TokenButler: Token Importance is Predictable

Yash Akhauri [1] **Ahmed F AbouElhamayed** [1] **Yifei Gao** [1] **Chi-Chih Chang** [1] **Nilesh Jain** [2]
Mohamed S. Abdelfattah [1]

Abstract

Large Language Models (LLMs) rely on the Key-Value (KV) Cache to store token history, enabling efficient decoding of tokens. As the KV-Cache grows, it becomes a major memory and computation bottleneck, however, there is an opportunity to alleviate this bottleneck, especially because prior research has shown that only a small subset of tokens contribute meaningfully to each decoding step. A key challenge in finding these *critical tokens* is that they are dynamic, and heavily input query-dependent. Existing methods either risk quality by evicting tokens permanently, or retain the full KV-Cache but rely on retrieving chunks (pages) of tokens at generation, failing at dense, context-rich tasks. Additionally, many existing KV-Cache sparsity methods rely on inaccurate proxies for token importance. To address these limitations, we introduce **TokenButler**, a high-granularity, query-aware predictor that learns to identify these critical tokens. By training a light-weight predictor with less than 1.2% parameter overhead, TokenButler prioritizes tokens based on their contextual, predicted importance. This improves perplexity & downstream accuracy by over 8% relative to SoTA methods for estimating token importance. We evaluate TokenButler on a novel synthetic small-context co-referential retrieval task, demonstrating near-oracle accuracy. Code, models and benchmarks: `[code]`

1. Introduction

As Large Language Models (LLMs) become more widely used (Thoppilan et al., 2022; Yuan et al., 2022; Wei et al., 2022; Zhang et al., 2023a), recent advances have extended their context lengths to 128k–1M tokens. However, recent research on long-context evaluation (Vodrahalli et al., 2024) reveal that model quality degrades noticeably as early as 8k tokens, even without token compression. Further, as input sequences grow, the memory footprint of the Key-Value (KV) cache—which stores intermediate key-value pairs to skip recomputation—scales linearly. This increases memory requirements and stresses the memory-bandwidth (to store and access the entire KV-Sequence). This raises important questions on how effectively existing token-pruning techniques address KV-cache size, especially in context-dense downstream tasks that go beyond retrieval or summarization. There have been several efforts at improving model quality while addressing KV-cache memory issues. Certain transformer variants aim at implicitly compressing the KV-cache via sparsity, quantization, efficient-attention, or low-rank compression (Child et al., 2019; Choromanski et al., 2020; Katharopoulos et al., 2020; Shazeer, 2019; Pope et al., 2022; Sun et al., 2024; Akhauri et al., 2024b; Chen et al., 2025).

The current literature on token pruning addresses this growing memory footprint in three ways. **(1)** *Purely static strategies* limiting KV-Cache to a fixed budget with fixed rules on removing tokens, naturally reducing bandwidth and storage (StreamingLLM (Xiao et al.), and Sliding Window Attention (Luong, 2015)), **(2)** *Adaptive strategies* that permanently sacrifice *less important* past-tokens effectively fixing the memory and bandwidth footprint (H_2O, SnapKV (Zhang et al., 2023b; Li et al., 2024)), and **(3)** *Adaptive dynamic strategies* that preserve the entire KV-Cache but access only a subset of the Key-Value entries (the *more important* past-tokens), incurring higher memory (storage) cost, but reducing memory bandwidth (accesses to memory) during the decode stage (generation) (Quest, FastGen, (Tang et al., 2024; Ge et al.))

Each of these strategies to limit storage and bandwidth costs have implications. Specifically, token preference has been shown to be highly dependent on the query (Tang et al., 2024), and vary significantly at generation. Purely static strategies do not have any query-awareness, and will fail at retrieving contextually relevant tokens. Additionally, *Adaptive strategies* that have permanently discarded tokens deemed less important at a prior decode step will not be able to fetch relevant tokens if the course of discussion is *co-referential* (Vodrahalli et al., 2024). A conversation is co-referential if text introduced earlier is referenced again

[1]Cornell University, New York, USA [2]Intel Labs, USA. Correspondence to: Mohamed S. Abdelfattah <mohamed@cornell.edu>, Yash Akhauri <ya255@cornell.edu>.

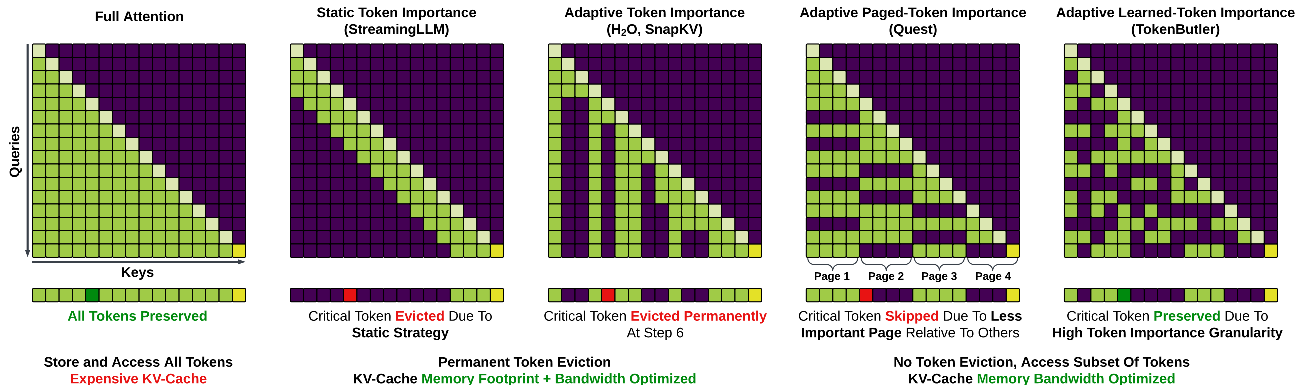

Figure 1. Full-Attention preserves all tokens, enabling access to the critical token (dark green) during the last decode step. Static strategies like StreamingLLM will not be able to access this token. Methods like H_2O may have evicted the token at an earlier decode step, if deemed unimportant. Paged-Token importance may cause a page-miss of a critical token in context dense tasks. **TokenButler** can effectively predict critical tokens, and can be leveraged by existing methods to offer both high-granularity and cheap importance estimation.

later, requiring accurate retrieval and reasoning over the earlier reference. For co-referential conversations *adaptive dynamic strategies* is the most reasonable solution. Current methods rely on *token grouping* to make the dynamic calculation of token relevance efficient (Tang et al., 2024).

There are several *metrics* to quantify token importance including recency, aggregate attention scores, and others listed in Table 1. Token Sparsity methods use these metrics to guide token eviction or retrieval decisions. There is an important interplay between methods and metrics—methods that permanently evict tokens based on strong metrics like the attention score. However, evicted tokens may become relevant later during generation. Token sparsity methods that preserve tokens but selectively retrieve a subset during generation cannot rely on strong metrics such as attention scores. This is because only a subset of the KV-cache is fetched during generation based on a token importance metric—the metric cannot be the result of the computation itself (attention score). To address this, we propose a novel *learned* metrics of token importance, called **TokenButler**, which provides fine-granularity estimates of token importance. Our contributions are summarized as:

- We show that it is possible to train a light-weight predictor ($< 1.2\%$ parameter overhead) for estimating token-importance, achieving up to 75% accuracy in identifying the top 50% tokens.
- We design a small-context (< 512 tokens) co-reference and decode-focused benchmark, which highlights the current gap in KV-Sparsity research, with TokenButler delivering near-oracle accuracy.
- TokenButler can improve the perplexity and downstream accuracy over existing token sparsity metrics by over 8%, identifying critical tokens with *near-oracle* accuracy.

Method	**Metric**
StreamLLM	Recency-based sliding window
H_2O	Attention Score for Token Eviction
SnapKV	Pooled Attention Score over a Fixed Window for Token Eviction
Quest	Query product with Per-Page Min-Max Token Magnitudes for Page Loading
TokenButler	Predicted Importance for Fine-Grained Token Loading

Table 1. Metrics for token importance

2. Related Work

Prior work has shown that transformers exhibit very strong contextual behavior, where head and neuron importance heavily depends on the query. (Liu et al., 2023; Akhauri et al., 2024a) leverage this behavior to contextually prune entire neurons and heads on a per-query basis. These methods train small neural networks to predict the relative importance–quantified using parameter magnitudes or gradients of neurons across the transformer. This *magnitude* can be considered as the *metric* of contextual importance. Further, these works explore techniques of using these metrics to then prune heads globally, or on a per-layer basis. This idea of pruning on a global or per-layer basis can be considered as the *method*, which leverages the *metric* to make decisions.

This contextual behavior applies to token importance by design, as the attention mechanism explicitly captures token relevant to a query. However, while *methods* to prune heads are simpler, as there is a fixed number of heads, *methods* to prune tokens are more expensive to realize. Specifically, for a transformer with N layers and H heads per-layer and L past-tokens, every head has to decide which *subset* S

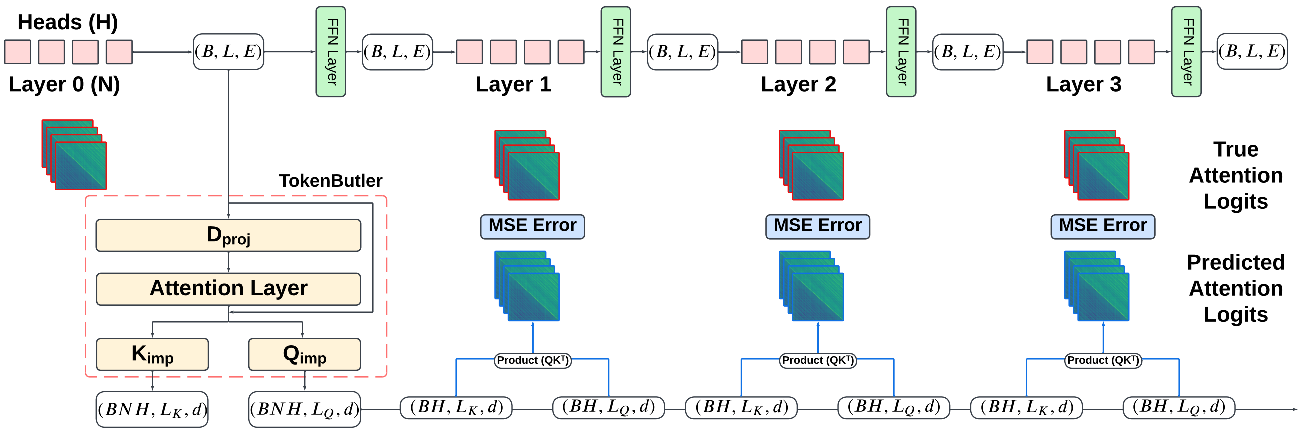

Figure 2. TokenButler is a light-weight predictor, with a down projection D_{proj} for cheaper attention, attention layer, and Key-Query projection *neural networks*. These $\{Q_{\text{imp}}, K_{\text{imp}}\}$ effectively map the output of the attention mechanism to $N \times H$ Key-Query projection tensors (N: Num. Layers, H: Num. Heads) on a small interaction-dimension $d \ll E$. The full-(pre-softmax) attention logits can then be computed for every head across all layers by taking Product(QK^T). At train-time, we simply minimize the MSE Error between true and prediction (pre-softmax) attention logits to learn the LLM behavior.

of L tokens are the most important at every decode step. This implies that any given *metric* has to be calculated for $N \times H \times L$ tokens, at *every decode step*.

As presented in Table 1, there have been significant efforts towards *co-designing* metrics with methods of token sparsity. The simplest methods are *purely static strategies*, StreamingLLM (Xiao et al.) relies on *recency* as a metric of token importance, with a sliding-window attention to fix a KV-Cache budget. More recently, methods like H_2O (Zhang et al., 2023b) and SnapKV (Li et al., 2024) avoid naïve sparsification of tokens, and instead rely on attention scores to permanently evict low-importance tokens. This can be a major limitation when tasks require synthesizing or reasoning over information distributed across the context (Vodrahalli et al., 2024), as a token important later in the decoding stage may be evicted due to low KV-Budget. To alleviate this issue, *Adaptive Dynamic Strategies* such as Quest (Tang et al., 2024) preserve all tokens, and dynamically decide which subset of tokens to fetch for a given query. Instead of calculating full attention scores to ensure the most important tokens are fetched (which can be prohibitively expensive), Quest relies on paging, preserving all tokens in paged memory, and selectively fetches important pages. Thus, they have to take the dot product of query with min-max token values within a page, and fetch pages of tokens. Essentially preserving tokens at the cost of granularity. This reduces memory bandwidth but does not optimize memory footprint, further, its sparsity is limited to the granularity of pages.

While *metrics* that rely on attention scores are an effective way to estimate token importance, its usefulness is limited as it is tied to the *method*, necessitating token-eviction, or paged-token fetching. By contrast, we propose to learn a lightweight token-importance predictor, *TokenButler*, which cheaply approximates token-level attention logits using QK projections from the first layer. This preserves fine-grained control over tokens (like full attention) while staying efficient as a fixed proportion of the transformer inference cost ($< 2\%$ latency overhead).

3. Methodology

We propose a method to prioritize most impactful tokens at each decode step. Our predictor avoids the expense of fine-grained token importance calculation for every head over all past tokens, instead leveraging a light-weight neural network to approximate the attention patterns. In this section, we describe the **TokenButler** predictor architecture.

3.1. Predictor Design

TokenButler is a lightweight transformer ($\approx 1\%$ of the LLM size), depicted in Figure 2. For each layer and head, TokenButler estimates token-importance. The predictor takes in the hidden-states from the attention mechanism of the first layer, down-projects it, adds an attention layer to process the sequence, and passes it to a query and key (QK) projection neural network (QK-NN). These QK-NNs capture the behavior of all heads from later layers in the LLM.

Concretely, given a batch of hidden states $\mathbf{I} \in \mathbb{R}^{B \times L \times E}$ (where B is the batch size, L is the sequence length, and E is the LLM's embedding dimension), the predictor reduces $\mathbf{I}$ via a attention sub-network. This sub-network includes:

- A dimensionality-reduction projection (Linear) that maps $\mathbf{I}$ into a smaller space for efficient self-attention.

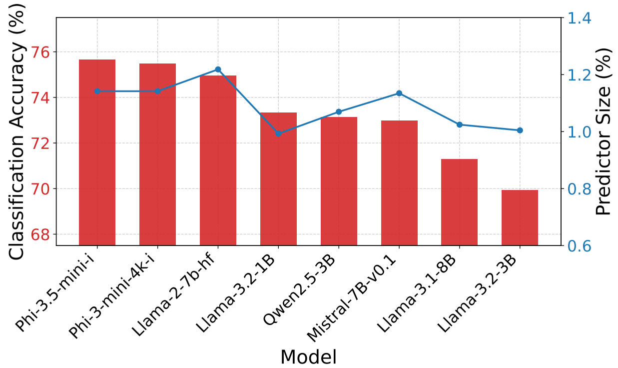

Figure 3. We train predictors within a $[1, 1.2]\%$ parameter count budget compared to the target LLM. Token Classification Accuracy indicates the classification accuracy in identifying the top 50% most important tokens. We achieve between 70-75% classification accuracy

- A single self-attention block over the reduced hidden states to capture context among tokens.
- A feed-forward block that up-projects reduced hidden states to $\mathbf{I}' \in \mathbb{R}^{B\times L\times E}$ and sums to $\mathbf{I}$ (residual).

Next, TokenButler uses two separate projection neural networks $\{Q_{\text{imp}}, K_{\text{imp}}\}$ to produce per-layer, per-head keys and queries for *importance*:

$$\mathbf{Q}_{\text{imp}} = Q_{\text{imp}}(\mathbf{I}'), \quad \mathbf{K}_{\text{imp}} = K_{\text{imp}}(\mathbf{I}'),$$

where each projection networks is implemented as two linear layers with a SiLU activation in between. Their outputs are reshaped to $\mathbb{R}^{B\times N\times H\times L\times d}$ (with N being the number of LLM layers, H the number of heads per layer, D the head dimension and $d \ll E$ a small *interaction* predictor-head dimension). Flattening N and H for efficiency yields $\mathbf{Q}_{\text{imp}}, \mathbf{K}_{\text{imp}} \in \mathbb{R}^{(B\,N\,H)\times L\times d}$. TokenButler then computes an approximate attention logit for each layer-head-token triplet:

$$\mathbf{A}_{\text{pred}} = \frac{\mathbf{Q}_{\text{imp}}\,\mathbf{K}_{\text{imp}}^{\top}}{\sqrt{d}}$$

giving an importance matrix $\mathbf{A}_{\text{pred}} \in \mathbb{R}^{(B\,N\,H)\times L\times L}$. These unnormalized logits mimic the pre-softmax attention maps of the LLM, and are used to predict how strongly each token attends to every other token, *per layer and head*, according to the learned behavior of token importance by TokenButler.

3.2. Predictor Training

The LLM is frozen and we train only the TokenButler predictor. We run a forward pass of the LLM on the C4-*realnewslike* training corpus and extract its (pre-softmax) attention logits $\mathbf{A}_{\text{true}} \in \mathbb{R}^{(B\,N\,H)\times L\times L}$ *before* causal masking and softmax. Meanwhile, TokenButler produces its approximate logits $\mathbf{A}_{\text{pred}}$. We then minimize a mean-squared-error (MSE) loss between the two as $\mathcal{L}_{\text{MSE}} = ||\mathbf{A}_{\text{pred}} - \mathbf{A}_{\text{true}}||_2^2$. In practice, for each training batch:

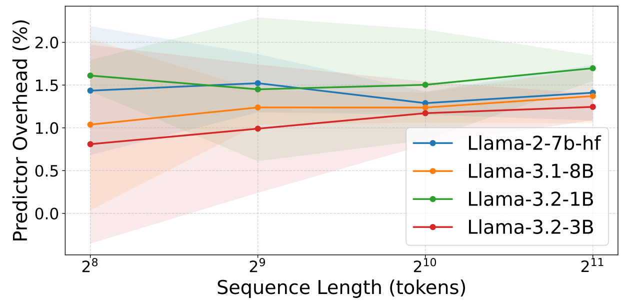

Figure 4. We measure the overhead of our predictor across sequence lengths, reporting the median over 100 iterations on a Nvidia A6000 GPU for Llama models. We do not implement any sparsity or optimizations, simply the overhead of prefill-token importance estimation. Our predictor scales with the model size ($\approx 1\%$), and thus adds $< 2\%$ latency overhead.

1. **Forward pass** Compute $\mathbf{A}_{\text{true}}$ for each layer $n = 1, \ldots, N$ and head $h = 1, \ldots, H$, pass the first-layer output of the LLM to the predictor to obtain $\mathbf{A}_{\text{pred}}$.
2. **Loss computation.** Accumulate MSE across all layers (except the first layer) and heads.
3. **Backward update (predictor only).** Update TokenButler's parameters; the LLM remains frozen.

The predictor learns to approximate attention patterns of the full model with minimal overhead. In downstream usage, it can thus rapidly identify which tokens are most critical—at per-token granularity—without performing expensive attention computations.

3.3. Efficient Training

A direct comparison of full attention logits $\mathbf{QK}^{\top}$ with TokenButler's predicted logits $\mathbf{Q}_{\text{imp}}\,\mathbf{K}_{\text{imp}}^{\top}$ would require materializing $\mathcal{O}(L^2)$ tensors and storing large intermediate gradients. To avoid this, we adapt a block-wise *flash attention*(Dao et al., 2022) kernel in Triton that computes partial products $\mathbf{Q}\,\mathbf{K}^{\top}$ in streaming form. At each block, we accumulate a partial MSE loss term before advancing to the next block as:

$$\Delta = \left\| \frac{\mathbf{QK}^{\top}}{\sqrt{D}} - \frac{\mathbf{Q}_{\text{imp}}\,\mathbf{K}_{\text{imp}}^{\top}}{\sqrt{d}} \right\|^2$$

This design matches the original flash-attention principle that avoids materializing the full attention matrix in mem-

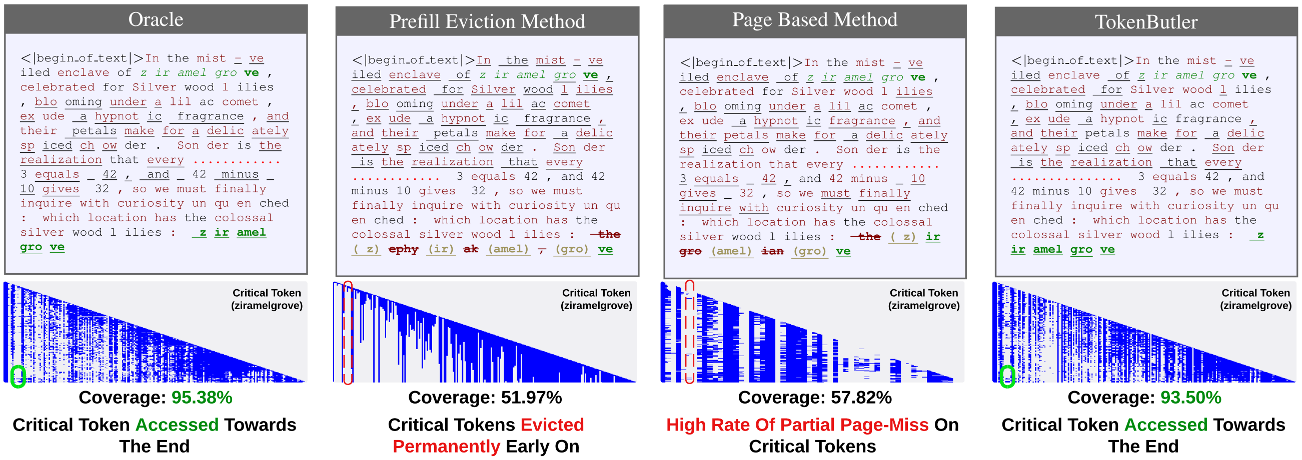

Figure 5. For different methods, we present the true decode-simulation of token access patterns for a sample head on Llama-3.2-3B. The tokens that are accessed at the decode step for location *ziramelgrove* are underlined. The tokens that were mis-predicted at the last step of the decode simulation are in red. The positions of the location is in green, and if the final location is mis-predicted, that token is striked-out in red. Coverage (number of correctly predicted tokens) is excellent for both Oracle and TokenButler.

ory, yet incorporates the pre-softmax MSE loss function computation.

During the backward pass, we leverage the kernel's streaming dot products to compute the gradients $\nabla \mathbf{Q}_{\text{imp}}$ and $\nabla \mathbf{K}_{\text{imp}}$ block by block, eliminating the need to store large intermediate activation maps. This approach ensures that the MSE-based training step remains memory efficient and scales effectively to long-context scenarios. Crucially, since our loss is computed directly on the logits **before** the softmax operation, the kernel seamlessly incorporates partial dot products with the MSE term within the same blockwise execution, enabling efficient gradient updates.

4. Evaluation

We train and evaluate the accuracy of our predictors on Llama-3.2-3B, Llama-3.1-8B (Grattafiori et al., 2024), Llama-2-7b-hf (Touvron et al., 2023), Mistral-7B-v0.1 (Jiang et al., 2023), Phi-3.5-mini-instruct, and Phi-3-mini-4k-instruct (Abdin et al., 2024). The predictors are trained on the same text-corpus, resulting in 80-100M tokens (due to tokenizer differences) using C4-*realnewslike* (Raffel et al., 2019).

4.1. Predictor Performance and Efficiency

Figure 3 illustrates two key results: (1) the *Token Classification Accuracy* of classifying the top 50% most important tokens (according to the LLM attention map) and (2) the relative size of our predictor compared to the base model. Across models, TokenButler attains 70-75% accuracy while amounting to roughly 1-1.2% of the LLM's total parameters.

Despite running alongside the LLM at each decoding step, the predictor imposes minimal runtime overhead. Figure 4 shows that TokenButler adds roughly 1-2% additional latency. This also indicates that the predictor can serve as a *learned metric* for token compression methods. For instance, for token eviction methods used by H2O/SnapKV, our predictor will get more efficient as the tokens get evicted and the KV-tensors of the predictor decrease in size. Additionally, for decode-time token access optimizations, our predictor can integrate with paging based methods, opening future work on new methods, for e.g., where prefill-tokens can be reordered in pages more effectively to have pages with more critical tokens for higher hit-rate, better token prefetching etc.

4.2. Evaluation On a Synthetic Task for Token Retrieval

We evaluate TokenButler on a difficult synthetic task inspired by Multi-Round Co-reference Resolution (Vodrahalli et al., 2024), using concise sequences (< 512 tokens). The model must recall a fictional location mentioned in a *contextual lead*, then referenced again after several distracting statements. By the time the location needs to be mentioned again after the location prelude, several tokens may have intervened, making it likely that the location tokens may have been evicted. Course-grained retrieval schemes risk not finding the entire location as it may be split across pages. This setup mimics conversation-like scenarios, it is especially challenging for token sparsity methods, since prematurely discarding or overlooking the location tokens can irreversibly break the final reference, leading to incorrect or incomplete retrieval of the location name.

We first use GPT-4o-mini to generate 100 fictional location names. We then generate 100 short contextual lead plus a matching prelude; then we generate 100 random math, culinary, and philosophical statements. At evaluation, we form 100 sequences adhering to the template. Note that every contextual lead is paired with a matching location prelude. Then, each test sequence is generated as a random sample as follows:

Synthetic Benchmark Template

```
<contextual lead>  <location> <philosophical
statement> <culinary statement> <math
problem>  <location prelude>  <location>
```

Summary Sample of our Synthetic Benchmark

```
Shrouded in luminescent fog, ...  color.  The
place is:   wraithspire In the spirit of ...
wisdom waiting to sprout.  Savor the delicate
...  home-cooked love.  If we compute 18 ...
7 gives us 16.   Which location is bathed ...
lights up the shore?   wraithspire
```

We evaluate two metrics under a uniform 50% token sparsity budget across all heads in the LLM, keeping the first layer dense. **Accuracy** measures the proportion of instances where the model correctly predicts the *entire mentioned* `<location>` given the contextual lead. **Coverage** accounts for the fact that locations span multiple tokens, we measure the fraction of location tokens correctly predicted across all examples.

Since every head may evict tokens based on their importance, we present the attention map for the first head of the 3rd layer (a random choice) for Figure 5. We observe there as well as Table 2 that (i) H_2O has low accuracy because it permanently evicts older tokens (the location name) once new context is being decoded. (ii) Quest (page-based) **very often** loses part of the location name if it straddles a page boundary in this context-dense example. *Coverage* gives a more detailed view on accuracy. Accuracy is binary, and locations are multiple tokens long, however, coverage counts the number of correctly-predicted tokens; for example, if the provided location is 4 tokens long, and a method gets 3 of those tokens correct, it is scored 0.75 in coverage and 0 in accuracy. We see that token eviction and page-based methods are still able to correctly predict around $30 - 50\%$ of the tokens, but not all of them, leading to low accuracy.

4.3. Accuracy Evaluation on Standard Benchmarks

We compare with several key works, such as H2O, SnapKV, StreamingLLM and Quest under a uniform token sparsity setup (applied to all layers except the first). We impose a token budget proportional to the input length (e.g. 50% sparsity retains half the tokens). In real-world generative use-cases, new tokens stream in while older tokens remain potentially important, whereas token-eviction based methods like H2O and SnapKV must decide at each step which tokens to discard. Meanwhile, TokenButler and Quest estimate token importance inexpensively for the full input without needing eviction, so they stay efficient even when preserving all tokens.

Metric	Oracle	Token Eviction	Page-Based	TokenButler
Llama-3.2-1B				
Accuracy	49.00	1.00	0	**49.00**
Coverage	84.32	32.50	19.78	**82.70**
Llama-3.2-3B				
Accuracy	81.00	10.00	6.00	**78.00**
Coverage	95.38	51.97	57.82	**93.50**
Llama-2-7b-hf				
Accuracy	77.00	18.00	1.00	**78.00**
Coverage	93.32	57.93	34.35	**94.00**
Llama-3.1-8B				
Accuracy	77.00	3.00	0	**73.00**
Coverage	93.47	37.50	46.98	**91.90**

Table 2. The accuracy and coverage of different KV-sparsity methods. As we can see, TokenButler approaches or sometimes exceeds Oracle in terms of accuracy and coverage.

To compare these approaches fairly, we simulate *token-by-token* decoding on the entire input to simulate generative tasks for standard benchmarks. This implies not having a prefill phase, and requiring H2O and SnapKV to apply their token eviction method at each step, rather than having access to the entire prefill attention map before generating a few tokens for the answer. This provides a more difficult task for token sparsity (for all methods equally) and more closely matches generative use-cases. It also tests whether TokenButler and Quest truly identify *and retain* the right tokens over the full sequence. Finally, we evaluate on perplexity and downstream tasks, revealing how token eviction can drop crucial context if done prematurely, and evaluating learned token importance metric in TokenButler.

StreamingLLM relies on *recency* as a guiding metric, maintaining attention only on the last W tokens within a fixed sliding window while discarding all older tokens, SnapKV on the other hand, determines token importance based on a rolling attention score magnitude, calculated over a small observation window of size 16/32 tokens. In our setup, this observation window is used solely for computing token importance but is not actively maintained unless the tokens it contains are deemed important by the metric itself. Similarly, H2O also employs QK-based importance but operates under a different mechanism—whenever a newly decoded token exhibits high attention magnitude, it evicts the least important token from its cache, provided the cache is full. Lastly, we include an Oracle baseline, which represents the best possible token sparsity achievable given full access to

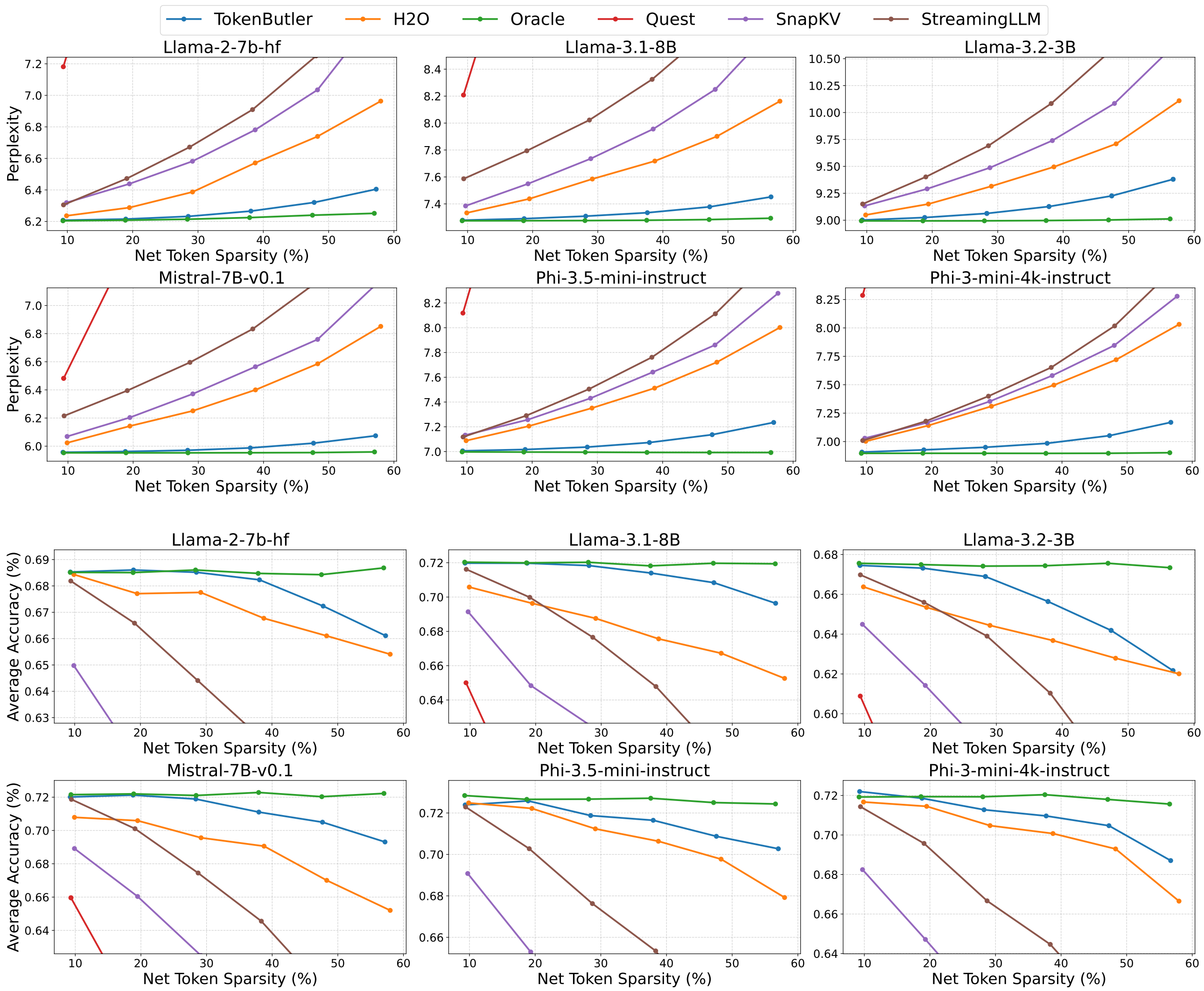

Figure 6. We evaluate TokenButler and other baselines in a uniform token pruning setting (method). We treat the entire input sequence as a *decode* task, and fix KV-Cache budget as a percentage of the *decoded sequence length*. Net Token Sparsity indicates the average observed token sparsity across all heads, with 4 anchor tokens and no sliding window attention. TokenButler outperforms other metrics at identifying critical tokens. H_2O and SnapKV evict crucial tokens during the decode simulation, and Quest incurs high rate of page-misses.

the LLM's attention logits. While this provides an upper bound on performance, it does not reduce computational costs, as requires a full attention pass to measure token importance before discarding unimportant tokens.

Our evaluation is done on perplexity and average of four downstream tasks (HellaSwag, ARC-Easy, PIQA and Wino-Grande) in zero-shot settings. Although these tasks are relatively simple, their critical tokens are often scattered across the entire context. Figure 6 shows the results. The *Oracle* baseline discards tokens *after* calculating their importance, and is thus nearly lossless even at 60% sparsity—revealing substantial redundancy. H2O also achieves decent results, but permanently discards tokens deemed unimportant early on, restricting later access when those tokens become relevant. Meanwhile, Quest's page-level metric underperforms on input lengths up to 1024, because a page size of 16 cannot flexibly capture tokens spread throughout the sequence on context dense tasks. By contrast, TokenButler accurately identifies important tokens in a fine-grained, query-dependent manner, consistently outperforming both eviction and page-based baselines in perplexity and downstream accuracy. From Figure 6, we can see that TokenButler in a fine-grained token access setting without prefill token eviction can offer up-to an 8% improvement in perplexity and downstream accuracy.

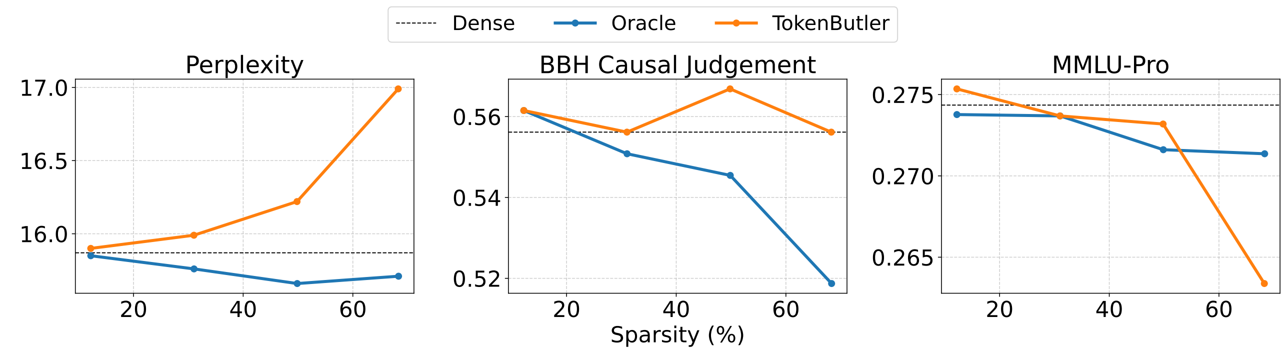

Figure 7. We train TokenButler on the `deepseek-ai/DeepSeek-R1-Distill-Llama-8B` model and evaluate its performance, comparing with a dense baseline and Oracle token pruning.

4.4. TokenButler On Reasoning Models

Reasoning models have been shown to have extremely long chain-of-thoughts. The generated CoT can significantly slow down decode, as well as cause significant increase in the KV-Cache, stressing the decode-time memory bandwidth. To reduce the memory-bandwidth overhead of excessive token-loading, we train TokenButler on the `DeepSeek-R1-Distill-Llama-8B` (DeepSeek-AI et al., 2025) model at 1% of the original model size, for 77M tokens using C4-*realnewslike*. We then evaluate TokenButlers perplexity, as well as two tasks from Open LLM Leaderboard (Fourrier et al., 2024) (BBH Causal Judgement (Kazemi et al., 2025) and MMLU Pro (Wang et al., 2024)) where the base reasoning model (`DeepSeek-R1-Distill-Llama-8B`) exhibits decent performance. From Figure 7, we can see that even at a very aggressive sparsity of 70%, TokenButler is able to preserve accuracy within 1%, and with a 2% increase in perplexity at 50% sparsity. This indicates that TokenButler can be used to reduce the memory and compute over-head of per-token decode on reasoning models well.

5. Discussion

Our experiments demonstrate that *fine-grained, per-head* token importance estimation can improve LLM performance on tasks that require retrieving previously referenced information. A key highlight is the stark difference between TokenButler's high-granularity, query-aware approach and existing token-eviction or page-level compression strategies. Methods like H2O and SnapKV tend to discard tokens prematurely under a size budget, limiting retrieval of critical context later. Page-based approaches (e.g. Quest) are better at retaining old tokens but cannot easily single out individually important tokens—particularly when references straddle page boundaries. Our synthetic co-reference benchmark highlights these issues: a single location name might be re-invoked well after it appears, yet it can get evicted or split across pages in favor of model performance.

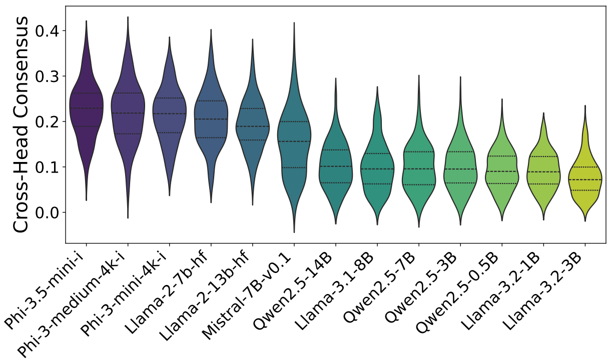

Figure 8. The plot depicts the mean correlation of attention logits between heads across the LLM. Different heads prefer different tokens, justifying the need for per-head TokenButler predictor design rather than a simpler per-layer approach.

An important observation from Figure 8 is that different heads can have drastically different token preferences. We test cross-head consensus, which is calculated by taking the attention logits from the *last* next-word prediction problem per sequence. We compute the correlation between attention logits across all heads in the LLMs. This gives us a $[NH, NH]$ correlation matrix, and we take the mean of the upper triangular matrix, giving us *mean cross-head* agreement (Cross-Head Consensus) in token preferences. The low correlation observed implies that preserving only a shared subset of tokens selected at the layer level (or from other heuristics) will lead to omission of tokens needed by other heads. TokenButler fixes this by dedicating a Q-K neural network to emulate all heads, ensuring that the tokens each head relies on for context remain accessible. While this slightly increases parameter count (by around 1%), we see a major improvement in perplexity and downstream performance at across token-sparsity levels.

In Figure 9, we first compare Oracle and *Oracle With Prefill Eviction*, which permanently evicts ”unimportant” tokens

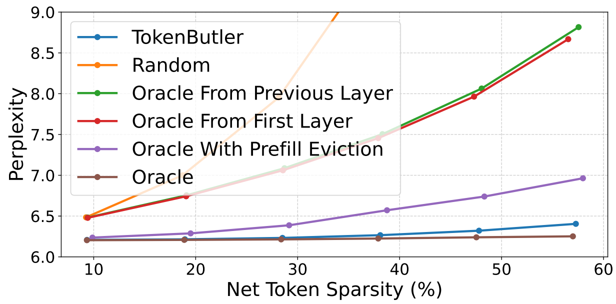

Figure 9. An ablation study on using Oracle token importance from the first layer vs. lookahead (previous layer) for the next layer. Both variants perform worse than Oracle with Prefill Eviction, however, TokenButler and Oracle is still significantly better on `Llama-2-7b-hf`.

after each next-word prediction. As previously seen, this degrades perplexity, but we also examine whether simpler signals—like reusing attention scores from the *first layer* or the *previous layer*—can guide subsequent layers' token choices without sacrificing tokens. Although such methods do beat a purely random token-dropping baseline, they still do not perform as well as even token eviction strategies. This is because of high cross-head disagreement, which means critical token choices vary widely. Further motivating our design of a decode-focused, fine-grained, per-head token importance prediction system.

6. Conclusion

In practice, our findings suggest that, to handle truly conversational or multi-round tasks—where new text keeps arriving and old tokens can become relevant again—LLMs benefit greatly from *retaining* rather than discarding. When memory limits necessitate a form of compression, it is important to do so in a query-aware, fine-grained manner. Our co-reference experiments show that all-or-nothing eviction or large-page retrieval strategies risk losing important information. TokenButler introduces a light predictor that tracks each head's token preference, preserving the tokens that *actually* matter. This results in up-to 8% gains in perplexity and downstream accuracy. Overall, our predictor substantially closes the gap to an oracle baseline at less than 2% latency overhead – paving the way for more precise token management techniques for large language models.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL `https://arxiv.org/abs/2404.14219`.

Akhauri, Y., AbouElhamayed, A. F., Dotzel, J., Zhang, Z., Rush, A. M., Huda, S., and Abdelfattah, M. S. ShadowLLM: Predictor-based contextual sparsity for large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19154–19167, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1068. URL `https://aclanthology.org/2024.emnlp-main.1068/`.

Akhauri, Y., Huda, S., and Abdelfattah, M. S. Attamba: Attending to multi-token states, 2024b. URL `https://arxiv.org/abs/2411.17685`.

Chen, Y., Dai, X., chih Chang, C., Akhauri, Y., and Abdelfattah, M. S. The power of negative zero: Datatype customization for quantized large language models, 2025. URL `https://arxiv.org/abs/2501.04052`.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., and Wolf, T. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Ge, S., Zhang, Y., Liu, L., Zhang, M., Han, J., and Gao, J. Model tells you what to discard: Adaptive kv cache compression for llms. In *The Twelfth International Conference on Learning Representations*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola,

B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Kazemi, M., Fatemi, B., Bansal, H., Palowitch, J., Anastasiou, C., Mehta, S. V., Jain, L. K., Aglietti, V., Jindal, D., Chen, P., Dikkala, N., Tyen, G., Liu, X., Shalit, U., Chiappa, S., Olszewska, K., Tay, Y., Tran, V. Q., Le, Q. V., and Firat, O. Big-bench extra hard, 2025. URL https://arxiv.org/abs/2502.19187.

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. Snapkv: Llm knows what you are looking for before generation. *arXiv preprint arXiv:2404.14469*, 2024.

Liu, Z., Wang, J., Dao, T., Zhou, T., Yuan, B., Song, Z., Shrivastava, A., Zhang, C., Tian, Y., Re, C., and Chen, B. Deja vu: Contextual sparsity for efficient LLMs at inference time. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22137–22176. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23am.html.

Luong, M.-T. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Levskaya, A., Heek, J., Xiao, K., Agrawal, S., and Dean, J. Efficiently scaling transformer inference. *arXiv preprint arXiv:2211.05102*, 2022.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Sun, H., Chang, L.-W., Bao, W., Zheng, S., Zheng, N., Liu, X., Dong, H., Chi, Y., and Chen, B. Shadowkv: Kv cache in shadows for high-throughput long-context llm inference, 2024. URL https://arxiv.org/abs/2410.21465.

Tang, J., Zhao, Y., Zhu, K., Xiao, G., Kasikci, B., and Han, S. Quest: Query-aware sparsity for efficient long-context llm inference, 2024.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Vodrahalli, K., Ontanon, S., Tripuraneni, N., Xu, K., Jain, S., Shivanna, R., Hui, J., Dikkala, N., Kazemi, M., Fatemi, B., Anil, R., Dyer, E., Shakeri, S., Vij, R., Mehta, H., Ramasesh, V., Le, Q., Chi, E., Lu, Y., Firat, O., Lazaridou, A., Lespiau, J.-B., Attaluri, N., and Olszewska, K. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024. URL https://arxiv.org/abs/2409.12640.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Yuan, A., Coenen, A., Reif, E., and Ippolito, D. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pp. 841–852, 2022.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*, 2023a.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023b.

A. Appendix

A.1. Synthetic Co-reference Benchmark

To rigorously evaluate token sparsity methods under retrieval-intensive scenarios, we developed a synthetic co-reference benchmark utilizing OpenAI's `gpt-4o-mini` model. The benchmark consists of 100 unique fictional location names, 100 paired location introductions and tieback questions, 100 philosophical reflections, 100 culinary descriptions, and 100 short math problems. Each data sample is constructed by randomly selecting one location introduction along with its corresponding tieback question, one location name, one philosophical statement, one culinary description, and one math problem. The resulting sequence is structured such that the location is introduced early in the context, followed by distractor content, and concludes with a prelude statement that prompts the retrieval of the original location name.

This modular generation approach allows for the creation of up to $100^4 = 10^8$ unique sequences by combining different components, ensuring extensive diversity. When a specific number of samples are requested, they are dynamically generated by randomly drawing from the respective pools of location introductions, location names, philosophical statements, culinary descriptions, and math problems. This on-the-fly sampling methodology ensures that each test instance presents a distinct retrieval challenge, effectively simulating real-world conversational dynamics where important tokens may reappear unpredictably after various interleaved topics. By designing the benchmark in this manner, we specifically target the capability of token sparsity methods to accurately retain and retrieve critical tokens between substantial contextual noise, thereby providing a robust assessment of their effectiveness in maintaining model performance on co-referential tasks.

We provide some *example* texts, and how they perform for different methods (Oracle, Prefill Eviction, Page Based and TokenButler) in the next few pages of the appendix. Each token is separated by a space for clarity. The tokens in *red* are the incorrectly predicted tokens at the last decode stage. The *underlined* tokens are the ones that the token-sparsity method decided to *preserve* when trying to decode the first token of the location word. The first occurence of the location tokens is in green, and their final appearance is bolded in green. Where the method failed to predict the tokens, their 'predicted' token is provided with a strike-through in red, and the *real* correct token is underlined in yellow in brackets.

Prefill Eviction Method

<|begin_of_text|>L oom ing above a shimmer ing sea of mist , a labyrinth of bi olum ines cent cavern s spark les like a hidden treasure chest beneath the surface . With each step , the walls pulse with vibrant colors , illumin ating the enchant ing mur als crafted by an unknown artisan from ages past . The place is : myst haven . In the manner of S ø ren K ier ke gaard , one might ponder that existence is not a mere state of being , but a relentless pursuit of authenticity amidst the anxiety of choice | each decision carving out the essence of the self , like a sculpt or unveiling a figure from un yield ing marble . Thus , we are et ern ally engaged in a dialogue with our own potential , whisper ing truth into the depths of uncertainty .. Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we calculate 26 + 15 - 9 , is the result 32 ? Indeed , it is 32 because 26 plus 15 equals 41 , and subtract ing 9 from 41 gives us 32 .. What mysterious location features a labyrinth of bi olum ines cent cavern s that pulse with vibrant colors and bear enchant ing mur als from an unknown artisan ?: ~~Myst~~ (myst) ~~ic~~ (haven)

Page Based Method

<|begin_of_text|>L oom ing above a shimmer ing sea of mist , a labyrinth of bi olum ines cent cavern s spark les like a hidden treasure chest beneath the surface . With each step , the walls pulse with vibrant colors , illumin ating the enchant ing mur als crafted by an unknown artisan from ages past . The place is : myst haven . In the manner of S ø ren K ier ke gaard , one might ponder that existence is not a mere state of being , but a relentless pursuit of authenticity amidst the anxiety of choice | each decision carving out the essence of the self , like a sculpt or unveiling a figure from un yield ing marble . Thus , we are et ern ally engaged in a dialogue with our own potential , whisper ing truth into the depths of uncertainty .. Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we calculate 26 + 15 - 9 , is the result 32 ? Indeed , it is 32 because 26 plus 15 equals 41 , and subtract ing 9 from 41 gives us 32 .. What mysterious location features a labyrinth of bi olum ines cent cavern s that pulse with vibrant colors and bear enchant ing mur als from an unknown artisan ?: ~~The~~ (myst) haven

Oracle

<|begin_of_text|>L oom ing above a shimmer ing sea of mist , a labyrinth of bi olum ines cent cavern s spark les like a hidden treasure chest beneath the surface . With each step , the walls pulse with vibrant colors , illumin ating the enchant ing mur als crafted by an unknown artisan from ages past . The place is : myst haven . In the manner of S ø ren K ier ke gaard , one might ponder that existence is not a mere state of being , but a relentless pursuit of authenticity amidst the anxiety of choice | each decision carving out the essence of the self , like a sculpt or unveiling a figure from un yield ing marble . Thus , we are et ern ally engaged in a dialogue with our own potential , whisper ing truth into the depths of uncertainty .. Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we calculate 26 + 15 - 9 , is the result 32 ? Indeed , it is 32 because 26 plus 15 equals 41 , and subtract ing 9 from 41 gives us 32 .. What mysterious location features a labyrinth of bi olum ines cent cavern s that pulse with vibrant colors and bear enchant ing mur als from an unknown artisan ?: myst haven

TokenButler

<|begin_of_text|>L oom ing above a shimmer ing sea of mist , a labyrinth of bi olum ines cent cavern s spark les like a hidden treasure chest beneath the surface . With each step , the walls pulse with vibrant colors , illumin ating the enchant ing mur als crafted by an unknown artisan from ages past . The place is : myst haven . In the manner of S ø ren K ier ke gaard , one might ponder that existence is not a mere state of being , but a relentless pursuit of authenticity amidst the anxiety of choice | each decision carving out the essence of the self , like a sculpt or unveiling a figure from un yield ing marble . Thus , we are et ern ally engaged in a dialogue with our own potential , whisper ing truth into the depths of uncertainty .. Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we calculate 26 + 15 - 9 , is the result 32 ? Indeed , it is 32 because 26 plus 15 equals 41 , and subtract ing 9 from 41 gives us 32 .. What mysterious location features a labyrinth of bi olum ines cent cavern s that pulse with vibrant colors and bear enchant ing mur als from an unknown artisan ?: myst haven

Prefill Eviction Method

<|begin_of_text|>B ene ath the lumin ous glow of a thousand glitter ing stars , a crystall ine lake sh immers with an eth ereal light that dances like fire flies on its surface . Sur rounded by mountains carved from colorful gem stones , whispers of long -lo st legends echo through the air , ming ling with the sweet fragrance of blo oming night flowers . The place is : *frost g len* . In the manner of K ier ke gaard , one might reflect : " True existence begins not in the relentless pursuit of outward cert ainties but in the brave plunge into the depth of one ’s own despair | where faith conce ives its true st essence .". Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we compute 8 + 15 − 5 , is the result 18 ? Indeed , it is 18 because 8 plus _ 15 equals 23 , and subtract ing 5 gives us 18 .. Which location boasts a crystall ine lake whose surface spark les like fire flies beneath a canopy of stars ?: ~~The~~ (frost) ~~y~~ (g) **len**

Page Based Method

<|begin_of_text|>B ene ath the lumin ous glow of a thousand glitter ing stars , a crystall ine lake sh immers with an eth ereal light that dances like fire flies on its surface . Sur rounded by mountains carved from colorful gem stones , whispers of long -lo st legends echo through the air , ming ling with the sweet fragrance of blo oming night flowers . The place is : *frost g len* . In the manner of K ier ke gaard , one might reflect : " True existence begins not in the relentless pursuit of outward cert ainties but in the brave plunge into the depth of one ’s own despair | where faith conce ives its true st essence .". Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we compute 8 + _ 15 − _ 5 , is the result 18 ? Indeed , it is 18 because _ 8 plus _ 15 equals _ 23 , and subtract ing _ 5 gives us _ 18 .. Which location boasts a crystall ine lake whose surface spark les like fire flies beneath a canopy of stars ?: ~~The~~ (frost) **g len**

Oracle

<|begin_of_text|>B ene ath the lumin ous glow of a thousand glitter ing stars , a crystall ine lake sh immers with an eth ereal light that dances like fire flies on its surface . Sur rounded by mountains carved from colorful gem stones , whispers of long -lo st legends echo through the air , ming ling with the sweet fragrance of blo oming night flowers . The place is : *frost g len* . In the manner of K ier ke gaard , one might reflect : " True existence begins not in the relentless pursuit of outward cert ainties but in the brave plunge into the depth of one ’s own despair | where faith conce ives its true st essence .". Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we compute 8 + 15 − 5 , is the result 18 ? Indeed , it is 18 because 8 plus _ 15 equals _ 23 , and subtract ing _ 5 gives us _ 18 .. Which location boasts a crystall ine lake whose surface spark les like fire flies beneath a canopy of stars ?: **frost g len**

TokenButler

<|begin_of_text|>B ene ath the lumin ous glow of a thousand glitter ing stars , a crystall ine lake sh immers with an eth ereal light that dances like fire flies on its surface . Sur rounded by mountains carved from colorful gem stones , whispers of long -lo st legends echo through the air , ming ling with the sweet fragrance of blo oming night flowers . The place is : *frost g len* . In the manner of K ier ke gaard , one might reflect : " True existence begins not in the relentless pursuit of outward cert ainties but in the brave plunge into the depth of one ’s own despair | where faith conce ives its true st essence .". Pl ump , sa ff ron -inf used Spanish pa ella gre ets the senses with its golden hue , as vibrant bell peppers , sweet peas , and tender m ors els of chor izo inter m ingle among a melody of br iny seafood | sh rimp , m ussels , and cock les gracefully nestled into a bed of al d ente rice . With each bite , the complex layers of smoked pap rika and lemon offer a taste of sun -d renched coast lines , echoing the spirited gatherings of festive Val enc ian fe asts , where laughter dances through the air like the enticing aroma rising from the pan .. If we compute 8 + 15 − 5 , is the result 18 ? Indeed , it is 18 because 8 plus 15 equals 23 , and subtract ing 5 gives us 18 .. Which location boasts a crystall ine lake whose surface spark les like fire flies beneath a canopy of stars ?: **frost g len**

Prefill Eviction Method

<|begin_of_text|>B ene ath a sky painted in eternal twilight , hills of lum ines cent moss und ulate like waves , whisper ing _secrets _to _the playful shadows of dancing fire flies . Cr yst all ine springs bubble _up , _their water shimmer ing _with hues that transcend _the spectrum , inviting weary _travelers _to drink from _their enchanted depths . The place _is : aer anth ium . In the spirit _of Conf uc ius , let _us ponder : " A _journey towards truth begins not _with certainty , _but _with the humility _to embrace our ignorance ; _for in every _question _lies _the _seed _of wisdom waiting _to spr out .". Golden sa ff ron -inf used bir y ani emerges _from _the pot , its frag rant _steam ming ling _with _the warm spices _of card am om and cloves , each pl ump bas mat i grain g listening _with _the essence _of mar inated chicken and tang y pick led lime ; a dish deeply rooted in South Asian heritage , it offers a harmon ious dance of sm oky , savory , and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 8 + 15 { 6 , is the result 17 ? Indeed , it is 17 because when we add 8 and _ 15 , we get 23 , and subtract ing 6 gives us 17 .. Which location features hills of lum ines cent moss that whisper secrets beneath an eternal twilight sky ?: ~~The~~ (aer) ~~ia~~ (anth) ~~os~~ (ium)

Page Based Method

<|begin_of_text|>B ene ath a sky painted in eternal twilight , hills of lum ines cent moss und ulate like waves , whisper ing secrets to _the playful shadows _of dancing fire flies . Cr yst all ine springs bubble up , their water shimmer ing _with hues that transcend the spectrum , inviting weary travelers _to drink from their enchanted depths . The place is : aer anth ium . In the spirit of Conf uc ius , let _us ponder : " A journey towards truth begins not _with certainty , _but _with the humility to embrace our ignorance ; for in every question _lies _the seed _of _wisdom waiting to spr out .". Golden sa ff ron -inf used bir y ani emerges from the pot , its frag rant _steam ming ling _with _the warm spices _of card am om and cloves , each pl ump bas mat i grain g listening _with _the essence _of mar inated chicken and tang y pick led lime ; a dish deeply rooted in South Asian heritage , it offers _a harmon ious dance of sm oky , savory , and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 8 + _ 15 { _ 6 , is the result 17 ? Indeed , _it _is _ 17 because when _we add _ 8 _and _ 15 , _we _get _ 23 , and subtract ing _ 6 gives _us _ 17 .. Which location features hills of lum ines cent moss that whisper secrets beneath an eternal twilight sky ?: ~~The~~ (aer) anth ium

Oracle

<|begin_of_text|>B ene ath a sky painted in eternal twilight , hills of lum ines cent moss und ulate like waves , whisper ing secrets _to _the playful shadows of dancing fire flies . Cr yst all ine springs bubble _up , _their water shimmer ing _with hues that transcend _the _spectrum , inviting weary travelers _to drink from _their enchanted depths . The place is : aer anth ium . In the spirit of Conf uc ius , let us ponder : " A journey towards truth begins not _with certainty , _but _with the humility _to embrace our ignorance ; _for in every _question _lies _the _seed _of wisdom waiting _to spr out .". Golden sa ff ron -inf used bir y ani emerges _from _the pot , _its frag rant _steam ming ling _with _the warm spices _of card am om _and cloves , each pl ump bas mat i _grain g listening _with _the essence _of mar inated chicken _and tang y pick led lime ; _a dish deeply rooted in South Asian heritage , it offers _a harmon ious dance _of sm oky , savory , _and z esty notes that transports din ers across bustling _b aza ars and op ulent fe asts of ancient kings .. If we calculate 8 + 15 { 6 , is the result 17 ? Indeed , it is 17 because when we add 8 _and _ 15 , _we _get _ 23 , _and subtract ing _ 6 gives _us _ 17 .. Which location features hills of lum ines cent moss that whisper secrets beneath an eternal twilight sky ?: _aer anth ium

TokenButler

<|begin_of_text|>B ene ath a sky painted in eternal twilight , hills of lum ines cent moss und ulate like _waves , whisper ing secrets _to _the playful shadows of dancing fire flies . Cr yst all ine springs bubble _up , _their water shimmer ing _with hues that transcend _the spectrum , inviting weary travelers _to drink from _their enchanted depths . _The place _is : aer anth ium . In the spirit _of Conf uc ius , let _us ponder : " A _journey towards truth begins not _with certainty , _but _with the humility _to embrace our ignorance ; _for in every question _lies _the seed _of wisdom waiting _to spr out .". Golden sa ff ron -inf used bir y ani emerges _from _the pot , _its frag rant steam ming ling _with _the warm spices _of card am om _and cloves , each pl ump bas mat i grain g listening _with _the essence _of mar inated chicken _and tang y pick led lime ; _a dish deeply _rooted _in South Asian heritage , it offers _a harmon ious dance _of sm oky , _savory , _and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 8 + 15 { 6 , is _the result 17 ? Indeed , it _is 17 because when we add 8 and 15 , we get 23 , and subtract ing 6 gives us 17 .. Which location features hills of lum ines cent moss that whisper secrets beneath an eternal twilight sky ?: _aer anth ium

Prefill Eviction Method

<|begin_of_text|>Ve iled in an everlasting twilight , this realm _is painted with lum ines cent flowers that bloom under _the soft _glow _of a silver sky . Rivers _of shimmer ing st ard ust me ander _through em erald fields , whisper ing _secrets _to those _who dare _to _listen . _The place _is : frost wyn . In the manner _of S ø ren _K ier ke gaard , one might ponder that existence _is not _a mere _state _of _being , _but _a relentless pursuit _of authenticity amidst _the anxiety _of _choice | each _decision carving out the essence _of the _self , like _a _sculpt or unveiling _a figure _from un yield ing marble . Thus , we are et ern ally engaged _in a dialogue _with _our _own potential , whisper ing truth into _the depths _of uncertainty .. Golden sa ff ron -inf used bir y ani emerges _from the pot , its frag rant steam ming ling _with _the warm spices of card am om and cloves , each pl ump bas mat i grain g listening _with _the essence _of mar inated chicken and tang y pick led lime ; a dish deeply rooted in South Asian heritage , it offers a harmon ious dance of sm oky , savory , and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 27 + 15 - 8 , is the result 34 ? Indeed , it is 34 because 27 plus _ 15 equals 42 , and subtract ing 8 gives us 34 .. Which realm is adorned with lum ines cent flowers and rivers of shimmer ing st ard ust that whisper secrets to those who listen ?: ~~Frost~~ (frost) wyn

Page Based Method

<|begin_of_text|>Ve iled in an everlasting twilight , this realm is painted with lum ines cent flowers that _bloom under _the soft glow _of a silver sky . Rivers _of shimmer ing st ard ust me ander through em erald fields , whisper ing secrets to those _who dare to listen . _The place is : frost wyn . In the manner of S ø ren K ier ke gaard , one might ponder that existence _is not _a mere state _of being , _but _a relentless pursuit of authenticity amidst the anxiety of choice | each decision carving _out the essence _of the _self , like _a sculpt or unveiling a figure from un yield ing marble . Thus , we are et ern ally engaged in a dialogue _with our own potential , whisper ing truth into _the depths of uncertainty .. Golden sa ff ron -inf used bir y ani emerges from the pot , its frag rant steam ming ling _with _the warm spices of card am om and cloves , each pl ump bas mat i grain g listening _with _the essence _of mar inated chicken and tang y pick led lime ; a dish deeply rooted in South Asian heritage , it offers _a harmon ious dance of sm oky , savory , _and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 27 + _ 15 - _ 8 , is _the result 34 ? Indeed , _it _is 34 because 27 plus _ 15 equals 42 , and subtract ing _ 8 gives us 34 .. Which realm is adorned with lum ines cent _flowers and rivers _of _shimmer ing st ard ust that whisper _secrets _to _those _who listen ?: ~~the~~ (frost) wyn

Oracle

<|begin_of_text|>Ve iled in an everlasting twilight , this realm is painted with lum ines cent flowers that bloom under _the soft _glow _of a silver sky . Rivers _of shimmer ing st ard ust me ander _through em erald fields , whisper ing _secrets _to those _who dare _to _listen . The place is : frost wyn . In the manner of S ø ren K ier ke gaard , one might ponder that existence _is not _a mere _state _of _being , _but _a relentless pursuit _of authenticity amidst _the anxiety _of _choice | each _decision carving out the essence _of the _self , like _a _sculpt or unveiling _a figure _from un yield ing marble . Thus , we are et ern ally engaged _in a dialogue _with _our _own potential , whisper ing truth into _the depths _of uncertainty .. Golden sa ff ron -inf used bir y ani emerges _from _the pot , its frag rant _steam ming ling _with _the warm spices _of card am om _and cloves , each pl ump bas mat i _grain g listening _with _the essence _of mar inated chicken _and tang y pick led lime ; _a dish deeply rooted in South Asian heritage , it offers _a harmon ious dance _of sm oky , savory , _and z esty notes that transports din ers across bustling b aza ars _and op ulent fe asts of ancient kings .. If we calculate 27 + 15 - _ 8 , is the result 34 ? Indeed , it is 34 because 27 plus _ 15 equals _ 42 , _and subtract ing _ 8 gives us _ 34 .. Which realm is adorned with lum ines cent flowers and rivers of shimmer ing st ard ust that whisper secrets to those who listen ?: _frost wyn

TokenButler

<|begin_of_text|>Ve iled in an everlasting twilight , this realm _is painted with lum ines cent flowers that bloom under _the soft glow _of a silver sky . Rivers _of shimmer ing st ard ust me ander _through em erald fields , whisper ing secrets to those _who _dare _to _listen . The place _is : frost wyn . In the manner _of S ø ren _K ier ke gaard , one might ponder that existence _is not _a mere _state _of _being , _but _a relentless pursuit _of authenticity amidst _the anxiety _of _choice | each _decision carving out the essence _of the _self , like _a sculpt or unveiling _a figure _from un yield ing _marble . Thus , we are et ern ally engaged _in a dialogue _with _our _own potential , whisper ing truth into _the depths _of uncertainty .. Golden sa ff ron -inf used bir y ani emerges _from _the pot , _its frag rant steam ming ling _with _the warm spices of card am om _and cloves , each pl ump bas mat i _grain g listening _with _the essence _of mar inated chicken _and tang y pick led lime ; _a dish deeply _rooted _in South Asian heritage , it offers _a harmon ious dance _of sm oky , savory , _and z esty notes that transports din ers across bustling b aza ars and op ulent fe asts of ancient kings .. If we calculate 27 + 15 - 8 , is the result 34 ? Indeed , it _is 34 because 27 plus 15 equals 42 , and subtract ing 8 gives us 34 .. Which realm is adorned with lum ines cent flowers and rivers of shimmer ing st ard ust that whisper secrets to those who listen ?: _frost wyn

Prefill Eviction Method

<|begin_of_text|>C rad led between swirling s apphire t ides , this breathtaking haven sh immers under a kale idos cope sky where the sun paints the horizon in vibrant hues of gold and rose . Myst ical creatures roam among phosph ores cent flora that emit a soft glow at twilight , illumin ating secret paths leading to hidden water falls that whisper secrets of the anc ients . The place is : myst ar idge . In the manner of S ø ren K ier ke gaard , one might say : \ To exist is to confront deeply personal truths that demand decision ; yet , in every choice made , a horizon not yet explored van ishes , revealing the paradox that our freedom shapes us even as we define ourselves through constraint ." . A ste aming plate of P anna C otta g list ens with a jewel -ton ed raspberry cou lis that casc ades over its vel v ety cream , ev oking the sun -d renched landscapes of Italy ; each spoon ful dances on the tongue with rich notes of vanilla and a whisper of citrus , a perfect end to a rustic meal , paying homage to centuries of culinary art istry .. If we calculate 18 + 22 { 7 , is the result 33 ? Indeed , it is 33 because 18 plus 22 equals 40 , and when we subtract 7 from 40 , we get 33 .. Which breathtaking haven is cr ad led between swirling s apphire t ides and features secret paths illuminated by phosph ores cent flora ?: ~~Myst~~ (myst) ar ~~.~~ (idge)

Page Based Method

<|begin_of_text|>C rad led between swirling s apphire t ides , this breathtaking haven sh immers under a kale idos cope sky where the sun paints the horizon in vibrant hues of gold and rose . Myst ical creatures roam among phosph ores cent flora that emit a soft glow at twilight , illumin ating secret paths leading to hidden water falls that whisper secrets of the anc ients . The place is : myst ar idge . In the manner of S ø ren K ier ke gaard , one might say : \ To exist is to confront deeply personal truths that demand decision ; yet , in every choice made , a horizon not yet explored van ishes , revealing the paradox that our freedom shapes us even as we define ourselves through constraint ." . A ste aming plate of P anna C otta g list ens with a jewel -ton ed raspberry cou lis that casc ades over its vel v ety cream , ev oking the sun -d renched landscapes of Italy ; each spoon ful dances on the tongue with rich notes of vanilla and a whisper of citrus , a perfect end to a rustic meal , paying homage to centuries of culinary art istry .. If we calculate 18 + _ 22 { _ 7 , is the result 33 ? Indeed , it is _ 33 because _ 18 plus _ 22 equals 40 , and when we subtract _ 7 from _ 40 , we get 33 .. Which breathtaking haven is cr ad led between swirling s apphire t ides and features secret paths illuminated by phosph ores cent flora ?: ~~The~~ (myst) ar idge

Oracle

<|begin_of_text|>C rad led between swirling s apphire t ides , this breathtaking haven sh immers under a kale idos cope sky where the sun paints the horizon in vibrant hues of gold and rose . Myst ical creatures roam among phosph ores cent flora that emit a soft glow at twilight , illumin ating secret paths leading to hidden water falls that whisper secrets of the anc ients . The place is : myst ar idge . In the manner of S ø ren K ier ke gaard , one might say : \ To exist is to confront deeply personal truths that demand decision ; yet , in every choice made , a horizon not yet explored van ishes , revealing the paradox that our freedom shapes us even as we define ourselves through constraint ." . A ste aming plate of P anna C otta g list ens with a jewel -ton ed raspberry cou lis that casc ades over its vel v ety cream , ev oking the sun -d renched landscapes of Italy ; each spoon ful dances on the tongue with rich notes of vanilla and a whisper of citrus , a perfect end to a rustic meal , paying homage to centuries of culinary art istry .. If we calculate 18 + 22 { _ 7 , is the result 33 ? Indeed , it is 33 because 18 plus _ 22 equals _ 40 , and when we subtract _ 7 from _ 40 , we get _ 33 .. Which breathtaking haven is cr ad led between swirling s apphire t ides and features secret paths illuminated by phosph ores cent flora ?: myst ar idge

TokenButler

<|begin_of_text|>C rad led between swirling s apphire t ides , this breathtaking haven sh immers under a kale idos cope sky where the sun paints the horizon in vibrant hues of gold and rose . Myst ical creatures roam among phosph ores cent flora that emit a soft glow at twilight , illumin ating secret paths leading to hidden water falls that whisper secrets of the anc ients . The place is : myst ar idge . In the manner of S ø ren K ier ke gaard , one might say : \ To exist is to confront deeply personal truths that demand decision ; yet , in every choice made , a horizon not yet explored van ishes , revealing the paradox that our freedom shapes us even as we define ourselves through constraint ." . A ste aming plate of P anna C otta g list ens with a jewel -ton ed raspberry cou lis that casc ades over its vel v ety cream , ev oking the sun -d renched landscapes of Italy ; each spoon ful dances on the tongue with rich notes of vanilla and a whisper of citrus , a perfect end to a rustic meal , paying homage to centuries of culinary art istry .. If we calculate 18 + 22 { _ 7 , is the result 33 ? Indeed , it is 33 because 18 plus 22 equals 40 , and when we subtract 7 from 40 , we get 33 .. Which breathtaking haven is cr ad led between swirling s apphire t ides and features secret paths illuminated by phosph ores cent flora ?: mystar idge