POSITION BIAS MITIGATES POSITION BIAS: Mitigate Position Bias Through Inter-Position Knowledge Distillation

Yifei Wang[1,2,3*], **Feng Xiong**[3*], **Yong Wang**[3†‡], **Linjing Li**[1,2†]
Xiangxiang Chu[3], **Daniel Dajun Zeng**[1,2]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]AMAP, Alibaba Group
wangyifei2022@ia.ac.cn, wangyong.lz@alibaba-inc.com
https://github.com/AMAP-ML/Pos2Distill

Abstract

Positional bias (PB), manifesting as non-uniform sensitivity across different contextual locations, significantly impairs long-context comprehension and processing capabilities. Previous studies have addressed PB either by modifying the underlying architectures or by employing extensive contextual awareness training. However, the former approach fails to effectively eliminate the substantial performance disparities, while the latter imposes significant data and computational overhead. To address PB effectively, we introduce **Pos2Distill**, a position to position knowledge distillation framework. Pos2Distill transfers the superior capabilities from advantageous positions to less favorable ones, thereby reducing the huge performance gaps. The conceptual principle is to leverage the inherent, position-induced disparity to counteract the PB itself. We identify distinct manifestations of PB under **R**etrieval and **R**easoning paradigms, thereby designing two specialized instantiations: *Pos2Distill-R*^{1} and *Pos2Distill-R*^{2} respectively, both grounded in this core principle. By employing our approach, we achieve enhanced uniformity and significant performance gains across all contextual positions in long-context retrieval and reasoning tasks. Crucially, both specialized systems exhibit strong cross-task generalization mutually, while achieving superior performance on their respective tasks.

1 Introduction

Who tied the bell could be the one to untie it.
— Chinese proverb

Large Language Models (LLMs) are increasingly proficient in handling long contexts, which has been unlocked by key innovations in efficient attention mechanisms (Dao et al., 2022; Ainslie et al., 2023) and length extrapolation techniques (He et al., 2024; Chi et al., 2023b; Press et al., 2021). These breakthroughs enable LLMs to tackle complex question answering over substantially larger context windows (Agarwal et al., 2024), marking a crucial step towards more capable and versatile natural language systems.

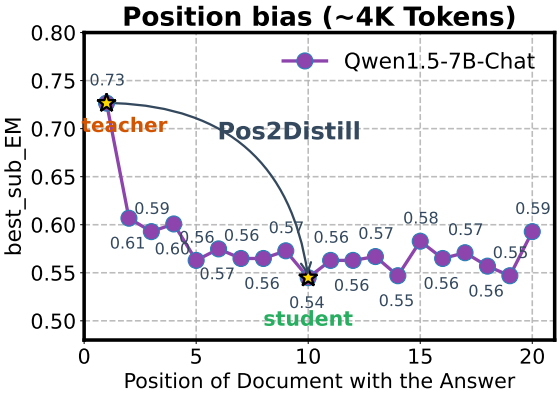


Figure 1: **Motivation for Pos2Distill**. Marked performance decline from the 1st to 10th position in multi-document QA underscores severe PB. From an alternative viewpoint, superior responses at advantageous positions provide effective supervisory signals for less optimal positions.

Nevertheless, recent studies point out a critical limitation: LLMs do not uniformly extract and utilize information across long context, consistently favoring information located at the context edges while neglecting that in the middle. This phenomenon, commonly termed the *lost in the middle* problem (Liu et al., 2024), underscores a pervasive and intrinsic PB inherent in long-context handling.

PB poses significant obstacles in information-rich settings, such as retrieval-augmented generation (Fang et al., 2025; Dong et al., 2025b), long-context reasoning (Li et al., 2024a; Kuratov et al., 2024), and LLM-as-a-judge (Wang et al., 2024b; Li et al., 2024c). When critical information is distributed arbitrarily throughout the input, LLMs fail to identify and integrate gold content (Baker et al., 2024a; Peng et al., 2025), culminating in unex-

[*] Equal contribution. Work done when Yifei's internship at AMAP, Alibaba Group.
[†] Corresponding author.
[‡] Project lead.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 1496–1513
November 4-9, 2025

pected model failures across various applications.

To alleviate PB, prior research has delved into its underlying causes, seeking to modify key architectural components or internal representations linked to uneven contextual sensitivity (Zhang et al., 2024c; Chen et al., 2024). Despite recent progress in narrowing the performance gap, a substantial disparity in information utilization between advantageous and disadvantaged positions still persists. Another line resorts to intensive contextual awareness training (An et al., 2024; Zhang et al., 2024a) by synthesizing training data with fine-grained information awareness (Zhang et al., 2024b). However, such data-driven approaches typically incur substantial costs in both data synthesis and computational resources. Consequently, there remains a critical need for *effective* and *efficient* strategies to mitigate PB that overcome these limitations.

Inspired by the proverb 1, we contend that PB not only imposes challenges but also implicitly reveals gold signals, which can be exploited to mitigate position-induced disparity (Fig. 1). Our analysis further reveals that PB exhibits distinct behavior under retrieval and reasoning paradigms. In retrieval tasks, PB predominantly manifests as token-shifting, whereas in reasoning tasks, PB interacts with Chain-of-Thought (CoT) processes (Wei et al., 2022), leading to thought-shifting, characterized by deviations in the reasoning trajectory.

To this end, we introduce **Pos2Distill**, a novel position to position knowledge distillation framework, transferring knowledge from advantageous positions to rectify responses at unfavorable ones. Customarily, we develop two systems: Pos2Distill-R^1 and Pos2Distill-R^2. Pos2distill-R^1 mitigates token-shifting in retrieval by incorporating Kullback-Leibler (KL) divergence loss (Kullback and Leibler, 1951), providing fine-grained corrective signals. Pos2Distill-R^2 addresses thought-shifting in reasoning tasks by distilling high-quality CoT responses from advantageously positioned inputs to guide and rectify reasoning trajectories at less favorable positions.

Extensive experiments demonstrate that Pos2Distill leads to more uniform and substantially improved performance both for in-context retrieval and reasoning tasks. Furthermore, data efficiency is a notable property of our method: with only 250 training instances, Pos2Distill increases the performance of Mistral-7B-v0.3 on the NQ dataset by 6.7%, as indicated in Fig. 5.

Our contributions can be summarized as follows:

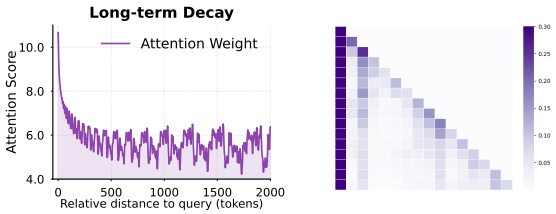


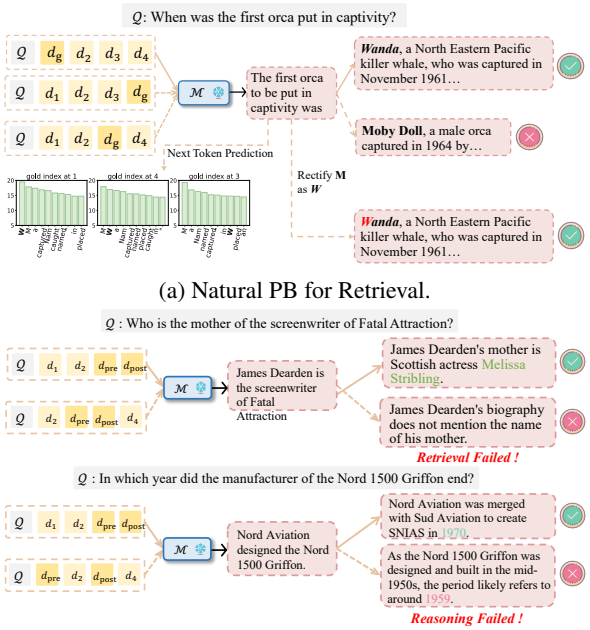
Figure 2: Left: long-term decay effect property of RoPE; Right: Attention sink (causal attention matrix).

- We uncover distinct manifestations of PB under retrieval and reasoning paradigms, namely *token-shifting* and *thought-shifting*, offering deeper insights into the nature of PB.
- We propose **Pos2Distill**, a novel position to position KD framework to mitigate PB. Given the different behavior of PB, we design two systems tailored to address the PB in both scenarios.
- Extensive experiments and ablation studies thoroughly demonstrate the efficacy in terms of performance, data efficiency, and superior generalization.

2 Related Work

Causes of Position Bias. Most LLMs adopt relative positional encodings (Peysakhovich and Lerer, 2023), such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2021), integrating token relative distances into attention score computation. This design induces a long-range decay effect in Fig. 2 (left), whereby LLMs preferentially attend to recent tokens. Concurrently, LLMs exhibit a notable bias towards the initial tokens, which can be largely attributed to the universal phenomenon, attention sink (Xiao et al., 2024; Gu et al., 2025), where disproportionate attention is allocated to early tokens in Fig. 2 (right), regardless of semantic significance. In addition, causal mask enforces a unidirectional flow of information, implicitly encoding positional information (Haviv et al., 2022; Chi et al., 2023a; Wang et al., 2024a). The interplay of these factors collectively contributes to the emergence of PB, as further elaborated in Appendix A.

Mechanistic Approaches. Current approaches predominantly focus on aforementioned underlying causes of PB. A broad spectrum of interventions has been investigated, ranging from modifications to position encodings (Zhang et al., 2024c; Chen et al., 2024; Lin et al., 2024) and alterations to causal masks (Wang et al., 2025b), to the manipulation of internal states, including attention re-



(a) Natural PB for Retrieval.

(b) Compound PB for Reasoning.

Figure 3: Different Behavior of Position Bias.

weighting (Hsieh et al., 2024; Tan et al., 2025) and hidden state manipulation (Yu et al., 2024). Despite these efforts, the substantial performance gap across token positions remains largely unmitigated.

Training Approaches. Standard next-token prediction pretraining often yields LLMs with inadequate contextual awareness (Shi et al., 2024). A research direction therefore explores specialized datasets designed to foster fine-grained information awareness (Zhang et al., 2024b), training LLMs to identify pertinent information within their full context. Nevertheless, the considerable data and computational overhead limit their scalability and practical viability.

3 Behavior of Position Bias

Natural PB for Retrieval. Given documents with identical constituent elements arranged in varying orders, our empirical observations indicate that the model's output variability is not uniform across the entire response, but rather highly concentrated at a few decisive turning positions[1]. This heterogeneous sensitivity leads to **token shifting**: the model produced erroneous tokens at these critical positions, consequently retrieval failed. Specifically, as illustrated in Fig. 3a, even with consistent prefix output, an erroneous generation of "M" instead of the correct "W" triggers retrieval failure. Moreover, we observe that manual correction of this misaligned token (e.g., changing "M" to "W") enables the model to successfully resume generation and complete the retrieval task. This finding uncovers a **token recovery** mechanism: once a misaligned token resulting from shifting is rectified, the model can realign its subsequent output. (More supported analysis experiments are provided in Appendix B .)

Compound PB for Reasoning. In-context reasoning fundamentally incorporates two processes: retrieval and reasoning. The two processes are deeply intertwined, creating a virtuous cycle and producing numerous outputs (Ren et al., 2025, 2024). Chain-of-Thought (CoT) reasoning guides the model to formulate more targeted queries, thereby enhancing the precision of the retrieval process (Wang et al., 2024c). Conversely, access to accurate, retrieved information provides the factual grounding needed to steer and validate the reasoning steps, thus ensuring the logical integrity of the reasoning chain. At the same time, the selection of reasoning path is very sensitive to irrelevant contexts (Yang et al., 2025b). Within this setting, PB is reflected both in variations during retrieval and in alterations that occur during reasoning. Examples of these two types of failures are presented in Fig. 3b. Therefore, it is crucial to reshape the overall response trajectory by integrating genuinely relevant information and reasoning chain.

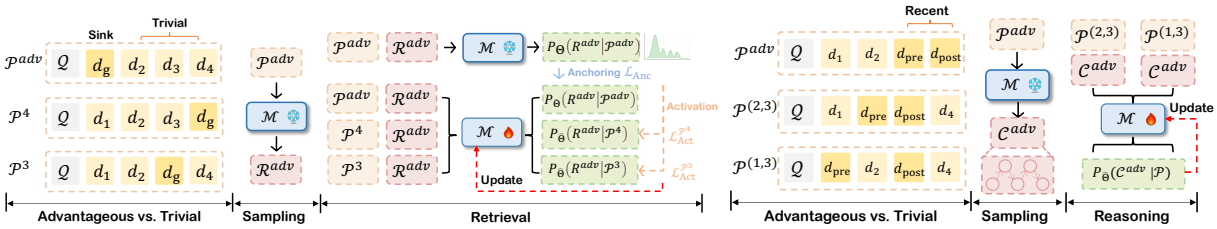
4 Methodology

PB behaves differently in retrieval and reasoning; thus, a unified approach fails to capture their inherent distinctions. Therefore, we propose two tailored position-aware distillation strategies: (1) For **R**etrieval (Pos2Distill-R^1), we directly calibrate token shifting. (2) For **R**easoning (Pos2Distill-R^2), we transfer high-quality reasoning patterns from advantageous to suboptimal positions, thereby effectively mitigating the compounded bias.

4.1 Preliminary

Task Definition. Following Wang et al. (2025b), we formally define a long-context task as follows. Given the task-specific instruction $\mathcal{I}$, a set of n retrieved documents $\mathcal{D}:=\{d_i\}_{i=1}^{n}$, and a context-dependent question $\mathcal{Q}$, a specific LLM $\mathcal{M}$ parameterized by P_Θ, generates a response conditioned on the corresponding prompt $\mathcal{P}:=\{\mathcal{I}|\ \gamma(\mathcal{D})|\ \mathcal{Q}\}$, where the function γ determines the specific ordering of documents in $\mathcal{D}$.

[1]Please refer to Appendix D for more details.



(a) **Pos2Distill-R**[1] for retrieval. (b) **Pos2Distill-R**[2] for reasoning.

Figure 4: The design of our **Pos2Distill**. The left subfigure illustrates **Pos2Distill-R**[1], highlighting its two key strategies: Activation of Trivial Positions and Anchoring of Advantageous Positions. The right subfigure depicts **Pos2Distill-R**[2], designed to reshape reasoning trajectories.

Retrieval vs. Reasoning. *Retrieval* tasks involve identifying an answer directly present in a single document $d_{\text{gold}} \in \mathcal{D}$ (Dong et al., 2025a; Liu et al., 2025). When d_{gold} is located at index i within $\mathcal{D}$, the associated prompt is denoted as $\mathcal{P}^i$, and the response to this prompt is $\mathcal{R}^* \sim \mathcal{M}(\mathcal{P}^*)$. In contrast, *reasoning* tasks require $\mathcal{M}$ to integrate information from multiple documents. For formulation brevity, we focus on the two-hop reasoning setting, which involves a first-hop document d_{pre} at index i and a second-hop document d_{post} at index j in $\mathcal{D}$. The reasoning prompt is denoted as $\mathcal{P}^{(i,j)}$, and the corresponding response is $\mathcal{C}^* \sim \mathcal{M}(\mathcal{P}^*)$.

Objective. The primary objective of Pos2Distill is to enable a model $\mathcal{M}$ such that, for any input $\mathcal{P}^i$ during generation, $\mathcal{M}$ emulates a scenario in which a gold document or a pair of sequentially relevant documents are placed at the most advantageous positions within the input context. Specifically, the advantageous position is designated as the "sink position" ($\mathcal{P}^1$) for *Retrieval* and as the "recent position" ($\mathcal{P}^{(n-1,n)}$) for *Reasoning* (shown in Fig. 4).

4.2 Pos2Distill-R[1] for Retrieval

Overall Framework. This section presents Pos2Distill-R[1], engineered to calibrate token-shifting behavior and hence mitigate PB for retrieval tasks. The framework consists of two core modules: *Activation of Trivial Positions* and *Anchoring of Advantageous Positions* in Fig. 4a. The former facilitates the transfer of effective processing capabilities from high-performing advantageous positions to underutilized trivial positions, while the latter ensures the preservation of established performance at advantageous positions, thereby narrowing the gap between trivial and advantageous positions.

Trivial vs. Advantageous Positions. For retrieval tasks, we formally define the advantageous position as the first position (sink position), with the remaining positions $\{2, \ldots, n\}$ designated as trivial ones. For d_{gold} occupies the sink position, the attention sink region overlaps with d_{gold}, thereby generating high-quality outputs that inherently imply optimal and robust attention patterns (discussed in 2) [2].

Activation of Trivial Positions. To rectify token-shifting behavior, we leverage KL divergence as a fine-grained alignment signal at each generation step. Concretely, we first sample responses $\mathcal{R}^{\text{adv}} \sim \mathcal{M}(\mathcal{P}^{\text{adv}})$ from advantageous positions. We then construct $\mathcal{K}$ prompts, each corresponding to a distinct trivial position, denoted as $\{\mathcal{P}^{n_k}\}_{k=1}^{\mathcal{K}}$ with $n_i \in \{2, \ldots, n\}$. Our objective is to align the $\mathcal{M}$'s predicted probability distribution over $\mathcal{R}^{\text{adv}}$, conditioned on each trivial prompt $\mathcal{P}^i$, with the native distribution over $\mathcal{R}^{\text{adv}}$ conditioned on $\mathcal{P}^{\text{adv}}$. Formally, for a given position n_i, knowledge distillation at n_i is defined as:

$$\mathcal{L}_{\text{Act}}^{\mathcal{P}^{n_i}} = \mathbb{E}\left[\text{KL}\left(P_\Theta(\mathcal{R}^{\text{adv}}|\mathcal{P}^{\text{adv}}) \parallel P_{\widehat{\Theta}}(\mathcal{R}^{\text{adv}}|\mathcal{P}^{n_i})\right)\right]$$

where $P_\Theta(\mathcal{R}^{\text{adv}}|\mathcal{P})$ denotes the probability distribution over outputs $\mathcal{R}^{\text{adv}}$ conditioned on input $\mathcal{P}$. Here, $\widehat{\Theta}$ represents the updated parameters at the current step. Notably, throughout training, we consistently treat $P_\Theta(\mathcal{R}^{\text{adv}}|\mathcal{P})$ as the teacher distribution rather than $P_{\widehat{\Theta}}$, avoiding the loss of advantages at the sink position during model updates.

Position-Aware Alignment. However, due to the impact on different trivial positions induced by PB varies, the alignment difficulty between $P_\Theta(\mathcal{R}^{\text{adv}})|\mathcal{P}^{\text{adv}})$ and $P_{\widehat{\Theta}}(\mathcal{R}^{\text{adv}}|\mathcal{P}^{n_i})$ is position-dependent. Intuitively, positions with higher alignment difficulty should be prioritized with gradient updates to better domain adaptation. Motivated by this intuition, we introduce *Position-Aware Alignment*, a dynamic learning strategy that adaptively adjusts learning based on alignment difficulty, ensuring balanced and effective training. Technically,

[2] Extensive empirical evidence also demonstrates superior performance at the sink position.

given a batch of $\mathcal{B}$ examples $\{\mathcal{Q}_b\}_{b=1}^{|\mathcal{B}|}$ and their corresponding $\mathcal{K}$ trivial prompts $\{\mathcal{P}_j^{n_k}\}_{k=1}^{\mathcal{K}}$ for each $\mathcal{Q}_j$, $\mathcal{K}\cdot\mathcal{B}$ prompts are divided into $n-1$ distinct trivial bins based on their trivial positions. Each bin is represented as $\widehat{\mathcal{B}}_i = \{\mathcal{P}_j^i \mid \mathcal{P}_j^i \in \mathcal{B}, j \in \{1,\ldots,|\mathcal{B}|\}\}$. To model alignment difficulty for each trivial position, we introduce an inter-position weighting scheme for position variations induced by PB while introducing an intra-position weight for differentiating instance variations in the same trivial bin $\widehat{\mathcal{B}}_i$. Therefore, a dynamic weight α_{ij} for each instance is defined as follows:

$$\alpha_{ij} = \underbrace{\frac{\exp\left(\frac{1}{|\widehat{\mathcal{B}}_i|}\sum_{j=1}^{|\widehat{\mathcal{B}}_i|}\mathcal{L}_{\text{Act}}^{\mathcal{P}_j^i}\right)}{\sum_{i=2}^{n}\exp\left(\frac{1}{|\widehat{\mathcal{B}}_i|}\sum_{j=1}^{|\widehat{\mathcal{B}}_i|}\mathcal{L}_{\text{Act}}^{\mathcal{P}_j^i}\right)}}_{\textit{inter-pos}} \cdot \underbrace{\frac{\mathcal{L}_{\text{Act}}^{\mathcal{P}_j^i}}{\max_k\left\{\mathcal{L}_{\text{Act}}^{\mathcal{P}_k^i}\right\}}}_{\textit{intra-pos}}.$$

The activation loss of $\mathcal{B}$ can be formulated as:

$$\mathcal{L}_{\text{Act}} = \sum_i^n \sum_j^{|\widehat{\mathcal{B}}_i|} \alpha_{ij}\mathcal{L}_{\text{Act}}^{\mathcal{P}_j^i}.$$

Anchoring of Advantageous Positions. During the distillation process, $\mathcal{M}$ becomes aware that the gold information possibly appear at any location within the context window, which can dilute significant attention to the sink position, potentially impairing the overall capability on diverse downstream tasks To prevent this, we introduce an anchoring loss to preserve the effectiveness of the advantageous position:

$$\mathcal{L}_{\text{Anc}} = \mathbb{E}\left[\text{KL}\left(P_\Theta(\mathcal{R}^{\text{adv}}|\mathcal{P}^{\text{adv}}) \| P_{\widehat{\Theta}}(\mathcal{R}^{\text{adv}}|\mathcal{P}^{\text{adv}})\right)\right].$$

Training Objective. Our approach optimizes a composite loss function that integrates activation loss and anchoring loss, formally:

$$\mathcal{L} = \mathcal{L}_{\text{Act}} + \lambda\mathcal{L}_{\text{Anc}},$$

where λ is a hyperparameter controlling the intensity of the anchoring loss in joint learning.

4.3 Pos2Distill-R^2 for In-context Reasoning

This section introduces the Pos2Distill-R^2 to reshape reasoning trajectories. The core principle underlying this method is to ensure that the correct reasoning process is consistently activated, irrespective of the positions of relevant documents.

Trivial vs. Advantageous Positions. Although PB in multi-hop reasoning scenario is sophisticated, insights from (Baker et al., 2024b) allow us to summarize several key patterns: PB is closely associated with the absolute position within the context window, the distance between relevant documents, and their relative order.[3] Empirically, placing d_{pre} and d_{post} at indices $n-1$ and n is expected to yield optimal performance. Consequently, we formally define the $\mathcal{P}^{\text{adv}}$ for reasoning tasks as $\mathcal{P}^{(n-1,n)}$.

Reshaping Reasoning Trajectories. We begin by sampling CoT reasoning trajectories from the advantageous positions $\mathcal{P}^{\text{adv}}$, denoted as $\mathcal{C}^{\text{adv}} \sim \mathcal{M}(\mathcal{P}^{\text{adv}})$. Similar to the procedure for *Retrieval* tasks, we construct $\mathcal{K}$ distinct prompts for each position set $\{n_k^{pre}, n_k^{post}\}$, denoted as $\{\mathcal{P}^{(n_k^{\text{pre}}, n_k^{\text{post}})}\}_{k=1}^{\mathcal{K}}$, where n_k^{pre} and n_k^{post} are selected from the set $\{1,\ldots,n\}$ subject to $n_k^{\text{pre}} \neq n_k^{\text{post}}$. The prompts $\mathcal{P}^{(n_k^{\text{pre}}, n_k^{\text{post}})}$ and their corresponding reasoning trajectory $\mathcal{C}^{\text{adv}}$ are subsequently optimized using the cross-entropy (CE) loss function to effectively capture the reasoning patterns. Formally,

$$\mathcal{L} = -\sum_{k=1}^{\mathcal{K}} \log \mathcal{M}\left(\mathcal{C}^{\text{adv}} \mid \mathcal{P}^{(n_k^{\text{pre}}, n_k^{\text{post}})}\right).$$

5 Experiments

5.1 Experimental Setup

Setup for Pos2Distill-R^1. We apply Pos2Distill-R^1 to three LLMs, including Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Qwen1.5-7B-Chat (Bai et al., 2023), and Llama-3-8B-Instruct (Touvron et al., 2023), all of which exhibit severe PB in retrieval tasks. The evaluation leverages three datasets: NaturalQuestions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQA) (Joshi et al., 2017) and WebQA (WebQ) (Berant et al., 2013), and a specilized task KV Retrieval (Liu et al., 2024). These datasets are setting as retrieval-augmented QA, with each question accompanied by 20 documents.

- **Baselines:** We compare our approach with base model, Ms-PoE (Zhang et al., 2024c), vanilla SFT and SeqKD (Kim and Rush, 2016).
- **Evaluation:** We assess PB by measuring task performance as d_{gold} is systematically placed at various positions.

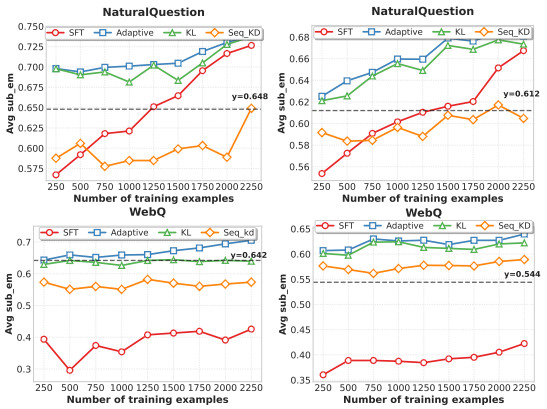
[3] Please refer to Table 12.

Methods	NQ						KV Retrieval						TQA						WebQ					
	1st	5th	10th	15th	20th	Avg.	0%	25%	50%	75%	100%	Avg.	1st	5th	10th	15th	20th	Avg.	1st	5th	10th	15th	20th	Avg.
MISTRAL-7B	**72.5**	62.3	60.7	63.9	64.9	64.8	99.8	97.6	62.0	35.6	78.0	74.6	85.2	81.8	81.8	82.6	82.2	82.7	82.7	69.7	64.5	62.4	62.5	68.4
+Ms-PoE	67.3	58.7	56.7	60.1	61.5	60.9	99.8	**97.7**	78.4	75.2	95.3	89.3	81.3	79.3	76.7	79.2	**83.4**	80.0	76.7	64.1	62.3	59.3	63.5	65.2
+SFT	68.3	64.5	66.5	66.7	62.7	65.7	**100.0**	89.0	89.2	75.6	81.8	87.1	78.2	78.0	76.8	77.2	76.2	77.3	52.7	51.1	52.3	50.7	53.9	52.1
+SeqKD	63.3	59.1	59.3	60.5	57.5	59.9	100.0	86.4	**93.0**	87.0	93.0	91.8	77.4	77.6	76.4	77.6	75.8	77.0	57.7	58.9	55.7	56.9	56.1	57.0
+Pos2Distill	70.5	**70.7**	**71.3**	**71.9**	**73.3**	**71.1**	99.0	95.4	92.6	**90.0**	**96.8**	**94.8**	**85.3**	**82.4**	**84.6**	**83.4**	81.8	**83.5**	**70.3**	**67.9**	**68.3**	**70.5**	**67.5**	**68.9**
QWEN1.5-7B	**73.6**	57.3	56.9	57.3	60.9	61.2	**100.0**	83.8	38.7	23.3	30.3	55.2	82.4	74.7	73.9	75.4	76.4	76.6	64.3	50.9	51.5	50.3	55.2	54.4
+Ms-PoE	67.4	54.8	54.2	57.4	61.3	59.0	97.4	76.5	4.7	6.3	3.2	37.6	76.4	75.2	69.4	67.4	75.1	72.7	**65.2**	53.4	54.5	49.7	55.6	55.7
+SFT	63.5	59.7	62.5	62.5	62.9	62.2	100.0	97.0	95.8	88.8	91.2	94.6	77.6	77.0	78.4	77.4	77.4	77.6	54.9	52.5	51.9	52.3	52.5	52.8
+SeqKD	63.9	58.5	62.5	61.1	57.7	60.7	100.0	91.6	66.9	50.5	54.5	72.7	80.2	77.0	80.2	78.8	75.6	78.4	56.1	54.5	55.3	54.5	54.9	55.1
+Pos2Distill	69.9	**67.3**	**68.1**	**69.1**	**67.5**	**68.4**	99.8	**97.3**	**96.5**	**97.5**	**93.2**	**96.9**	**82.6**	**79.8**	**79.0**	**80.6**	**78.1**	**80.0**	**64.9**	**61.5**	**61.5**	**61.3**	**61.8**	**62.2**
LLAMA-3-8B	**67.9**	56.7	53.7	57.9	60.8	59.4	98.0	85.4	70.3	83.2	68.5	81.1	85.6	83.4	82.2	84.0	83.2	83.7	**57.9**	51.8	50.7	50.7	52.3	52.8
+Ms-PoE	65.7	58.5	57.3	58.2	62.5	60.4	97.5	87.3	78.3	81.2	73.4	83.5	**86.4**	84.2	81.2	81.5	82.7	83.2	56.2	52.3	52.1	49.8	52.3	52.5
+SFT	65.1	60.7	62.1	65.3	66.7	63.9	98.6	93.0	97.0	98.0	96.8	96.7	82.8	83.4	**84.8**	83.2	83.2	83.5	56.8	54.7	55.1	55.3	54.5	55.3
+SeqKD	61.7	58.7	58.9	59.9	61.9	60.2	**100.0**	95.6	95.2	**98.2**	97.6	97.3	82.4	**84.6**	83.6	**84.2**	82.6	83.5	54.5	53.7	52.9	51.9	52.5	53.1
+Pos2Distill	67.7	**64.1**	**68.3**	**66.7**	**68.1**	**67.0**	98.8	**96.2**	**98.2**	97.0	**97.8**	**97.6**	85.6	84.2	84.0	84.1	**83.8**	**84.3**	57.7	**56.3**	**57.1**	**56.3**	**56.2**	**56.7**

Table 1: Main results of Pos2Distill-R[1] on both Retrieval-Augmented QA datasets and KV retrieval (140 KV pairs).

		Retrieval-Augmented QA							
Num.	Meth.	0%	25%	50%	75%	100%	Avg.↑	GAP.↓	LEN.
20	BASE.	72.3	63.3	61.2	63.5	65.1	65.1	11.1	3.3k
	ours	72.3	69.5	67.5	68.5	69.7	**69.5**	**4.8**	
30	BASE.	73.7	59.5	60.5	61.3	64.1	63.8	14.2	4k
	ours.	70.3	67.9	69.3	67.3	70.9	**69.1**	**3.6**	
40	BASE.	72.7	60.1	61.3	61.7	65.7	64.3	12.6	6k
	ours.	67.1	66.7	68.3	66.5	68.3	**67.4**	**1.8**	
50	BASE.	74.7	56.9	58.	59.5	66.3	63.1	17.8	8k
	ours.	69.1	66.1	66.7	66.1	67.9	**67.2**	**3.0**	

Table 2: Generalization of Pos2Distill-R[1] on longer context. (Mistral-7B-v0.3 trained on 20 docs).



(a) Mistral-7B (b) Qwen1.5-7B

Figure 5: Ablation for Pos2Distill-R[1]. Each point shows mean accuracy, averaged across gold document positions 1,5,10,15,20. The x-axis represents the total size of training data, which increases with the number of questions, while $\mathcal{K}$ is fixed. "KL" refers to word-level KD; adaptive refers to our approach.

Setup for Pos2Distill-R[2]. For in-context reasoning, we utilize two capable LLMs: Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. The evaluation is performed on three long-context multi-hop reasoning datasets: Hotpot QA (Yang et al., 2018), MusiQue (Trivedi et al., 2022) and 2WikiMultiHopQA (Ho et al., 2020). (All details in Appx. C)

- **Baselines:** We compare two self-training baselines utilizing CoT data: SEALONG (Li et al., 2024a) and LONGFAITH-SFT and LONGFAITH-DPO (Yang et al., 2025a)[4].
- **Evaluation:** We evaluate from two aspects: (1) performance gains in long-context reasoning, and (2) performance gap induced by the positions.

5.2 Main Results for Pos2Distill-R[1]

Pos2Distill-R[1] obviously mitigates PB. Tab. 1 summarizes the performance of various methods across different benchmarks. Our analysis yields two key findings: First, Pos2Distill-R[1] demonstrates robust and uniform performance irrespective of the position of d_{gold}, markedly reducing position-induced performance disparities. For example, on WebQ dataset, Pos2Distill-R[1] enables Llama-3-8B to achieve an average accuracy of 56.7% across 20 positions. This performance, comparable to 57.9% attained when d_{gold} is situated at an optimal sink position, illustrates successful knowledge transfer from advantageous to unfavorable positions, the core principle of Pos2Distill-R[1]. Notably, our method significantly outperforms standard SFT, which suggests that the closer distributions of $\mathcal{R}^{adv}$ and $\mathcal{R}^{trivial}$ allows LLMs to adapt more readily than when learning from markedly different distributions, leading to the high data efficiency detailed in Sec. 5.2.2. Furthermore, Pos2Distill-R[1] effectively generalizes to longer contexts. When evaluated with contexts containing 20 to 50 documents, Mistral-7B-v0.3 trained on 20-document contexts maintains both high overall accuracy and positional uniformity in Tab. 2. Crucially, it exhibits significantly narrowed performance gaps across positions.

5.2.1 Ablation for Pos2Distill-R[1]

To verify the effectiveness of our design, we conducted comprehensive ablation studies in Tab. 3. Our findings yield several important insights:

[4]See more illustrations about baselines in Appendix C

Method	1	5	10	15	20	AVG.
Base	73.6	57.3	56.9	57.3	60.9	61.2
SeqKD	61.7	58.7	58.9	59.9	61.9	60.2
KL	61.7	64.3	65.6	66.4	67.6	65.1
KL+Align	64.5	66.5	67.8	67.3	67.4	66.7
KL+Align+Anc.	69.9	67.3	68.1	69.1	67.5	68.4

Table 3: Ablation study of two core modules for Qwen1.5-7B-Chat: the adaptive strategy targeting trivial positions and the anchoring strategy for advantageous positions. The hyperparameter λ is set to 1.

KL is effective for token-shifting correction. When trained on identical data compositions, the inferior performance of SeqKD highlights the shortcomings of hard-label supervision in scenarios involving token-shifting. In contrast, KL offer a superior mechanism for correcting such shifts compared to the rigid guidance This property makes KL loss particularly well-suited for token-recovery.

Position-aware alignment ensures balance and better learning. Integrating our Position-Aware Alignment strategy with KL divergence leads to significantly more balanced and robust performance, increasing the average score to 66.7.

Anchoring strategy reinforces robustness via key position focus. Incorporating anchoring not only addresses attention dilution at sink positions, but also yields performance gains across other positions. With an average score of 68.4, the strategy is particularly effective at position 1 while maintaining strong performance at trivial positions.

5.2.2 Analysis Results

We investigate the property of Pos2Distill-R^1, presenting comprehensive results in Fig. 5. **High Data Efficiency.** As shown in Fig. 5, our positional awareness metrics achieve superior performance with minimal training data (e.g., Mistral-7B achieves 70.2% accuracy with just 250 examples) and consistently outperform other strategies as the dataset grows. This highlights the data efficiency of our approach. As discussed before, this efficiency stems from the distribution similarity in responses, enabling rapid adaptation to in-domain data, avoiding reliance on entirely out-of-domain samples.

The Impact of $\mathcal{K}$. We conduct studies on $\mathcal{K}$ by varying its value from 1 to 10 while keeping other settings fixed. Increasing $\mathcal{K}$ from 0 to 6 improves LLM performance by 4.6% improvement in LLM performance, but further increase to 10 shows marginal gains with performance saturation. To balance effectiveness and computational efficiency, we select $\mathcal{K} = 4 \sim 6$ as the optimal configuration.

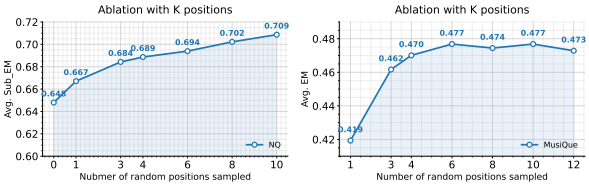


Figure 6: Ablation for the number of sampled positions $\mathcal{K}$ for Pos2Distill-R^1 (left) and Pos2Distill-R^2 (right).

Mechanistic Insights. Since PB emerges from the architectures and parameters of LLMs, we seek to uncover the internal model dynamics following Pos2Distill-R^1 and provide an interpretable explanation. We record the attention distribution over 20 documents as d_{gold} moves from 1 to 20 in Fig. 7. Pos2Distill-R^1 strengthens contextual fidelity by dynamically shifting the focus of attention to consistently align with the relevant document, thereby facilitating more accurate retrieval.

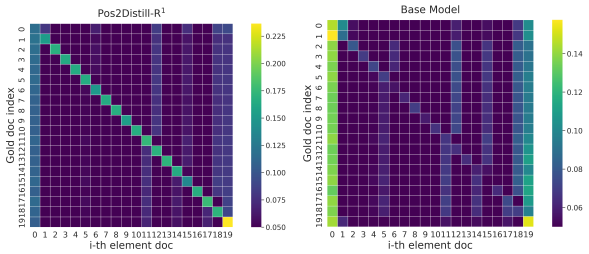


Figure 7: Attention distribution of each doc across total 20 docs, as the position of d_{gold} varies from 1 to 20.

5.3 Main Results for Pos2Distill-R^2

Pos2Distill-R^2 strengthens in-context reasoning. Pos2Distill-R^2 surpasses existing self-training approaches in both in-domain performance and out-of-domain generalization. As detailed in Tab. 4, when trained on the MusiQue dataset, Pos2Distill-R^2 achieves an Exact Match (EM) score of 42.8, outperforming all leading baselines. Furthermore, our method exhibits robust cross-domain generalization; for instance, on the HotpotQA dataset, it attains an EM score of 58.3, compared to 50.9 from the strongest baseline. Our findings suggest that *training LLMs to reason across diverse, scattered gold positions potentially enhances their long-context reasoning more effectively than conventional instance-by-instance training.* This insight offers a new perspective for improving reasoning capabilities in complex, long-context tasks.

To assess PB within the reasoning paradigm, we evaluate performance on two-hop data from

	MuSiQue				2WikiMultiHopQA			HotpotQA		
LLAMA-3.1-8B	Overall	2-Hop	3-Hop	4-Hop	Overall	2-Hop	4-Hop	Overall	Bridge	Comparison
+CoT	11.8	11.9	11.9	7.7	27.4	23.4	40.8	19.4	19.3	23.7
+CoC	13.2	13.3	14.7	11.1	33.8	38.0	50.5	23.2	21.8	29.5
+SEALONG	25.5	31.6	24.5	15.3	52.6	48.6	66.3	49.6	47.7	56.7
+LONGFAITH-SFT (CoT)	39.9	43.5	37.5	29.1	55.0	49.6	72.9	49.8	50.5	46.7
+LONGFAITH-SFT (CoC)	39.6	44.4	38.3	29.9	56.6	50.9	75.1	50.9	51.3	50.7
+LONGFAITH-DPO	34.2	38.8	32.9	25.8	51.2	48.8	59.8	49.7	49.8	42.2
+POS2DISTILL	**42.8**$_{+2.9}$	**47.4**$_{+3.0}$	**38.4**$_{+0.1}$	**31.6**$_{+1.7}$	**61.8**$_{+5.2}$	**57.0**$_{+6.1}$	**78.0**$_{+2.9}$	**58.3**$_{+7.4}$	**56.5**$_{+5.2}$	**65.8**$_{+9.1}$
QWEN2.5-7B	Overall	2-Hop	3-Hop	4-Hop	Overall	2-Hop	4-Hop	Overall	Bridge	Comparison
+CoT	28.7	30.4	29.1	25.2	49.5	41.9	76.6	52.1	49.1	62.5
+CoC	25.9	26.3	28.6	23.7	45.9	38.0	74.3	47.3	43.7	60.6
+SEALONG	33.0	34.5	31.1	28.6	47.6	42.3	68.0	48.6	48.6	42.1
+LONGFAITH-SFT(CoT)	43.3	45.9	41.1	37.0	51.1	46.2	71.2	53.6	55.7	46.0
+LONGFAITH-SFT(CoC)	42.5	43.3	41.1	38.0	51.1	46.6	68.9	53.6	55.4	47.5
+LONGFAITH-DPO	38.4	39.9	35.7	31.9	59.6	52.5	85.0	56.5	54.2	67.4
+POS2DISTILL	**46.2**$_{+2.9}$	**47.6**$_{+1.7}$	**43.5**$_{+2.4}$	**39.2**$_{+1.2}$	**63.4**$_{+3.8}$	**56.7**$_{+4.2}$	76.5	**58.7**$_{+5.5}$	**61.2**$_{+7.0}$	67.2

Table 4: Main experiment results on long-context multi-hop reasoning datasets using the **EM** metric. The training set has 2K samples. The dataset of PosDistill consists 500 questions, setting $\mathcal{K}$ as 4. CoC refers an effective prompting strategy named *Chain-of-Citation* (Li et al., 2024b).

	Connected					Disconnected					Reversed					
POSITIONS	[0,1]	[5,6]	[12,13]	[17,18]	Avg.↑	[0,8]	[5,13]	[6,14]	[8,16]	Avg.↑	[8,0]	[13,5]	[14,6]	[16,8]	Avg.	GAP. ↓
QWEN2.5-7B	37.4	33.2	32.9	38.4	**35.5**	34.8	26.8	28.2	27.7	**29.4**	33.3	30.2	29.6	29.8	**30.7**	11.6
+LONGFAITH-SFT	51.0	46.6	46.7	50.1	**48.6**	47.1	44.3	44.2	43.8	**44.9**	45.4	42.6	41.7	43.0	**43.2**	9.3$_{-2.3}$
+LONGFAITH-DPO	39.0	31.1	35.5	40.5	**36.5**	37.3	32.6	31.5	30.4	**33.0**	39.0	31.1	35.5	40.5	**36.5**	10.1$_{-1.5}$
+POS2DISTILL	50.6	49.8	49.1	51.3	**50.3**	48.5	48.1	47.3	46.7	**47.4**	47.7	46.5	47.1	47.3	**47.2**	4.8$_{-6.8}$
LLAMA3.1-8B	14.3	13.4	15.4	16.0	**14.8**	15.8	15.5	13.7	14.5	**14.9**	14.3	11.6	12.9	12.3	**12.8**	4.4
+LONGFAITH-SFT	45.4	44.4	44.9	47.4	**45.5**	44.3	45.2	43.5	42.7	**43.9**	42.9	42.7	42.9	41.5	**42.5**	5.9$_{+1.5}$
+LONGFAITH-DPO	37.4	38.4	40.7	40.1	**39.2**	37.7	37.4	36.6	36.2	**37.0**	34.8	34.2	36.4	36.0	**35.4**	6.5$_{+2.1}$
+POS2DISTILL	47.2	48.4	47.8	49.4	**48.2**	47.5	49.0	47.2	46.6	**47.6**	45.6	46.0	46.1	47.0	**46.2**	3.4$_{-1.0}$

Table 5: PB assessment for Pos2Distill-R^2 on two-hop data from MusiQue.

MusiQue, considering three relative position configurations for the two gold documents: (i) *connected* (hops are adjacent), (ii) *disconnected* (distracting content separates the hops), and (iii) *reversed* (hops are logically inverted and disconnected). Illustratively, the accuracy of Qwen2.5-7B peaks at 38.4% in the connected mode (hops at positions 17 and 18) but declines to 26.8% in the disconnected mode (hops at positions 5 and 13), exhibiting a obvious performance gap 11.6%. As detailed in Tab. 5, our method not only effectively mitigates the performance gaps to 4.1% across these three modes but also enhances overall reasoning performance irrespective of hop configuration. Conversely, conventional self-training methods, which typically rely on instance-by-instance learning, struggle to eliminate these inter-mode disparities and may even exacerbate this trend, which underscores a fundamental limitation of their training paradigm in handling such PB.

5.3.1 Analysis for Pos2Distill-R^2

The Impact of $\mathcal{K}$ on Pos2Distill-R^2. In Fig. 6 (right), a similar trend to Pos2Distill-R^1 is observed: initially increasing $\mathcal{K}$ leads to a notable improvement in EM scores, but subsequently, but further increases result in diminishing returns, with performance eventually reaching a saturation point. Therefore, an optimal range for $\mathcal{K}$ exists, beyond which further increases yield marginal benefits.

5.4 In-depth Exploration

	Retrival		Reasoning	
Task	**NQ**	**WebQ**	**MusiQue**	**HotpotQA**
BASE	62.5	63.9	41.8	66.7
Pos2Distill-R^1	74.3$_{+11.8}$	68.1$_{+4.2}$	45.1$_{+3.3}$	69.2$_{+2.5}$
Pos2Distill-R^2	64.1$_{+1.6}$	66.7$_{+2.8}$	48.9$_{+7.1}$	72.3$_{+5.6}$

Table 6: Performance of Qwen1.5-7B fine-tuned with Pos2Distill-R^1 versus with Pos2Distill-R^2 on retrieval and reasoning tasks, utilizing the same size of data.

Discussion on Two Systems. As presented in Tab. 6, both systems exhibit notable generalization to their mutual tasks. Specifically, Pos2Distill-R^1, primarily optimized for retrieval, demonstrates that its enhanced contextual retrieval capabilities also improve reasoning over long contexts, yielding a

Model	NQ								Webq							
	0	5	10	15	20	25	30	Gap. ↓	0	5	10	15	20	25	30	Gap. ↓
Qwen2.5-14B	72.8	68.9	64.7	67.5	66.5	64.7	69.7	8.1	64.2	59.6	59.2	60.8	59.2	57.0	59.1	7.2
+Pos2Distill-R^1	71.5	68.9	70.5	71.1	71.6	72.1	72.3	**3.4**	62.0	64.4	62.2	62.4	62.6	63.8	63.2	**2.4**
Qwen2.5 32B	70.3	65.1	65.5	65.9	64.3	66.1	71.3	7.0	64.2	59.6	59.2	60.8	59.2	57.0	59.0	7.2
+Pos2Distill-R^1	71.5	67.7	70.3	70.1	70.9	70.7	70.7	**3.8**	63.2	63.2	61.6	60.6	61.2	62.2	62.8	**2.6**

Table 7: Generalization results on model size for Pos2Distill-R^1.

Musique	Connected				Disconnected				Reversed				Gap ↓
	[0,1]	[5,6]	[12,13]	[17,18]	[0,8]	[5,13]	[6,14]	[8,16]	[8,0]	[13,5]	[14,6]	[16,8]	
Qwen2.5 14B	56.6	54.1	54.7	59.5	55.0	49.5	50.0	51.8	57.7	51.4	52.0	51.9	10.0
+PosDistill R^2	60.1	58.4	60.9	63.2	59.5	56.7	57.7	58.1	59.0	58.6	56.7	56.5	**6.7**
Qwen2.5 32B	61.7	59.8	59.1	63.2	59.4	54.7	54.2	54.7	60.3	56.6	55.4	57.8	9.0
+PosDistill R^2	64.2	65.2	63.1	65.7	63.4	61.8	60.8	61.0	62.8	62.9	61.8	62.4	**4.9**

Table 8: Generalization results on model size for Pos2Distill-R^2.

3.3% increase on the MusiQue task. Conversely, Pos2Distill-R^2, optimized for reasoning, shows that its acquired proficiency in reasoning over long context also bolsters contextual awareness, thereby benefiting retrieval performance. Despite this cross-task generalization, each system excels in its primary domain: Pos2Distill-R^2 achieves superior performance on complex long-context reasoning tasks where Pos2Distill-R^1 lags, and vice versa for retrieval. This suggests distinct underlying dynamics for mitigating PB, potentially influenced by the presence or absence of CoT. Consequently, the development of these two specialized Pos2Distill designs proves both necessary and effective.

Generalization to Larger LLMs To further ascertain the robustness and broad applicability of our findings, we extended our investigations to larger-scale LLMs. The comprehensive evaluations presented in Tab. 7 and 8, conducted across 30 docs, consistently reveal that even significantly larger models exhibit a pronounced prevalence of PB. This critical observation underscores the universal nature and persistence of PB across model scales. Notably, our proposed method, Pos2Distill-R^1 and -R^2, proved remarkable effectiveness in mitigating PB within both 14 and 32B. Specifically, for the Qwen2.5-32B, Pos2Distill-R^1 significantly reduced the performance gap: decreasing it from 7.2% to 2.6% in retrieval tasks and from 9% to 4.9% in reasoning tasks. These compelling quantitative results affirm the scalability and efficacy of our method when applied to larger LLMs.

6 Conclusion

This work introduces a novel paradigm to mitigate PB by leveraging the performance disparity it creates. Specifically, our method distills knowledge from privileged positions to unfavored ones, thereby reducing the disparty induced by PB. PB manifests as token-shifting in retrieval and as thought-shifting in reasoning. To address distinct facets of PB dynamics, we develop two specialized frameworks: Pos2Distill-R^1 and Pos2Distill-R^2. Extensive experiments validate the efficacy of our approach in reducing PB and robust generalization in both in-context retrieval and reasoning tasks.

Limitations

While our proposed methods demonstrate substantial improvements in performance and data efficiency, we acknowledge certain limitations exist. Specifically, for Pos2Distill-R^2, there is scope for further refinement. The current design, while effective, could benefit from more granular mechanisms to precisely calibrate the mitigation of PB. For instance, future work could explore adaptive strategies that adjust the positional distillation process based on the complexity of the reasoning chain or the specific configuration of supporting documents. Such enhancements might lead to even more nuanced control over positional influences in complex reasoning scenarios.

Potential Social Impacts

Enhancing positional robustness in large language models (LLMs) fosters more reliable, fair, and consistent information processing, especially in scenarios requiring long-context retrieval and reasoning. By mitigating PB, our approach encourages equitable model behavior and reduces spurious disparities in output quality that could disadvantage critical content occurring in less prominent positions. These improvements are especially significant for real-world applications where fair and accurate comprehension of lengthy documents is essential, such as in education, law, healthcare, and scientific research. In these settings, a focus on understanding and reasoning supports the development of more inclusive and trustworthy AI systems, enabling better information access and more dependable, model-assisted decision making. Ultimately, our work advances large language models as robust and ethical tools for the benefit of society.

Acknowledgement

This work was supported in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA0480301.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie C.Y. Chan, Biao Zhang, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. In *ICML 2024 Workshop on In-Context Learning*.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your LLM fully utilize the context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024a. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa. *arXiv preprint arXiv:2412.10079*.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024b. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. 2024. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11160–11174, Bangkok, Thailand. Association for Computational Linguistics.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander Rudnicky, and Peter Ramadge. 2023a. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1183–1193, Toronto, Canada. Association for Computational Linguistics.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. 2023b. Dissecting transformer length extrapolation via the lens of receptive field analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13522–13537, Toronto, Canada. Association for Computational Linguistics.

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

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. 2025a. Mmdocir: Benchmarking multi-modal retrieval for long documents. *arXiv preprint arXiv:2501.08828*.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025b. Benchmarking retrieval-augmented multimomal generation for document question answering. *arXiv preprint arXiv:2505.16470*.

Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. 2025. Attentionrag: Attention-guided context pruning in retrieval-augmented generation. *arXiv preprint arXiv:2503.10720*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhenyu He, Guhao Feng, Shengjie Luo, Kai Yang, Liwei Wang, Jingjing Xu, Zhi Zhang, Hongxia Yang, and Di He. 2024. Two stones hit one bird: Bilevel positional encoding for better length extrapolation. In *International Conference on Machine Learning*, pages 17858–17876. PMLR.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. BABILong: Testing the limits of LLMs with long context reasoning-in-a-haystack. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024b. Making long-context language models better multi-hop reasoners. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024c. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA. Association for Computational Linguistics.

Hongzhan Lin, Ang Lv, Yang Song, Hengshu Zhu, Rui Yan, et al. 2024. Mixture of in-context experts enhance llms' long context awareness. *Advances in Neural Information Processing Systems*, 37:79573–79596.

Aofan Liu, Shiyuan SONG, haoxuan li, Cehao Yang, and Yiyan Qi. 2025. Beyond function-level search: Repository-aware dual-encoder code retrieval with adversarial verification. In *Knowledgeable Foundation Models at ACL 2025*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Han Peng, Jinhao Jiang, Zican Dong, Wayne Xin Zhao, and Lei Fang. 2025. Cafe: Retrieval head-based coarse-to-fine information seeking to enhance multi-document qa capability. *arXiv preprint arXiv:2505.10063*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

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*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Xuan Ren, Qi Chen, and Lingqiao Liu. 2025. Efficient response generation strategy selection for fine-tuning large language models through self-aligned perplexity. *arXiv preprint arXiv:2502.11779*.

Xuan Ren, Biao Wu, and Lingqiao Liu. 2024. I learn better if you speak my language: Understanding the superior performance of fine-tuning large language models with llm-generated responses. *arXiv preprint arXiv:2402.11192*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Tao Tan, Yining Qian, Ang Lv, Hongzhan Lin, Songhao Wu, yongbo wang, Feng Wang, Jingtong Wu, xin lu, and Rui Yan. 2025. PEAR: Position-embedding-agnostic attention re-weighting enhances retrieval-augmented generation with zero inference overhead. In *THE WEB CONFERENCE 2025*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. 2024a. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14024–14040, Bangkok, Thailand. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. 2025a. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024c. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2025b. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Feng Xiong, Hongling Xu, Yifei Wang, Runxi Cheng, Yong Wang, and Xiangxiang Chu. 2025. Hs-star: Hierarchical sampling for self-taught reasoners via difficulty estimation and budget reallocation. *arXiv preprint arXiv:2505.19866*.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025a. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Wang, and Liangming Pan. 2025b. How is llm reasoning distracted by irrelevant context? an analysis using a controlled benchmark.

Zhaohui Yang, Chenghua He, Xiaowen Shi, Linjing Li, Qiyue Yin, Shihong Deng, and Daxin Jiang. 2025c. Beyond the first error: Process reward models for reflective mathematical reasoning. *arXiv preprint arXiv:2505.14391*.

Zhaohui Yang, Yuxiao Ye, Shilei Jiang, Chen Hu, Linjing Li, Shihong Deng, and Daxin Jiang. 2025d. Unearthing gems from stones: Policy optimization with negative sample augmentation for llm reasoning. *arXiv preprint arXiv:2505.14403*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiangyu Yin, Chuqiao Shi, Yimo Han, and Yi Jiang. 2024. Pear: a robust and flexible automation framework for ptychography enabled by multiple large language model agents. *arXiv preprint arXiv:2410.09034*.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2024. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024a. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

Zheng Zhang, Fan Yang, Ziyan Jiang, Zheng Chen, Zhengyang Zhao, Chengyuan Ma, Liang Zhao, and Yang Liu. 2024b. Position-aware parameter efficient fine-tuning approach for reducing positional bias in llms.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024c. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jie Zhao, Wanting Ning, Yuxiao Fei, Yubo Feng, and Lishuang Li. 2025. Gdllm: A global distance-aware modeling approach based on large language models for event temporal relation extraction. *arXiv preprint arXiv:2508.20828*.

A Related Work

On the Emergence of PB. This section elucidates the fundamental mechanisms governing position bias. Position bias describes the tendency to prioritize information differentially based on its position within an input. A notable example of this effect is the "lost-in-the-middle" phenomenon, characterized by a pronounced performance gap depending on the placement of crucial information. Specifically, models tend to achieve higher accuracy when key information is positioned at the sequence boundaries but experience significant performance degradation when it is located in the middle. This pattern emerges despite strong performance at both ends of the sequence, driven by a combination of primary bias, which favors early content, and recency bias, which enhances reliance on recent information.

Primary Bias. The enhanced utilization of initial context can be attributed to a universal phenomenon known as "attention sinks". This concept refers to the inherent tendency of models to allocate substantial attention to initial tokens, regardless of their semantic relevance. Consequently, LLMs often give disproportionate emphasis to the first few tokens in a sequence. The emergence of attention sinks can be traced back to the intricate interplay of pre-training dynamics, encompassing factors such as optimization processes, data distribution, and the model's loss function (Gu et al., 2025).

$$\textbf{RoPE:} \quad \mathbf{q}_m = \mathcal{R}(\mathbf{q}, m), \quad \mathbf{k}_n = \mathcal{R}(\mathbf{k}, n)$$
$$\mathbf{q}_m \mathbf{k}_n^{\mathrm{T}} = \mathcal{G}(\mathbf{q}, \mathbf{k}, m-n)$$
$$\textbf{Attention:} \quad \mathbf{a}_{m,n} = \mathrm{Softmax}\left(\frac{\mathbf{q}_m \mathbf{k}_n^{\mathrm{T}} + \mathbf{mask}}{\sqrt{d}}\right)$$

Recency Bias. The preferential attention allocation to terminal positions arises from two key architectural components: **Causal Mask** and **Relative Position Encoding**. First, the causal mask enforces a unidirectional flow of information ($m > n$), implicitly embedding positional information and producing context-dependent token embeddings that vary with sequential permutations. Simultaneously, Rotary Position Embeddings (RoPE) encode relative positional relationships ($m - n$ in equations) within attention calculations, inherently biasing the model toward recent tokens. While the intricate interplay between these components warrants further investigation, this preliminary analysis highlights the emergence of asymmetric attention patterns, providing insights that inform both the design and broader understanding of the framework.

The advent of the long-context era for Large Language Models (LLMs) has been notably advanced by progress in two pivotal directions: (1) efficient attention mechanisms such as FlashAttention (Dao et al., 2022; Ainslie et al., 2023), which drastically reduce the computational overhead of processing extended sequences, and (2) length extrapolation techniques (He et al., 2024; Chi et al., 2023b; Press et al., 2021), which enable LLMs to generalize beyond their training context length (Su et al., 2024; Raffel et al., 2020). Collectively, these breakthroughs empower LLMs to perform complex question answering over much larger context windows (Agarwal et al., 2024), representing a significant step toward more capable and flexible natural language systems (Yang et al., 2025d).

Mechanistic Approaches. For instance, to alleviate intrinsic long-range decay (Zhao et al., 2025), Zhang et al. propose *Ms-PoE*, which assigns distinct rescaling factors to each attention head, compressing relative distance $m - n$ by a factor of $1/n$. The work *Attention Buckets* (Chen et al., 2024) exploits the approximate periodicity observed in the attention waveform at distal positions, shifting key information away from the waveform's trough regions. Furthermore, *MoICE* (Lin et al., 2024) designs a router within each attention head to dynamically select among multiple RoPE angles, effectively avoiding trough zones during generation. Yu et al. reveals that PB is reflected in positional hidden states, and mitigates this by scaling certain dimensions of these representations. However, these component-level modifications without continual training present two limitations: (1) obtaining optimal hyperparameters requires multiple forward passes (Chen et al., 2024); and (2) such interventions disrupts the parallelism of multi-head attention (Zhang et al., 2024c).

Training Approach. Therefore, FILM (An et al., 2024) systematically synthesizes long-context QA datasets to promote fine-grained information awareness, encompassing a range of in-context retrieval and reasoning tasks over varying context lengths. Although such training allows LLMs to better utilize extended contexts, it incurs considerable computational overhead. Zhang et al. (2024b) introduce a position-aware adapter module to remove intrinsic preference bias. However, this approach is limited to single-segment retrieval and does not gen-

eralize well to more complex scenarios involving integration and reasoning over multiple segments.

B Preliminary Results

B.1 Token Shifting Phenomenon

Gold index 1: The first orca to be put in captivity was Wanda, a North Eastern Pacific killer whale, which was captured in November 1961 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 2: The first orca to be put in captivity was Wanda, a North Eastern Pacific killer whale, who was captured in November 1961 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 3: The first orca to be put in captivity was Wanda, a North Eastern Pacific killer whale, who was captured in November 1961 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 4: The first orca to be put in captivity was Wanda, a North Eastern Pacific killer whale, who was captured in November 1961 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 5: The first orca to be put in captivity was Moby Doll, a male orca captured in November 1961 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 6: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by Ted Griffin. Moby Doll was held at the Vancouver Aquarium for three months before dying.
Gold index 7: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by Ted Griffin. Moby Doll was held at the Vancouver Aquarium for three months before dying.
Gold index 8: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by Ted Griffin. Moby Doll was held at the Vancouver Aquarium for three months before dying.
Gold index 9: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 10: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by a collecting crew from Marineland of the Pacific in Los Angeles. He was placed in a tank at the aquarium.
Gold index 11: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 12: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 13: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 and displayed at the Vancouver Aquarium for three months before dying.
Gold index 14: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 by a collecting crew from Marineland of the Pacific in Los Angeles.
Gold index 15: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 and displayed at the Vancouver Aquarium in British Columbia.
Gold index 16: The first orca to be put in captivity was Moby Doll, who was captured in 1964 and held at the Vancouver Aquarium for three months before dying.
Gold index 17: The first orca to be put in captivity was Moby Doll, who was captured in 1964 and displayed at the Vancouver Aquarium for three months.
Gold index 18: The first orca to be put in captivity was Moby Doll, a male orca captured in 1964 and displayed at the Vancouver Aquarium in British Columbia.
Gold index 19: The first orca to be put in captivity was Moby Doll, who was captured in 1964 and displayed at the Vancouver Aquarium in British Columbia.
Gold index 20: The first orca to be put in captivity was Moby Doll, who was captured in 1964 and displayed at the Vancouver Aquarium for three months.

Phenomena Related to PB To verify the phenomenon mentioned in Sec. 3, we provide more proof related to them: (1) **Token shifting:** We analyze token-level KL divergence between responses generated from trivial and advantageous positions to verify the existence of token-shifting via case study in Tab. 9. Our results on Mistral and Qwen demonstrate that token shifting is indeed a universal phenomenon: KL divergence values at specific token positions is extremely high compared to normal level, indicating that the severe divergence happens at this decoding step, which is named token-

shifting in our paper. According to the table, it is obvious token-shifting only take places on very few tokens; (2) **Token recovery:** Our Pos2Distill-R[1] framework especially focus on these key tokens (those exhibiting higher KL divergence at token-level) and fix them, in order to recover the optimal decoding trajectory like sink positions again, thus mitigating PB. This core concept parallels recent findings in mathematical reasoning tasks (Wang et al., 2025a), where tokens with high entropy are labeled as "forking tokens." Interventions that specifically correct model predictions at these forking tokens have been shown to outperform even full gradient updates; (3) **Compound PB:** This occurs in reasoning tasks in long-context scenarios and involves the interplay of retrieval and actual thinking processes, leading to very sophisticated compound effects. Therefore, instead of just focusing retrieval or reasoning, we choose to reshape CoT process. Therefore, Pos2Distill-R[2] can be seen as the form of Reinforced Self-Training (Gulcehre et al., 2023), Self-Taught Reasoner (Zelikman et al., 2022) or HS-STAR (Xiong et al., 2025).

C Experiments Details.

Implementation Details. All experiments are with the following hyperparameters: (1) learning rate of $\alpha = 3 \times 10^{-6}$; (2) a batch size of $b = 32$; and (3) $n = 2$ training epochs. We employ DeepSpeed ZeRO-3 optimization and FlashAttention (Dao et al., 2022) to accelerate the training process, utilizing the bfloat16 data format. Furthermore, for inference acceleration, we adopt the vLLM (Kwon et al., 2023) framework. For Retrieval, we conducted training with 300 samples at four distinct locations, each randomly selected from the range of 1 to 20. For Reasoning, we conducted training on 500 samples. These samples were distributed across four randomly selected pairs of locations, where each individual location within a pair could range from 1 to 20. For both retrieval and reasoning tasks, we only collected a greedy-search response from anchor positions. Commonly, we set the hyperparameter λ controlling the intensity of the anchoring loss as 1.0. All experiments are conducted on NVIDIA H20 GPUs.

Evaluation Metrics. For retrieval tasks, we adopt the Sub_EM metric, while for reasoning tasks, we use the EM metric. Specifically, in reasoning tasks, we first use regular expressions to identify the tokens where the answer appears, and then perform exact match (EM) evaluation on the subsequent part. We test TQA[5], WebQ [6]and NQ [7]under retrieval-argumented QA settings and the data sources are from Huggingface.

Baselines. For Pos2Distill-R[1], we introduce vanilla SFT and SeqKD as baselines. SFT directly fine-tunes the student model on data supervised by gold responses, whereas SeqKD fine-tunes the student model on data generated by the teacher model. For Pos2Distill R[1], as discussed in related work, mainstream state-of-the-art approaches addressing PB in retrieval tasks include MsPoE (Zhang et al., 2024c), Attention Buckets (Chen et al., 2024), MoICE (Lin et al., 2024), and PEAR (Yin et al., 2024). Although these methods improve performance across positions, PB still persists. All of them are based on mechanistic approaches, and following (Yu et al., 2024), we mainly compare with MsPoE in the main paper. To further demonstrate the effectiveness of our method, we also provide additional comparisons on Llama2-7B-chat-4k 10. Our approach not only better balances performance across positions but also achieves the highest average score of 68.18, which illustrates the advantage compared to previous methods. For Pos2Distill R[2], which follows a self-training paradigm using CoT data distilled from advantageous positions, we consider the most recent and relevant baselines, including Longfaith-SFT (Yang et al., 2025a), longfaith-DPO (Yang et al., 2025a), and Sealong (Li et al., 2024a). These works involves how to generate high-quality and faithful Chain-of-Thought data for self-improvement and achieving good performance on long-context reasoning tasks. Therefore, we adopt these methods as baselines against to Pos2Distill R[2] in our experiments.

Prompt Template The prompt template can be found in Tab. 11. The Retrieval prompt template instructs the LLM to provide a high-quantify answer (likely meaning a high-quality, precise, or well-supported answer) by exclusively using information from provided documents, explicitly noting that some documents might be irrelevant. This template emphasizes factual accuracy and direct extraction from given sources, limiting the model's ability to introduce external knowledge. The Reasoning prompt template outlines a structured, multi-step approach. It first directs the LLM to identify

[5]https://huggingface.co/datasets/vsearch/tqa
[6]https://huggingface.co/datasets/vsearch/webq
[7]https://huggingface.co/datasets/vsearch/nq

Response	Token1	Token2	Token3	Token4	Token5	Token6	Token7	Token8	Token9	Token10
Mistral+NQ	1.6e-03	1.4e-02	5.7e-06	7.7e-05	1.87e-03	**1.4e-01**	1.6e-03	4.8e-06	2.0e-07	2.6e-07
Mistral+Webq	5.7e-03	5.5e-03	5.4e-05	1.0e-05	1.4e-04	3.6e-03	3.3e-06	1.2e-04	**1.7e-02**	4.4e-03
Llama3+NQ	5.1e-03	6.6e-04	6.0e-03	2.2e-04	7.3e-04	1.0e-03	2.9e-04	**1.0e-02**	7.3e-06	5.3e-05
Llama3+Webq	9.6e-03	2.6e-04	9.1e-03	7.7e-08	2.9e-04	1.6e-04	2.1e-06	**1.2e-02**	1.9e-06	2.1e-06

Table 9: Phenomenon related token-shifting.

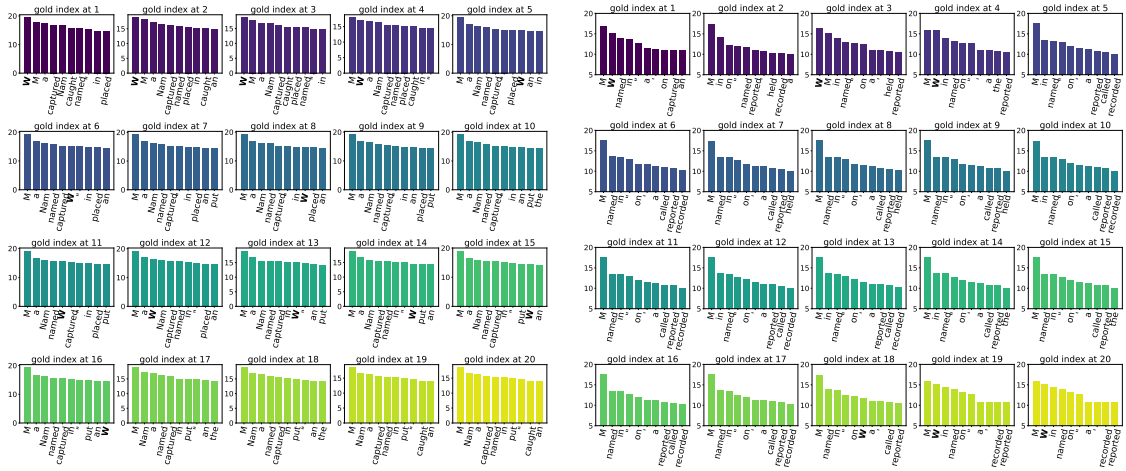


Figure 8: Visualization of the top 10 tokens in the logit distributions at the critical divergence points for Meta-Llama-3-8B-Instruct (left) and Mistral-7B-Instruct-v0.3 (right).

the relevant information from a long context. Next, it requires "step-by-step reasoning based on that information." Finally, it specifies that "The final answer must end with: 'The answer is:'". This template promotes logical deduction, information synthesis, and a clear, conclusive output format (Yang et al., 2025c).

Method	1	3	5	7	10	Avg. ↑	GAP ↓
base	64.14	65.95	64.97	62.67	67.53	65.05	4.86
+ Ms-PoE	66.06	64.29	63.99	62.22	64.75	64.34	3.84
+ AB	66.36	66.14	65.25	63.20	64.93	65.18	3.16
+ MoICE	65.50	66.33	65.61	64.11	65.84	65.48	2.22
+ PEAR	62.71	67.01	68.32	66.44	69.57	66.81	6.86
+ PosDistill R1	**67.27**	**68.46**	**69.06**	**68.66**	**67.47**	**68.18**	**1.79**

Table 10: Comparison of mainstream state-of-the-art approaches addressing PB in retrieval tasks.

D Behavior of PB

For retrieval tasks, we statistically compute the perplexity (PPL) of each response sampled from both the sink and recent positions, conditioned on prompts corresponding to trivial positions (see Fig. 9). Interestingly, the PPL is much lower compared to that of gold labels from SFT, which strongly indicates a high degree of similarity in the response space when conditioned on different permutations of the same set of documents. A closer examination reveals that LLMs primarily diverge at a few decisive tokens, a phenomenon we refer to as token-shifting. At these key generation steps, as shown in the logit distributions in Fig. 8, the correct token still appears among the top-10 predictions. This observation suggests that LLMs possess the potential to recover from initially incorrect decoding paths, and that responses from the sink position serve as natural gold signals for correction. In contrast, the dynamics in reasoning tasks are more complex. As shown by the PPL values in Fig. 9, perplexity increases substantially, indicating that token-shifting is not an appropriate assumption for reasoning scenarios.

E Additional Results

LCLMs with More Documents Under longer-context settings, we conducted additional experiments with Mistral-7B-Instruct trained on 20 and 50-document contexts respectively, and evaluated them on context lengths ranging from 20 (≈4k tokens) to 80 documents (≈14k tokens), as indicated in Tab. 13. **Takeaways:** While training on longer contexts helps further improve the model's retrieval performance, training with relatively shorter contexts can also yield competitive results. Trained on

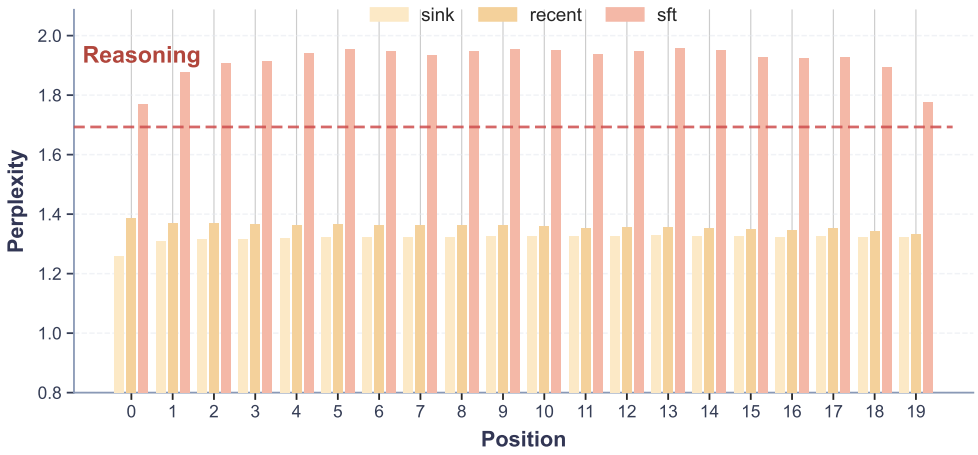


Figure 9: Average PPL over 500 examples for responses collected from sink,recent positions and SFT labels.

Category	Prompt Template
Retrieval	Please write a high-quantify answer for the given question using only the provided search documents (some of which might be irrelevant).
Reasoning	Let's first identify the relevant information from the long context and list it. Then, carry out step-by-step reasoning based on that information, and finally, provide the answer. The final answer must end with "The answer is:".

Table 11: Prompt Templates.

	Connected					**Disconnected**					**Reversed**					
POSITIONS	[0,1]	[5,6]	[12,13]	[17,18]	**Avg.**↑	[0,8]	[5,13]	[6,14]	[8,16]	**Avg.**↑	[8,0]	[13,5]	[14,6]	[16,8]	**Avg.**	**GAP.** ↓
QWEN2.5-3B	31.2	27.5	28.1	30.2	**29.3**	28.7	24.3	22.2	23.4	**24.7**	25.5	24.9	22.6	20.9	**23.5**	10.3
QWEN2.5-14B	56.6	54.1	54.7	59.5	**56.2**	55.0	49.5	50.0	51.8	**51.6**	57.7	51.4	52.0	51.9	**53.3**	10.0
QWEN2.5-32B	61.7	59.8	59.1	63.2	**61.0**	59.4	54.7	54.2	54.7	**55.8**	60.3	56.6	55.4	57.8	**57.5**	9.0

Table 12: PB in reasoning tasks.

50 documents, the model not only generalizes well to fewer documents (achieving high and stable performance, e.g., 77% accuracy on 20 docs), but also generalizes effectively to more documents (maintaining 70% accuracy on 80 docs). In addition, models trained on 20 documents still achieve 67% accuracy when evaluated on 50 docs, exhibiting minimal variance across positions and no significant drop compared to the 69.5% accuracy observed when tested on 20 docs. This strong out-of-domain generalization, even when trained with fewer docs (e.g., 20), further demonstrates the superiority of our approach.

F Statement on AI Usage

In the preparation of this manuscript, ChatGPT by OpenAI was utilized solely for the purpose of language refinement and stylistic enhancement. All scientific ideas, methodologies, analyses, and conclusions presented in this work are entirely the authors' own and were developed independently without reliance on AI-generated content.

Test.	**Train.**	**0%**	**25%**	**50%**	**75%**	**100%**	**AVG.** ↑
20docs	20	72.3	69.5	67.5	68.5	69.7	69.5
	50	**77.8**	76.2	76.6	77.2	77.4	**77.0**
30docs	20	70.3	67.9	69.3	67.3	70.9	69.1
	50	74.9	75.4	75.1	76.2	77.2	**75.8**
40docs	20	67.1	66.7	68.3	66.5	68.3	67.4
	50	74.1	75.2	76.2	77.1	77.6	**75.8**
50docs	20	69.1	66.1	66.7	66.1	67.2	67.0
	50	73.5	73.9	74.5	75.2	74.9	**74.4**
70docs	20	64.9	63.9	63.9	65.1	65.1	64.6
	50	69.3	69.1	71.7	70.1	70.1	**70.1**
80docs	20	64.9	63.9	63.9	65.1	65.1	64.6
	50	68.1	68.3	70.1	70.9	71.9	**69.9**

Table 13: Performance comparison on different document numbers. The column **AVG.** ↑ shows the average score across percentages. Bold values indicate the best result in each group.