# Towards Dynamic KV-Cache Compression: Fine-Grained Evaluation of Key/Value Ranks in LLMs

Jian Chen, Zhuoran Wang, Jiayu Qin, Ming Li, Meng Wang,
Changyou Chen, Yin Chen, Qizhen Weng, Yirui Liu

*TeleAI, University at Buffalo, Delft University of Technology, University of Maryland*

## 1. Motivation: KV-Cache as the Real Bottleneck

Large language models repeatedly read and write a growing key–value (KV) cache during autoregressive decoding. As context length increases, **memory bandwidth—not compute—becomes the dominant bottleneck**. Existing KV compression approaches typically assume: (1) **a *fixed* compression ratio for all layers; (2) ignoring data-dependent low-rank structure; (3) ignoring layer-wise heterogeneity**

## 2. Method: Dataset-Level Rank Analysis via Incremental SVD

While SVD yields the optimal low-rank approximation of a matrix, most prior KV-cache compression methods apply SVD only to the **projection weights**—minimizing

$$||W_{down} \cdot W_{up} - W_k||_F$$

which does **not** minimize the true compression error of the KV-cache,

$$||W_{down} \cdot W_{up} \cdot X - W_k \cdot X||_F$$

The latter depends on the **data-induced activations** X and therefore requires a **data-dependent** formulation.

We use an **incremental SVD** to compute the singular value spectrum of key/value activations for a layer, directly over large datasets. This produces the **optimal** low-rank approximation of the KV-cache itself and allows us to use the resulting singular values to estimate the **compressibility** of each layer's KV-cache.

**Algorithm 1** An adaptive algorithm to compute $\Sigma$ and $\mathcal{V}$ of $K$

**Input:** Dataset containing $l$ tokens in total; LLM model weights $W^K$
**Output:** singular values $\Sigma$ and right singular vectors $\mathcal{V}$ of $K$

$C \leftarrow m_h d_h \times m_h d_h$ zero matrix ▷ Initialize covariance matrix to zero
**for** $t = 1, \ldots, l$ **do**
  $\mathbf{k}_t \leftarrow \mathbf{x}_t W^K$
  $C \leftarrow C + \mathbf{k}_t^T \mathbf{k}_t$ ▷ Update covariance matrix in each step
**end for**
$\mathcal{V}, \Sigma^2, \mathcal{V}^T \leftarrow$ eigen-decomposition$(C)$ ▷ perform eigen-decomposition on the final covariance matrix
**return** $\Sigma, \mathcal{V}$

## 3. Metric: Normalized Effective Rank (NER)

- Measures the intrinsic dimensionality of K/V activations
- Strongly correlates with perplexity degradation under truncation

$$\mathrm{NER}(M) = \frac{1}{d} \exp\left( -\sum_{i=1}^{d} \frac{\sigma_i}{\sum_{j=1}^{d} \sigma_j} \log \frac{\sigma_i}{\sum_{j=1}^{d} \sigma_j} \right)$$

## 4. Findings: Systematic KV Compressibility Patterns

Key results:

- Keys are consistently more compressible than values
- Language effects > domain effects
- Older models (e.g., LLaMA-2) show lower NER → easier to compress
- Low-resource languages exhibit rank collapse (e.g., Arabic)

*Average NER of keys and values across all layers of 7 models on diverse datasets*

| Datasets | Qwen3-4B | | Qwen3-8B | | Gemma-2B | | Gemma-7B | | Mistral-7B | | Phi-3 | | LLaMA-2-7B | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | K | V | K | V | K | V | K | V | K | V | K | V | K | V |
| *Multi-domain Datasets* | | | | | | | | | | | | | | |
| Alpaca | 0.428 | 0.724 | 0.452 | 0.753 | 0.612 | 0.900 | 0.359 | 0.469 | 0.449 | 0.773 | 0.409 | 0.616 | 0.023 | 0.464 |
| MedAlpaca | 0.429 | 0.723 | 0.452 | 0.752 | 0.594 | 0.889 | 0.344 | 0.455 | 0.441 | 0.771 | 0.403 | 0.604 | 0.176 | 0.310 |
| CodeAlpaca | 0.421 | 0.708 | 0.443 | 0.737 | 0.589 | 0.869 | 0.321 | 0.429 | 0.420 | 0.733 | 0.380 | 0.571 | 0.028 | 0.337 |
| WizardCoder | 0.425 | 0.726 | 0.447 | 0.753 | 0.597 | 0.889 | 0.329 | 0.445 | 0.420 | 0.750 | 0.385 | 0.587 | 0.265 | 0.304 |
| FunctionCall | 0.432 | 0.731 | 0.451 | 0.756 | 0.608 | 0.900 | 0.342 | 0.458 | 0.432 | 0.762 | 0.397 | 0.604 | 0.135 | 0.449 |
| *Multilingual Question in VisR-Bench Datasets* | | | | | | | | | | | | | | |
| English | 0.424 | 0.717 | 0.446 | 0.745 | 0.597 | 0.884 | 0.337 | 0.448 | 0.430 | 0.753 | 0.393 | 0.593 | 0.088 | 0.141 |
| German | 0.383 | 0.640 | 0.400 | 0.666 | 0.536 | 0.802 | 0.305 | 0.400 | 0.392 | 0.676 | 0.345 | 0.523 | 0.092 | 0.138 |
| Italian | 0.377 | 0.632 | 0.392 | 0.655 | 0.529 | 0.793 | 0.299 | 0.393 | 0.387 | 0.669 | 0.339 | 0.515 | 0.084 | 0.135 |
| Swedish | 0.371 | 0.620 | 0.387 | 0.645 | 0.526 | 0.794 | 0.296 | 0.389 | 0.381 | 0.656 | 0.322 | 0.486 | 0.078 | 0.128 |
| Spanish | 0.367 | 0.629 | 0.384 | 0.654 | 0.523 | 0.790 | 0.299 | 0.396 | 0.381 | 0.665 | 0.334 | 0.513 | 0.064 | 0.095 |
| Japanese | 0.361 | 0.619 | 0.380 | 0.648 | 0.506 | 0.775 | 0.276 | 0.369 | 0.364 | 0.636 | 0.321 | 0.467 | 0.079 | 0.125 |
| Arabic | 0.337 | 0.582 | 0.354 | 0.609 | 0.503 | 0.769 | 0.270 | 0.361 | 0.338 | 0.594 | 0.288 | 0.413 | 0.052 | 0.173 |

## 5. Layer-Wise Rank Dynamics

Layer-wise NER for key and value representations in Qwen3-4B, evaluated on five datasets and three VisR-Bench languages. The pronounced non-uniformity across layers demonstrates substantial heterogeneity in intrinsic rank structure. Consequently, uniform compression ratios are fundamentally suboptimal: they risk degrading high-rank layers disproportionately while failing to exploit the considerable low-rank structure present in earlier and mid-depth layers.