Get More with LESS: Synthesizing Recurrence with KV Cache Compression for Efficient LLM Inference

Harry Dong [1] **Xinyu Yang** [1] **Zhenyu Zhang** [2] **Zhangyang Wang** [2] **Yuejie Chi** [1] **Beidi Chen** [1 3]

Abstract

Many computational factors limit broader deployment of large language models. In this paper, we focus on a memory bottleneck imposed by the key-value (KV) cache, a computational shortcut that requires storing previous KV pairs during decoding. While existing KV cache methods approach this problem by pruning or evicting large swaths of relatively less important KV pairs to dramatically reduce the memory footprint of the cache, they can have limited success in tasks that require recollecting a majority of previous tokens. To alleviate this issue, we propose LESS, a simple integration of a (nearly free) constant sized cache with eviction-based cache methods, such that all tokens can be queried at later decoding steps. Its ability to retain information throughout time shows merit on a variety of tasks where we demonstrate LESS can help reduce the performance gap from caching everything, sometimes even matching it, all while being efficient. Relevant code can be found at https://github.com/hdong920/LESS.

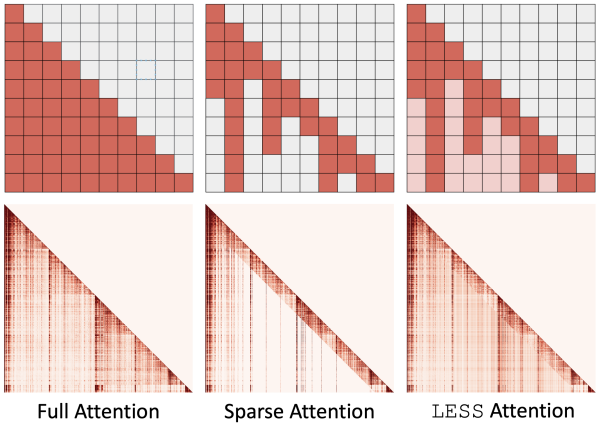

Figure 1. Toy (top row) and Llama 2 7B (bottom row) example decoder attention maps with H_2O as the underlying sparse policy. In the top row, red/pink and grey squares are positive and zero attention probabilities, respectively. In the bottom row, darker colors indicate larger attention probabilities. Sparse attention policies zero out many positive attention probabilities. Our method, LESS, ensures all previous tokens will have some contribution to the attention layer output to better retain information.

1. Introduction

Throughout its lifetime, the transformer architecture (Vaswani et al., 2017) has made strides in natural language processing (Lin et al., 2022), computer vision (Khan et al., 2022), healthcare (Nerella et al., 2023), and many other domains. Large language models (LLMs) (Zhang et al., 2022; Scao et al., 2022; Fedus et al., 2022; Anil et al., 2023; OpenAI, 2023; Touvron et al., 2023; Team et al., 2023; Jiang et al., 2024) take transformers to the extreme by scaling the model, data, and context lengths to extraordinary levels. This has been remarkably useful for complex tasks such as chatbots, long document tasks, and biological sequences. However, during deployment, these tasks require generating long sequences or inputting large batch sizes, which places an immense computational burden on the key-value (KV) cache (Pope et al., 2023), the storage of all previous keys and values at each layer to bypass recomputing them at future decoding steps. While this significantly saves computation, the tradeoff is an explosion of memory consumption. In such scenarios, the KV cache size often eclipses the model size. For instance, the Llama 2 7B model (Touvron et al., 2023) occupies about 26 GB of memory, but the KV cache for an input of batch size 64 and sequence length 1024 occupies 64 GB of memory, nearly 2.5 times the model size. Hence, addressing this accessibility issue is imperative as LLMs continue to scale and break tight deployment constraints.

Thankfully, there have been initiatives to reduce the KV cache size. A line of work, in which we refer to as *sparse policies or algorithms*, explores the selection of the best

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, USA [2]Department of Electrical and Computer Engineering, University of Texas at Austin, USA [3]Meta AI (FAIR), USA. Correspondence to: Harry Dong <harryd@andrew.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

subset of KV pairs to cache (Zhang et al., 2023; Liu et al., 2023b; Han et al., 2023; Xiao et al., 2023). Although very promising, these methods are inevitably and irrecoverably discarding KV pairs deemed, in one way or another, less important than others, leading to gaps in attention maps as shown in Figure 1. Consequently, they are boldly assuming tokens that are unimportant now will not hold significance at future decoding steps, a faulty conjecture for tasks that deviate from this pattern. For instance, using sparse policy H_2O (Zhang et al., 2023) on Falcon 7B (Almazrouei et al., 2023) to summarize an article (BBC, 2015; Narayan et al., 2018) produces a *factually incorrect* summary in Figure 2.

Article: ...ban on fracking in Wales...support for oil and gas applications is contrary to the approach of the Welsh Government of promoting renewable and low carbon forms of energy...

(Incorrect) Sparse Policy Summary: The Welsh government has announced measures to speed up the development of fracking.

Figure 2. Incorrect summary by Falcon 7B with sparse policy H_2O.

For the full article, see Figure 13 in Appendix D. One way to combat information loss is to cache more tokens, but this is far from memory efficient. An ideal KV cache policy should **1)** minimize performance degradation from a full cache, **2)** scale at a much slower rate than the full KV cache, and **3)** be cheap to integrate into existing pretrained LLMs.

Fortunately, with some investigation into the *residual* between full and sparse attention outputs, a better strategy emerges. First, define the residual as $\boldsymbol{\Delta}_{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{A}_{\text{sparse}}$, where $\boldsymbol{A}$ and $\boldsymbol{A}_{\text{sparse}}$ are the full and sparse attention outputs, respectively. Using top-k selection as our sparse policy, we observe the residuals $\boldsymbol{\Delta}_{\boldsymbol{A}}$ are in fact *low-rank* — more so than $\boldsymbol{A}$ — based on Figure 3, a similar observation to Chen et al. (2021). Even a very low-rank approximation can nearly negate the performance degradation from sparse caching. *In turn, this finding motivates the use of low-rank methods to approximate the residuals for efficient caches*.

We propose LESS (**L**ow-rank **E**mbedding **S**idekick with **S**parse policy) to learn the *residual* between the original attention output and the attention output approximated by a sparse policy. LESS does this by accumulating information that would have been discarded by sparse policies into a constant-sized low-rank cache or state, allowing for queries to still access information to recover previously omitted regions in attention maps (see Figure 1). We show that LESS makes significant progress towards an ideal cache:

1. **Performance Improvement:** LESS synthesizes sparse KV policies with low-rank states to bridge the performance gap on a variety of tasks where these sparse algorithms show cracks of weakness. In fact, LESS improves the performance much more than simply dedicating that memory to storing more KV pairs.

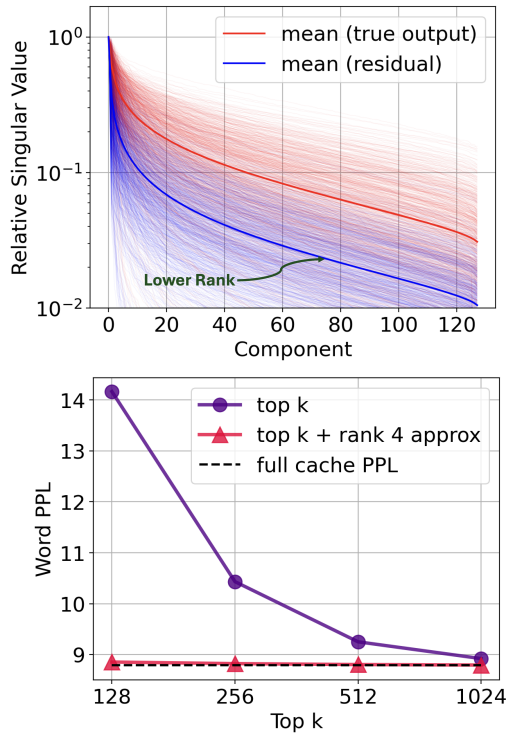

Figure 3. Attention residuals exploration in Llama 2 7B on WikiText (Merity et al., 2016). Mean and 1000 sample relative singular value plots of true attention outputs and residuals from top-512 sparse policy, showing the residual is much lower rank (left). End-to-end performance (lower is better) using top-k caching with and without low-rank approximations (right). A rank-4 approximation virtually recovers the original performance.

2. **Constant Low-rank Cache Size:** Low-rank caches in LESS occupy constant memory with respect to the sequence length, and in our experiments, the extra storage to accommodate LESS is *nearly free*, taking up the equivalent space of only 4 extra KV pairs in our experiments. Inspired by recurrent networks, the low-rank state stores new information by recursive updates rather than concatenation. As each sample has its own cache, LESS provides the same proportional cache reduction for small and large batch sizes.

3. **Cheap Integration:** Changes to the LLMs' architectures are small and do not perturb the original weights. The only modifications to LLMs will be the addition of tiny multilayer perceptions (MLPs) at each attention layer. For example, using LESS with Llama 2 13B adds fewer than 2% of the total number of parameters. In addition, we can train LESS at each attention layer independently from all others, bypassing expensive end-to-end training. Trained once, LESS can transfer

to more relaxed settings while maintaining comparable performance, further extending its applicability.

Our comprehensive experiments on Llama 2 (Touvron et al., 2023) and Falcon (Almazrouei et al., 2023) with different sparse policies (Zhang et al., 2023; Han et al., 2023; Xiao et al., 2023) on a variety of tasks demonstrates LESS as a highly performative method that reduces KV cache memory. For instance, LESS recovers more than 40% of the Rouge-1 degradation caused by a sparse policy on the CNN/DailyMail dataset (Hermann et al., 2015; See et al., 2017) with Falcon 7B. Finally, we provide an implementation of LESS that reduces the latency by up to $1.3\times$ and increases the throughput by $1.7\times$ from the full cache.

Notation. We use unbolded letters (e.g. a, A), bold lowercase letters (e.g. $\boldsymbol{a}$), bold uppercase letters (e.g. $\boldsymbol{A}$) for scalars, row vectors, and matrices, respectively. Let $[\boldsymbol{A}]_{i,\cdot}$ and $[\boldsymbol{A}]_{\cdot,i}$ refer to the i-th row and column of $\boldsymbol{A}$, respectively. Additionally, define $\mathbf{0}_{n\times m}$ as a matrix of zeros and $\mathbf{1}_{n\times m}$ as a matrix of ones, both having shape $n \times m$.

2. Background & Intuition

We start by building the intuition behind LESS. Sparse and low-rank caches individually have noteworthy advantages but also severe drawbacks. Understanding the mechanisms of both (Section 2.1 and 2.2) allows us to effectively synthesize sparse and low-rank structures to create LESS. In Section 2.3, we show that this type of synthesis is a principled approach which has also found success in other areas.

2.1. KV Cache Policies

Many current methods to reduce the KV cache footprint involve keeping a tiny subset of the keys and values either with some pruning policy (Liu et al., 2023b; Zhang et al., 2023; Han et al., 2023; Xiao et al., 2023; Ge et al., 2023; Oren et al., 2024) or a local attention mechanism (Child et al., 2019; Parmar et al., 2018). The former method can be applied directly to trained models whereas the latter typically cannot, so with limited compute, deploying a KV cache pruning policy is more practical. Such methods take advantage of the observation that many tokens are irrelevant for attention in some tasks and thus omitting them leads to negligible performance loss. For instance, one of our baselines, H_2O (Zhang et al., 2023), continuously accumulates attention probabilities at each generation step to identify a set of heavy-hitting tokens to be cached together with the most recent tokens. Not explicitly designed for KV cache compression, algorithms for infinite inference (Han et al., 2023; Xiao et al., 2023) maintain a full cache, but as the input sequence exceeds the maximum context length of a model, KV pairs in the middle of the sequence are dropped. Staying within the maximum context length, this results in a cache that maintains the most recent and first few tokens. Regardless of the sparse method, maintaining a tight KV cache budget can seriously impair model performance, as we will see in Section 4.

There also exist promising non-eviction based methods. DMC involves using a large amount of data to fine tune models to choose between accumulating or appending each KV pair (Nawrot et al., 2024). CacheGen's KV cache compression at the bit-level takes a query-agnostic approach (Liu et al., 2023a). In vision tasks, token merging is an effective way to cut down the number of tokens to process (Bolya et al., 2022; Renggli et al., 2022).

2.2. Low-rank Attention

Low-rank structures in attention have been explored extensively (Tay et al., 2022), namely from the lens of recurrent neural networks (RNNs). Unlike transformers, RNNs integrate information from all previous tokens into *hidden states*, analogous low-rank structures to KV caches that organically occupy constant memory. In fact, this feature of RNNs over transformers has motivated research in alternative architectures (Dao et al., 2022; Poli et al., 2023; Peng et al., 2023; Sun et al., 2023; Gu & Dao, 2023), but for now, their adoption in LLMs is very limited compared to transformers. Though not as performative as these alternative architectures, linear transformers that break apart the attention operation into kernels also maintain a constant sized KV cache (Tsai et al., 2019; Katharopoulos et al., 2020; Choromanski et al., 2020; Peng et al., 2021) by reformulating the cache into an RNN hidden state. These types of caching mechanisms are *low-rank* since information is condensed along the sequence axis, rather than explicitly maintaining individual tokens. This is possible when we replace the softmax with a separable similarity metric $\phi(\boldsymbol{q}_t)\psi(\boldsymbol{K}_t)^\top$ for some row-wise functions ϕ and ψ, letting $\boldsymbol{q}_t \in \mathbb{R}^{1\times D}$ and $\boldsymbol{K}_t \in \mathbb{R}^{t\times D}$ be the query and keys at step t, respectively. To elaborate, if ϕ and ψ are such that

$$\boldsymbol{a}_t = \text{softmax}\left(\frac{\boldsymbol{q}_t\boldsymbol{K}_t^\top}{\sqrt{D}}\right)\boldsymbol{V}_t \approx \frac{\phi(\boldsymbol{q}_t)\psi(\boldsymbol{K}_t)^\top\boldsymbol{V}_t}{\phi(\boldsymbol{q}_t)\psi(\boldsymbol{K}_t)^\top\mathbf{1}_{S\times 1}},$$

we just need to cache hidden states $\boldsymbol{H}_t = \psi(\boldsymbol{K}_t)^\top\boldsymbol{V}_t \in \mathbb{R}^{R\times D}$ and $\boldsymbol{z}_t = \sum_{s=1}^{t}\psi([\boldsymbol{K}_t]_s) \in \mathbb{R}^{1\times R}$ for inference which can be expressed recursively as

$$\begin{aligned}\boldsymbol{H}_{t+1} &= \boldsymbol{H}_t + \psi(\boldsymbol{k}_t)^\top\boldsymbol{v}_t,\\ \boldsymbol{z}_{t+1} &= \boldsymbol{z}_t + \psi(\boldsymbol{k}_t)\end{aligned}$$

for each new KV pair $(\boldsymbol{k}_t, \boldsymbol{v}_t)$. At initialization, $\boldsymbol{H}_0 = \mathbf{0}_{R\times D}$ and $\boldsymbol{z}_0 = \mathbf{0}_{1\times R}$. This is a clear improvement from having to store ever increasing sizes of $\boldsymbol{K}_t$ and $\boldsymbol{V}_t$, as the memory consumption is independent from t. Note that our presentation differs slightly since we do not constrain $\phi = \psi$ (Chen et al., 2023). With this formulation, transformers act

like RNNs which occupy constant memory during generation by not appending but updating hidden states during each generation step. Since LLMs are not typically trained in this fashion, a major challenge is to induce this property without significant computation or adjustment to the original weights (Kasai et al., 2021). While its dilution of information restricts its performance when specific tokens need to be recalled with strong signals (Khandelwal et al., 2018), this is exactly what a sparse KV cache algorithm can do, so we can fully take advantage of its infinite compression capability to obtain some high level representation of the less important tokens, meaning kernelized attention is a good candidate method for LESS to learn the residual.

2.3. Sparse and Low-rank Decomposition

LESS follows a rich history of decomposing structures into sparse and low-rank components. Particularly, the study of robust principal component analysis (Candès et al., 2011; Chandrasekaran et al., 2011) has shown this type of decomposition greatly enhances approximation accuracy and expressibility beyond just sparse or low-rank matrices alone. Its success has spread to deep learning areas such as efficient attention (Chen et al., 2021), model compression (Li et al., 2023), and fine-tuning (Nikdan et al., 2024). Likewise, we take inspiration from these works in our design.

3. Method

When we convert the intuition in Section 2 into an algorithm, a couple technical challenges arise. One challenge is finding an effective way to mix attention probabilities produced by sparse policies and low-rank kernels. Second, we need to design a framework general enough to work with a broad class of sparse policies. In some cases, different sparse policies may be preferable, so our method should be compatible with many sparse policies. Third, our method should be cheap compute to develop. We show that LESS overcomes all these challenges in a two step process: attention computation followed by cache updates.

3.1. KV Caching with LESS

We propose LESS, a general method to synthesize low-rank caches with *any* eviction-based sparse KV cache policy, $\mathfrak{C}$, to close the performance gap from full KV caching while being efficient. Notably, our method only adds a constant sized cache which does not scale with the sequence length. For the sparse policy, $\mathfrak{C}$, we require that it can output the cached keys $\boldsymbol{K}_{\mathfrak{C},t} \in \mathbb{R}^{B_t \times D}$, the cached values $\boldsymbol{V}_{\mathfrak{C},t} \in \mathbb{R}^{B_t \times D}$, and the set of discarded KV pairs $\mathcal{D}_t$ at iteration t where $B_t \in \mathbb{N}$ is the number of cached pairs.

Letting $\cdot$ denote both ϕ and ψ, we define our kernels as

$$\phi(\boldsymbol{q}) = |\sigma_\phi(\sigma_\phi(\boldsymbol{q}\boldsymbol{W}_{\phi,1})\boldsymbol{W}_{\phi,2})| \tag{1}$$

$$\psi(\boldsymbol{k}) = |\sigma_\psi(\sigma_\psi(\boldsymbol{k}\boldsymbol{W}_{\psi,1})\boldsymbol{W}_{\psi,2})\boldsymbol{W}_{\psi,3}| \tag{2}$$

for activation functions $\sigma_\cdot$, $\boldsymbol{W}_{\cdot,1} \in \mathbb{R}^{D \times R'}$, $\boldsymbol{W}_{\cdot,2} \in \mathbb{R}^{R' \times R}$, and $\boldsymbol{W}_{\psi,3} \in \mathbb{R}^{R \times R}$. The element-wise absolute values ensure the inner product $\phi(\boldsymbol{q})\psi(\boldsymbol{k})^\top > 0$ to preserve the nonnegativity of attention probabilities. In the ideal case, if $\phi(\boldsymbol{q})\psi(\boldsymbol{k})^\top = e^{\boldsymbol{q}\boldsymbol{k}^\top/\sqrt{D}}$ for all $\boldsymbol{q}, \boldsymbol{k}$, then the result would be the original attention probabilities.

Attention Calculation. Now, we describe the attention calculation procedure summarized in Algorithm 1. At step t, we find the query-key-value triplet $(\boldsymbol{q}_t, \boldsymbol{k}_t, \boldsymbol{v}_t)$ from the input token as usual. Recalling that we have cached $\boldsymbol{K}_{\mathfrak{C},t}$, $\boldsymbol{V}_{\mathfrak{C},t}$, $\boldsymbol{H}_t$, and $\boldsymbol{z}_t$ from the previous generation step, append $\boldsymbol{k}_t$ to $\boldsymbol{K}_{\mathfrak{C},t}$ and $\boldsymbol{v}_t$ to $\boldsymbol{V}_{\mathfrak{C},t}$ to obtain $\boldsymbol{K}'_{\mathfrak{C},t} \in \mathbb{R}^{(B_t+1) \times D}$ and $\boldsymbol{V}'_{\mathfrak{C},t} \in \mathbb{R}^{(B_t+1) \times D}$, respectively. Then, we can find $\hat{\boldsymbol{a}}_t$, our approximation of the original attention $\boldsymbol{a}_t$, by computing

$$\hat{\boldsymbol{a}}_t = \frac{\phi(\boldsymbol{q}_t)\boldsymbol{H}_t + \exp(\boldsymbol{q}_t(\boldsymbol{K}'_{\mathfrak{C},t})^\top/\sqrt{D})\boldsymbol{V}'_{\mathfrak{C},t}}{\phi(\boldsymbol{q}_t)\boldsymbol{z}_t^\top + \exp(\boldsymbol{q}_t(\boldsymbol{K}'_{\mathfrak{C},t})^\top/\sqrt{D})\boldsymbol{1}_{B \times 1}}. \tag{3}$$

During the prompting phase (i.e. $t = 0$), it is just regular attention since $\boldsymbol{H}_0 = \boldsymbol{0}_{R \times D}$ and $\boldsymbol{z}_0 = \boldsymbol{0}_{1 \times R}$.

Cache Updates. With the attention computed, we need to prepare the necessary ingredients for iteration $t+1$ by finding $\boldsymbol{K}_{\mathfrak{C},t+1}$, $\boldsymbol{V}_{\mathfrak{C},t+1}$, $\boldsymbol{H}_{t+1}$, and $\boldsymbol{z}_{t+1}$. The first two are simple since the sparse policy will return $\boldsymbol{K}_{\mathfrak{C},t+1}$, $\boldsymbol{V}_{\mathfrak{C},t+1}$, and $\mathcal{D}_{t+1}$. Before freeing $\mathcal{D}_{t+1}$ from memory, we embed its information into $\boldsymbol{H}_{t+1}$ and $\boldsymbol{z}_{t+1}$:

$$\boldsymbol{H}_{t+1} = \boldsymbol{H}_t + \sum_{(\boldsymbol{k},\boldsymbol{v}) \in \mathcal{D}_{t+1}} \psi(\boldsymbol{k})^\top \boldsymbol{v}, \tag{4}$$

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \sum_{(\boldsymbol{k},\boldsymbol{v}) \in \mathcal{D}_{t+1}} \psi(\boldsymbol{k}). \tag{5}$$

After this, $\mathcal{D}_{t+1}$ can be deleted, and we are prepared for the following generation step. Intuitively, $\boldsymbol{H}_{t+1}$ and $\boldsymbol{z}_{t+1}$ are updated recursively by keys and values that have been newly pruned at each decoding step. As such, they are constant size repositories of information from all deleted KV pairs which becomes clear when we expand the recursion:

$$\boldsymbol{H}_{t+1} = \sum_{(\boldsymbol{k},\boldsymbol{v}) \in \bigcup_{i=1}^{t+1} \mathcal{D}_i} \psi(\boldsymbol{k})^\top \boldsymbol{v}, \tag{6}$$

and similarly for $\boldsymbol{z}_{t+1}$.

3.2. Implementation Details

Inexpensive Training. With our inference-time protocol outlined, we now describe how we can cheaply train our kernel functions ϕ and ψ. Because training end-to-end is time

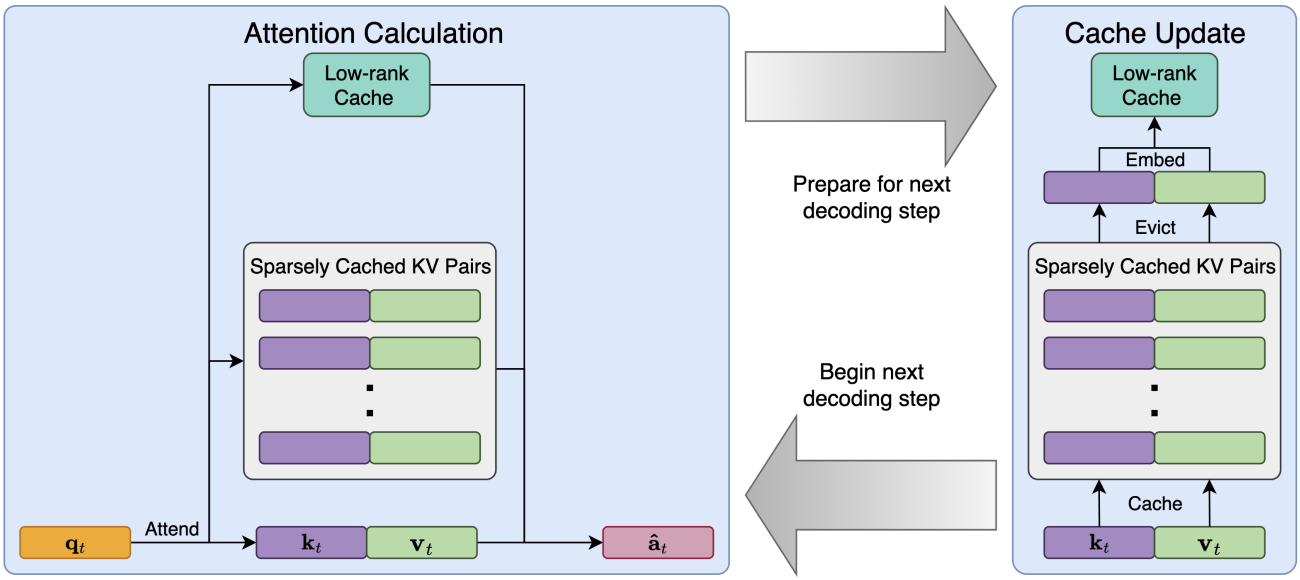

Figure 4. LESS algorithm during inference. At each decoding step, attention is calculated as in (3). To prepare for the next decoding step, the cache is updated by placing the most recent KV pair into the sparse policy cache, and if it has exceeded capacity, a KV pair will be evicted and integrated into the low-rank cache $\boldsymbol{H}_t$ before being deleted.

Algorithm 1 Generation Step with LESS

Input: $\mathfrak{C}, \boldsymbol{q}_t, \boldsymbol{k}_t, \boldsymbol{v}_t$
Load $\boldsymbol{K}_{\mathfrak{C},t}, \boldsymbol{V}_{\mathfrak{C},t}, \boldsymbol{H}_t, \boldsymbol{z}_t$ from memory.
$\boldsymbol{K}'_{\mathfrak{C},t} \leftarrow \text{concatenate}(\boldsymbol{K}_{\mathfrak{C},t}, \boldsymbol{k}_t)$
$\boldsymbol{V}'_{\mathfrak{C},t} \leftarrow \text{concatenate}(\boldsymbol{V}_{\mathfrak{C},t}, \boldsymbol{v}_t)$
Obtain $\hat{\boldsymbol{a}}_t$ via (3).
Obtain $\boldsymbol{K}_{\mathfrak{C},t+1}, \boldsymbol{V}_{\mathfrak{C},t+1}, \mathcal{D}_{t+1}$ from sparse KV cache algorithm $\mathfrak{C}$.
Update $\boldsymbol{H}_{t+1}$ via (4).
Update $\boldsymbol{z}_{t+1}$ via (5).
Save $\boldsymbol{K}_{\mathfrak{C},t+1}, \boldsymbol{V}_{\mathfrak{C},t+1}, \boldsymbol{H}_{t+1}, \boldsymbol{z}_{t+1}$.
Delete $\mathcal{D}_{t+1}$.
Return: $\hat{\boldsymbol{a}}_t$

consuming and resource intensive, we choose to train ϕ and ψ at each layer independent of all other layers which already surprisingly gives strong results. The training objective is to minimize the ℓ_2 distance to the output projection of the original attention layer using that layer's inputs. All weights except for those in ϕ and ψ are frozen. As a result, the only computational requirements are the abilities to backpropagate through a single attention layer and run inference on the full model to collect a dataset of attention layer inputs and outputs, which for all models we experiment with, *can be done on a single GPU*. With more devices, training each layer can be parallelized. While inference follows recursive updates of $\boldsymbol{H}_t$ and $\boldsymbol{z}_t$, this does not impede parallelism along the sequence axis because we can just construct the full attention matrix where entries not computed by sparsely cached KV pairs, as determined by whichever sparse policy we train on, will be found by the kernel functions.

All training runs used identical hyperparameters for simplicity. LESS was trained using Adam (Kingma & Ba, 2014) for 40 epochs with an initial learning rate of 0.001 which halved every 10 epochs. We fixed the hidden layer dimension $R' = 512$, used a dropout rate of 0.3 within the kernels, and let all nonlinear functions σ_ϕ and σ_ψ to be GELUs. None of the original model's weights are updated. First, we sampled 512 sequences for Llama 2 models (Touvron et al., 2023) and 1024 sequences for Falcon (Almazrouei et al., 2023) from the C4 training set (Raffel et al., 2019). Since Falcon's context length is half of Llama 2's, the training sets have the same number of tokens. Next, queries, keys, and values at each layer would be collected as each sample propagated through the models. These collected features (fed in batches of 2) would be used to train the kernels at each layer independently using some sparse policy at some sparsity level. For multi-query attention (Shazeer, 2019), we extend H_2O to aggregate attention scores across all query attention heads to determine KV pairs to evict.

We find that the kernel initialization is critical. As we will show in our experiments (Section 4), the sparse policies already have decent performance which we want to use as a starting point. As such, we add learnable scalars between layers in ψ which are initially set to 10^{-4}, so the influence of LESS during the first few gradient steps is small. In this way, the sparse policy acts as a warm start, and we can immediately reduce the sparse policy's residual.

Efficient Generation. We also develop an implementation that enhances throughput and reduces the latency of LLM generation of LESS. For the sparse cache, we adapt the implementation from Zhang et al. (2023) to support any KV cache eviction algorithm efficiently. To avoid data movement in memory, we directly replace the evicted KV pair with the newly-added one. As our kernels are small MLPs with GELUs, we implement a fused linear kernel that absorbs the activation into the layer before to avoid writing the intermediate results to DRAM for the low-rank cache.

4. Experiments

Here, we demonstrate the impressive performance of LESS across multiple datasets, models (Llama 2 and Falcon), sparse policies (Zhang et al., 2023; Han et al., 2023; Xiao et al., 2023; Oren et al., 2024), and sparsity levels, despite allocating only approximately 4 tokens of storage to the low-rank state. In Section 4.1, LESS achieves the closest performance to the full cache in language modeling and classification tasks. For example, evaluated with 2% H_2O in Llama 2 7B, LESS reduces the word perplexities on WikiText and PG-19 by over 20% from H_2O alone, relative to the full cache performance. Section 4.2 shows similar gains in summarization. For example, LESS reduces Rouge-1 degredation by 10% H_2O in Falcon 7B on CNN/DailyMail by 41.4%. In Section 4.3, we note the lower latency ($1.1-1.3\times$ reduction) and higher throughput of LESS ($1.7\times$ higher) compared to full caching. Finally, in Section 4.4, we discuss different characteristics of LESS, namely the recovery of true attention probabilities, kernel size scaling, and capabilities for long sequences.

We explore three sparse policies: H_2O (Zhang et al., 2023), Λ-masking from the infinite generation literature (Han et al., 2023; Xiao et al., 2023), and TOVA (Oren et al., 2024). When using H_2O, the sparse KV cache is equally split between the heavy hitter tokens and the recent tokens (e.g. 5% H_2O cache consists of 2.5% heavy hitters and 2.5% recent tokens). For Λ-masking, the cache consists of the first 4 and most recent tokens. The percentages represent how much of the model's max context length is cached, so regardless of input length, the cache size remains the same for fairness. Since the sparsity level can translate to different numbers of tokens among models based on the max input lengths, we include Table 1 as a quick reference for the models we evaluate on, Llama 2 and Falcon. The token count is rounded down to the nearest even number to make sure H_2O can have an even split.

For our experiments, we set the kernel size $R = 8$, unless otherwise stated. Thus, while minuscule, the size of $\boldsymbol{H}$ is nonzero, equivalent to caching 4 extra tokens. We ignore the influence of $\boldsymbol{z}$ since it only has R entries. As such, when evaluating on a task at $\alpha\%$ sparsity, we compare LESS with the sparse policy $\mathfrak{C}$ at $\alpha\%$ sparsity and at $\alpha\%$ sparsity plus additional tokens to match the extra space taken by $\boldsymbol{H}$ (e.g. 4 tokens in experiments where $R = 8$), which we denote as Baseline and Baseline+, respectively. Both are inherently sparse-only policies. A visual representation of the different baselines can be found in Figure 5. Note that the sparsity level and policy $\mathfrak{C}$ will vary throughout our experiments depending on the context. The purpose of evaluating Baseline is to compare the performance gain from extra tokens and the low-rank state $\boldsymbol{H}$. Additionally, we evaluate the full KV cache to observe how far we are from the unconstrained potential of the original model. For our method, we denote it as LESS ($\beta\%$) where β is the percent cache sparsity LESS was trained on with some sparse policy depending on the context.

Table 1. Token counts at different sparsity levels.

MODEL	MAX LENGTH	# TOKENS AT 2%/5%/10%
LLAMA 2	4096	80 / 204 / 408
FALCON	2048	40 / 102 / 204

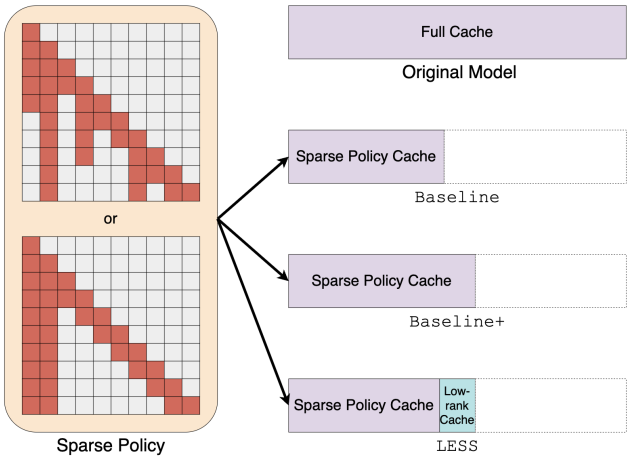

Figure 5. Experimental setup. First, a sparse policy is chosen as the underlying policy behind all methods. Then, we compare performance among the full cache model, Baseline, Baseline+, and LESS. Baseline+ and LESS use the same amount of storage which is slightly larger than the requirements of Baseline.

4.1. Language Modeling & Classification

We start with validating our method trained at different sparsity levels on some language modeling and classification tasks at different sparsity levels using Language Modeling Evaluation Harness (Gao et al., 2023). For these tasks, we use the same setup as in training by masking out query-key interactions depending on the sparse policy and having LESS capture the masked probabilities. In addition, we purposefully mismatch training and testing sparsity levels to

Table 2. Llama 2 7B WikiText and PG-19 word perplexities with H_2O as the primary underlying sparse policy. Numeric column names indicate the sparsity levels during test time. Lower is better.

H_2O METHOD	2% H_2O	5% H_2O	10% H_2O
WikiText			
FULL CACHE	8.791	8.791	8.791
BASELINE	13.333	9.863	9.295
BASELINE+	12.718	9.842	9.288
H_2O+PERFORMER	13.332	9.863	9.296
LESS (2%)	**10.745**	9.658	9.261
LESS (5%)	11.321	**9.657**	9.239
LESS (10%)	14.577	9.693	**9.230**
PG-19			
FULL CACHE	23.787	23.787	23.787
BASELINE	37.013	27.939	25.451
BASELINE+	35.832	27.829	25.429
H_2O+PERFORMER	36.996	27.938	25.451
LESS (2%)	**32.157**	27.887	26.322
LESS (5%)	33.195	**27.089**	25.979
LESS (10%)	41.204	27.201	**25.134**

uncover insight on the transferability between test sparsity levels. To illustrate why a learned kernel is necessary, we also evaluate H_2O with Performer kernels (Choromanski et al., 2020) based on random Fourier features (Rahimi & Recht, 2007), which we denote as H_2O+Performer.

Table 2 shows Llama 2 7B performance on WikiText (Merity et al., 2016) and PG-19 (Rae et al., 2019; Gao et al., 2020) using H_2O. Looking at the scenarios where training sparsity is equal to the test sparsity, our method is able to achieve much lower word perplexities than the baselines. Notably, LESS beats Baseline by a wider margin than Baseline+ and H_2O+Performer, indicating that LESS uses the space of 4 extra tokens most effectively. The lackluster performance of H_2O+Performer suggests that learned kernels are needed to make a noticeable improvement. Moreover, LESS trained at one sparsity level can often generalize reasonably to higher sparsity levels especially on WikiText, even sometimes matching the performance of ones trained at the test sparsity level. The reverse is less effective but can still be better than the baselines. However, all methods are still quite far from the full cache performance.

Evaluation results (Clark et al., 2019; Cui et al., 2020) with Λ-masking in Table 3 show LESS's benefit to a different sparse policy (though less performative than H_2O). Similar to the case with H_2O, LESS closes the gap from full caching but cannot match the performance completely. TOVA also observes similar benefits in language modeling, shown Appendix B. While LESS is efficacious for language modeling and classification, we also want to assess its utility for generation where the KV cache storage becomes a critical bottleneck.

Table 3. Llama 2 7B performance on WikiText (word perplexity), MuTual (16-shot R@1), and BoolQ (10-shot accuracy) with 5% Λ-masking as the primary underlying sparse policy.

Λ METHOD	WIKITEXT (↓)	MUTUAL	BOOLQ
FULL CACHE	8.79	55.08	80.40
BASELINE	10.66	53.50	77.28
BASELINE+	10.64	53.27	77.46
LESS (5%)	**10.12**	**54.51**	**78.65**

Table 4. Llama 2 13B and Falcon 7B generation quality comparison on CNN/DailyMail and XSum with 408 sparse tokens (10% and 20% of the context lengths of Llama 2 and Falcon, respectively) with H_2O as the primary underlying sparse policy. Llama 2 13B is given 5 shots while Falcon 7B is given 3 shots due to its shorter context length. Values are in the format [Rouge-1/2/L].

H_2O METHOD	CNN/DAILYMAIL	XSUM
Llama 2 13B		
FULL CACHE	27.55/9.96/25.80	33.14/13.05/27.33
BASELINE	23.57/7.35/22.04	33.09/**13.09/27.44**
BASELINE+	23.40/7.31/21.88	33.09/13.06/27.41
LESS (2%)	**25.27/7.76/23.64**	**33.40**/12.98/27.41
LESS (5%)	24.45/7.70/22.87	33.15/13.02/27.39
Falcon 7B		
FULL CACHE	25.92/8.52/24.15	27.17/8.83/22.67
BASELINE	21.26/5.95/19.73	24.50/7.65/20.50
BASELINE+	21.31/6.16/19.75	24.55/7.66/20.56
LESS (5%)	23.00/6.28/21.28	24.94/8.17/20.94
LESS (10%)	**23.22/6.37/21.53**	**25.21/8.28/21.17**

4.2. Summarization

Now, we move on to generation, specifically summarization, to test the ability to generate longer and coherent sequences by synthesizing numerous tokens. Unlike in our language modeling evaluations, the model will have access to all tokens during the prompting phase with the sparse policy and LESS only kicking in during the subsequent generation steps. Consequently, generation sparse policies are fundamentally different from the language modeling masks LESS is trained on, yet despite this, we show that our method maintains its superior performance.

In Tables 4 and 5, we see LESS achieves better ROUGE (Lin, 2004) scores than purely H_2O on the CNN/DailyMail (Hermann et al., 2015; See et al., 2017), MultiNews (Fabbri et al., 2019), and XSum (Narayan et al., 2018) datasets. Even at exceptionally low sparsity levels, H_2O can capture a significant amount of the full cache's performance. This is even more surprising for Falcon models which already cache many times fewer tokens than Llama 2 due to the multi-query attention mechanism. Despite this, we observe LESS surpasses the already strong performance of H_2O across the board where H_2O underperforms compared to the full cache. Like in language modeling, we again see that

Table 5. Llama 2 7B performance on MultiNews (1-shot), CNN/DailyNews (5 shot), and XSum (5-shot) with 5% and 10% H_2O as the primary underlying test sparse policies. Values are in the format [Rouge-1]/[Rouge-2]/[Rouge-L].

H_2O METHOD	5% H_2O	10% H_2O
MultiNews		
FULL CACHE	23.79/6.87/21.35	23.79/6.87/21.35
BASELINE	13.38/3.25/12.25	19.44/4.97/17.73
BASELINE+	13.58/3.32/12.41	19.44/4.96/17.72
LESS (2%)	15.31/3.73/14.03	20.32/5.24/18.51
LESS (5%)	**15.42/3.80/14.14**	**20.55/5.29/18.70**
CNN/DailyMail		
FULL CACHE	26.25/9.34/24.40	26.25/9.34/24.40
BASELINE	18.18/4.92/16.89	20.04/6.09/18.66
BASELINE+	18.24/4.91/16.85	20.15/6.21/18.73
LESS (2%)	18.71/5.40/17.34	20.76/6.40/19.32
LESS (5%)	**19.21/5.44/17.80**	**22.29/6.85/20.69**
XSum		
FULL CACHE	30.65/11.11/25.40	30.65/11.11/25.40
BASELINE	29.03/10.77/24.28	30.68/**11.54**/25.58
BASELINE+	28.94/10.78/24.15	30.64/11.49/**25.59**
LESS (2%)	**30.72/11.53/25.57**	30.34/10.98/25.31
LESS (5%)	30.03/11.19/25.03	**30.82**/11.17/25.56

the improvement from Baseline to Baseline+ pales in comparison to the improvement induced by LESS, sometimes even matching the full cache performance as in XSum. Again, we also see the transferability of LESS to other sparsity levels. See Appendix D for example generation outputs.

4.3. Latency and Throughput

Following Sheng et al. (2023), we benchmark the generation throughput and latency of LESS on an NVIDIA A100 80G GPU using FP16 precision. We focus on the Llama 2 7B and 13B models, with all speedup results tested end-to-end with both prompting and generation phases. To measure its performance when generating long sequences or inputting large batch sizes, we use synthetic datasets where all prompts are padded to the same length and batched together. The same number of tokens are generated for each prompt. We test different combinations of prompt and generation lengths.

Table 6 shows results with sequence lengths from 4K to 10K. With the same batch size, LESS reduces the latency by $1.1-1.3\times$ compared to the full cache, though slightly slower than H_2O. Moreover, LESS saves memory to allow larger batch sizes with a $1.7\times$ improvement on generation throughput for Llama 2 7B, closely matching the performance of H_2O.

4.4. Empirical Analysis and Ablations

Now that we have shown that LESS is simple and effective, we share some interesting characteristics of our method.

Reconstructing Attention Probabilities. Sparse KV cache policies can delete tokens that may be needed later on. A way to see this is to construct the sparse attention matrix and compare with the full one. In Figure 1, H_2O zeroes out many relatively high attention probabilities with a bias towards keeping early tokens. More examples are in Appendix A. Visually, LESS provides a sketch of the deleted tokens which appears to reasonably reconstruct trends.

Numerically, we measure the similarity of each row in the attention matrix with corresponding rows produced by H_2O and LESS with the Hellinger distance, which for two discrete probability vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, is defined as

$$\mathcal{H}(\boldsymbol{p}, \boldsymbol{q}) := \|\sqrt{\boldsymbol{p}} - \sqrt{\boldsymbol{q}}\|_2/\sqrt{2} \tag{7}$$

where the square root is elementwise. The value of $\mathcal{H}(\boldsymbol{p}, \boldsymbol{q})$ ranges from 0 to 1, where a lower value indicates greater similarity. In Figure 6, we see that our method more accurately replicates the original attention probability distributions as measured by the Hellinger distance. We choose to aggregate each layer separately since the attention distribution patterns tend to vary dramatically throughout the model.

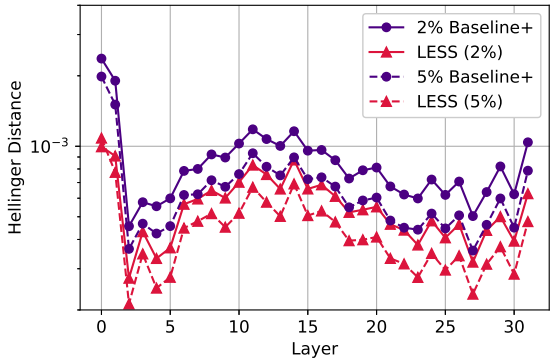

Figure 6. Layer-wise Llama 2 7B mean Hellinger distance from original attention probabilities, aggregated across WikiText evaluation samples. The underlying sparse policy is H_2O. Here, LESS is evaluated based on their training sparsity percentages.

Larger Kernels. In our experiments, we fixed $R = 8$, and as we show in Figure 7, the performance generally increases as R increases. However, at a certain point, the marginal benefit derived from increasing R is less than shifting more of the KV cache to the sparse policy, suggesting that a small low-rank cache is enough.

Providing Hope for Long Sequences. Model performance appears to be highly correlated with the input sequence length regardless of the caching method. As shown in Figure 8, even the full cache model performance drops dramatically and immediately as the prompt length increases. Baseline+ and LESS (1% H_2O) appear to perform similarly for shorter sequences but diverge for longer sequences where we see LESS is more performative.

Table 6. Llama 2 7B and 13B's generation throughput and latency on an A100 GPU. In the sequence length column, we use "5000 + 5000" to denote a prompt length of 5000 and a generation length of 5000. "OOM" stands for out-of-memory.

SEQ. LENGTH	MODEL SIZE	BATCH SIZE	METRIC	FULL CACHE	BASELINE+	LESS (5%)
5000+5000	13B	4	LATENCY (S)	257.3	185.2	204.7
2048+2048	7B	24	LATENCY (S)	116.7	78.3	95.1
2048+2048	7B	24	THROUGHPUT (TOKENS/S)	421.2	627.7	516.9
2048+2048	7B	64	THROUGHPUT (TOKENS/S)	OOM	819.2	699.2

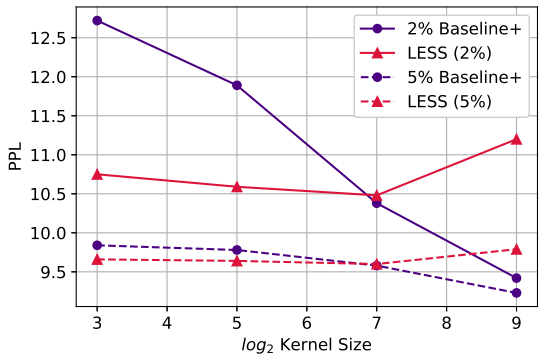

Figure 7. Llama 2 7B WikiText word perplexity (lower is better) as the kernel size quadruples, compared against Baseline+ which occupies the same space. The sparse KV cache policy is H_2O.

This follows our intuition since for sparse cache policies, a smaller fraction of KV pairs is saved as the sequence length increases, so more information is omitted. This is where a low-rank state can help to recover some of this information.

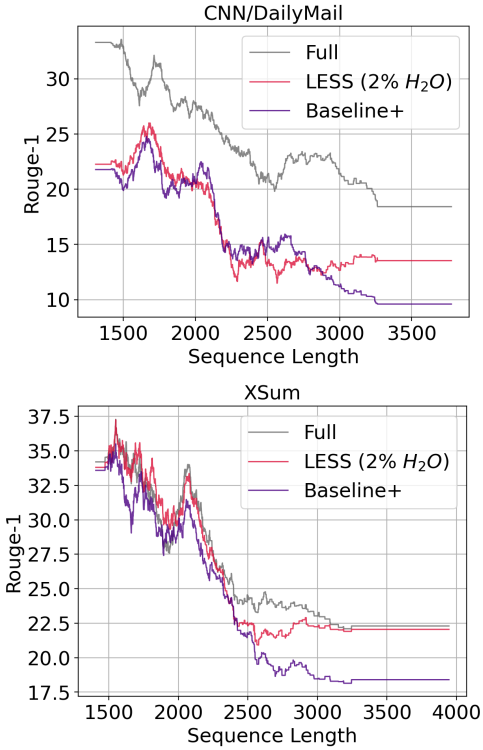

Figure 8. Relationship between Rouge-1 score and prompt length for Llama 2 7B with different cache methods on CNN/DailyMail (top) and XSum (bottom). The test sparse KV cache policy is 5% H_2O for all models. As these results can be fairly noisy, the lines are k-nearest regression lines where k is 10% of the dataset size.

5. Conclusion and Future Work

To tackle the KV cache bottleneck, we introduce LESS which has demonstrated itself to be an effective way to boost eviction-based KV cache algorithms. Motivated by the necessity to maintain information that would have been discarded, the constant-sized LESS recovers a significant portion of the performance lost due to maintaining a small cache across a variety of scenarios and intensities, despite being cheap to train and deploy. There are many exciting avenues of work that can enhance LESS or build upon it, such as improving kernel design and investigating the residual of LESS. Such directions will further push the performance of a condensed KV cache to that of a complete cache, allowing LLMs to accomplish the same tasks with less.

Acknowledgements

The work of H. Dong is supported in part by the Liang Ji-Dian Graduate Fellowship, the Michel and Kathy Doreau Graduate Fellowship in Electrical and Computer Engineering, and the Wei Shen and Xuehong Zhang Presidential Fellowship at Carnegie Mellon University. Z. Wang is in part supported by a Google Research Scholar Award and the NSF AI Institute for Foundations of Machine Learning (IFML). The work of Y. Chi is supported in part by the grants NSF DMS-2134080 and ONR N00014-19-1-2404. B. Chen is supported in part by MOFFETT AI gift funding.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. As our work improves LLM efficiency, there is the potential risk that it may be easier to use these models for unethical or unlawful purposes, such as spreading misinformation. While this is a notable concern, the benefits of making LLMs more accessible are plenty,

such as expediting research progress in multiple domains and improving technological equity. Thus, we think the net benefit is overwhelmingly positive and warrants publication.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G. H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C. A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P., Ros, A. C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D. R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., and Wu, Y. Palm 2 technical report, 2023.

BBC. Fracking still opposed in wales, ministers tell councils. *The British Broadcasting Corporation*, 2015. URL https://www.bbc.com/news/uk-wales-politics-33935224.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Brumfield, B. Death toll rises quickly as conflict rages in yemen. *The Cable News Network*, 2015. URL https://www.cnn.com/2015/04/06/middleeast/yemen-conflict-houthis-saudi-arabia/index.html.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chen, B., Dao, T., Winsor, E., Song, Z., Rudra, A., and Ré, C. Scatterbrain: Unifying sparse and low-rank attention. *Advances in Neural Information Processing Systems*, 34: 17413–17426, 2021.

Chen, Y., Tao, Q., Tonin, F., and Suykens, J. A. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. *arXiv preprint arXiv:2305.19798*, 2023.

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.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Cui, L., Wu, Y., Liu, S., Zhang, Y., and Zhou, M. Mutual: A dataset for multi-turn dialogue reasoning. *arXiv preprint arXiv:2004.04494*, 2020.

Dao, T., Fu, D. Y., Saab, K. K., Thomas, A. W., Rudra, A., and Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li,

H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

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. *arXiv preprint arXiv:2310.01801*, 2023.

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

Han, C., Wang, Q., Xiong, W., Chen, Y., Ji, H., and Wang, S. Lm-infinite: Simple on-the-fly length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. In *NIPS*, pp. 1693–1701, 2015.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Kasai, J., Peng, H., Zhang, Y., Yogatama, D., Ilharco, G., Pappas, N., Mao, Y., Chen, W., and Smith, N. A. Finetuning pretrained transformers into rnns. *arXiv preprint arXiv:2103.13076*, 2021.

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.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, Y., Yu, Y., Zhang, Q., Liang, C., He, P., Chen, W., and Zhao, T. Losparse: Structured compression of large language models based on low-rank and sparse approximation. *arXiv preprint arXiv:2306.11222*, 2023.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Lin, T., Wang, Y., Liu, X., and Qiu, X. A survey of transformers. *AI Open*, 2022.

Liu, Y., Li, H., Du, K., Yao, J., Cheng, Y., Huang, Y., Lu, S., Maire, M., Hoffmann, H., Holtzman, A., et al. Cachegen: Fast context loading for language model applications. *arXiv preprint arXiv:2310.07240*, 2023a.

Liu, Z., Desai, A., Liao, F., Wang, W., Xie, V., Xu, Z., Kyrillidis, A., and Shrivastava, A. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *arXiv preprint arXiv:2305.17118*, 2023b.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Nawrot, P., Łańcucki, A., Chochowski, M., Tarjan, D., and Ponti, E. M. Dynamic memory compression: Retrofitting llms for accelerated inference. *arXiv preprint arXiv:2403.09636*, 2024.

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., et al. Transformers in healthcare: A survey. *arXiv preprint arXiv:2307.00067*, 2023.

Nikdan, M., Tabesh, S., and Alistarh, D. Rosa: Accurate parameter-efficient fine-tuning via robust adaptation. *arXiv preprint arXiv:2401.04679*, 2024.

OpenAI. Gpt-4 technical report, 2023.

Oren, M., Hassid, M., Adi, Y., and Schwartz, R. Transformers are multi-state rnns. *arXiv preprint arXiv:2401.06104*, 2024.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. Image transformer. In *International conference on machine learning*, pp. 4055–4064. PMLR, 2018.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Cao, H., Cheng, X., Chung, M., Grella, M., GV, K. K., et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., and Kong, L. Random feature attention. *arXiv preprint arXiv:2103.02143*, 2021.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S., and Dean, J. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5, 2023.

Rae, J. W., Potapenko, A., Jayakumar, S. M., Hillier, C., and Lillicrap, T. P. Compressive transformers for long-range sequence modelling. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1911.05507.

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.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Renggli, C., Pinto, A. S., Houlsby, N., Mustafa, B., Puigcerver, J., and Riquelme, C. Learning to merge tokens in vision transformers. *arXiv preprint arXiv:2202.12015*, 2022.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

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

Sheng, Y., Zheng, L., Yuan, B., Li, Z., Ryabinin, M., Fu, D. Y., Xie, Z., Chen, B., Barrett, C. W., Gonzalez, J., Liang, P., Ré, C., Stoica, I., and Zhang, C. High-throughput generative inference of large language models with a single gpu. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:257495837.

Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., and Wei, F. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey, 2022.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., and Salakhutdinov, R. Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. Opt: Open pre-trained transformer language models, 2022.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. H _2 o: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv preprint arXiv:2306.14048*, 2023.

A. Attention Matrix Visualizations

This section provides some qualitative results on attention matrix approximations by sparse policies and LESS. While low-rank caches LESS cannot perfectly recover all the missing information, it visually is able to reconstruct a patterns that are completely ignored by sparse policies. We can also see the idiosyncrasies of the sparse policies and LESS, such as H_2O's bias towards keeping early tokens, as shown in Figures 9 and 10, and Λ-masking's tendency to miss influential tokens which are captured by LESS, as show in Figure 11.

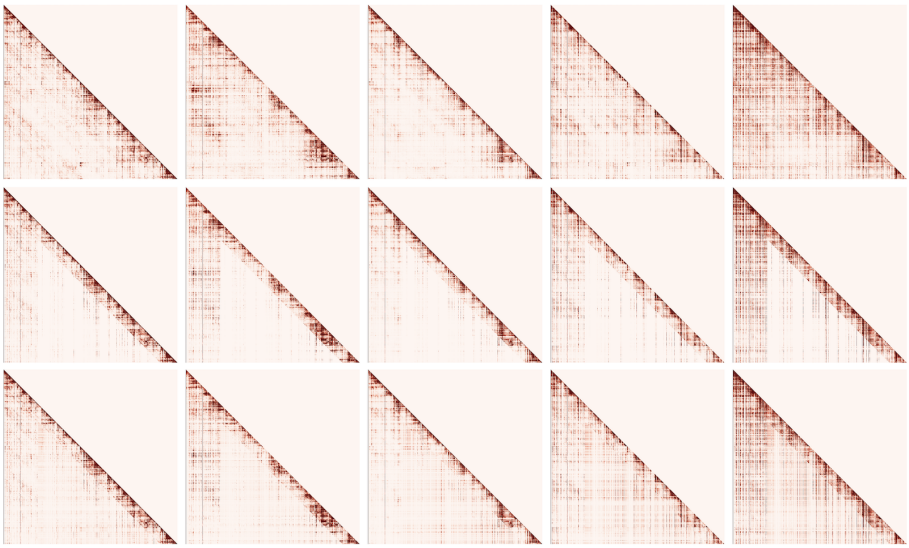

Figure 9. Example attention probability matrices from passing a single input into Falcon 7B. From top to bottom, the rows consist of attention maps from the original model, 10% H_2O (204 tokens), and LESS (10% H_2O). Darker pixels indicate larger probability weights. Only the first 1024 tokens are displayed.

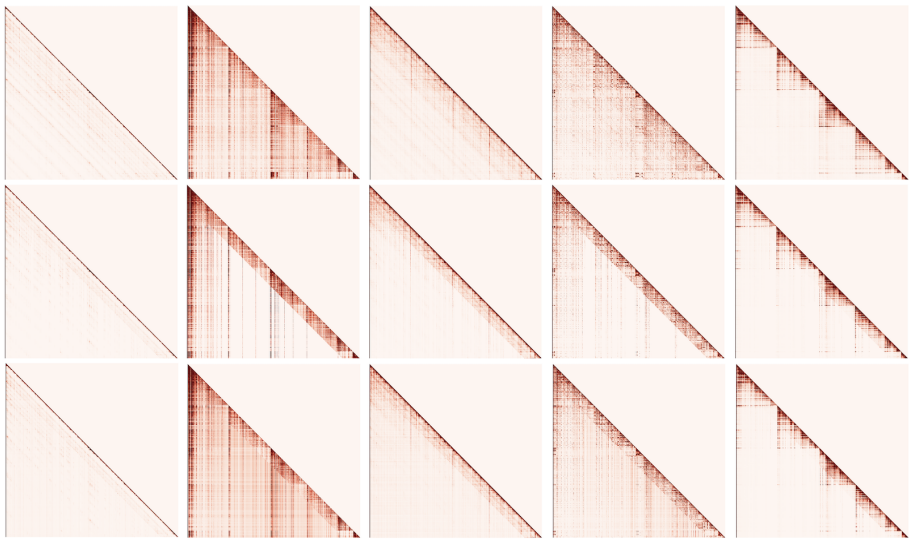

Figure 10. Example attention probability matrices from passing a single input into Llama 2 7B. From top to bottom, the rows consist of attention maps from the original model, 5% H_2O (204 tokens), and LESS (5% H_2O). Darker pixels indicate larger probability weights. Only the first 1024 tokens are displayed.

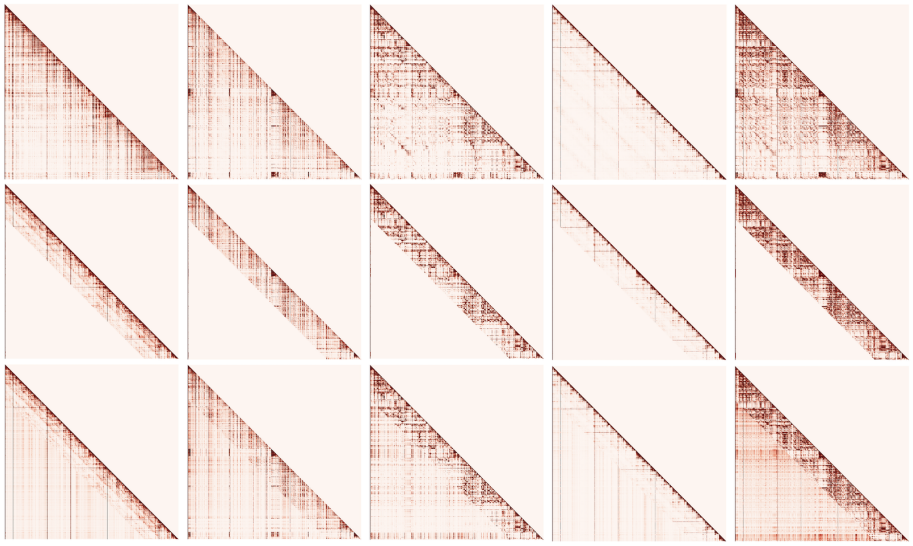

Figure 11. Example attention probability matrices from passing a single input into Llama 2 7B. From top to bottom, the rows consist of attention maps from the original model, 5% Λ-masking (204 tokens), and LESS (5% Λ). Darker pixels indicate larger probability weights. Only the first 1024 tokens are displayed.

B. Language Modeling Comparison

In Table 7, we compare language modeling using LESS with H_2O, Λ-masking, and TOVA. We see that LESS makes a greater improvement on word perplexity than simply caching the equivalent number of extra tokens, though none of the methods can fully recover the original model's performance.

Table 7. Llama 2 7B performance (word PPL) on WikiText and PG-19 at 5% token sparsity using different sparse policies. Lower is better.

METHOD	H_2O	Λ	TOVA
WikiText			
FULL CACHE	8.79	8.79	8.79
BASELINE	9.86	10.66	9.97
BASELINE+	9.84	10.64	9.95
LESS (5%)	**9.66**	**10.12**	**9.72**
PG-19			
FULL CACHE	23.79	23.79	23.79
BASELINE	27.94	22K	27.88
BASELINE+	27.83	22K	27.79
LESS (5%)	**27.09**	**3.9K**	**27.34**

C. Latency Breakdown

Here, we show that maintaining and using the low-rank state in LESS has little overhead by determining the latencies of different operations in LESS. Shown in Table 8, the total latency of LESS is much smaller than the full cache but slightly higher than Baseline+ due to the additional operations associated with the low-rank state. This overhead represents about 15% of the decoding time, meaning it does not significantly impact the overall efficiency.

Table 8. Latency breakdown for Llama 2 7B on an A100 GPU, with a batch size of 64, prompt length of 512, and generation length of 512 with 5% H_2O as the underlying sparse policy. "Eviction" refers to H_2O's KV eviction algorithm, "Kernels" refers to (1) and (2), "Attention Synthesis" refers to (3), and "Low-rank Cache Update" refers to (4) and (5).

	DECODING	EVICTION	KERNELS	ATTENTION SYNTHESIS	LOW-RANK CACHE UPDATE	TOTAL
FULL CACHE	50.71	N/A	N/A	N/A	N/A	50.71
BASELINE+	23.53	4.52	N/A	N/A	N/A	28.05
LESS (5%)	23.61	4.39	0.87	1.35	1.51	32.96

D. Generation Outputs

We include a couple examples of generation outputs in Figure 12 and Figure 13. In both cases, the full cache, LESS, and Baseline+ models attempt to summarize news articles. We see in Figure 12 that LESS is able to produce the same concise summary as the full cache while Baseline+ produces rambling text. In Figure 13, we observe that LESS completely changes the meaning of the summary from H_2O alone–Baseline+ is factually incorrect based on the article.

Prompt

Article: (CNN)Residents of central Sanaa, the Yemeni capital, have learned the hard way that key strategic bombing targets are located in their neighborhoods: Detonating ordnance has been shattering their windows and doors. … Hadi's government had cooperated with the United States to fight AQAP, but with the Houthi takeover, that arrangement has evaporated, and the terror group operates generally unchecked.

Summarize the above article in 3 sentences.

Llama 2 7B: Full Cache

Saudi-led coalition bombs Yemen's Defense Ministry Central Command in capital Sanaa . Fighting has killed hundreds of people in less than two weeks . Yemen's deposed President Abdu Rabu Mansour Hadi heaped scorn on top of the airstrikes .

Llama 2 7B: LESS (2% H_2O)

Saudi-led coalition bombs Yemen's Defense Ministry Central Command in capital Sanaa . Fighting has killed hundreds of people in less than two weeks . Yemen's deposed President Abdu Rabu Mansour Hadi heaped scorn on top of the airstrikes .

Llama 2 7B: Baseline+

The Houthis have taken over Yemen's capital of Sanaa . The Saudi-led coalition has been bombing the Houthis and their allies . The fighting has killed hundreds of people . The Houthis are Shiite Muslims and the Saudis are Sunni Muslims . The Houthis have been fighting the government of Yemen's former President, Abed Rabbo Mansour Hadi, who is now in Saudi Arabia . The Saudi-led coalition is fighting to restore Hadi to power . The Houthis have been fighting the government of Yemen'*<cut off before reaching the stop token>*

Figure 12. Example 5-shot (not shown) CNN/DailyMail summary generation results produced by variations of Llama 2 7B with an underlying sparse policy of 2% H_2O. For brevity, only the start and end of the article are shown with the middle omitted with an ellipsis. LESS produces the same concise summary as the full cache while Baseline+ produces rambling text, exceeding the 3 sentence requirement by the prompt. The original article is from Brumfield (2015).

Prompt

Article: The UK government has announced measures to speed up development.
Natural Resources Minister Carl Sargeant has written to councils reminding them of the temporary ban on fracking in Wales.
He said fracking technology was "unproven", saying Welsh ministers preferred to look at renewable energy.
Fracking is a process of using high pressure water to break up rocks deep underground to release gas and pipe it to the surface.
It is a major industry in the United States, but there are concerns about safety, the environment and underground water.
There have been protests against applications for test drilling for gas in both the Vale of Glamorgan and Wrexham, as well as sites in England.
Energy Secretary Amber Rudd has stressed the UK government's support for fracking, and frustration at delays in its development.
She has told councils in England to make decisions on applications within 16 weeks, to stop the process being "dragged out for months".
In his letter to Welsh councils, Mr Sargeant said the new guidance only applied to England and that the moratorium in Wales remained in place.
"The UK government's general support for oil and gas applications is contrary to the approach of the Welsh Government of promoting renewable and low carbon forms of energy through the planning system and other measures," he said.
"We still see renewable energy as a key element in ensuring that Wales achieves sustainable development for the benefit of future generations.
“Local planning authorities must ensure that planning applications for renewable energy projects are determined within statutory timescales.”
Plans to devolve control over fracking were confirmed in the Queen's Speech in June following the Conservatives' general election victory.

Summarize the above article in 1 sentences.

Falcon 7B: Full Cache

The UK government has announced measures to speed up development of fracking in Wales.

Falcon 7B: LESS (10% H_2O)

The Welsh government has announced that it will not allow fracking in Wales.

Falcon 7B: Baseline+

The Welsh government has announced measures to speed up the development of fracking.

Figure 13. Example 3-shot (not shown) XSum summary generation results produced by variations of Falcon 7B. Models were evaluated with 20% H_2O. The summary by `Baseline+` is factually incorrect based on the article, while `LESS` preserves the meaning better. The original article is from BBC (2015).