in models. Attention scores of ChatGLM2-6B-32K and Our model over the input are depicted in Figure 3.

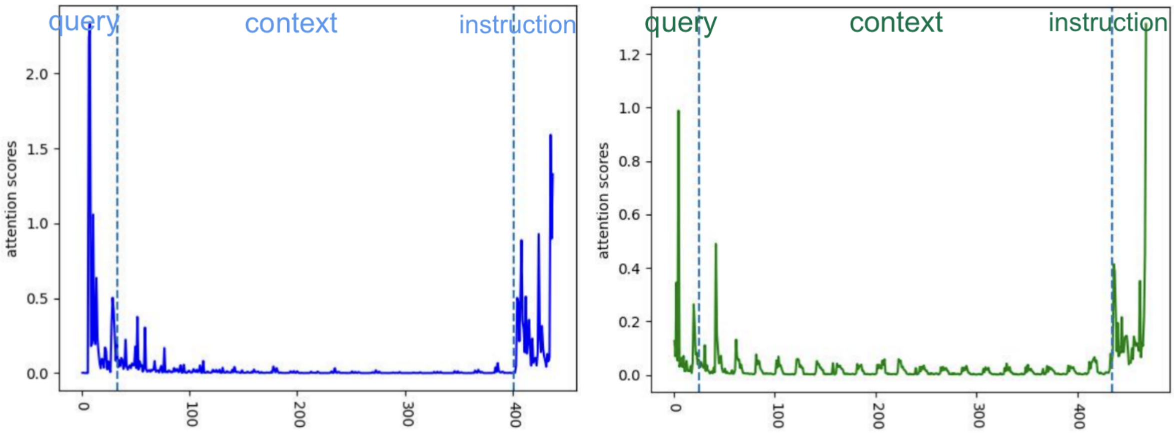
We can see the attention scores on documents are fading away in ChatGLM2-6B-32K, as the context after the first 100 tokens is almost neglected. The situation is quite different when it comes to our model. 20 peaks of attention scores are observed (the last one is next to the beginning of instruction), corresponding to the answers in sentences. It reveals that attention to related tokens is the key to the performance gap between models. The models struggle to precisely focus on the correct tokens, paying tremendous attention to the beginning and the ending tokens (where instruction and query are frequently located), which is the culprit of the "lost in the middle" problem.

### 5.4 Competitive results observed in short text Multi-doc QA

As reported in Table 3, our model has a competitive performance among open-source models on short-text multi-doc QA although not trained with any short texts. Even compared with the latest popular Chinese LLMs, Qwen-14B-Chat and ChatGLM3-6B-32K, the results of our model are higher under the setting of noise rate in [0,0.4,0.6].

### 5.5 General ability is preserved with PAM QA Training

A side-by-side (SBS) comparison was performed by 3 human annotators to check the general ability of our model. General capabilities including commonsense, math, reasoning, QA, writing, harmless-




(a) ChatGLM2-6B-32K

(b) Our model

Figure 3: The attention scores over the input tokens in the self-attention procedure within ChatGLM2-6B-32K and our model on a document repeated 20 times. Length differs with tokenizers.

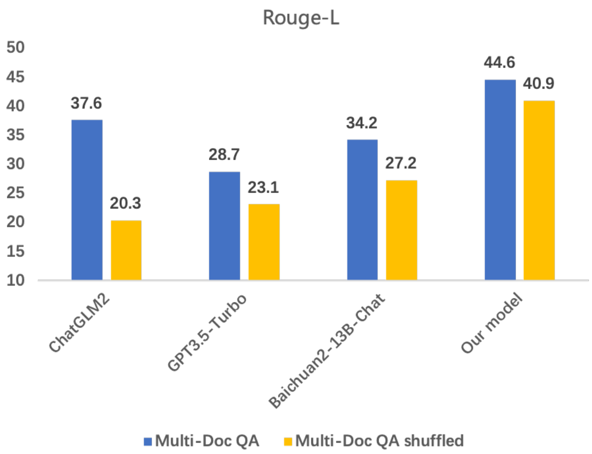



Figure 4: Performance on Multi-doc QA before and after shuffling. ChatGLM2 is short for ChatGLM2-6B-32K, GPT3.5-Turbo is short for GPT3.5-Turbo-16K. Scores are in percentage.

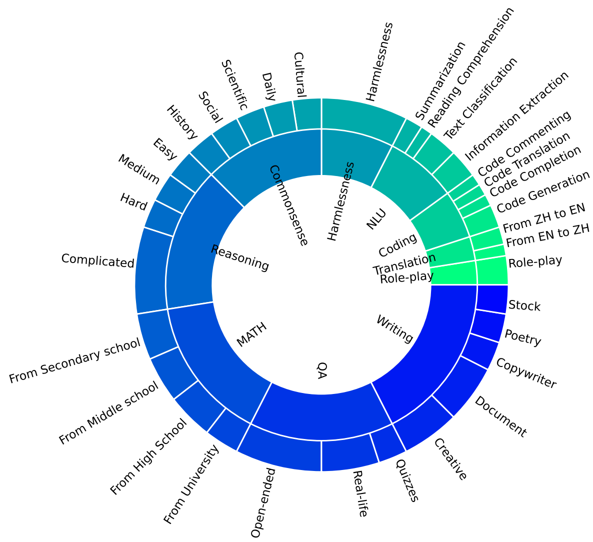



Figure 5: The distribution of tasks in the general ability test.

ness, etc. are examined in the test, as shown in Figure 5. The annotators are asked to choose a better answer among two given answers unless the answers are both bad or the same, as in (Zheng et al., 2023). Annotators are all master students. They are blind to the models and other information. Results compared with similar size models, Ziya-LLaMa-13B-v1.1[3] and Baichuan2-13B-Chat respectively are illustrated in Figure 6. We also compare our model with the same base model after full SFT training, Ziya2-13B-SFT.

Figure 6 summarizes the human preference between our model and other open-source LLMs. Although our model was trained only with PAM QA data, it performs slightly inferior to Baichuan2-13B-Chat but better than Ziya-LLaMa-13B-v1.1 and Ziya2-13B-SFT significantly. Thus, the general capabilities are maintained after the PAM QA training.

### 5.6 Ablation Study

**Each step in PAM QA matters.** Here we inspect the contribution of each step in PAM QA. The variants are evaluated on Multi-doc QA and Synthetic tasks. Results of this ablation study are listed in Table 4.

[3] https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1.1

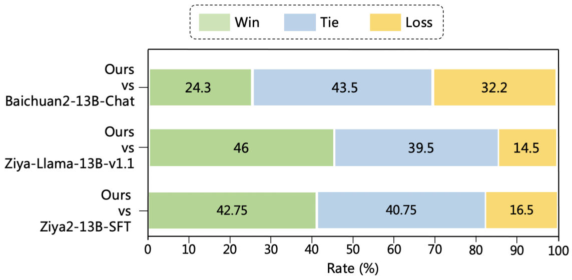



Figure 6: SBS results on general ability evaluation that contains a wide range of tasks. Ziya2-13B-SFT is the full SFT version based on the same pre-trained model.

| Variants | Multi-doc QA | Synt. |
|---|---|---|
| Our model | **44.6** | **98.5** |
| - QR | 38.8 | 98.0 |
| - QR - IP | 37.8 | 1.3 |
| Only-CWE | 8.7 | 7.5 |

Table 4: Synt. is short for Synthetic tasks. Results are in percentage. QR is short for question repetition. IP is short for index prediction. Only-CWE represents the finetuned model only with context window expansion.

Without question repetition, the first step in PAM QA, a 5.8% decrease can be observed in Multi-doc QA, showing its inevitable contribution to high performance. It strengthens the attention of the question by repeating the question first. Then the model can directly attend to the question in the subsequent steps without going through a long context, reducing the distraction of context when performing self-attention.

When the index prediction (IP) step is removed, the pronounced decrease in Synthetic tasks emphasizes its importance. It not only teaches LLMs to distinguish between pertinent and irrelevant information but also changes the model's prior behaviour (i.e., seeking information from the beginning and the end of context). Meanwhile, it streamlines the process by allowing models to concentrate on relevant abstracted information, instead of repeatedly scanning extensive input tokens. A performance drop in multi-doc QA also shows the contribution of IP. Since the scale of the attention scores decays as the distance grows (Su et al., 2021), models with rotary position embeddings (RoPE) struggle to remember the remote tokens without training. With the former two steps, the question and the potentially correct evidence are listed just a few tokens ahead. This reduces the probability of forgetting questions and context by decreasing the distance.

An enormous gap between the results of Our model and that without QR and IP, indicates the substantial improvement from PAM QA training. We visualize the attention scores when predicting the first token and discover the generated questions and indexes are highlighted, shown in Figure 7 in Appendix D.

Compared with Only-CWE, the variant model without QR and IP steps also gains 29.1% improvement, which shows the effect of position-agnostic and challenging negative candidates. By transforming the Multi-doc QA into PAM QA, the same data can boost the performance by 6.8% in Multi-doc QA, and 97.2% in Synthetic Tasks, which reveals the strength of task decomposition training.

**Necessity of Training.** To investigate whether training (fine-tuning) is necessary, we performed multi-step COT prompting in the style of "first predict the indexes of relevant documents" and "according to the information, the final answer is" on different models. We removed the question repetition step in COT for better performance. Results are in Table 5.

| Model | Strategy | Multi-doc. | Synt. |
|---|---|---|---|
| GPT3.5-Turbo-16K | w/o COT | 28.7 | 77.0 |
| GPT3.5-Turbo-16K | w/ COT | 28.3 | 63.9 |
| Yi-34B-Chat | w/o COT | 14.9 | 35.3 |
| Yi-34B-Chat | w/ COT | 2.5 | 58.6 |
| Yi-34B-Reader (Ours) | w/o COT | **45.1** | 50.4 |
| Ziya2-13B-SFT | w/o COT | 11.0 | 6.3 |
| Ziya2-13B-SFT | w/ COT | 1.1 | 2.1 |
| Ziya2-Reader (Ours) | w/o COT | 44.6 | **98.5** |

Table 5: Comparison of results from different strategies: models with (w/), without (w/o) multi-step COT inference and with PAM training (Ours).

As demonstrated by the superior results of our models over the multi-step COT inference, training is essential to optimize performance. Especially in Multi-doc QA, LLMs tend to produce an answer with more hallucination after predicting a list of indexes of related documents without fine-tuning. We found models with multi-step COT struggle to handle complex instructions and maintain long-context memory.

**Generalizability of Method.** To illustrate the generalization of the approach, we also performed identical training on another Chinese pre-trained model, Yi-34B-Base[4], a top 34B pre-trained model in LLM Benchmarks. Table 6 lists the comparison

[4]https://huggingface.co/01-ai/Yi-34B

| Model | Multi-doc QA | Synthetic Task | Summarization |
|---|---|---|---|
| Yi-34B-Chat | 14.9 | 35.3 | 13.8 |
| Yi-34B-Reader (Ours) | **45.1** | 50.4 | 14.2 |
| Ziya2-13B-SFT | 11.0 | 6.3 | 12.6 |
| Ziya2-13B-Reader (Ours) | 44.6 | **98.5** | **15.6** |

Table 6: Comparison of models trained with PAM and official full SFT models based on the same pre-trained models. Ziya2-13B-SFT is the model trained on the same pre-trained model, Ziya2-13B-Base.

of the model we trained (Yi-34B-Reader) and the official instruction tuning version, Yi-34B-Chat[5].

Results show our method can be generalized to other LLMs. However, it is harder to change the behaviour of Yi-34B-Base using only 100K PAM QA data compared to the 13B model since it was pre-trained rather (maybe over) sufficiently. It results in lower results in the synthetic task than the 13B Ziya2-Reader.

## 6 Related Works

### 6.1 Retrieval-Augmented Language Models

Retrieval-Augmented Language Models (RALMs) mark notable progress in NLP by merging the capabilities of expansive LMs with the precision and intricacy offered by external knowledge sources. (Guu et al., 2020; Lewis et al., 2020a; Izacard et al., 2022). These models use a retriever to search through a large body of evidence, like Wikipedia, to find specific documents related to the user's query. Afterwards, a reader component is utilized to carefully examine these documents and generate a response. This two-step process guarantees both relevance and depth in the produced answers. Recent research efforts have concentrated on enhancing the performance of the retriever (Karpukhin et al., 2020; Sachan et al., 2023) or the reader(Izacard and Grave, 2020; Cheng et al., 2021), training the system end-to-end (Lewis et al., 2020a; Sachan et al., 2021), and integrating the retrieval systems with black-box large language models (Shi et al., 2023b; Yu et al., 2023; Trivedi et al., 2023)

### 6.2 RALMs Adapted to Long and Noisy Context

Recent research emphasizes the influence of contextual length and the position of related context on the performance of LLMs (Krishna et al., 2023; Bai et al., 2023; Liu et al.). The research closely aligned with ours is the study by (Yoran et al., 2023), training RALMs to disregard irrelevant contexts. A homothetic COT-like training approach was proposed to solve math and coding problems, emitting intermediate computation steps into a "scratchpad" (Nye et al., 2021). However, they overlooked long context scenarios, specifically the "lost in the middle" issue, a key consideration in our work.

An earlier work that considered multi-doc modelling in training is proposed by Caciularu et al.. After splitting long context into pieces and generating QA pairs based on picked salient ones, they asked models to predict the masked salient sentences and answers, given other pieces and the generated questions. Significant improvement in multi-hop QA benchmarks after fine-tuning with the training set is reported at the expensive cost of pre-training. However, there is no training set in most benchmarks nowadays and it fails to perform diverse tasks in the zero-shot setting.

## 7 Conclusion

In this paper, we assume that the widely recognized "lost in the middle" phenomenon may caused by weak attention to target information. We discover popular Chinese LLMs are "lost" both in the middle and tail. A novel approach is proposed to address the deficiency in LLMs by training models with Posistion-Agnostic Multi-step (PAM) QA. Experimental results show the superiority and effectiveness of our method, surpassing SOTA LLMs in Multi-doc QA and passage retrieval significantly, with only 1/4 context window size. By shuffling the candidate documents in open benchmarks, degraded performance is observed in all models, among which our model is the most robust one. The ablation study also reveals the significant effect of PAM QA and the positive contribution of its components. Our study also finds that LMs with extremely long context windows do not ensure better performance on Multi-doc QA and passage retrieval tasks. We hope our study provides profound insight into the "lost in the middle" problem at a broader scale and sheds light on developing more

[5]https://huggingface.co/01-ai/Yi-34B-Chat

intelligent LLMs.

## Limitations

Our work covers the important "lost in the middle" issue and experiments with Chinese Benchmarks on popular Chinese and English LLMs with long context capability. The improvements in tested tasks do not imply similar upgrades in all aspects, like math and reasoning. The constructed PAM QA data were used after or during SFT, with the effect in pre-training and RLHF period unexplored.

The data construction method is mainly based on multi-doc QA and shows substantial gains in related tasks. Improvements in other long-context tasks are not as impressive as multi-doc QA and synthetic tasks since other abilities are more required than discriminating and focusing. Those situations are not considered in this paper.

The proposed approach is language-independent and could be applied to datasets of other languages. The core of the method lies in constructing samples with challenging related negative documents, diverse positions of positive samples and multi-step reasoning answers. Therefore, the method can potentially alleviate the "lost in the middle" issue in other languages.

## References


2023. Dynamically scaled rope further increases performance of long context llama with zero fine-tuning.

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

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

DaniilA. Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models.

Avi Caciularu, Matthew E. Peters, Jacob Goldberger, Ido Dagan, and Arman Cohan. 2023. Peek across: Improving multi-document modeling via cross-document question-answering.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023a. Walking down the memory maze: Beyond context limit through interactive reading.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation.

Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2021. UnitedQA: A hybrid approach for open domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3080–3090, Online. Association for Computational Linguistics.

Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. 2015. Break it down: A comparison of macro-and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4061–4064.

Kunming Cheng, Qiang Guo, Yongbin He, Yanqiu Lu, Shuqin Gu, and Haiyang Wu. 2023. Exploring the potential of gpt-4 in biomedical engineering: the dawn of a new era. *Annals of Biomedical Engineering*, pages 1–9.

Carlos G Correa, Mark K Ho, Frederick Callaway, Nathaniel D Daw, and Thomas L Griffiths. 2023. Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, 19(6):e1011087.

Li* Dacheng, Shao* Rulin, Xie Anze, Sheng Ying, Zheng Lianmin, Gonzalez Josep E., Stoica Ion, Ma Xuezhe, and Hao Zhang. 2023. How long can open-source llms truly promise on context length?

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Ruyi Gan, Ziwei Wu, Renliang Sun, Junyu Lu, Xiaojun Wu, Dixiang Zhang, Kunhao Pan, Ping Yang, Qi Yang, Jiaxing Zhang, et al. 2023. Ziya2: Data-centric learning is all llms need. *arXiv preprint arXiv:2311.03301*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. *ACL 2018*, page 37.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

PatrickS.H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Michael Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv: Computation and Language,arXiv: Computation and Language*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

NelsonF Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts.

Shengjie Luo, Shanda Li, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. 2022. Your transformer may not be as powerful as you expect. *Advances in Neural Information Processing Systems*, 35:4301–4315.

Michael McCloskey and Neal J. Cohen. 1989. *Catastrophic interference in connectionist networks: the sequential learning problem*, page 109–165.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. 2023. Long sequence modeling with xgen: A 7b llm trained on 8k input sequence length. *Salesforce AI Research Blog*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models.

OpenAI OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Xiangru Tang, Bolun Wang, Johan S. Wind, Stansilaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. Rwkv: Reinventing rnns for the transformer era.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988, Toronto, Canada. Association for Computational Linguistics.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2023. Questions are all you need to train a dense passage retriever. *Transactions of the Association for Computational Linguistics*, 11:600–616.

Devendra Singh Sachan, Siva Reddy, William Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and ScottWen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023b. Replug: Retrieval-augmented black-box language models.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, pages 1–4.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. *arXiv preprint arXiv:2304.03679*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models.

Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Towards a better understanding of human reading comprehension with brain signals. In *Proceedings of the ACM Web Conference 2022*, pages 380–391.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2023. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

## A Hyper-parameters

We used the following settings in the inference phrase in the LongBench test: $do\_sample = True$, $top_p = 0.85$, $temperature = 0.8$, $repetition\_penalty = 1.0$, $early\_stopping = True$.
As for the RGB noise robustness testbed, the temperature was set to 0.2.

## B Evaluation of Index Prediction

We apply human annotation to 140 samples in Multi-doc QA to investigate the precision and recall of the index prediction (IP) step. The macro precision is 43.0% and the recall is 70.9%. They result from more indexes predicted than ground truths. The LLMs may have the reflection ability to check the indexes and finally produce correct answers. Therefore, the IP step benefits the answer generation even in low precision.

## C Demonstration of PAM QA

A case of Multi-doc QA and the answer produced by Our model and ChatGPT3.5 are shown in Table 7.

## D Visualization of attention scores on tokens in answer summarization step

The heatmap of attention scores of each token when generating the final answer is displayed in Figure 7.

## E Annotation System

The screenshot of the annotation system in SBS general ability evaluation is shown in Figure E.

Chinese:

<s><human> :给定问题：联合国2022年对人口展望中，预计世界人口峰值达到多少?
文章:
[1] 联合国发布的《2022 年世界人口展望》报告中预测，印度将在明年超过中国，成为世界上人口最多的国家。
[2] 根据联合国经济和社会事务部发布的《2022 年世界人口展望》报告，全球人口预计将在 2022 年 11 月 15 日达到 80 亿。报告同时预测，世界人口预计将在 2080 年代达到约 104 亿的峰值，并在 2100 年之前保持这一水平。
[3] 11月15日，根据统计网站Worldometers实时统计数据，全球人口已超80亿。
[4] 联合国(UN)核监督机构周二批准了日本的“核排污”计划，不顾当地渔民和该地区其他国家的反对，允许日本将受损的福岛第一核电站逾100万吨核废水排入海洋。
[5] 据联合国官网消息，近日，联合国193个成员国一致通过了一项具有法律约束力的协定，旨在加强海洋生物多样性养护和可持续发展，尤其是对公海的保护。
请阅读理解上面多个文章，正确地回答问题。若检索结果都不相关，请回答都不相关。
<bot> : 对于问题“联合国2022年对人口展望中，预计世界人口峰值达到多少? ”，根据上面的编号为1,2,3的信息，我的答案是根据联合国经济和社会事务部发布的《2022 年世界人口展望》报告，全球人口预计将在 2022 年 11 月 15 日达到 80 亿。报告同时预测，世界人口预计将在 2080 年代达到约 104 亿的峰值，并在 2100 年之前保持这一水平。因此，联合国预计世界人口峰值将达到约 104 亿。</s>

English:

<s><human>: Given question: What is the expected peak world population in the UN's 2022 population outlook?
Essays:
[1] The World Population Prospects 2022 report released by the United Nations predicts that India will surpass China as the world's most populous country next year.
[2] According to the World Population Prospects 2022 report published by the United Nations Department of Economic and Social Affairs, the global population is expected to reach 8 billion by 15 November 2022. The report also predicts that the world's population is expected to peak at about 10.4 billion in the 2080s and remain at that level until 2100.
[3] On November 15, according to real-time statistics from the statistics website Worldometers, the global population has exceeded 8 billion.
[4] The United Nations (UN) nuclear watchdog on Tuesday approved Japan's plan to "dismantle nuclear wastewater," allowing Japan to discharge more than 1 million tons of nuclear wastewater from the damaged Fukushima Daiichi nuclear power plant into the ocean, despite opposition from local fishermen and other countries in the region.
[5] According to the official website of the United Nations, the 193 member states of the United Nations recently unanimously adopted a legally binding agreement aimed at strengthening the conservation and sustainable development of marine biodiversity, especially the protection of the high seas.
Please read and understand many of the passages above and answer the questions correctly. If the search results are not relevant, please answer that they are not relevant.

<bot> : In response to the question "What is the expected peak world population in the United Nations 2022 population outlook?" Based on the information numbered 1,2,3 above, my answer is that according to the World Population Prospects 2022 report published by the United Nations Department of Economic and Social Affairs, the global population is expected to reach 8 billion on November 15, 2022. The report also predicts that the world's population is expected to peak at about 10.4 billion in the 2080s and remain at that level until 2100. As a result, the United Nations expects the world's population to peak at about 10.4 billion.</s>

ChatGPT3.5: The expected peak world population in the UN's 2022 population outlook is not explicitly mentioned in the provided essay passages.

Table 7: A Multi-doc QA sample and the output of Our model, ChatGPT3.5 in the last two cells respectively. The answer of our model starts after the <bot> :. Special tokens are removed when testing ChatGPT3.5. We can see that Our model produces a correct answer while ChatGPT3.5 fails to answer the question.

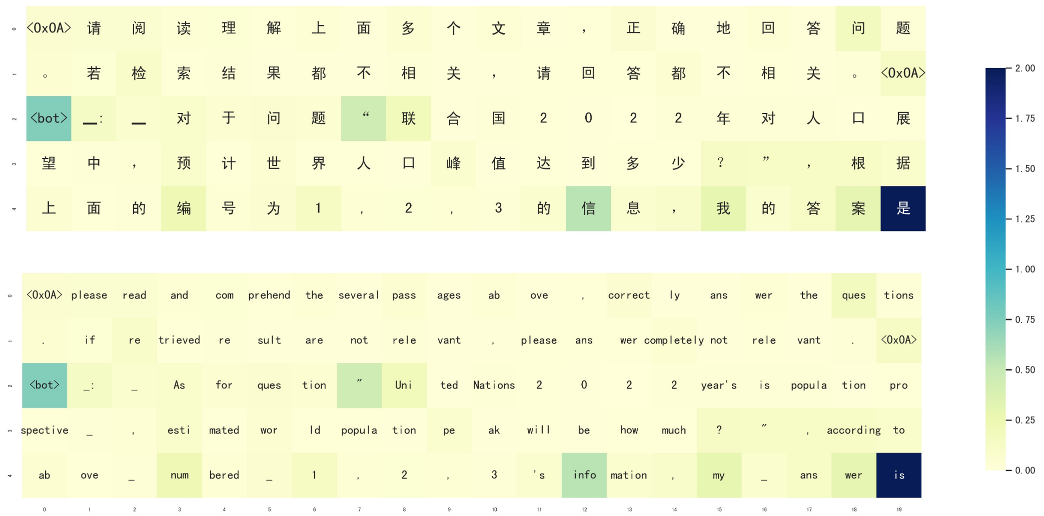



Figure 7: The heatmap of attention scores on the last 100 tokens when our model begins to generate an answer. The darker represents the higher score. The English version is translated from the Chinese heatmap token by token. The first token of the generated question and the first predicted index are attended to, indicating the utility of the question repetition and index prediction in PAM QA.


Figure 8: Interface of SBS evaluation system. Users can only rank answers without access to the names of models.