

Judge Q: Trainable Queries for Optimized Information Retention in KV Cache Eviction

Yijun Liu, Yixuan Wang, Yuzhuang Xu, Shiyu Ji,
Yang Xu, Qingfu Zhu, Wanxiang Che*

Research Center for Social Computing and Interactive Robotics,
Harbin Institute of Technology, Harbin, China
{yijunliu, car}@ir.hit.edu.cn

Abstract

Large language models (LLMs) utilize key-value (KV) cache to store historical information during sequence processing. The size of KV cache grows linearly as the length of the sequence extends, which seriously affects memory usage and decoding efficiency. Current methods for KV cache eviction typically utilize the last window from the pre-filling phase as queries to compute the KV importance scores for eviction. Although this scheme is simple to implement, it tends to overly focus on local information, potentially leading to the neglect or omission of crucial global information. To mitigate this issue, we propose **Judge Q**, a novel training method which incorporates a soft token list. This method only tunes the model’s embedding layer at a low training cost. By concatenating the soft token list at the end of the input sequence, we train these tokens’ attention map to the original input sequence to align with that of the actual decoded tokens. In this way, the queries corresponding to the soft tokens can effectively capture global information and better evaluate the importance of the keys and values within the KV cache, thus maintaining decoding quality when KV cache is evicted. Under the same eviction budget, our method exhibits less performance degradation compared to existing eviction approaches. We validate our approach through experiments conducted on models such as Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3, using benchmarks including LongBench, RULER, and Needle-in-a-Haystack. Results indicate an improvement of approximately 1 point on the LongBench and over 3 points on RULER. This proposed methodology can be seamlessly integrated into existing open-source models with minimal training overhead, thereby enhancing performance in KV cache eviction scenarios.

Code — <https://github.com/Mambaaaaaaa/Judge-Q>

1 Introduction

Large language models (LLMs) have demonstrated remarkable performance across a wide range of domains, including language modeling (Hendrycks et al. 2020; Suzgun et al. 2022), text comprehension (Rajpurkar et al. 2016; Joshi et al. 2017), and code understanding (Chen et al. 2021; Jain et al. 2024). Currently, many scenarios (Liu et al. 2025a,b;

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

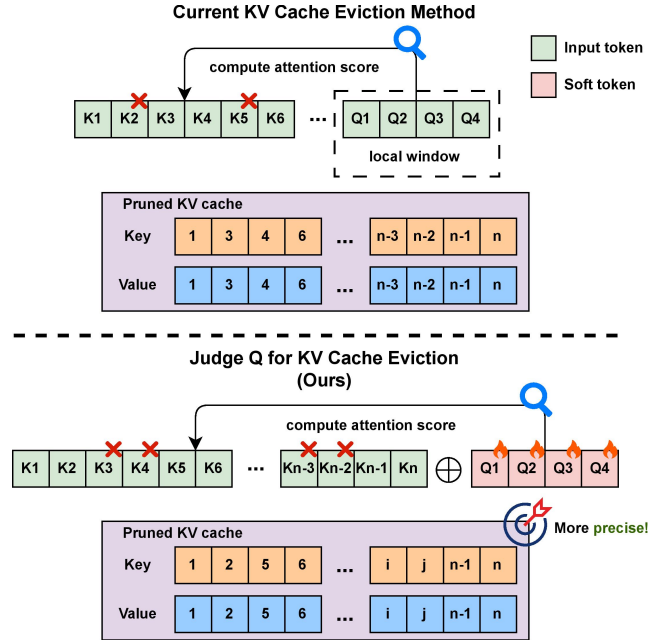


Figure 1: Difference between current KV cache eviction methods and ours when calculating the importance score of key-value pairs during prefilling stage. We utilize trainable soft tokens to effectively leverage global information.

Chen et al. 2025; Lu et al. 2025a,b) involve input sequences with over ten thousand tokens, such as books, code repositories (Kočíský et al. 2018; Zhang et al. 2023a), and dialogue systems with long-term histories (Li et al. 2024a). LLMs employ KV cache to store historical information during the pre-filling stage, and reuse them during decoding. The size of KV cache increases linearly with input sequence length. In long sequence scenarios, KV cache often consumes substantial memory space, posing a significant bottleneck for inference, particularly in resource-limited environments.

To tackle the challenge of excessive memory consumption by KV cache, a series of KV cache eviction methods (Zhang et al. 2023b; Li et al. 2024b; Cai et al. 2024) are proposed from the perspective of directly reducing the elements within KV cache during the pre-filling stage. Li et al. (2024b)

observe that only a small portion of tokens in the prompt have high attention scores which contain crucial information to generate satisfactory responses. These strategies leverage the characteristics of auto-regressive language models and the observation that questions often appear towards the end of inputs. They typically choose queries from the last window during pre-filling to calculate the importance scores of key-value pairs based on attention. By retaining the top-k most important portions and discarding the rest, they effectively manage the size of the KV cache.

This straightforward approach of directly selecting the last window is simple to implement, however, it has limitations. Firstly, it tends to overly focus on local information within the last window while neglecting global content, which can result in potential information loss. Additionally, if the last window does not contain question, the performance of these methods often significantly degrades. In fact, the primary objective of KV cache eviction is to reduce memory usage while maintaining decoding quality. Therefore, identifying the key-value pairs that are crucial for decoding the subsequent tokens is logical and ideal.

In this paper, we discover that directly using the original decoded tokens to select critical key-value pairs preserves high decoding quality, representing the theoretical upper bound of KV cache eviction methods. However, during the prefilling stage, the actual decoded tokens are not available. Inspired by this phenomenon, we propose a training method called **Judge Q**. This approach involves introducing a list of soft tokens into the model’s vocabulary and leveraging them to perform KV cache eviction during the pre-filling stage. Specifically, during the training phase, we concatenate these soft tokens directly to the end of original input sequence. The soft tokens are trained to align the attention map to input sequence with that of the actual decoded tokens, enabling the soft tokens to capture global information effectively. This process incurs minimal training costs, as it involves training only the parameters associated with the soft tokens in the model’s embedding layer, while keeping the rest of the model’s weights frozen. During inference, we similarly append these soft tokens to the end of the original input sequence in the pre-filling stage. The queries corresponding to soft tokens are used to calculate the importance scores of the key-value pairs in KV cache across the entire input sequence. Based on these scores, we retain the top-k key-value pairs with the highest scores and discard the rest. After the pre-filling stage, the soft tokens are also removed. The pruned KV cache is then used for the subsequent decoding process. We conduct experiments using two models, Llama-3.1-8B-Instruct (Dubey et al. 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al. 2023), trained with self-generated data from the ShareGPT dataset (shareAI 2023). Our method is evaluated on LongBench (Bai et al. 2023), RULER (Hsieh et al. 2024), and Needle-in-a-Haystack (Kamradt 2023). Compared to other methods performing KV cache eviction during the pre-filling stage, our approach consistently achieves superior results under the same budget and more closely approaches the theoretical upper bound of using actual decoded tokens.

The main contributions of this paper are as follows:

- We propose Judge Q, a novel training method that alleviates the limitations of existing KV cache eviction methods, which tend to focus excessively on local information while overlooking global content during pre-filling.
- The proposed method can be adapted to any open-source model with minimal training costs. By fine-tuning only the embedding layer of the model, it enhances the model’s performance in KV cache eviction scenarios.
- Experimental results on three benchmarks demonstrate that our method outperforms existing KV cache eviction methods during the pre-filling stage under the same budget, with significant gains in low-resource settings. Notably, our approach achieves nearly 1 point improvement on LongBench and over 3 points on RULER.

2 Related Works

2.1 KV Cache Eviction

To address the challenge of reducing the memory footprint of KV cache, a series of methods are proposed from the perspective of reducing the number of elements in KV cache. These strategies can effectively control KV cache size by evaluating the importance of key-value pairs within the input sequence during pre-filling, retaining only those with high importance scores and discarding less critical ones. This approach is grounded in the observation of the sparsity of attention (Zhang et al. 2023b) and that only a small portion of tokens convey essential information for response generation during the pre-filling stage (Li et al. 2024b). Most of these methods compute the cumulative attention scores of key-value pairs with respect to queries within the local window during the prefilling stage, which are used as the corresponding importance scores for eviction. Methods like H2O (Zhang et al. 2023b) and Scissorhands (Liu et al. 2023) perform dynamic KV cache eviction, while approaches (Li et al. 2024b; Chen et al. 2024; He et al. 2025) conduct a static eviction. Additionally, some work (Feng et al. 2024; Cai et al. 2024; Yang et al. 2024a) focuses on dynamically allocating budgets to optimize memory usage. Although these methods achieve effective KV cache eviction during pre-filling stage. As shown in Fig 1, their directly relying on the queries of local window to score the importance of key-value pairs often results in an overemphasis on local information while neglecting global content, thereby limiting performance. Therefore, it is essential to develop a KV cache eviction method that can retain more information during the prefilling stage to support subsequent generation effectively.

Besides the mainstream KV cache eviction methods mentioned above, there are also other KV cache eviction approaches from different perspectives. Devoto et al. (2024) use the L2 norm of the key vector as the importance score of key-value pair. Ge et al. (2023) adaptively select and apply different eviction strategies for different layers and heads. Lookahead Q-Cache (Wang et al. 2025) utilizes decoded pseudo queries to guide KV cache eviction. And Xiao et al. (2023) propose StreamingLLM, which only retains the tokens from the beginning and the local segments. While Xiao et al. (2024) train the heads of model to selectively utilize streaming attention and full attention.

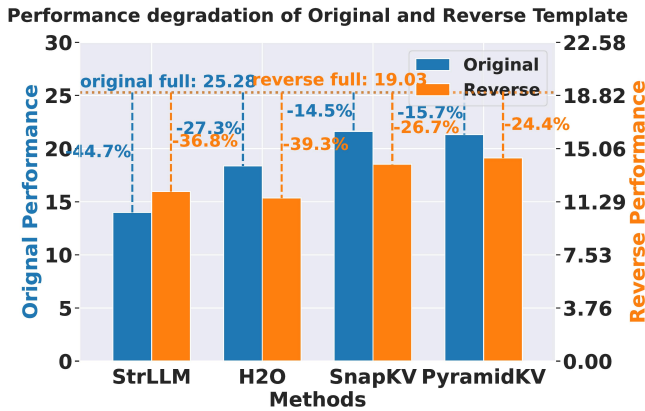


Figure 2: Performance degradation of different pruning methods on Single-QA tasks before and after prompt adjustment, with all methods having a KV budget of 512 tokens.

2.2 KV Cache Compression

In addition to KV cache eviction methods, memory space can also be conserved by compressing the representation of KV cache. Xu et al. (2024) identify substantial redundancy in the channel dimension of KV cache, and propose THINK, which reduces memory usage by retaining only the essential channels of KV cache vectors. Besides direct channel discarding for dimension reduction, some methods (Singhania et al. 2024; Zhang et al. 2024; Chang et al. 2025) utilize techniques such as Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) to effectively reduce the dimension of key and value vectors in KV cache.

Moreover, Bolya et al. (2022); Wan et al. (2024) focus on merging KV cache representations at token level, enabling key-value vectors to store more information per element, thus shrinking KV cache size. Furthermore, several studies (Sun et al. 2024; Yang et al. 2024b; Wu and Tu 2024) exploit the similarities of KV cache representations across different layers and employ cross-layer KV cache sharing to reduce its memory footprint.

3 Observations

Although existing KV cache eviction methods achieve memory reduction at the expense of small performance degradation, our empirical analysis reveals several fundamental limitations that constrain their practical effectiveness.

(1) Directly utilizing queries from the last window for importance calculation may lead to information loss. Current KV cache eviction methods during the pre-filling phase are predominantly designed based on the observation that the problem contexts tend to occur at the end of inputs. These methods typically compute importance scores for KV cache using the queries corresponding to the tokens in the last window. However, they are problem-specific and do not effectively preserve essential information. When the last window does not contain the question, such methods often experience significant performance degradation. To evaluate this limitation, we conduct experiments using Single-QA tasks from LongBench, systematically modifying the

Method	KV Cache Budget		
	128	256	512
Full KV-gen	38.27	39.29	39.80
StreamingLLM	30.50	31.79	32.64
H2O	33.67	34.27	35.30
SnapKV	34.31	36.56	38.31
PyramidKV	34.08	36.00	37.58

Table 1: The performance of different pruning methods on LongBench using Llama-3.1-8B-Instruct, all methods are under the same settings in Sec 5. The result of Full KV is 41.23. The optimal results are highlighted in bold.

prompt template by moving the question components from their conventional terminal position to the initial segment. Meanwhile, the prompt is adjusted to ensure semantic coherence. Under this modified setting, we evaluate the relative performance degradation of existing eviction methods before and after the prompt adjustment. As illustrated in Figure 2, the performance degradation of these pruning methods becomes more serious after prompt adjustment. This indicates that these methods suffer from information loss during importance calculation and are unable to adequately capture and process global contextual information.

(2) KV cache eviction guided by queries corresponding to actual decoded tokens yields excellent results. On the other hand, our findings suggest that directly using the queries corresponding to the actual decoded tokens to compute the importance scores of key-value pairs in KV cache for pruning can achieve highly effective results. We evaluate on the LongBench with Llama-3.1-8B-Instruct, leveraging FullKV-generated responses to guide KV cache eviction. The results are compared against other eviction methods. The experimental results, as shown in Table 1, demonstrate that under the same budget, the pruning method based on the actual response outperforms other pruning methods. This indicates that the decoded tokens are better at capturing global information. This observation is also logically consistent: since the pruned KV cache is ultimately used for decoding subsequent tokens, directly retaining the key-value pairs most important to the decoded tokens yields highly effective results, representing the theoretical upper bound of KV cache pruning methods during the pre-filling stage. However, during the pre-filling stage, the content of decoded tokens is not yet known in advance. Therefore, it is essential to develop an approximate pruning method that can utilize tokens from the pre-filling stage to identify critical key-value pairs that closely align with those selected based on actual decoded tokens.

4 Method

4.1 Framework of Judge Q

Motivated by the shortcomings of existing KV cache eviction methods during the pre-filling stage mentioned in Section 3, we propose the framework ‘‘Judge Q’’ to address these

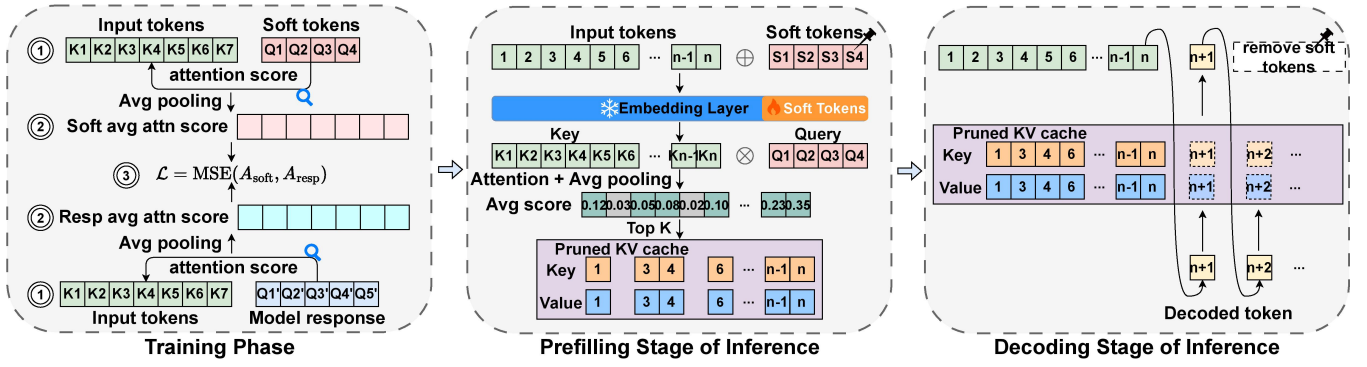


Figure 3: The framework of Judge Q. We specifically show the details of training, prefilling, and decoding stages. The number of soft tokens in the figure is set $n = 4$.

issues. The framework introduces a set of learnable soft tokens into model’s vocabulary, requiring fine-tuning of the embedding layer. It is illustrated in Figure 3. The training procedure is carefully designed to equip these tokens with enhanced global information awareness, while corresponding modifications during inference ensure these tokens actively guide optimized KV cache eviction.

4.2 Training Model With Soft Tokens

Before training, we define the number of soft tokens as n and expand the model’s vocabulary by appending these tokens. During the training process, we exclusively fine-tune the subset of parameters in the model’s embedding layer that correspond to these soft tokens, while keeping the remaining parameters frozen. We construct our training set using 50,000 samples from the ShareGPT dataset, with processing details provided in the repository. Each sample contains two fields: prompt and response, where the response represents the model’s actual output for the given prompt.

To train the soft tokens to effectively capture global information, we optimize their attention patterns to align with those of the response tokens. Specifically, the soft tokens and the response tokens are individually concatenated with the prompt tokens:

$$\mathbf{Input}_{\text{soft}} = \text{Concat}(\mathbf{Prompt}, \mathbf{Soft}) \quad (1)$$

$$\mathbf{Input}_{\text{resp}} = \text{Concat}(\mathbf{Prompt}, \mathbf{Resp}) \quad (2)$$

Next, we calculate their attention scores with respect to the prompt, and average these attention scores across the token dimension to derive the attention map. The training objective is to minimize the Mean Squared Error (MSE) between these two attention maps.

$$\mathbf{A}_{\text{soft}} = \frac{1}{n} \sum_{i=1}^n \text{Attention}(\mathbf{Prompt}, \mathbf{Soft}_i) \quad (3)$$

$$\mathbf{A}_{\text{resp}} = \frac{1}{m} \sum_{j=1}^m \text{Attention}(\mathbf{Prompt}, \mathbf{Resp}_j) \quad (4)$$

$$\mathcal{L} = \text{MSE}(\mathbf{A}_{\text{soft}}, \mathbf{A}_{\text{resp}}) = \frac{1}{d} \|\mathbf{A}_{\text{soft}} - \mathbf{A}_{\text{resp}}\|_2^2 \quad (5)$$

This approach ensures that the soft tokens are trained to focus on prompt regions critical for generation, achieving more informed KV cache eviction during inference.

4.3 Optimizing KV Cache Eviction via Soft Tokens during Inference

During inference, we leverage the trained soft tokens to perform KV cache eviction. Similarly to training procedure, we append n soft tokens sequentially to the end of original input. In the pre-filling stage, we compute the attention map of soft tokens with respect to the input, as formulated in Eq 3. Based on this attention map, we retain the key-value pairs corresponding to the top- k highest scores while discarding the rest. The soft tokens act as probes, which are removed after this pruning step, and the model continues decoding from the end of the original input using the pruned KV cache.

5 Experiments

5.1 Experimental Setting

Training data and configuration. We train models using a subset extracted from ShareGPT dataset (shareAI 2023). As described in Sec 4.2, this subset comprises 50,000 samples, with 45,000 from the common domain and 5,000 from computer. Experimental validation indicates that choosing $n = 32$ provides an effective balance between training efficiency and generalization. We conduct experiments on Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3.

Baselines. To validate the effectiveness of our method, we select three commonly used and effective KV cache eviction methods as baselines. During the evaluation, all baselines are tested under the same settings, i.e., local window size and pooling configurations, etc.

- **StreamingLLM** (Xiao et al. 2023) heuristically retains the key-value pairs corresponding to the beginning of the input as well as those from the local window.
- **H2O** (Zhang et al. 2023b) utilizes cumulative attention scores as importance scores. For ease of implementation, we adopt the approach outlined in (Cai et al. 2024), where the tokens from last window during the pre-filling stage are used as queries to compute importance scores.

Method	Single-Doc QA			Multi-Doc QA			Summ.			Few-shot			Synth.		Code		Avg.	
	NQA	Qasp	MF	HQA	2Wiki	Musq	GovR	QM	MN	TREC	Triv	SAM	PCnt	PRE	Lcc	RB-P		
Llama-3.1-8B-Instruct	Full KV	31.85	15.59	28.39	29.93	22.98	18.20	34.18	23.79	27.14	72.50	92.14	43.66	8.45	97.59	61.36	51.88	41.23
	KV Cache Size = 128																	
	StrLLM	15.35	4.96	14.83	14.68	17.76	7.55	20.56	19.66	20.55	42.50	83.12	32.14	10.34	96.58	46.78	40.56	30.50
	H2O	25.11	7.00	18.28	18.15	20.05	9.21	22.06	22.71	21.49	40.00	91.05	40.50	7.76	93.54	55.44	46.39	33.67
	SnapKV	24.14	7.59	21.56	17.74	19.25	10.15	19.91	22.13	20.40	48.50	91.63	40.26	10.00	93.70	55.29	46.75	<u>34.31</u>
	PyraKV	25.39	6.89	21.59	16.67	17.66	8.85	20.29	22.45	20.37	50.50	89.92	40.41	9.25	93.46	55.43	46.20	34.08
	Ours	28.21	10.45	24.47	19.51	18.38	12.13	23.85	23.94	22.97	51.50	92.20	41.58	9.52	92.44	57.00	46.22	35.90
	KV Cache Size = 256																	
	StrLLM	16.57	5.61	15.74	15.34	16.10	6.96	22.96	19.69	22.99	55.50	83.18	32.54	10.00	95.55	48.18	41.77	31.79
	H2O	25.89	8.64	19.57	17.34	18.19	10.02	23.35	22.95	23.00	41.50	91.29	41.72	7.96	94.57	56.87	46.98	34.37
	SnapKV	27.77	10.12	22.94	22.48	19.84	11.15	22.64	22.94	23.23	58.00	91.32	41.46	8.20	95.42	59.22	48.30	<u>36.56</u>
	PyraKV	26.52	8.55	22.66	17.94	18.56	9.95	22.34	23.16	22.52	59.00	91.69	41.35	9.35	96.93	58.33	47.18	36.00
	Ours	29.76	12.87	25.59	22.02	21.76	13.82	26.07	23.53	24.46	59.00	92.15	41.84	7.76	95.94	58.56	47.84	37.69
	KV Cache Size = 512																	
	StrLLM	17.99	6.38	17.60	13.97	16.71	8.03	25.26	20.17	24.88	60.00	83.85	33.14	10.17	92.08	50.00	42.03	32.64
	H2O	25.42	8.70	20.98	20.25	19.26	9.67	24.11	22.97	24.24	44.00	92.10	41.24	7.93	96.48	59.18	48.33	35.30
	SnapKV	28.53	11.33	25.00	25.86	21.13	13.05	25.05	23.13	24.44	66.00	92.32	41.56	8.70	96.23	60.24	50.31	<u>38.31</u>
	PyraKV	27.46	11.31	25.19	20.62	18.88	10.95	24.37	23.08	24.42	67.00	91.82	42.50	8.69	96.83	59.56	48.60	37.58
Ours	32.48	12.98	26.40	24.63	22.49	14.96	27.62	23.91	25.38	68.50	92.44	42.30	7.10	95.20	60.65	49.66	39.17	
Mistral-7B-Instruct-v0.3	Full KV	29.20	41.58	53.11	49.27	39.01	28.58	31.95	25.67	23.89	76.00	88.59	47.50	5.50	98.00	51.97	46.28	46.01
	KV Cache Size = 128																	
	StrLLM	21.42	20.42	30.56	40.83	32.77	18.26	18.12	19.70	16.42	46.00	81.07	31.78	6.00	81.00	37.61	34.88	33.55
	H2O	24.48	28.97	43.61	42.98	35.31	24.03	17.46	23.04	17.60	38.50	87.67	43.57	6.50	95.50	47.60	41.52	38.65
	SnapKV	23.86	28.02	51.44	47.71	35.90	24.16	16.53	21.79	16.70	47.50	89.54	42.98	6.00	93.50	48.87	41.56	<u>39.75</u>
	PyraKV	24.21	27.04	49.20	47.52	34.81	23.51	16.78	21.39	16.87	48.00	88.16	43.05	5.50	94.00	49.08	40.93	39.38
	Ours	27.86	35.71	54.33	48.14	38.69	26.97	20.43	24.23	19.96	56.50	90.16	46.29	5.50	93.00	48.74	41.55	42.38
	KV Cache Size = 256																	
	StrLLM	22.00	21.43	33.36	41.14	32.38	16.27	21.43	19.71	20.59	56.50	80.08	31.47	6.50	81.00	38.74	34.84	34.84
	H2O	25.15	28.31	47.49	46.61	35.39	24.52	18.81	23.65	19.37	39.50	88.32	43.70	6.00	97.50	50.15	43.08	39.85
	SnapKV	26.91	33.15	54.14	48.05	36.03	26.88	18.68	23.35	19.07	61.50	89.03	44.61	5.50	95.00	51.11	43.99	<u>42.31</u>
	PyraKV	25.67	31.84	54.15	48.43	36.38	26.62	18.76	23.35	19.07	62.00	89.46	44.42	6.00	96.00	50.88	44.00	<u>42.31</u>
	Ours	28.42	36.32	54.67	48.39	38.00	27.50	22.02	23.49	21.33	64.50	89.21	46.03	6.50	94.50	50.80	44.47	43.51
	KV Cache Size = 512																	
	StrLLM	22.22	22.98	37.29	40.29	32.89	16.47	24.58	19.98	23.00	65.50	80.33	31.06	5.50	83.50	40.35	35.90	36.37
	H2O	25.61	31.24	50.52	46.62	36.14	24.19	19.91	24.36	21.02	41.50	89.18	45.67	5.50	97.50	51.56	44.43	40.93
	SnapKV	28.46	35.53	55.18	49.30	37.40	27.59	21.03	23.87	20.75	70.50	89.57	44.72	5.50	97.50	52.01	45.67	<u>44.04</u>
	PyraKV	27.05	36.16	55.78	48.88	36.85	27.44	21.13	23.78	20.95	70.50	89.36	45.18	3.50	96.50	51.93	45.23	43.76
Ours	28.08	38.16	54.08	48.96	38.91	27.76	24.28	24.68	22.06	68.00	89.36	47.06	7.00	96.00	51.65	45.68	44.48	

Table 2: Performance comparison of different methods on LongBench under different budgets across various LLMs. The best result is in bold and the second best result is underlined. The local window size of all baseline is 32, and the length of soft tokens in our method is also 32. In the method, StrLLM and PyraKV correspond to StreamingLLM and PyramidKV respectively.

- **SnapKV** (Li et al. 2024b) calculates attention scores using tokens within the local window and applies a pooling mechanism to derive the final importance scores, ensuring consistency in KV cache eviction.
- **PyramidKV** (Cai et al. 2024) builds on SnapKV by dynamically allocating budgets across layers, with the budget gradually decreasing as the layer depth increases.

Benchmarks. We evaluate methods using three common benchmarks for KV cache eviction: LongBench (Bai et al. 2023), RULER (Hsieh et al. 2024), and Needle-in-a-Haystack (Kamradt 2023), and compare the performance of our method against baselines under multiple budgets.

5.2 Main Results

Results on LongBench. As shown in Table 2, under the same budget, our method outperforms these KV cache eviction baselines by approximately 1 point and even by up to 2.6 points. Moreover, the lower the budget, the greater the improvement achieved by our method. In addition, in certain tasks, the performance of our method even surpasses that of Full KV. This demonstrates the effectiveness of our method.

Results on RULER. We evaluate the performance of the methods under different budgets with input lengths of 8192 and 32768, respectively. The experimental results are presented in Table 3. Our method significantly outperforms all baselines, consistently maintaining an advantage of over 3

	Budget	Method					Ours
		Full KV	StrLLM	H2O	SnapKV	PyraKV	
Sequence Length = 8192							
Llama-3.1-8B-Ins	256	87.18	14.45	28.71	<u>57.83</u>	56.86	63.13
	512	87.18	16.95	41.61	<u>62.76</u>	61.19	69.24
	1024	87.18	21.89	53.22	<u>68.21</u>	66.30	74.12
Sequence Length = 32768							
Llama-3.1-8B-Ins	256	78.82	13.08	23.83	<u>53.51</u>	53.25	56.31
	512	78.82	14.00	34.02	<u>57.87</u>	57.01	63.08
	1024	78.82	16.29	43.56	<u>61.64</u>	60.60	67.55
Sequence Length = 8192							
Mistral-7B-Ins-v0.3	256	81.87	10.55	15.06	<u>42.44</u>	40.35	52.72
	512	81.87	12.81	21.28	<u>53.12</u>	50.14	61.95
	1024	81.87	17.98	30.09	<u>61.74</u>	58.69	71.24
Sequence Length = 32768							
Mistral-7B-Ins-v0.3	256	73.46	9.95	11.32	<u>26.25</u>	24.67	33.48
	512	73.46	10.69	13.16	<u>33.05</u>	31.69	40.83
	1024	73.46	11.92	16.20	<u>40.63</u>	39.13	49.97

Table 3: Performance on RULER under different budgets across various LLMs. The number of samples in the test dataset is 100, and all other settings are the same as Table 2.

points and peaking at nearly 10 points. Additional results under other settings are provided in the repository.

Results on Needle-in-a-Haystack. The experimental results are shown in Figure 4. It is evident that our method significantly exceeds all baselines, indicating its robust performance in retrieval scenarios. Detailed visualization results are available in the repository.

6 Discussion

6.1 The Effect of Soft Tokens

Based on the test results on benchmarks in Sec 5.2, we can observe the effectiveness of our method. We further conduct experiments to explain the advantages of the soft tokens.

Soft tokens can improve the critical key-value hit rate. As discovered in Section 3, utilizing actual decoded tokens to guide KV cache eviction can yield excellent results, and it can be considered the theoretical upper bound of KV cache eviction methods. To evaluate this, we introduce a new metric called the critical key-value hit rate. As defined in Equation 6, this metric measures the overlap between the key-value pairs selected by the current eviction method and those chosen using real decoded tokens under the same budget. By measuring the critical key-value hit rate, we can directly compare the discrepancy between the current eviction method and the theoretical upper bound.

$$\text{HitRate}(\mathbf{I}_{\text{cur}}, \mathbf{I}_{\text{resp}}) = \frac{|\mathbf{I}_{\text{cur}} \cap \mathbf{I}_{\text{resp}}|}{|\mathbf{I}_{\text{resp}}|} \quad (6)$$

According to the experimental results in Sec 5.2, SnapKV performs the best among all baselines. Therefore, we only compare the hit rate of Judge Q and SnapKV under different budgets. We preform evaluation using Llama-3.1-8B-Instruct on LongBench. As shown in Table 4, the critical

key-value hit rate of our method consistently exceeds that of SnapKV by approximately 8 points across various budgets.

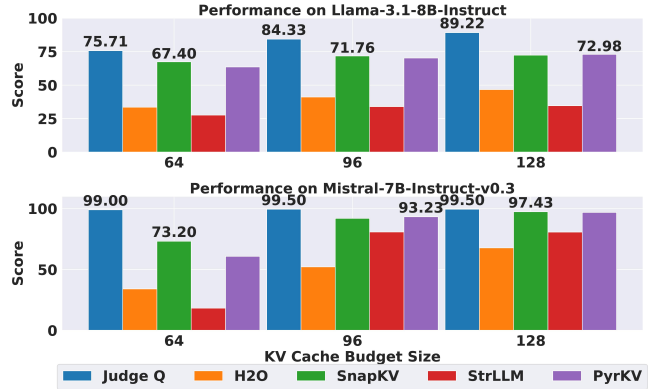


Figure 4: Performance on Needle-in-a-Haystack. The settings are the same as previous experiments. We set the minimum length to 4000, the maximum length to 32000, and the intervals for the document depth percent to 10.

This indicates that during the KV cache eviction process, our method is more effective at retaining important key-value pairs that meet decoding requirements, preserving more information and enhancing model’s performance. It is also consistent with the training objective of Judge Q, which is to train soft tokens to identify key-value pairs that closely approximate those selected by actual decoded tokens.

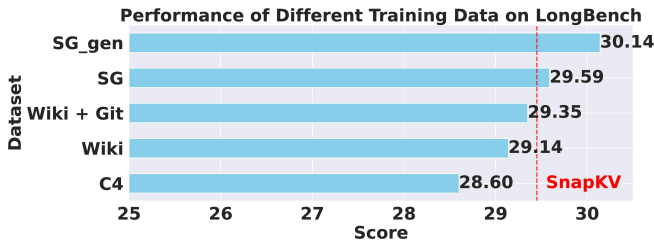
Method	KV Cache Budget		
	128	256	512
SnapKV	53.44	55.23	58.46
Ours	61.37	62.34	65.06

Table 4: Critical key-value hit rate of our method and SnapKV on LongBench under different budgets.

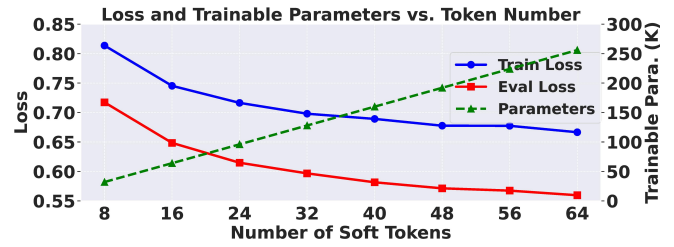
Using soft tokens for eviction can ensure the quality of generated content. We further evaluate the effectiveness of Judge Q. Following the prompt template adjustment settings described in Section 3, we test the proportion of performance degradation when the question is placed at the start of the prompt. Under the same settings, our method exhibits less than 7% performance loss, which is lower compared to the approximately 10% loss observed in baselines.

Method	Dataset	
	MATH-500	AIME24
SnapKV	52.4	31.1
Ours	55.0	37.8

Table 5: Performance on text continuation tasks, under a budget of 25% of average output length. Specifically, the budget is set to 1024 for MATH-500 and 3072 for AIME24.



(a) Performance of Llama-3.2-1B-Instruct on LongBench across training datasets (SG = ShareGPT, Wiki = Wikipedia.en, Git = GitHub repos), using consistent setup and a budget of 512.



(b) Loss and trainable parameters under different numbers of soft tokens during training on Llama-3.1-8B-Instruct. The hyperparameters used for training are the same as prior experiments.

Figure 5: The impact of data quality and the number of soft tokens.

Additionally, we conduct experiments on text continuation tasks. Specifically, we apply the Judge Q framework to DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI 2025) and evaluate its performance on the MATH-500 (Lightman et al. 2023) and AIME24 (MAA 2024) datasets. For each dataset, the model initially generates responses for individual questions. Subsequently, we extract the last 768 tokens from the generated responses, and perform KV cache eviction before continuing to generate responses. Finally, we evaluate the results of the regenerated outputs. As shown in Table 5, our method exhibits strong performance in KV cache eviction under long-output scenarios. Notably, in the case of AIME24, where the output exceeds 10,000 tokens and only one-quarter of the tokens are retained, our method significantly outperforms SnapKV.

Case Study

Prompt: You are given a story, which can be either a novel or a movie script, ... Question: What is Saltram’s living situation? Answer: (31406 tokens)

SnapKV Response: He is living with the Saltrams, who have a large house. ✗

Judge Q Response: He is living with the Mulvilles in their drawing room. ✓

Full KV Response: He is living with the Mulvilles at Wimbledon. ✓

Gold Response: He is a guest in the home of the Mulvilles.

Furthermore, we analyze specific case to compare the differences in the generated content between Judge Q and other eviction methods. We select an example from Narrative QA using Llama-3.1-8B-Instruct as follows with a KV cache budget set to 512. It is evident that the response generated by Judge Q closely resembles the of Full KV and correctly answers the question, whereas SnapKV deviates significantly from the Full KV output, resulting in an incorrect response. This validates that Judge Q can effectively maintain high generation quality.

6.2 The Impact of Data Quality

Training with model-generated responses yields better results. The quality of training data can impact the ef-

fectiveness of our method to some extent. During experiments, we observe that using model-generated responses outperforms directly using the original responses from the datasets. For efficiency, we conduct data quality explorations using Llama-3.2-1B-Instruct and evaluate on LongBench. From results in Figure 5a, it is evident that utilizing model-generated responses as training data is more effective.

Training data content crucial for performance. We also extract and combine training data from various pre-training datasets, including C4 (Raffel et al. 2020), Wikipedia (Foundation 2023) and GitHub. As Figure 5a shows, these results are inferior to ShareGPT, because the data content is limited and does not cover a comprehensive range of domains.

6.3 The Impact of the Number of Soft Tokens

Additionally, during training, we construct a validation set comprising 5000 samples to explore the impact of the number of soft tokens. As illustrated in Figure 5b, when the number of soft tokens is around 32, there is an optimal balance between training cost and generalization performance.

7 Conclusion

In this paper, we first identify that existing KV cache eviction methods during the pre-filling stage, which directly select the local window for importance calculation, tend to overly focus on local information while neglecting global information. To address this issue, we propose Judge Q, a lightweight training framework that introduces a small number of soft tokens into the vocabulary, and only fine-tunes a small portion of the model’s embedding parameters. This approach enhances the consideration of global information.

We conduct experiments using various models across multiple benchmarks, demonstrate that our method consistently outperforms existing KV cache eviction methods under different budgets. Judge Q outperforms existing methods by approximately 1 point on LongBench, exceeds by more than 3 points on RULER, and significantly surpasses the baselines on Needle-in-a-Haystack. We further validate the effect of soft tokens as well as the impact of data quality and the number of soft tokens.

In the future, we plan to conduct experiments with a wider range of models, use longer data for training and implement streaming KV cache eviction during the decoding process.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China (NSFC) via grant 62236004, 62206078 and 62476073.

References

- Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.
- Bolya, D.; Fu, C.-Y.; Dai, X.; Zhang, P.; Feichtenhofer, C.; and Hoffman, J. 2022. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*.
- Cai, Z.; Zhang, Y.; Gao, B.; Liu, Y.; Li, Y.; Liu, T.; Lu, K.; Xiong, W.; Dong, Y.; Hu, J.; et al. 2024. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.
- Chang, C.-C.; Lin, W.-C.; Lin, C.-Y.; Chen, C.-Y.; Hu, Y.-F.; Wang, P.-S.; Huang, N.-C.; Ceze, L.; Abdelfattah, M. S.; and Wu, K.-C. 2025. Palu: KV-Cache Compression with Low-Rank Projection. In *The Thirteenth International Conference on Learning Representations*.
- Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.
- Chen, Y.; Wang, G.; Shang, J.; Cui, S.; Zhang, Z.; Liu, T.; Wang, S.; Sun, Y.; Yu, D.; and Wu, H. 2024. Nacl: A general and effective kv cache eviction framework for llms at inference time. *arXiv preprint arXiv:2408.03675*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948*.
- Devoto, A.; Zhao, Y.; Scardapane, S.; and Minervini, P. 2024. A Simple and Effective L_2 Norm-Based Strategy for KV Cache Compression. *arXiv preprint arXiv:2406.11430*.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints, arXiv:2407*.
- Feng, Y.; Lv, J.; Cao, Y.; Xie, X.; and Zhou, S. K. 2024. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *arXiv preprint arXiv:2407.11550*.
- Foundation, W. 2023. Wikimedia Downloads. <https://dumps.wikimedia.org>.
- Ge, S.; Zhang, Y.; Liu, L.; Zhang, M.; Han, J.; and Gao, J. 2023. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*.
- He, Z.; Yuan, J.; Bai, H.; Leng, J.; and Jiang, B. 2025. TreeKV: Smooth Key-Value Cache Compression with Tree Structures. *arXiv preprint arXiv:2501.04987*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Hsieh, C.-P.; Sun, S.; Krizan, S.; Acharya, S.; Rekish, D.; Jia, F.; Zhang, Y.; and Ginsburg, B. 2024. RULER: What’s the Real Context Size of Your Long-Context Language Models? *arXiv preprint arXiv:2404.06654*.
- Jain, N.; Han, K.; Gu, A.; Li, W.-D.; Yan, F.; Zhang, T.; Wang, S.; Solar-Lezama, A.; Sen, K.; and Stoica, I. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B.
- Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints, arXiv:1705.03551*.
- Kamradt, G. 2023. Needle in a Haystack - Pressure Testing LLMs. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.
- Kočíšký, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.
- Li, J.-N.; Tu, Q.; Mao, C.; Yu, Z.; Wen, J.-R.; and Yan, R. 2024a. Streamingdialogue: Prolonged dialogue learning via long context compression with minimal losses. *Advances in Neural Information Processing Systems*, 37: 86074–86101.
- Li, Y.; Huang, Y.; Yang, B.; Venkitesh, B.; Locatelli, A.; Ye, H.; Cai, T.; Lewis, P.; and Chen, D. 2024b. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37: 22947–22970.
- Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let’s Verify Step by Step. *arXiv preprint arXiv:2305.20050*.
- Liu, J.; Zhu, D.; Bai, Z.; He, Y.; Liao, H.; Que, H.; Wang, Z.; Zhang, C.; Zhang, G.; Zhang, J.; et al. 2025a. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.
- Liu, Y.; Yu, J.; Xu, Y.; Li, Z.; and Zhu, Q. 2025b. A survey on transformer context extension: Approaches and evaluation. *arXiv preprint arXiv:2503.13299*.
- Liu, Z.; Desai, A.; Liao, F.; Wang, W.; Xie, V.; Xu, Z.; Kyrillidis, A.; and Shrivastava, A. 2023. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36: 52342–52364.
- Lu, S.; Chen, Y.; Feng, W.; Fan, J.; Li, F.; Zhang, Z.; Lv, J.; Shen, J.; Law, C.; and Liang, J. 2025a. Uni-layout: Integrating human feedback in unified layout generation and

- evaluation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7709–7718.
- Lu, S.; Wang, Y.; Sheng, L.; He, L.; Zheng, A.; and Liang, J. 2025b. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2): 1–39.
- MAA. 2024. American Invitational Mathematics Examination - AIME. In *American Invitational Mathematics Examination - AIME 2024*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.
- shareAI. 2023. ShareGPT-Chinese-English-90k Bilingual Human-Machine QA Dataset. <https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k>.
- Singhania, P.; Singh, S.; He, S.; Feizi, S.; and Bhatele, A. 2024. Loki: Low-rank keys for efficient sparse attention. *Advances in Neural Information Processing Systems*, 37: 16692–16723.
- Sun, Y.; Dong, L.; Zhu, Y.; Huang, S.; Wang, W.; Ma, S.; Zhang, Q.; Wang, J.; and Wei, F. 2024. You only cache once: Decoder-decoder architectures for language models. *Advances in Neural Information Processing Systems*, 37: 7339–7361.
- Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- Wan, Z.; Wu, X.; Zhang, Y.; Xin, Y.; Tao, C.; Zhu, Z.; Wang, X.; Luo, S.; Xiong, J.; Wang, L.; et al. 2024. D2o: Dynamic discriminative operations for efficient long-context inference of large language models. *arXiv preprint arXiv:2406.13035*.
- Wang, Y.; Ji, S.; Liu, Y.; Xu, Y.; Xu, Y.; Zhu, Q.; and Che, W. 2025. Lookahead Q-Cache: Achieving More Consistent KV Cache Eviction via Pseudo Query. *arXiv preprint arXiv:2505.20334*.
- Wu, H.; and Tu, K. 2024. Layer-condensed kv cache for efficient inference of large language models. *arXiv preprint arXiv:2405.10637*.
- Xiao, G.; Tang, J.; Zuo, J.; Guo, J.; Yang, S.; Tang, H.; Fu, Y.; and Han, S. 2024. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.
- Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.
- Xu, Y.; Jie, Z.; Dong, H.; Wang, L.; Lu, X.; Zhou, A.; Saha, A.; Xiong, C.; and Sahoo, D. 2024. Think: Thinner key cache by query-driven pruning. *arXiv preprint arXiv:2407.21018*.
- Yang, D.; Han, X.; Gao, Y.; Hu, Y.; Zhang, S.; and Zhao, H. 2024a. Pyramidinfer: Pyramid kv cache compression for high-throughput llm inference. *arXiv preprint arXiv:2405.12532*.
- Yang, Y.; Cao, Z.; Chen, Q.; Qin, L.; Yang, D.; Zhao, H.; and Chen, Z. 2024b. Kvsharer: Efficient inference via layer-wise dissimilar kv cache sharing. *arXiv preprint arXiv:2410.18517*.
- Zhang, F.; Chen, B.; Zhang, Y.; Keung, J.; Liu, J.; Zan, D.; Mao, Y.; Lou, J.-G.; and Chen, W. 2023a. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*.
- Zhang, R.; Wang, K.; Liu, L.; Wang, S.; Cheng, H.; Zhang, C.; and Shen, Y. 2024. Lorc: Low-rank compression for llms kv cache with a progressive compression strategy. *arXiv preprint arXiv:2410.03111*.
- Zhang, Z.; Sheng, Y.; Zhou, T.; Chen, T.; Zheng, L.; Cai, R.; Song, Z.; Tian, Y.; Ré, C.; Barrett, C.; et al. 2023b. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36: 34661–34710.