# How Does KV-Cache Eviction Shape Output Divergence?

**Anonymous authors**
Paper under double-blind review

## Abstract

KV-cache eviction reduces inference memory while altering subsequent generation. Across the interventions studied, lower cumulative token mismatch with full-cache continuations is associated mainly with later and less frequent first divergence; exposure-weighted post-divergence mismatch rates are also lower but remain high. Under a specified stepwise maximal coupling, we decompose the expected mismatch fraction into a first-mismatch term plus post-divergence exposure multiplied by its mismatch rate. Across an exploratory cohort and an independent 90%-retention cohort of two language models, exposure accounts for approximately 85–92% of six pooled mismatch-gap point estimates, including matched-byte selector comparisons at 50% and 90% retention. In the 90% cohort, prespecified comparisons within fixed early-divergence cohorts show higher total variation in late than early post-divergence windows for both models and selectors. We characterize the sharp range of cumulative mismatch compatible with an exact profile of finite divergence-aligned observations over unrestricted autoregressive kernel pairs and construct unbiased estimators via conditional Monte Carlo.

## 1 Introduction

KV-cache eviction reduces the memory needed for language-model inference by retaining selected past states. Attention, retrieval structure, and forward-looking signals guide this selection (Li et al., 2024; Fu et al., 2025; Tang et al., 2025; Kai et al., 2026; Ahn et al., 2026). The resulting distributional changes unfold along the continuation. Lower cumulative mismatch with a full-cache continuation can arise from delaying or avoiding the first divergence, reducing mismatch afterward, or both. Distinguishing these contributions is necessary to understand what an eviction policy preserves over time.

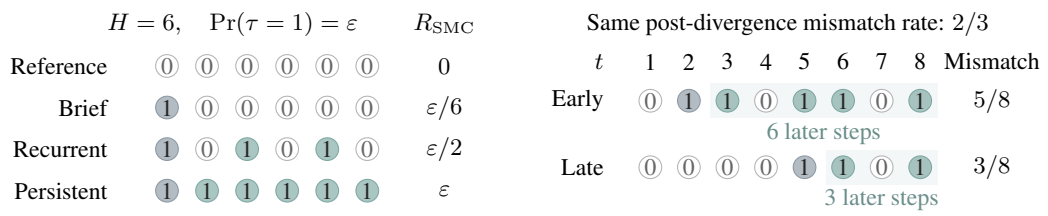
A binary example isolates the distinction. Let a reference always emit zero. Two interventions emit a first-position one with probability $\varepsilon$. Conditional on that departure, they continue with zeros or ones, respectively; otherwise both continue with zeros. Both have the same first-divergence law and sequence total variation (TV), $\varepsilon$, but their expected mismatch fractions are $\varepsilon/H$ and $\varepsilon$. The first departure therefore leaves the subsequent occupation of disagreement undetermined (Figure 1). This distinction connects our question to error accumulation and self-recovery in autoregressive models (He et al., 2021; Arora et al., 2022). We formalize the trajectory comparison using recursive overlap–residual maximal coupling (Goldstein, 1979; Levin & Peres, 2017). Reference kernels $P$ and intervention kernels $Q^a$ generate paired continuations, each with its original marginal law. Their cumulative positional disagreement is

$$R_{\mathrm{SMC}}(a) = \mathbb{E}\left[\frac{1}{H}\sum_{t=1}^{H}\mathbf{1}\{X_t \neq Y_t^a\}\right]. \tag{1}$$

Our decomposition separates the first mismatch from post-divergence exposure multiplied by its mismatch rate. Their relative contributions are empirical. We study two language models in an exploratory cohort and an independent cohort at 90% retention. Across six pooled contrasts, exposure accounts for approximately 85–92% of the mismatch gaps in descriptive decompositions. The comparisons include matched-byte SnapKV versus recent-token selection at 50% and 90% retention, plus SnapKV-512 versus SnapKV-half at different budgets. Smaller cumulative mismatch accompanies fewer and later departures, with a smaller improvement in a still-high post-divergence mismatch rate.



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.




(a) Same entry, different mismatch

(b) Later entry, unchanged mismatch rate

**Figure 1:** Lower cumulative mismatch need not reflect a lower post-divergence mismatch rate. Constructed binary continuations are compared with an all-zero reference; gray marks first divergence at $\tau$, and shading marks later positions. (a) Each displayed intervention path occurs with probability $\varepsilon$; otherwise all tokens are zero. The right column gives expected mismatch fractions. (b) Both paths have post-divergence mismatch rate $2/3$ ($4/6$ and $2/3$), yet later divergence reduces total mismatch from $5/8$ to $3/8$.

The temporal profiles reveal a further distinction. Pooled gaps between interventions are smaller in late than early generation windows, while average cross-history TV increases from early to late windows within fixed early-divergence cohorts. The independent 90% study supports this post-divergence pattern in prespecified comparisons. Generation-position averages and divergence-aligned averages therefore answer different questions about the same continuation law.

The theory also characterizes what finite observations determine about the cumulative outcome. Over unrestricted autoregressive kernel pairs, every risk in $[R_L, R_L + B_L]$ is compatible with an exact profile of the divergence law and the first $L$ mismatch marginals of each branch. Here $R_L$ is the observed contribution and $B_L$ the remaining probability-weighted trajectory capacity. A residual-branch conditional Monte Carlo construction supplies unbiased measurements of the cumulative loss and its components. Ordinary coupled rollouts used in the empirical studies target this same law. This framework separates uncertainty about an unobserved suffix from where disagreement accumulates.

## 2 Related Work

KV-cache compression uses attention sinks, heavy hitters, retrieval heads, observation windows, and heterogeneous budgets to select retained state (Xiao et al., 2024; Zhang et al., 2023; Li et al., 2024; Fu et al., 2025; Feng et al., 2025; Qin et al., 2025). Quantization and low-rank projection offer complementary compression axes (Liu et al., 2024; Chang et al., 2025); serving evaluations study their quality–memory–latency trade-offs (Gao et al., 2025). InfoKV evaluates the influence of KV removal on future predictive distributions along a fixed token sequence, while LookaheadKV predicts future attention without generating the continuation (Kai et al., 2026; Ahn et al., 2026). Worst-case space bounds characterize compression limits under general attention inputs (Haris & Onak, 2025).

Feedback, error accumulation, and self-recovery have been studied in sequential prediction and language generation (Ross et al., 2011; Bengio et al., 2015; He et al., 2021; Arora et al., 2022). Sequence TV estimation (Price et al., 2026) measures distance between marginal sequence laws; maximal agreement couplings (Völlering, 2016; Ernst et al., 2019) maximize the probability of agreement through each time. One-step maximal coupling realizes TV as mismatch probability (Goldstein, 1979); token couplings also support speculative decoding (Leviathan et al., 2023; Sun et al., 2023) and control generation randomness in LLM evaluation (Benz et al., 2026).

Partial identification characterizes the target values compatible with available observations (Manski, 2003; Stoye, 2007; Khan et al., 2024). Conditional Monte Carlo reduces simulation variance by integrating out components of the sampling randomness (Blackwell, 1947; Casella & Robert, 1996; Glasserman & Staum, 2001). The present study examines how first-divergence timing and subsequent mismatch jointly determine cumulative discrepancy under persistent KV-cache eviction.

## 3 From Divergence Timing to Cumulative Disagreement

We decompose cumulative mismatch into entry and post-entry contributions, then characterize the risks compatible with a finite branch window.

### 3.1 DIVERGENCE TIMING, EXPOSURE, AND OCCUPATION

Let $\mathcal{V}$ be a finite token alphabet (the model vocabulary), and let $H \geq 1$ be an integer generation horizon. Fix a prompt and an intervention $a$. For $t \in \{1, \ldots, H\}$ and histories $x_{<t}, y_{<t} \in \mathcal{V}^{t-1}$, write

$$p = P_t(\cdot \mid x_{<t}), \qquad q = Q_t^a(\cdot \mid y_{<t}),$$
$$c(z) = \min\{p(z), q(z)\}, \quad e = 1 - \sum_{z \in \mathcal{V}} c(z) = \mathrm{TV}(p, q).$$

With probability $1 - e$, draw a common token from $c/(1-e)$. With probability $e$, draw the two tokens independently from the disjoint residual laws $(p-c)/e$ and $(q-c)/e$. This is the standard overlap–residual maximal coupling (Levin & Peres, 2017, Proposition 4.7), applied recursively at each pair of histories. A normalized law on a zero-probability branch can be chosen arbitrarily.

The sample loss is normalized Hamming loss, $H^{-1}\sum_{t=1}^{H} I_t$, where $I_t = \mathbf{1}\{X_t \neq Y_t\}$. For histories $x, y$ of length $t-1$, define $\delta_t(x, y) = \mathrm{TV}(P_t(\cdot \mid x), Q_t^a(\cdot \mid y))$. One-step maximality gives the first identity below; iterated expectation (Levin & Peres, 2017, Appendix A.2) gives the second

$$\mathbb{E}[I_t \mid X_{<t}, Y_{<t}] = \delta_t(X_{<t}, Y_{<t}), \qquad R_{\mathrm{SMC}} = \frac{1}{H}\sum_{t=1}^{H} \mathbb{E}[\delta_t(X_{<t}, Y_{<t})]. \tag{2}$$

Thus the expectation integrates local TV along the actual paired histories. On a common prefix it reduces to $e_t(h) = \delta_t(h, h)$; after divergence both histories enter the next-token comparison. This identity connects a token-level loss to the recursively specified model distributions.

To separate entry from subsequent disagreement, put $\tau = \inf\{t \leq H : I_t = 1\}$, with $\tau = \infty$ for full agreement, and $p_s = \mathbb{P}(\tau = s)$. On positive-mass branches let $\kappa_s = \mathbb{E}[H^{-1}\sum_t I_t \mid \tau = s]$, assigning zero otherwise. The loss is zero on $\{\tau = \infty\}$. Applying total expectation to the first-divergence partition gives

$$R_{\mathrm{SMC}} = \sum_{s=1}^{H} p_s \kappa_s, \qquad \frac{1}{H} \leq \kappa_s \leq \frac{H - s + 1}{H} \quad (p_s > 0). \tag{3}$$

On $\{\tau = s\}$, $I_t = 0$ for $t < s$ and $I_s = 1$, which gives the displayed bounds. The first-mismatch contribution depends only on the divergence law; later mismatches contribute post-divergence occupation, giving

$$O_H = \frac{1}{H}\sum_{s=1}^{H} p_s, \qquad \Pi_H = \frac{1}{H}\mathbb{E}\left[\sum_{t>\tau} I_t\right], \qquad R_{\mathrm{SMC}} = O_H + \Pi_H. \tag{4}$$

The occupation term further separates the amount of time after divergence from the mismatch rate during that time. Write

$$E_H = \frac{1}{H}\sum_{s} p_s(H - s), \qquad C_H = \frac{\mathbb{E}[\sum_{t>\tau} I_t]}{H E_H}, \qquad R_{\mathrm{SMC}} = O_H + E_H C_H, \tag{5}$$

with $C_H = 0$ if $E_H = 0$. $E_H$ is the expected fraction of positions strictly after divergence; it changes with both divergence probability and timing. $C_H$ is the mismatch rate weighted by this exposure, rather than an average of per-trajectory rates. For two actions, let $\Delta$ denote their difference and bars their arithmetic mean. Then

$$\Delta R_{\mathrm{SMC}} = \Delta O_H + \underbrace{\Delta E_H\, \overline{C}_H}_{\text{exposure term}} + \underbrace{\Delta C_H\, \overline{E}_H}_{\text{rate term}}. \tag{6}$$

Either term may dominate. Under persistent eviction, occupation can reflect both retained-state changes and history feedback; the decomposition describes their temporal outcome. Appendix A.1 proves the identities, and Appendix A.3 gives their interpretation.

### 3.2 RISKS COMPATIBLE WITH A FINITE BRANCH WINDOW

The divergence law fixes the available exposure; a finite branch window also reveals part of the subsequent occupation. Partial identification (Manski, 2003) characterizes the cumulative risks consistent with these observations. Fix $1 \leq L \leq H$, put $\ell_s = \min\{L, H - s + 1\}$, and define

$$\rho_{s,j} = \mathbb{P}(I_{s+j} = 1 \mid \tau = s), \qquad 0 \leq j < \ell_s,$$

on positive-mass branches, with value zero when $p_s = 0$. The first divergence is included, so $\rho_{s,0} = 1$ when $p_s > 0$. The window contribution and residual capacity are

$$R_L(p, \rho) = \frac{1}{H} \sum_s p_s \sum_{j=0}^{\ell_s - 1} \rho_{s,j}, \qquad B_L(p) = \frac{1}{H} \sum_s p_s (H - s - L + 1)_+. \tag{7}$$

Thus $R_1 = O_H$, $R_H = R_{\mathrm{SMC}}$, and $B_H = 0$.

**Definition 3.1** (Admissible depth-$L$ profile)**.** Write $\mathcal{S}_L(P, Q) = (p, \rho)$. A profile is admissible when $p_s \geq 0$, $\sum_s p_s \leq 1$, $0 \leq \rho_{s,j} \leq 1$ for $0 \leq j < \ell_s$, $\rho_{s,0} = 1$ on positive-mass branches, and $\rho_{s,j} = 0$ on zero-mass branches.

**Theorem 3.2** (Sharp identified risk interval)**.** *For every admissible profile on an alphabet with at least two symbols, allowing both $P$ and $Q$ to vary over all autoregressive kernels subject to that profile, the compatible full-horizon risks are exactly*

$$\{R_{\mathrm{SMC}}(P, Q) : \mathcal{S}_L(P, Q) = (p, \rho)\} = [R_L(p, \rho), R_L(p, \rho) + B_L(p)]. \tag{8}$$

*Proof of Theorem 3.2.* For every compatible pair of kernels,

$$0 \leq R_{\mathrm{SMC}} - R_L = \frac{1}{H} \sum_{s : p_s > 0} p_s \sum_{j=\ell_s}^{H-s} \mathbb{E}[I_{s+j} \mid \tau = s] \leq \frac{1}{H} \sum_s p_s (H - s + 1 - \ell_s) = B_L.$$

To attain every point, fix $\alpha \in [0, 1]$ and symbols $0, 1$. Set $P_t = \delta_0$, $\bar{S}_t = 1 - \sum_{j<t} p_j$, and $\eta_t = p_t / \bar{S}_t$ when $\bar{S}_t > 0$, with $\eta_t = 0$ otherwise. Admissibility implies $0 \leq \eta_t \leq 1$. For a binary history let $s(y) = \min\{j : y_j = 1\}$, with $\min \varnothing = \infty$, and define

$$Q_t^\alpha(1 \mid y_{<t}) = \begin{cases} \eta_t, & s(y) = \infty, \\ \rho_{s,t-s}, & s = s(y) < t,\ t - s < \ell_s, \\ \alpha, & s = s(y) < t,\ t - s \geq \ell_s. \end{cases} \qquad Q_t^\alpha(0 \mid y_{<t}) = 1 - Q_t^\alpha(1 \mid y_{<t}).$$

Other symbols have probability zero and kernels on nonbinary histories are arbitrary. Since $X_t = 0$ and $I_t = \mathbf{1}\{Y_t = 1\}$ almost surely, $\bar{S}_t(1 - \eta_t) = \bar{S}_{t+1}$ yields $\mathbb{P}(\tau \geq t) = \bar{S}_t$ and $\mathbb{P}(\tau = t) = \bar{S}_t \eta_t = p_t$. On $\{\tau = s\}$, the first 1 stays at position $s$ in every subsequent history. Conditional mismatch probabilities therefore equal $\rho_{s,j}$ in the observed window and $\alpha$ afterwards. Thus $\mathcal{S}_L(P, Q^\alpha) = (p, \rho)$ and $R_{\mathrm{SMC}}(P, Q^\alpha) = R_L + \alpha B_L$, attaining the whole interval, including when $B_L = 0$. □

Sharpness allows both kernels to vary; fixing the reference may narrow the compatible set (Appendix A.2). The profile includes the divergence law over all $H$ positions and up to $L$ observations *from each divergence*; it is not a prefix of the first $L$ generation steps. Its width $B_L$ measures incomplete observation even with an exact profile. Estimating that profile adds sampling error (Appendix A.5). An intervention's actual tail, $R_H - R_L$, can be smaller than this available capacity.

## 4 Measuring Divergence-Aligned Occupation

The two main trajectory studies use ordinary coupled rollouts. For the supplementary depth measurements, Survival-Weighted Residual Branching (SWRB) conditions on a divergent suffix and integrates out full agreement, a zero-loss event. This conditional Monte Carlo construction follows the Rao–Blackwell principle (Blackwell, 1947; Casella & Robert, 1996; Robert & Casella, 2004).

At a common prefix let $c_s = \min(P_s, Q_s)$, $e_s = 1 - \sum_z c_s(z)$, and sample an auxiliary overlap path $U$ with kernel $K_s^\cap = c_s / (1 - e_s)$, extended arbitrarily when $e_s = 1$. Define

$$w_s = e_s(U_{<s}) \prod_{j<s} (1 - e_j(U_{<j})), \qquad M = \sum_s w_s = 1 - \prod_{s=1}^{H} (1 - e_s(U_{<s})). \tag{9}$$

For $M > 0$, draw $J$ with probabilities $w_s / M$, sample a fresh residual token pair at $U_{<J}$, and continue the specified coupling. Write $D_J = H^{-1} \sum_{t=J}^{H} I_t$; all estimates below are zero at $M = 0$.

**Proposition 4.1** (Unbiased conditional measurement)**.** *The estimates*

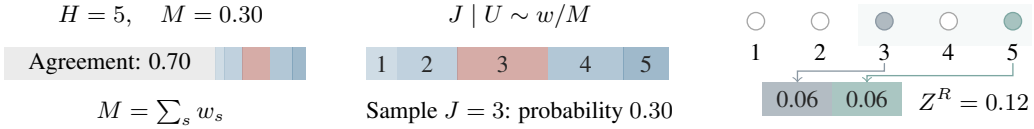
$$Z^R = MD_J, \qquad Z^O = M/H, \qquad Z^\Pi = M(D_J - 1/H) \tag{10}$$

*have expectations* $R_{\mathrm{SMC}}, O_H, \Pi_H$*. The variance of* $Z^R$ *is no greater than that of one ordinary coupled loss, and is strictly smaller exactly when* $\mathbb{P}(0 < M < 1) > 0$*.*

For $D_{s,L} = H^{-1} \sum_{j=0}^{\min(L,H-s+1)-1} I_{s+j}$, put $Z_L = MD_{J,L}$ and $W_L = H^{-1} \sum_s w_s(H - s - L + 1)_+$. Then

$$\mathbb{E}[Z_L] = R_L, \qquad \mathbb{E}[Z_H - Z_L] = R_H - R_L, \qquad \mathbb{E}[W_L] = B_L. \tag{11}$$

One suffix thus measures the window and actual tail jointly, while the scan estimates unresolved capacity. Figure 2 illustrates the mass weighting; Algorithm 1 specifies the sampler. Appendix A.4 gives the proof, and Appendix A.5 separates sampling error from capacity. The variance result compares single replicates; it does not establish a compute-normalized advantage.




(a) Scan branch mass (b) Select one branch (c) Weight its loss

**Figure 2:** Conditional sampling removes the zero-loss agreement event; weighting by $M$ preserves the unconditional mean. In this illustration, $H = 5$ and the overlap scan gives $M = 0.30$. The sampled branch $J = 3$ has mismatches at positions 3 and 5 (filled circles), giving $D_3 = 2/5$, $Z^R = 0.12$, and entry/post-entry contributions $Z^O = Z^\Pi = 0.06$.

**Algorithm 1** One conditional Monte Carlo replicate

**Require:** Kernels $P, Q$, horizon $H$, depths $\mathcal{L} \subseteq \{1, \ldots, H\}$
**Ensure:** $(Z^R, Z^O, Z^\Pi)$ and $(Z_L, Z_H - Z_L, W_L)_{L \in \mathcal{L}}$
Initialize $u \leftarrow ()$, $S \leftarrow 1$, and $w_1, \ldots, w_H \leftarrow 0$.
**for** $s = 1, \ldots, H$ **do**
  Compute $e_s(u)$; set $w_s \leftarrow Se_s(u)$, $S \leftarrow S(1 - e_s(u))$.
  If $S = 0$, **break**; all later branch weights are zero.
  Draw $U_s \sim K_s^{\cap}(\cdot \mid u)$ and append it to $u$.
**end for**
Set $M \leftarrow \sum_s w_s$, $W_L \leftarrow H^{-1} \sum_s w_s(H - s - L + 1)_+$.
If $M = 0$, **return** zero for every output.
Draw $J$ with probabilities $w_s/M$.
Draw a fresh residual pair at $U_{<J}$; continue the specified coupling, recording $I_J, \ldots, I_H$.
Compute $Z_L \leftarrow (M/H) \sum_{j=0}^{\min(L,H-J+1)-1} I_{J+j}$, including $L = H$.
**return** $(Z_H, M/H, Z_H - M/H), (Z_L, Z_H - Z_L, W_L)_{L \in \mathcal{L}}$.

## 5 How Eviction Shapes the Disagreement Trajectory

### 5.1 Paired trajectories and study design

We study Llama-3.1-8B-Instruct (Llama) and Qwen2.5-7B-Instruct (Qwen) in six HELMET/RULER benchmark–length strata (4K, 16K, 32K) (Yen et al., 2025; Hsieh et al., 2024). The first study uses 288 instances per model (48 per stratum), with $K = 64$ trajectories per action. Each instance contains a task input and document context. Actions retain 512 prompt entries with SnapKV, 50% with SnapKV, or the same bytes with a recent-token selector. A second study uses an independent cohort of 288 instances, $K = 32$, and SnapKV versus recent-token retention at 90%.

Both studies use $H = 128$, temperature-one full-softmax sampling, persistent compact caches, and same-path no-deletion controls ($K = 2$); the controls yield zero TV. Eviction follows prefill and preserves the first-token distribution. Generated entries are retained, and EOS is absorbing. The recent-token selector retains four initial sink positions and the most recent remaining entries. Budget comparisons refer to persistent cache bytes at intervention start. Sample sizes were selected

using separate development instances before evaluation. Appendix B specifies benchmark sampling, realized lengths, and the development/evaluation separation.

Ordinary coupled rollouts measure token mismatch and cross-history TV, whose expectations agree by Equation (2). Their estimates can differ at finite $K$. Actions are paired by document; their realized reference continuations may differ. Inference resamples documents with equal stratum weights: trajectories are repeated measurements, not independent statistical units.

The first study's decomposition and temporal analyses are exploratory, with pointwise 95% stratified document-bootstrap intervals. The independent 90% study prespecifies pooled late-versus-early TV and saturation endpoints, adjusting for eight comparisons; its stratified and EOS analyses are descriptive.

### 5.2 Reduced exposure accounts for most of the mismatch gap

**Table 1:** Later entry accompanies smaller, still-high post-divergence mismatch rates. $p_{\le H} = \mathbb{P}(\tau \le H)$ (%), $\bar{\tau}_H = \mathbb{E}[\tau \mid \tau \le H]$ (steps), and $C_H$ is the exposure-weighted mismatch rate (%). Entries are pooled point estimates; each cohort has $N = 288$ documents per model and $H = 128$. Selectors match bytes within each retention level; the two cohorts are independent.

| | Llama | | | Qwen | | |
|---|---|---|---|---|---|---|
| Action | $p_{\le H}$ | $\bar{\tau}_H$ | $C_H$ | $p_{\le H}$ | $\bar{\tau}_H$ | $C_H$ |
| First cohort, $K = 64$ | | | | | | |
| SnapKV-512 | 98.14 | 11.69 | 93.10 | 99.91 | 12.18 | 95.61 |
| SnapKV-50% | 92.29 | 30.37 | 89.74 | 96.76 | 29.37 | 93.59 |
| Recent-50% | 97.66 | 10.59 | 93.68 | 99.75 | 12.14 | 96.20 |
| Independent 90% cohort, $K = 32$ | | | | | | |
| SnapKV-90% | 74.24 | 49.61 | 86.56 | 80.88 | 47.74 | 91.96 |
| Recent-90% | 94.88 | 20.19 | 92.65 | 97.75 | 21.47 | 95.51 |

At identical cache bytes, SnapKV enters divergence later and less often than recent-token retention in both models, in both the 50% and independent 90% cohorts (Table 1). SnapKV-half also delays and reduces divergence relative to the smaller SnapKV-512 budget. Exposure-weighted post-divergence mismatch rates also improve, but remain above 86% under SnapKV at both retention levels.

Exposure accounts for approximately 85–92% of the six pooled mismatch gaps; rate contributions are positive but smaller (Figure 3). Exposure is the majority component in all 24 first-cohort and 12 higher-retention stratified point decompositions (Appendix B.3). The independent cohorts support a repeated pattern; between-study differences are unpaired.

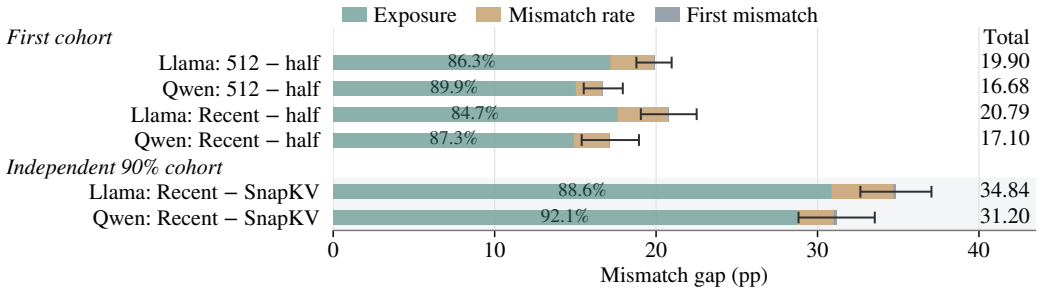



**Figure 3: Exposure dominates all six comparisons, including equal-byte selectors.** Stacks decompose mismatch gaps by Equation (6); labels give exposure shares and total gaps (pp). Components and shares are point estimates; total error bars are pointwise 95% document-bootstrap intervals. Both cohorts use $N = 288$ documents per model and $H = 128$; $K = 64$ in the first cohort and $K = 32$ in the independent 90% cohort.

Exposure is the discrete area under the first-divergence CDF (Appendix B.3): with $(H - \infty)_+ = 0$, and when $\mathbb{P}(\tau \le H) > 0$,

$$E_H = \frac{\mathbb{E}[(H-\tau)_+]}{H} = \frac{1}{H}\sum_{t=1}^{H-1}\mathbb{P}(\tau \le t) = \mathbb{P}(\tau \le H)\left(1 - \frac{\mathbb{E}[\tau \mid \tau \le H]}{H}\right). \tag{12}$$

### 5.3 THE TWO CLOCKS SEPARATE INTERVENTION GAPS FROM BRANCH EVOLUTION

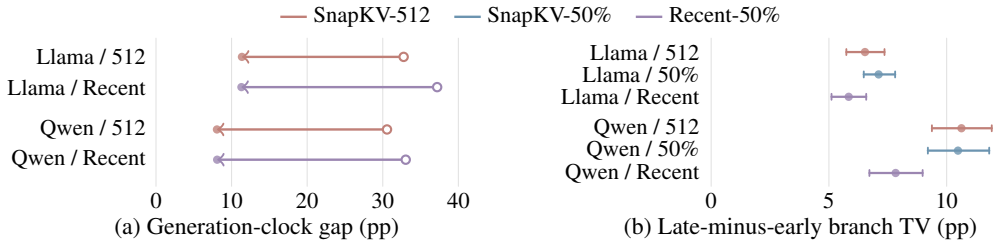



**Figure 4:** Smaller gaps between interventions coexist with rising discrepancy within divergent branches. (a) TV gap relative to SnapKV-half: arrows connect early (positions 1–32, open) and late (65–128, filled) point estimates. (b) All six branch-TV contrasts, lags 49–64 minus 1–16 in each action's fixed $\tau \leq 64$ cohort; whiskers are paired, pointwise 95% document-bootstrap intervals. Exploratory first-cohort analysis: $N = 288$ documents per model, $K = 64$.

Generation-position averages align continuations by distance from the prompt; divergence-aligned averages align them by distance from the first mismatch. Each action's fixed $\tau \leq 64$ cohort supplies all 64 subsequent lags within $H = 128$, excluding new arrivals as an explanation for its late-window increase.

Pooled intervention gaps narrow over generation time, while branch TV rises from early to late windows (Figure 4). All six pooled action–model contrasts have positive pointwise 95% lower bounds, as do all 36 stratified contrasts (Appendix B.4). This cohort-average pattern does not imply monotonic increase on each trajectory.

For each action, write $\delta_t = \delta_t(X_{<t}, Y_{<t})$. One-step maximality gives the mixture identity

$$\mathbb{E}[\delta_t] = \underbrace{p_t}_{\text{new departures}} + \underbrace{\sum_{s<t} p_s \, \mathbb{E}[\delta_t \mid \tau = s]}_{\text{occupation of existing branches}} . \tag{13}$$

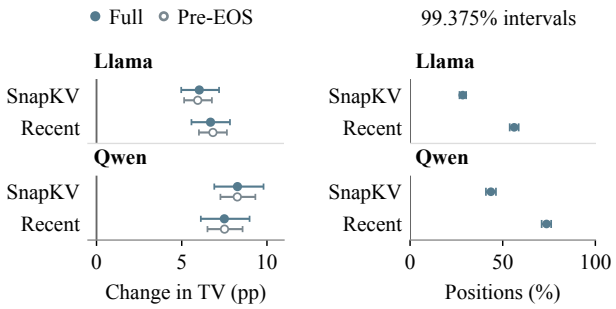
Here $\mathbb{E}[\delta_t \mathbf{1}\{\tau \geq t\}] = p_t$; zero-mass terms vanish. Generation averages mix branch ages according to each action's entry law, allowing smaller intervention gaps to coexist with rising within-branch discrepancy.

For $1 \leq j \leq 64$, branch TV also measures expected token mismatch: $\mathbb{E}[I_{\tau+j} \mid \tau \leq 64] = \mathbb{E}[\delta_{\tau+j} \mid \tau \leq 64]$. The conditioning event precedes the sampled token, permitting conditional one-step maximality (Appendix B.4). Cohorts are action-specific. Token matches after divergence need not restore similar next-token laws because histories remain different.

### 5.4 THE BRANCH PATTERN PERSISTS AT HIGH RETENTION AND BEFORE EOS

The independent 90% study confirms the branch pattern when most prompt state is retained: all four prespecified late-minus-early TV contrasts have 99.375% intervals above zero (Figure 5a). Far fewer SnapKV trajectories enter the early-divergence cohort, yet their branch TV is already high and rises further (Figure 5c). Reduced entry thus coexists with substantial conditional discrepancy. These cohort means differ from the exposure-weighted mismatch rate in Table 1; action-specific cohorts preclude interpreting between-action differences as effects on a matched population. SnapKV-90 also reduces the fraction of all positions with TV above 0.99 to 28.44%/43.63% on Llama/Qwen, compared with 56.18%/73.65% for recent-90 (Figure 5b). Fewer nearly disjoint next-token distributions coexist with rising branch TV, although substantial saturation remains.

Strict pre-EOS TV also rises for both selectors and models (Figure 5a). These descriptive means weight surviving steps, so their populations can change with lag. Late windows retain 94.5–94.9% of possible steps on Llama and approximately 99.8% on Qwen, showing an increase with little EOS exclusion in Qwen. This does not eliminate survivor selection. Appendix B.5 partitions fixed-history versus cross-history discrepancy by EOS status without causal attribution.




(a) Branch TV (b) Near-saturation

| Action | Cohort (%) | Branch TV (%) early → late |
|---|---|---|
| **Llama** | | |
| SnapKV | 51.25 | 82.64 → 88.68 |
| Recent | 88.10 | 87.44 → 94.13 |
| **Qwen** | | |
| SnapKV | 57.24 | 85.87 → 94.14 |
| Recent | 90.07 | 89.56 → 97.06 |

(c) Entry and branch TV

**Figure 5:** Fewer early divergences coexist with high and rising branch TV at 90% retention. (a) Lags 49–64 minus 1–16 in each action's fixed $\tau \leq 64$ cohort. Hollow markers show strict pre-EOS contrasts with pointwise 95% intervals, using surviving-step means. (b) The fraction of all $H$ positions with $\delta_t > 0.99$. Its intervals and the filled full-branch intervals in (a) are 99.375% (eight prespecified endpoints, family $\alpha = .05$). (c) Point estimates of $\mathbb{P}(\tau \leq 64)$ and full-branch TV at lags 1–16 → 49–64 within each action's fixed cohort. This cohort fraction differs from $\mathbb{P}(\tau \leq 128)$ in Table 1. Independent cohort: $N = 288$ documents per model, $K = 32, H = 128$.

Generation-clock gaps narrow in every HELMET cell and widen in every RULER cell (Figure 6a). Negative values mean that the Recent-minus-SnapKV gap shrinks from early to late positions; positive values mean it grows. This between-selector comparison differs from each selector's own branch evolution. All 24 pre-EOS branch contrasts have positive pointwise 95% interval lower endpoints (minimum 2.47 pp; Figure 6b). Rising branch discrepancy therefore persists across the evaluated strata even where the generation-clock trend reverses. Entry laws and branch-age weights differ between selectors (Equation (13)), so the evolution of their gap cannot substitute for divergence-aligned measurement. The full-branch contrasts share the pre-EOS direction (Appendix B.6); the pointwise intervals provide neither simultaneous coverage nor a formal test of task differences.

narrows (−) widens (+)

| | 4K | 16K | 32K |
|---|---|---|---|
| Llama / HELMET | −21.72 | −18.14 | −14.34 |
| Llama / RULER | +3.25 | +0.50 | +3.74 |
| Qwen / HELMET | −25.84 | −26.36 | −19.54 |
| Qwen / RULER | +0.27 | +0.28 | +7.84 |

Late minus early gap (pp)

| | SnapKV-90 | | | Recent-90 | | |
|---|---|---|---|---|---|---|
| | 4K | 16K | 32K | 4K | 16K | 32K |
| Llama / HELMET | +2.47 | +5.02 | +6.64 | +4.77 | +3.93 | +5.06 |
| Llama / RULER | +2.87 | +3.55 | +3.60 | +4.46 | +6.05 | +5.61 |
| Qwen / HELMET | +4.85 | +8.19 | +7.40 | +3.86 | +3.39 | +4.79 |
| Qwen / RULER | +4.42 | +5.12 | +5.92 | +6.67 | +6.70 | +6.59 |

95% interval lower endpoint (pp)

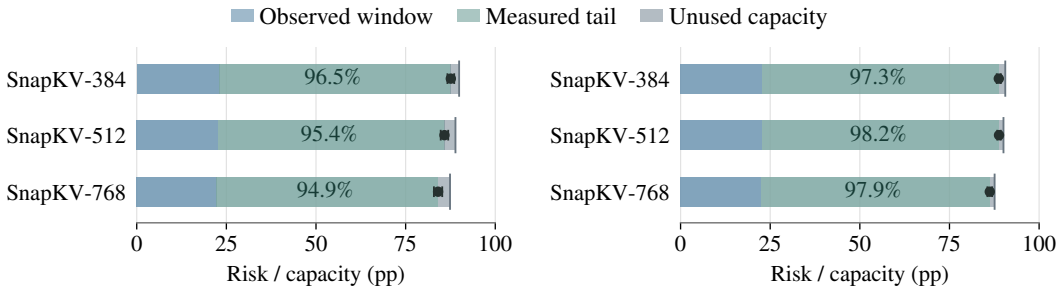
All 24 endpoints are positive

(a) Selector gap shrinks or widens (b) Pre-EOS TV: lower interval endpoints

**Figure 6:** Observed generation-clock gap trends differ across benchmark–length strata; all pre-EOS branch contrasts have positive pointwise lower bounds. (a) Descriptive recent-minus-SnapKV gap changes from positions 1–32 to 65–128; blue denotes narrowing and orange widening. (b) Lower endpoints of the two-sided, pointwise 95% document-bootstrap intervals for pre-EOS TV at lags 49–64 minus 1–16. Intervals are not simultaneous. Each action uses its own $\tau \leq 64$ cohort and surviving-step window means. Independent 90% study: $N = 288$ documents per model, $K = 32, H = 128$.

### 5.5 Measured tails occupy most of the remaining capacity

Theorem 3.2 bounds unobserved occupation by $B_L$. In a supplementary SnapKV study, tails beyond $L = 32$ occupy 94.9–98.2% of the estimated remaining capacity across six model–budget conditions (Figure 7; Appendix C.2). Under SnapKV-512, the window contributes approximately 22.7/22.8 pp on Llama/Qwen, whereas full-horizon risk reaches 85.8/88.8 pp. All quantities use the same denominator $H$: the small window contribution leaves substantial disagreement unobserved. All depths reuse complete traces, so they do not establish the cost of a shorter simulation. Proximity to the plug-in upper endpoint concerns these interventions; theoretical sharpness permits both kernels to vary.




(a) Llama (b) Qwen

**Figure 7: Measured tails occupy most of the remaining capacity.** At branch depth $L = 32$, labels show the point ratio $(\widehat{R}_{128} - \widehat{R}_{32})/\widehat{B}_{32}$. Bars end at the plug-in endpoint $\widehat{R}_{32} + \widehat{B}_{32}$; dots mark $\widehat{R}_{128}$ with one conditional MC SE. Estimates use 24 evaluation documents per model and $K = 8$ SWRB replicates. Neither the endpoint nor the conditional SE is a population confidence bound.

For nested exact profiles with the same divergence law, increasing depth from $L$ to $L + 1$, for $1 \le L < H$, gives

$$\begin{aligned} B_L - B_{L+1} &= \frac{\mathbb{P}(\tau \le H - L)}{H}, \\ R_{L+1} - R_L &= \frac{1}{H} \sum_{s \le H-L} p_s \rho_{s,L}. \end{aligned} \tag{14}$$

The first equality follows by differencing Equation (7); since $0 \le \rho_{s,L} \le 1$, the compatible intervals nest. Entry mass determines their width reduction, while newly observed mismatch determines how much the lower endpoint rises and the upper endpoint falls. Thus observation replaces unresolved capacity with measured occupation. Increasing $K$ improves Monte Carlo precision at a fixed depth, whereas increasing $L$ contracts the exact identification interval. Estimated endpoints carry additional sampling error (Appendix A.5).

Development data can support a narrower predictive target. In the first cohort, frozen ridge predictors using generation positions 1–32 recover the population-average SnapKV-512 minus SnapKV-half full-horizon rollout-TV gap within the prespecified $\pm 2$ pp equivalence margin on both models (adjusted 98.75% intervals; Appendix B.2). These generation-prefix observations differ from the theorem's divergence-aligned profile, and development data supply information beyond that profile. Aggregate prediction therefore coexists with unresolved trajectory structure: mean-gap equivalence does not establish document-level accuracy or recovery of the branch evolution.

## 6 Conclusion

We study how persistent KV-cache eviction shapes the onset and evolution of output disagreement. Across the tested interventions, lower cumulative mismatch is associated mainly with reduced post-divergence exposure, including at identical cache bytes. Exposure-weighted mismatch rates also improve but remain high. An independent study at 90% retention finds higher late- than early-window TV within fixed early-divergence cohorts in both models. These findings distinguish avoiding or delaying divergence from reducing disagreement after divergence as complementary aspects of positional agreement.

Our technical contributions comprise an exact mismatch decomposition, a sharp characterization of risks compatible with exact finite divergence-aligned profiles over unrestricted kernel pairs, and unbiased conditional Monte Carlo estimators of observed occupation, actual tail, and unresolved capacity. Together, they separate incomplete observation from sampling error and clarify how trajectory structure informs aggregate risk. However, our empirical conclusions concern the evaluated models and interventions, with positional disagreement measured under the specified coupling. Future work should examine how these diagnostics relate to task performance and test the causal role of history feedback through controlled interventions.

## Reproducibility Statement

Section 3.2 proves the identification theorem, and Algorithm 1 specifies the conditional sampler. Appendix A provides the remaining proofs and scope examples. Appendix B specifies sampling, study separation, statistical units, primary endpoints, and supporting results for the trajectory studies. Appendices C and D report supplementary measurements and applications.

## AI Use Statement

AI assistants supported manuscript drafting, code development, literature search, and checks of mathematical arguments. The authors are responsible for verifying the claims, citations, analyses, and final presentation.

## References

Jinwoo Ahn, Ingyu Seong, Akhil Kedia, Junhan Kim, Hyemi Jang, Kangwook Lee, and Yongkweon Jeon. LookaheadKV: Fast and accurate KV cache eviction by glimpsing into the future without generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/746222c1871d3fa7a94bdacabc34e26c-Abstract-Conference.html.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, 2022. doi: 10.18653/v1/2022.findings-acl.58. URL https://aclanthology.org/2022.findings-acl.58/.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Nina L. Corvelo Benz, Stratis Tsirtsis, Eleni Straitouri, Ivi Chatzi, Ander Artola Velasco, Suhas Thejaswi, and Manuel Gomez Rodriguez. Evaluation of large language models via coupled token generation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2215–2223, 2026. URL https://proceedings.mlr.press/v300/benz26a.html.

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 1947. doi: 10.1214/aoms/1177730497.

George Casella and Christian P. Robert. Rao–Blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996. doi: 10.1093/biomet/83.1.81.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: KV-cache compression with low-rank projection. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7da6e0e00702c60607a6ae05c802ef85-Abstract-Conference.html.

Philip A. Ernst, Wilfrid S. Kendall, Gareth O. Roberts, and Jeffrey S. Rosenthal. MEXIT: Maximal un-coupling times for stochastic processes. *Stochastic Processes and their Applications*, 129(2): 355–380, 2019. doi: 10.1016/j.spa.2018.03.001.

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.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. Not all heads matter: A head-level KV cache compression method with integrated retrieval and reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f649556471416b35e60ae0de7c1e3619-Abstract-Conference.html.

Wei Gao, Xinyu Zhou, Peng Sun, Tianwei Zhang, and Yonggang Wen. Rethinking key-value cache compression techniques for large language model serving. In *Proceedings of Machine Learning and Systems*, volume 7, 2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/26289c647c6828e862e271ca3c490486-Abstract-Conference.html.

Paul Glasserman and Jeremy Staum. Conditioning on one-step survival for barrier option simulations. *Operations Research*, 49(6):923–937, 2001. doi: 10.1287/opre.49.6.923.10018. URL https://pubsonline.informs.org/doi/10.1287/opre.49.6.923.10018.

Sheldon Goldstein. Maximal coupling. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 46(2):193–204, 1979. doi: 10.1007/BF00533259.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785, 2025. URL https://proceedings.mlr.press/v291/haris25a.html.

Tianxing He, Jingzhao Zhang, Zhiming Zhou, and James Glass. Exposure bias versus self-recovery: Are distortions really incremental for autoregressive text generation? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5087–5102, 2021. doi: 10.18653/v1/2021.emnlp-main.415. URL https://aclanthology.org/2021.emnlp-main.415/.

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.

Jushi Kai, Zhuiri Xiao, Alexandra Birch, and Zhouhan Lin. Information-aware KV cache compression for long reasoning. *arXiv preprint arXiv:2606.26875*, 2026. URL https://arxiv.org/abs/2606.26875.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: Sharp partial identification under smoothness. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23734–23757, 2024. URL https://proceedings.mlr.press/v235/khan24b.html.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19274–19286, 2023. URL https://proceedings.mlr.press/v202/leviathan23a.html.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2 edition, 2017. URL https://pages.uoregon.edu/dlevin/MARKOV/mcmt2e.pdf. With contributions by Elizabeth L. Wilmer.

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.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32332–32344, 2024. URL https://proceedings.mlr.press/v235/liu24bz.html.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003. doi: 10.1007/b97478. URL https://link.springer.com/book/10.1007/b97478.

Eric Price, Kevin Tian, Zhiyang Xun, and Yusong Zhu. Total variation distance estimation in autoregressive models. *arXiv preprint arXiv:2607.19510*, 2026. URL https://arxiv.org/abs/2607.19510.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. CAKE: Cascading and adaptive KV cache eviction with layer preferences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dfae940651f3e690a12e19c874edad7c-Abstract-Conference.html.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2 edition, 2004. doi: 10.1007/978-1-4757-4145-2. URL https://link.springer.com/book/10.1007/978-1-4757-4145-2.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Jörg Stoye. Minimax regret treatment choice with incomplete data and many treatments. *Econometric Theory*, 23(1):190–199, 2007. doi: 10.1017/S0266466607070089.

Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. SpecTr: Fast speculative decoding via optimal transport. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6034a661584af6c28fd97a6f23e56c0a-Abstract-Conference.html.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Danning Ke, Shikuan Hong, Yiwu Yao, and Gongyi Wang. RazorAttention: Efficient KV cache compression through retrieval heads. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2a98af4fea6a24b73af7b588ca95f755-Abstract-Conference.html.

Florian Völlering. On maximal agreement couplings. *arXiv preprint arXiv:1608.01511*, 2016. URL https://arxiv.org/abs/1608.01511.

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

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. HELMET: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f5332c8273d02729730a9c24dec2135e-Abstract-Conference.html.

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. H2O: 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 Proofs and Conditional Measurement

All probabilities use the stepwise overlap–residual coupling in Section 3, unless an auxiliary sampling law is specified. Empty sums and conditional quantities on null events are zero.

## A.1 First-divergence and occupation identities

Let $\mathcal{F}_{t-1} = \sigma(X_{<t}, Y_{<t})$, $\delta_t = \mathrm{TV}(P_t(\cdot \mid X_{<t}), Q_t(\cdot \mid Y_{<t}))$, and $S_t = \mathbb{P}(\tau \geq t)$. Stepwise maximality gives $\mathbb{E}[I_t \mid \mathcal{F}_{t-1}] = \delta_t$. Both $\{\tau \geq t\}$ and $\{\tau < t\}$ are $\mathcal{F}_{t-1}$-measurable, so

$$p_t = \mathbb{E}[\delta_t \mathbf{1}\{\tau \geq t\}] = S_t \lambda_t, \qquad S_{t+1} = S_t(1 - \lambda_t),$$
$$HO_H = \sum_t \mathbb{E}[\delta_t \mathbf{1}\{\tau \geq t\}], \qquad H\Pi_H = \sum_t \mathbb{E}[\delta_t \mathbf{1}\{\tau < t\}], \tag{15}$$

where $\lambda_t = \mathbb{E}[\delta_t \mid \tau \geq t]$. The first equality uses $I_t \mathbf{1}\{\tau \geq t\} = \mathbf{1}\{\tau = t\}$; the last uses $\sum_t I_t \mathbf{1}\{\tau < t\} = \sum_{t>\tau} I_t$. Adding the two occupation terms proves Equation (4) and the TV representation (2). The first-divergence partition gives $R_{\mathrm{SMC}} = \sum_s p_s \kappa_s$, since loss vanishes on $\{\tau = \infty\}$. On $\{\tau = s\}$, exactly one mismatch is forced and at most $H - s + 1$ are possible, proving the bounds in Equation (3). The exposure definition gives $\Pi_H = E_H C_H$; the identity $E_a C_a - E_b C_b = (E_a - E_b)(C_a + C_b)/2 + (C_a - C_b)(E_a + E_b)/2$ proves Equation (6).

## A.2 A fixed reference can narrow the interval

Theorem 3.2 varies both kernels. To see why this qualification matters, take $\mathcal{V} = \{0, 1\}$, $H = 2$, $L = 1$, and the fixed reference $P_{0,t}(1 \mid h) = 1/2$ at both steps. Let the profile have $p = (\varepsilon, 0)$, $0 < \varepsilon < 1/2$, and its required first-mismatch probabilities. Every compatible $Q$ satisfies $|Q_1(1) - 1/2| = \varepsilon$; hence both overlap masses $c_1(z) = \min\{1/2, Q_1(z)\}$ are positive. Since

$$0 = p_2 = \sum_{z=0}^{1} c_1(z) \left| \tfrac{1}{2} - Q_2(1 \mid z) \right|,$$

we must have $Q_2(1 \mid 0) = Q_2(1 \mid 1) = 1/2$. Consequently $I_2 = 0$ almost surely, including after a first-step divergence, and every compatible risk equals $\varepsilon/2$. The choice $Q_1(1) = 1/2 + \varepsilon$, $Q_2(1 \mid z) = 1/2$ attains it. Thus the fixed-reference compatible set is $\{\varepsilon/2\}$, a strict subset of the unrestricted interval $[\varepsilon/2, \varepsilon]$.

## A.3 Temporal occupation and history feedback

Post-divergence occupation need not be caused by feedback from differing histories. For $e \in (0, 1)$, take history-independent kernels $P_t = \delta_0$, $Q_t = (1 - e)\delta_0 + e\delta_1$. Fresh coupling randomness gives independent $I_t \sim \mathrm{Bernoulli}(e)$, so

$$R_{\mathrm{SMC}} = e, \qquad O_H = \frac{1 - (1 - e)^H}{H}, \qquad \frac{\Pi_H}{R_{\mathrm{SMC}}} = 1 - \frac{1 - (1 - e)^H}{He} \longrightarrow 1.$$

Both kernels have zero sensitivity to history, yet almost all occupation is post-divergence for large $H$. Separating the direct effects of persistent KV changes from history feedback therefore requires additional intervention controls.

**Equal entry laws need not imply equal occupation.** With $P_t = \delta_0$, set $Q_1(1) = \varepsilon$ for two actions. At $t \geq 2$, let one action always output 0 and the other output $\mathbf{1}\{y_1 = 1\}$. Both have common-prefix discrepancy $\varepsilon \mathbf{1}\{t = 1\}$, first-divergence mass $p_1 = \varepsilon$, and sequence TV $\varepsilon$; their cumulative risks are $\varepsilon/H$ and $\varepsilon$, respectively. Also, a later token match cannot erase the earlier differing tokens from the histories: $X_{<t} \neq Y_{<t}$ for every $t > \tau$. The pattern $(I_1, I_2, I_3) = (1, 0, 1)$, for example, follows from deterministic outputs $X = 000$, $Y = 101$.

## A.4 Conditional measurement: correctness and variance

The auxiliary path uses the normalized overlap kernel

$$K_s^{\cap}(z \mid h) = \frac{\min\{P_s(z \mid h), Q_s(z \mid h)\}}{1 - e_s(h)} \quad (e_s(h) < 1), \tag{16}$$

with any extension at $e_s = 1$. A residual branch at $s$ samples its unequal token pair from the two residual laws and then applies the specified coupling with fresh randomness. Let $D_s = H^{-1}\sum_{t=s}^{H} I_t$, $m_s = \mathbb{E}[D_s \mid U]$, and $\nu_s = \mathbb{E}[D_s^2 \mid U]$, all zero when $w_s = 0$. Conditional on $U$, select $J$ with mass $w_s/M$ and generate $D_J$; if $M = 0$, set $(J, D_J) = (1, 0)$.

*Proof of Proposition 4.1.* Writing $c_j = \min(P_j, Q_j)$, the overlap identity gives

$$\begin{aligned}\mathbb{P}_U(U_{<s} = u_{<s})w_s(u) &= \prod_{j<s} c_j(u_j \mid u_{<j})\, e_s(u_{<s}) \\ &= \mathbb{P}(X_{<s} = Y_{<s} = u_{<s}, \tau = s).\end{aligned} \tag{17}$$

Likewise, $\mathbb{P}_U(U = u)(1 - M(u)) = \prod_{j=1}^{H} c_j(u_j \mid u_{<j})$ is the probability of full agreement on $u$. Take $B \mid U \sim \mathrm{Bernoulli}(M)$, conditionally independent of $(J, D_J)$. For any bounded measurable $f$,

$$\mathbb{E}[f(BD_J)] = \mathbb{E}_U\left[(1-M)f(0) + \sum_s w_s \mathbb{E}[f(D_s) \mid U]\right] = \mathbb{E}\left[f\left(H^{-1}\sum_t I_t\right)\right].$$

Thus $BD_J$ has the ordinary coupled-loss law and $\mathbb{E}[BD_J \mid U, J, D_J] = MD_J$. Also $\mathbb{E}[w_s] = p_s$, so $\mathbb{E}[Z^O] = \sum_s p_s/H = O_H$ and $\mathbb{E}[Z^\Pi] = \mathbb{E}[Z^R - Z^O] = \Pi_H$. Conditional second moments are $\mathbb{E}[(BD_J)^2 \mid U] = \sum_s w_s\nu_s$ and $\mathbb{E}[(MD_J)^2 \mid U] = M\sum_s w_s\nu_s$; hence

$$\mathrm{Var}(Z_{\mathrm{naive}}) - \mathrm{Var}(Z^R) = \mathbb{E}\left[(1-M)\sum_s w_s\nu_s\right] \geq 0. \tag{18}$$

On every positive-weight branch $H^{-1} \leq D_s \leq 1$. Therefore

$$\frac{\mathbb{E}[M(1-M)]}{H^2} \leq \mathrm{Var}(Z_{\mathrm{naive}}) - \mathrm{Var}(Z^R) \leq \mathbb{E}[M(1-M)],$$

proving that strict improvement holds exactly when $\mathbb{P}(0 < M < 1) > 0$. The same argument applied to $D_J - 1/H$ gives variance dominance for the post-divergence component. All conclusions include $M = 0$ and $M = 1$. □

**Joint depth identities.** For $m_{s,L} = \mathbb{E}[D_{s,L} \mid U]$, Equation (17) implies $\mathbb{E}[w_s m_{s,L}] = (p_s/H)\sum_{j<\ell_s}\rho_{s,j}$. Since $\mathbb{E}[Z_L \mid U] = \sum_s w_s m_{s,L}$, summation proves $\mathbb{E}[Z_L] = R_L$. Linearity gives $\mathbb{E}[Z_H - Z_L] = R_H - R_L$ and $\mathbb{E}[W_L] = H^{-1}\sum_s p_s(H - s - L + 1)_+ = B_L$, proving Equation (11). Reusing one suffix retains the covariance in $\mathrm{Var}(Z_H - Z_L) = \mathrm{Var}(Z_H) + \mathrm{Var}(Z_L) - 2\,\mathrm{Cov}(Z_H, Z_L)$.

### A.5 Sampled profiles and interval uncertainty

The averages $\widehat{R}_L$ and $\widehat{R}_L + \widehat{B}_L$ estimate the identified-set endpoints. Identification width $B_L$ and sampling error are distinct. For a fixed action and a fixed set of $n$ documents, assume that all $nK$ replicate streams are independent conditional on the documents. Both $Z_L$ and $Z_L + W_L$ lie in $[0, 1]$, because $Z_L \leq ML/H$ and $W_L \leq M(H - L)/H$. Hoeffding's inequality and a union bound then enclose the mean risk over these documents, with probability at least $1 - \alpha$, in

$$\left[\max\{0, \widehat{R}_L - \epsilon_\alpha\}, \quad \min\{1, \widehat{R}_L + \widehat{B}_L + \epsilon_\alpha\}\right], \qquad \epsilon_\alpha = \sqrt{\frac{\log(4/\alpha)}{2nK}}.$$

This conditional Monte Carlo statement does not quantify generalization to new documents; document-level inference treats each document as one statistical unit. Simultaneous actions or depths require further error allocation.

## B Complete-Trajectory Studies

### B.1 Documents, interventions, and measurement paths

The first study evaluates Llama and Qwen on 288 independent documents, with 48 per task–length cell (HELMET/RULER at 4K, 16K, and 32K). Actions comprise SnapKV-512, SnapKV-half, recent-half,

and a same-path no-deletion control. The half-budget actions retain $\lfloor n/2 \rfloor$ prompt slots per KV head and layer, where $n$ is the actual tokenized prompt length. Recent-half retains four initial sink positions and the latest remaining positions, matching SnapKV-half in persistent resident bytes. SnapKV uses a 64-token scoring window and kernel size five. Each document has $K = 64$ trajectories per compressed action and $K = 2$ per control, totaling 111,744 trajectories across 576 model–document pairs.

The higher-retention study uses an independent cohort of 288 documents (48 per cell), separate from the first study and its development data. SnapKV-90 and recent-90 retain $\lfloor 9n/10 \rfloor$ slots per KV head and layer, with the same recent-token selector and equal persistent resident bytes. With $K = 32$ per compressed action and $K = 2$ per control, it yields 38,016 trajectories. Retention comparisons are between independent cohorts.

Both studies use $H = 128$, temperature-one full-softmax distributions, and the stepwise maximal coupling defined in Section 3. Eviction follows prefill and preserves the first-token distribution. Generated KV entries are appended; prompt deletion persists. EOS becomes absorbing after its output step. No-deletion controls have zero measured TV.

Within each action and replicate, the full cache generates $X$, the compressed cache generates $Y$, and a second compressed cache consumes $X$. Writing

$$v_t^a = \mathrm{TV}\big(P_t(\cdot \mid X_{<t}), Q_t^a(\cdot \mid Y_{<t})\big), \qquad f_t^a = \mathrm{TV}\big(P_t(\cdot \mid X_{<t}), Q_t^a(\cdot \mid X_{<t})\big),$$

we estimate $R(a) = H^{-1} \sum_t \mathbb{E}[v_t^a] = R_{\mathrm{SMC}}(P, Q^a)$ and $F(a) = H^{-1} \sum_t \mathbb{E}[f_t^a]$ by ordinary coupled rollouts. The three paths are paired within actions; across actions, pairing is by document, and action-specific coupling may yield different $X$ sequences. The difference $R - F$ compares history-conditioned distributions and does not identify a causal history-feedback effect.

Each study uses disjoint implementation-check, development, and evaluation sets, with six implementation-check and 36 development documents (six per cell). With development $K = 32$, the first selects calibrators and $(N, K) = (288, 64)$ from prespecified grids $N \in \{120, 180, 288\}$ and $K \in \{16, 32, 64\}$. The second uses development $K = 16$ and an eight-endpoint precision criterion to select $N = 288$ from the same grid at evaluation $K = 32$. Some development precision targets exceeded the grids; sample sizes were fixed before evaluation without adaptive extension.

Documents are the statistical units, with repeated trajectories per document and equal weights across cells. Bootstrap resampling is stratified by cell and paired across actions. Prespecified comparisons use 10,000 draws; supporting analyses use 2,000 in the first study and 10,000 in the second, with pointwise 95% intervals retaining finite-$K$ measurement noise. First-cohort decomposition and temporal analyses are exploratory, having followed its aggregate validation outcomes.

Recorded generation depths are $1, 4, 8, 16, 32, 64, 128$; calibration uses positions 1–32. The theoretical depth-$L$ profile starts at and includes the first mismatch. Empirical branch profiles use strictly later lags 1–64 within the fixed cohort $\{\tau \leq 64\}$.

**Benchmark sampling and length bands.** The candidate pool comprises six HELMET subsets (NQ, TriviaQA, HotpotQA, PopQA, ASQA, and QAMPARI) and 13 RULER tasks. Previously used identities and content hashes are excluded; the 90% study also excludes the first study's entire source inventory. Within each benchmark–band cell, fixed hash ranking and subtask round-robin selection select documents without using model outcomes. A separate identity ranking assigns one implementation-check, six development, and 48 evaluation instances per cell; subtask counts need not be equal. The labels 4K, 16K, and 32K denote nominal budgets of 4,096, 16,384, and 32,768 tokens. Both tokenizers must fit the budget minus 128 generation tokens. Shorter prompts are retained; over-budget contexts are truncated while preserving the question. Tables 2 and 3 report the first study's realized composition and lengths; the higher-retention cohort is sampled separately.

### B.2 Aggregate prediction from development-selected calibrators

The four prespecified calibration comparisons predict the population-average SnapKV-512 minus SnapKV-half gap in $R$. For each model and observation type, six-fold development validation selects among cell means, flat predictors, and ridge by document-level paired-gap MSE.

**Table 2:** First-study instance counts by subtask and final data split, pooled across nominal length bands. Each benchmark–band cell contains one implementation-check, six development, and 48 evaluation instances.

| Benchmark | Subtask | Check | Dev. | Eval. |
|---|---|---|---|---|
| HELMET | NQ | 0 | 4 | 29 |
| | TriviaQA | 0 | 1 | 32 |
| | HotpotQA | 1 | 3 | 28 |
| | PopQA | 0 | 0 | 1 |
| | ASQA | 0 | 7 | 26 |
| | QAMPARI | 2 | 3 | 28 |
| RULER | `cwe` | 1 | 2 | 10 |
| | `fwe` | 1 | 4 | 7 |
| | `niah_multikey_1` | 0 | 2 | 10 |
| | `niah_multikey_2` | 0 | 3 | 9 |
| | `niah_multikey_3` | 0 | 3 | 9 |
| | `niah_multiquery` | 1 | 1 | 12 |
| | `niah_multivalue` | 0 | 0 | 13 |
| | `niah_single_1` | 0 | 0 | 14 |
| | `niah_single_2` | 0 | 1 | 13 |
| | `niah_single_3` | 0 | 2 | 10 |
| | `qa_1` | 0 | 0 | 12 |
| | `qa_2` | 0 | 0 | 12 |
| | `vt` | 0 | 0 | 13 |
| Total | | 6 | 36 | 288 |

**Table 3:** Actual prompt-token ranges in the first study's evaluation cohort, with 48 instances per row. Lengths include tokenizer-added special tokens. No prompt in this cohort required truncation.

| Benchmark | Nominal band | Llama tokens | Qwen tokens |
|---|---|---|---|
| HELMET | 4K | 1,898–2,618 | 2,023–2,770 |
| | 16K | 9,352–14,436 | 9,845–15,454 |
| | 32K | 23,397–31,122 | 24,312–32,618 |
| RULER | 4K | 1,606–3,702 | 2,079–3,701 |
| | 16K | 10,296–15,595 | 12,089–15,612 |
| | 32K | 21,700–31,284 | 26,061–31,287 |

Cell means group observations by stratum and action. Short-window ridge augments stratum and action indicators with

$$z_{32}(d,a) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{H}\sum_{t=1}^{32} I_{dakt},\, \mathbf{1}\{\tau_{dak} \le 32\},\, \mathbf{1}\{e^X_{dak} \le 32\},\, \mathbf{1}\{e^Y_{dak} \le 32\}\right),$$

where $e^X$ and $e^Y$ are the first EOS positions, taken as infinity if absent. Fixed-history ridge uses the full-horizon $F(d,a)$ instead of $z_{32}(d,a)$. Ridge uses training-only standardization, $\alpha = 10$, and an unpenalized intercept. All candidate predictions are clipped to $[0,1]$ separately for each action before computing the predicted gap. Error is the full-$H$ rollout-TV gap minus its prediction. Mean-gap equivalence requires the entire adjusted interval to lie within the prespecified $[-2,2]$ percentage-point (pp) margin.

Table 4 distinguishes aggregate equivalence from unresolved evidence: the Llama fixed-history interval establishes neither equivalence nor bias outside the margin. Qwen's selected cell predictor uses no document-specific $F$, so equivalence does not establish its incremental predictive value. Document-contrast MAE/RMSE against finite-$K$ targets is 4.09/5.16 pp (Llama) and 4.92/7.09 pp (Qwen) for short-window predictors, versus 8.73/11.04 and 10.03/13.18 pp for fixed-history predictors. Aggregate equivalence therefore does not imply document-level accuracy; branch-curve recovery, stopping guarantees, and net computation savings were not tested.

**Table 4: Calibrated population gaps.** Error is the full-$H$ rollout-TV gap minus its development-selected prediction. Bold intervals establish mean-gap equivalence within $[-2, 2]$ pp; the remaining comparison is unresolved. Intervals use $\alpha = 0.05/4$.

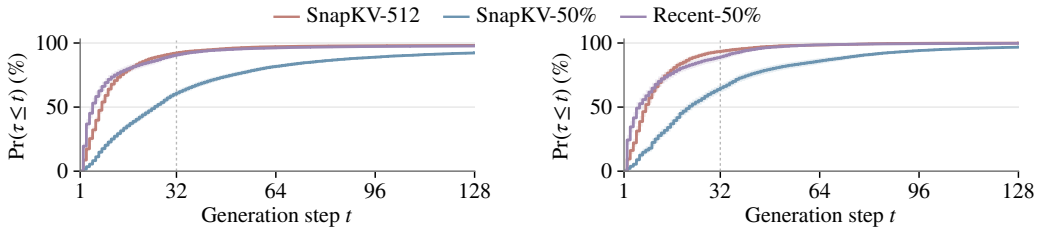
| Model / observation | Fit | Error, pp [98.75% CI] |
|---|---|---|
| Llama fixed history | Ridge | 1.127 $[-0.304, 2.570]$ |
| Llama positions 1–32 | Ridge | **−0.388** $\mathbf{[-1.062, 0.301]}$ |
| Qwen fixed history | Cell | **0.370** $\mathbf{[-1.162, 1.965]}$ |
| Qwen positions 1–32 | Ridge | **0.114** $\mathbf{[-0.867, 1.135]}$ |

**Table 5: Uncalibrated population gaps.** All gaps compare SnapKV-512 with SnapKV-half, as in Table 4. Flat extrapolation uses the mean token mismatch at positions 1–32; it differs from the calibrated ridge predictor. Difference intervals are pointwise 95%.

| Gap, pp [95% CI] | Llama | Qwen |
|---|---|---|
| Full $H$ | 19.89 | 16.68 |
| Fixed history | 9.50 | 12.63 |
| Flat 1–32 | 32.78 | 30.57 |
| Full − fixed history | 10.39 $[9.26, 11.55]$ | 4.06 $[2.60, 5.46]$ |
| Full − flat 1–32 | −12.89 $[-13.97, -11.83]$ | −13.89 $[-15.39, -12.40]$ |

### B.3 Entry, exposure, and mismatch occupation

We apply the entry–exposure identity (Equation (5)) and its symmetric contrast (Equation (6)) to empirical mismatch moments $M = H^{-1}\mathbb{E}[\sum_t I_t]$, where $I_t = \mathbf{1}\{X_t \neq Y_t\}$. All reported compressed conditions have positive exposure. The rate $C_H$ weights trajectories by exposure, which combines divergence occurrence and timing. TV and token mismatch agree in expectation under the coupling, although their sample means may differ (see also Equation 19). Component shares divide by the empirical mismatch gap; they are descriptive, not causal.




(a) Llama (b) Qwen

**Figure 8: First-divergence CDFs in the first cohort.** Curves show $\mathbb{P}(\tau \leq t)$ over all trajectories, including those agreeing through $H = 128$. There are $N = 288$ documents per model and $K = 64$ replicates per document and action. Shading gives pointwise 95% stratified document-bootstrap intervals.

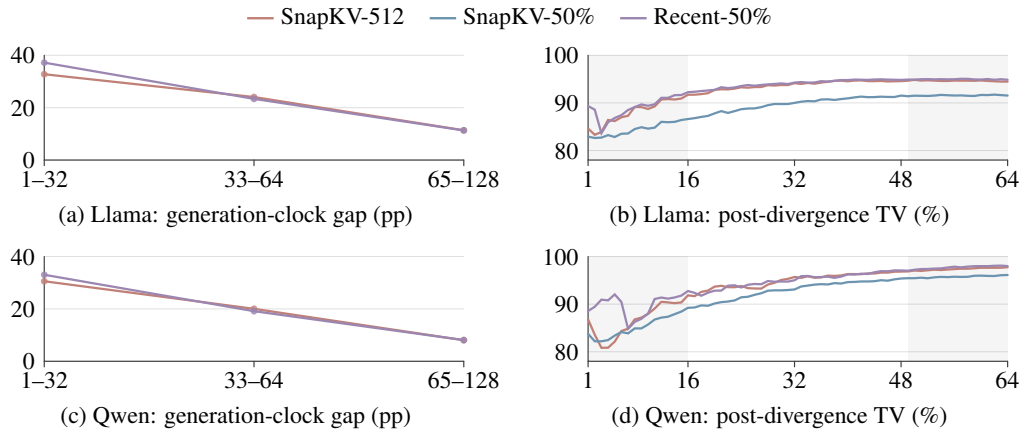
**Table 6:** Mismatch-gap decomposition (pp) by Equation (6). Entry is $\Delta O_H$; exposure and rate are $\Delta E_H \, \overline{C}_H$ and $\Delta C_H \, \overline{E}_H$. Here 512 and half denote SnapKV-512 and SnapKV-50%. Values are pooled point estimates; rounding can change sums by 0.01 pp.

| | Llama | | | | Qwen | | | |
|---|---|---|---|---|---|---|---|---|
| Contrast | Total | Entry | Exposure | Rate | Total | Entry | Exposure | Rate |
| 512 − half | 19.90 | 0.05 | 17.17 | 2.68 | 16.68 | 0.02 | 14.99 | 1.67 |
| Recent-half − half | 20.79 | 0.04 | 17.60 | 3.15 | 17.10 | 0.02 | 14.92 | 2.15 |
| Recent-90 − SnapKV-90 | 34.84 | 0.16 | 30.87 | 3.82 | 31.20 | 0.13 | 28.72 | 2.35 |

The pooled decomposition appears in Table 6. Exposure is the majority component in every model–stratum–contrast point estimate: 78.1%–96.1% across 24 first-cohort comparisons and 86.13%–95.36% across 12 higher-retention comparisons.

### B.4 Post-divergence time and higher retention

Conditioning on $\tau \leq 64$ supplies all 64 post-divergence positions within $H = 128$. Each action's cohort is fixed across lags, although cohorts differ between actions.




(a) Llama: generation-clock gap (pp) (b) Llama: post-divergence TV (%)

(c) Qwen: generation-clock gap (pp) (d) Qwen: post-divergence TV (%)

**Figure 9:** Narrowing intervention gaps coexist with increasing within-branch discrepancy. Left: intervention-minus-SnapKV-half TV gaps over generation-position blocks. Right: TV at strictly post-divergence lags within each action's fixed $\tau \leq 64$ cohort. Shading marks the early (1–16) and late (49–64) windows used in the paired contrast. Curves are exploratory point estimates, $N = 288, K = 64$; TV and token mismatch have equal conditional means at these lags but can differ in finite samples.

For a fixed action and $1 \leq j \leq 64$, put $\mathcal{F}_{t-1} = \sigma(d, X_{<t}, Y_{<t})$, where $d$ is the document. Since $\{\tau = s\} \in \mathcal{F}_{s+j-1}$, one-step maximality gives

$$\begin{aligned} \sum_{s=1}^{64} \mathbb{E}[I_{s+j}\mathbf{1}\{\tau = s\}] &= \sum_{s=1}^{64} \mathbb{E}[v_{s+j}\mathbf{1}\{\tau = s\}], \\ \mathbb{E}[I_{\tau+j} \mid \tau \leq 64] &= \mathbb{E}[v_{\tau+j} \mid \tau \leq 64]. \end{aligned} \tag{19}$$

The second line follows by division by $\mathbb{P}(\tau \leq 64) > 0$. Lag-zero and strict pre-EOS selection depend on current tokens, so the same conditional equality need not hold.

In the first study the cohort fractions for SnapKV-512, SnapKV-half, and recent-half are respectively 97.18%, 81.81%, and 96.19% on Llama, and 98.48%, 86.07%, and 98.57% on Qwen. Their post-divergence late-minus-early TV contrasts are $6.53\,[5.74, 7.36]$, $7.11\,[6.48, 7.81]$, and $5.84\,[5.11, 6.58]$ pp on Llama, and $10.63\,[9.37, 11.91]$, $10.48\,[9.20, 11.81]$, and $7.83\,[6.72, 8.98]$ pp on Qwen (exploratory pointwise 95% intervals). All 36 corresponding stratum–model–action point estimates and pointwise interval lower bounds are positive.

**Table 7:** The independent 90% study ($N = 288, K = 32, H = 128$). Mean entry time conditions on divergence by $H$. Early/late TV uses lags 1–16/49–64 in each action's fixed $\tau \leq 64$ cohort. Saturation is the fraction of all $H$ positions with $\delta_t > 0.99$. The eight saturation and late-minus-early endpoints have prespecified 99.375% bootstrap intervals; other entries are descriptive point estimates. Contributing documents refer to the branch contrast.

| | Llama | | Qwen | |
|---|---|---|---|---|
| Metric | SnapKV-90 | Recent-90 | SnapKV-90 | Recent-90 |
| Divergence by $H$ (%) | 74.24 | 94.88 | 80.88 | 97.75 |
| Mean entry time (steps) | 49.61 | 20.19 | 47.74 | 21.47 |
| Early branch TV (%) | 82.64 | 87.44 | 85.87 | 89.56 |
| Late branch TV (%) | 88.68 | 94.13 | 94.14 | 97.06 |
| Saturation (%) | 28.44 | 56.18 | 43.63 | 73.65 |
| [99.375% CI] | [26.68, 30.24] | [53.77, 58.61] | [40.94, 46.35] | [71.02, 76.14] |
| Late − early (pp) | 6.03 | 6.69 | 8.27 | 7.50 |
| [99.375% CI] | [4.97, 7.19] | [5.57, 7.83] | [6.91, 9.80] | [6.12, 8.98] |
| Cohort $\tau \leq 64$ (%) | 51.25 | 88.10 | 57.24 | 90.07 |
| Contributing documents | 288 | 288 | 286 | 287 |

The 90% study prespecifies eight pooled endpoints: the full-$H$ fraction with $v_t > 0.99$ and the late-minus-early branch TV contrast, for each model and compressed action (Table 7). Intervals use $\alpha = 0.05/8$. All four branch comparisons meet the prespecified minimum of 24 documents and 128 trajectories, with defined denominators in every bootstrap draw. Saturation is measured continuously, without an outcome-selected threshold for declaring a regime.

The branch contrasts use 4,723/8,119 Llama trajectories and 5,275/8,301 Qwen trajectories (SnapKV/recent). Table 7 reports cohort sizes, absolute TV means, and all primary endpoints. High TV remains frequent in both studies: the first cohort's full-$H$ $v_t > 0.99$ fractions are 58.15%/44.92%/62.41% for Llama and 81.23%/65.53%/82.48% for Qwen (512/half/recent-half). Higher retention does not by itself isolate a post-divergence regime free from the bounded-TV ceiling.

Within the first study's 64-lag cohort, at least 16 consecutive pre-EOS agreements occur in 2.60%/7.60%/2.85% of Llama trajectories and 3.05%/7.32%/3.52% of Qwen trajectories (512/half/recent-half). A pre-EOS mismatch after an earlier agreement occurs in 76.88%/76.86%/74.34% and 71.20%/68.48%/65.11%, respectively. Denominators are the entire selected cohorts. These local agreements neither imply equal cache histories nor contradict increasing cohort-average TV.

### B.5 Where the discrepancy accumulates: EOS and task strata

Let $e_X, e_Y$ be the first EOS output positions. Four disjoint states partition generation positions: before $\min(e_X, e_Y)$, the first EOS output, one side already absorbed, and both sides absorbed. Their contributions $H^{-1}\mathbb{E}[\sum_t \mathbf{1}\{t \in S\}(v_t - f_t)]$ sum to $R - F$; selector differences give the additive partition in Table 8. The common $H$ denominator distinguishes these contributions from conditional pre-EOS means. The difference $v_t - f_t$ is zero through first divergence.

Equivalently, let $A_t$ indicate that neither path has emitted EOS at or before $t$. Then

$$R - F = \frac{1}{H}\sum_t \mathbb{E}[(v_t - f_t)A_t] + \frac{1}{H}\sum_t \mathbb{E}[(v_t - f_t)(1 - A_t)]. \tag{20}$$

The second term combines EOS-output and post-EOS positions; both terms retain the original horizon denominator.

**Table 8:** Additive EOS partition for recent-90 minus SnapKV-90, in pp. The both-absorbed contribution is zero. EOS share is the sum of the three termination-related contributions divided by $\Delta(R - F)$; shares are point estimates, without constructed ratio intervals.

| Model | $\Delta(R - F)$ | Pre-EOS | EOS output | One absorbed | EOS share |
|---|---|---|---|---|---|
| Llama | 28.5625 | 26.7181 | 0.0214 | 1.8230 | 6.46% |
| Qwen | 21.6271 | 21.5786 | 0.00095 | 0.04754 | 0.22% |

The pointwise 95% intervals for $\Delta(R - F)$ are $[26.66, 30.48]$ pp on Llama and $[19.58, 23.66]$ pp on Qwen; the pre-EOS components have intervals $[25.11, 28.31]$ and $[19.52, 23.61]$ pp. All 12 model–stratum gaps and their pre-EOS components have positive pointwise lower bounds (Table 9). EOS shares can be signed because they compare actions; they do not identify an EOS intervention effect. In the first cohort, the pooled pre-EOS components of 512-minus-half $\Delta(R - F)$ are 8.83 of 10.39 pp on Llama and 4.08 of 4.06 pp on Qwen; for recent-half minus half they are 10.65 of 12.14 pp and 4.70 of 4.70 pp. The Qwen HELMET–32K 512-minus-half gap is $-0.35\,[-2.84, 2.05]$ pp: its direction is unresolved despite the positive pooled comparison.

Within each action's fixed cohort $\mathcal{C} = \{(d, r) : \tau_{dr} \le 64\}$, define $A_{drj} = \mathbf{1}\{\tau_{dr} + j < \min(e_{X,dr}, e_{Y,dr})\}$. For lag window $W$, the strict pre-EOS mean is

$$\widehat{\mu}_W^{\mathrm{pre}} = \frac{\sum_{(d,r)\in\mathcal{C}} \sum_{j\in W} A_{drj} v_{dr,\tau_{dr}+j}}{\sum_{(d,r)\in\mathcal{C}} \sum_{j\in W} A_{drj}}, \qquad W \in \{1{:}16, 49{:}64\}.$$

Here $d, r$ index documents and replicates, and $v_{dr,t}$ is rollout TV. These means weight surviving steps rather than lag-specific means equally; document-bootstrap draws recompute both numerator and denominator. At 90% retention, the pooled late-minus-early contrasts are $5.94\,[5.14, 6.77]$ and $6.83\,[6.01, 7.65]$ pp for Llama, and $8.26\,[7.27, 9.32]$ and $7.51\,[6.51, 8.57]$ pp for Qwen

(SnapKV/recent; pointwise 95%). The effective early/late step fractions are 97.95%/94.87% for Llama SnapKV-90 and 98.63%/94.54% for recent-90; RULER late fractions are approximately 90%–92%. Qwen's late fractions are 99.77%/99.82% for SnapKV/recent. Thus pre-EOS means describe surviving steps, not a fixed set of survivors or an intervention preventing EOS.

**Table 9:** All 90% model–stratum recent-minus-SnapKV gaps in $R - F$ and their additive pre-EOS contributions. The intervals are pointwise 95%; EOS shares are point estimates without ratio intervals.

| Model / stratum | $\Delta(R - F)$, pp [95% CI] | Pre-EOS, pp [95% CI] | EOS share |
|---|---|---|---|
| Llama / HELMET–4K | 29.02 [25.63, 32.22] | 27.95 [24.74, 31.08] | 3.680% |
| Llama / HELMET–16K | 29.53 [27.12, 31.97] | 29.27 [26.90, 31.62] | 0.863% |
| Llama / HELMET–32K | 31.18 [28.67, 33.60] | 30.16 [27.63, 32.62] | 3.277% |
| Llama / RULER–4K | 27.43 [21.47, 33.34] | 24.24 [19.46, 28.83] | 11.647% |
| Llama / RULER–16K | 28.75 [22.35, 35.07] | 26.01 [20.88, 31.05] | 9.521% |
| Llama / RULER–32K | 25.46 [19.86, 30.93] | 22.67 [18.06, 27.20] | 10.955% |
| Qwen / HELMET–4K | 24.95 [21.53, 28.39] | 24.95 [21.53, 28.39] | 0.000% |
| Qwen / HELMET–16K | 22.41 [18.63, 26.22] | 22.41 [18.63, 26.22] | 0.000% |
| Qwen / HELMET–32K | 21.74 [18.81, 24.73] | 21.74 [18.80, 24.73] | 0.024% |
| Qwen / RULER–4K | 16.65 [11.18, 22.02] | 16.44 [10.95, 21.83] | 1.251% |
| Qwen / RULER–16K | 21.87 [14.44, 29.54] | 21.78 [14.35, 29.48] | 0.380% |
| Qwen / RULER–32K | 22.15 [16.59, 27.72] | 22.15 [16.53, 27.74] | -0.024% |

### B.6 Temporal structure across task strata

Table 10 gives the full-branch counterparts and numerical values of the pre-EOS intervals in Figure 6. Denominators are defined in all 10,000 bootstrap draws. These pointwise intervals do not provide simultaneous coverage.

**Table 10:** All 90% stratum-specific late (lags 49–64) minus early (lags 1–16) TV contrasts. Full-branch and strict pre-EOS columns use action-specific $\tau \leq 64$ cohorts; pre-EOS means weight surviving steps within each window. Intervals are descriptive, pointwise 95%.

| Model / stratum | Action | Full branch, pp [95% CI] | Pre-EOS, pp [95% CI] |
|---|---|---|---|
| Llama / HELMET–4K | SnapKV | 4.21 [2.51, 6.10] | 4.18 [2.47, 6.10] |
| Llama / HELMET–4K | recent | 6.18 [4.73, 7.74] | 6.21 [4.77, 7.79] |
| Llama / HELMET–16K | SnapKV | 7.43 [5.27, 9.87] | 7.21 [5.02, 9.67] |
| Llama / HELMET–16K | recent | 5.01 [3.90, 6.16] | 5.02 [3.93, 6.15] |
| Llama / HELMET–32K | SnapKV | 8.48 [6.69, 10.18] | 8.44 [6.64, 10.12] |
| Llama / HELMET–32K | recent | 6.28 [4.96, 7.71] | 6.36 [5.06, 7.77] |
| Llama / RULER–4K | SnapKV | 5.01 [3.13, 7.03] | 4.85 [2.87, 6.94] |
| Llama / RULER–4K | recent | 7.26 [4.36, 10.38] | 7.38 [4.46, 10.56] |
| Llama / RULER–16K | SnapKV | 5.39 [3.59, 7.52] | 5.41 [3.55, 7.61] |
| Llama / RULER–16K | recent | 8.54 [6.10, 11.10] | 8.63 [6.05, 11.35] |
| Llama / RULER–32K | SnapKV | 5.49 [4.07, 7.02] | 5.16 [3.60, 6.81] |
| Llama / RULER–32K | recent | 7.16 [5.18, 9.16] | 7.54 [5.61, 9.54] |
| Qwen / HELMET–4K | SnapKV | 6.60 [4.85, 8.45] | 6.60 [4.85, 8.45] |
| Qwen / HELMET–4K | recent | 6.08 [3.86, 8.87] | 6.08 [3.86, 8.87] |
| Qwen / HELMET–16K | SnapKV | 11.46 [8.19, 15.12] | 11.46 [8.19, 15.12] |
| Qwen / HELMET–16K | recent | 4.86 [3.39, 6.47] | 4.86 [3.39, 6.47] |
| Qwen / HELMET–32K | SnapKV | 9.12 [7.40, 10.88] | 9.12 [7.40, 10.88] |
| Qwen / HELMET–32K | recent | 7.23 [4.79, 10.25] | 7.23 [4.79, 10.25] |
| Qwen / RULER–4K | SnapKV | 7.25 [4.48, 10.57] | 7.21 [4.42, 10.54] |
| Qwen / RULER–4K | recent | 8.81 [6.63, 11.28] | 8.85 [6.67, 11.32] |
| Qwen / RULER–16K | SnapKV | 7.01 [5.10, 9.03] | 7.04 [5.12, 9.07] |
| Qwen / RULER–16K | recent | 10.00 [6.71, 13.81] | 10.00 [6.70, 13.81] |
| Qwen / RULER–32K | SnapKV | 7.81 [5.94, 9.59] | 7.80 [5.92, 9.58] |
| Qwen / RULER–32K | recent | 8.68 [6.60, 10.92] | 8.67 [6.59, 10.91] |

Generation-position blocks (1–32, 33–64, 65–128) are normalized by their respective lengths. Pooled recent-90 minus SnapKV-90 gaps are 37.89, 41.35, and 30.11 pp on Llama and 36.09, 37.63, and

25.53 pp on Qwen. The final gap is smaller than the initial gap in every HELMET model–length cell and larger in every RULER cell (Table 11); these descriptive contrasts have no intervals or heterogeneity test. Even in the first cohort, whose pooled profiles decline (Figure 9), the RULER–4K gap rises in the middle block. Pooled trends therefore need not characterize each task or each time interval.

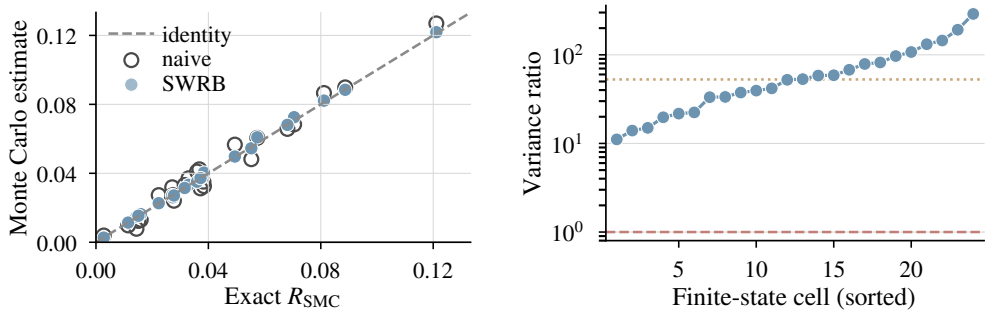
**Table 11:** Observed generation-clock gap changes across benchmark–length strata at 90% retention: the recent-minus-SnapKV gap at positions 65–128 minus its gap at positions 1–32, in pp. These are descriptive point estimates.

| Stratum | Llama | Qwen |
|---|---|---|
| HELMET–4K | −21.72 | −25.84 |
| HELMET–16K | −18.14 | −26.36 |
| HELMET–32K | −14.34 | −19.54 |
| RULER–4K | +3.25 | +0.27 |
| RULER–16K | +0.50 | +0.28 |
| RULER–32K | +3.74 | +7.84 |

# C Measurement Validation and Supplementary Depth Profiles

## C.1 Exact finite-state validation and computational cost

The conditional estimator can be checked against exact expectations in small systems. Figure 10 compares ordinary coupled Monte Carlo and SWRB using 512 replicates in each of 24 binary, horizon-three systems. Mean absolute error decreases from 0.003928 to 0.000659, and Spearman correlation with exact risk increases from 0.9513 to 0.9974. The naive-to-SWRB variance ratio ranges from 11.1 to 289, with median 52.7. Mean kernel-pair cost is 1.37 times that of ordinary coupling; the ratio-of-means variance–cost advantage is 20.0. These results concern enumerable systems and kernel evaluations, whereas Transformer execution includes different costs for scans, replay, and suffix generation.




(a) Monte Carlo means versus exact risk.

(b) Variance comparison across systems.

**Figure 10:** Exact finite-state validation, with 512 replicates per estimator in each of 24 binary systems at $H = 3$. Conditional measurement reduces variance in every system; these per-replicate results do not determine Transformer efficiency.

A separate development screen uses Llama-3.1-8B and Qwen3-8B, 48 documents per model (eight per environment–length cell), SnapKV budgets 256/512/1024, and $H = 128$ at temperature one. Each document–action cell has four replicates per estimator and an independent 16-replicate SWRB reference mean. The statistic averages each replicate's squared error to this noisy reference multiplied by its estimator wall time; it differs from the product of separately averaged error and time. FullKV controls are excluded. SWRB-to-naive ratios are 2.38 on Llama and 0.897 on Qwen3. Timing excludes prefill and initial cache compaction and includes the SWRB scan, prefix replay, and residual suffix. Timing did not explicitly synchronize GPU execution. The mixed results therefore do not establish a compute-normalized efficiency advantage.

### C.2 Real-model observation-depth measurements

This supplementary study uses FullKV as reference and persistent SnapKV actions retaining 384, 512, or 768 prompt entries. The six HELMET/RULER environment–length cells span nominal 4K, 16K, and 32K contexts (Yen et al., 2025; Hsieh et al., 2024). Each cell contributes four training and four evaluation documents, giving 24 documents per role, shared across the same Llama and Qwen Instruct checkpoints used in the trajectory studies. Training, evaluation, and implementation-check subsets are disjoint; historical-study identities and exact-text duplicates are excluded. The training role supports the supplementary application; the profiles below use only the evaluation documents.

Generation uses temperature-one full-softmax distributions, $H = 128$, and $K = 8$ paired SWRB replicates. EOS is absorbing, preserving a fixed $H$-position outcome. Each cache advances on its own generated history and absolute positions, retaining the selected prompt entries and appending generated entries. Each sampled branch records the full suffix $(I_J, \ldots, I_H)$, including its first mismatch. Thus all depths use the same complete overlap scan, suffix, and denominator $H$. Varying $L$ changes the observed portion of a full trace; it is not a comparison of separately run, shorter simulations.

Figure 7 locates the observed window and actual tail within the estimated capacity at $L = 32$. Under SnapKV-512, estimated divergence probabilities are $0.98455$ and $0.99965$ for Llama and Qwen. At $L = 32$, their window risks are $22.702 \pm 0.299$ and $22.827 \pm 0.210$ pp, remaining capacities $66.194 \pm 0.552$ and $67.234 \pm 0.126$ pp, and full risks $85.830 \pm 1.042$ and $88.818 \pm 0.492$ pp (one MC SE). Across all three actions, window risks span 22.3–23.1 pp and full risks 84.0–88.8 pp. Most measured occupation lies beyond this window. Shared traces induce covariance, so the marginal SEs alone do not determine uncertainty in window–tail differences. The accompanying anonymous research materials provide complete numerical profiles, study specifications, and implementation details.

## D Complementary Application and Scope

This application uses predicted mismatch reductions to allocate a fixed cache budget across documents, comparing heterogeneous budgets rather than equal-budget selectors.

**Design and target.** The historical assignment study uses 174 development and 360 previously unevaluated documents, balanced across six HELMET/RULER environment–length cells (4K, 16K, and 32K). Persistent SnapKV actions retain 384, 512, or 768 prompt tokens; FullKV is the reference. Measurements use $H = 128$, temperature-one full-softmax generation, absorbing EOS, and eight paired SWRB replicates per document and action. Predictors and checkpoint eligibility are selected before evaluation. Llama and Qwen receive adaptive assignments; Qwen3-8B and Mistral-7B-Instruct-v0.3 retain uniform-512. Each 30-document batch exactly matches uniform-512's intervention-start resident bytes (compact K/V, persistent indices, and metadata); temporary buffers and generated K/V are accounted for separately. Documents are the statistical units.

For fixed documents, predictors, batches, and assignments $A_{md}$, the equally weighted $C$-checkpoint target is

$$\theta_A = \frac{1}{CN} \sum_{m=1}^{C} \sum_{d=1}^{N} \big\{ \mu_m(x_d, 512) - \mu_m(x_d, A_{md}) \big\}, \qquad \mu_m(x, a) = R_{\mathrm{SMC}}(P_m, Q_m^a \mid x). \quad (21)$$

Positive values indicate reduced mismatch. Conditional Monte Carlo (MC) intervals average paired action/checkpoint differences within replicates across independent document streams, holding training and assignment fixed. Document-composition reweighting changes cohort weights; its descriptive endpoints are not MC intervals.

**Fixed-assignment effect and controlled comparisons.** The four-checkpoint reduction is $+0.2016$ percentage points (pp) (approximate conditional 95% MC interval $[+0.0699, +0.3333]$ pp). Two uniform branches contribute exact zeros; adaptive effects are $+0.6124$ pp on Llama and $+0.1941$ pp on Qwen2.5. The finite-sample Hoeffding interval for independent checkpoint–document streams is $[-1.2731, +1.6763]$ pp. Post-divergence occupation accounts for 99.6% of the pooled point estimate, combining exposure and conditional mismatch rate without attributing improvement to either.

**Table 12:** Assignment and controlled-comparison effects on full-$H$ mismatch (pp). "State/$H$ vs. comparator" is the difference between their reductions relative to uniform-512. Intervals are approximate conditional 95% MC intervals. Historical effects average four checkpoints (two adaptive); depth-study effects average two adaptive checkpoints.

| Study | Contrast | Effect | 95% MC interval |
|---|---|---|---|
| Historical | State/$H$ vs. uniform-512 | +0.2016 | [+0.0699,+0.3333] |
| | State/$H$ vs. Cell/$H$ | -0.003 | [-0.098,+0.091] |
| | State/$H$ vs. Mass/$H$ | +0.018 | [-0.069,+0.106] |
| | State/$H$ vs. State/occurrence | +0.039 | [-0.040,+0.118] |
| Depth | State/$H$ vs. uniform-512 | +0.618 | [-0.091,+1.327] |
| | State/$H$ vs. Mass/$H$ | +0.379 | [-0.368,+1.126] |
| | State/$H$ vs. State/$L = 8$ | +0.376 | [-0.603,+1.355] |
| | State/$H$ vs. State/$L = 32$ | +0.927 | [-0.084,+1.937] |

Controlled candidates share ridge ($\alpha = 10$), training-only standardization, assignment algorithm, byte budget, and full-$H$ evaluation. Cell features encode environment and action; Mass adds pre-decision removed-attention information; State adds state and selector summaries. Historical feature comparisons are retrospective. The depth study uses 24 training and 24 evaluation documents, disjoint from the historical cohorts and shared across Llama and Qwen2.5. Nine candidates are fixed before evaluation; the three focal comparisons are State/$H$ against Mass/$H$, State/$L = 8$, and State/$L = 32$. The three intervals span zero (Table 12), without establishing incremental assignment benefits from state features or full-horizon supervision. Depth-study State/$H$ effects differ across checkpoints ($-0.545$ pp on Llama; $+1.781$ pp on Qwen2.5), and concentrated variance makes its small-cohort normal approximation tentative.

**Scope of the application.** An earlier task-table strategy on a separate 120-document cohort had a positive pooled estimate ($+0.4589$ pp) but failed its prespecified criterion (composition lower endpoint $-0.0812$ pp). Under the later fixed assignment, greedy generation on the same 360 documents yields a pooled native-task score change of $+0.2703$ pp (composition lower endpoint $-0.1143$ pp). Scores combine RULER string match and HELMET task-specific F1/ROUGE-L. This meets the prespecified $-1$-pp preservation margin, but not the secondary superiority criterion; the two zero branches make the margin equivalent to $-2$ pp on the adaptive-checkpoint mean. The anonymous research materials contain complete candidate results, earlier positive and negative outcomes, composition and label sensitivities, and implementation specifications.