SAME MEMORY, BETTER ANSWERS: CHOOSING KV CACHES BY THEIR OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

KV-cache compression research asks which tokens to keep, and judges each compressor by the one cache it builds. This misses a larger lever. At a fixed memory budget, different compressors succeed on different questions: choosing the best of their caches for each question matches or beats the uncompressed model while keeping one-tenth of the document. Existing methods cannot make this choice, because every signal they rely on—attention mass, reconstruction error, hidden-state distance—measures what a cache preserves inside the model, not what the model will say. On question answering over research papers, retained attention mass ranks candidate caches worse than a coin flip. We select caches by their outputs instead: each candidate is scored by how far it moves the next-token distribution from the full cache's. The principle follows from the geometry of the unembedding, which weights every internal change by its effect on the tokens the model is about to emit—information that internal distances discard. At the same memory and compute, output-based selection picks better caches than internal-state features on question answering, beats fixed policies tuned on training data, and transfers without retraining from 7B to 70B models and to 128K-token documents. Internal features win only where the answer is a verbatim copy: synthetic needle retrieval.

1 INTRODUCTION

The KV cache bounds long-context inference: on our 70K–128K-token books, Llama-3.1-8B's document cache alone takes a median 11.2 GiB per sequence. Compressors cut this cost by evicting tokens, ranking them by accumulated, windowed, last-token, or expected attention (Zhang et al., 2023; Li et al., 2024; Oren et al., 2024; Devoto et al., 2025). Each compressor produces one cache, and the field compares compressors against each other. That framing leaves quality on the table, because no compressor wins everywhere. At 10% retention, TOVA scores 48.8 F1 on RULER where EXPECTEDATTENTION scores 2.9, yet trails EXPECTEDATTENTION by 10.0 F1 on Qasper. Per question, the spread is wider still: picking the best of eleven compressed caches, each holding 10% of the document, would match the uncompressed cache on Qasper and beat it by 11.4 and 8.0 F1 on 2WikiMQA and RULER. At this budget, memory does not limit answer quality; the choice of cache does.

Capturing that quality requires a score that predicts, before decoding, which cache will answer well. The obvious candidates are the quantities compressors already optimize—retained attention mass and attention-reconstruction error—and distances between compressed and full-cache hidden states. All of them measure how much of the model's computation survives, not what the model will say. On Qasper, retained attention mass ranks the worse of two caches higher in 54.4% of pairs: worse than a coin flip.

We score each candidate cache by its effect on the output. The question runs once against the full cache and once against each candidate, and two features compare the resulting next-token distributions: KL divergence and top-10 token overlap. The same passes yield eight internal-state features, so the comparison is exact: same eleven candidates, same memory, same compute, different features. Selectors trained on the two output features choose caches that score 4.64 F1 higher on Qasper and 3.70 F1 higher on 2WikiMQA than selectors trained on the eight state features (Figure 1), and they beat every fixed policy chosen on training documents on both QA workloads. KL cuts Qasper ranking errors by a third. Frozen after training on 7–8B models, the output selectors choose better caches than

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

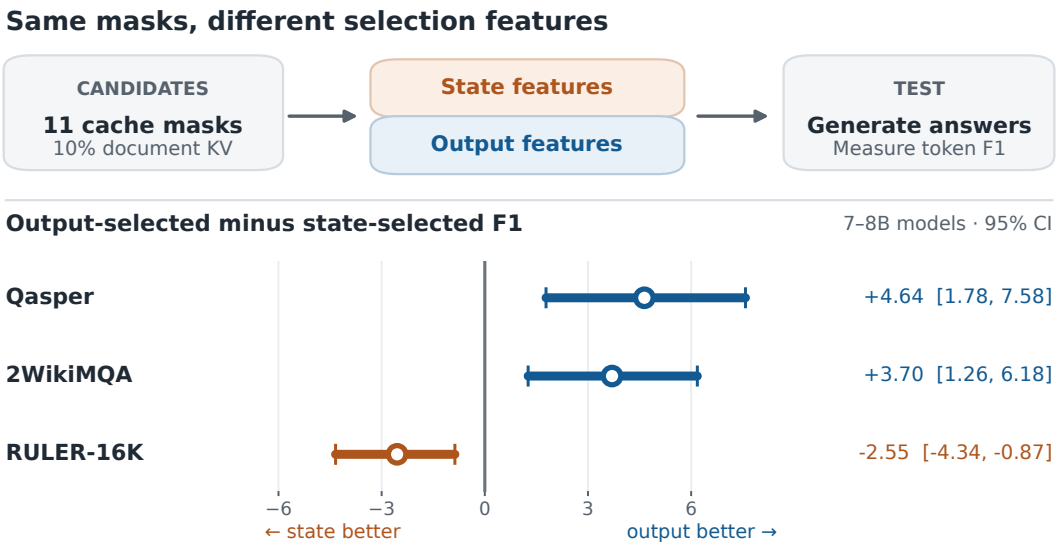


Figure 1: Same eleven candidate caches, same memory, same forward passes; only the selection features differ. Each candidate retains 10% of the document cache. The plot shows the paired held-out F1 difference, output-selected minus state-selected, averaged over three 7–8B models, with 95% document-bootstrap intervals. Output features win both QA workloads; state features win literal retrieval (RULER).

EXPECTEDATTENTION on Qwen3-32B and Llama-3.1-70B in all six model–task cells, and gain 6.68 F1 over EXPECTEDATTENTION on books of up to 128K tokens. On RULER, where every answer is a literal copy of a number, state features lead by 2.55 F1; a ranker that adds eight reference-trajectory tokens matches the best fixed policy there.

A second-order expansion of KL explains the split (Equation 2). The unembedding weights each direction of hidden-state change by its effect on the probabilities of the tokens the model is likely to emit; a state norm weights every direction equally and so conflates harmless changes with harmful ones. The worst-case consequence is stark: a norm-based certificate of top-token agreement fires in none of 7,805 measured cases, although the top token agrees in 3,276 of them.

The study spans 31,632 generated answers—28,996 from compressed caches and 2,636 full-cache references—from five checkpoints in three model families (Mistral-7B, Llama-3.1-8B/70B, Qwen3-8B/32B) on Qasper, packed 2WikiMQA, and RULER-16K, plus baseline extensions and full-length InfiniteBench runs. Our contributions are:

1. **Cache selection is a first-order lever for KV compression.** At a fixed 10% budget, the per-question best of eleven caches matches or beats the uncompressed cache on all three workloads, and no single compressor, including TOVA, wins all of them (§4.1, §6).

2. **Output features select better caches than internal-state features, at identical cost.** With the same candidates and forward passes, two output features beat eight state features by 4.64 and 3.70 F1 on QA and beat training-selected fixed policies. Retained attention mass misranks Qasper caches worse than chance; KL removes a third of those errors. The selectors transfer unchanged to 32B and 70B models and to 128K-token inputs (§4–§6).

3. **A geometric account of why output features win, and a workload where they do not.** The KL expansion shows what state distance discards; RULER shows where state features lead and how reference trajectories recover the gap (§2, §4).

4. **An open testbed for cache selection.** We release every generated answer, mask hash, feature, saved selector, and analysis script, so new selectors can be trained and scored on the same 31,632 outcomes without generating new answers (Appendix F).

Selection runs entirely before decoding: it costs one prefill per candidate cache plus question-length scoring passes, and it leaves decoding with 10% of the document cache (§6).

2 CHOOSING A CACHE BY ITS OUTPUTS

The selection problem. Let d be a document of N tokens, q a question, and m a mask that retains, for every layer l and KV head h, exactly $k = \lfloor 0.1N \rfloor$ distinct positions $m_{l,h} \subset \{0, \ldots, N-1\}$. Gathering the retained keys and values, with their original rotary positions, gives a compressed cache that holds 10% of the document; question and generated tokens append to it during decoding. The outcome $Y(d, q, m)$ is the token-F1 (0–100) of the greedy answer, maximized over accepted references. Given K candidate masks built by different compressors, a selector picks one before decoding and without seeing the answer. All candidates occupy the same memory, so every F1 point separating the chosen cache from the best candidate is quality the budget already paid for.

State features. We run the question once against the full cache (the reference) and once against each candidate cache. Eight internal-state features compare the candidate with the reference at four sites—attention input, attention output, post-attention residual, and post-MLP residual—by relative ℓ_2 distance and cosine similarity, averaged over the final four question positions and over layers.

Output features. The same passes give the next-token distributions p_F (full cache) and p_m (candidate) at the final question position. The two output features are $D_{\mathrm{KL}}(p_F \| p_m)$ and the overlap between the ten most probable tokens under p_F and p_m. Output and state features therefore cost exactly the same: $(K+1)T_q$ token positions for a question of T_q tokens, one reference pass plus one pass per candidate, with no decoding. Section 6 reports the measured wall-clock time and memory.

Frozen-attention features. Four further features need no candidate pass and measure what attention-based compressors optimize. For one query position and head, let a_i be the full-cache attention probability of document token i, v_i its value, o the full attention output, and B the always-retained question tokens, with $b = \sum_{i \in B} a_i (v_i - o)$ and $c = \sum_{i \in B} a_i$. With queries, keys, and values held fixed, the attention output of retained set S deviates from o by exactly

$$e(S) = \frac{\left\| b + \sum_{i \in S} a_i (v_i - o) \right\|_2}{c + \sum_{i \in S} a_i}. \tag{1}$$

The question terms b and c matter: question tokens carry attention mass and enter the renormalization. The frozen features are the mean and upper-tail response error e and the mean and minimum retained attention mass $c + \sum_{i \in S} a_i$, computed from the reference pass (Appendix A).

Why output divergence separates caches that state distance cannot. The model maps its final normalized hidden state to logits through the unembedding W_U. For a candidate-minus-reference state change Δh,

$$D_{\mathrm{KL}}(p_F \| p_m) = \frac{1}{2} \Delta h^\top W_U^\top \left[\mathrm{diag}(p_F) - p_F p_F^\top \right] W_U \Delta h + O(\|W_U \Delta h\|_\infty^3). \tag{2}$$

The quadratic term is half the variance, under p_F, of the logit change $W_U \Delta h$ (Appendix D). It weights each direction of state change by its effect on the probabilities of the tokens the model is likely to emit. A state norm weights every direction equally. Only the part of a change that moves the logits of likely tokens relative to one another counts: two changes of the same length can produce very different KL, and moving every logit equally changes nothing. The worst-case alternative confirms the point: a sufficient condition that bounds logit changes by hidden-state distance (Appendix C) certifies top-token agreement in none of 7,805 measured candidate cases, although the top token agrees in 3,276 of them (Table 14). State distance measures how far the cache moves the model; output divergence measures whether that movement reaches the output.

Learning a selector. A selector scores the K candidates for a question and keeps the highest-scoring one. We fit linear selectors on the eight state features or the two output features; further comparisons use the four frozen-attention features, all fourteen features, and a small neural ranker. Nonnegative features are log-transformed, and all features are standardized across the candidates

of each question. Training minimizes a regularized pairwise logistic loss over candidate pairs with unequal F1, weighted by their F1 difference. Five outer document folds give held-out predictions; three inner folds choose hyperparameters by generated F1. All questions and model runs for a document stay in one fold. Answers supervise training only; applying a fitted selector uses candidate features alone (Appendix B).

Reference trajectories. A richer variant scores candidates along the greedy full-cache continuation at 4, 8, or 16 token positions. The first position comes from the question pass; later positions require sequential decoding with the full cache. The trajectory is model-generated and uses no gold answers.

Ranking errors. Orient every score s so that higher predicts higher F1 (attention mass keeps its sign; error and KL are negated). A pair of physically distinct candidates m, m' is a *reversal* when

$$[s(m) - s(m')][Y(d, q, m) - Y(d, q, m')] < 0. \tag{3}$$

We count reversals over pairs with nonzero score and quality differences, after removing duplicate masks by index hash, and report the excluded ties (Table 8). A score that orders caches at random reverses 50% of pairs; a perfect score reverses none.

3 EXPERIMENTAL SETUP

Eleven candidate caches. Every candidate retains exactly $\lfloor 0.1N \rfloor$ document tokens per layer and KV head; no candidate uses answer labels. We group them by the information used to build the mask (Appendix A):

- *Prefix-only (query-blind)*: EXPECTEDATTENTION (Devoto et al., 2025), SNAPKV (Li et al., 2024), and a four-sink Start–Recent window (Xiao et al., 2024). These see only the document prefill.
- *Query-conditioned*: question-specific scalar and vector response-error objectives, built from full-cache attention at the active question's final four positions.
- *Document-shared*: mean observed attention, shared scalar response error, and three shared vector objectives, pooling attention traces over all benchmark questions of the document.
- *Random*: seeded random retention, independent of the questions.

FULLKV retains the whole document and serves as the uncompressed reference. The ranking-error analysis uses a five-compressor pool of strong, distinct candidates—EXPECTEDATTENTION, SNAPKV, mean attention, shared scalar, and question-specific vector—and we also report the full eleven-candidate grid.

Models and workloads. The primary grid runs Llama-3.1-8B-Instruct, Qwen3-8B, and Mistral-7B-Instruct-v0.3 (Grattafiori et al., 2024; Qwen Team, 2025; Jiang et al., 2023) in bfloat16 with PyTorch SDPA on three workloads with 339 evaluation documents. Qasper (Dasigi et al., 2021; Bai et al., 2024) contributes 139 papers with 173 questions, after excluding 9 documents used during development. Packed 2WikiMQA (Ho et al., 2020) concatenates two LongBench multi-hop contexts into each of 100 documents; each document carries two questions, each answerable from one half (200 questions). RULER-16K (Hsieh et al., 2024) splits 100 multi-key needle-in-a-haystack contexts of 16K tokens into four single-key questions each (400 questions). The executed grid covers 348 documents, 1,044 model–document runs, and 28,800 generated answers; the 339 evaluation documents contribute 1,017 runs and 27,828 answers. Section 5 applies the trained selectors, unchanged, to Qwen3-32B and Llama-3.1-70B-Instruct. Two packed 2WikiMQA documents exceed Mistral's configured 32K-token limit (Table 18); excluding them moves no Mistral selector result by more than 0.15 F1 (Table 19).

Execution and uncertainty. All runs use one node with eight NVIDIA B300 GPUs. Scores average questions within each document and model, then average models, so documents with many questions do not dominate. All 95% intervals come from 20,000 paired document-level bootstrap draws that resample whole documents with every question and model paired.

Table 1: Held-out F1 change from EXPECTEDATTENTION (95% paired document-bootstrap intervals). "Frozen" uses four frozen-attention features, "state" eight internal-state features, "output" two next-token features, and "all" their 14-feature union. Absolute F1 for every candidate is in Table 24.

Features (count)	Ranker	Qasper	2WikiMQA	RULER-16K
States (8)	Linear	−1.89 [−5.74, 1.99]	+2.91 [−0.60, 6.41]	+41.05 [38.76, 43.38]
States (8)	MLP	−0.66 [−4.42, 3.01]	+4.08 [0.31, 7.84]	+43.09 [40.79, 45.39]
Output (2)	Linear	+2.75 [−0.04, 5.53]	+6.62 [3.51, 9.73]	+38.50 [36.34, 40.59]
Output (2)	MLP	+2.72 [−0.08, 5.50]	+6.58 [3.47, 9.69]	+38.51 [36.39, 40.57]
Frozen + states (12)	Linear	−4.32 [−8.43, −0.26]	+4.35 [0.82, 7.88]	+42.18 [40.17, 44.25]
Frozen + states (12)	MLP	−1.45 [−5.10, 2.15]	+5.28 [1.86, 8.72]	+47.00 [44.81, 49.25]
Frozen + output (6)	Linear	+3.04 [0.15, 5.91]	+5.83 [2.58, 9.10]	+40.74 [38.75, 42.76]
Frozen + output (6)	MLP	+3.24 [0.40, 6.09]	+6.11 [2.93, 9.25]	+47.99 [45.83, 50.11]
All (14)	Linear	+3.15 [0.28, 6.02]	+6.62 [3.38, 9.79]	+40.93 [38.86, 43.08]
All (14)	MLP	+3.00 [0.11, 5.89]	+6.97 [3.86, 10.04]	+48.98 [46.76, 51.20]

Table 2: Strict quality-preference reversals on the five-compressor pool (%; lower is better; 50% is random ordering). Intervals resample documents. Duplicate physical masks are removed; excluded ties are in Table 8.

Workload	Mass	Local error	Next-token KL
Qasper	54.38 [50.63, 58.14]	52.20 [48.71, 55.80]	35.63 [32.52, 38.86]
2WikiMQA	42.21 [37.32, 47.10]	40.77 [35.80, 45.82]	38.85 [34.10, 43.71]
RULER-16K	14.54 [12.11, 16.93]	11.67 [8.72, 14.66]	13.49 [11.76, 15.19]

4 OUTPUT FEATURES PICK BETTER CACHES

Same candidates, same passes, better answers. Output-feature selection beats state-feature selection by +4.64 F1 on Qasper and +3.70 F1 on packed 2WikiMQA; state features lead by 2.55 F1 on RULER. The paired 95% intervals are [1.78, 7.58], [1.26, 6.18], and [−4.34, −0.87]. Both selectors see the same held-out questions, the same eleven candidates, and features from the same forward passes; only the features differ. The neural ranker reproduces the ordering: +3.38, +2.50, and −4.58 F1 (Table 9). Two output features also carry nearly all the QA signal in the fourteen-feature union: adding the eight state and four frozen features to them moves Qasper by +0.40 F1 and 2WikiMQA by 0.00 (Table 1). Relative to EXPECTEDATTENTION, output selection gains +2.75 [-0.04, 5.53] F1 on Qasper and +6.62 [3.51, 9.73] on 2WikiMQA, reaching 90% of FULLKV F1 on 2WikiMQA with 10% of the document cache.

Attention mass ranks caches worse than chance on Qasper. On Qasper's five-compressor pool, retained attention mass reverses 54.38% of quality preferences—significantly worse than random ordering—while next-token KL reverses 35.63%, one third fewer. The intervals are [50.63, 58.14] and [32.52, 38.86] (Table 2). KL also ranks best on 2WikiMQA (38.85%). On RULER, where every answer is a literal copy of a number, all three scores rank well and frozen response error ranks best (11.67%). The ranking diagnostic mirrors the selector results. Where attention quantities misrank caches, output features win selection; where they rank well, state features win. The four frozen-attention features alone lose 3.29 F1 to EXPECTEDATTENTION on Qasper and gain +44.20 on RULER (Table 6).

Adding features and reference tokens closes the RULER gap. A neural ranker over all fourteen features keeps the QA gains and trails the model-specific fixed policy on RULER by only 1.59 F1. Scoring candidates along an eight-token reference trajectory then adds +9.05 [7.48, 10.67] F1 on RULER for the linear ranker and +1.99 [0.80, 3.24] for the neural ranker, against at most +0.53 on the QA workloads (Table 22). With these features, the neural ranker beats the model-specific fixed policy by +4.71 F1 on Qasper and +4.72 on 2WikiMQA and matches it on RULER (+0.40; Table 21). Output features carry the QA signal; literal retrieval rewards scoring candidates on more of the full-cache continuation. Observed trajectory lengths are in Table 20.

Stable across partitions and held-out domains. Across twenty document repartitions, the output selector's Qasper gain over EXPECTEDATTENTION stays between +2.47 and +2.76 F1 (mean +2.72),

Table 3: Oracle ceilings and baselines, in F1 relative to EXPECTEDATTENTION. Per-question and shared-policy ceilings use all eleven compressed candidates; the shared-mask ceiling excludes the two question-specific candidates. Sibling-question transfer and its matched ceilings use multi-question documents. Uniform choice averages nine policies, excluding random tokens and start–recent.

Estimand	Qasper	2WikiMQA	RULER-16K
Per-question ceiling	+16.02 [13.06, 19.24]	+22.38 [19.28, 25.61]	+56.78 [54.48, 59.09]
Shared-policy ceiling	+15.14 [12.26, 18.31]	+19.60 [16.61, 22.72]	+55.33 [53.07, 57.63]
Shared-mask ceiling	+14.95 [12.06, 18.13]	+19.10 [16.14, 22.23]	+52.13 [49.78, 54.56]
Query ceiling (multi-Q)	+19.45 [13.85, 25.59]	+22.38 [19.28, 25.61]	+56.78 [54.48, 59.09]
Shared-policy ceiling (multi-Q)	+15.85 [10.72, 21.69]	+19.60 [16.61, 22.72]	+55.33 [53.07, 57.63]
Sibling-question transfer	−1.01 [−5.61, 4.91]	−0.37 [−2.83, 2.31]	+46.82 [44.29, 49.37]
Uniform nine-policy choice	−3.58 [−7.27, 0.16]	+3.17 [0.42, 5.95]	+34.08 [32.55, 35.61]
Random-token mask	−25.35 [−29.19, −21.74]	−25.42 [−29.61, −21.33]	−2.80 [−3.90, −1.85]

while the state selector falls below EXPECTEDATTENTION in all twenty (mean −1.91); Table 11 reports each split. Output features also lead state features on both QA workloads when an entire model family or an entire source dataset is held out of training (Table 10).

4.1 HEADROOM AND FIXED-POLICY BASELINES

The candidate pool already holds full-cache quality. Choosing the best of the eleven 10% caches for each question would reach 43.2, 52.7, and 59.6 F1 on Qasper, 2WikiMQA, and RULER—matching FULLKV (42.6) on Qasper and beating it (41.2 and 51.6) on the other two. These per-question ceilings lie +16.02, +22.38, and +56.78 F1 above EXPECTEDATTENTION (Table 3). Output selection already captures 17%, 30%, and 68% of that headroom. Choosing one policy per document gives up little (+15.14, +19.60, +55.33), yet the policy that scored best on a document's *other* questions gains nothing: −1.01 F1 on Qasper, where sibling questions ask about the same paper. The best cache depends on the question, so selection must read each question's own evidence—which is what candidate features provide. Appendix B defines each ceiling (Equation 11); document counts are in Table 16.

Per-question selection beats fixed policies on QA. A fixed policy applies the single candidate that scored best on the training documents to every test question. Output selection beats the model-specific fixed policy by +4.11 F1 on Qasper and +3.84 on 2WikiMQA (Table 21), and the common fixed policy by +2.75 and +2.15 (Tables 1 and 23). On RULER, question-conditioned candidates nearly match FULLKV—question-specific scalar reaches 50.61 F1 against 51.62—so a fixed rule is hard to beat: output selection finishes 12.08 F1 behind the model-specific policy, and the trajectory-augmented neural ranker matches it (§4).

EXPECTEDATTENTION loses the needle on RULER. EXPECTEDATTENTION falls to 2.80 F1 on RULER, against 51.62 for FULLKV and 48.85 for mean attention pooled over the document's questions. It returns the target number in 3.08% of answers, against 99.17% for FULLKV and 61.08% for mean attention (Table 17). All three query-blind candidates in the pool—EXPECTEDATTENTION, SNAPKV, and start–recent—stay below 6 F1; output selection recovers +38.50 [36.34, 40.59] F1 over EXPECTEDATTENTION by choosing among the other candidates.

5 SELECTORS TRANSFER TO 32B AND 70B MODELS WITHOUT RETRAINING

Output selectors trained only on 7–8B models choose better caches than EXPECTEDATTENTION on Qwen3-32B and Llama-3.1-70B in all six model–task cells, by +9.18 to +84.38 F1. On Llama-3.1-70B RULER, EXPECTEDATTENTION and the frozen fixed policy both score 0.00 F1; state and output selection reach 85.94 and 84.38 (Table 4). On Llama-3.1-70B 2WikiMQA, EXPECTEDATTENTION collapses to 7.03 F1 and output selection reaches 48.64. **Output features add the most on Qwen3-32B: on 2WikiMQA, output selection beats state selection by +8.22 F1 [1.19, 17.12]**, reaching 52.96 F1 against 44.75 for state selection, 42.60 for the fixed policy, and 57.91 for FULLKV. Output features also lead on Qwen3-32B Qasper (+6.56) and RULER (+1.82). On Llama-3.1-70B the two feature sets perform alike: 15 of 16 Qasper documents tie, and the three differences lie within 1.6 F1.

Table 4: Saved 7–8B selectors applied unchanged to larger checkpoints. Each row averages questions within 16 documents; all values are token-F1. "Fixed" is the policy selected on the document's original outer-training fold. The final column is the paired output-minus-state difference with a 95% document-bootstrap interval.

Model	Task	Full	EA	Fixed	State	Output	Output − state [95% CI]
Qwen 32B	Qasper	44.77	16.44	16.44	19.05	25.62	+6.56 [-1.00, 16.82]
Qwen 32B	2WikiMQA	57.91	34.67	42.60	44.75	52.96	+8.22 [1.19, 17.12]
Qwen 32B	RULER-16K	75.31	0.00	32.16	24.09	25.91	+1.82 [-2.73, 7.03]
Llama 70B	Qasper	50.79	15.36	15.36	37.41	37.98	+0.57 [0.00, 1.70]
Llama 70B	2WikiMQA	71.30	7.03	16.63	48.85	48.64	-0.21 [-9.38, 9.16]
Llama 70B	RULER-16K	100.00	0.00	0.00	85.94	84.38	-1.56 [-9.38, 4.69]

Protocol. Each checkpoint runs the first sixteen eligible non-development documents per workload, on the same texts as the original grid, with FULLKV and all eleven candidates for every question: 96 model–document runs and 2,832 generated answers. Each document uses the linear selectors and fixed policy saved from its original outer fold, fit on 7–8B training documents that exclude it; no 32B or 70B label enters any fit or policy choice. Every prompt and generation allowance fits both models' native limits without truncation. Appendix E lists model revisions, mask checks, the fold mapping, and the bootstrap.

6 STRONGER BASELINES, 128K-TOKEN INPUTS, AND COST

No single eviction score wins every workload. On the first 32 non-development documents of each workload and model, we add further eviction scores at the same 10% budget. TOVA's last-token attention score (Oren et al., 2024) is the strongest query-blind compressor we test on RULER: 48.77 F1, against 2.89 for EXPECTEDATTENTION and 52.02 for FULLKV (Table 5). It beats the saved output selector on RULER by 9.08 F1 [4.92, 13.37], yet trails it by 14.15 F1 on Qasper and 10.98 on 2WikiMQA (Table 25). Selectors that add eight reference-trajectory tokens beat TOVA on all three workloads: by +13.37, +11.89, and +4.40 F1 for the neural ranker, and by +14.25, +11.56, and +4.46 for the linear one. TOVA lies outside the eleven-candidate pool, and we added it after the selectors were fixed; every selector choice in this comparison is the saved held-out choice. An AnDPro projection score (Geng et al., 2025), adapted to our per-head budget, scores below EXPECTEDATTENTION on both QA workloads and on passkey retrieval. Five rank fusions of EXPECTEDATTENTION and SNAPKV—three linear mixtures, reciprocal-rank, and max-rank—stay within the bootstrap noise of EXPECTEDATTENTION: all fifteen intervals include zero (Table 7). Merging two compressors' rankings is no substitute for choosing between their caches.

Transfer to 128K-token books. A separate Llama-3.1-8B experiment runs full-length InfiniteBench inputs (Zhang et al., 2024) with no truncation: 32 passkey documents above 125K tokens and 112 questions over 22 English books of 70K–128K tokens. **A linear output selector trained only on the three primary workloads gains +6.68 F1 [2.83, 11.11] over EXPECTEDATTENTION on book QA**, choosing among five compressed candidates with no book-QA answers used for fitting or tuning (Table 12). On passkey retrieval, FULLKV, SNAPKV, and TOVA reach 100.00% accuracy and EXPECTEDATTENTION 93.75%.

Table 5: Baseline extensions on the same 0–100 scale. Upper block: 32 documents per workload and original model; RULER presence is the rate of answers that contain the target number. Lower block: full-length InfiniteBench inputs on Llama-3.1-8B; passkey accuracy and F1 are separate outcomes.

Arm	Qasper F1	2Wiki F1	RULER F1	RULER presence
FullKV	40.77 [32.23, 49.36]	42.26 [33.73, 51.24]	52.02 [49.14, 55.11]	98.96 [97.92, 99.74]
ExpectedAttention	21.77 [16.63, 27.21]	34.41 [27.00, 41.52]	2.89 [1.04, 5.55]	3.12 [1.04, 6.25]
SnapKV	16.23 [10.44, 22.70]	34.01 [24.65, 43.63]	6.90 [1.04, 14.38]	9.90 [1.04, 21.35]
Start–recent	12.49 [8.42, 16.99]	28.44 [20.84, 36.37]	4.71 [2.40, 7.35]	10.42 [5.47, 16.15]
TOVA score	11.75 [6.80, 17.63]	31.26 [22.36, 40.60]	48.77 [45.55, 51.87]	71.88 [68.23, 75.26]
AnDPro-matched	15.44 [9.93, 21.78]	28.94 [20.83, 37.32]	6.50 [0.00, 14.54]	8.85 [0.00, 20.31]

Arm	Passkey accuracy	Passkey F1	Book QA F1
FullKV	100.00 [100.00, 100.00]	79.38 [70.00, 88.75]	35.91 [28.44, 42.94]
ExpectedAttention	93.75 [84.38, 100.00]	91.88 [81.88, 100.00]	26.63 [19.18, 33.96]
SnapKV	100.00 [100.00, 100.00]	94.38 [86.88, 100.00]	24.78 [17.79, 32.20]
Start–recent	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	19.25 [13.10, 26.24]
TOVA score	100.00 [100.00, 100.00]	94.38 [86.88, 100.00]	28.06 [20.26, 36.00]
AnDPro-matched	9.38 [0.00, 21.88]	5.62 [0.00, 13.12]	24.19 [17.20, 31.51]

What selection costs. Our implementation prefills the document once for FULLKV and once per candidate, then runs one question pass per cache, all before decoding; each question pass covers only the question's tokens, a median of 13–27 per workload (Table 15). On book QA, prefilling and scoring FULLKV plus five candidates takes a median 23.71 s per question, 6.1× the 3.91 s of a single EXPECTEDATTENTION prefill and question pass. Sharing each prefill across a document's questions lowers the pool to 5.22 s per question, against 0.84 s for EXPECTEDATTENTION under the same sharing. Both figures exclude answer decoding, which runs on the selected 10% cache. The payoff is decode memory: on book QA the document cache drops from 11.21 GiB to 1.12 GiB per sequence (Table 13). Peak prefill memory stays at the FULLKV level, because selection needs the full-cache reference. For question and generated lengths T_q and T_g, the selected cache holds $(\lfloor 0.1N \rfloor + T_q + T_g)/(N + T_q + T_g)$ of the FULLKV tokens. Reference-trajectory features add sequential decoding steps with the full cache before the choice.

7 RELATED WORK

Building compressed caches. Eviction methods score tokens by accumulated attention (Zhang et al., 2023; Liu et al., 2023), an observation window at the end of the prompt (Li et al., 2024), last-token attention (Oren et al., 2024), attention expected under future queries (Devoto et al., 2025), key norms (Devoto et al., 2024), or the projection of each token's attention-weighted value onto the full attention output (Geng et al., 2025); StreamingLLM keeps attention sinks and a recent window (Xiao et al., 2024). Other methods allocate the budget across heads or layers (Feng et al., 2025; Cai et al., 2025; Wang et al., 2025), split heads into full and streaming caches (Xiao et al., 2025), or keep the full cache and load query-relevant pages (Tang et al., 2024). Each produces one cache per input; we take such caches as candidates and ask which one to keep.

Choosing among compression policies. FastGen profiles each attention head on the prompt and keeps the cheapest policy whose compressed attention map recovers the full map within a threshold (Ge et al., 2024)—an attention-reconstruction criterion of the kind our frozen-attention features measure. Concurrent work chooses one retention policy and budget per deployment from calibration data, with a finite-sample risk guarantee (Kang et al., 2026). We choose a cache per question with a learned scorer and show that output features beat attention and hidden-state features for that choice.

Output divergence and compression risk. Dutta et al. (2024) show that accuracy hides behavioral changes in quantized and pruned models and propose KL divergence and answer flips as metrics. For KV eviction, Xie & Alanwar (2026) certify attention-output error and find output log-probability the stronger failure predictor, and Haverbeck et al. (2026) characterize minimax compression risk through cache compressibility. We turn output divergence from an evaluation metric into a selection signal and measure its effect on generated answers at a fixed budget.

8 CONCLUSION

KV-cache compression has treated the compressor as the unit of progress. Our measurements point to a second lever at the same budget: the choice among compressed caches. At 10% retention, the per-question best of eleven caches matches or beats the uncompressed cache on three workloads, and on question answering, scoring caches by their effect on the next-token distribution captures more of that quality than the attention and hidden-state quantities compressors optimize. Output features win on question answering, transfer unchanged from 7–8B to 70B models and to books of up to 128K tokens, and the geometry of the unembedding explains why; state features win on literal retrieval, where richer reference trajectories close the gap.

Operating envelope. The experiments cover five checkpoints from three model families, a 10% document budget with a uniform per-head allocation, and the candidate pools of Section 3. Selectors learn from answers on training documents and, at inference, score candidates against a full-cache reference, so peak prefill memory matches FULLKV while decoding runs on the 10% cache.

Next steps. Every candidate is a subset of one full cache, so a single shared prefill can produce all masks; our measured implementation prefills each candidate separately, and sharing that prefill is the first target for cutting selection time. Cheaper output proxies, variable layer budgets as in PyramidKV (Cai et al., 2025), and selection repeated during long generations extend the same comparison to serving.

AI USE STATEMENT

We used generative AI tools to implement and debug experiment and analysis scripts, to give feedback on the experimental methodology, to check and correct mathematical derivations in the appendices, to identify related work and verify bibliography entries, and to draft and edit sections of the manuscript and documentation. We did not use generative AI to generate datasets; the benchmarks are existing public datasets. All empirical results come from executed model runs and statistical analysis of recorded outputs, and the released artifact regenerates every table and figure from those records. The authors reviewed and verified all AI-assisted code, derivations, citations, and text, and take full responsibility for the content of this paper.

ETHICS STATEMENT

The experiments use existing public benchmarks and open-weight language models. Cache compression can change answers in ways aggregate scores obscure. We release per-example predictions and describe each method's information access so these results are not mistaken for reliability guarantees in other settings.

REPRODUCIBILITY STATEMENT

The artifact provides pinned model and dataset revisions, frozen input manifests, source code, generated predictions, grouped folds, tuning scores, confidence-interval scripts, and all five-checkpoint analysis records. The README separates full GPU reproduction from analysis of released records. Large intermediate tensors can be regenerated; no simulated model outcomes enter the paper.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. PyramidKV: Dynamic KV cache compression based on pyramidal information funneling. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2406.02069.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of NAACL-HLT*, pp. 4599–4610, 2021. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

Alessio Devoto, Yu Zhao, Simone Scardapane, and Pasquale Minervini. A simple and effective L_2 norm-based strategy for KV cache compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18476–18499, 2024. doi: 10.18653/v1/2024.emnlp-main.1027. URL https://aclanthology.org/2024.emnlp-main.1027/.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: Kv cache compression by estimating attention from future queries distribution, 2025. URL https://arxiv.org/abs/2510.00636.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S. Kevin Zhou. Ada-KV: Optimizing KV cache eviction by adaptive budget allocation for efficient LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL

https://proceedings.neurips.cc/paper_files/paper/2025/hash/a40ff56daab9f4808b1e18350c8a11ce-Abstract-Conference.html.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Zijie Geng, Jie Wang, Ziqi Liu, Feng Ju, Yiming Li, Xing Li, Mingxuan Yuan, Jianye Hao, Defu Lian, Enhong Chen, and Feng Wu. Accurate KV cache eviction via anchor direction projection for efficient LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 10420–10456, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0f29157c7613e137834a607075c55e17-Abstract-Conference.html.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Lukas Haverbeck, Carmen Amo Alonso, Andres Felipe Posada-Moreno, Sebastian Trimpe, and Marco Pavone. The risk of KV cache compression, 2026. URL https://arxiv.org/abs/2607.01520.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Albert Q. Jiang et al. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Beomgu Kang, SoJin Yun, Hojoon Kim, and Hyunseok Seo. Risk-controlled KV-cache eviction: From memory budgets to risk targets, 2026. URL https://arxiv.org/abs/2609.27981.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for LLM KV cache compression at test time. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a452a7c6c463e4ae8fbdc614c6e983e6-Abstract-Conference.html.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. Transformers are multi-state rnns. In *Proceedings of EMNLP*, 2024. URL https://aclanthology.org/2024.emnlp-main.1043/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 47901–47911, 2024. URL https://proceedings.mlr.press/v235/tang24l.html.

Zihao Wang, Bin Cui, and Shaoduo Gan. SqueezeAttention: 2D management of KV-cache in LLM inference via layer-wise optimal budget. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3b0a8df568ec496a717566a7f8158aaa-Abstract-Conference.html.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5c1ddd2e59df46fd2aa85c833b1b36ed-Abstract-Conference.html.

Peng Xie and Amr Alanwar. Error certificates for KV-cache eviction via randomized design, 2026. URL https://arxiv.org/abs/2607.21475v3.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞Bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6ceefa7b15572587b78ecfcebb2827f8-Abstract-Conference.html.

A Candidate objectives and exact local identities

Budget and gathering. The budget $k = \lfloor 0.1N \rfloor$ is computed by integer truncation, which avoids the off-by-one losses of evaluating $\lfloor N(1-0.9) \rfloor$ in floating point. The budget applies to document tokens only. Retained keys and values are gathered into contiguous buffers that keep their original rotary positions. Implementations of Equation 1 clamp the denominator at 10^{-12}.

Probe layout. The reference trace applies each complete question to an untouched full document cache. For each layer and KV head, it stores attention from the last four available positions of the templated question input and every associated grouped-query head. Probes are ordered by question, then position, then query-head group. Values and keys are captured from the full document cache; question-prefix attention contributes the fixed background in Equation 1. This avoids renormalizing attention over document tokens alone.

The frozen mean-error and mean-mass features average all probes of the active question across layers and KV heads. The tail-error feature first averages the largest quarter of that question's probe errors within each layer/head, rounding the tail count up, then averages those layer/head values. Minimum mass is the minimum across all those probes, layers, and heads.

Scalar and vector candidates. Let p index probes, $r_{pi} = a_{pi}(v_i - o_p)$, and P_q the probes belonging to question q. The scalar score is

$$u_i = \frac{1}{|P|} \sum_{p \in P} \|r_{pi}\|_2. \tag{4}$$

The shared scalar arm retains its top-k tokens; the question-specific scalar arm restricts P to P_q. This score measures the magnitudes of individual contributions to attention error.

For a vector candidate, define $e_p(S)$ by Equation 1, $w_p(S) = c_p + \sum_{i \in S} a_{pi}$, and $v_p(S) = [0.2 - w_p(S)]_+$. Let $T(x)$ denote the mean of the largest $\max(1, \lceil 0.25|x| \rceil)$ components of a vector. The

three shared objectives are

$$J_{\mathrm{pool}}(S) = T((e_p(S))_{p\in P}) + 100T((v_p(S))_{p\in P}), \tag{5}$$

$$J_{\max}(S) = \max_q T((e_p(S))_{p\in P_q}) + 100 \max_q T((v_p(S))_{p\in P_q}), \tag{6}$$

$$J_{\mathrm{cap}}(S) = \max_q \frac{1}{|P_q|} \sum_{p\in P_q} \min\{e_p(S)/\tau_q, 1\} + 100 \max_q \frac{1}{|P_q|} \sum_{p\in P_q} v_p(S). \tag{7}$$

Here τ_q is the upper-tail error of the initial shared scalar mask on question q, clamped below at 10^{-6}. The arm named `question_shortfall_shared` implements J_{cap}, taking the maximum question-average capped relative error plus 100 times the maximum question-average mass shortfall. The question-specific vector arm minimizes the pooled objective using only P_q.

Restricted search. Each vector search starts from the corresponding scalar set. At an iteration, the code evaluates the objective after temporarily adding every outside token and retains the 128 best additions. It separately evaluates temporary removal of every retained token and retains the 64 best removals. It evaluates the Cartesian product of these pools as exact one-swaps, accepts an improving swap, and otherwise stops. At most ten swaps are performed. An improvement must exceed 10^{-8}. Candidate-pool ties follow the pinned PyTorch `topk` implementation; among the evaluated swaps, `argmin` chooses the first minimum in removal-major order. Selected indices are sorted before gathering. On singleton documents the shared vector objectives are represented by the common pooled-vector alias and deduplicated in physical-mask analyses.

Other masks. Mean attention retains the largest mean reference attention probabilities over the supplied question probes. Start–recent retains the first $\min(4, k)$ document tokens and fills the remainder with the most recent tokens. Random retention uses independently sampled scores for each layer/head, with seed $42 + 1009\ell$ at layer ℓ, then takes top-k. EXPECTEDATTENTION and SNAPKV use their released KVPress scorers[1] and record the exact gathered indices. The full-cache baseline retains all document tokens. All masks preserve the original position of each selected token.

Why the centered-response identity is exact. For one probe, the compressed response with frozen queries, keys, and values is

$$o_S = \frac{\sum_{i\in S} a_i v_i + \sum_{i\in B} a_i v_i}{w(S)}, \qquad w(S) = \sum_{i\in S} a_i + \sum_{i\in B} a_i. \tag{8}$$

Subtracting o gives

$$o_S - o = \frac{\sum_{i\in S} a_i (v_i - o) + \sum_{i\in B} a_i (v_i - o)}{w(S)}, \tag{9}$$

which proves Equation 1 for $w(S) > 0$. It also provides an independent numerical check: compute the renormalized response directly and compare its difference from o to the centered-residual calculation. This identity computes exact response error at the frozen attention head with queries, keys, and values held fixed.

Rank-fusion extension. For a source ranking's top-k tokens, zero-based rank r receives normalized score $(k - r)/k$; unlisted tokens receive zero. Linear fusion uses weights 0.25, 0.5, or 0.75 on ExpectedAttention and the complementary weight on SnapKV. Reciprocal-rank fusion sums $1/(60 + r)$ using one-based ranks of listed tokens. Max-rank fusion takes the larger normalized source score at each token. The top k fused scores define the mask, with token order breaking ties. Every configuration is evaluated through actual generation.

AnDPro-matched score. For observation position t and query head h, let $\widetilde{a}_{hti}$ be causal attention after width-seven max pooling. For each four-token chunk G_j, define $c_{htj} = \sum_{i\in G_j} \widetilde{a}_{hti} v_{hi}$. Let U exclude the first chunk and the final eight chunks. The anchor and chunk score are

$$a_{ht} = \sum_{j\in U} c_{htj}, \qquad s_{hj} = \sum_{t=1}^{32} c_{htj}^{\top} a_{ht}. \tag{10}$$

[1] `https://github.com/NVIDIA/kvpress`; the library asks users to cite Devoto et al. (2025).

We average these scores across the query heads associated with each KV head. The protected boundary chunks receive priority; remaining chunks are ranked by score, then expanded in token order until exactly k valid tokens are retained. Padding is removed, so the final retained chunk can be partial. Attention and weighted chunk construction follow the bfloat16 model computation; anchors and projections are accumulated in float32. These precision, grouped-query, and per-head-budget choices define the fixed-budget adaptation evaluated at the same per-head token budget.

TOVA score in this harness. We use KVPress's `TOVAPress` at prefill. It averages last-token attention across query heads and applies the resulting ranking to every KV head. The final document token receives the maximum score. This evaluates TOVA's last-token attention metric as a prefill eviction scorer under matched per-head retention budgets.

B Training, statistical analysis, and additional results

Training details. Within a question, every candidate has the same ordered feature schema. The code applies $\log(1+x)$ to nonnegative errors, KL, and relative ℓ_2 distances, then subtracts each feature's across-candidate mean and divides by its across-candidate standard deviation. A standard deviation below 10^{-12} is replaced by one. This transformation is part of the selector and uses only the candidate features available for that question. No answer-based feature is included. Training and selection retain the eleven named policies, including policy aliases that sometimes produce the same mask. Physical duplicates are collapsed for the reversal diagnostic only. The neural ranker selects the candidate with the largest sum of signed pairwise scores. The candidate list determines both the training pairs and this sum.

Let x_i be the transformed feature vector of candidate i and Y_i its generated F1. For unequal-quality pairs, the positive direction is $D = \operatorname{sign}(Y_i - Y_j)(x_i - x_j)$ and the weight is $|Y_i - Y_j|$, normalized to mean one over training pairs. The pairwise logistic loss is $n^{-1}\sum_{r=1}^{n} w_r \log(1+\exp(-D_r^\top \theta))$, where D stacks the n pair differences and w contains their weights. The linear optimizer runs 600 full-batch steps with a fixed learning rate $\min(1, 0.9/L)$, where $L = \lambda_{\max}(D^\top \operatorname{diag}(w) D)/(4n) + \lambda/F$ bounds the loss Hessian, and penalty $\lambda\|\theta\|_2^2/(2F)$ for feature count F. The inner grid is $\lambda \in \{0.01, 0.1, 1, 10\}$. The neural ranker uses $g(D) = w_2^\top \tanh(W_1^\top D)$ without biases, 1,200 full-batch steps, learning rate 0.1, widths $\{8, 16\}$, and penalties $\{0.001, 0.01\}$. Initialization uses seed zero. Hyperparameters are selected by inner-fold generated F1, not by pairwise training loss. Deterministic grid order resolves tuning ties. Candidates are ordered by method name; a selection tie chooses the first candidate in that order.

Document grouping. The nine excluded Qasper document ordinals are 0, 4, 11, 18, 30, 59, 75, 86, and 99. They are identified by stable source-row groups, then mapped to this manifest's ordering; the older release's ordinal ordering is not reused. These exclusions identify documents used in earlier development. Candidate and feature choices also draw on earlier benchmark work; nested validation separates subsequent fitting and tuning from held-out prediction. The remaining document identifiers are shuffled with the reported seed and partitioned into five outer folds. All questions and model measurements from an identifier stay together. Inner folds are formed only within the outer-training documents. The fixed-arm comparator maximizes training-document mean F1 using the same candidate family, then applies that arm to the outer-test questions.

Intervals and paired contrasts. For a quality difference, each document contributes its mean question difference under each fixed model. We average models and resample documents 20,000 times with replacement. Percentiles 2.5 and 97.5 give the interval. For reversal rates, each draw instead resamples the document-level counts of reversed and eligible pairs, forms a rate for each model, then averages model rates. The artifact includes the raw counts and weights for both calculations. Quality ties use tolerance 10^{-9} on the raw 0–1 F1 values; metric ties use 10^{-12}. Overlapping paired contrasts use the same resampling indices.

The bootstrap holds the fitted selectors and their held-out choices fixed. Its intervals describe document-sampling variation conditional on those fits. We report pointwise 95% intervals without adjustment for multiple comparisons. The twenty split seeds measure sensitivity to repartitioning the same documents. Model-family holdouts allow shared texts across training and test models; dataset holdouts separate the source and target workloads.

Table 6: Additional feature sets under the same nested validation. Reference-trajectory features add sequential computation beyond the first output distribution.

Features	Ranker	Qasper	2WikiMQA	RULER-16K
Frozen (4)	Linear	-3.29 [-7.70, 1.15]	$+4.47$ [0.95, 8.00]	$+44.20$ [41.94, 46.46]
Frozen (4)	MLP	-3.00 [-7.07, 1.11]	$+4.32$ [0.82, 7.82]	$+47.78$ [45.50, 50.02]
All + trajectory 4 (16)	Linear	$+3.45$ [0.66, 6.24]	$+6.37$ [3.28, 9.48]	$+49.46$ [47.16, 51.76]
All + trajectory 4 (16)	MLP	$+2.78$ [-0.00, 5.56]	$+7.33$ [4.09, 10.53]	$+49.74$ [47.35, 52.13]
All + trajectory 8 (16)	Linear	$+3.48$ [0.69, 6.26]	$+6.84$ [3.58, 10.07]	$+49.98$ [47.64, 52.30]
All + trajectory 8 (16)	MLP	$+3.35$ [0.62, 6.07]	$+7.50$ [4.27, 10.67]	$+50.98$ [48.49, 53.45]
All + trajectory 16 (16)	Linear	$+3.91$ [1.18, 6.65]	$+6.75$ [3.55, 9.94]	$+48.13$ [45.71, 50.50]
All + trajectory 16 (16)	MLP	$+3.54$ [0.82, 6.27]	$+7.56$ [4.23, 10.85]	$+50.32$ [47.96, 52.67]

Table 7: Paired change in generated F1 versus ExpectedAttention for every rank-fusion configuration on the primary-workload subsets. Each configuration uses the same strict token count; intervals resample paired documents.

Fusion	Qasper	2WikiMQA	RULER-16K
0.25	-3.55 [-8.69, 1.32]	-3.91 [-9.14, 1.04]	$+4.08$ [-0.51, 9.95]
0.5	-0.68 [-5.13, 3.87]	-0.66 [-5.22, 3.80]	$+3.58$ [-0.73, 9.33]
0.75	$+0.40$ [-3.03, 4.33]	$+0.02$ [-4.08, 3.95]	$+3.99$ [-0.78, 10.19]
rrf	-1.24 [-5.64, 3.37]	-0.59 [-4.84, 3.61]	$+4.06$ [-0.47, 9.87]
max	-1.33 [-5.80, 3.29]	-0.55 [-4.89, 3.64]	$+3.48$ [-1.04, 9.56]

Oracle reporting. Let Q_d be the questions of document d and let m index candidate policies, each of which builds a mask for (d, q). With deterministic decoding, $\max_m Y(d, q, m)$ is the exact finite-pool ceiling, and

$$\underbrace{\frac{1}{|Q_d|}\sum_{q\in Q_d}\max_m Y(d,q,m)}_{\text{one choice per question}} \;\geq\; \underbrace{\max_m \frac{1}{|Q_d|}\sum_{q\in Q_d} Y(d,q,m)}_{\text{one policy per document}}, \tag{11}$$

with equality when one policy is best for every question of the document. For question-specific arms, "one policy" means one eviction algorithm, not one physical cache. The shared-mask ceiling restricts the pool to the nine arms that build a single cache per document. Sibling-question transfer picks the policy that scores best on a document's other questions and applies it to the held-out question. The per-question oracle, best document-level policy, and sibling-question transfer values are all calculated from measured arm outcomes. The latter is defined only on multi-question documents, so its comparison ceilings are recomputed on that identical subset. Uniform policy choice is an analytic average of the nine policy outcomes. It differs from the seeded random-token arm, which is a real mask with an actual generated answer.

Table 8: Competitive-pool comparison counts after physical-mask deduplication. Ordered pairs exclude both quality and surrogate ties; the two tie counts can overlap.

Workload	Metric	Pairs	Ordered	Quality ties	Signal ties
Qasper	Mass	5,190	2,987	2,203	0
Qasper	Local error	5,190	2,987	2,203	0
Qasper	Next-token KL	5,190	2,987	2,203	1
2WikiMQA	Mass	6,000	1,485	4,515	0
2WikiMQA	Local error	6,000	1,485	4,515	0
2WikiMQA	Next-token KL	6,000	1,485	4,515	0
RULER-16K	Mass	12,000	5,014	6,986	0
RULER-16K	Local error	12,000	5,014	6,986	0
RULER-16K	Next-token KL	12,000	5,014	6,986	0

Table 9: Paired F1 difference: output-only selection minus state-only selection on identical outer-test questions. Positive values favor the two-feature output selector.

Ranker	Qasper	2WikiMQA	RULER-16K
Linear	$+4.64$ $[1.78, 7.58]$	$+3.70$ $[1.26, 6.18]$	-2.55 $[-4.34, -0.87]$
MLP	$+3.38$ $[0.71, 6.12]$	$+2.50$ $[0.19, 4.92]$	-4.58 $[-6.36, -2.88]$

Table 10: F1 change relative to ExpectedAttention under model-family and dataset holdout. Both ranker classes and both feature sets are reported, including negative transfer. Model holdout allows the same text in source-family training; dataset holdout does not.

Held out	Features	Ranker	Qasper	2WikiMQA	RULER-16K
model	state	Linear	-3.45 $[-7.58, 0.72]$	$+1.71$ $[-1.59, 5.00]$	$+41.30$ $[39.03, 43.59]$
model	state	MLP	-2.50 $[-6.59, 1.62]$	$+2.35$ $[-1.04, 5.67]$	$+41.53$ $[39.23, 43.85]$
model	output	Linear	$+2.22$ $[-0.77, 5.16]$	$+5.94$ $[2.72, 9.21]$	$+36.68$ $[34.42, 38.91]$
model	output	MLP	$+2.30$ $[-0.51, 5.03]$	$+5.72$ $[2.55, 8.93]$	$+36.06$ $[33.89, 38.25]$
dataset	state	Linear	-3.71 $[-7.88, 0.51]$	$+1.87$ $[-1.40, 5.12]$	$+42.16$ $[39.86, 44.45]$
dataset	state	MLP	-2.52 $[-6.66, 1.67]$	$+4.22$ $[0.69, 7.75]$	$+42.18$ $[40.19, 44.15]$
dataset	output	Linear	$+2.73$ $[-0.05, 5.51]$	$+6.32$ $[3.14, 9.52]$	$+37.96$ $[35.84, 40.07]$
dataset	output	MLP	$+2.62$ $[-0.19, 5.43]$	$+6.32$ $[3.14, 9.52]$	$+37.82$ $[35.68, 39.90]$

Table 11: Qasper partition sensitivity over 20 grouped split seeds. Positive counts refer to point estimates above ExpectedAttention; CI > 0 counts intervals entirely above zero.

Features	Ranker	Splits	Positive	CI > 0	Mean	Range
output	Linear	20	20	0	+2.72	[+2.47, +2.76]
output	MLP	20	20	0	+2.64	[+2.23, +2.77]
state	Linear	20	0	0	-1.91	[-2.78, -1.23]
state	MLP	20	7	0	-0.38	[-1.62, +0.14]

Table 12: Fixed-rule stress test: linear rankers trained on the three primary workloads, then applied to the five compressed long-context arms. Regularization is fixed at one; no long-context answer labels tune it. Values are F1 changes relative to ExpectedAttention.

Features	Passkey >100K	Book QA >32K
state	$+2.50$ $[-6.87, 13.12]$	$+3.83$ $[-0.10, 8.29]$
output	$+1.25$ $[-5.62, 9.38]$	$+6.68$ $[2.83, 11.11]$

Table 13: Instrumented B300 execution costs. Prefill and complete-question scoring time are medians over the recorded instances; peak allocated memory is the maximum and document-cache memory the median. Primary rows pool three models; long-context rows use Llama-3.1-8B only. Per-model measurements and generation timings are in the released artifact.

Workload	Arm	Prefill (s)	Scoring (s)	Peak (GiB)	KV (GiB)
Qasper	FullKV	0.16	0.076	19.54	0.66
Qasper	EA	0.25	0.064	17.32	0.07
Qasper	AnDPro-matched	0.39	0.035	17.32	0.07
2WikiMQA	FullKV	0.43	0.100	22.80	1.74
2WikiMQA	EA	0.55	0.062	18.88	0.17
2WikiMQA	AnDPro-matched	0.83	0.035	18.88	0.17
RULER-16K	FullKV	0.34	0.087	19.08	2.11
RULER-16K	EA	0.53	0.040	17.10	0.21
RULER-16K	AnDPro-matched	0.92	0.035	17.10	0.21
Passkey >100K	FullKV	4.57	0.291	44.21	15.30
Passkey >100K	EA	5.81	0.057	30.44	1.53
Passkey >100K	AnDPro-matched	8.22	0.056	30.44	1.53
Book QA >32K	FullKV	4.07	0.258	44.80	11.21
Book QA >32K	EA	3.84	0.076	30.76	1.12
Book QA >32K	AnDPro-matched	5.55	0.049	30.76	1.12

Table 14: Post-normalization sufficient condition for first-token preservation. Instances exclude zero reference margins; agreement is observed argmax agreement. A bound below one certifies exact next-token argmax agreement.

Workload	Instances	Certified	Agreement (%)	Minimum bound
2WikiMQA	1,880	0	47.1	8.04
Book QA >32K	540	0	58.3	16.34
Passkey >100K	145	0	47.6	117.87
Qasper	1,450	0	43.7	14.46
RULER-16K	3,790	0	36.3	24.72

Table 15: Cost accounting for the FullKV reference and five non-fusion compressed candidates. Values are medians of sums of measured prefill and complete-question scoring times. Cold cost assigns all six prefills to one question; amortized cost divides prefill by the number of questions on that document. EA cold uses one prefill and one scoring pass; with the same prefill sharing, EA's amortized cost is 0.21, 0.34, 0.18, 5.87, and 0.84 s for Qasper, 2WikiMQA, RULER-16K, passkey, and book QA. Decode time is excluded to isolate the prefill and scoring costs directly attributable to cache compression selection.

Workload	Question tokens	EA cold (s)	Pool cold (s)	Pool amortized (s)
2WikiMQA	27	0.62	3.51	1.94
Book QA >32K	24	3.91	23.71	5.22
Passkey >100K	13	5.87	33.16	33.16
Qasper	19	0.31	1.51	1.09
RULER-16K	24	0.58	3.07	0.99

Table 16: Document counts for the oracle comparisons. All three models share these documents. Sibling-question transfer and its matched oracle ceilings use only the final column.

Workload	All documents	Multi-question documents
Qasper	139	34
2WikiMQA	100	100
RULER-16K	100	100

Table 17: RULER target-presence evaluation. The secondary answer-presence rate measures whether a generated response contains an accepted target number delimited by non-digits, measuring literal target retrieval alongside token-F1. Full per-arm results are released.

Arm	Token F1	Answer present (%)
FullKV	51.62 [49.92, 53.38]	99.17 [98.67, 99.58]
ExpectedAttention	2.80 [1.85, 3.90]	3.08 [2.00, 4.33]
SnapKV	5.25 [2.35, 8.66]	7.17 [3.00, 12.08]
Start–recent	4.30 [3.08, 5.59]	9.50 [6.83, 12.33]
Random tokens	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Mean attention	48.85 [46.65, 51.09]	61.08 [59.08, 63.08]
Shared scalar	46.79 [44.73, 48.87]	62.00 [60.33, 63.67]
Question-specific vector	46.00 [43.72, 48.31]	65.33 [63.42, 67.25]

Table 18: Primary context lengths and configured model limits. Exceedance uses the complete templated context, question, and requested generation allowance, not only the displayed document length. The frozen primary inputs are retained and their actual outputs are all reported.

Model	Workload	Max document tokens	Configured limit	Exceeding documents
Qwen	2WikiMQA	33,320	40,960	0
Qwen	Qasper	22,035	40,960	0
Qwen	RULER-16K	15,983	40,960	0
Llama	2WikiMQA	31,896	131,072	0
Llama	Qasper	21,293	131,072	0
Llama	RULER-16K	15,908	131,072	0
Mistral	2WikiMQA	36,447	32,768	2
Mistral	Qasper	24,282	32,768	0
Mistral	RULER-16K	17,275	32,768	0

Table 19: Length-defined sensitivity for each affected model/workload. Values are F1 changes versus ExpectedAttention using the frozen out-of-fold selections, before and after restricting evaluation to prompts within the configured limit. Each row reports one model and workload.

Model/workload	Features	Ranker	All inputs	Within configured limit
Mistral 2WikiMQA	state	Linear	−2.42 [−7.85, 2.86]	−2.47 [−7.90, 2.81]
Mistral 2WikiMQA	state	MLP	+2.03 [−3.92, 7.93]	+2.07 [−4.06, 8.04]
Mistral 2WikiMQA	output	Linear	+6.96 [1.91, 12.14]	+7.10 [1.87, 12.31]
Mistral 2WikiMQA	output	MLP	+6.85 [1.81, 12.01]	+6.99 [1.77, 12.17]
Mistral 2WikiMQA	all	Linear	+6.19 [1.28, 11.32]	+6.32 [1.22, 11.42]
Mistral 2WikiMQA	all	MLP	+7.03 [1.94, 12.41]	+7.17 [1.80, 12.54]

Table 20: Actual available reference-trajectory lengths. FullKV was replayed on every primary question to record token IDs and available distributions; decoded predictions and F1 were checked against the primary run. Counts here exclude the Qasper development documents and pool reference questions across the three models. Requested horizon includes the first distribution, which requires no continuation.

Workload	Requested	Median used	Range used	Stopped early
2WikiMQA	4	4	2–4	129/600
2WikiMQA	8	6	2–8	421/600
2WikiMQA	16	6	2–16	539/600
Qasper	4	4	2–4	63/519
Qasper	8	8	2–8	142/519
Qasper	16	14	2–16	281/519
RULER-16K	4	4	4–4	0/1200
RULER-16K	8	8	5–8	400/1200
RULER-16K	16	16	5–16	516/1200

Table 21: Change in F1 relative to a model-specific fixed policy selected on the same outer-training documents. The policy can depend on the known model family but cannot adapt its policy choice to the held-out question. All choices use training labels only. This is a more flexible fixed-policy control than one common policy across model families.

Features	Ranker	Qasper	2WikiMQA	RULER-16K
States (8)	Linear	−0.53 [−4.20, 3.12]	+0.14 [−2.12, 2.47]	−9.52 [−11.59, −7.37]
States (8)	MLP	+0.69 [−2.86, 4.13]	+1.30 [−0.63, 3.34]	−7.48 [−9.39, −5.51]
Output (2)	Linear	+4.11 [1.28, 6.94]	+3.84 [1.35, 6.47]	−12.08 [−14.27, −9.90]
Output (2)	MLP	+4.08 [1.25, 6.92]	+3.81 [1.35, 6.39]	−12.07 [−14.28, −9.87]
Frozen + output (6)	Linear	+4.40 [1.57, 7.16]	+3.06 [0.98, 5.19]	−9.83 [−11.51, −8.19]
Frozen + output (6)	MLP	+4.60 [1.82, 7.36]	+3.34 [1.28, 5.45]	−2.58 [−3.76, −1.37]
All (14)	Linear	+4.50 [1.72, 7.28]	+3.85 [1.59, 6.13]	−9.65 [−11.55, −7.72]
All (14)	MLP	+4.35 [1.59, 7.11]	+4.19 [1.92, 6.53]	−1.59 [−3.00, −0.16]
All + trajectory 4 (16)	Linear	+4.81 [2.07, 7.52]	+3.59 [1.37, 5.87]	−1.11 [−2.48, 0.30]
All + trajectory 4 (16)	MLP	+4.14 [1.45, 6.82]	+4.56 [2.10, 7.10]	−0.83 [−2.33, 0.66]
All + trajectory 8 (16)	Linear	+4.83 [2.15, 7.49]	+4.07 [1.97, 6.23]	−0.60 [−1.98, 0.83]
All + trajectory 8 (16)	MLP	+4.71 [2.05, 7.30]	+4.72 [2.44, 7.07]	+0.40 [−1.08, 1.92]
All + trajectory 16 (16)	Linear	+5.26 [2.63, 7.87]	+3.98 [1.91, 6.12]	−2.44 [−4.14, −0.68]
All + trajectory 16 (16)	MLP	+4.89 [2.26, 7.52]	+4.79 [2.52, 7.10]	−0.25 [−1.92, 1.52]

Table 22: Paired incremental F1 from adding reference-trajectory features to the 14 first-pass features. Each difference uses the identical outer-test questions and ranker class, isolating the incremental performance contribution of multi-step reference trajectories beyond first-pass features.

Requested horizon	Ranker	Qasper	2WikiMQA	RULER-16K
4	Linear	+0.30 [−0.52, 1.14]	−0.26 [−1.01, 0.44]	+8.54 [7.01, 10.08]
4	MLP	−0.22 [−0.98, 0.52]	+0.37 [−0.90, 1.66]	+0.76 [−0.30, 1.82]
8	Linear	+0.33 [−0.66, 1.33]	+0.22 [−0.65, 1.09]	+9.05 [7.48, 10.67]
8	MLP	+0.35 [−0.53, 1.28]	+0.53 [−0.45, 1.50]	+1.99 [0.80, 3.24]
16	Linear	+0.76 [−0.22, 1.76]	+0.13 [−0.80, 1.03]	+7.21 [5.70, 8.79]
16	MLP	+0.54 [−0.46, 1.58]	+0.60 [−0.51, 1.68]	+1.34 [0.02, 2.71]

Table 23: Held-out F1 differences between training-selected fixed policies and ExpectedAttention. The shared fixed policy uses one rule across models; the model-specific fixed policy chooses a rule for each model.

Training-selected control	Qasper	2WikiMQA	RULER-16K
Common policy	+0.00 [0.00, 0.00]	+4.47 [0.98, 8.01]	+47.82 [45.50, 50.07]
Model-specific policy	−1.36 [−2.53, −0.18]	+2.77 [−0.62, 6.20]	+50.57 [48.25, 52.90]

Table 24: Absolute F1 of all twelve primary arms, using the full evaluation documents and the same document/model weighting as the selection results. Qasper development documents are excluded. These scores provide the baseline scale for the reported gains; they are distinct from the smaller baseline-extension subset.

Arm	Qasper	2WikiMQA	RULER-16K
FullKV	42.64 [37.85, 47.47]	41.22 [36.10, 46.45]	51.62 [49.92, 53.38]
ExpectedAttention	27.20 [23.58, 31.05]	30.28 [25.83, 34.86]	2.80 [1.85, 3.90]
SnapKV	18.44 [14.76, 22.46]	26.73 [22.02, 31.52]	5.25 [2.35, 8.66]
Start–recent	15.95 [12.88, 19.21]	22.90 [18.88, 27.03]	4.30 [3.08, 5.59]
Random tokens	1.86 [1.34, 2.51]	4.87 [3.71, 6.15]	0.00 [0.00, 0.00]
Mean attention	23.02 [19.47, 26.79]	33.72 [29.10, 38.36]	48.85 [46.65, 51.09]
Shared scalar	23.44 [20.02, 27.03]	35.14 [30.50, 39.93]	46.79 [44.73, 48.87]
Shared pooled vector	24.18 [20.65, 27.88]	34.92 [30.39, 39.49]	43.87 [41.84, 45.94]
Shared max-question vector	24.20 [20.68, 27.91]	35.36 [30.81, 40.03]	43.89 [41.81, 46.01]
Shared capped-error vector	23.90 [20.37, 27.58]	34.59 [30.15, 39.15]	43.84 [41.70, 45.96]
Question-specific scalar	23.82 [20.43, 27.37]	35.29 [30.66, 40.07]	50.61 [48.31, 52.87]
Question-specific vector	24.45 [20.89, 28.12]	35.07 [30.50, 39.79]	46.00 [43.72, 48.31]

Table 25: Unchanged outer-test primary selector choices versus the TOVA prefill-score baseline on the same preselected extension documents. Entries are paired F1 differences (selector minus TOVA). This comparison was added after the TOVA outcomes were available; the saved selector choices remain fixed.

Primary features	Ranker	Qasper	2WikiMQA	RULER-16K
States (8)	Linear	$+9.37$ [4.22, 14.58]	$+7.23$ [1.98, 12.42]	-4.84 [-8.35, -1.07]
States (8)	MLP	$+9.55$ [4.42, 14.95]	$+7.66$ [1.44, 13.79]	-2.71 [-6.05, 0.86]
Output (2)	Linear	$+14.15$ [9.09, 19.36]	$+10.98$ [5.70, 16.36]	-9.08 [-13.37, -4.92]
Output (2)	MLP	$+14.03$ [9.00, 19.18]	$+10.98$ [5.75, 16.34]	-9.04 [-13.20, -4.98]
All (14)	Linear	$+14.64$ [9.52, 19.93]	$+10.73$ [5.38, 16.33]	-5.45 [-8.48, -2.25]
All (14)	MLP	$+14.51$ [9.46, 19.72]	$+10.23$ [5.10, 15.65]	$+2.53$ [-0.61, 5.64]
All + trajectory 4 (16)	Linear	$+14.65$ [9.68, 19.83]	$+10.06$ [4.41, 15.82]	$+3.00$ [-0.56, 6.68]
All + trajectory 4 (16)	MLP	$+12.66$ [7.22, 17.97]	$+11.49$ [6.69, 16.53]	$+3.13$ [-0.80, 7.03]
All + trajectory 8 (16)	Linear	$+14.25$ [8.79, 19.91]	$+11.56$ [5.45, 17.47]	$+4.46$ [0.74, 8.30]
All + trajectory 8 (16)	MLP	$+13.37$ [7.65, 18.98]	$+11.89$ [6.97, 17.11]	$+4.40$ [0.25, 8.49]
All + trajectory 16 (16)	Linear	$+14.66$ [9.51, 19.99]	$+11.21$ [5.28, 16.97]	$+1.86$ [-2.23, 6.20]
All + trajectory 16 (16)	MLP	$+13.89$ [8.15, 19.48]	$+11.41$ [6.23, 16.86]	$+4.88$ [0.82, 9.18]

C A SUFFICIENT CONDITION FOR NEXT-TOKEN AGREEMENT

A bound on the final hidden-state change gives a sufficient condition for preserving the highest-probability next token. Write logits as $z = Wh + b$, where h denotes the *post-normalization* final hidden state. Define $M = \max_j \|W_{j,:}\|_2$ and let $\Delta = z^F_{(1)} - z^F_{(2)} > 0$ be the full-cache top-two logit margin. Then

$$\|z^S - z^F\|_\infty \le M\|h^S - h^F\|_2. \tag{12}$$

If $2M\|h^S - h^F\|_2 < \Delta$, the original top logit strictly dominates all competitors: the leading coordinate can decrease by at most the bound while competing logits can increase by at most the bound, certifying exact argmax preservation at the measured token prefix.

Using a pre-normalization state x requires an additional factor. For RMSNorm $f(x) = \mathrm{diag}(g)x/\sqrt{\|x\|_2^2/d + \epsilon}$, a valid pathwise bound between two states is

$$L \le \frac{\|g\|_\infty}{\min\limits_{t\in[0,1]} \sqrt{\|x^F + t(x^S - x^F)\|_2^2/d + \epsilon}}. \tag{13}$$

The Jacobian of $x/\sigma(x)$ is $I/\sigma - xx^\top/(d\sigma^3)$; its operator norm is bounded by $1/\sigma$. Integrating along the segment $[x^F, x^S]$ and multiplying by the diagonal gain yields the pathwise bound. For LayerNorm, centering states forms an orthogonal projection with norm at most one, yielding an analogous pathwise integral.

In our evaluation, post-normalization hidden states are recorded directly across all model families. The artifact contains the resulting bound, full-cache margin, empirical top-token agreement, and generated quality for each arm. Across 7,805 compressed-candidate cases with positive reference margins, the condition certifies 0 (Table 14); the smallest bound is 8.04, against the value below one that certification requires. The observed highest-probability token agrees in 3,276 cases (41.97%).

D WHAT OUTPUT KL RETAINS ABOUT A STATE PERTURBATION

Let $p_i = e^{z_i}/\sum_j e^{z_j}$ and $q_i = e^{z_i+\delta_i}/\sum_j e^{z_j+\delta_j}$. Direct substitution gives

$$D_{\mathrm{KL}}(p\|q) = \sum_i p_i \left[-\delta_i + \log \frac{\sum_j e^{z_j+\delta_j}}{\sum_j e^{z_j}}\right] \tag{14}$$

$$= \log \sum_i p_i e^{\delta_i} - \sum_i p_i \delta_i. \tag{15}$$

Expanding the log moment-generating function at zero yields

$$D_{\mathrm{KL}}(p\|q) = \frac{1}{2}\delta^\top(\mathrm{diag}(p) - pp^\top)\delta + O(\|\delta\|_\infty^3). \tag{16}$$

For post-normalization states u and $u + \Delta u$ with affine vocabulary projection $z = Wu + b$, the logit change is exactly $\delta = W\Delta u$. Thus, locally,

$$D_{\mathrm{KL}}(p\|q) = \frac{1}{2}\Delta u^\top W^\top(\mathrm{diag}(p) - pp^\top)W\Delta u + O(\|W\Delta u\|_\infty^3). \tag{17}$$

The matrix $W^\top(\mathrm{diag}(p) - pp^\top)W$ weights each direction of state change by its local effect on token probabilities. Two changes with equal Euclidean norm can therefore have different KL divergence. Directions with small curvature contribute little to the local quadratic term. If $W\Delta u$ is constant across vocabulary entries, softmax is exactly unchanged for any magnitude of that shift.

E LARGER-CHECKPOINT PROTOCOL AND CHECKS

Table 4 reports the six complete model–workload cells. The two checkpoints are Qwen/Qwen3-32B at revision `9216db5781bf21249d130ec9da846c4624c16137` and meta-llama/Llama-3.1-70B-Instruct at revision `1605565b47bb9346c5515c34102e054115b4f98b`. Both run in

BF16 with SDPA. Qwen's chat template has thinking disabled. The frozen protocol takes the first 16 eligible documents in source-manifest order after excluding all nine Qasper development documents. Every complete prompt and generation allowance must fit both models' native limits; no input is shortened to fit.

Each selected document retains its original context identity and original outer-test fold. The applied eight-state-feature and two-output-feature linear rankers were fitted only on that fold's 7–8B training documents. The fixed arm is chosen from those same training documents. A released mapping records the fold, weight files, hashes, training-document sets, candidate order, and tie rule for every selected document. Replaying these saved rankers on the corresponding original 7–8B records exactly reproduces all 708 held-out choices and fixed-policy assignments. No 32B or 70B label enters a fit or policy choice.

Each model runs 22 Qasper, 32 packed 2WikiMQA, and 64 RULER questions. Every question has actual FullKV and eleven compressed generations, giving 1,416 outcomes per model. All compressed masks have exactly $\lfloor 0.1N \rfloor$ unique, in-range document positions per layer and KV head. Validation checks the physical index tensors, context hashes, question identity, prediction records, finite features, and all 96 model–document jobs. Reference-trajectory features were disabled and saved as unavailable, not as measured zeros. For each model and workload, we average questions within document and compute 20,000 paired bootstrap resamples of the 16 documents with seed 42.

The comparison with 8B checkpoints reuses their primary outcomes on these same texts. The combined artifact provides the full protocol, selected manifests, flat records, per-question choices, index-check results, validation hashes, and execution receipts under `large_models/`. The complete execution procedure and model-access requirements are documented in `large_models/README.md` and `large_models/docs/LARGE_MODEL_PROTOCOL.md`.

F Artifact and execution guide

The combined artifact root contains `src/`, `scripts/`, `data/manifests/`, `data/configs/`, `results/`, and `paper/`. `README.md` gives the expected output counts, and `REPRODUCIBILITY.md` lists the exact entry points and commands. Model weights remain externally distributed under their original licenses; the revision identifiers and download commands are included. The anonymous package excludes local account paths, private review screenshots, Git credentials, and model-weight copies. Its measurements preserve provenance through content hashes and relative artifact paths.

The primary execution proceeds in three stages per model and dataset: full-cache trace capture, scalar/vector mask construction, and actual greedy generation plus feature measurement. Eight workers partition document ordinals. A resumed run reuses outputs with matching input and configuration hashes. Historical measurements are stored separately. The collection script checks 1,044 model–document runs, 28,800 rows, twelve arms per question, unique record keys, finite numeric features, and exact index counts. Final aggregation requires these checks to pass. To overlap CPU and GPU work, selection fits may start once every record of one complete workload is available; the final stage reuses them only after checking hashes of both their inputs and analysis code.

Extension execution loads the same model revisions and runs its frozen manifest on every declared arm. The released measurements separate prefill and selection, question forward passes, generation, and cache storage. Token counts refer to the actual tokenizer and model input. InfiniteBench eligibility is checked before generation; eligible examples retain their full input.

The statistical scripts write raw out-of-fold choices, tuning scores, aggregate tables, and vector figures. A manuscript-generation script reads these outputs to produce numerical macros and table sources. Rebuilding the PDF is the final stage of reproduction. The revision ledger `docs/REVISION_CHANGES.md` records exact before/after source changes and the reason for each conceptual, numerical, and methodological correction. Large intermediate traces and mask tensors can be regenerated with the provided commands; compact raw predictions and per-mask hashes support independent analysis without those intermediates.

Generation caps are 148 tokens for Qasper, 52 for packed 2WikiMQA, 32 for RULER, 16 for passkey retrieval, and 40 for book QA. The book-QA cap matches the pinned InfiniteBench helper; the

passkey cap expands beyond that helper's six-token default to allow complete integer capture. All candidates for an example use the same frozen prompt, generation cap, and scoring rule.

The 32B/70B confirmation has a separate complete source and record bundle under `large_models/`. Its frozen manifests, held-out-fold ranker mapping, actual predictions, execution receipts, physical-mask checks, and compact reanalysis are included there. The root paper and its six-row table use the validated summary in `large_models/results/analysis/large_model_summary.csv`. The original 7–8B records remain at the root. The larger-checkpoint analysis uses its own summaries and the selected original texts described in Appendix E.