INFORMATION LIMITS AND OUTPUT GEOMETRY FOR QUANTIZED ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A certificate can reject a quantized attention output even when its actual error is small. We characterize how retained probability and value information determines the tightest attainable error bound. Our exact scalar output envelope combines range, mean, and variance constraints: support bounds clip the classical variance-only optimizer, and two-point laws attain every regime. Joint constraints can certify budgets that separate moment bounds cannot. A distinct construction gives a $\Theta(n)$ separation from the output conversion of the optimal total-variation (TV) bound. We also derive attainable TV envelopes for selected exact probabilities and lower bounds, with saturation and stopping criteria. Outward moments and a sharp correction for perturbed-distribution value error connect the envelope to quantized attention. On a fixed GPT-2 replay, an exact CPU reference certifies 357 of 1,008 case/budget pairs, versus 249 for midpoint-centered TV and 352 for coordinatewise mixing of variance-only and range/mean bounds. Independent numerical enclosures validate every certified pair. These comparisons distinguish the benefit of value information from the additional gain of joint constraints.

1 INTRODUCTION

Consider a language model generating the next token from a conversation. Each attention head reads cached keys to assign weights to earlier tokens, then combines their cached value vectors. Storing the cache at lower precision saves memory, but can change both the weights and the values being combined. For a head whose output must meet an error tolerance, this creates a concrete decision: can the system use the quantized result for the current query, or must it return to the original cache?

A runtime certificate makes this decision by bounding the output error. It accepts the quantized result when the bound fits the tolerance and otherwise uses the original cache (Calver, 2026; Wei et al., 2026). Yet a valid bound can exceed the tolerance even when the actual error is small. The question is therefore how much error can be ruled out using the information retained about the attention weights and values.

One approach first bounds changes in attention probabilities, then converts that bound to output error using the value range. Total variation (TV) measures this probability change. Its sharp probability-independent bound is known (Cohen & Fausti, 2024), and selected probabilities can tighten it when they exclude the worst case. But output error also depends on which values receive the changed weights: changes can cancel in the weighted sum. Thus a tight probability bound can still give a conservative output bound. This leads to two questions: what is the strongest output bound supported by a compact value summary, and when can further probability information improve a certificate?

We answer these questions by considering all configurations consistent with a retained summary, an *information class*. Its exact envelope is the smallest bound valid for every configuration in that class. We first derive this envelope from value range, mean, and variance, then characterize the limits of probability refinement. Figure 1 connects these results to the certification pipeline. Our contributions are:

- **An attainable output envelope.** Joint range, mean, and variance constraints yield an exact scalar optimum and a rational budget test. One construction shows a strict gain over separate moment bounds; another gives a $\Theta(n)$ separation from centered TV.

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.

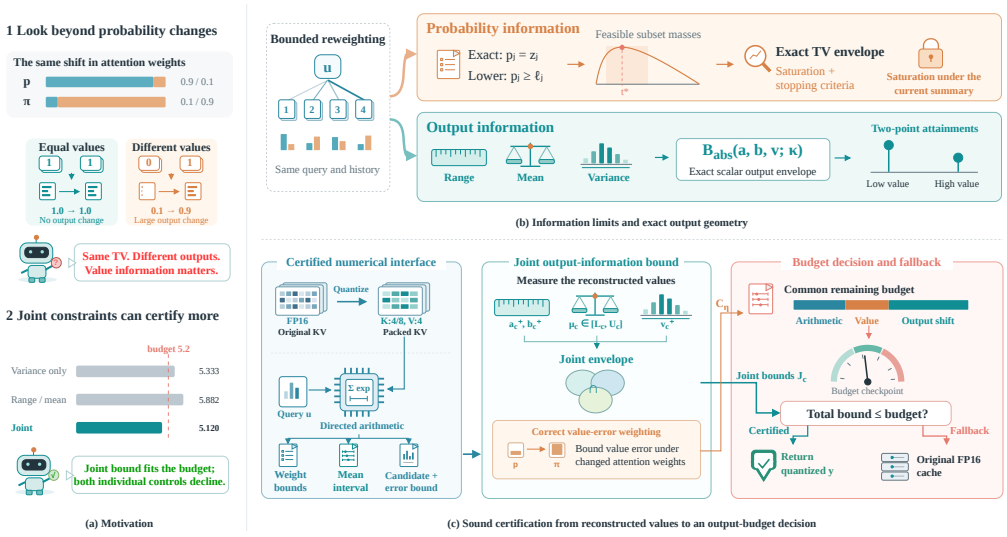

Figure 1: **Information-aware certification for quantized attention.** (a) Value structure reduces output uncertainty beyond probability-based bounds. (b) Probability summaries and joint value moments define the information available to the certificate. (c) Outward numerical bounds yield the acceptance test $E_{\text{arith}} + C_\eta + \max_c J_c \le \varepsilon$; the original-cache fallback is checked separately.

- **Limits of probability refinement.** Exact masses and lower bounds yield distinct attainable TV envelopes. Saturation and dominant-mass criteria identify when retained information cannot tighten the bound.
- **A certified attention interface.** Outward moments and a corrected value-error term connect the envelope to numerical attention. A fixed CPU replay certifies 357/1,008 pairs, versus 249 for centered TV and 352 for coordinatewise mixing.

Related work. Cohen & Fausti (2024) establish the fixed-probability TV expression and critical subset mass. Calver (2026) develop compressed storage, metadata, and fallback; Wei et al. (2026) also study probability-weighted witnesses and a probabilistic dither tier. Top-k truncation certificates bound omitted mass (Tzachristas et al., 2025), whereas our perturbation model retains all atoms and changes their relative weights. Response-geometric minimax compression reduces the number of key–value pairs (Haverbeck et al., 2026). The variance-only output bound is the worst-case expectile of Jakobsons & Vanduffel (2015, Theorem 5); the bounded-variable variance inequality of Bhatia & Davis (2000) provides a complementary control. We characterize the joint support-constrained optimum and exact decisions under partial probability information.

2 Problem Setup and Probability Perturbations

We study the decision for one query $\mathbf{u}$, one attention head, and n cached tokens. Let K, V be the original cache and $\widehat{K}, \widehat{V}$ its numerical reconstruction, with per-head dimension d. We certify a returned vector y against the exact-real output $O = \sum_i \pi_i V_i$ on the original FP16 inputs, requiring $\|y - O\|_\infty \le \varepsilon$. Define the exact-real scores and probabilities

$$s_i = \mathbf{u}^\top \widehat{K}_i / \sqrt{d}, \qquad p_i = \frac{e^{s_i}}{\sum_j e^{s_j}}, \qquad \pi_i = \frac{e^{s_i+\delta_i}}{\sum_j e^{s_j+\delta_j}}, \quad |\delta_i| \le \Delta. \tag{1}$$

Here π uses the original keys; p uses the reconstructed keys. Both refer to the same query and history. Setting $\kappa = e^{2\Delta}$ and removing a common factor gives the uncertainty model

$$T_x(p)_i = \frac{p_i x_i}{\sum_j p_j x_j}, \qquad x \in [1, \kappa]^n, \qquad \mathrm{TV}(p, q) = \tfrac{1}{2} \sum_i |p_i - q_i|. \tag{2}$$

This model is equivalent to perturbations in $[-\Delta, \Delta]^n$. Optimality below refers to (2); a physical cache may impose further coupling. A certified upper bound may replace κ.

From probability vectors to subset masses. The worst TV change reduces to subset masses. For $t \in [0, 1]$, define

$$f_\kappa(t) = \frac{(\kappa - 1)t(1 - t)}{1 + (\kappa - 1)t}, \qquad \mathcal{E}(p, \kappa) = \max_{x \in [1,\kappa]^n} \mathrm{TV}(p, T_x(p)). \tag{3}$$

The fixed-probability formula of Cohen & Fausti (2024, Lemma 5.2), in the half-ℓ_1 convention, is

$$\mathcal{E}(p, \kappa) = \max_{A \subseteq [n]} f_\kappa\left(\sum_{i \in A} p_i\right). \tag{4}$$

Indeed, $\mathrm{TV}(p, q) = \max_A(q(A) - p(A))$. For fixed A, its largest excess puts $x_i = \kappa$ in A and $x_i = 1$ outside, giving (4). For $\kappa > 1$, $f'_\kappa(t) = \kappa/[1 + (\kappa - 1)t]^2 - 1$, so its unique maximum occurs at

$$t_* = (1 + \sqrt{\kappa})^{-1}, \qquad U_\kappa = f_\kappa(t_*) = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}. \tag{5}$$

At $\kappa = 1$, both envelopes are zero. The following sections determine how value moments and partial probability information refine this baseline.

3 An Exact Output-Information Boundary

The same probability change can produce different output errors depending on the values it reweights. We first bound this interaction using value scatter, then derive an exact bound from joint moment constraints. Define

$$R_* = \tfrac{1}{2} \max_c(\max_i V_{ic} - \min_i V_{ic}), \qquad S = \max_c \sum_i (V_{ic} - \overline{V}_c)^2, \quad \overline{V}_c = \tfrac{1}{n} \sum_i V_{ic}. \tag{6}$$

Here R_* is the midpoint-centered radius and S is the unweighted value scatter. A TV bound B gives the baseline output bound $2R_*B$.

Theorem 1 (Output scatter control). *For $q = T_x(p)$, $x_i \in [1, \kappa]$ and $\omega \geq \max_i p_i$,*

$$\left\| \sum_i (q_i - p_i) V_i \right\|_\infty^2 \leq \omega S(\kappa - 1)^2/(4\kappa). \tag{7}$$

Every rational $\kappa > 1$ has a finite uniform-probability equality example.

This universally sharp coefficient need not be exact for a fixed value matrix. The runtime normalizer bound Z^- supplies $\omega^+ = \min\{1, 1/Z^-\}$; outward R_*^+, S^+ use eight metadata bytes. Appendix G gives the proof and write-time construction.

The joint moment problem. For one coordinate, center values at their probability-weighted mean: $Y = V_{ic} - \sum_j p_j V_{jc}$ with probability p_i, and let $X = x_i$ on the same atom. Then $\mathbb{E}Y = 0$ and the output change is $\mathbb{E}[XY]/\mathbb{E}X$. We maximize it given support bounds a, b and variance bound v; clip projects onto the specified interval.

Theorem 2 (Joint scalar output envelope). *Let $a, b, v \geq 0$ and $\kappa \geq 1$. Over all finite joint laws satisfying $\mathbb{E}Y = 0$, $-a \leq Y \leq b$, $\mathbb{E}Y^2 \leq v$ and $1 \leq X \leq \kappa$, the maximum of $\mathbb{E}[XY]/\mathbb{E}X$ is zero if $abv(\kappa - 1) = 0$. Otherwise put*

$$w = \min\{v, ab\}, \quad t_b = \frac{w}{b^2 + w}, \quad t_a = \frac{a^2}{a^2 + w}, \quad \widehat{t} = \mathrm{clip}\left(\frac{1}{\kappa + 1}, [t_b, t_a]\right). \tag{8}$$

The exact positive and absolute envelopes are

$$B_+(a, b, v; \kappa) = \frac{(\kappa - 1)\sqrt{w\widehat{t}(1 - \widehat{t})}}{1 + (\kappa - 1)\widehat{t}}, \qquad B_{\mathrm{abs}} = \max\{B_+(a, b, v; \kappa), B_+(b, a, v; \kappa)\}. \tag{9}$$

Each direction is attained by a two-point law of variance w.

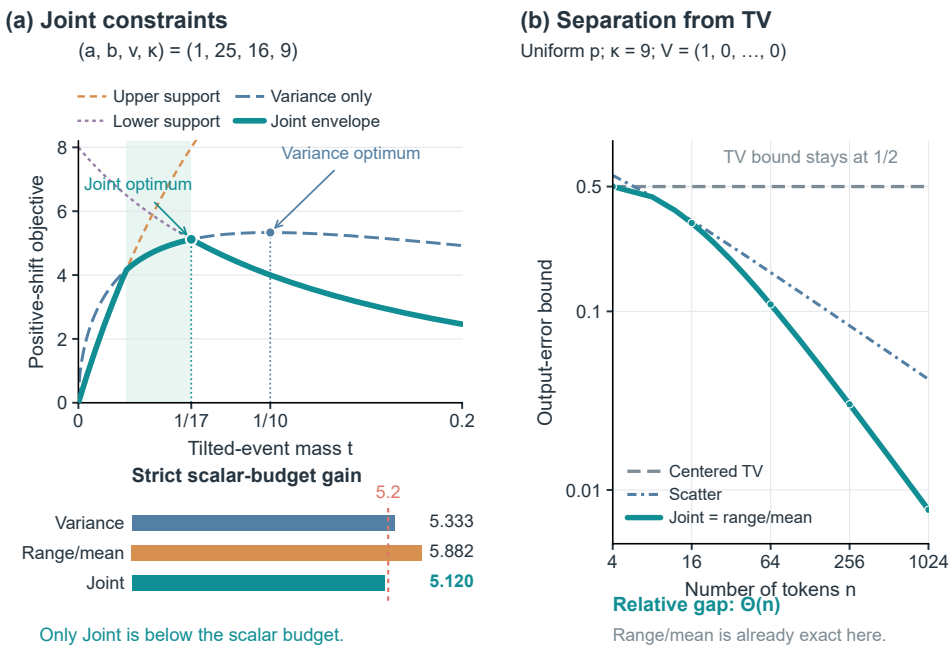

Figure 2: **Joint output bounds: mechanism and separation.** (a) For $(a, b, v, \kappa) = (1, 25, 16, 9)$, support constraints move the optimizer from $1/10$ to $1/17$; only Joint meets budget $26/5$. Shading marks variance-saturating two-point laws compatible with the support. (b) For uniform p, $4 \mid n$, $\kappa = 9$, and $V = (1, 0, \ldots, 0)$, the exact output bound separates from centered TV by $\Theta(n)$. Range/mean is already exact in (b). All curves are analytical.

The optimum allows zero mass at support endpoints and can exceed that for fixed (p, V). An endpoint tilt sets $X = \kappa$ on an event A of mass t. Writing $m = \mathbb{E}[Y 1_A]$ gives

$$m \le \min\{bt, a(1-t), \sqrt{wt(1-t)}\}. \tag{10}$$

For $0 < t < 1$, centering gives $\mathbb{E}[Y 1_{A^c}] = -m$. Conditional Cauchy–Schwarz yields

$$\mathbb{E}Y^2 \ge m^2/t + m^2/(1-t) = m^2/[t(1-t)]. \tag{11}$$

Combining this with the support inequalities proves (10). After multiplication by $(\kappa-1)/[1+(\kappa-1)t]$, the upper-support branch increases, the variance branch peaks at $1/(\kappa+1)$, and the lower-support branch decreases. Enforcing all three constraints at the same mass gives the clipping rule; maximizing them separately can require incompatible masses. Appendix H gives the full proof.

Strict improvement over separate moment bounds. For a fixed law, the largest tilted expectation is its expectile at $\alpha = \kappa/(\kappa+1)$. Without binding support constraints, (9) recovers the worst-case variance bound $\sqrt{v}(\kappa-1)/(2\sqrt{\kappa})$ of Jakobsons & Vanduffel (2015). Clipping captures the gain from support information. For $(a, b, v, \kappa) = (1, 25, 16, 9)$, the absolute joint bound $128/25$ passes budget $26/5$, while variance alone gives $16/3$ and range/mean alone gives $100/17$; both decline (Figure 2a). Mass $1/17$ at 16 and $16/17$ at -1 attains the joint bound.

Separation from probability-based bounds. For uniform p, $4 \mid n$, $\kappa = 9$, and values $V_1 = 1$, $V_{i>1} = 0$, full-information TV and its centered output conversion both equal $1/2$. Here $(a, b, v) = (1/n, (n-1)/n, (n-1)/n^2)$ and the joint bound equals the actual maximum $8(n-1)/(n(n+8))$, an order-n relative separation. Scatter gives $4\sqrt{n-1}/(3n)$. Here $v = ab$, so range/mean alone is exact: this family isolates the benefit of output information, while the preceding example isolates the benefit of joint constraints (Figure 2).

Rational budget decision. To test a budget $r \geq 0$, first return zero in the degenerate cases of Theorem 2. Otherwise let $d_\kappa = \kappa - 1$. In each direction,

$$B_+ = \begin{cases} d_\kappa bw/(b^2 + \kappa w), & b^2 < \kappa w, \\ d_\kappa aw/(w + \kappa a^2), & w > \kappa a^2, \\ d_\kappa \sqrt{w}/(2\sqrt{\kappa}), & \text{otherwise.} \end{cases} \tag{12}$$

The first two conditions cannot both hold since $w \leq ab$. Compare rational products in those branches, or $d_\kappa^2 w \leq 4\kappa r^2$ in the third; equality accepts. Repeat with a, b swapped. Class inclusion proves monotonicity in a, b, v, κ, permitting valid outward bounds. The exact reference uses arbitrary-precision rational arithmetic; its cost also depends on operand bit length.

4 Probability information and stopping criteria

We now ask how far probability information alone can tighten the bound, without constructing query-weighted value moments. The answer depends on what is retained: exact masses fix $p_j = z_j$ on h selected coordinates, whereas lower bounds $\ell_j \leq p_j$ allow extra mass on those coordinates. For the lower-bound class, put $\eta = 1 - \sum_j \ell_j$ and $s_J = \sum_{j \in J} \ell_j$. Every compatible subset mass lies in $[s_J, s_J + \eta]$ and every point of that interval is attainable in the nonnegative class. Consequently,

$$\mathcal{B}_{\geq} = \max_{J \text{ admissible}} f_\kappa\big(\mathrm{clip}(t_*, [s_J, s_J + \eta])\big). \tag{13}$$

Here $n \geq 2$ and $\kappa > 1$. All J are admissible when $h < n$; otherwise omit the empty and full subsets. Theorems 6 and 7 give the full information classes and exact-mass rules. For example, with two atoms and $\kappa = 9$, $p_1 = 1/2$ fixes the uniform distribution and gives TV $2/5$, whereas $p_1 \geq 1/2$ permits $p = (3/4, 1/4)$ and TV $1/2$ (Figure 3a).

Saturation and refinement. An interval containing t_* permits the unrestricted worst case: the bound *saturates*. Additional constraints may remove that case. The next two results distinguish saturation from a regime where other lower bounds cannot help (Figure 3b,c).

Corollary 3 (Saturation)**.** *Under Theorem 7, $\mathcal{B}_{\geq}(\ell, n; \kappa) = U_\kappa$ if and only if an admissible interval $[s_J, s_J + \eta]$ contains t_*. Equivalently, with*

$$D_\kappa(t) = \kappa - [1 + (\kappa - 1)t]^2, \tag{14}$$

there is an admissible J for which $D_\kappa(s_J) \geq 0$ and $D_\kappa(s_J + \eta) \leq 0$. If $h < n$ and $\eta \geq t_*$, saturation is unavoidable.

The unique maximizer of f_κ proves the criterion; $J = \varnothing$ proves the last case. Total lower mass above $1 - t_*$ is necessary but insufficient for tightening: a nonempty subset may still contain t_* in its feasible interval.

Corollary 4 (Stop collecting other lower masses)**.** *Under Theorem 7, suppose a retained coordinate j has $\ell_j > 1 - t_*$. Then, for any valid lower bounds on the other coordinates,*

$$\mathcal{B}_{\geq}(\ell, n; \kappa) = f_\kappa(1 - \ell_j). \tag{15}$$

Thus adding or tightening other coordinatewise lower bounds cannot improve this optimum while ℓ_j and κ remain fixed.

Subsets omitting j have mass at most $1 - \ell_j < t_*$; those containing it have mass at least ℓ_j. Taking $p_j = \ell_j$ and tilting its complement attains the bound (Appendix C.3). The premise equivalently reads $D_\kappa(1 - \ell_j) > 0$. Tightening ℓ_j or adding an upper constraint can still help. Outside this regime, the lower constraint $\ell_1 = 2/5$ at $\kappa = 9$ permits TV $1/2$, whereas the complete lower vector $(2/5, 3/10, 3/10)$ gives $42/85$.

For $p = (99/100, 1/200, 1/200)$ and $\kappa = 9$, retaining $\ell_1 = 99/100$ gives $11/150 < 1/10$, certifying the TV budget in Figure 3c. A uniform four-atom vector instead attains U_9 even with full information.

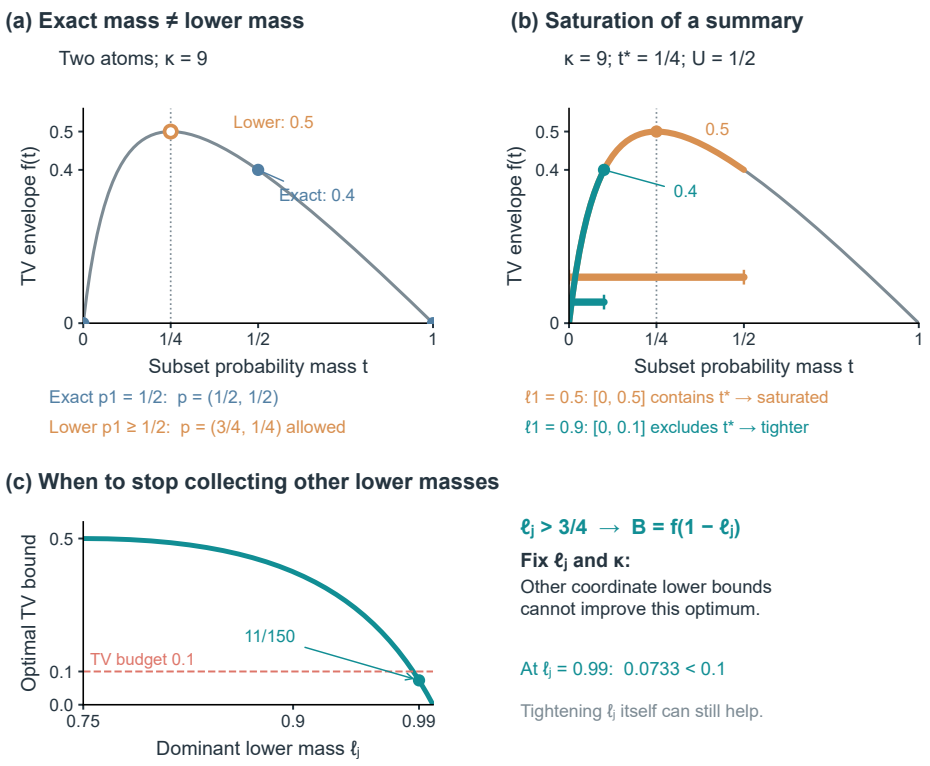

Figure 3: **Probability information and stopping criteria** ($\kappa = 9$). (a) An exact mass of $1/2$ gives TV $2/5$ for two atoms; a lower bound of $1/2$ permits TV $1/2$. (b) A feasible interval containing $t_* = 1/4$ saturates the current summary. Horizontal segments show mass intervals; their heights are illustrative. (c) With $\ell_j > 3/4$, other coordinate lower bounds cannot improve $f_9(1 - \ell_j)$ while ℓ_j remains fixed. At $\ell_j = 0.99$, the bound meets budget $1/10$.

4.1 A RATIONAL BUDGET GATE

For a target $\tau \geq 0$, positivity of the denominator in (3) gives

$$f_\kappa(t) \leq \tau \quad \Longleftrightarrow \quad Q_\tau(t) := (\kappa - 1)t(1 - \tau - t) - \tau \leq 0. \tag{16}$$

The concave quadratic Q_τ has vertex $v_\tau = (1 - \tau)/2$. Although this differs from t_*, it answers the same budget question without a square root.

Proposition 5 (Exact budget gate)**.** *For $n \geq 2$ and $\kappa > 1$, $\mathcal{B}_{\geq}(\ell, n; \kappa) \leq \tau$ if and only if, for every admissible J,*

$$Q_\tau\big([v_\tau]_{[s_J, s_J + \eta]}\big) \leq 0. \tag{17}$$

All comparisons are rational for rational inputs, including equality. The cases $n = 1$, $\kappa = 1$, or $\tau \geq 1$ are immediate.

Depth-first traversal assigns lower masses to included or excluded sums, testing (17) on $[s_J, 1 - s_{J^c}]$ at each leaf. It uses $O(2^h)$ scalar operations and $O(h)$ auxiliary memory, excluding summary construction and rational bit costs. The implementation first tests $[0, 1]$ and declines unresolved comparisons (Appendix D). Timings cover this recursion; the stopping corollary guides summary collection.

Complete exact evaluation on rational probabilities is weakly NP-hard; Appendix B gives the reduction. The bounded-h gate above instead exploits the size of the retained summary.

5 From information bounds to certified attention

Certifying a computed output also requires bounds on arithmetic and value-reconstruction errors. The reference stores INT4/INT8 keys and INT4 values with FP16 scales. Upward metadata bounds key and value residuals after clipping and scale rounding. Directed score, exponential, and output calculations provide $\bar{\kappa}$, a denominator lower bound Z^-, and coordinate intervals $[L_c, U_c]$ for $\widehat{O} = \sum_i p_i \widehat{V}_i$. An FP32 representative y has arithmetic error at most E_{arith} relative to $\widehat{O}$. Appendix F specifies the numerical path and the probability gate.

Let $\eta_i \geq \|V_i - \widehat{V}_i\|_\infty$ and obtain $E_{\mathrm{val}} \geq \sum_i p_i \eta_i$ from the same outward exponential reduction. The original-value centered-TV control inserts $\sum_i p_i V_i$ and uses $E_{\mathrm{arith}} + E_{\mathrm{val}} + 2R_* B$. The moment bound instead uses reconstructed values, so its residual term must account for the change from p to π.

Correcting the value-error term. For $z_i = \widehat{V}_{ic}$, let $\mu_c = \sum_i p_i z_i \in [L_c, U_c]$. The range and outward second moment supply a_c^+, b_c^+, v_c^+ for Theorem 2. Decomposing the error through $\sum_i \pi_i \widehat{V}_i$ requires a bound on $\sum_i \pi_i \eta_i$. With $H = \max_i \eta_i$ and $e = \min\{H, E_{\mathrm{val}}\}$, the sharp scalar correction is

$$C_\eta = \frac{\bar{\kappa} e H}{H + (\bar{\kappa} - 1)e}, \qquad \|y - O\|_\infty \leq E_{\mathrm{arith}} + C_\eta + \max_c B_{\mathrm{abs}}(a_c^+, b_c^+, v_c^+; \bar{\kappa}), \tag{18}$$

with $C_\eta = 0$ when $H = 0$. Appendix H proves the correction and its attainability.

Outward moments and cost. Intersecting the mean enclosure with the reconstructed range gives $a^+ = U - z_-$ and $b^+ = z_+ - L$. A centered second-moment upper bound, minus a lower bound on the squared mean offset, gives v^+; Appendix H specifies the construction. The CPU reference adds an $O(nd)$ weighted-value pass and $O(d)$ rational moment fields and decisions; its diagnostic materializes the reconstructed matrix. Unsupported, nonfinite, or inconsistent enclosures decline.

Coordinatewise controls and the shared budget. Write $J_c = B_{\mathrm{abs}}(a_c^+, b_c^+, v_c^+; \bar{\kappa})$, $W_c = \sqrt{v_c^+}(\bar{\kappa} - 1)/(2\sqrt{\bar{\kappa}})$ and $M_c = B_{\mathrm{abs}}(a_c^+, b_c^+, a_c^+ b_c^+; \bar{\kappa})$. The strongest direct mixture of the individual moment controls is

$$B_{\mathrm{mix}} = \max_c \min\{W_c, M_c\}, \qquad \max_c J_c \leq B_{\mathrm{mix}} \leq \min\{\max_c W_c, \max_c M_c\}. \tag{19}$$

Mix chooses a control per coordinate rather than per head. Comparing Joint against Mix isolates the gain from enforcing both constraints together. All moment rules use the same remainder $r = \varepsilon - E_{\mathrm{arith}} - C_\eta$. Joint accepts when $r \geq 0$ and every $J_c \leq r$, evaluated by the rational tests in (12).

Certified fallback. If certification fails, the original FP16 head is transferred from CPU to GPU and evaluated with interval arithmetic. Its FP32 result is accepted if its own enclosure fits the budget. Both paths target the same exact-real output defined in Section 2.

6 Certification gains on quantized attention

We test the gain from value information over TV and from joint constraints over separate moment controls. All 144 cases from an existing packed-cache capture are replayed. The analyses are post-hoc; cases and budgets stay fixed.

Setup and controls. A pinned GPT-2 model (Radford et al., 2019) processes a fixed 512-token WikiText-2 validation prefix (Merity et al., 2017). One CPU FP32 prefill captures layers 0, 6, and 11, rounded once to FP16 as reference inputs. All 12 heads at positions 255 and 511, with INT4 and INT8 key formats, yield 144 head/format cases; values use INT4. Each case is tested at the original binary64 budgets $\varepsilon = \rho R_0$, where $\rho \in \{.001, .01, .05, .1, .25, .5, 1\}$ and $R_0 \geq \max_i \|V_i\|_\infty$, giving 1,008 case/budget pairs. CTV uses original-value midpoint centering and its valid E_{val} term. Weighted variance (WV), range/mean (RM), coordinatewise Mix, and Joint use reconstructed values and the same corrected C_η. Appendices J and H detail the protocol and checks.

Certification across budgets. Joint certifies 357 pairs, including all 249 CTV admissions and 108 additional pairs (Table 1; Figure 4a,b). WV certifies 342 and RM 270. Their head-level union certifies 350 pairs; coordinatewise Mix certifies 352. Joint adds five beyond Mix: two INT4 cases at $\rho = .5$ and three at $\rho = 1$. For INT8, WV, Mix, and Joint have identical admission counts at every budget. Reconstructed-value centered TV with the same C_η certifies 244 pairs, supporting the role of moment information in the observed gain. Joint dominates WV and RM by class inclusion; its inclusion of every original-value CTV admission is an empirical result.

Table 1: **Admissions out of 72 cases per key format.** Methods and value-error terms are defined in the text. Extra counts Joint admissions rejected by Mix. Each budget in a grouped row has the displayed count.

Bits	ρ	CTV	WV	RM	Mix	Joint	Extra
4	$.001, .01, .05, .1, .25$	0	0	0	0	0	0
4	.5	1	8	1	8	10	2
4	1	16	40	38	50	53	3
8	$.001, .01$	0	0	0	0	0	0
8	.05	0	15	0	15	15	0
8	.1	21	63	20	63	63	0
8	.25	67	72	67	72	72	0
8	$.5, 1$	72	72	72	72	72	0

A strict gain from joint information. In the first strict case (layer 6, head 3, position 255, INT4, $\rho = 1$), the budget 2.98828125 minus the common value-plus-arithmetic term of approximately 0.210452 leaves $r \approx 2.777830$. Joint gives $2.522542 < r$, whereas Mix gives $2.964915 > r$. The separate head bounds are 5.186063 for WV and 3.047412 for RM. Joint thus changes the decision even against Mix. Table 2 lists all five cases, and Figure 4c compares their margins.

Table 2: **All five strict gains over Mix.** Position p, layer l, and head h identify INT4-key cases. Every row satisfies Joint $< r <$ Mix. Values are rounded for display; certification uses exact rational comparisons.

p	l	h	ρ	Remainder r	Joint	Mix
255	6	3	1	2.777830	2.522542	2.964915
255	6	7	1	2.869769	2.789822	3.299556
511	0	0	.5	.778552	.752279	.789318
511	11	0	.5	4.290162	4.262614	4.420941
511	11	11	1	7.519041	6.999065	8.428816

Coordinate 18 exposes the mechanism: the variance-only optimum uses event mass 0.00630, below the support-compatible threshold 0.03215. Enforcing both constraints at one mass gives Joint's 2.522542 (Appendix K). The independently enclosed actual error is at most 0.106806; Joint certifies the budget despite remaining conservatism.

Probability refinement and numerical validation. Midpoint centering raises $h = 0$ admissions from 215 to 249; $h = 1, 4, 8$ lower masses add none at any budget. Scatter certifies 145 pairs, all already admitted by CTV. For $h = 8$, 69/72 INT4 and 59/72 INT8 summaries saturate the unrestricted TV bound. Of the INT8 saturations, 29 arise through the empty subset and 30 through retained subsets. The saturation criterion thus explains why probability refinement has limited benefit on this capture. All methods decline at the two smallest budgets.

An independent oracle recomputes exact reconstructed scores, 100-digit directed Decimal exponentials, means, and raw second moments. All 9,216 coordinate enclosures pass. An exact-rational checker independently reconstructs the event-envelope extrema and reproduces all moment and mixed decisions. Every certified pair also passes the independent enclosure of actual original-FP16 output error.

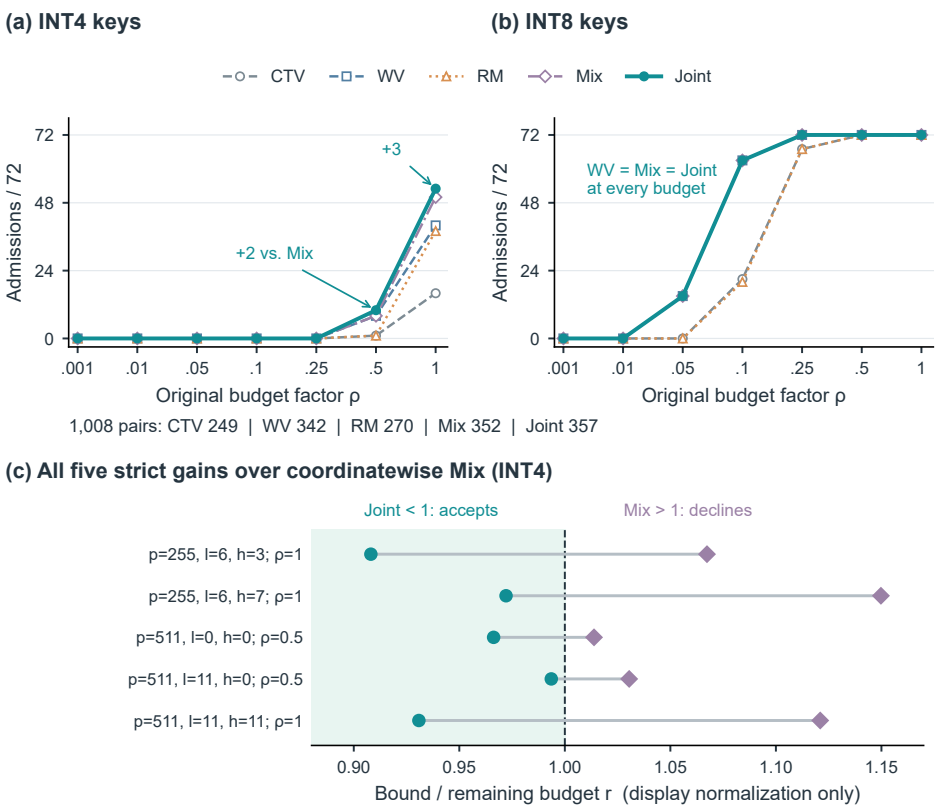

Figure 4: **Certification gains at fixed budgets.** (a,b) Admissions per key format at seven tested budget factors, shown as equally spaced categories. Joint certifies 357 pairs versus 352 for Mix and 249 for CTV. (c) All five additional admissions are INT4 cases with Joint $< r <$ Mix. Dividing the displayed bounds by their remaining budget r places the acceptance threshold at 1; certification budgets are unchanged. Case labels give position, layer, head, and ρ.

Computational cost. The CPU reference scans 3,538,944 reconstructed value elements over 144 queries, with rational operands up to 205 bits. Range extrema and H can be stored at cache write; weighted second moments require a query-dependent pass. Appendices G and J give separate centered-TV CUDA timings and memory/protocol details.

7 Discussion and Conclusion

Probability summaries bound changes in attention weights; value moments capture their effect on the output. Joint constraints exclude worst cases permitted by separate bounds, yielding five admissions beyond coordinatewise mixing on the fixed replay. The criteria clarify when to refine probabilities and when value information can resolve uncertainty left by TV conversion. Richer score constraints and fused certified moment reductions are natural next steps.

Scope and limitations. Optimality holds within the stated information classes; ordering, heterogeneous score errors, or cache coupling could restrict them further. A fixed cache may have a smaller optimum, and separately sharp key and value bounds need not attain their sum. Evaluation covers one GPT-2 prefix, correlated heads, and fixed FP16 inputs. Generalization to other models, contexts, and autoregressive trajectories remains untested. Joint moments add a weighted-value pass; certified GPU implementation and end-to-end timing remain open. Reported CUDA timings cover centered TV.

AI USE STATEMENT

Generative AI tools assisted with reviewing mathematical proofs for rigor, writing and debugging code, and improving the clarity and readability of the manuscript. The authors take full responsibility for the final manuscript, all mathematical derivations, experimental results, and code.

REPRODUCIBILITY STATEMENT

Appendices C, G, and H provide the proofs; Appendix E specifies the arithmetic assumptions and output enclosures. Appendix J records model and data identities, case selection, quantization formats, budgets, and measurement procedures. Appendix K describes the portable certificate replay and distinguishes it from the original capture and numerical checks.

ETHICS STATEMENT

The study uses public model weights and public text, with no human-subject study or deployed intervention. Its guarantees concern the numerical error of a single attention output under specified input and arithmetic assumptions. Assessing the safety, fairness, or factual accuracy of generated responses requires separate evaluation. Retained public inputs preserve their source attribution and license.

REFERENCES

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000. doi: 10.1080/00029890.2000.12005203. URL https://repository.ias.ac.in/2603/.

Dean Calver. Runtime-certified bounded-error quantized attention. *arXiv preprint arXiv:2605.20868v1*, 2026. URL https://arxiv.org/abs/2605.20868v1.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. *arXiv preprint arXiv:2309.02413v2*, 2024. URL https://arxiv.org/abs/2309.02413v2.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

Lukas Haverbeck, Carmen Amo Alonso, Andres Felipe Posada-Moreno, Sebastian Trimpe, and Marco Pavone. The risk of KV cache compression. *arXiv preprint arXiv:2607.01520v1*, 2026. URL https://arxiv.org/abs/2607.01520v1.

Edgars Jakobsons and Steven Vanduffel. Dependence uncertainty bounds for the expectile of a portfolio. *Risks*, 3(4):599–623, 2015. doi: 10.3390/risks3040599. URL https://www.mdpi.com/2227-9091/3/4/599.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

NVIDIA. CUDA math API: Double precision intrinsics, 2023. URL https://docs.nvidia.com/cuda/archive/12.1.0/cuda-math-api/group__CUDA__MATH__INTRINSIC__DOUBLE.html. CUDA Toolkit 12.1.

Python Software Foundation. Decimal fixed-point and floating-point arithmetic, 2024. URL https://docs.python.org/3.12/library/decimal.html. Python 3.12 library reference.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Georgios Tzachristas, Lei Deng, Ioannis Tzachristas, Gong Zhang, and Renhai Chen. A mathematical theory of top-k sparse attention via total variation distance. *arXiv preprint arXiv:2512.07647v1*, 2025. URL https://arxiv.org/abs/2512.07647v1.

Fanzhe Wei, Li Liu, Ziyang Wang, and Chenyu Wang. WitCert: Sound runtime risk observability and gating for KV-cache quantization. *arXiv preprint arXiv:2607.28699v3*, 2026. URL https://arxiv.org/abs/2607.28699v3.

A Detailed probability information contracts

A.1 Selected exact masses

Suppose h distinct probabilities $z_1, \dots, z_h > 0$ are supplied exactly, and only the atom count $r = n - h$ and total $\eta = 1 - \sum_j z_j$ are supplied for the rest. Require $\eta > 0$ if $r > 0$ and $\eta = 0$ if $r = 0$. There is no ordering constraint or per-atom cap. Write $s_J = \sum_{j \in J} z_j$ and $[t]_{[a,b]} = \min\{b, \max\{a, t\}\}$.

Theorem 6 (Exact-mass information optimum). *For $\kappa > 1$ and $r \geq 2$, the largest possible $\mathcal{E}(p, \kappa)$ over all positive probabilities compatible with this information is*

$$\mathcal{B}_{=}(z, r; \kappa) = \max_{J \subseteq [h]} f_\kappa\big([t_*]_{[s_J, s_J + \eta]}\big)\,. \tag{20}$$

For $r = 1$, replace each interval maximum by $\max\{f_\kappa(s_J), f_\kappa(s_J + \eta)\}$; for $r = 0$, use $f_\kappa(s_J)$. The maximum is attained. Revealing an additional exact mass cannot increase this optimum, and revealing all masses recovers $\mathcal{E}(p, \kappa)$.

Every token subset contributes its selected mass s_J plus a mass in $[0, \eta]$. When at least two unknown atoms remain, any interior mass can be allocated to one atom, with the remainder split positively over the others. The endpoints select none or all of those atoms. Unimodality then proves the formula and its attainability. With only one unknown atom there are only two choices, which explains the separate endpoint rule. The complete argument, including refinement, is in Appendix C.1.

If the retained coordinates happen to be the largest h probabilities, their ordering is extra information. Equation (20) remains valid, but the theorem does not claim optimality among certificates that also exploit that ordering. This distinction matters when interpreting a negative result as an information limit.

A.2 Probability lower bounds

Stored probabilities generally cannot be treated as exact softmax values. The implementable interface instead retains $\ell_j \geq 0$ with $\ell_j \leq p_j$ on h distinct coordinates. Set $L = \sum_{j=1}^{h} \ell_j \leq 1$, $\eta = 1 - L$, and define the compact class

$$\mathcal{P}(\ell) = \{p \in \mathbb{R}^n_+ : \sum_i p_i = 1,\ p_j \geq \ell_j \text{ for } j \leq h\}. \tag{21}$$

Unretained coordinates have lower bound zero. Thus η is unallocated mass, including mass that may be assigned to retained coordinates; it is not restricted to unretained atoms.

Theorem 7 (Lower-mass information optimum). *Let $n \geq 2$, $\kappa > 1$ and $s_J = \sum_{j \in J} \ell_j$. If $h < n$, every $J \subseteq [h]$ is admissible. If $h = n$, only nonempty proper J are admissible. Then*

$$\mathcal{B}_{\geq}(\ell, n; \kappa) := \max_{p \in \mathcal{P}(\ell)} \mathcal{E}(p, \kappa) = \max_{J \text{ admissible}} f_\kappa\big([t_*]_{[s_J, s_J + \eta]}\big)\,. \tag{22}$$

The maximum is attained in the nonnegative class. If the strictly positive class is nonempty, the same number is its supremum. For $n = 1$ or $\kappa = 1$, the answer is zero. Increasing valid lower bounds, or supplying more of them, cannot increase the optimum.

Why it is exact. For a nonempty proper token set A, the possible mass is precisely $[\ell(A), 1 - \ell(A^c)]$: start from the lower vector and allocate the two remaining amounts to one atom inside A and one outside. Interchanging the finite maximum over A with the maximum over p reduces the problem to these intervals. Their selected-coordinate descriptions are exactly the admissible J above. A positive feasible vector can be mixed with a boundary maximizer to establish the supremum statement. Appendix C.2 details all cases.

A necessary distinction. With $n = 2$, $\ell_1 = 1/2$ and $\kappa = 9$, the true vector $(3/4, 1/4)$ is compatible and has $\mathcal{E} = 1/2$. Substituting ℓ_1 into the exact-mass $r = 1$ rule would give $f_9(1/2) = 2/5$, an underestimate. The lower-mass formula instead permits the interval $[0, 1/2]$ and returns $1/2$. Downward rounding is safe only together with the corresponding lower-bound information model.

B Rational probability decisions and exact evaluation

B.1 Why complete exact evaluation is combinatorial

Proposition 8 (Exact-evaluation obstruction)**.** *Given positive rational probabilities p, deciding whether $\mathcal{E}(p, 9) = 1/2$ is NP-complete. This is a weak NP-hardness statement about rational probabilities and a constant rational multiplier bound.*

Equality requires a subset of mass $1/4$. For a positive-integer Partition instance with total W, use atoms $w_i/(2W)$ and one extra atom of mass $1/2$. A quarter-mass subset exists exactly when the integers partition equally. Appendix C.4 proves the reduction and membership in NP. This obstruction applies to exact evaluation on rational probabilities. It neither rules out inexpensive certificates at bounded h nor establishes approximation hardness.

C Complete information and decision proofs

Throughout this appendix $\kappa \geq 1$, probabilities sum to one and TV is half the ℓ_1 distance. For $\kappa = 1$ every tilt is the identity. That case can be removed before any formula involving a critical point is used. Nonnegative vectors in the lower-information class are mathematical closure points; the actual softmax of finite scores is positive.

C.1 Exact-mass optimum and refinement

Proof of Theorem 6. Fix a positive probability vector consistent with the supplied masses. A token subset consists of selected coordinates J and some subset of the unrevealed atoms. Its mass is $s_J + t$ for $t \in [0, \eta]$. By (4), its contribution to the worst-case TV is $f_\kappa(s_J + t)$. Unimodality bounds this by the value at the clipped critical point. If $r = 1$, the only unrevealed subset masses are 0 and η; if $r = 0$, only 0 is possible. This proves all upper bounds.

For sharpness with $r \geq 2$, select a maximizing J and let $t = [t_*]_{[s_J, s_J+\eta]} - s_J$. If $t = 0$, split η positively among the r unrevealed atoms and select none. If $t = \eta$, make any positive split and select all of them. If $0 < t < \eta$, give one unrevealed atom mass t and split $\eta - t$ positively among the other $r - 1$ atoms. Select the first of these atoms. The resulting token subset A has the desired mass. Use multiplier κ on A and 1 outside. Its probability excess is the proposed value, and the established upper bound prevents a larger TV. Every revealed mass and the remaining atom count are preserved. The $r = 1$ and $r = 0$ cases use the same endpoint construction with their allowed unrevealed subsets. There are finitely many J, so a maximizing choice exists.

Finally, reveal an actual unrevealed probability w and change η to $\eta - w$. The set of compatible probability vectors shrinks, so its maximum cannot increase. Equivalently, when interval descriptions remain applicable, the new intervals are $[s_J, s_J + \eta - w]$ and $[s_J + w, s_J + \eta]$, both contained in the old interval. The endpoint cases are also subsets of the old feasible choices. Revealing all masses leaves the original fixed vector and recovers $\mathcal{E}(p, \kappa)$. □

No ordered-top-h constraint is used in this construction. In particular, the maximizing completion can assign a large mass to an unrevealed atom. If that violates an ordering observation available to

another algorithm, that algorithm is operating on a stricter information class. This does not invalidate either the stated upper bound or its optimality within the stated class.

C.2 Lower-information optimum and all atom-count cases

Extend the retained lower vector by zeros on its unretained coordinates and continue to denote the resulting vector by ℓ. Let $\ell(A) = \sum_{i \in A} \ell_i$. The nonnegative feasible class is nonempty whenever $\ell \geq 0$ and $\sum_i \ell_i \leq 1$; the leftover mass can be placed on any atom.

Lemma 9 (Attainable subset masses)**.** *For every nonempty proper $A \subset [n]$, the set $\{p(A) : p \in \mathcal{P}(\ell)\}$ is exactly $[\ell(A), 1 - \ell(A^c)]$.*

Proof. Both inequalities follow from the coordinatewise lower constraints. For any t in the interval, start at $p = \ell$. Choose $a \in A$ and $b \in A^c$, add $t - \ell(A)$ to p_a and add $1 - t - \ell(A^c)$ to p_b. These quantities are nonnegative and sum to $1 - \sum_i \ell_i$. The resulting vector is feasible and has $p(A) = t$. □

Proof of Theorem 7. The class is compact, and $\mathcal{E}$ is a finite maximum of continuous functions. For $n \geq 2$, the empty and full token subsets contribute zero. Interchange the maximum over the class with the finite maximum over the nonempty proper subsets, and apply Lemma 9:

$$\max_{p \in \mathcal{P}(\ell)} \mathcal{E}(p, \kappa) = \max_{\varnothing \neq A \subsetneq [n]} \max_{t \in [\ell(A), 1 - \ell(A^c)]} f_\kappa(t) \tag{23}$$

$$= \max_{\varnothing \neq A \subsetneq [n]} f_\kappa\left([t_*]_{[\ell(A), 1 - \ell(A^c)]}\right). \tag{24}$$

If $h = n$, the set A is its selected-coordinate subset J, so J must be nonempty and proper. If $h < n$ and $h \geq 1$, the empty J is realized by choosing one unretained atom for A. The full J is realized by choosing all retained atoms for A and excluding an unretained atom. Any other J is already a nonempty proper token subset. If $h = 0$, the condition $n \geq 2$ supplies an atom on either side. Thus every stated admissible J is realized, and every A induces such a J. In every case, the interval is $[s_J, s_J + \eta]$. Equation (24) proves the formula and compact attainment.

For strict positivity, suppose a feasible $p^0 > 0$ exists, and let p^* attain the compact maximum. Then $p^{(\epsilon)} = (1-\epsilon)p^* + \epsilon p^0$ is positive and feasible for $0 < \epsilon < 1$. Continuity gives $\mathcal{E}(p^{(\epsilon)}, \kappa) \to \mathcal{E}(p^*, \kappa)$. The maximum is therefore the supremum in the positive class. If $\sum_i \ell_i = 1$ with a zero coordinate, that positive class is empty; no positive-attainment claim is made in that case. For $n = 1$, every probability vector and its tilt have a single unit atom and TV zero. For $\kappa = 1$, the tilt is the identity.

Increasing lower bounds or adding a valid coordinatewise lower constraint shrinks the feasible class. The associated maximum cannot increase, which proves monotonic refinement. □

C.3 Saturation and the dominant-mass stopping condition

Proof of Corollary 3. For $\kappa > 1$, f_κ has the unique maximizer t_* on $[0, 1]$. There are finitely many admissible closed intervals. Their union has the same maximum as $[0, 1]$ exactly when it contains t_*. The sign of the derivative is the sign of $D_\kappa(t)$, whose unique zero in $[0, 1]$ is t_*. This proves the rational criterion. For $h < n$, the empty J is admissible and supplies $[0, \eta]$, proving the last statement. □

Proof of Corollary 4. Write $u = \ell_j$. The assumption implies $1-u < t_* < u$ and $u > 1/2$. For every compatible vector p, a subset omitting j has mass in $[0, 1 - u]$, while a subset containing j has mass in $[u, 1]$. Unimodality gives upper bounds $f_\kappa(1 - u)$ and $f_\kappa(u)$ on these two types. The numerators at u and $1 - u$ agree, and the denominator at $1 - u$ is smaller. Consequently $f_\kappa(u) \leq f_\kappa(1 - u)$, including the zero endpoints when $u = 1$. The classical subset formula therefore bounds every feasible $\mathcal{E}(p, \kappa)$ by $f_\kappa(1 - u)$.

For attainment, begin with the full lower vector and keep $p_j = u$. Its nonnegative unallocated mass can all be placed on any coordinate other than j, since $n \geq 2$. This preserves every lower bound and makes the complement of j have mass exactly $1 - u$. Giving that complement multiplier κ and j multiplier 1 attains $f_\kappa(1 - u)$. This proves equality for arbitrary consistent lower bounds on other

coordinates, including $h = n$. The positive-class supremum statement follows by the same feasible mixing argument as Theorem 7.

The strict condition $1 - u < t_*$ is equivalent to $D_\kappa(1-u) > 0$. For the counterexample outside the regime, a single lower bound $2/5$ leaves $[0, 3/5]$ attainable and hence saturates at $1/2$. With lower vector $(2/5, 3/10, 3/10)$ the total is one, so the distribution is fixed. Its possible nontrivial subset masses are $3/10, 2/5, 3/5, 7/10$. Their largest f_9 value is $f_9(3/10) = 42/85 < 1/2$. □

C.4 Budget equivalence, traversal and hardness

Proof of Proposition 5. Multiplying $f_\kappa(t) \le \tau$ by its positive denominator and collecting terms gives (16). For $\kappa > 1$, Q_τ is a concave quadratic with vertex $(1-\tau)/2$. Its maximum on an interval is the value at the clipped vertex. Therefore the inequalities hold for every possible subset mass exactly when (17) holds for every admissible interval. The $n = 1$ and $\kappa = 1$ cases have zero TV; $\tau \ge 1$ is trivial since TV is at most one. The argument includes equality and uses only rational operations on rational inputs. □

A depth-first binary traversal assigns ℓ_j either to an included sum or to an excluded sum. It uses constant work per internal node, then constant scalar work for each leaf's interval and quadratic. A full tree has $2^{h+1} - 1$ nodes and depth h, giving the stated time and auxiliary memory. Exact rational numerator and denominator lengths are not constant; their bit costs are excluded from the scalar-operation count. Likewise, deriving probabilities from n scores is a separate operation.

Proof of Proposition 8. The identity is

$$\frac{1}{2} - f_9(t) = \frac{(4t-1)^2}{2(1+8t)}. \tag{25}$$

Identity (25) follows by putting both sides over $2(1 + 8t)$. Hence $f_9(t) = 1/2$ exactly at $t = 1/4$, and $\mathcal{E}(p, 9) = 1/2$ exactly when such a subset mass exists. Starting from the standard positive-integer Partition problem (Garey & Johnson, 1979), map weights of sum $W > 0$ to $(w_1/(2W), \ldots, w_m/(2W), 1/2)$. All probabilities are positive and sum to one. A quarter-mass subset excludes the last atom and exists precisely when $\sum_{i \in A} w_i = W/2$. The rational encoding size is polynomial in the original input size. A subset is a polynomially checkable witness using exact rational addition, so the decision problem is in NP. The reduction establishes weak NP-hardness; it gives neither a strong hardness statement nor an approximation lower bound. □

C.5 Output propagation

For any center c and any norm, assume $R \ge \max_i \|V_i - c\|$ and $\eta_i \ge \left\|V_i - \widehat{V}_i\right\|$. Then

$$\left\|O - \widehat{O}\right\| \le \left\|\sum_i (\pi_i - p_i)(V_i - c)\right\| + \left\|\sum_i p_i (V_i - \widehat{V}_i)\right\| \tag{26}$$

$$\le R \sum_i |\pi_i - p_i| + \sum_i p_i \eta_i \le 2RB + \sum_i p_i \eta_i. \tag{27}$$

The center cancels because both probability vectors sum to one. Adding $\left\|y - \widehat{O}\right\| \le E_{\text{arith}}$ proves (33) in the chosen infinity norm with either zero or midpoint centering. It is valid for an adaptively selected current query whenever its deterministic metadata are valid. The guarantee compares outputs for the same query and history; it does not provide a recursion for errors along an autoregressive trajectory.

D Complete rational decision procedure

Decision procedure. For exact rational inputs, the following procedure specifies the complete test. Let $\text{pass}(a, b)$ mean $Q_\tau([v_\tau]_{[a,b]}) \le 0$.

1. Validate $n \geq 1$, $0 \leq h \leq n$, $\kappa \geq 1$, $\tau \geq 0$, $\ell_j \geq 0$ and $\sum_j \ell_j \leq 1$. Invalid inputs do not certify.
2. If $n = 1$, $\kappa = 1$ or $\tau \geq 1$, return *certified*. If $\mathrm{pass}(0, 1)$ holds, also return *certified*.
3. Set $\eta = 1 - \sum_j \ell_j$. Traverse the include/exclude tree with initial included sum $a = 0$, excluded sum $b = 0$ and included count $c = 0$.
4. At coordinate j, recurse on $(a + \ell_j, b, c + 1)$ and $(a, b + \ell_j, c)$, retaining only the current depth-first stack.
5. At a leaf, skip it if $h = n$ and $c \in \{0, n\}$. Otherwise test $\mathrm{pass}(a, 1 - b)$. Return *uncertified* as soon as one admissible test fails.
6. If every admissible leaf passes, return *certified*.

The included count distinguishes empty or full sets even when some lower bounds are zero. For $h = 0 < n$, the root is a single admissible leaf. This procedure is the exact reference for the outward-rounded implementation: an unresolved numerical comparison declines certification. The dominant-mass stopping corollary is an analytical consequence, not an additional shortcut used in the reported timing.

E Arithmetic contract and implementation details

The CPU gate assumes binary64 round-to-nearest basic operations with gradual underflow, together with neighboring values from `math.nextafter`. The GPU uses explicit directed binary64 CUDA intrinsics with fast math and flush-to-zero disabled. These numerical contracts define the scope of the guarantee. Environment probes and finite fixtures can expose implementation errors; they cannot establish the semantics of arbitrary hardware, compilers, or third-party kernels.

E.1 Exact reconstruction metadata

Finite FP16 numbers are integer multiples of 2^{-24} with magnitude less than 2^{16}. An INT8 code times an FP16 scale has magnitude less than 2^{23} and is still a multiple of 2^{-24}. Its difference from an FP16 original therefore fits within 49 significant bits, below binary64's 53. Products, residual subtraction, absolute value and maxima used at cache write time are exact in this format range. The implementation then converts to FP32 and advances one FP32 neighbor upward if the nearest conversion fell below the exact residual. Nonfinite source inputs or stored scales are rejected.

Keys are grouped by 32 tokens with separate scales for all 64 channels. Values are grouped by 32 channels within each token. A zero group receives a positive unit scale and zero codes. A nonzero group whose nearest FP16 scale would underflow receives the smallest positive FP16 scale. Codes are rounded against the actual stored scale and clipped. The residual metadata therefore remain valid even if a rounded scale causes clipping. INT4 storage adds eight to each signed code, placing even flat indices in the low nibble and odd indices in the high nibble. The primary experiment uses signed code ranges $[-7, 7]$ or $[-127, 127]$; the nibble decoder supports $[-8, 7]$.

The GPU decodes FP16 directly from sign, exponent and mantissa bits. A query coordinate times an integer-times-half reconstructed key also fits exactly in binary64, but the kernel encloses products and every accumulated sum with directed operations anyway. Since $d = 64$, the $1/\sqrt{d}$ multiplier is exactly $1/8$. This yields valid s_i^-, s_i^+ without relying on a nominal FP32 matrix multiplication or a summation-order assumption.

E.2 A bounded exponential without a libm error assumption

For an exact input $x \in [-4096, 700]$, repeatedly halve $|x|$ until $z = |x|/2^k \leq 1/8$. All halving steps occur in the normal binary64 range and are exact; an already tiny argument requires no halving. Define $a_j = z^j/j!$ and $S_{12} = \sum_{j=0}^{12} a_j$. Positivity and the bound on successive term ratios give

$$S_{12} \leq e^z \leq S_{12} + \frac{a_{13}}{1 - z/14}. \tag{28}$$

Indeed, for every term after a_{13} its ratio to the preceding term is at most $z/14 < 1$, so the remaining positive tail is bounded by a geometric series beginning with a_{13}.

The recurrence $a_j = a_{j-1}z/j$ is evaluated down and up, as are the two partial sums. The upper tail uses an upper numerator and a downward positive denominator. If $x < 0$, invert the interval with directed division. Square the positive interval k times with directed multiplication. Monotonicity of reciprocal on positive numbers and of squaring on nonnegative numbers proves the final enclosure of e^x. Directed underflow can produce a zero lower endpoint or a smallest-subnormal upper endpoint; both remain valid. The exact zero input returns $[1, 1]$. Unsupported arguments or nonfinite intermediates are not silently replaced by finite certifying values.

For score intervals, the lower exponential is computed from $\mathrm{down}(s_i^- - M)$ and the upper from $\mathrm{up}(s_i^+ - M)$. Exponential monotonicity proves their validity. The denominator is accumulated in both directions. At least one upper score equals M; the code additionally requires the summed lower denominator to be positive. Pointwise division by the upper denominator gives valid lower probabilities, regardless of how their distinct indices were chosen.

E.3 Output and value-error enclosures

For a reconstructed value v, an exponential interval $[e^-, e^+]$ gives $[e^- v, e^+ v]$ when $v \geq 0$ and $[e^+ v, e^- v]$ otherwise, with directed products. Summing these intervals yields a signed numerator interval $[N^-, N^+]$. For a positive denominator $[Z^-, Z^+]$, its lower quotient is N^-/Z^+ when $N^- \geq 0$ and N^-/Z^- otherwise. The upper quotient is N^+/Z^- when $N^+ \geq 0$ and N^+/Z^+ otherwise. Directed division therefore encloses every exact dequantized output coordinate.

The kernel rounds an interval-midpoint representative to FP32. Irrespective of whether that rounded value lies inside its interval, the maximum of $\mathrm{up}(y_c - L_c)$ and $\mathrm{up}(U_c - y_c)$ bounds its distance to every possible exact output coordinate. Their maximum over coordinates is E_{arith}. Nonfinite values make the result unusable.

Since $p_i \leq e_i^+/Z^-$ and $\eta_i \geq 0$, an upward evaluation of $\sum_i e_i^+ \eta_i / Z^-$ bounds the weighted value error. The alternative $\max_i \eta_i$ is also valid because p sums to one. Their minimum is therefore a valid upper bound. Together with Appendix C.5 these facts prove the entire local implication of an accepted output gate.

E.4 Outward binary64 decisions

The CPU interprets supplied float metadata as exact dyadic numbers and validates their nonnegativity and total lower mass using rational arithmetic. It accumulates each subset's included and excluded lower sums downward. Thus $[a, b] = [s_J^{\mathrm{down}}, \mathrm{up}(1 - s_{J^c}^{\mathrm{down}})]$, clipped to $[0, 1]$, contains the exact feasible mass interval. Compute an enclosure $[v^-, v^+]$ of $(1 - \tau)/2$ and set

$$W = \left[[v^-]_{[a,b]}, [v^+]_{[a,b]} \right]. \tag{29}$$

Monotonicity of clipping shows that W contains the maximizer of Q_τ on the expanded interval. An outward interval evaluation of $Q_\tau(W)$ bounds its value there, which bounds its maximum on the original interval. Acceptance requires an upper endpoint at most zero for every admissible interval. The proof follows inductively from the enclosure of each basic operation. A conservative interval may fail an exactly feasible boundary case; such a failure is reported as unresolved, not an actual error violation.

For the output budget, the implementation subtracts the stored error bounds and divides by $2R$ in exact Fraction arithmetic, then converts the positive result downward to binary64. Its default subset cap is $h \leq 16$; the fixed GPU study uses at most eight masses. Nonfinite arithmetic or an exceeded enumeration cap declines. The complete path is specific to this reference and does not confer a numerical guarantee on other kernels.

F Packed interval path and the probability gate

Packed storage and write-time metadata. The fixed path has $d = 64$. Keys use symmetric signed INT4 or INT8 codes with one stored FP16 scale per channel per block of 32 tokens. Values

use INT4 codes with one FP16 scale per token and group of 32 channels. INT4 codes occupy two nibbles of one byte; there is no resident unpacked cache. Quantization rounds and clips against the *stored* scale. To account for scale rounding and clipping, we compute upward FP32 residual metadata from the actual reconstruction:

$$a_c \geq \max_i |K_{ic} - \widehat{K}_{ic}|, \quad \eta_i \geq \left\| V_i - \widehat{V}_i \right\|_\infty, \quad R_0 \geq \max_i \|V_i\|_\infty . \tag{30}$$

The reference originals are FP16. Their residuals against integer-times-FP16 reconstructions fit exactly in binary64 before upward FP32 storage (Appendix E). A directed evaluation of $\sum_c |u_c| a_c/8$ then bounds Δ. Original FP16 keys and values remain on the CPU for whole-head fallback; their storage is charged separately.

Scores and lower masses. The GPU decodes each packed code and scale and encloses the dot product using explicit round-down and round-up CUDA intrinsics (NVIDIA, 2023). Let $s_i^- \leq s_i \leq s_i^+$ be those score endpoints and $M = \max_i s_i^+$. We compute

$$e_i^- \leq e^{s_i - M} \leq e_i^+, \qquad 0 < Z^- \leq \sum_i e^{s_i - M} \leq Z^+, \qquad \ell_i = \mathrm{down}(e_i^- / Z^+). \tag{31}$$

A positive Taylor sum, a bounded remainder and directed squaring enclose exponentials on $[-4096, 700]$ (Appendix E). Unsupported or nonfinite arithmetic cannot certify.

The h largest lower unnormalized exponentials are selected, with index order breaking ties. Selection visits $O(hn)$ values in this reference; the scalar denominator is common. The theorem uses the resulting lower bounds and identities, without exploiting extra ordering. The unrestricted control requests $h = 0$ and skips selection. A CPU Decimal implementation and an independent Decimal output enclosure provide numerical references (Python Software Foundation, 2024); they are validation costs, not GPU timing results.

The returned vector. Directed summation of $e_i \widehat{V}_{ic}$ and division by $[Z^-, Z^+]$ give an interval $[L_c, U_c]$ for each coordinate of exact dequantized attention $\widehat{O}$. The kernel returns an FP32 representative y_c and computes an upper bound E_{arith} on $\max_c \max\{|y_c - L_c|, |y_c - U_c|\}$. It also obtains

$$E_{\mathrm{val}} \geq \sum_i p_i \eta_i, \qquad E_{\mathrm{val}} = \min\left\{ \max_i \eta_i, \mathrm{up}\left(\frac{\sum_i e_i^+ \eta_i}{Z^-} \right) \right\}, \tag{32}$$

where the numerator is itself accumulated upward. For any valid TV bound B, choose a valid enclosing radius R (either R_0 or R_*^+). Centering gives

$$\|y - O\|_\infty \leq E_{\mathrm{arith}} + E_{\mathrm{val}} + 2RB, \qquad O = \sum_i \pi_i V_i. \tag{33}$$

Insert and subtract $\sum_i p_i V_i$. The value difference contributes $\sum_i p_i \eta_i$, while the key difference equals $\sum_i (\pi_i - p_i)(V_i - c)$ and is bounded by $2RB$. The triangle inequality adds the returned vector's arithmetic error. Thus a sharper key certificate cannot conceal an exhausted value budget.

For $R > 0$, the CPU forms $\tau = \mathrm{down}((\varepsilon - E_{\mathrm{val}} - E_{\mathrm{arith}})/(2R))$ and applies the lower-mass gate, declining if the numerator is negative. The $R = 0$ case is direct. The downward gate rounds all interval operations outward; near a boundary it can decline even when the rational gate passes. No positive residual is waived by a numerical tolerance.

Proposition 10 (Local implication of acceptance)**.** *Let $R > 0$ enclose the original values about a common center. Suppose the metadata in (30) and (31) are valid, $\bar{\kappa} \geq e^{2\Delta}$, and the output intervals enclose $\widehat{O}$ and its returned vector's arithmetic error as above. If the outward gate certifies with inputs $(\ell, n, \bar{\kappa}, \tau)$, then $\|y - O\|_\infty \leq \varepsilon$.*

The true probability and multiplier vectors belong to the declared classes; sound comparisons and the downward budget give $2R\tau + E_{\mathrm{val}} + E_{\mathrm{arith}} \leq \varepsilon$. Equation (33) proves the implication.

Outward evaluation encloses the quadratic maximum on an expanded mass interval; an unresolved sign declines. Appendix E.4 gives the construction. The CPU fixtures retain 125 exactly feasible declines among 4,604 comparisons, with no false acceptance. Arithmetic refinement can resolve numerical ambiguity, not information saturation.

G VALUE GEOMETRY: PROOFS, ARITHMETIC AND FIXED VALIDATION

G.1 COMPLETE PROOF AND ATTAINMENTS

Let X take $x_i \in [1, \kappa]$ with probabilities p_i and let $a = \sum_i p_i x_i$. The nonnegative product $(X - 1)(\kappa - X)$ gives $\mathbb{E}X^2 \le (\kappa + 1)a - \kappa$. Consequently

$$\chi^2(T_x(p)\|p) = \sum_{i:p_i>0} p_i(x_i/a - 1)^2 = \mathrm{Var}(X)/a^2 \tag{34}$$

$$\le \frac{(\kappa - a)(a - 1)}{a^2} \le \frac{(\kappa - 1)^2}{4\kappa}, \tag{35}$$

where the last step has the exact nonnegative remainder

$$\frac{(\kappa - 1)^2}{4\kappa} - \frac{(\kappa - a)(a - 1)}{a^2} = \frac{((\kappa + 1)a - 2\kappa)^2}{4\kappa a^2}. \tag{36}$$

For each coordinate, the weights sum to zero, so weighted Cauchy–Schwarz gives

$$\left|\sum_i (q_i - p_i)V_{ic}\right|^2 = \left|\sum_i p_i(x_i/a - 1)(V_{ic} - \overline{V}_c)\right|^2 \tag{37}$$

$$\le \chi^2(q\|p)\sum_i p_i(V_{ic} - \overline{V}_c)^2 \le \chi^2(q\|p)\omega S. \tag{38}$$

Taking the maximum over coordinates proves Theorem 1, including $n = 1$, $\kappa = 1$ and constant values.

Write rational $\kappa = A/B > 1$ for positive integers $A > B$. Take $n = A + B$ equal probabilities, B multipliers equal to κ and A equal to one, and scalar $V_i = x_i$. Then $a = 2\kappa/(\kappa + 1)$, X takes only endpoints, $\omega = 1/n$, and every inequality above is an equality. Replicating each atom any number of times preserves equality. This construction establishes the asserted universal sharpness without claiming optimality for arbitrary fixed metadata.

For any center c, the centered TV argument gives $\|\sum_i (q_i - p_i)V_i\|_\infty \le 2B \max_i \|V_i - c\|_\infty$. Each coordinate range requires at least half its length; its midpoint attains that lower bound simultaneously. This proves the formula for R_*. The scatter argument instead uses the coordinate means, which minimize the unweighted squared deviations. The two centers may differ: each is valid because the distribution differences sum to zero. Squaring the positive centered and scatter bounds and cancelling $(\sqrt{\kappa} - 1)^2$ shows that scatter is strictly better precisely when $\omega S/R_*^2 < 16\kappa/(\sqrt{\kappa} + 1)^4$, for $\kappa > 1, R_* > 0$.

For the separation in Section 3, $R_* = 1/2$, $S = (n - 1)/n$ and $\omega = 1/n$. The full probability vector has a subset of mass $1/4$, so its exact TV envelope is $U_9 = 1/2$. The scatter output bound is $4\sqrt{n - 1}/(3n)$ and the ratio of the centered TV bound to this bound is $3n/(8\sqrt{n - 1})$. It exceeds one for the asserted multiples of four with $n \ge 8$. The actual worst output change is the largest change to the first probability, namely $8(n - 1)/(n(n + 8))$. It is smaller than the scatter bound; no equality claim is made on this separation family.

G.2 EXACT WRITE-TIME METADATA AND RUNTIME GATE

Every finite FP16 value has the form $I_i 2^{-24}$ for an integer I_i. For a coordinate, with $T = \sum_i I_i$, its exact squared scatter and half-range are

$$\frac{n\sum_i I_i^2 - T^2}{n2^{48}}, \qquad \frac{\max_i I_i - \min_i I_i}{2 \cdot 2^{24}}. \tag{39}$$

Arbitrary-precision integers avoid cancellation between rounded moments. The implementation takes the coordinate maxima and rounds upward to FP32, checking the resulting dyadic values against the exact fractions. Unsupported, nonfinite or unrepresentable metadata cannot certify. This reference uses $O(nd)$ write-time arithmetic operations plus integer bit costs; it does not establish an efficient online update algorithm.

Since the runtime sets $M = \max_i s_i^+$, each exact shifted exponential is at most one. Dividing by the valid positive lower normalizer proves the cap

$$\max_i p_i \le \omega^+ := \min\{1, 1/Z^-\}. \tag{40}$$

An arbitrary scaled softmax denominator would not suffice: the numerator bound is part of the contract. With $r = \varepsilon - E_{\mathrm{val}} - E_{\mathrm{arith}} \ge 0$, the scatter gate accepts iff

$$4\bar{\kappa} r^2 \ge \omega^+ S^+ (\bar{\kappa} - 1)^2. \tag{41}$$

All stored floats are interpreted as exact dyadic rationals. The comparison uses a constant number of rational operations, with bit costs charged separately, and requires no square root. Equality accepts; negative remaining budgets, invalid arithmetic flags and nonpositive lower normalizers decline. The function $(\kappa - 1)^2/\kappa$ is increasing for $\kappa \ge 1$, so the stored upper factor $\bar{\kappa}$ remains sound. Combining this bound with the unchanged value and output-arithmetic terms proves the same local output implication as Proposition 10. The union of the centered-TV and scatter admissions is sound.

The geometry module stores R_*^+, S^+ as two FP32 values (eight numeric payload bytes). The timed centered-only reference retains an additional four-byte host radius for comparison with its original radius; it does not execute the rational scatter gate. Both consume the existing 128-byte summary. These payload counts exclude object overhead and cache identity. The gate is evaluated by an exact CPU reference.

G.3 Finite checks and all retained model cases

The new protocol is a post-hoc analysis of the original data. Its rule and seven unchanged binary64 output budgets were recorded before geometry outcomes, after the old aggregate results and one old row had been inspected. It uses all 144 saved cases; no model, quantizer or favorable head is added. Source, trace and packet identities are checked, original caches are repacked, and every accepted decision is checked against the saved independent Decimal absolute-error enclosure. Centered binary64 gates agree with their exact rational controls on all 1,008 case/budget pairs. The subsequent necessary comparison evaluates $h = 0, 1, 4, 8$ with the same centered radius: all four have identical counts at every original budget, totaling 996 admissions among 4,032 decisions, with zero false acceptance.

The geometry fixtures contain 11,624 exact pointwise inequalities, nine equality constructions, five TV-saturation separations and five exact metadata fixtures. They also check constant values, invalid inputs and an exactly feasible dyadic budget whose immediately preceding float declines. The fixed diagnostic yields 249 centered admissions and 145 scatter admissions among 1,008 pairs; the union is still 249. The analytical separation therefore does not translate into additional admissions over centering on this grid. The reference metadata construction totals 268.402 ms across 72 unique value snapshots; this is untimed-path write work, not an online gate benchmark.

G.4 Centered-only runtime comparison

The follow-up retains the two original timing slices, layer 6/head 0/position 511, for INT4 and INT8 keys and all seven original budgets. It alternates the first method in paired old/centered $h = 0$ executions, with three warmups and 15 retained repetitions. The unchanged CUDA reference is actually executed; nonselection summary values, quantized outputs and dense outputs agree bitwise with the saved packets. CPU-original transfers occur only on actual fallback. Dense resident and dense offloaded controls are timed separately.

At $\rho = 0.25$, both radii decline for INT4 (median wall times 1.508 and 1.506 ms). For INT8, every old-radius repetition falls back, whereas all 15 centered repetitions admit, changing median wall time from 1.483 to 0.757 ms and removing the 131,072-byte original-cache upload. The separately measured resident and offloaded dense medians are 0.652 and 0.728 ms. The local fallback saving therefore does not establish overall acceleration. All raw samples and budgets remain in the artifact; this follow-up does not replace the original sequential timing observations.

Table 3: Centered-only follow-up: full-path wall-time medians in ms over 15 repetitions. All seven original budgets and both original timing slices are retained. Dense controls are measured separately in this phase.

Key bits	ρ	Zero-centered $h = 0$	Centered $h = 0$	Dense offload	Dense resident
4	0.001	1.699	1.692	0.753	0.655
4	0.01	1.704	1.686	0.753	0.655
4	0.05	1.728	1.740	0.753	0.655
4	0.1	1.497	1.496	0.753	0.655
4	0.25	1.508	1.506	0.753	0.655
4	0.5	1.505	1.498	0.753	0.655
4	1	1.494	1.513	0.753	0.655
8	0.001	1.454	1.450	0.728	0.652
8	0.01	1.446	1.477	0.728	0.652
8	0.05	1.480	1.486	0.728	0.652
8	0.1	1.479	1.507	0.728	0.652
8	0.25	1.483	0.757	0.728	0.652
8	0.5	0.760	0.757	0.728	0.652
8	1	0.764	0.778	0.728	0.652

H JOINT MOMENT OPTIMUM AND ITS NUMERICAL INTERFACE

H.1 FULL PROOF OF THEOREM 2

The parameters bound support and variance. The information class allows any finite atom count, does not require mass at either support endpoint, and does not prescribe a value matrix. If $a = 0$ or $b = 0$, the support and zero mean force $Y = 0$ almost surely. The same holds at $v = 0$; at $\kappa = 1$ the multiplier is constant. All asserted envelopes are then zero. Assume henceforth that $a, b, v > 0$ and $\kappa > 1$.

Since $(Y + a)(b - Y) \geq 0$, taking expectations gives $\mathbb{E}Y^2 \leq ab$. Thus $w = \min\{v, ab\}$ is a valid variance bound. For any admissible finite joint law, the fractional objective as a function of each multiplier is monotone or constant when the other multipliers are held fixed. A maximum therefore uses endpoints. Equivalently, enlarge the probability space by a conditional Bernoulli event A with $P(A \mid X, Y) = (X - 1)/(\kappa - 1)$. Replacing X by $1 + (\kappa - 1)1_A$ preserves $\mathbb{E}X$ and $\mathbb{E}[XY]$ and leaves the law of Y unchanged. The upper-bound argument therefore also applies when X is not initially a deterministic function of Y.

Let $t = P(A)$ and $m = \mathbb{E}[Y1_A]$. Zero mean and the support give $m \leq bt$ and $m = -\mathbb{E}[Y1_{A^c}] \leq a(1 - t)$. Since $\mathbb{E}Y = 0$, covariance Cauchy–Schwarz gives

$$m^2 = \big(\mathbb{E}[Y(1_A - t)]\big)^2 \leq \mathbb{E}Y^2\,\mathbb{E}(1_A - t)^2 \leq wt(1 - t).$$

The positive shift is $(\kappa - 1)m/(1 + (\kappa - 1)t)$. For $0 < t < 1$, every value of the nonnegative upper envelope in (10) is attained: assign $Y = m/t$ on A and $Y = -m/(1 - t)$ on its complement. Its mean is zero, its variance is $m^2/[t(1 - t)] \leq w$, and both support constraints hold. At $t = 0, 1$ the shift is zero.

The transitions of the active minimum occur at $t_b = w/(b^2 + w)$ and $t_a = a^2/(a^2 + w)$. The inequality $w \leq ab$ implies $t_b \leq t_a$. The minimum is bt on $[0, t_b]$, the square-root term on $[t_b, t_a]$, and $a(1 - t)$ on $[t_a, 1]$. After division by $1 + (\kappa - 1)t$, the first term increases and the last decreases. The derivative of the square of the middle term has the sign of $1 - (\kappa + 1)t$. Its maximizer is consequently the clipped $\widehat{t}$ in (8). At that point use

$$Y_+ = \sqrt{w(1 - \widehat{t})/\widehat{t}}, \qquad Y_- = -\sqrt{w\widehat{t}/(1 - \widehat{t})}$$

with probabilities $\widehat{t}$ and $1 - \widehat{t}$. These values lie in the required support, have mean zero and variance w, and attain B_+. Swapping the roles of a, b and replacing Y by $-Y$ proves the absolute statement. Splitting either atom into identical copies yields any larger finite representation with positive probabilities.

For rational inputs, the upper endpoint binds when $b^2 < \kappa w$, and the lower endpoint binds when $w > \kappa a^2$, with boundary equalities also satisfying the interior formula. Both strict conditions would imply $w > ab$, so cannot coexist. Substitution proves (12). All denominators are positive; squaring only compares nonnegative quantities. Hence its rational test is exact, including equality. Increasing any of a, b, v, κ enlarges the class, which proves monotonicity independently of branch changes.

Mechanism of the joint bound. The event reduction can be read as a short construction. Holding all other multipliers fixed, the tilted expectation is fractional-linear in one multiplier, so an endpoint choice is optimal. Repeating over atoms yields an event A tilted by κ. Centering gives $\mathbb{E}[Y1_{A^c}] = -m$. Conditional Cauchy–Schwarz gives the second-moment inequality (11) in the main text. The support separately requires $m/t \le b$ and $m/(1-t) \le a$. Conversely, assign $Y = m/t$ on A and $Y = -m/(1-t)$ on its complement. Every feasible m on the lower envelope in (10) is realized by this explicit construction. The middle, variance branch meets the support branches at t_b and t_a; its peak is feasible only between them. This explains both the clipping rule and why taking the minimum of two separately maximized bounds can lose information: those two maxima may require incompatible event probabilities.

H.2 Expectile relation and the two different separations

For a fixed centered law let $F(z) = \kappa\mathbb{E}[(Y-z)_+] - \mathbb{E}[(z-Y)_+]$. The largest possible $\mathbb{E}[X(Y-z)]$ equals $F(z)$, by choosing $X = \kappa$ above z and $X = 1$ below z. The unique zero of F is therefore the largest tilted expectation. It is the expectile at $\alpha = \kappa/(\kappa+1)$. In the unrestricted mean/variance class, Jakobsons & Vanduffel (2015, Section 3.1, Theorem 5) give the maximum $\sqrt{v}(2\alpha-1)/(2\sqrt{\alpha(1-\alpha)})$ and a two-point attaining law. Substitution gives the interior branch here. The bounded support restricts the optimizing mass; the proof above extends the classical variance envelope to this support-constrained information class.

For $(a, b, v, \kappa) = (1, 25, 16, 9)$, the positive direction binds the lower support endpoint, giving $128/25$; the reverse gives $128/145$. Variance alone gives $16/3$. Omitting the variance constraint sets $w = ab$ and gives $B_+ = (\kappa-1)ab/(b+\kappa a)$; its absolute maximum here is $100/17$. The strict budget $26/5$ lies between the joint and both individual bounds. The stated two-point law attains the joint result although the allowed upper support bound 25 is not occupied. These are exact information-class statements.

For the one-active-value family, $p_i = 1/n$, $V_1 = 1$ and $V_{i>1} = 0$. Its centered support parameters are $a = 1/n$, $b = (n-1)/n$ and $v = ab$. At $\kappa = 9$ and $n \ge 4$, the larger direction gives $8(n-1)/(n(n+8))$. The first atom tilted by nine attains that change; its opposite change is smaller. For $4 \mid n$, a subset of mass $1/4$ also attains TV $1/2$, so the ratio of the centered TV output bound to the exact output bound is $n(n+8)/(16(n-1)) = \Theta(n)$. Because $v = ab$, range/mean alone is already exact here. The two examples isolate the benefit of value information over TV and the additional benefit of enforcing support and variance jointly.

H.3 Sharp correction for the distribution of value error

The original-value decomposition is

$$\pi V - y = (\pi - p)V + p(V - \widehat{V}) + (p\widehat{V} - y).$$

Its residual term is bounded by $E_{\mathrm{val}} \ge \mathbb{E}_p\eta$. Applying the scalar theorem to accessible $\widehat{V}$ instead uses

$$\pi V - y = (\pi - p)\widehat{V} + \pi(V - \widehat{V}) + (p\widehat{V} - y). \tag{J1}$$

The residual must now bound $\mathbb{E}_\pi\eta$. Set $H = \max_i \eta_i$, $e = \min\{H, E_{\mathrm{val}}\}$ and $\kappa = \bar{\kappa}$. At $H = 0$ it vanishes. Otherwise put $R = \eta/H$ and $m = \mathbb{E}_p R \le e/H$. Define $z = \kappa m/(1+(\kappa-1)m)$. For $1 \le X \le \kappa$ and $0 \le R \le 1$,

$$X(R-z) \le \kappa R(1-z) - (1-R)z.$$

Indeed the maximum over the two endpoints of $X(R-z)$ is convex in R, and the right side is its chord from zero to one. The expectation of that chord is $\kappa m(1-z) - (1-m)z = 0$. Therefore $\mathbb{E}[XR]/\mathbb{E}X \le z$, and monotonicity in m proves $\mathbb{E}_\pi\eta \le C_\eta$ in (18). Mass e/H at residual H and

the rest at zero, with high tilt on the former, attains it. It is at most $\min\{H, \kappa e\}$ and reduces to e at $\kappa = 1$. For example $e = 1/2, H = 1, \kappa = 9$ permits residual mean $9/10$ after the tilt, directly disproving uncorrected reuse of e. Separate sharpness of this term and the key term does not assert joint attainability of their triangle sum on a fixed cache.

H.4 Exact outward weighted moments and sound admission

For a coordinate, let $z_i = \widehat{V}_{ic}$ be the exact reconstructed dyadic values, $z_- = \min_i z_i$, $z_+ = \max_i z_i$, and $c = (z_- + z_+)/2$. Intersect the valid output interval $[L, U]$ with $[z_-, z_+]$. A disjoint intersection is an invalid arithmetic contract. Using the saved exponential upper endpoints and positive Z^-, compute

$$T^+ = \frac{\sum_i e_i^+ (z_i - c)^2}{Z^-}, \qquad d_\mu = \min_{\mu \in [L,U]} |\mu - c|, \tag{42}$$

$$v^+ = T^+ - d_\mu^2, \qquad a^+ = U - z_-, \qquad b^+ = z_+ - L. \tag{43}$$

The reference interprets every input float as an exact dyadic and evaluates these expressions with arbitrary-precision rationals. The distance is zero when $c \in [L, U]$ and otherwise the closer endpoint distance. Since $T^+ \geq \mathbb{E}_p(z - c)^2$ and $d_\mu^2 \leq (\mathbb{E}_p z - c)^2$, we have $v^+ \geq \mathrm{Var}_p\, z$. A negative exact result signals an invalid contract rather than a reason to waive an error. The other two bounds follow from the mean enclosure. Class inclusion then permits these outward quantities in the scalar theorem.

Given $\varepsilon \geq 0$, compute $r = \varepsilon - C_\eta - E_{\mathrm{arith}}$ exactly. If $r < 0$, decline. Otherwise both directional tests in (12) must pass for every coordinate. Theorem 2 bounds $\|(\pi - p)\widehat{V}\|_\infty$ by r, the residual lemma bounds the middle term of (J1), and the existing output box bounds its last term. The triangle inequality proves $\|\pi V - y\|_\infty \leq \varepsilon$. Nonfinite inputs, invalid interval dimensions or orders, a nonpositive lower normalizer, a nonzero runtime unusable flag or negative error bounds cannot certify. Structural checks alone do not establish that a supplied interval encloses its target; this is the numerical contract independently verified in the replay.

Constructing weighted second moments requires $O(nd)$ rational arithmetic operations and $O(d)$ retained rational fields in a streaming reduction, plus integer bit costs. The saved-packet reference also materializes the reconstructed matrix and the existing exponential arrays, requiring $O(nd)$ diagnostic storage. The extrema can be stored at cache write and H is one existing FP32 residual maximum, optionally a four-byte scalar. Query-weighted moments require an online reduction. The exact reference serializes arbitrary-precision rational fields. In the fixed replay, their largest numerator or denominator uses 205 bits.

H.5 Discriminating replay and finite checks

The rule and controls were written before this replay, after the previous case outcomes and one packet had been inspected. All 144 packets and seven original budgets were retained. Source hashes, packet identities, the original error enclosures and old control records are checked unchanged. The certificate uses saved exponential endpoints and a centered second moment; an independent oracle derives fresh exact scores, 100-digit directed Decimal exponentials and raw second moments about zero. All 9,216 mean and variance enclosures and all residual-mean bounds pass. Each admission is also checked against the previously independently enclosed actual error.

The exact theorem checker covers 13,420 finite distributions, 142,585 direct vertex tilts and 4,300 two-value attainments. A separate grid covers 225 moment cases, 18,225 mass points, 144 moment attainments and 1,800 rational budget comparisons, including equality and all degeneracies. The interface checker adds 4,782 direct residual tilts, 48 sharp residual examples, nine exact metadata coordinates and 768 envelope/gate comparisons. It explicitly tests the incorrect-distribution counterexample, the strict joint budget and invalid numerical contracts. These checks validate finite instances, degeneracies, and implementation branches alongside the analytical proofs.

The moment control comparison always uses the same C_η and E_{arith}. Its variance-only squared bound is $(\kappa - 1)^2 v^+/(4\kappa)$; range/mean omits the variance bound, setting $w = a^+ b^+$. Reconstructed centered TV uses the actual reconstructed coordinate ranges with the unrestricted TV bound.

Original-value centered TV and scatter retain their original correct E_{val} decomposition. The full case-level record includes all controls, both rational directions, the metadata and independent moment enclosures. The 357 joint admissions contain the 350 admissions in the union of the two moment controls. That historical head-level union leaves seven cases at INT4 budgets .5 (two) and 1 (five). The additional coordinatewise-mixture audit in Section 6 admits 352, reducing the strict remainder to five: two at .5 and three at 1. This comparison reuses the captured model cases and existing timing records.

I Additional zero-centered baseline

For completeness, the initial zero-centered-radius study evaluated $h = 0, 1, 4, 8$ and certified 866 of 4,032 decisions. At INT8 $\rho = .25$, retaining $h > 0$ increases admissions from 59/72 to 61/72. The stronger midpoint-centered comparison in the main text uses 1,008 case/budget pairs for each method and isolates gains beyond that baseline.

J Fixed validation protocol and retained measurements

J.1 Model inputs and selection

The model is `openai-community/gpt2`, at revision

`607a30d783dfa663caf39e06633721c8d4cfcd7e`.

The source is the public safetensors checkpoint, with SHA256

`248dfc3911869ec493c76e65bf2fcf7f615828b0254c12b473182f0f81d3a707`.

The dataset is `Salesforce/wikitext`, with configuration `wikitext-2-raw-v1`, at revision

`b08601e04326c79dfdd32d625aee71d232d685c3`.

Its validation parquet has SHA256

`204929b7ff9d6184953f867dedb860e40aa69c078fc1e54b3baaa8fb28511c4c`.

The dataset's CC BY-SA 3.0 / GFDL attribution is retained with the input.

Exactly the first 64 physical text rows are joined with newline; blank rows are not filtered. Tokenization adds no special tokens, and the first 512 tokens are retained. The longer token stream triggers a tokenizer warning, but the model receives only the saved 512-token slice. Capture uses CPU FP32 evaluation with eager attention, four CPU threads and seed 90260912. Hooks read the `c_attn` projections of layers 0, 6 and 11, reshape to $512 \times 12 \times 64$ and round to FP16 once. All heads and positions 255 and 511 are used, with no selection based on their certificate outcomes. The captured trace defines the exact reference arrays. Bitwise identity across other model runtimes is not assumed.

J.2 Numerical fixtures and model-case checking

The first rational fixtures cover 128 fixed probability settings, 8,160 endpoint distributions, 704 summary/refinement/attainment checks, 1,920 interior-tilt checks, 1,920 scalar output checks and 340 small reduction instances. Their normalized probability and ℓ_1 reference is independent of the f_κ evaluation.

The lower-information suite checks 648 contracts with 12,852 full-atom attainment constructions and 1,944 rational budget comparisons. It compares 4,604 binary64 decisions to exact dyadic-input decisions: 3,221 accept, none falsely accept, and 125 exactly feasible cases remain unresolved. There are also 7,276 finite interval-operation enclosures, 527 numerical-range declines, 20 exponential enclosures checked by an independent rational Taylor remainder, 226 probability comparisons from 15 score-box configurations, and 13 invalid or unsupported cases. These are finite diagnostics, not probabilistic coverage estimates or runtime benchmarks.

The dominant-mass stopping corollary has a separate finite rational check: 540 valid contracts, 2,970 feasible completions, 110,376 direct normalized endpoint-tilt TV comparisons and 540 exact attainments. The check also retains the counterexample outside its premise. It does not import the numerical gate or substitute finite coverage for the proof.

The GPU fixtures check all 63,488 finite FP16 bit patterns, 20 exponential enclosures, ten complete packed cases, 280 output-gate decisions and ten dense-reference cases. A cost-control derivative allows $h = 0$ to skip selection and other h values to collect only their requested masses. Thirty comparisons preserve the nonselection output bits and requested summary prefixes. The original successful implementation and report remain as a predecessor; the fixed model case uses this derivative.

For each model slice, the CPU oracle forms the dot products with exact Fractions. A 100-digit Decimal calculation encloses exponentials and attention output with independent directed operations. It checks both dequantized and original output intervals against the GPU intervals. Every accepted output is compared to the original-output enclosure, and every dense fallback vector is checked separately. Raw packets retain codes, scales, error metadata, scores, exponentials, summary, selected indices, FP32 output and dense-reference output. Verification runs on the same host using independently implemented CPU arithmetic.

One display-only normalized-error field in the original case rows used Decimal's default nearest rounding despite an upper-bound name. A retained derivative recomputes all 144 ratios with upward rounding and exact rational checks. Absolute error enclosures and all admission decisions are unchanged. Reported normalized bounds use this display correction to the saved results.

J.3 Saturation at each summary size

Table 4: Exact rational analysis of the saved lower-mass metadata. Each denominator is 72 model slices. "Empty subset" is the sufficient saturation condition $\eta \geq t_*$; "Any subset" applies the full exact criterion.

Key bits	h	Empty subset	Any subset
4	1	72	72
4	4	70	70
4	8	66	69
8	1	61	61
8	4	36	59
8	8	29	59

These post-run comparisons use the saved κ upper bound and probability lower bounds as exact dyadic values. The counts characterize their information class, not the failure rate of every richer certificate. At $h = 8$ with INT8 keys, $72 - 29 = 43$ slices meet the necessary total-mass condition, and $72 - 59 = 13$ escape saturation. The admission difference is then tested on the previously fixed output-budget grid.

J.4 Storage accounting

Scratch includes two score and two exponential arrays, each of length n in binary64, 16 binary64 state/summary values, eight 32-bit indices, 64 FP32 output values and 128 binary64 output endpoints. Its size is $32n + 1440$ bytes. The algorithm does not create a full dequantized KV array on the GPU. The independent CPU oracle does reconstruct arrays for validation, outside the timed path. Framework allocator changes are retained separately from these logical array counts, which measure allocated data structures rather than total process memory or serving capacity.

J.5 Timing protocol and all budget-level medians

Timing uses only layer 6, head 0, position 511, fixed before outcomes, for each key precision. Three warmups precede 15 retained repetitions. The quantized wall-time path includes kernel launches, the 128-byte state/summary readback, the CPU gate, and, when required, the original-cache transfer

Table 5: Named array bytes for one $n = 512, d = 64$ head. The FP32 radius is stored on the host. A packed host mirror is also retained by this reference; neither it nor the original host cache is omitted from the full record.

Array	INT4 keys	INT8 keys
Key codes on GPU	16,384	32,768
INT4 value codes on GPU	16,384	16,384
FP16 key scales on GPU	2,048	2,048
FP16 value scales on GPU	2,048	2,048
FP32 key residual maxima on GPU	256	256
FP32 value residuals on GPU	2,048	2,048
Packed GPU cache and metadata	39,168	55,552
Runtime scratch on GPU	17,824	17,824
FP16 query on GPU	128	128
Named live quantized GPU arrays	57,120	73,504
Original FP16 KV on CPU	131,072	131,072
Packed mirror on CPU	39,168	55,552
FP32 radius on CPU	4	4

and another 128-byte readback. The return is synchronized before elapsed time is recorded. Dense offload transfers the original cache each time. Dense resident timing places originals on the GPU first and removes unused packed device arrays, so the resident baseline incurs no host-to-device transfer.

Table 6: All full-path timing medians in milliseconds for the two fixed timing cases. Each cell uses 15 repetitions. F denotes fallback in every repetition; Q denotes a certified quantized return in every repetition. These are local reference timings.

Key bits	ρ	$h = 0$	$h = 1$	$h = 4$	$h = 8$
4	0.001	9.535 F	1.529 F	1.641 F	2.105 F
4	0.01	1.677 F	1.513 F	1.635 F	1.937 F
4	0.05	1.722 F	1.545 F	1.722 F	1.984 F
4	0.1	1.744 F	1.538 F	1.694 F	1.970 F
4	0.25	1.719 F	1.637 F	1.586 F	1.958 F
4	0.5	1.806 F	1.573 F	1.603 F	2.100 F
4	1	1.932 F	1.541 F	1.765 F	2.420 F
8	0.001	1.551 F	1.542 F	1.631 F	1.884 F
8	0.01	1.596 F	2.208 F	1.649 F	1.892 F
8	0.05	1.514 F	1.843 F	1.683 F	2.086 F
8	0.1	1.527 F	1.517 F	1.748 F	1.950 F
8	0.25	1.483 F	1.597 F	1.703 F	2.042 F
8	0.5	0.772 Q	0.780 Q	0.943 Q	1.191 Q
8	1	0.763 Q	0.794 Q	0.955 Q	1.179 Q

The CUDA-event measurement records a stream span across the kernel sequence, including launch gaps; it is not the sum of isolated instruction times. The unexplained first INT4 $h = 0$ event group is retained alongside all other raw samples; its unexplained variation limits comparisons of kernel speed across summary sizes. The centered follow-up in Appendix G uses the same heads in a separate measurement phase and preserves the original timing record.

The measured host is Windows 11 with an NVIDIA GeForce RTX 3060 Laptop GPU, compute capability 8.6, Python 3.12.7 and PyTorch 2.5.1+cu121. NVRTC uses `compute_86`, C++14 and the flags `--fmad=false`, `--ftz=false`, `--prec-div=true`, `--prec-sqrt=true`. Generated PTX, compiler options and hashes are retained. The loader uses the CUDA 12.1 runtime bundled with PyTorch and is Windows-specific.

K REPRODUCTION WORKFLOW

Portable certificate replay. The accompanying artifact contains the complete 144-case moment and packed-case records, with all seven original budgets, and two independent certificate checks. Both use Python's standard library, make no model or network calls, and report input SHA-256 checksums. The first optimizes the rational event-envelope extrema at the support transitions and stationary mass. It checks all 1,008 decisions and reproduces Joint, RM, WV, Mix, and head-union admission totals of 357, 270, 342, 352, and 350, respectively. It retains every strict gain. The supplied expected reports allow direct comparison.

Checking the strict-gain mechanism. The second check inspects coordinate 18 at position 255, layer 6, head 3, with INT4 keys and $\rho = 1$. The outward parameters are $(a, b, v) \approx (1.460421, 3.022978, 0.303604)$ and $\bar{\kappa} = 157.764047$. The unconstrained variance optimizer uses mass $1/(\bar{\kappa} + 1) \approx 0.00630$, below the support-compatible threshold $w/(b^2 + w) \approx 0.03215$. The upper-support branch of Equation (12) yields 2.522542 after clipping to that threshold. The checker verifies this branch, the saved squared Joint bound, and the shared remaining budget with exact rational arithmetic.

Retained runtime observations. The artifact also supplies the original centered-radius timing samples and configuration: two key precisions, all seven budgets, fifteen repetitions per gate and the resident and offloaded dense controls. An offline summary check recomputes the medians, admission counts and transfer volumes from all 480 retained timings, and verifies the budgets against the packed-case records. It reproduces Table 3 from the recorded observations without executing a GPU workload.

Replay and capture have different inputs. This artifact starts from the saved outward moments and packed-case records. It verifies the certificate calculations and budget comparisons; it does not regenerate the model capture, independently recompute the Decimal moment enclosures, or repeat GPU timings. Those original procedures and their checks are specified in Appendices H, G, and J. Recapture requires the pinned public GPT-2 checkpoint, the fixed WikiText-2 prefix, Transformers, safetensors, and PyArrow; the reported environment used NumPy 1.26.4 and Transformers 4.57.3. GPU measurements additionally require PyTorch and CUDA. Fresh captures should record their own input hashes and environment; timing is not a bitwise replay target.

Manuscript build. The source archive contains the current figure PDFs and tables. Figure 4 displays Tables 1 and 2; the other analytical figures illustrate the stated information classes and decision rules. The manuscript builds independently of the experimental replay. The artifact documentation provides the replay and build instructions.