# OBCache: Optimal Brain KV Cache Pruning for Efficient Long-Context LLM Inference

Yuzhe Gu[1], Xiyu Liang[2], Jiaojiao Zhao[3], Enmao Diao[4]
[1]University of Pennsylvania [2]University of Electronic Science and Technology of China
[3]Duke Kunshan University [4]DreamSoul

ICML International Conference On Machine Learning

Paper Code

## Motivation

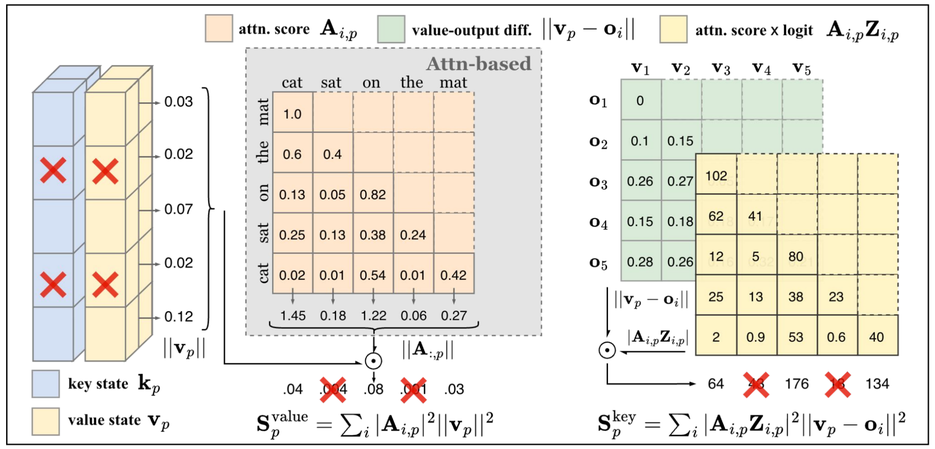
- **Efficiency Bottleneck**: KV cache memory in LLMs grows linearly with sequence length & batch size, creating a key bottleneck in long-context deployment.
- **Attention-Only Heuristics**: Existing KV cache eviction methods mainly estimate token saliency using cumulative attention weights, overlooking value states and the true eviction impact on model outputs.
- **Weak Theoretical Grounding**: Recent KV scoring mechanisms are often empirically driven, lacking a principled theoretical framework for characterizing how token eviction affects model outputs.

## Contributions

- **Structured Pruning View**: Building on Optimal Brain Damage (OBD) theory, we formalize KV cache eviction as a layer-wise structured pruning problem, providing a new theoretical framework for cache eviction.
- **Output-Aware Saliency**: We measure token saliency from historical attention output perturbation, deriving three second-order closed-form saliency scores that generalize existing attention-only scores.
- **Plug-and-play Score**: OBCache can be seamlessly integrated into any score-based KV cache eviction pipeline to improve their long-context accuracy, while incurring negligible additional overhead.

***OBCache is a Theoretically Grounded Scoring Framework for KV Cache Compression.***

## Optimal Brain Cache (OBCache)




Token Saliency via *Pruning-Induced Eviction Error*

$$\boldsymbol{S}_p := \mathcal{L}_{\boldsymbol{e}_p^\top[\widehat{\mathbf{V}}\ \widehat{\mathbf{K}}]=\mathbf{0}}(\widehat{\mathbf{V}},\widehat{\mathbf{K}}) = \left\| \text{softmax}\left(\frac{\mathbf{Q}_{w:s}\widehat{\mathbf{K}}^\top}{\sqrt{d}}\right)\widehat{\mathbf{V}}\Big|_{\boldsymbol{e}_p^\top[\widehat{\mathbf{V}}\ \widehat{\mathbf{K}}]=\mathbf{0}} - \text{softmax}\left(\frac{\mathbf{Q}_{w:s}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \right\|_F^2$$

Second-order Taylor Appx.

$$\boldsymbol{S}_p \overset{\text{second}}{=\!=}_{\text{order}} \underbrace{\frac{1}{2}\mathbf{v}_p^\top \mathbf{H}_{pp}^{vv}\mathbf{v}_p}_{①} + \underbrace{\frac{1}{2}\mathbf{k}_p^\top \mathbf{H}_{pp}^{kk}\mathbf{k}_p}_{②} + \underbrace{\mathbf{v}_p^\top \mathbf{H}_{pp}^{vk}\mathbf{k}_p}_{③}$$

OBCache (output-aware) scores

① $\boldsymbol{S}_p^{\text{value}} = \sum_i |\mathbf{A}_{i,p}|^2 \|\mathbf{v}_p\|^2$

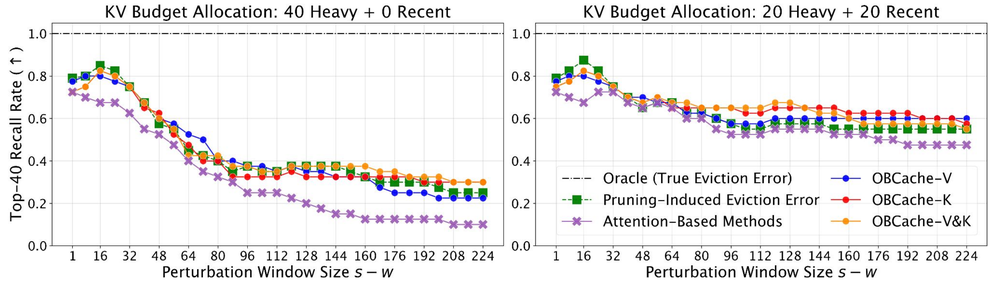
② $\boldsymbol{S}_p^{\text{key}} = \sum_i |\mathbf{A}_{i,p}\mathbf{Z}_{i,p}|^2 \|\mathbf{v}_p - \mathbf{o}_i\|^2$

③ $\boldsymbol{S}_p^{\text{joint}} = \boldsymbol{S}_p^{\text{value}} + \boldsymbol{S}_p^{\text{key}} + 2\sum_i |\mathbf{A}_{i,p}|^2 \mathbf{Z}_{i,p}(\|\mathbf{v}_p\|^2 - \mathbf{v}_p^\top \mathbf{o}_i)$

Attn-only scores

$$\boldsymbol{S}_p^{\text{attn}} = \sum_i |\mathbf{A}_{i,p}|$$

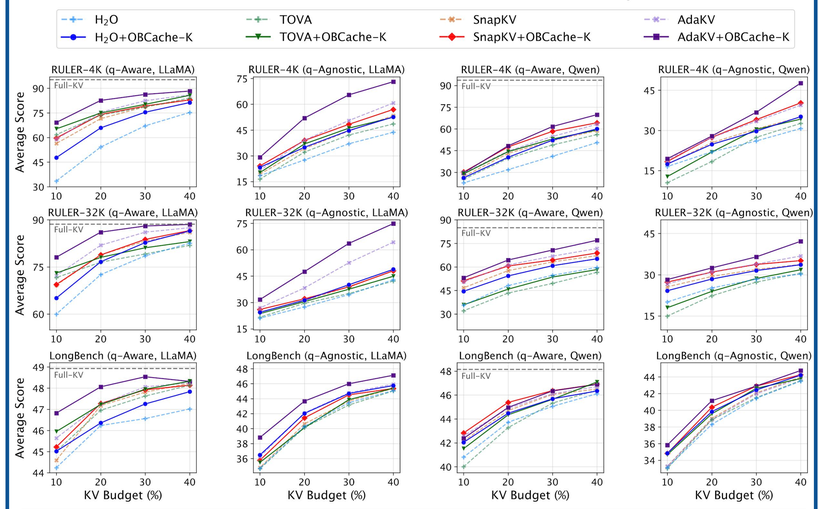
## Effectiveness Analysis




- True Eviction Error (unobservable) → Pruning-Induced Eviction Error (Proxy Objective)
- Exact Proxy (computationally infeasible) → OBCache scores ✅ Attention-only scores ❌

## Experimental Results

*Static Prefill Cache Eviction*




| KV Budget (%) | RULER-4K (Q-Aware) 10 | 20 | 30 | 40 | Avg. | RULER-4K (Q-Agnostic) 10 | 20 | 30 | 40 | Avg. | RULER-32K (Q-Aware) 10 | 20 | 30 | 40 | Avg. | RULER-32K (Q-Agnostic) 10 | 20 | 30 | 40 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full KV | 95.3 | | | | | | | | | | 88.6 | | | | | | | | | |
| $H_2O$ | 33.5 | 54.2 | 67.1 | 75.3 | 57.5 | 18.5 | 27.6 | 37.0 | 43.7 | 31.7 | 60.0 | 72.5 | 78.5 | 82.5 | 73.4 | 21.1 | 27.5 | 34.4 | 42.9 | 31.5 |
| + OBCache-V | 37.7 | 57.1 | 68.6 | 76.0 | 59.9 | 19.6 | 28.2 | 36.5 | 44.9 | 32.3 | 62.7 | 73.5 | 80.4 | 84.3 | 75.2 | 20.0 | 27.7 | 36.7 | 46.0 | 32.6 |
| + OBCache-K | 47.7 | 65.9 | 75.4 | 81.3 | 67.6 | 23.3 | 34.9 | 44.7 | 52.7 | 38.9 | 65.1 | 76.6 | 82.7 | 86.5 | 77.7 | 24.4 | 31.3 | 40.1 | 48.8 | 36.2 |
| + OBCache-V&K | 46.3 | 67.5 | 75.6 | 82.0 | **67.8** | 23.3 | 35.4 | 46.4 | 54.7 | **40.0** | 66.0 | 76.8 | 83.0 | 86.7 | **78.1** | 24.9 | 31.9 | 40.6 | 49.5 | **36.7** |
| TOVA | 61.8 | 73.6 | 78.6 | 84.0 | 74.5 | 16.4 | 32.2 | 42.1 | 48.6 | 34.8 | 71.6 | 76.5 | 79.1 | 81.9 | 77.3 | 21.6 | 29.5 | 35.1 | 42.2 | 32.1 |
| + OBCache-V | 64.9 | 74.8 | 80.0 | 85.4 | 76.3 | 18.8 | 34.6 | 44.3 | 50.7 | 37.1 | 72.3 | 76.9 | 79.9 | 83.2 | 78.1 | 22.3 | 30.9 | 36.8 | 44.0 | 33.5 |
| + OBCache-K | 65.3 | 75.0 | 80.2 | 85.7 | 76.5 | 20.3 | 36.9 | 46.1 | 52.5 | 39.0 | 73.0 | 78.0 | 81.1 | 83.1 | **78.8** | 23.7 | 30.5 | 37.5 | 44.9 | 34.1 |
| + OBCache-V&K | 65.5 | 75.3 | 80.1 | 85.8 | **76.7** | 21.5 | 37.0 | 46.0 | 52.3 | **39.2** | 72.7 | 77.9 | 81.4 | 83.2 | **78.8** | 23.7 | 30.9 | 37.8 | 44.7 | **34.3** |
| SnapKV | 56.4 | 71.5 | 78.8 | 83.0 | 72.4 | 19.4 | 34.2 | 44.5 | 53.4 | 37.9 | 69.2 | 78.9 | 83.2 | 86.0 | 79.3 | 24.5 | 31.1 | 39.0 | 47.9 | 35.6 |
| + OBCache-V | 55.8 | 72.2 | 77.6 | 82.0 | 71.9 | 18.6 | 32.1 | 41.7 | 50.7 | 35.8 | 68.5 | 78.0 | 83.7 | 86.6 | 79.2 | 24.3 | 30.3 | 37.7 | 46.3 | 34.7 |
| + OBCache-K | 59.7 | 74.0 | 79.2 | 82.9 | **73.9** | 24.1 | 38.9 | 48.3 | 57.0 | **42.1** | 69.4 | 78.8 | 83.8 | 86.6 | 79.7 | 26.0 | 32.1 | 38.8 | 48.0 | 36.2 |
| + OBCache-V&K | 59.0 | 73.6 | 78.9 | 82.9 | 73.6 | 22.4 | 38.5 | 48.8 | 58.0 | 41.9 | 69.7 | 79.3 | 83.6 | 86.5 | **79.8** | 26.2 | 31.9 | 39.3 | 48.2 | **36.4** |
| AdaKV | 58.9 | 75.4 | 82.4 | 86.0 | 75.7 | 21.5 | 39.1 | 50.5 | 60.7 | 43.0 | 72.2 | 81.9 | 86.1 | 87.5 | 81.9 | 26.9 | 38.2 | 52.6 | 64.3 | 45.5 |
| + OBCache-V | 66.2 | 81.4 | 86.0 | 87.6 | 80.3 | 23.9 | 49.0 | 61.9 | 70.9 | 51.4 | 76.5 | 85.4 | 87.7 | 88.6 | 84.6 | 30.3 | 46.1 | 60.9 | 72.4 | 52.4 |
| + OBCache-K | 69.2 | 82.6 | 86.3 | 88.3 | 81.6 | 29.2 | 52.0 | 65.5 | 73.2 | 55.0 | 78.1 | 86.1 | 88.0 | 88.5 | **85.2** | 31.7 | 47.5 | 63.6 | 74.8 | 54.4 |
| + OBCache-V&K | 70.1 | 82.7 | 86.3 | 88.4 | **81.9** | 30.0 | 52.2 | 65.1 | 73.6 | **55.2** | 78.2 | 86.1 | 88.1 | 88.6 | **85.2** | 32.6 | 48.1 | 64.2 | 75.4 | **55.1** |

- **Integration into 4 Attn-only Baselines**: $H_2O$, TOVA, SnapKV, AdaKV.
- **Static Prefill Cache Eviction**: consistently improves RULER (4K, 32K) & LongBench performance under both query-aware & -agnostic settings, for both LLaMA-3.1-8B & Qwen-2.5-7B, at all compression rates.
- **Dynamic Decoding Eviction**: consistently reduces PG19 language modeling ppl.
- **Comparison of OBCache Scores**:
  - value score: simplest, minimal overhead.
  - key-aware scores: more output-sensitive, more accurate token saliency estimation.

*Dynamic Decoding Eviction*