

Operator-Based Activation Compression: Query Sensitivity and a KV-Cache Prototype Audit

Daniel Schmitter

Abstract

Temporal recurrences and graph bases suggest compact representations of transformer activations, but tensor reconstruction error alone does not qualify a usable key-value cache. We formulate an attention-output perturbation bound that separates key-induced weight changes from value reconstruction error, and give a two-token counterexample with vanishing relative key error but nonvanishing output error. We then audit an operator-compression prototype containing 259,200 rows across ten saved CSV files. The implementation explores temporal polynomial and exponential bases, low-rank factors, graph transforms, and quantization controls. Its attention diagnostic omits a causal mask, reconstructed GPT-2 queries omit the block’s pre-attention normalization, and several payload estimates charge reduced precision without decoding rounded factors. These findings prevent the archived scores from establishing end-to-end cache quality or realized memory savings. Bounded arithmetic checks reproduce the masking and query-sensitivity failures without loading a language model. The contribution is a reproducible analysis of the representation-to-query boundary and a negative qualification result, not a competitive cache codec.

1 Introduction

The key-value cache of an autoregressive transformer grows with retained tokens. A low-dimensional operator trajectory is an appealing alternative to storing every activation independently: temporal exponentials, polynomial trends, or graph modes may describe repeated structure. However, a cache exists to answer future attention queries. Its useful distortion measure depends on that computation, not merely on the Euclidean geometry of its stored tensors.

We examine this distinction through a deterministic perturbation analysis and a source audit of an existing exploratory prototype. The analysis identifies quantities that control a fixed attention query. The audit asks whether the saved experiments actually measure those quantities with the model’s causal semantics and a realized storage format. It finds that they do not yet establish that application. No new model search or transformer execution is performed.

2 Related work and representation family

Low-rank cache projection and asymmetric quantization are established approaches, exemplified by Palu [1] and KIVI [2]. They make important comparators for an operator-based proposal, although we do not reproduce their complete systems. A uniform scalar quantizer in the prototype is not a reproduction of KIVI, and a standalone SVD of observed activations is not a reproduction of Palu’s complete inference method.

A scalar sequence satisfying a constant-coefficient recurrence has an exponential-polynomial representation, with polynomial factors for repeated roots. A vector activation matrix can combine a low-rank spatial factor with such temporal functions. This is classical recurrence and modal analysis; token index is an ordering coordinate, not evidence that a transformer obeys a physical differential equation.

$$x_{t+K} + \sum_{j=0}^{K-1} h_j x_{t+j} = 0, \quad (1)$$

$$x_t = \sum_k \sum_{m=0}^{m_k-1} c_{km} t^m z_k^t, \quad (2)$$

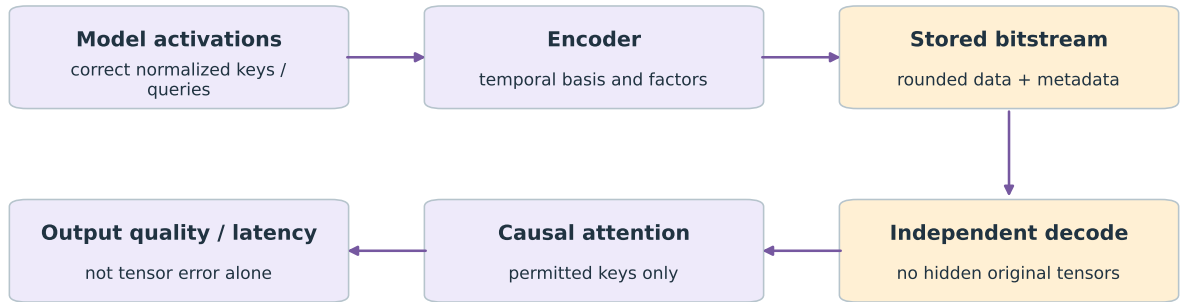
$$X \approx B_{\text{time}} C W^T. \quad (3)$$

The prototype includes fixed polynomial and exponential-polynomial bases, DMD-style modal fitting, piecewise windows, SVD, and graph-spectral variants. A deterministic fixed basis may be regenerated from metadata. A data-dependent basis must be stored or reconstructed from retained information. Counting its coefficients while silently retaining the original attention graph would not be a valid cache compression ratio.

3 Architecture and information flow

The relevant architecture includes a communication boundary even inside one process: after encoding, attention must consume only information recoverable from the declared retained object. A learned basis, graph transform, scale, or factor must therefore be stored or deterministically regenerated. Decoding rounded factors is part of the measured operation. The original cache is not available to repair the reconstruction at evaluation time.

A compressed cache must answer the real query



Read left to right above, then right to left below. Arrows show data or computation flow.

Figure 1: A compressed cache must answer the real query. Boxes distinguish supplied information, fitted components, and the quantity evaluated. Arrows show computation or data dependence, not a newly trained deep network.

The query-sensitive bound explains why this boundary matters statistically as well as computationally. A small key error can create a large change in relative logits when aligned with a large query. A causal mask changes the normalization set itself. An unmasked metric with incorrectly extracted queries can therefore rank representations for a different operation than the transformer performs. More exploratory rows do not fix the wrong customer for the compressed data.

4 From key and value errors to attention error

For one query and its permitted keys, define logits, attention probabilities, and output as follows. Masked-out positions are excluded from the index set, and the same permitted set is used by both calculations.

$$\ell_j = q^T k_j / \sqrt{d}, \quad p_j = \frac{e^{\ell_j}}{\sum_i e^{\ell_i}}, \quad o = \sum_j p_j v_j, \quad (4)$$

$$\hat{o} = \sum_j \hat{p}_j \hat{v}_j. \quad (5)$$

Let ε_V be the largest value-vector reconstruction error in a chosen norm, $V_{\max}$ the largest original value norm, and Ω the range of the logit perturbations. Decomposing output error into changed values at fixed approximate weights and changed weights at fixed original values yields

$$\|\hat{o} - o\| \leq \epsilon_V + V_{\max} \|\hat{p} - p\|_1 \leq \epsilon_V + 2V_{\max} \tanh(\Omega/4), \quad (6)$$

$$\Omega = \max_j (\hat{\ell}_j - \ell_j) - \min_j (\hat{\ell}_j - \ell_j). \quad (7)$$

The second inequality follows by bounding the exponential likelihood ratios of the two normalized categorical distributions. An endpoint distribution maximizes total variation under a bounded log-ratio range. If every key error has Euclidean norm at most ϵ_K , Cauchy-Schwarz gives Omega at most twice the query norm times ϵ_K divided by the square root of d . This classical perturbation argument is stated as a diagnostic, not a newly discovered softmax theorem. Relative output error additionally needs a nonzero lower bound on the original output norm.

5 A two-token counterexample

Take scalar keys one and one plus epsilon, reconstruct both as one, use query one divided by epsilon, and choose values minus one and one. The relative Frobenius key error tends to zero, while the logit gap stays one. The exact output is $\tanh(1/2)$ and the reconstructed output is zero. Therefore small relative tensor error alone cannot guarantee a small query-output error over unrestricted query norms.

$$K = \begin{pmatrix} 1 \\ 1 + \epsilon \end{pmatrix}, \quad \hat{K} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad q = 1/\epsilon, \quad V = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad (8)$$

$$\frac{\|K - \hat{K}\|_F}{\|K\|_F} \rightarrow 0, \quad |o - \hat{o}| = \tanh(1/2). \quad (9)$$

This deliberately adverse construction does not assert that actual GPT-2 queries have unbounded norm. It shows the missing assumption required to turn a reconstruction score into a query guarantee. The same principle applies to an attention-weighted distortion metric: its usefulness depends on which queries, masks, and execution stage generated the weights.

6 Experimental methods

The source audit reads all ten saved benchmark CSV files in place and retains their hashes, schemas, method counts, prompt identifiers, token lengths, nonfinite counts, and numerical ranges. A row is a prompt/layer/head/method-setting diagnostic, not an independent language task. The collector’s five default prompts include prose, code, and repeated tokens. The maximum-token command-line parameter truncates rather than pads or extends them; a filename containing 256 or 512 does not establish that many tokens were observed.

The audit follows the query collector, attention metric, method dispatch, reconstruction functions, and byte estimators. It does not alter the original code or replace historical scores with results from a repaired implementation. Deterministic tests use the two-token counterexample at epsilon values 0.1, 0.01, 0.001, and 0.0001. A second two-position test compares causal and unmasked attention for identical zero-gap logits and values zero and one. No model weights, prompts, or external datasets are downloaded.

7 Results

The query issue is visible by comparing the collector with the GPT-2 block implementation: normalization precedes attention in the model [3], while the auxiliary collector applies the query projection to the stored block input directly. The metric then compares full all-to-all attention without a causal mask. These two discrepancies affect interpretation even if every arithmetic operation in a compressor is correct.

The four query-sensitivity fixtures produce absolute output discrepancy approximately 0.462117 while relative key error falls to approximately 0.00007071. All satisfy the displayed perturbation bound. The masking fixture has a Frobenius output discrepancy of 0.5: the first position incorrectly uses a future value in the unmasked calculation. These are reproducible arithmetic counterexamples, not new language-model benchmark results.

Table 1: Source-audit findings and the claim each prevents.

Finding	Consequence
Query extraction applies the attention projection to raw block input	Not the model’s pre-normalized query
Attention-output metric has no triangular mask	Not causal prefill attention for the supplied query rows
SVD factors reconstructed before reduced-precision rounding	Error does not match the advertised factor precision
Quantizer retains int32 codes while pricing 4/8-bit payloads	Estimated bit count, not serialized or peak memory
learned_exp_poly dispatches to the fixed exponential basis	Not an independently learned-pole algorithm
Reconstructions replace nonfinite values before scoring	Sanitized scores do not establish numerical stability

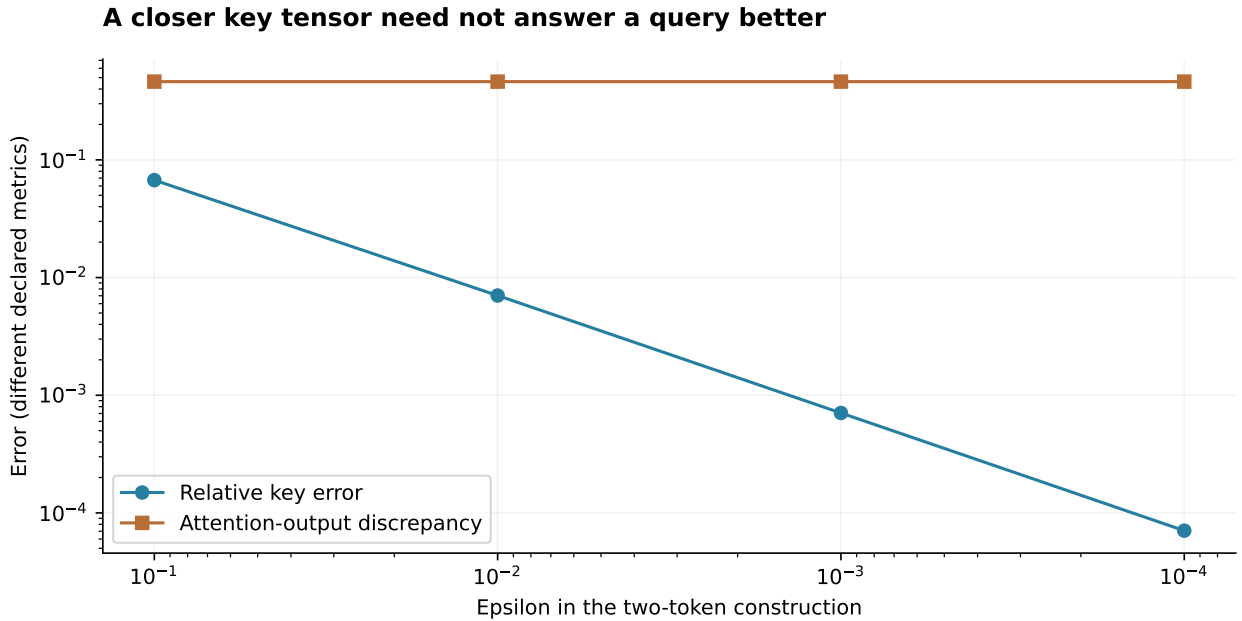


Figure 2: Analytic two-token counterexample from the stated equations. Relative key error shrinks, but attention-output discrepancy stays at $\tanh(1/2)$. Query magnitude changes with epsilon; this is not measured transformer performance.

The archive contains 259,200 rows, but row count does not repair these semantic defects or provide independent replication. Numerical summaries are retained for traceability, not ranked into a winning method. In particular, selecting the best layer, prompt, graph, rank, and basis after inspecting the full exploratory sweep would not constitute a frozen application evaluation. No causal-generation quality, decoded-cache latency, persistent-byte measurement, or end-to-end peak-memory advantage is established by this record.

8 Discussion

The operator hypothesis remains mathematically coherent: a sequence with a useful recurrence can be represented compactly. The missing bridge is a decoded representation that preserves the actual attention queries under a truthful storage budget. A fixed deterministic temporal basis has different metadata requirements from a data-dependent graph basis. A cached spectral transform is not free merely because it can be reused inside one diagnostic.

The source-level failures also distinguish a research prototype from a codec. A codec must serialize, decode, and use the rounded representation in the measured computation. The prototype estimates several such costs but does not yet implement that boundary consistently. Our audit preserves the useful exploratory machinery and documents why its original attention-error scores cannot be promoted into a deployment claim.

The supplement supplies source fingerprints, complete file-level accounting, all bounded counterexamples, and the audited source modules. This technical report is a negative qualification study. It does not establish

a novel competitive compression algorithm, and editorial reconstruction cannot supply that missing evidence. A future repaired evaluation would be a separately named experiment rather than a silent correction of this historical result.

9 Application boundary and research implication

The useful surviving result is a testable codec interface and a counterexample to a misleading distortion criterion. A future operator codec must demonstrate end-to-end generation and actual bytes under that interface. The historical sweep is not repaired by editorial relabeling, and no best setting from it is promoted as a deployed result.

10 Conclusion

Activation compression must be judged through the queries the retained representation serves. A small key reconstruction error can coexist with a substantial attention error, and an unmasked surrogate with incorrect queries cannot qualify a causal KV cache. The prototype offers an operator-representation research testbed but no established end-to-end compression advantage. Explicit query sensitivity and serialization accounting are necessary before such an advantage can be claimed.

References

- [1] C.-C. Chang et al. Palu: Compressing KV-Cache with Low-Rank Projection. arXiv:2407.21118, 2024.
<https://arxiv.org/abs/2407.21118>
- [2] Z. Liu et al. KIVI: A Tuning-Free Asymmetric 2bit Quantization for KV Cache. ICML, 2024.
<https://arxiv.org/abs/2402.02750>
- [3] Hugging Face Transformers developers. GPT-2 model implementation, GPT2Block forward computation. Source inspected September 2026. https://github.com/huggingface/transformers/blob/main/src/transformers/models/gpt2/modeling_gpt2.py