ONE SPECTRUM, TWO RESOURCES: DATA–MEMORY SCALING IN AUTOREGRESSIVE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

How much learned memory is needed to benefit from more data? We show that the two resources are governed by one predictive-energy spectrum in a positive-entropy autoregressive retrieval source. Each coordinate contributes its query probability times the squared radius of its unknown logit. Writing μ for the resulting energy spectrum, we prove the minimax law $\mathfrak{R}^*_{\text{value}}(n, B) \asymp_R \Phi_\mu(n^{-1}) + \Phi_\mu(\tau_B), \Phi_\mu(t) = \int \min\{x, t\}\, \mu(\mathrm{d}x)$, for n prediction blocks and a learned state with at most 2^B values. Data set the resolution $1/n$; memory sets the level τ_B reached by optimal bit allocation. The complete curve also recovers the positive spectrum. Energy–dimension pairing is essential: two causal sources with identical block-energy and block-dimension marginals have different data and memory exponents. A masked query–key attention head learns the route and values, realizing the law with explicit routing, format, and arithmetic errors. Further results give exponent-adaptive allocation, finite-precision realization, and compute–precision laws under two-sided arithmetic assumptions. Experiments test held-out resource prediction and resource selection, recover coupling exponents, and examine routing, allocation, and deployment precision across six pretrained-model scales.

1 INTRODUCTION

More data reveal predictive structure; learned memory limits its retention. When should we increase one rather than the other? Empirical laws relate loss to data, model size, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). We seek their source-dependent mechanism: parameters differ in query frequency and predictive influence, so equal parameter accuracy need not minimize loss. What sets another observation's or bit's value?

Each independent unknown scalar parameter is a *coordinate*. Coordinate i has query frequency p_i and a logit—binary target log-odds—in $[-R_i, R_i]$. Its *predictive energy* $q_i = p_i R_i^2$ measures potential prediction error. Energies counted with multiplicity form the energy-per-coordinate spectrum: data and memory read *the same curve at different resolutions*.

Let n count independent examples and B retained bits. Excess loss subtracts the loss of the predictor knowing all parameters. For decreasing energies $q_{(i)} \asymp i^{-(1+\chi)}$, with decay exponent $\chi > 0$ and $\asymp$ denoting equality up to constants, revealing the relevant coordinate gives optimal excess loss $n^{-\chi/(1+\chi)} + B^{-\chi}$. The terms balance at $B \asymp n^{1/(1+\chi)}$: below this scale memory limits retention; above it observations limit learning. Section 3 defines independent causal-sequence blocks and extends the law beyond power laws: the complete curve determines both terms and identifies all positive energies and multiplicities.

Pairing matters: energies $0.8, 0.2$ spread over $1, 4$ coordinates give $(0.8, 0.05, 0.05, 0.05, 0.05)$; reversing the pairing gives five energies 0.2. This finite example illustrates information lost by separate marginals. Our causal construction turns this distinction into different resource exponents.

In blocks of candidates, a query, and a stochastic next token, calibration teaches fixed-key, trainable-query attention the relevant candidate; prediction gradients teach its value. Adaptive allocation and finite-precision evaluation connect the law to implementation under explicit arithmetic assumptions. Synthetic resource tests and pretrained-model quantization probe these mechanisms.

Contributions.

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.

- **One spectrum, two resources.** One truncated-energy curve expresses statistical and finite-state minimax risk and their exponents. A masked-attention learner realizes the law with explicit routing, format, and arithmetic errors.
- **Why separate marginals fail.** Same-marginal causal sources have different data and memory exponents: energy–dimension pairing supplies the information missing from separate scaling descriptors.

2 Related Work

Scaling laws. Empirical laws relate loss to data, parameters, and compute (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022), while spectral and dynamical theories derive task-dependent learning curves (Bordelon et al., 2020; Bahri et al., 2024; Bordelon et al., 2024; Yang, 2025). Knowledge-capacity studies infer retained bits from factual loss and measure bits per parameter (Allen-Zhu & Li, 2025; Morris et al., 2025); compression models derive data and model laws from frequency-ranked knowledge (Pan et al., 2025). We identify the coordinate-energy spectrum governing joint finite-data, finite-state minimax risk and separate its pairing from its marginals.

Finite-rate representation. Rate–distortion, functional quantization, and quantized sequence estimation connect spectra to finite-codebook risk (Shannon, 1959; Luschgy & Pagès, 2004; Zhu & Lafferty, 2014; 2018b). Neural quantization studies heterogeneous precision and extreme low-bit scaling (Micikevicius et al., 2018; Dong et al., 2020; Daliri et al., 2025; Dettmers et al., 2022; Frantar et al., 2023). Using classical testing and bit allocation, we connect rare-coordinate sampling to arbitrary finite-state encoding; energy–dimension pairing then changes both resource exponents.

Minimax and attention learning. Compression and distribution-specific minimax bounds connect concise predictors or source regularity to generalization (Arora et al., 2018; Lotfi et al., 2022; Schmidt-Hieber, 2020). Transformer theory studies prefix learning, next-token generalization, attention dynamics, retrieval, and associative memory (Liang et al., 2025; Yüksel & Flammarion, 2025; Ke et al., 2025; Li et al., 2024; Cabannes et al., 2024; Goel et al., 2026). Sparse approximation and context compression study which token state prediction should retain (Deng et al., 2024; Xiong et al., 2025; Yang & Li, 2026). We quantify routing samples, value-learning samples, and retained bits in one causal source. Appendix A gives theorem-level comparisons and optimization context.

3 The Energy-per-Coordinate Spectrum: Minimax Risk and Identifiability

We first define the source and finite-state value-estimation problem; Section 4 additionally learns the relevant route.

Problem setup

Coordinates and unknown knowledge. A coordinate is an independent scalar prediction parameter. Mode j groups d_j coordinates $i = (j, r)$, $r \in [d_j]$, with $[m] = \{1, \dots, m\}$. The finite/countable set $\mathcal{I}$ has public frequencies $p_i > 0$, $\sum_i p_i = 1$, and radii $0 < R_i \le R < \infty$. Unknown logits $\theta \in \Theta := \prod_{i \in \mathcal{I}} [-R_i, R_i]$ specify binary target probabilities.

One causal prediction block. Sample candidates $C_1, \dots, C_K$ independently from p, then an independent query row Q, uniform on $[K]$, followed by target Y. A fixed unknown matching map $\pi : [K] \to [K]$ selects $I = C_{\pi(Q)}$, with $Y \mid I \sim \mathrm{Bernoulli}(\sigma(\theta_I))$. Here $\sigma(t) = (1 + e^{-t})^{-1}$ is the sigmoid, θ_i is the log-odds at $I = i$, and $\Pr(I = i) = p_i$. Delimiters separate independent causal blocks; n counts blocks, not tokens. Calibration reveals $(Q, \pi(Q))$; prediction targets teach logits. Independent blocks test prediction. Revealing routes/coordinates gives value data $\mathcal{D}_n = ((I_t, Y_t))_{t=1}^n$, separating estimation from retrieval.

Loss and the retained-state constraint. For coordinate predictions $f = (f_i)$, define $\ell(y, z) = -y \log \sigma(z) - (1 - y) \log(1 - \sigma(z))$ and $\mathcal{L}_\theta(f) = \mathbb{E}_\theta[\ell(Y, f_I)]$. Loss is in nats, storage in bits. Bounded logits give positive conditional entropy, Bayes loss $\mathcal{L}_\infty = \mathcal{L}_\theta(\theta) > 0$, and excess loss comparable to frequency-weighted squared logit error, with R-dependent constants. A learner encodes $\mathcal{D}_n$ into $S = \phi(\mathcal{D}_n) \in \mathcal{S}$, $|\mathcal{S}| \le 2^B$, and predicts with a fixed decoder $g(S, i) \in [-R, R]$. Arbitrary computation and joint coordinate coding are allowed, with all data-dependent output in S. For integer $B \ge 0$, minimax excess risk is

$$\mathfrak{R}^*_{\mathrm{value}}(n, B) = \inf_{\phi, g:\, |\mathcal{S}| \le 2^B} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\mathcal{L}_\theta\big(g(\phi(\mathcal{D}_n), \cdot)\big) - \mathcal{L}_\theta(\theta) \right].$$

Expectations average training, learner, and independent test randomness. Side information is fixed/public; learned bits exclude fixed network storage. Constants in $\asymp_R$ depend only on R.

Predictive energy and its spectrum. Define $q_i = p_i R_i^2$, with $\sum_i q_i < \infty$. Frequency scales energy linearly, logit radius quadratically; different frequencies and ranges can give equal energies. With δ_x a unit point mass,

$$\mu = \sum_i \delta_{q_i}, \quad \mathcal{N}_\mu(t) = \mu((t,\infty)), \quad \Phi_\mu(t) = \int \min\{x,t\}\,\mu(\mathrm{d}x) = \int_0^t \mathcal{N}_\mu(s)\,\mathrm{d}s.$$

Thus $\mathcal{N}_\mu(t)$ counts energies above t, including repeats; $\Phi_\mu(t) = \sum_i \min\{q_i, t\}$ caps unresolved contributions. For water level $\tau > 0$, let $\mathcal{B}_\mu(\tau) = \frac{1}{2}\int \log_2(x/\tau)_+ \mu(\mathrm{d}x)$ and define τ_B as the generalized solution of $\mathcal{B}_\mu(\tau_B) = B$, with $\tau_0 = \max_i q_i$. Here $(x)_+ = \max\{x, 0\}$ applies after the logarithm: coordinates above τ receive $\frac{1}{2}\log_2(q_i/\tau)$ bits, others zero. $\mathcal{B}_\mu$ counts continuous bits needed to reach τ; finitely many energies exceed each positive level, even in countable sets.

Energy groups and scaling profiles. An energy block is a coordinate group, distinct from a sampled token block. For total energy a_j, dimension d_j, and equal-energy coordinates, $q_{jr} = a_j/d_j$ and $\mu = \sum_j d_j \delta_{a_j/d_j}$: dimension sets both atom count and size. Profiles $a_j \asymp j^{-(1+\chi)}$, $d_j \asymp j^\gamma$, with $\chi > 0$, $\gamma \geq 0$, describe energy decay and dimension growth; the theorem itself needs no power law.

3.1 One Spectrum Controls Data and Learned State

Theorem 3.1 (One spectrum controls data and memory; informal version of Theorem D.1)**.** *Let $q_i > 0$, $\sum_i q_i < \infty$, and let μ, Φ_μ, and τ_B be defined above. Then*

$$\mathcal{S}_q(n) := \sum_i \min\{q_i, n^{-1}\} = \Phi_\mu(n^{-1}), \tag{3.1}$$

$$D_q(B) := \inf_{\substack{b_i \geq 0 \\ \sum_i b_i \leq B}} \sum_i q_i 2^{-2b_i} = \Phi_\mu(\tau_B). \tag{3.2}$$

For the bounded-logit value experiment,

$$\boxed{\mathfrak{R}^*_{\mathrm{value}}(n, B) \asymp_R \Phi_\mu(n^{-1}) + \Phi_\mu(\tau_B).} \tag{3.3}$$

Moreover, for every $t > 0$,

$$\Phi'_{\mu,+}(t) = \mu((t,\infty)), \qquad \Phi'_{\mu,-}(t) = \mu([t,\infty)). \tag{3.4}$$

Hence the exact curve Φ_μ recovers the positive spectrum, including atom multiplicities, up to coordinate relabeling.

Observations resolve energy at $1/n$; storage resolves it at τ_B. Binary testing gives the statistical lower bound $\min\{q_i, 1/n\}$, and entropy bounds representation error even for joint codes. Estimation and quantization attain their sum, including integer rounding of the continuous allocation. Water filling spends $\frac{1}{2}\log_2(q_i/\tau_B)$ bits above the threshold, equalizing remaining error there. For one coordinate the terms are $\min\{q, 1/n\}$ and $q2^{-2B}$; for the whole spectrum, $\mathcal{N}_\mu$ counts resolved coordinates and Φ_μ accumulates unresolved energy. Thus the spectrum compares the error reductions available from more observations and more retained bits.

Corollary 3.2 (Spectral scaling exponents; informal version of Corollary D.2)**.** *If $\mathcal{N}_\mu(t) \sim t^{-\rho} L(1/t)$ for $0 < \rho < 1$, where L is slowly varying, then $\Phi_\mu(n^{-1}) \in \mathrm{RV}_{-(1-\rho)}$, $\Phi_\mu(\tau_B) \in \mathrm{RV}_{-(1-\rho)/\rho}$. For $a_j \asymp j^{-(1+\chi)}$ and $d_j \asymp j^\gamma$, $\rho = (1+\gamma)/(1+\chi+\gamma)$, giving rates $n^{-\chi/(1+\chi+\gamma)}$ and $B^{-\chi/(1+\gamma)}$, up to slowly varying factors.*

Here $g \in \mathrm{RV}_\alpha$ means $g(cx)/g(x) \to c^\alpha$ as $x \to \infty$, for every $c > 0$. One small-atom tail index therefore determines both exponents.

3.2 Why Separate Marginals Cannot Identify Scaling

A marginal descriptor lists group energies and dimensions but forgets their pairing. Concentrating high energy in one-coordinate groups makes it easier to learn and retain than spreading it over larger groups. The next theorem compares these assignments exactly at each dyadic level ℓ, where group counts are proportional to 2^ℓ.

Theorem 3.3 (Same block marginals, different scaling laws; informal version of Theorem D.6)**.** *For every $\chi, \gamma > 0$, there are two positive-conditional-entropy causal-retrieval sources with exactly the same block-energy multiset and the same block-dimension multiset at every dyadic level. They differ only in the pairing. All displayed dimensions in this construction are rounded upward to integers. With the common fixed-K route revealed,*

$$\mathfrak{R}_{\text{easy}}(n, B) \asymp n^{-\chi/(1+\chi)} + B^{-\chi}, \text{ whereas } \mathfrak{R}_{\text{hard}}(n, B) \asymp n^{-\chi/(1+\chi+\gamma)} + B^{-\chi/(1+\gamma)}.$$

At dyadic level ℓ, both sources have 2^ℓ high- and low-energy blocks with energies proportional to $2^{-(1+\chi)\ell}$ and $2^{-(1+\chi+\delta)\ell}$, and 2^ℓ copies each of dimensions 1 and $2^{\gamma\ell}$. Choose $\delta > \gamma\chi$. Pair high energy with dimension one (easy) or $2^{\gamma\ell}$ (hard), preserving both block marginals exactly.

Block-energy exponent s and dimension exponent g give tail index $(1 + g)/(1 + s + g)$: high energies yield $1/(1+\chi)$ (easy) versus $(1+\gamma)/(1+\chi+\gamma)$ (hard), while δ makes low energies lower order. Corollary 3.2 gives all four exponents; Appendix J supplies the construction and converse. Pairing is thus indispensable to a scaling descriptor, and Equation (3.4) recovers its coordinate-energy spectrum.

4 Learning the Causal Self-Attention Predictor

Can learned routing preserve the resource exponents? Exact-arithmetic stochastic gradient descent (SGD) learns routes (Stage A), then values (Stage V). Their combined sample and state costs retain the exponents under the conditions below.

From a route table to query–key attention. Selectors $P_Q, P_C \in \mathbb{R}^{K\times d}$ extract fixed orthogonal addresses from query/candidate tokens $x_q, x_s \in \mathbb{R}^d$: $P_Q x_q = e_q$ identifies the query row, $P_C x_s = e_s$ the candidate position; e_s is a standard basis vector. For head dimension $d_h \geq K$, embed addresses by $J \in \mathbb{R}^{d_h\times K}$, $J^\top J = I_K$. Trainable query–position scores $A \in \mathbb{R}^{K\times K}$ give fixed keys and trainable queries:

$$W_K = \sqrt{d_h} J P_C, \qquad W_Q(A) = J A^\top P_Q, \qquad \frac{\langle W_Q(A) x_q, W_K x_s\rangle}{\sqrt{d_h}} = A_{qs}. \tag{4.1}$$

Proposition 4.1 (Exact query–key realization; informal version of Proposition D.3)**.** *With a strict causal mask over the K preceding candidates, Equation (4.1) exactly realizes the route scores and attention. Fix W_K and train only W_Q's query-address block. Euclidean SGD then gives the same A update, attended feature, and output-vector gradient.*

Unknown routes/logits reside in trained query-address/output coordinates, not fixed embeddings, keys, or value extraction.

Learning the route. Calibration raises a query row's matching score above zero-initialized, equal distractors. After t visits, β_t is their score gap and δ_t their total attention mass.

Proposition 4.2 (Online routing dynamics; informal version of Proposition D.4)**.** *For $K \geq 2$, zero initialization and softmax cross-entropy steps of size $0 < \eta \leq 1$ give the margin and distractor mass after t row visits: $\beta_{t+1} = \beta_t + \eta \frac{K}{e^{\beta_t}+K-1}$, $\delta_t = \frac{K-1}{e^{\beta_t}+K-1}$. After $b_K = \Theta_\eta(\log(eK))$ visits, $\delta_t \asymp_\eta [1+(t-b_K)_+]^{-1}$. For S_A independent calibration blocks and a held-out row, $\mathfrak{r}_K(S_A) = \mathbb{E}[\delta_N^2]$, $N \sim$ Binomial$(S_A, 1/K)$. For $0 < \varepsilon \leq 1/16$, the time to $\mathfrak{r}_K \leq \varepsilon$ is $\Theta_\eta(K\{\varepsilon^{-1/2} + \log(eK)\})$; at $K = 1$, leakage and calibration cost vanish.*

Learning and storing the values. Freeze the calibrated route and fit resolvable coordinates $A_n = \{i : q_i \geq n^{-1}\}$ in the natural box $\mathcal{W} = \prod_{i\in A_n}[-R_i, R_i]$, setting others to zero. The attended feature h_t contains attention-weighted coordinate indicators on A_n. With $\mu_{\text{sgd}} = \frac{9}{16}\min_{|z|\leq R}\sigma'(z) > 0$,

start at $w_1 = 0$, use $\eta_t = 2/[\mu_{\mathrm{sgd}}(t+1)]$, and return $\bar{w}_n = 2\sum_{t=1}^{n} t w_t/[n(n+1)]$ (zero if A_n is empty). Restrict D to A_n; projection $\Pi^D_{\mathcal{W}}$ uses $\|v\|_D^2 = \sum_{i\in A_n} p_i v_i^2$. The update is

$$w_{t+1} = \Pi^D_{\mathcal{W}}\left[w_t - \eta_t D^{-1}\{\sigma(w_t^\top h_t) - Y_t\}h_t\right], \qquad D = \mathrm{diag}(p). \tag{4.2}$$

Encode fitted values in scalar formats using B_V bits; the stored-range/logit-radius ratio κ_i amplifies error (unit ratios recover Section 3). For the causal predictor $\widehat{F}$, let $\mathcal{R}_\theta(\widehat{F})$ denote complete-block population cross-entropy and $\mathcal{E}_{\mathrm{arithmetic}}$ deployment error under the formal theorem's numerical assumptions.

Theorem 4.3 (Causal self-attention realization; informal version of Theorem D.5)**.** *The weighted-average SGD with scalar formats in Theorem D.5 satisfies*

$$\mathbb{E}[\mathcal{R}_\theta(\widehat{F}) - \mathcal{L}_\infty] \lesssim_{R,K,\eta} \Phi_\mu(n^{-1}) + D^{\mathrm{fmt}}_{q,\kappa}(B_V) + \mathfrak{r}_K(S_A) + \mathcal{E}_{\mathrm{arithmetic}}, \tag{4.3}$$

where B_V is the value-state budget and $\kappa_i \geq 1$ the public format-range ratio, with

$$D^{\mathrm{fmt}}_{q,\kappa}(B_V) = \inf_{b_i\geq 0,\, \sum_i b_i \leq B_V} \sum_i q_i \min\{1, \kappa_i^2 2^{-2b_i}\}.$$

For $\kappa_i = 1$, this is $\Phi_\mu(\tau_{B_V})$, and the core surface holds at total budget B for power-law spectra, fixed K, $o(B)$ routing bits and lower-order routing/arithmetic errors.

Preconditioning gives conditional strong convexity in the frequency-weighted norm (gradient, state, and format details: Appendix C).

The exponents survive routing. With B_A routing bits, total state/blocks are $B = B_V + B_A$, $N = n + S_A$. For decreasing energies $q_{(r)} \asymp r^{-(1+\xi)}$, $\xi > 0$, fixed K, unit midpoint formats, and the deployment-error condition of Proposition T.1, risk ε requires at most $N = n + S_A = O(\varepsilon^{-(1+\xi)/\xi})$ blocks and $B = B_V + B_A = O(\varepsilon^{-1/\xi})$ learned bits. Calibration and route-state costs are lower order, preserving both exponents.

5 Adaptive Representation and Finite-Precision Realization

How much must an implementation know about the spectrum? The first result allocates bits without its decay exponent; the second evaluates a supplied logit table in finite precision. Scalar j, Δ_j denote i, R_i of Section 3; population loss $\mathcal{L}_\theta$ is unchanged.

Adaptive representation.

Theorem 5.1 (Adaptation with a known ordering; informal version of Theorem I.1)**.** *Fix $0 < \chi_{\min} \leq \chi_{\max} < \infty$, and suppose the canonical mode order is known. Assume $q_j \asymp j^{-(1+\chi)}$ with common comparison constants and unknown $\chi \in [\chi_{\min}, \chi_{\max}]$. For every $n, B \in \mathbb{N}$, one clipped likelihood estimator and one deterministic integer bit schedule, both independent of χ, satisfy simultaneously throughout this interval*

$$\sup_{\theta_j \in [-\Delta_j, \Delta_j]} \mathbb{E}_\theta\left[\mathcal{L}_\theta(\widehat{f}) - \mathcal{L}_\theta(\theta)\right] \lesssim n^{-\chi/(1+\chi)} + B^{-\chi}. \tag{5.1}$$

Choose $A > \chi_{\max}/2$ and the largest prefix m satisfying

$$b_j = \lfloor A \log_2(m/j) \rfloor_+ \quad (j \leq m), \qquad \sum_{j\leq m} b_j \leq B. \tag{5.2}$$

Then $m \asymp B$. The schedule favors leading coordinates and matches quantization to omitted-tail error at $O(m^{-\chi})$, uniformly over the exponent interval.

Finite-precision self-attention. Retain a supplied table's first M logits and use zero elsewhere. Let $t_\star$ be the matching position, $\beta \geq 0$ its score, and j_t candidate t's table index. With public basis addresses $a_t \in \mathbb{R}^K$, mode vectors $e_j^{(M)} = e_j$ for $j \leq M$ (zero otherwise), and $d = K + M + 1$, tokens are $x_t = (a_t, e_{j_t}^{(M)}, 0)$ and $x_q = (a_{t_\star}, 0, 1)$. The fixed-key construction in Proposition 4.1 gives

$$\frac{\langle W_Q x_q, W_K x_t\rangle}{\sqrt{d_h}} = \beta\langle a_{t_\star}, a_t\rangle. \tag{5.3}$$

Project values and output vector w onto mode coordinates and apply a strict causal mask:

$$h_q = \sum_{t=1}^{K} \alpha_t \bar{e}_{j_t}^{(M)}, \qquad \alpha_t = \frac{\exp(\beta \mathbf{1}\{t = t_\star\})}{e^\beta + K - 1}, \qquad z = w^\top h_q, \tag{5.4}$$

where $\bar{e}_j^{(M)} = (0, e_j^{(M)}, 0)$, $w_j^{\mathrm{mode}} = w_{K+j}$, and $\vartheta_j = w_j^{\mathrm{mode}}$ for $j \le M$ (zero otherwise). Unit roundoff u bounds scalar relative errors. The bound's four terms measure distractor selection, rounding, logit storage, and omitted coordinates.

Theorem 5.2 (Finite-precision realization; informal version of Theorem K.1)**.** *For the construction above, fix $K, M \ge 1$, $R < \infty$, and $\beta \ge 0$.*

1. *In exact arithmetic, $|w_j^{\mathrm{mode}}| \le R$ for all $j \in [M]$ gives, for any matching mode $j \ge 1$,*

$$|z - \vartheta_j| \le 2R(K-1)e^{-\beta}. \tag{5.5}$$

2. *Uniformly quantize $w_j^{\mathrm{mode}} = \theta_j$, $j \in [M]$, using $b_j \in \mathbb{N}_0$ bits on $[-\Delta_j, \Delta_j]$. Use max-shifted softmax with exact shifts $0, -\beta$ and exponent range containing their exponentials. All exponentials and arithmetic return finite values with relative error $\le u$, preserving exact zeros. Both denominator and K-candidate output sums are balanced pairwise. If $(\lceil \log_2 K \rceil + 4)u \le 1/4$, the resulting predictor f_{fp} has held-out excess cross-entropy*

$$\mathcal{L}_\theta(f_{\mathrm{fp}}) - \mathcal{L}_\infty \le C_R \left\{ (K-1)^2 e^{-2\beta} + u^2(1 + \log K)^2 + \sum_{j=1}^{M} q_j 2^{-2b_j} + \sum_{j>M} q_j \right\}. \tag{5.6}$$

3. *For separable scalar tables with at most 2^{b_j} values at coordinate j, the bit-dependent term in Equation (5.6) is tight up to constants.*

Here and below, C_R depends only on the logit bound R.

Pairwise summation makes arithmetic error logarithmic in K before squaring. Theorem 4.3 learns the logits evaluated here; dense storage and proofs are in Appendices C, I, K.

6 Bit Complexity and Compute–Precision Scaling

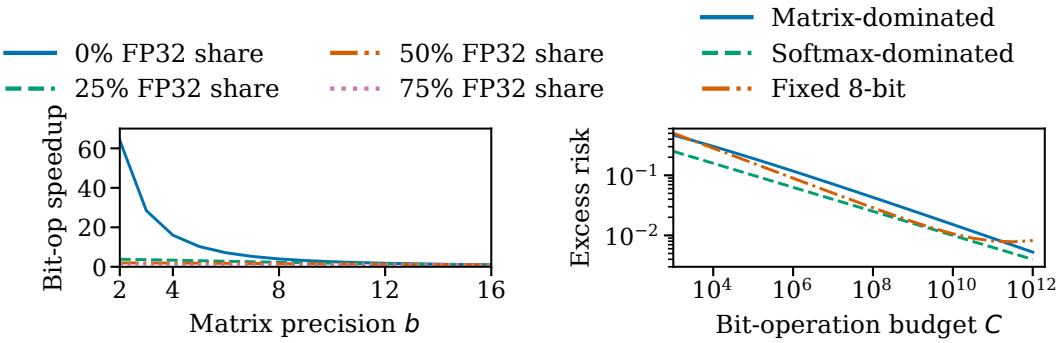


Figure 1: **Retained precision sets resource ceilings and error floors.** Left: analytic bit-operation ratio with retained high-precision work. Right: risk–cost laws under the stated arithmetic model; fixed softmax precision creates a numerical floor.

We compare growing-prefix precision, dense-graph cost, and fixed-shape savings.

Error and cost as model size grows. Let M count represented modes, b their arithmetic bits, and $\kappa(M)$ the amplification of stored perturbations. Assume model error $\Theta(M^{-\chi})$, $\chi > 0$; per-block work $\Theta(M^f)$, $f \ge 0$; necessary and sufficient sample count $\Theta(M^{1+\chi})$; scalar bit-operation cost $\Theta(b^\nu)$, $\nu \ge 0$; and perturbation error $\Theta(\kappa(M)^2 2^{-2b})$. Under these two-sided laws, $\mathcal{R}(C)$ is optimal held-out risk at arithmetic budget C.

Counting a dense forward–backward pass. For dense Transformers of length ℓ, width d, feed-forward width m_{ff}, H heads, and L layers, matrix/softmax precisions $b_{\mathrm{mm}}, b_{\mathrm{sm}}$ give bit cost, up to constants, $L\left[(\ell d^2 + \ell d m_{\mathrm{ff}} + \ell^2 d)\mathsf{M}(b_{\mathrm{mm}}) + H\ell^2\mathsf{S}(b_{\mathrm{sm}})\right]$. M, S are bit costs per matrix multiply–accumulate and softmax scalar operation. Excluding optimizer updates, recomputation saves attention state without changing the dense arithmetic order.

Theorem 6.1 (Precision-aware compute phases; informal version of Theorem N.1)**.** *For the two-sided risk–cost family of Theorem N.1, with $\chi > 0$, $f, \nu \geq 0$, and $n(M) \asymp M^{1+\chi}$:*

1. *If $1 \lesssim \kappa(M) \lesssim M^{\rho}$, then the least model-error-matching precision is $b = \Theta(\log M)$, and*
$$\mathcal{R}(C) - \mathcal{L}_\infty \asymp C^{-\chi/(f+1+\chi)}(\log C)^{\chi\nu/(f+1+\chi)} \tag{6.1}$$
up to other slowly varying factors specified in the assumptions.
2. *If $\kappa(M) = \exp(\Theta(M^{\lambda}))$, $\lambda > 0$, then $b = \Theta(M^{\lambda})$, and the loss exponent is $\chi/(f+1+\chi+\lambda\nu)$.*
3. *If $q_j \asymp j^{-(1+\chi)}$, adaptive per-mode precision has error $\Theta(B^{-\chi})$, whereas a uniform-width prefix has optimal error*
$$\Theta((B/\log B)^{-\chi}). \tag{6.2}$$

Each converse uses the representation, operation set, storage layout, and bit-operation cost specified above.

Here ρ describes conditioning, distinct from Section 3's spectral tail index. Matching numerical error to $M^{-\chi}$ requires

$$b \geq \log_2 \kappa(M) + \frac{\chi}{2}\log_2 M - O(1). \tag{6.3}$$

Savings at a fixed shape. At fixed dimensions, let G, G_{sm} count matrix/softmax operations, ν, ν_{sm} be their precision-cost exponents, N_q count quantized scalars, and R_{hp} retained high-precision bits. Lowering matrix and stored-scalar precision b_0 to b gives

$$\mathrm{BitOpsRatio} = \frac{Gb_0^{\nu} + G_{\mathrm{sm}}b_{\mathrm{sm}}^{\nu_{\mathrm{sm}}}}{Gb^{\nu} + G_{\mathrm{sm}}b_{\mathrm{sm}}^{\nu_{\mathrm{sm}}}}, \qquad \mathrm{Compression}_{\mathrm{bit}} = \frac{N_q b_0 + R_{\mathrm{hp}}}{N_q b + R_{\mathrm{hp}}}.$$

Ratios above one mean savings, capped by retained fixed-precision softmax, master weights, optimizer state, or activations as matrix bits fall (an Amdahl-type ceiling).

Appendix C derives dense pass counts, storage, and precision–risk laws.

7 Experiments

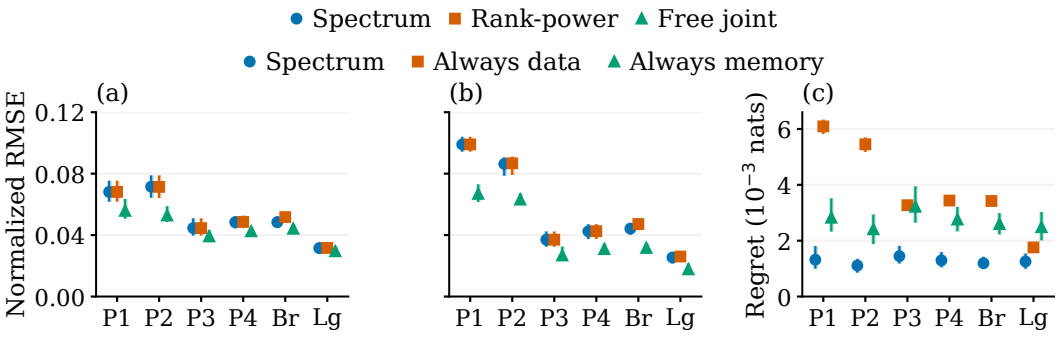


Figure 2: **Predicting risk and choosing resources.** (a,b) Data/storage extrapolation RMSE divided by zero-predictor risk; upper legend compares full-spectrum, rank-power, and free joint responses. (c) Regret for fourfold data or memory expansion, with staying allowed; lower legend compares selection rules. P1–P4: the four power profiles; Br/Lg: broken-power/log-corrected sources. All use weighted SGD, $K = 32$, 256 energy groups; intervals: 95% BCa over 16 paired seeds.

Experimental design. Spectral, coupling, and learning tests compare excess cross-entropy (CE) above Bayes loss on independent training/test blocks, paired by seed or item. Fixed-source intervals use 95% bias-corrected and accelerated (BCa) seed bootstraps; mini-batch and terminal-iterate variants test robustness. Appendices R, S give algorithms, units, and uncertainty procedures.

Table 1: **Deployment, pairing, and learning mechanisms.** (a) HQQ-g32 minus BF16 CE: 96 WikiText-103 windows, 512 LAMBADA items. (b) Storage slopes, 16 seeds; E/H: easy/hard. (c,d) SGD decomposition/interventions: 10^{-3} nats, eight seeds; (d) has $(\chi, \gamma) = (1, .5)$, $M = 64$, $K = 32$, bit ratio .5 (Table 3: 1). $\pm$: rounded 95% BCa envelopes (Unif.: constant across seeds). Bold/underline: lowest/second-lowest in (a,d).

(a) Pythia weight-only PTQ; lower is better.

Model	WikiText	LAMBADA
70M	3.009 $\pm$0.037	3.012 $\pm$0.386
160M	1.325 $\pm$0.020	0.589 $\pm$0.225
410M	0.156 $\pm$0.003	0.319 $\pm$0.082
1B	<u>0.054</u> $\pm$0.002	<u>0.021</u> $\pm$0.028
1.4B	0.066 $\pm$0.003	0.149 $\pm$0.045
2.8B	**0.048** $\pm$0.003	**0.007** $\pm$0.034

(b) Same-marginal coupling.

$(\chi, \mathrm{E/H})$	Theory	Held-out CE
$(1, E)$	1.000	0.972 $\pm$0.0016
$(1, H)$	0.500	0.503 $\pm$0.0003
$(2, E)$	2.000	2.016 $\pm$0.0011
$(2, H)$	1.000	1.002 $\pm$0.0002

(c) Signed full-source excess-CE terms.

Data	Tail	Route	Format
3.584 $\pm$.219	1.218 $\pm$.029	.0270 $\pm$.0031	1.491 $\pm$.275

(d) Interventions; excess CE, lower is better.

Full	Known	No pre.	Frozen	Unif.
19.01 $\pm$.27	<u>19.06</u> $\pm$.26	22.47 $\pm$.21	60.50 $\pm$.74	89.65 $\pm$.00

Data–memory prediction and resource choice. Six sources cross five sample and six state budgets. FP64 weighted SGD restarts per sample budget; storage budgets reuse the fitted vector. Validation excludes the largest budgets. Free joint power laws best predict single-axis test CE; the spectrum best predicts joint expansion (Figures 2 and 12). Unfitted G_μ selects fourfold expansion with lower mean regret than either fixed-resource choice on every source. Rank-power closely matches single-axis predictions. Appendix U gives all baselines; Figures 11 and 14 and Table 4b give collapse, unbinned diagnostics, and exponents.

Same-marginal coupling. On one normalized dyadic source, easy/hard data exponents are $0.509/0.339$ ($\chi = 1$) and $0.681/0.489$ ($\chi = 2$), against $1/2, 1/3$ and $2/3, 1/2$. Fits use the same last four budgets for all 16 seeds; Figure 3(a,d) shows every budget. Together with storage separation in (b,e) and Table 1b, this shows identical block marginals can hide different exponents, not just constants.

Learned routing and interventions. With more visits, mini-batch calibration reduces leakage toward known routing; frozen uniform routes lag (Figure 3(c,f)). Tables 1d and 3 vary one factor (No pre.: no preconditioning; Unif.: uniform precision). On identical full-source contexts, theorem-matched runs evaluate four counterfactuals, fixing all other quantities. Learned-minus-known routing and quantized-minus-unquantized losses plus unquantized known-route loss sum to deployed excess CE. Splitting the last term over active/omitted targets yields the signed terms in Table 1c (Proposition T.2).

Stored formats and softmax precision. Three format-range/logit-radius regimes compare shared, group, and coordinate scales and adaptive widths; softmax tests track precision, underflow, and matching mass at $K \in \{8, 32, 128, 512\}$. Eight-bit adaptive distortion is 7.82×10^{-10}, versus 8.46×10^{-6} (uniform block widths) and 1.08×10^{-5} (shared scale). Appendix Figure 8 shows allocation controls distortion, while quantized fraction and retained high-precision work determine savings.

Profile robustness. Broken-power, log-corrected, and perturbed-order tests use exact finite spectra and eight paired seeds. Count-corrected fitted energies guide empirical allocation; known-energy allocation minimizes true-energy distortion. These objectives differ from Theorem 5.1's deterministic unknown-exponent schedule (Appendix S). Empirical allocation improves on uniform widths in all three cases (Figure 6).

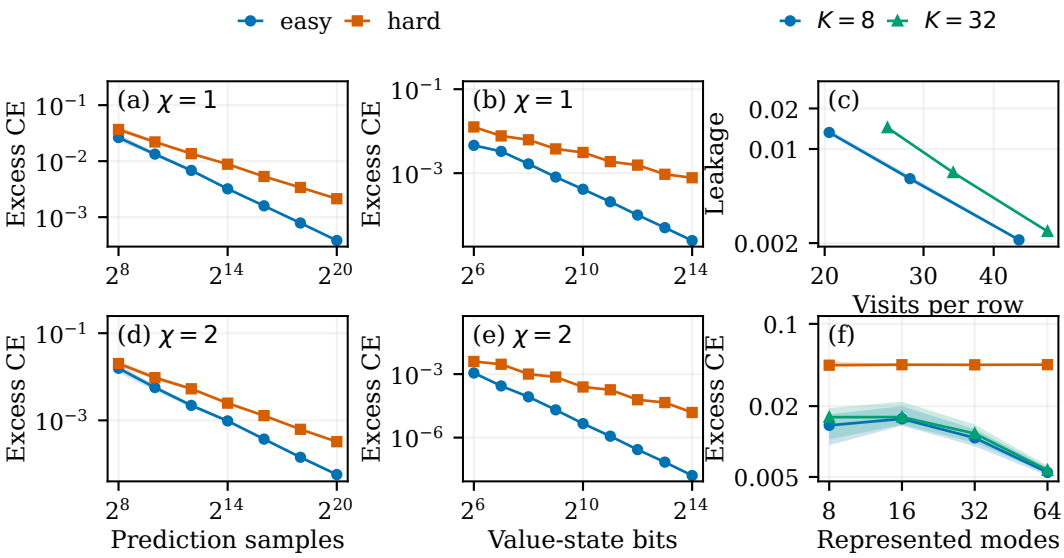


Figure 3: **Pairing and learned routing.** (a,d) Fixed-source data and (b,e) population-CE storage, $\chi = 1, 2$ (blue/orange: easy/hard). (c) Squared leakage versus visits, $K = 8/32$ (blue/green). (f) Finite-source excess CE without added tail: learned/known/frozen routes (blue circles/green triangles/orange squares). 95% BCa: 16 data seeds (a,d), 16 amplitudes (b,e), eight paired seeds (c,f).

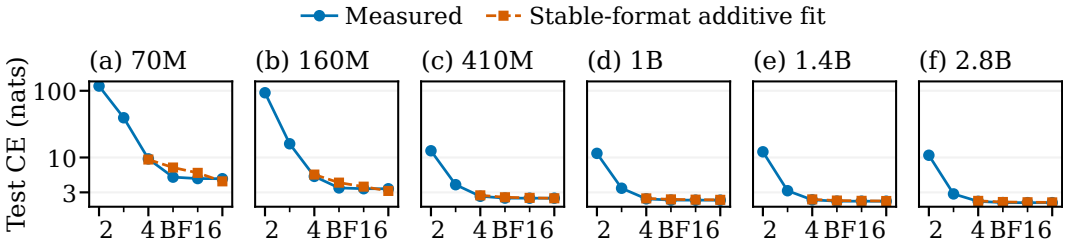


Figure 4: **Data and deployed storage in Pythia.** (a–f) 70M, 160M, 410M, 1B, 1.4B, 2.8B test CE after 2^{26} extra tokens. Blue: 2/3/4/6/8-bit and BF16 (left to right); dashed orange: additive fits to first-three-budget 4/8-bit/BF16 validation. All points hold out data; 6-bit also holds out format. Shared log-y; bands: 95% BCa, eight paired training seeds.

Pretrained Transformers. Six-scale weight-only post-training quantization (PTQ) with half-quadratic 32-weight groups (HQQ-g32) has decreasing penalties except at 1.4B; 2.8B's LAMBADA interval includes zero (Table 1a). Continued training scans four token budgets and six formats (Figure 4). Additive 4/8-bit/BF16 fits reduce extrapolation root-mean-square error versus data-only fits at every scale ($0.910/0.380$ at 70M/160M). 6-bit interpolation improves at 410M–2.8B, but worsens at 70M/160M. 2-bit calibration fits 410M–2.8B; 70M/160M losses exceed its $[0, 20]$-nat range. Softmax stays FP32. All formats and errors appear in Figures 7 and 9 and Table 4a and Tables 5 to 8. A fixed-penalty baseline improves 70M extrapolation; additive fits improve the other five scales and every format interpolation (Appendix U.1).

8 Conclusion

Finite data and learned memory resolve the same spectrum at different levels, giving causal-source minimax risk, exponents, and the resource crossover. The complete curve identifies positive energies beyond a fitted exponent; energy–dimension pairing gives identical marginals different resource laws. Masked attention learns routes and values; adaptive allocation, finite-precision realization, and conditional compute laws explain representation costs. Held-out prediction and resource choices test this organizing principle. The distribution of $q_i = p_i R_i^2$ determines what data reveal and what finite memory retains.

AI USE STATEMENT

Generative artificial-intelligence tools were used for proof exploration, code assistance, experiment orchestration, figure and table layout, and manuscript editing. The authors reviewed the resulting proofs, claims, citations, code, experimental records, and rendered paper, and take responsibility for the final content and artifacts.

ETHICS STATEMENT

The theoretical sources are synthetic, and the pretrained-model study uses public language-model checkpoints and standard public evaluation corpora. We did not collect human-subject or sensitive personal data.

REPRODUCIBILITY STATEMENT

The source model, assumptions, and complete proofs are provided in the appendices. Appendix R specifies every synthetic grid, model revision, dataset split, numerical format, uncertainty estimator, software stack, and execution environment. The accompanying code preserves seed- and item-level outputs from which all figures and tables are generated.

REFERENCES

Ibrahim M. Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263, 2018.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), 2024. doi: 10.1073/pnas.2311878121.

Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Tomer Berg, Or Ordentlich, and Ofer Shayevitz. Deterministic finite-memory bias estimation. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 566–585, 2021.

Yaniv Blumenfeld, Dar Gilboa, and Daniel Soudry. A mean field theory of quantized deep networks: The quantization–depth trade-off. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4345–4382, 2024.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5114–5134, 2024.

Jerry Chee, Arturs Backurs, Rainie Heck, Li Zhang, Janardhan Kulkarni, Thomas Rothvoss, and Sivakanth Gopi. DiscQuant: A quantization method for neural networks inspired by discrepancy theory. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 924–951, 2025.

Jianfei Chen, Yu Gai, Zhewei Yao, Michael W. Mahoney, and Joseph E. Gonzalez. A statistical framework for low-bitwidth training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 883–894, 2020.

Timothy Chu, Zhao Song, and Chiwun Yang. How to protect copyright data in optimization of large language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (16):17871–17879, 2024. doi: 10.1609/aaai.v38i16.29741. URL https://ojs.aaai.org/index.php/AAAI/article/view/29741.

Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George Bm Van Den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J. Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc'Aurelio Ranzato, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 4057–4086, 2022.

Majid Daliri, Zhao Song, and Chiwun Yang. Unlock the theory behind scaling 1-bit neural networks. In *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 545–598, 2025. URL https://proceedings.mlr.press/v280/daliri25a.html.

Amit Daniely and Elad Granot. Generalization bounds for neural networks via approximate description length. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yichuan Deng, Zhao Song, Jing Xiong, and Chiwun Yang. How sparse attention approximates exact attention? your attention is naturally n^C-sparse, 2024.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Vitaly Feldman, Guy Kornowski, and Xin Lyu. Trade-offs in data memorization via strong data processing inequalities. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1935–1973, 2025.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Gautam Goel, Mahdi Soltanolkotabi, and Peter L. Bartlett. Training dynamics of softmax self-attention: Fast global convergence via preconditioning, 2026.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44 (6):2325–2383, 1998. doi: 10.1109/18.720541.

Hassan Hafez-Kolahi, Behrad Moniri, Shohreh Kasaei, and Mahdieh Soleymani Baghshah. Rate-distortion analysis of minimum excess risk in bayesian learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3998–4007, 2021.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Yekun Ke, Yingyu Liang, Zhenmei Shi, Zhao Song, and Chiwun Yang. Curse of attention: A kernel-based perspective for why transformers fail to generalize on time series forecasting and beyond. In *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 675–738, 2025. URL https://proceedings.mlr.press/v280/ke25a.html.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 685–693, 2024.

Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34943–34975, 2025.

Yingyu Liang, Zhenmei Shi, Zhao Song, and Chiwun Yang. Towards infinite-long prefix in transformer. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11127–11191. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.563. URL https://aclanthology.org/2025.emnlp-main.563/.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G. Wilson. PAC-bayes compression bounds so tight that they can explain generalization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Harald Luschgy and Gilles Pagès. Sharp asymptotics of the functional quantization problem for gaussian processes. *The Annals of Probability*, 32(2):1574–1599, 2004. doi: 10.1214/009117904000000324.

Harald Luschgy, Gilles Pagès, and Benedikt Wilbertz. Asymptotically optimal quantization schemes for gaussian processes on hilbert spaces. *ESAIM: Probability and Statistics*, 14:93–116, 2010. doi: 10.1051/ps:2008026.

Alexander Maloney, Daniel A. Roberts, and James Sully. A solvable model of neural scaling laws, 2022.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize?, 2025.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

NVIDIA. Transformer engine 2.18.0: Attention is all you need!, 2026. Software documentation, accessed 2026-08-31.

Zhixuan Pan, Shaowen Wang, Liao Pengfei, and Jian Li. Understanding LLM behaviors via compression: Data generation, knowledge acquisition and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, pp. 168404–168454, 2025. doi: 10.52202/085713-5613. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f61d7778e89b9221d1ea0ce8428b7014-Abstract-Conference.html.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Pedro Henrique Pamplona Savarese, Xin Yuan, Yanjing Li, and Michael Maire. Not all bits have equal value: Heterogeneous precisions via trainable noise. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, volume 4, pp. 142–163, 1959.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold, 2020.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1564–1587, 2015.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1405–1413, 2024.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099, 2023.

Jing Xiong, Jianghan Shen, Chuanyang Zheng, Zhongwei Wan, Chenyang Zhao, Chiwun Yang, Fanghua Ye, Hongxia Yang, Lingpeng Kong, and Ngai Wong. ParallelComp: Parallel long-context compressor for length extrapolation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68998–69016, 2025. URL https://proceedings.mlr.press/v267/xiong25b.html.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Tight sample complexity of transformers. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6887–6923, 2026.

Chiwun Yang. Unifying learning dynamics and generalization in transformers scaling law, 2025.

Chiwun Yang and Xiaoyu Li. Beyond sparse weights: When is attention compressible?, 2026.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V3: Dyadic neural network quantization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11875–11886, 2021.

Oğuz Kaan Yüksel and Nicolas Flammarion. On the sample complexity of next-token prediction. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 694–702, 2025.

Yuancheng Zhu and John Lafferty. Quantized estimation of gaussian sequence models in euclidean balls. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Yuancheng Zhu and John Lafferty. Distributed nonparametric regression under communication constraints. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 6009–6017, 2018a.

Yuancheng Zhu and John Lafferty. Quantized minimax estimation over sobolev ellipsoids. *Information and Inference: A Journal of the IMA*, 7(1):31–82, 2018b. doi: 10.1093/imaiai/iax007.

APPENDIX

A More Related Work

Empirical and mechanistic scaling laws. Extended power-law regimes have been reported across data, model size, and compute (Hestness et al., 2017; Kaplan et al., 2020; Henighan et al., 2020), with compute-optimal allocation separating parameters and tokens (Hoffmann et al., 2022). Routed models, broken-power-law fits, and replications show that apparent exponents depend on the resource definition, model family, and training prescription (Clark et al., 2022; Alabdulmohsin et al., 2022; Caballero et al., 2023; Porian et al., 2024). Theoretical accounts derive learning curves from task dimension, spectral decay, or simplified dynamics (Sharma & Kaplan, 2020; Bordelon et al., 2020; Bahri et al., 2024), while random-feature and dynamical models separate data-, width-, and time-limited phases (Maloney et al., 2022; Bordelon et al., 2024; Paquette et al., 2024). A discrete-skill model obtains aggregate laws from decreasing skill frequencies (Michaud et al., 2023). Factual-knowledge studies infer retained bits from predictive loss and measure capacity per parameter (Allen-Zhu & Li, 2025; Morris et al., 2025); a compression model with power-law knowledge frequencies derives data and model scaling (Pan et al., 2025). Transformer optimization theory separately derives resource-dependent generalization phases (Yang, 2025). We identify the joint source statistic that determines both sample and retained-state exponents: predictive energy per independent coordinate.

Finite-rate estimation and memory. Classical rate–distortion theory characterizes the minimum coding rate at a given distortion (Shannon, 1959; Gray & Neuhoff, 1998). Quantized Gaussian sequence estimation gives arbitrary-codebook storage–risk trade-offs and adaptive or distributed extensions (Zhu & Lafferty, 2014; 2018b;a). Related work bounds working memory, sequential estimator states, retained sample information, or Bayesian minimum excess risk (Steinhardt & Duchi, 2015; Berg et al., 2021; Feldman et al., 2025; Hafez-Kolahi et al., 2021). Functional quantization is especially close: regularly varying covariance eigenvalues determine Hilbert-space quantization rates (Luschgy & Pagès, 2004; Luschgy et al., 2010). We connect this spectral machinery to positive-entropy held-out loss in a causal source, distinguish routing from value learning, and determine what separate block marginals reveal about the resulting rates.

Neural quantization and precision. Mixed-precision training retains high precision where it is numerically useful (Micikevicius et al., 2018); mean-field, Hessian-aware, and learned precision methods model sensitivity and heterogeneous allocation (Blumenfeld et al., 2019; Dong et al., 2020; Savarese et al., 2022). Per-tensor analyses and integer programs contain the familiar dynamic-range-squared times 4^{-b} distortion (Chen et al., 2020; Yao et al., 2021); kernel-regime analysis gives a width-scaling theory for 1-bit networks (Daliri et al., 2025). Practical large language model (LLM) methods use outlier handling, second-order rounding, activation rescaling, and data-dependent discrepancy control (Dettmers et al., 2022; Frantar et al., 2023; Xiao et al., 2023; Chee et al., 2025). Our finite-precision analysis connects local numerical distortion to finite-sample uncertainty and learned-state bits, then quantifies the associated arithmetic cost.

Minimax generalization and attention optimization. Worst-case neural bounds use capacity, norms, or margins (Bartlett et al., 2019; 2017); compression and PAC–Bayes analyses connect concise predictors to generalization (Arora et al., 2018; Daniely & Granot, 2019; Lotfi et al., 2022). Distribution-specific minimax theory instead ties rates to smoothness or intrinsic dimension (Schmidt-Hieber, 2020; Suzuki, 2019; Nakada & Imaizumi, 2020). For sequence models, generalization analyses address architecture and dependence (Trauger & Tewari, 2024; Yüksel & Flammarion, 2025; Li et al., 2025; Yang et al., 2026), while attention optimization studies token selection, retrieval learning, and associative memory (Ataee Tarzanagh et al., 2023; Li et al., 2024; Cabannes et al., 2024; Goel et al., 2026). Prefix learning and time-series prediction provide complementary analyses of attention training and generalization (Liang et al., 2025; Ke et al., 2025), and structured softmax-regression objectives admit efficient gradient methods (Chu et al., 2024). Sparse approximation, parallel context compression, and value-aware KV compression characterize which token state preserves an attention computation or its downstream prediction (Deng et al., 2024; Xiong et al., 2025; Yang & Li, 2026). We connect online causal routing to finite-state value learning and stored precision, identifying how each stage contributes to prediction error.

Table 2: **Closest results and the resource question they answer.** Here m is Gaussian sequence dimension, N_c is codebook size, and b is bits per Gaussian coordinate, B is total retained bits, and n is the number of training blocks in our source. CE denotes cross-entropy. The last row identifies the observation model and finite-resource conclusion needed here.

Study	Observations and parameters	Code and risk	Conclusion
Zhu & Lafferty (2014)	Gaussian sequence; deterministic means in a Euclidean ball	Arbitrary codebook, mb bits; normalized squared error	Sharp asymptotic estimation–storage trade-off as $m \to \infty$.
Luschgy & Pagès (2004)	Gaussian random element; covariance eigenvalues	N_c-point quantizer; expected squared Hilbert error	High-resolution distortion from eigenvalue decay; $B = \log_2 N_c$.
Allen-Zhu & Li (2025)	Random discrete factual tuples; trained language model weights	Finite weight state; factual generation negative log-likelihood	Achieved loss lower-bounds the information that the weights must retain.
Pan et al. (2025)	Hierarchical syntax–knowledge mixture; frequency-ranked knowledge	Bayesian/universal code; predictive redundancy under data or model capacity	Power-law knowledge frequencies generate data and model scaling laws.
Goel et al. (2026)	Gaussian linear regression; softmax attention	Expected empirical-gradient oracle; squared loss; no bit constraint	Geometric optimization with a finite-context statistical term.
This work	Random coordinate observations; independently bounded scalar logits	Any 2^B-state learner; held-out excess Bernoulli CE	Uniform (n, B) risk through one spectrum; pairing changes both exponents.

B Discussion, Scope, and Closest Precedents

One curve for two resources. The central object is the spectrum of predictive energy per independent coordinate. Its integrated tail answers both what data cannot resolve and what a finite learned state cannot retain. Data evaluate it at resolution $1/n$; storage chooses a water level. The same-marginal construction shows why the energy–dimension pairing is indispensable.

From a risk curve to a resource decision. The two terms distinguish unresolved knowledge from knowledge that cannot be retained. This distinction makes the law useful even when the finite-sample loss of a particular learner is not exactly proportional to its minimax curve. The held-out tests in Figure 2 compare response prediction and action choice separately. A flexible joint power law better fits single-axis expansion in this grid; the spectrum extrapolates better when both budgets grow and selects a useful resource without fitted constants. The exact spectrum and its rank-power approximation give similar predictions on the tested finite profiles. These observations connect the theoretical resource balance to a decision while keeping the exact curve's mathematical identifiability distinct from estimating an unknown spectrum through noisy loss measurements.

Knowledge capacity as an inverse resource curve. To express the information relation, place a fixed prior on the source parameters Z, including the route when it is unknown. Let S be the learned state formed from training blocks, and let X contain all candidates and the query of an independent held-out block. Conditional on (X, Z), the target Y is independent of the training state. Entropy and mutual information, denoted by H and I, are measured in nats. The conditional mutual-information identity and the finite-state constraint give

$$H(Y \mid X, S) - H(Y \mid X, Z) = I(Y; Z \mid X, S), \qquad I(Z; S) \le H(S) \le B \log 2.$$

The first difference compares the Bayes predictive losses given the learned state and given the source parameters; a particular decoder can incur additional loss. The prior is used for this information comparison, whereas the main minimax theorem takes a worst case over fixed parameters. This identity places knowledge-capacity and prediction-risk laws on opposite directions of the same information path. Allen-Zhu and Li (Allen-Zhu & Li, 2025) turn factual generation loss into a lower bound on the information retained by model weights—a loss-to-bits question. Our minimax law turns n observations and B retained bits into the smallest held-out excess loss—a bits-to-risk question.

Deterministic factual recall and positive-entropy prediction yield complementary geometries: the former prices discrete facts, while local KL curvature in the latter produces the coordinate distortion $q_i 2^{-2b_i}$, reverse water filling, and the common spectrum. Frequency-ranked knowledge also yields data and model scaling under a compression model (Pan et al., 2025); here the paired predictive energies determine the complete finite-(n, B) curve, and the same-marginal construction isolates the pairing information required for its exponents. Empirical capacity estimates (Allen-Zhu & Li, 2025; Morris et al., 2025) calibrate how efficiently a particular architecture stores information per parameter; the spectrum supplies the source-dependent risk curve to which such an architectural calibration can be attached. Sparse attention, prefix learning, and context compression (Liang et al., 2025; Deng et al., 2024; Xiong et al., 2025; Yang & Li, 2026) then connect the abstract state budget to the practical question of which predictive coordinates a Transformer should retain.

Learning the routing and value parameters. Public addresses determine where comparisons occur, unknown routing is learned in the query projection, and unknown values are learned in the output coordinates. Fixing the key projection preserves the route-table scores and routing-gradient trajectory. This identifies the trainable parameters through which self-attention acquires the source's predictive structure.

A hierarchy of resource laws. The arbitrary-state lower bound applies to every learner with at most 2^B final value states. The routing lower bound belongs to the stated row-symmetric online trajectory; the format lower bound applies to the required stored interface; the bit-operation lower bound concerns the specified dense graph and arithmetic model. Each result identifies the cost of a distinct bottleneck and makes the progression from statistical information to numerical computation explicit.

From learning to finite-precision prediction. The two-stage learning theorem trains routing and output coordinates using public coordinate frequencies, supervised calibration, and exact-arithmetic updates. The finite-precision theorem starts from supplied mode logits and quantifies their realization by the network; its converse concerns the separable scalar-table representation. The compute theorem studies precision allocation under a two-sided numerical response assumption. Together, these results form a three-step account of acquiring parameters, representing them numerically, and allocating arithmetic resources. A unified low-precision training theorem would complete this chain at the level of the optimization trajectory.

Precision in practical Transformer stacks. The benefit of a low-bit matrix path depends on the fraction of storage and arithmetic it changes. FP32 softmax, accumulation, master weights, optimizer states, and saved activations contribute retained high-precision work. Streaming attention reduces retained state while preserving the dense arithmetic order. Device-level speedups are governed by the corresponding throughput and memory-traffic costs.

Open directions. Three extensions would carry the spectrum principle into broader sequence models: jointly learning frequencies, ordering, queries, and keys; analyzing low-precision stochastic updates along the routing and value trajectories; and estimating predictive-energy spectra from unrestricted corpora. These directions turn the present exact source model into a program for predicting resource tradeoffs in learned representations.

C Extended Representation and Compute Results

This appendix gives the complete statements and derivations summarized in Sections 5 and 6.

C.1 Adaptive Schedule and Information Used

The schedule in Equation (5.2) knows the canonical index order and the interval $[\chi_{\min}, \chi_{\max}]$, but not χ. It adapts to the decay rate while treating energy ordering as side information. Data-driven allocation with an unknown ordering is related to adaptive quantized Sobolev estimation (Zhu & Lafferty, 2018b); learning that ordering together with the bit schedule is a separate statistical problem. The full proof and restatement are in Appendix I.

C.2 Finite-Precision Forward Error and Storage

The matching attention weight in Equation (5.4) is $\alpha_\star = e^\beta/(e^\beta + K - 1)$. Since the output coordinates lie in $[-R, R]$, $|z - \vartheta_{j_\star}| \le 2R(1 - \alpha_\star) \le 2R(K-1)e^{-\beta}$. Uniformly quantizing $[-\Delta_j, \Delta_j]$ with 2^{b_j} values gives

$$|\widetilde{w}_j^{\mathrm{mode}} - w_j^{\mathrm{mode}}| \le 2\Delta_j 2^{-b_j}. \tag{C.1}$$

Under the finite-range arithmetic conditions of Theorem 5.2, let u be the unit roundoff. Max shifting produces unnormalized weights in $\{1, e^{-\beta}\}$. Balanced pairwise summation has depth $h = \lceil \log_2 K \rceil$; put $\gamma_h = hu/(1-hu)$. Positive pairwise summation incurs relative error at most γ_h. Composing exponential, denominator, and division errors gives, when $(h+4)u \le 1/4$,

$$\|\widehat{\alpha} - \alpha\|_1 \le 8(h+4)u. \tag{C.2}$$

The final sum of K candidate-weighted scalars can have mixed signs, so an absolute bound is needed:

$$|\mathrm{fl}(\widetilde{w}^\top \widehat{\alpha}) - \widetilde{w}^\top \alpha| \le C_R u(1 + \log K). \tag{C.3}$$

The absolute bound also handles cancellation. Combining routing, quantization, and arithmetic errors, squaring and averaging over the mode, and applying bounded log-loss curvature yields Equation (5.6), including the unrepresented tail. The detailed proof and restatement of Theorem 5.2 are in Appendix K.

Learned state versus physical storage. Here $d = K + M + 1$ is the token width of the particular realization in Section 5, with M retained scalar output coordinates. If the score matrix, value projection, and output head are dense arrays, their physical scalar-slot count is

$$N = 2d^2 + d. \tag{C.4}$$

Only the M output coordinates depend on the source, and their description length is $\sum_{j=1}^{M} b_j$. Thus N, M, and B are different resources even in this fixed-depth, one-head realization with public orthogonal addresses. The finite-precision theorem initializes those coordinates with supplied logits; Theorem 4.3 separately analyzes gradient training of routing and output values.

C.3 Forward–Backward Bit Complexity

Consider eager, non-checkpointed reverse-mode differentiation (standard backpropagation) through a dense Transformer with sequence length ℓ, hidden width d, feed-forward width m_{ff}, H attention heads whose widths sum to d, and L layers. Matrix multiplications use a format indexed by b_{mm}, the stable max-shifted softmax uses compute precision b_{sm}, and a materialized attention probability uses b_{st} storage bits. The index b_{mm} determines all operand, output, and accumulator widths in this format family. Let $\mathsf{M}(b_{\mathrm{mm}})$ be the bit-operation cost of one such multiply–accumulate. Let $\mathsf{S}(b_{\mathrm{sm}})$ be the amortized per-attention-entry cost of stable softmax and its reverse-mode Jacobian–vector product, including exponentials, reductions, division, and pointwise arithmetic. We assume normalization, activation, and cast costs per entry do not asymptotically exceed these matrix and softmax costs. We do not count the optimizer update. This conventional query/key/value/output stack is a separate computational graph from the smaller realization in Section 5. Let b_w and b_a denote aggregate bits per stored weight and activation slot at peak memory, and let $\mathrm{Mem}_{\mathrm{hp}}$ denote retained high-precision state.

Lemma C.1 (Transformer forward–backward bit complexity; informal version of Lemma L.1)**.** *For the graph above, with positive dimensions and conventional dense matrix multiplication,*

$$\mathrm{BitOps}_{\mathrm{fb}} \asymp L\Big[\big(\ell d^2 + \ell d m_{\mathrm{ff}} + \ell^2 d\big)\mathsf{M}(b_{\mathrm{mm}}) + H\ell^2 \mathsf{S}(b_{\mathrm{sm}})\Big]. \tag{C.5}$$

$$\mathrm{Mem}_{\mathrm{fb}} \asymp b_w L(d^2 + d m_{\mathrm{ff}}) + b_a L\ell(d + m_{\mathrm{ff}}) + b_{\mathrm{st}} LH\ell^2 + \mathrm{Mem}_{\mathrm{hp}}\,. \tag{C.6}$$

The memory $\mathrm{Mem}_{\mathrm{fb}}$ is measured in bits. The values b_w, b_a, and b_{st} count aggregate bits per logical weight, activation, and saved-attention slot at peak memory; they include any same-shaped gradient or backward-workspace buffer materialized by this eager implementation. The term $\mathrm{Mem}_{\mathrm{hp}}$ contains master weights, optimizer state, and all other explicitly retained high-precision data. A streaming, recomputing attention implementation replaces the quadratic attention state saved for backward by linear online-softmax state plus temporary tile workspace, while leaving the dense attention arithmetic order unchanged.

The three matrix terms in Equation (C.5) come respectively from the input/output projections, the two feed-forward projections, and the score–value attention products. Reverse mode performs two matrix products for each forward matrix product, so it changes constants rather than orders. For a softmax row p and incoming vector v, its derivative is applied as $(\mathrm{diag}(p) - pp^\top)v = p \odot (v - \langle p, v\rangle \mathbf{1})$. Thus the Jacobian–vector product costs $\Theta(H\ell^2)$, without forming an $\ell \times \ell$ Jacobian for every row. The proof and a shape-by-shape operation count are in Section L.

C.4 The Compute–Precision Pareto Law

Let M denote the number of represented modes. Under a specified shape schedule $(\ell, d, m_{\mathrm{ff}}, H, L)(M)$, Lemma C.1 gives $G(M) \asymp L(\ell d^2 + \ell d m_{\mathrm{ff}} + \ell^2 d)$ and $S(M) \asymp LH\ell^2$. We write $G(M) \asymp M^g$ for quantizable matrix work and $S(M) \asymp M^s$ for softmax work, where $g, s \geq 0$. Reaching model error $M^{-\chi}$ requires $n(M) \asymp M^{1+\chi}$ independent blocks. Let $\mathcal{E}$ denote held-out excess risk, and let κ_{mm} and κ_{sm} measure how matrix and softmax rounding errors are amplified. We assume the following two-sided error and cost relations, which separate effective matrix precision b from effective softmax precision b_{sm}:

$$\mathcal{E}(M, b, b_{\mathrm{sm}}) \asymp M^{-\chi} + \kappa_{\mathrm{mm}}(M)^2 2^{-2b} + \kappa_{\mathrm{sm}}(M)^2 2^{-2b_{\mathrm{sm}}}. \tag{C.7}$$

$$C \asymp M^{1+\chi}\left[G(M)b^{\nu} + S(M)b_{\mathrm{sm}}^{\nu_{\mathrm{sm}}}\right]. \tag{C.8}$$

The exponents $\nu, \nu_{\mathrm{sm}} \geq 0$ belong to the stated arithmetic model. The abstract format family has total width proportional to its effective precision, equivalently unit roundoff $\Theta(2^{-b})$; concrete floating-point formats are discussed separately below.

Theorem C.2 (Compute–precision Pareto law; informal version of Theorem M.1)**.** *Assume Equations (C.7) and (C.8), $\kappa_{\mathrm{mm}}(M) \asymp M^r$, and $\kappa_{\mathrm{sm}}(M) \asymp M^{r_{\mathrm{sm}}}$, with all displayed relations two-sided and $r, r_{\mathrm{sm}} \geq 0$. For the specified dense graph, $H \leq d$ implies $G \gtrsim S$ and hence $g \geq s$. Item 3 concerns general operation counts satisfying the same two-sided relations; items 1, 2, and 4 apply to both settings.*

1. *The smallest matrix precision that keeps matrix rounding at the model error is*

$$b_{\mathrm{match}}(M) = \left(r + \frac{\chi}{2}\right)\log_2 M + O(1). \tag{C.9}$$

2. *If the matrix term dominates the bracket in Equation (C.8) at matched precisions—in particular in the case $g \geq s$ when the remaining precision factors do not reverse that dominance—then*

$$\inf_{M, b, b_{\mathrm{sm}}:\, \mathrm{cost} \leq C} \mathcal{E} \asymp C^{-\chi/(1+\chi+g)}(\log C)^{\chi\nu/(1+\chi+g)}, \tag{C.10}$$

   up to other declared slowly varying factors.

3. *For general operation counts, if $s > g$, b_{sm} is fixed, and the softmax term dominates computation, then for budgets in the pre-floor range*

$$C^{1/(1+\chi+s)} \lesssim 2^{2b_{\mathrm{sm}}/(\chi+2r_{\mathrm{sm}})},$$

$$\inf_{M, b:\, \mathrm{cost} \leq C} \mathcal{E} \asymp C^{-\chi/(1+\chi+s)}. \tag{C.11}$$

   Reducing matrix precision cannot improve this leading exponent.

4. *At fixed matrix precision b, with softmax precision optimized, the model–rounding crossover and the infimum error over model scales are*

$$M_b \asymp 2^{2b/(\chi+2r)}, \qquad \mathcal{E}_{\mathrm{floor}}(b) \asymp 2^{-2\chi b/(\chi+2r)}. \tag{C.12}$$

   A fixed b_{sm}-bit softmax similarly supports the scaling phase only while

$$M \lesssim 2^{2b_{\mathrm{sm}}/(\chi+2r_{\mathrm{sm}})}. \tag{C.13}$$

The displayed two-sided risk and arithmetic assumptions define the class over which the infima and matching bounds are taken.

The theorem distinguishes increasing precision from fixed-precision deployment. Polynomial conditioning requires only logarithmically increasing precision, so adaptive precision changes a logarithmic factor, not the power exponent. Conversely, permanently fixing the precision creates a numerical floor. For a concrete 32-bit floating-point (FP32) softmax with unit roundoff u_{sm}, provided the stated exponent-range and relative-error assumptions remain valid, Equation (C.13) reads $M \lesssim u_{\mathrm{sm}}^{-2/(\chi+2r_{\mathrm{sm}})}$; the total format width is not substituted for effective accuracy bits. Exponential conditioning, treated in Theorem 6.1, is the case in which precision itself creates a new power-law phase.

Limits imposed by high-precision work. At a fixed model shape, lowering a matrix path from baseline precision b_0 to b gives the relative bit-operation count

$$\mathrm{BitOpsRatio} = \frac{Gb_0^{\nu} + Sb_{\mathrm{sm}}^{\nu_{\mathrm{sm}}}}{Gb^{\nu} + Sb_{\mathrm{sm}}^{\nu_{\mathrm{sm}}}}. \tag{C.14}$$

If Q stored scalars can be quantized and R_{hp} bits must remain at high precision, the corresponding storage ratio is

$$\mathrm{Compression}_{\mathrm{bit}} = \frac{Qb_0 + R_{\mathrm{hp}}}{Qb + R_{\mathrm{hp}}}. \tag{C.15}$$

These formulas expose the ceiling imposed by high-precision normalization, master weights, optimizer states, or saved activations. Actual graphics processor speed also depends on memory traffic, kernel support, and hardware throughput.

The two analytic consequences appear in Figure 1: its left panel shows the retained-high-precision ceiling in Equation (C.14), and its right panel compares the three risk–cost regimes above.

D Proofs for the Main Results

The proofs follow the paper's mechanism order: the spectral identities and their exponents, the query–key realization and online route, the network upper bound, and the same-marginal separation.

Throughout this appendix, i indexes the scalar coordinates defined in Section 3. Their energies are $q_i = p_i R_i^2$, the spectrum is $\mu = \sum_i \delta_{q_i}$, and $\Phi_\mu(t) = \sum_i \min\{q_i, t\}$. The water level τ_B is selected by the storage budget in that setup. Each restatement below uses the same objects as its corresponding main-text result.

Theorem D.1 (One spectrum controls data and memory; formal version of Theorem 3.1)**.** *Let $q_i > 0$, $\sum_i q_i < \infty$, and let μ, Φ_μ, and τ_B be defined above. Then*

$$\mathcal{S}_q(n) := \sum_i \min\{q_i, n^{-1}\} = \Phi_\mu(n^{-1}), \tag{D.1}$$

$$D_q(B) := \inf_{\substack{b_i \geq 0 \\ \sum_i b_i \leq B}} \sum_i q_i 2^{-2b_i} = \Phi_\mu(\tau_B). \tag{D.2}$$

For the bounded-logit value experiment,

$$\boxed{\mathfrak{R}^*_{\mathrm{value}}(n, B) \asymp_R \Phi_\mu(n^{-1}) + \Phi_\mu(\tau_B).} \tag{D.3}$$

Moreover, for every $t > 0$,

$$\Phi'_{\mu,+}(t) = \mu((t, \infty)), \qquad \Phi'_{\mu,-}(t) = \mu([t, \infty)). \tag{D.4}$$

Hence the exact curve Φ_μ recovers the positive spectrum, including atom multiplicities, up to coordinate relabeling.

Proof. Summability gives $\mathcal{N}_\mu(t) < \infty$ for every $t > 0$, and the largest positive energy is attained. Tonelli's theorem yields

$$\Phi_\mu(t) = \sum_i \int_0^t \mathbf{1}\{s < q_i\}\, \mathrm{d}s = \sum_i \min\{q_i, t\},$$

which proves Equation (3.1). For a finite storage budget, reverse water filling has rates

$$b_i^* = \frac{1}{2}\log_2(q_i/\tau_B)_+.$$

Here the positive part applies to the logarithm. To justify the optimizer without interchanging a limit and an infimum, write

$$B(\tau) = \frac{1}{2\ln 2}\sum_i [\ln(q_i/\tau)]_+.$$

On every compact subinterval of $(0, q_{\max}]$, only finitely many summands can be nonzero. Thus $B(\tau)$ is continuous, strictly decreasing below $q_{\max}$, equals zero at $q_{\max}$, and tends to infinity as $\tau \downarrow 0$. This proves existence and uniqueness of τ_B for $B > 0$; for $B = 0$ take $\tau_0 = q_{\max}$. For $\lambda = 2(\ln 2)\tau_B$, scalar convexity gives, for every $b_i \geq 0$,

$$q_i 2^{-2b_i} + \lambda b_i \geq q_i 2^{-2b_i^*} + \lambda b_i^*.$$

All distortion sums are bounded by $\sum_i q_i$, and the two bit sums are finite. Summing the inequality over all coordinates shows that every feasible allocation has distortion at least $\sum_i q_i 2^{-2b_i^*} + \lambda(B - \sum_i b_i) \geq \sum_i q_i 2^{-2b_i^*}$. This proves optimality directly in the countable problem. Since $q_i 2^{-2b_i^*} = \min\{q_i, \tau_B\}$, this proves Equation (3.2). The bounded-logit curvature, Assouad, and arbitrary codebook bounds, proved in Theorem H.2, give $\mathfrak{R}^*_{\text{value}}(n, B) \asymp_R \mathcal{S}_q(n) + D_q(B)$, proving Equation (3.3). Finally, for $0 < h < t/2$, both one-sided difference quotients of $\min\{q_i, t\}$ lie in $[0, 1]$ and vanish whenever $q_i < t/2$. There are only finitely many remaining coordinates. Taking their limits gives respectively $\mathbf{1}\{q_i > t\}$ and $\mathbf{1}\{q_i \geq t\}$, proving Equation (3.4). Their difference is exactly the multiplicity of the atom at t. □

Corollary D.2 (Spectral scaling exponents; formal version of Corollary 3.2)**.** *If $\mathcal{N}_\mu(t) \sim t^{-\rho}L(1/t)$ for $0 < \rho < 1$, where L is slowly varying, then*

$$\Phi_\mu(n^{-1}) \in \mathrm{RV}_{-(1-\rho)}, \qquad \Phi_\mu(\tau_B) \in \mathrm{RV}_{-(1-\rho)/\rho}.$$

For $a_j \asymp j^{-(1+\chi)}$ and $d_j \asymp j^\gamma$, $\rho = (1+\gamma)/(1+\chi+\gamma)$, giving rates $n^{-\chi/(1+\chi+\gamma)}$ and $B^{-\chi/(1+\gamma)}$, up to slowly varying factors.

Proof. We first prove the regular-variation assertion under its stated hypothesis, then treat power comparisons separately. Abbreviate $N(t) = \mathcal{N}_\mu(t)$. For fixed $x > 0$, $N(tx)/N(t) \to x^{-\rho}$ as $t \downarrow 0$. Choose $0 < \epsilon < \min\{\rho, 1-\rho\}$. The ratio at $x = 1/2$, iterated over dyadic intervals and combined with monotonicity, gives

$$\frac{N(tx)}{N(t)} \leq C_\epsilon x^{-\rho-\epsilon} \quad (0 < x \leq 1), \qquad \frac{N(tx)}{N(t)} \leq C_\epsilon x^{-\rho+\epsilon} \quad (1 \leq x \leq t_0/t)$$

for a fixed sufficiently small t_0. For example, choose t_0 so that the ratio $N(s/2)/N(s)$ lies between $2^{\rho-\epsilon}$ and $2^{\rho+\epsilon}$ for all $s \leq t_0$, and telescope; rounding x to its neighboring powers of two changes only C_ϵ. Consequently dominated convergence in $\Phi_\mu(t)/(tN(t)) = \int_0^1 N(tx)/N(t)\,\mathrm{d}x$ gives

$$\Phi_\mu(t) \sim \frac{t^{1-\rho}L(1/t)}{1-\rho}.$$

Layer integration of the storage budget gives

$$\mathcal{B}_\mu(\tau) = \frac{1}{2\ln 2}\int_\tau^{q_{\max}} \frac{\mathcal{N}_\mu(t)}{t}\,\mathrm{d}t \sim \frac{\tau^{-\rho}L(1/\tau)}{2\rho\ln 2}.$$

Indeed, after substituting $t = \tau x$, the second domination bound makes the integral converge to $\int_1^\infty x^{-\rho-1}\,\mathrm{d}x = 1/\rho$. The part above t_0 is finite and negligible relative to $N(\tau)$. Thus $B(c\tau)/B(\tau) \to c^{-\rho}$. For $a > 0$, sandwiching the inverse between $(a^{-1/\rho} - \varepsilon)\tau_B$ and $(a^{-1/\rho} + \varepsilon)\tau_B$, and then sending $\varepsilon \downarrow 0$, proves $\tau_{aB}/\tau_B \to a^{-1/\rho}$. The same monotone sandwich for Φ_μ gives $\Phi_\mu(\tau_{aB})/\Phi_\mu(\tau_B) \to a^{-(1-\rho)/\rho}$. This proves both regular-variation indices without assuming a differentiable slowly varying factor.

For the block profile, dimensions are positive integers, so the stipulated comparison $d_j \asymp j^\gamma$ has $\gamma \geq 0$. Coordinate energy $a_j/d_j \asymp j^{-(1+\chi+\gamma)}$ occurs $d_j \asymp j^\gamma$ times. Bounds on the two comparison constants show that every index below $ct^{-1/(1+\chi+\gamma)}$ is active and no index above $Ct^{-1/(1+\chi+\gamma)}$ is active. Summing the multiplicities therefore gives $N(t) \asymp t^{-\rho}$ with $\rho = (1+\gamma)/(1+\chi+\gamma)$. Integrating these two-sided bounds directly yields

$$\Phi_\mu(t) \asymp t^{1-\rho}, \qquad B(\tau) \asymp \tau^{-\rho}, \qquad \Phi_\mu(\tau_B) \asymp B^{-(1-\rho)/\rho}.$$

These power orders do not require regular variation of the individual profile or of an atomic counting tail. □

Proposition D.3 (Exact query–key realization; formal version of Proposition 4.1)**.** *Under the strict causal mask exposing exactly the K preceding candidates, Equation (4.1) reproduces the route-table scores and attention probabilities exactly. If only the query-address block of W_Q is trained and W_K is fixed, Euclidean SGD on W_Q induces the same update of A; the attended feature and output-vector gradient are unchanged.*

Proof. Because $P_Q x_q = e_q$, $P_C x_s = e_s$, and $J^\top J = I_K$, direct substitution gives Equation (4.1). If $G = \nabla_A \ell$, the physical query-projection gradient is $\nabla_{W_Q} \ell = JG^\top P_Q$. Restricting training to this address block therefore maps one Euclidean step to $W_Q(A - \eta G)$. The scores, candidate-only masked softmax, and attended value feature coincide. The output loss sees the same feature, so its gradient with respect to w also coincides. Orthogonality of the residual subspace to w makes the residual logit zero. □

Proposition D.4 (Online routing dynamics; formal version of Proposition 4.2)**.** *For $K \geq 2$, zero initialization and softmax cross-entropy steps of size $0 < \eta \leq 1$ give the margin and distractor mass after t row visits:*

$$\beta_{t+1} = \beta_t + \eta \frac{K}{e^{\beta_t} + K - 1}, \qquad \delta_t = \frac{K-1}{e^{\beta_t} + K - 1}.$$

After $b_K = \Theta_\eta(\log(eK))$ visits, $\delta_t \asymp_\eta [1 + (t - b_K)_+]^{-1}$. For S_A independent calibration blocks and a held-out row,

$$\mathfrak{r}_K(S_A) = \mathbb{E}[\delta_N^2], \quad N \sim \mathrm{Binomial}(S_A, 1/K),$$

and the hitting time to $\mathfrak{r}_K(S_A) \leq \varepsilon$, $0 < \varepsilon \leq 1/16$, is $\Theta_\eta(K\{\varepsilon^{-1/2} + \log(eK)\})$. For $K = 1$, leakage and calibration cost are zero.

Proof. Zero initialization preserves equality among distractor scores. If u_t is the correct score and v_t a distractor score, one softmax cross-entropy update gives

$$u_{t+1} = u_t + \eta(1 - p_\star), \qquad v_{t+1} = v_t - \eta p_0,$$

which proves the margin recurrence. Introduce

$$\Psi_K(\beta) = \frac{e^\beta - 1 + (K-1)\beta}{K}.$$

Put $d_t = \beta_{t+1} - \beta_t$. Since $\beta_t \geq 0$, $0 < d_t \leq \eta$, and $\Psi'_K(\beta_t) d_t = \eta$. Moreover $1 \leq \Psi'_K(\beta_t + v)/\Psi'_K(\beta_t) \leq e^\eta$ for $0 \leq v \leq d_t$. Integrating and summing gives $\eta t \leq \Psi_K(\beta_t) \leq e^\eta \eta t$. Define the positive burn-in time

$$b_K = \min\{t : \beta_t \geq \ln(2K)\}.$$

The preceding bounds, and the overshoot $d_t \leq \eta$, imply $b_K \asymp_\eta \ln(eK)$, including $K = 2$. Set $x_t = e^{\beta_t}/(K-1)$. Its increment satisfies

$$x_t d_t \leq x_{t+1} - x_t \leq e^\eta x_t d_t, \qquad x_t d_t = \frac{\eta K}{K-1} \frac{x_t}{1 + x_t}.$$

At burn-in $2 \leq x_{b_K} \leq 4e^\eta$, and thereafter each increment is between two positive constants depending only on η. Hence $\delta_{b_K + k} = 1/(1 + x_{b_K + k}) \asymp_\eta (1+k)^{-1}$. Before burn-in, $x_t < 4e^\eta$, so $\delta_t \asymp_\eta 1$. This proves the shifted reciprocal bound at every integer time.

For a held-out row, independence and uniform query addresses give $N \sim \mathrm{Binomial}(S_A, 1/K)$; write $m = S_A/K$. If $m \geq 4b_K$, the binomial exponential-moment bound gives $\Pr\{N < m/2\} \leq e^{-m/8}$. On the complementary event, $\delta_N^2 \leq C_\eta(1+m)^{-2}$. Thus

$$\mathbb{E}\delta_N^2 \leq C_\eta(1+m)^{-2} + e^{-m/8},$$

which is at most ε when $m \geq C_\eta\{\ln(eK) + \varepsilon^{-1/2}\}$. For the reverse bound, the increment estimate holds globally from $x_0 \leq 1$, giving $x_t \leq 1 + 2\eta e^\eta t$. Convexity of $(2 + 2\eta e^\eta t)^{-2}$ and Jensen's inequality imply $\mathbb{E}\delta_N^2 \geq (2 + 2\eta e^\eta m)^{-2}$. Since $\varepsilon \leq 1/16$, this forces $m \geq c_\eta \varepsilon^{-1/2}$. For the logarithmic term, $\beta_t \leq 2\eta t$. If $m \leq \ln(K-1)/(8\eta)$, Markov's inequality gives probability at least $1/2$ to $N \leq 2m$, where $\beta_N \leq \frac{1}{2}\ln(K-1)$ and $\delta_N \geq 1/2$. Then $\mathbb{E}\delta_N^2 \geq 1/8$, a contradiction. This forces $m \gtrsim_\eta \ln K$; bounded K, including $K = 2$, is already covered by the first lower bound. Taking the maximum of the two lower bounds proves the claimed order of the hitting time. With a single candidate routing is exact without calibration. □

In the following restatement, $\Delta_i = R_i$ denotes the same public logit radius as in the main-text setup.

Theorem D.5 (Causal self-attention realization; formal version of Theorem 4.3)**.** *For the source in Section 3, let Stage A start from zero and use the routing update in Proposition 4.2 on S_A independent calibration blocks. Freeze the route before Stage V, which uses $n \geq 1$ independent labeled blocks. With public p_i, Δ_i, set*

$$A_n = \{i : q_i \geq n^{-1}\}, \quad \mathcal{W} = \prod_{i \in A_n} [-\Delta_i, \Delta_i], \quad D = \mathrm{diag}(p_i)_{i \in A_n}, \quad w_1 = 0.$$

Apply Equation (4.2) to the attended feature restricted to A_n, with zero output coordinates elsewhere, and use

$$\mu_{\mathrm{sgd}} = \frac{9}{16} \min_{|z| \leq R} \sigma'(z), \qquad \eta_t = \frac{2}{\mu_{\mathrm{sgd}}(t+1)}, \qquad \bar{w}_n = \frac{2}{n(n+1)} \sum_{t=1}^{n} t\, w_t.$$

If A_n is empty, take $\bar{w}_n = 0$. For each public format-range ratio $\kappa_i \geq 1$ and integer width $k \geq 0$, the scalar decoder is defined as follows. If $\kappa_i 2^{-k} \geq 1$, its output is zero. Otherwise quantize uniformly with 2^k midpoint values on $[-\kappa_i\Delta_i, \kappa_i\Delta_i]$, then clip the decoded value to $[-\Delta_i, \Delta_i]$. The allocation is public and deterministic; quantize $\bar{w}_n$ with total width at most B_V, choosing an allocation within an arbitrarily small error of the infimum below and flooring its widths. Let $\mathcal{E}_{\mathrm{arithmetic}}$ bound the mean squared difference between the computed and exact logits of this stored network. Then

$$\mathbb{E}[\mathcal{R}_\theta(\widehat{F}) - \mathcal{L}_\infty] \lesssim_{R,K,\eta} \Phi_\mu(n^{-1}) + D^{\mathrm{fmt}}_{q,\kappa}(B_V) + \mathfrak{r}_K(S_A) + \mathcal{E}_{\mathrm{arithmetic}}, \tag{D.5}$$

where B_V is the value-state bit budget and

$$D^{\mathrm{fmt}}_{q,\kappa}(B_V) = \inf_{b_i \geq 0,\, \sum_i b_i \leq B_V} \sum_i q_i \min\{1, \kappa_i^2 2^{-2b_i}\}.$$

For $\kappa_i = 1$, this equals $\Phi_\mu(\tau_{B_V})$, and the network attains the core surface at total budget B for power-law spectra, fixed K, $o(B)$ routing bits, and lower-order routing/arithmetic errors.

Proof. The finite active set has size $m = |A_n| \leq n \sum_i q_i$. We first analyze the exact output before storage, conditioning on Stage A.

Feature covariance and curvature. For $K \geq 2$, write I for the matching candidate's mode and $I_1, \ldots, I_{K-1}$ for the distractor modes. They are independent with law p, also conditional on the query row and its calibration history. Restricting the coordinate vectors to A_n, the feature is

$$h = (1 - \delta_q)e_I + \frac{\delta_q}{K-1} \sum_{s=1}^{K-1} e_{I_s}.$$

Expanding its second moment separates equal-candidate and distinct-candidate products:

$$\mathbb{E}[hh^\top \mid Q = q] = c_{d,q} D + (1 - c_{d,q}) p_A p_A^\top, \qquad c_{d,q} = (1-\delta_q)^2 + \frac{\delta_q^2}{K-1},$$

where $p_A = (p_i)_{i \in A_n}$. With $r_{\mathrm{cal}} = K^{-1} \sum_q \delta_q^2$ and $\bar{c}_d = K^{-1} \sum_q c_{d,q}$, the event $\mathcal{G} = \{r_{\mathrm{cal}} \le 1/16\}$ gives

$$\bar{c}_d \ge \left(1 - \frac{1}{K} \sum_q \delta_q\right)^2 \ge \frac{9}{16}.$$

Let $F(w) = \mathbb{E}[\log(1 + e^{w^\top h}) - Y w^\top h \mid \text{Stage A}]$. Every $w \in \mathcal{W}$ has $|w^\top h| \le R$. Hence, on $\mathcal{G}$, its Hessian satisfies $\nabla^2 F(w) \succeq \mu_{\mathrm{sgd}} D$, with μ_{sgd} as in the statement. The stochastic gradient $g_t = (\sigma(w_t^\top h_t) - Y_t) h_t$ is unbiased conditional on the past, and, for $m \ge 1$,

$$\mathbb{E}[\|g_t\|_{D^{-1}}^2 \mid w_t, \text{Stage A}] \le \mathrm{tr}(D^{-1} \mathbb{E} h h^\top) = \bar{c}_d m + (1 - \bar{c}_d) \sum_{i \in A_n} p_i \le m.$$

For $K = 1$, use $h = e_I$, $\bar{c}_d = 1$, and $r_{\mathrm{cal}} = 0$; all the same inequalities hold.

Projected SGD and its output. The compact box has a minimizer w^* of F. Nonexpansiveness of projection in the D-norm and strong convexity give, conditional on $\mathcal{G}$ and the calibration history,

$$2\eta_t \mathbb{E}[F(w_t) - F(w^*)] \le (1 - \mu_{\mathrm{sgd}} \eta_t) V_t - V_{t+1} + \eta_t^2 m, \qquad V_t = \mathbb{E}\|w_t - w^*\|_D^2.$$

Multiply by $t/(2\eta_t)$ and substitute the specified step size:

$$t\mathbb{E}[F(w_t) - F(w^*)] \le \frac{\mu_{\mathrm{sgd}}}{4} \{t(t-1)V_t - t(t+1)V_{t+1}\} + \frac{mt}{\mu_{\mathrm{sgd}}(t+1)}.$$

The distance terms telescope, with zero initial coefficient and a nonpositive final term. Convexity of F therefore yields

$$\mathbb{E}[F(\bar{w}_n) - F(w^*)] \le \frac{2m}{\mu_{\mathrm{sgd}}(n+1)}.$$

Comparison with the source. Let $\theta_i^A = \theta_i \mathbf{1}\{i \in A_n\}$. It belongs to $\mathcal{W}$, and

$$|\theta_I - (\theta^A)^\top h| \le |\theta_I| \mathbf{1}\{I \notin A_n\} + 2R\delta_q.$$

Squaring, averaging, and applying upper logit curvature give

$$F(w^*) - \mathcal{L}_\infty \le F(\theta^A) - \mathcal{L}_\infty \le C_R \left(\sum_{i \notin A_n} q_i + r_{\mathrm{cal}}\right).$$

On $\mathcal{G}^c$, every exact output remains in $[-R, R]$, so its excess loss is bounded by a constant depending only on R. Furthermore $\Pr(\mathcal{G}^c) \le 16 \mathbb{E} r_{\mathrm{cal}}$, and independence of a held-out query gives $\mathbb{E} r_{\mathrm{cal}} = \mathfrak{r}_K(S_A)$. Combining the events proves

$$\mathbb{E}[F(\bar{w}_n) - \mathcal{L}_\infty] \le C_R \left\{\frac{m}{n} + \sum_{i \notin A_n} q_i + \mathfrak{r}_K(S_A)\right\} = C_R\{\Phi_\mu(n^{-1}) + \mathfrak{r}_K(S_A)\}.$$

When $m = 0$, the zero output pays at most $C_R \sum_i q_i = C_R \Phi_\mu(n^{-1})$, so the same bound holds.

Finite-state representation. For any $x \in [-\Delta_i, \Delta_i]$, the stated k-bit decoder obeys

$$|Q_{i,k}(x) - x|^2 \le \Delta_i^2 \min\{1, \kappa_i^2 2^{-2k}\}.$$

Indeed, the zero branch has error at most Δ_i; in the grid branch the midpoint error is at most $\kappa_i \Delta_i 2^{-k}$, and clipping decreases it. Both the branch and the codebook are public. Rounding b_i down to k_i multiplies this upper bound by at most four. Only finitely many k_i are positive, and their product codebook has at most $2^{\sum_i k_i} \le 2^{B_V}$ entries. A zero-width coordinate is decoded as zero because $\kappa_i \ge 1$.

Extend $\bar{w}_n$ by zero outside A_n, and apply the public scalar decoder on every allocated coordinate. Let $e_i = Q_{i,k_i}(\bar{w}_{n,i}) - \bar{w}_{n,i}$ and let $H = \sum_s a_{Qs} e_{C_s}$ be the full attended coordinate feature. Conditional on the fitted state, Jensen's inequality and the iid candidate marginals give

$$\mathbb{E}[(e^\top H)^2 \mid \text{fitted state}] \le \mathbb{E}\left[\sum_s a_{Qs} e_{C_s}^2 \mid \text{fitted state}\right] = \sum_i p_i e_i^2.$$

It follows that quantizing the average costs at most $4\sum_i q_i \min\{1, \kappa_i^2 2^{-2b_i}\}$ in mean squared logit error. Lower curvature converts the preceding learning bound to squared logit error; the squared triangle inequality adds storage and arithmetic error, and global upper curvature converts back to cross-entropy. Taking the infimum over allocations proves Equation (4.3).

Finally, a calibrated row is determined by its visit count and its matching position, with one additional symbol for an unvisited row. It therefore uses at most

$$B_A = K\{\lceil \log_2(S_A + 1)\rceil + \lceil \log_2(K + 1)\rceil\}$$

bits: the fixed decoder reconstructs its scores from the recurrence. The retained value and routing state together use at most $B_V + B_A$ bits; the stated arithmetic error accounts for numerical reconstruction. For $K = 1$ the route uses no state. When $\kappa_i = 1$, the format functional is exactly $D_q(B_V) = \Phi_\mu(\tau_{B_V})$, giving the final claim at value budget B_V. For ranked coordinate energies $q_{(r)} \asymp r^{-(1+\xi)}$, $D_q(B_V) \asymp B_V^{-\xi}$; in the equal-energy block profile, $\xi = \chi/(1+\gamma)$. If $B_A = o(B)$ and $B_V = B - B_A$, then $B_V \asymp B$, proving the total-budget conclusion. □

Theorem D.6 (Same block marginals, different scaling laws; formal version of Theorem 3.3)**.** *For every $\chi, \gamma > 0$, there are two positive-conditional-entropy causal-retrieval sources with exactly the same block-energy multiset and the same block-dimension multiset at every dyadic level. They differ only in the pairing. All displayed dimensions in this construction are rounded upward to integers. With the common fixed-K route revealed,*

$$\mathfrak{R}_{\text{easy}}(n, B) \asymp n^{-\chi/(1+\chi)} + B^{-\chi},$$

whereas

$$\mathfrak{R}_{\text{hard}}(n, B) \asymp n^{-\chi/(1+\chi+\gamma)} + B^{-\chi/(1+\gamma)}.$$

Proof. Fix $\delta > \gamma\chi$. At each level $\ell \geq 1$, take 2^ℓ blocks of energy $c2^{-(1+\chi)\ell}$ and 2^ℓ blocks of energy $c2^{-(1+\chi+\delta)\ell}$. The dimension multiset consists of 2^ℓ copies each of 1 and $d_\ell = \lceil 2^{\gamma\ell}\rceil$. In the easy pairing the higher energy has dimension one; in the hard pairing it has dimension d_ℓ. Divide a block's energy equally among its coordinates. Both multisets agree exactly at every level, including after the integer rounding of dimensions.

To realize these spectra as sources, choose any fixed radius $R_0 > 0$ within the logit bound, and choose $c > 0$ so that total energy is R_0^2. Give a coordinate of energy q_i probability $p_i = q_i/R_0^2$ and radius $\Delta_i = R_0$. These probabilities sum to one. Independent candidate draws and the same fixed-K retrieval map give the source in Section 3; bounded Bernoulli logits give strictly positive conditional entropy.

A generic component with exponent $s > 0$ and dimension exponent $g \geq 0$ has $\Theta(2^{(1+g)\ell})$ coordinates of energy $\Theta(2^{-(1+s+g)\ell})$ at level ℓ. The active levels at threshold t lie between two cutoffs differing by a constant number of levels, each of size $\log_2(1/t)/(1+s+g)$. A geometric sum therefore gives

$$N_{s,g}(t) \asymp t^{-\rho(s,g)}, \qquad \rho(s, g) = \frac{1+g}{1+s+g}.$$

Use the two-sided integral comparison at the end of the preceding spectral proof, rather than regular variation: dyadic spectra can have log-periodic fluctuations. This gives data and storage exponents $s/(1+s+g)$ and $s/(1+g)$, respectively. The easy components have $(s, g) = (\chi, 0), (\chi+\delta, \gamma)$. The inequality $\delta > \gamma\chi$ makes the second component's two exponents strictly larger. The hard components have $(s, g) = (\chi, \gamma), (\chi+\delta, 0)$, again with the second faster. For data, component errors add exactly. For storage, each full distortion is at least that of the slower component at budget B, while allocating $B/2$ to each component gives the matching upper bound. Applying Theorem D.1 now proves all four displayed rates. □

E Preliminaries and Notation

We begin by specifying what is observed, what is predicted, and which resources are counted. Section E.1 introduces the regenerative autoregressive source. Section E.2 defines held-out loss and learners with finitely many data-dependent states. Section E.3 adds mode-dependent representation costs, and Section E.4 distinguishes learned bits from physical storage and arithmetic cost.

E.1 A REGENERATIVE AUTOREGRESSIVE SOURCE

This section writes the revealed-coordinate value problem with scalar index j: Δ_j is the logit radius denoted R_i in the main text, and $\mathfrak{E}^*(n, B)$ below is $\mathfrak{R}^*_{\mathrm{value}}(n, B)$. For a positive integer m, write $[m] = \{1, \dots, m\}$, and let $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. We write $\mathrm{Ber}(p)$ for the Bernoulli distribution with success probability p.

One regeneration block contains K candidate tokens followed by a query and a binary target token. Candidate $t \in [K]$ carries a public address a_t and an observed mode label j_t. The query carries one address $a_{t_\star}$, so its causally relevant mode is

$$J \coloneqq j_{t_\star}, \qquad \mathbb{P}(J = j) = p_j.$$

Conditionally on $J = j$, the next token obeys

$$Y \mid J = j \sim \mathrm{Ber}(\sigma(\theta_j)), \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

The parameter is $\theta = (\theta_j)_{j \geq 1}$ with $\theta_j \in [-\Delta_j, \Delta_j] \subseteq [-R, R]$, where $R < \infty$ is fixed. We define the predictive energy

$$q_j \coloneqq p_j \Delta_j^2$$

and assume $\sum_j q_j < \infty$. We write $Z = (\{(a_t, j_t)\}_{t=1}^K, a_{t_\star}, Y) \in \mathcal{Z}$ for the observed training block; the source parameters $(\theta_j)_j$ are unknown. A regeneration delimiter resets the random variables, making distinct blocks independent. Tokens inside a block are causally ordered, but the statistical sample size n counts independent blocks, not tokens. Bounded logits imply $0 < \sigma(-R) \leq \mathbb{P}(Y = 1 \mid J) \leq \sigma(R) < 1$, so the conditional entropy and Bayes cross-entropy are strictly positive.

E.2 HELD-OUT RISK AND FINITE LEARNED STATES

For a logit predictor $f = (f_j)_j$, define held-out cross-entropy

$$\mathcal{L}_\theta(f) \coloneqq \mathbb{E}_\theta[-Y \log \sigma(f_J) - (1 - Y) \log(1 - \sigma(f_J))] .$$

The Bayes predictor is $f_j = \theta_j$; its loss is denoted $\mathcal{L}_\infty \coloneqq \mathcal{L}_\theta(\theta) > 0$. Training and evaluation regeneration blocks are independent.

For an integer $B \geq 0$, a B-bit learner consists of an encoder

$$\phi : \mathcal{Z}^n \longrightarrow [2^B]$$

and a fixed decoder $g : [2^B] \times \mathbb{N} \to [-R, R]$. The decoder may share arbitrary computation across modes; only its data-dependent state is limited to 2^B possibilities. Randomized learners are allowed, although conditioning on their random seed shows that they do not improve the Bayes lower bounds used below. Because the Bayes loss depends on θ, we define minimax *excess* risk directly and use the notation

$$\mathfrak{E}^*(n, B) \coloneqq \inf_{\phi, g} \sup_{\theta_j \in [-\Delta_j, \Delta_j]} \mathbb{E}_\theta[\mathcal{L}_\theta(g(\phi(Z^n), \cdot)) - \mathcal{L}_\theta(\theta)] .$$

We use $x \asymp y$ when positive constants bound their ratio in both directions. These constants may depend on R and on stated regular-variation constants, but not on n, B, or the number of active modes.

E.3 HETEROGENEOUS REPRESENTATION COSTS

The statistical theorem uses q_j. The structural theorem allows a more general representable energy $a_j > 0$, a per-bit difficulty $c_j > 0$, and a numerical amplification $\kappa_j > 0$. Assigning $b_j \geq 0$ bits to mode j incurs distortion $a_j 2^{-2b_j}$ and cost $c_j b_j$. Its optimal distortion under budget B is

$$D(B) \coloneqq \inf_{\substack{b_j \geq 0 \\ \sum_j c_j b_j \leq B}} \sum_j a_j 2^{-2b_j}. \tag{E.1}$$

Let $v_j \geq 0$ record a separately specified statistical residual for mode j. This auxiliary quantity is used in the complete tuple summary below, not in the spectral formula for $D(B)$. For a finite mode collection, the unlabeled joint measure

$$\eta \coloneqq \sum_j \delta_{(v_j, a_j, c_j, \kappa_j)} \tag{E.2}$$

records statistical residual v_j, representable energy, difficulty, and amplification together. The cost-weighted difficulty-normalized spectrum is

$$\mu \coloneqq \sum_j c_j \, \delta_{a_j/c_j}. \tag{E.3}$$

For unit costs and $a_j = q_j$, this reduces to the coordinate spectrum in Section 3; general costs weight each atom by c_j. We will compare unlabeled summaries of these mode-dependent quantities. Such a summary is sufficient for the complete budget–distortion curve if equal summary values imply equal $D(B)$ for every $B \geq 0$.

E.4 Neural and arithmetic resources

In Section 5, $X \in \mathbb{R}^{(K+1)\times d}$ denotes the matrix of candidate and query token vectors, and $U \in \mathbb{R}^{d\times d}$ denotes the bilinear attention score matrix. The symbol N counts physically stored dense scalar parameter slots. It is not identified with B, the number of bits in the data-dependent learned state.

The symbol C denotes a non-statistical cost budget: representation cost in Section G and arithmetic computation in Sections C.3 and 6. These meanings occur in separate results and are never equated. When C denotes computation, its arithmetic cost model is stated explicitly. In Section 6, one b-bit scalar operation costs $\Theta(b^\nu)$, where $\nu \geq 0$ is fixed. In Section C.3, ℓ, d, m_{ff}, H, and L denote sequence length, hidden width, feed-forward width, attention-head count, and layer count. The symbol b_{mm} indexes a matrix-arithmetic format family whose operand, output, and accumulator widths are prescribed functions of b_{mm}; b_{sm} specifies the effective softmax precision; and b_{st} is the number of bits used to store an attention probability. Unit roundoff is denoted u. These variables specify arithmetic work and storage; Section B discusses their relation to implementations.

F Learning after Data: A Finite-Bit Autoregressive Minimax Law

Two irreducible losses govern finite-bit prediction. We first lower-bound the distortion of any learned state with at most 2^B possible values, even when its coordinates are decoded jointly. We then combine this bound with statistical uncertainty across regeneration blocks and evaluate the resulting risk law for power-law energies. Full proofs appear in Section H.

Lemma F.1 (Arbitrary joint finite-state representations; informal version of Lemma H.1)**.** *Let Θ_j be independent and uniform on $[-\Delta_j, \Delta_j]$, let $p_j > 0$, and put $q_j = p_j\Delta_j^2$. For $B \in \mathbb{N}_0$, every encoder–decoder pair with at most 2^B reconstruction vectors satisfies*

$$\sum_j p_j \mathbb{E}(\Theta_j - \widehat{\Theta}_j)^2 \geq \frac{2}{\pi e} \inf_{\substack{b_j \geq 0 \\ \sum_j b_j \leq B}} \sum_j q_j 2^{-2b_j}. \tag{F.1}$$

The reconstruction vectors may couple all coordinates.

Here the *codebook* is simply the set of at most 2^B reconstruction vectors available to the decoder. The entropy proof controls total codebook cardinality, so it covers shared representations that compress several modes jointly as well as coordinatewise quantizers.

Theorem F.2 (Finite-bit autoregressive minimax law; informal version of Theorem H.2)**.** *Fix $R < \infty$. For every $n \in \mathbb{N}$, $B \in \mathbb{N}_0$, and source family specified by $(p_j, \Delta_j)_{j\geq 1}$ as in Section E.1, there are constants $0 < c_R \leq C_R < \infty$ such that*

$$c_R\{S_q(n) + D_q(B)\} \leq \mathfrak{E}^*(n, B) \leq C_R\{S_q(n) + D_q(B)\}, \tag{F.2}$$

where

$$S_q(n) \coloneqq \sum_j \min\{q_j, n^{-1}\}, \qquad D_q(B) \coloneqq \inf_{\substack{b_j \geq 0 \\ \sum_j b_j \leq B}} \sum_j q_j 2^{-2b_j}. \tag{F.3}$$

The lower bound holds for arbitrary 2^B-state joint encoders and arbitrary fixed shared decoding computation.

Why the two terms add. On bounded logits, excess Bernoulli cross-entropy is equivalent to squared logit error. A clipped modewise likelihood estimate contributes $\min\{q_j, 1/n\}$: a rare or weak mode is no harder to ignore than to estimate. Quantizing the resulting coordinates according to a feasible (b_j) contributes $q_j 2^{-2b_j}$. For the converse, an Assouad hypercube gives the statistical term. Giving the encoder the true parameter can only make representation easier; Lemma F.1 then gives the bit term. The actual risk exceeds the maximum of these two lower bounds, which is at least half their sum. This proves the constant-factor equivalence in Equation (F.2).

Corollary F.3 (Power-law energies; informal version of Corollary H.3)**.** *Suppose that $n, B \geq 1$ and, for some $\chi > 0$ and constants $0 < c_- \leq c_+ < \infty$,*

$$c_- j^{-(1+\chi)} \leq q_j \leq c_+ j^{-(1+\chi)} \qquad (j \geq 1).$$

Then

$$\mathfrak{E}^*(n, B) \asymp n^{-\chi/(1+\chi)} + B^{-\chi}. \tag{F.4}$$

The data–bit crossover occurs at $B \asymp n^{1/(1+\chi)}$.

To see the two exponents, split $S_q(n)$ where q_j crosses $1/n$. For $D_q(B)$, reverse water filling equalizes $q_j 2^{-2b_j}$ on active modes. If m modes are active, Stirling's formula gives $B \asymp m$, while active distortion and the inactive tail are both $\asymp m^{-\chi}$.

What the law identifies. The theorem identifies the data and retained-bit dependence for the stated positive-entropy autoregressive source. Its distinctive role is to couple held-out log-loss with arbitrary shared decoding; the next structural results show what additional information determines the resulting exponent.

G Why Marginal Summaries Do Not Determine Scaling

This section asks which information about mode difficulty identifies a scaling law. Here cost is charged directly to selecting or representing a coordinate. The main-text construction instead realizes difficulty through the number of independent coordinates within an energy group. Two problems with identical energy and cost marginals first show what separate summaries lose. The joint spectrum then recovers exactly the complete budget–distortion curve and gives the smallest exact summary for separable allocation.

G.1 Same marginals, different exponents

Consider first a binary selection rule. Selecting mode j pays c_j and removes energy q_j; unselected energy remains as error. Under total representation-cost budget C, its optimum is

$$\mathcal{A}(C) := \inf_{\substack{S \subseteq \mathbb{N} \\ \sum_{j \in S} c_j \leq C}} \sum_{j \notin S} q_j. \tag{G.1}$$

Theorem G.1 (Marginals do not identify scaling; informal version of Theorem J.1)**.** *For every $\chi, \gamma > 0$, there are two instances of Equation (G.1) with identical energy multisets and identical cost multisets at every dyadic level, but*

$$\mathcal{A}_{\text{easy}}(C) \asymp C^{-\chi}, \qquad \mathcal{A}_{\text{hard}}(C) \asymp C^{-\chi/(1+\gamma)}. \tag{G.2}$$

The same two exponents hold when selection is replaced by continuous bits with distortion $q_j 2^{-2b_j}$ and cost $c_j b_j$. Hence no summary that retains only the two marginal multisets identifies the optimal exponent in either representation class.

The construction makes the missing information visible. At level ℓ, create 2^ℓ high-energy modes of size $2^{-(1+\chi)\ell}$ and 2^ℓ lower-energy modes of size $2^{-(1+\chi+\delta)\ell}$, with $\delta > \gamma\chi$. The costs are 2^ℓ copies of 1 and 2^ℓ copies of $2^{\gamma\ell}$. The easy instance pairs high energy with unit cost; the hard instance pairs it with the larger cost. All marginals agree level by level. A component with multiplicity 2^ℓ, energy $2^{-(1+s)\ell}$, and cost $2^{g\ell}$ has exponent $s/(1+g)$. This yields Equation (G.2); the full lower bound also permits non-prefix selections through a fractional relaxation.

The causal-source coupling experiment is shown in Figure 3(b,e), together with analytic distortion; Table 1b compares theoretical and held-out slopes.

G.2 THE JOINT SPECTRUM

The full unlabeled joint measure η in Equation (E.2) retains the pairing by construction. For the separable exponential response Equation (E.1), it can be compressed further.

Theorem G.2 (Joint energy–difficulty characterization; informal version of Theorem J.2)**.** *For a finite collection of modes, η is a maximal invariant under permutations of mode labels: it retains all information except the arbitrary names of the modes. For a finite or countable collection, consider the separable problem Equation (E.1) and assume*

$$\sum_j a_j < \infty, \qquad \sum_{j:a_j/c_j>t} c_j < \infty \quad \textit{for every } t > 0.$$

Then the spectrum $\mu = \sum_j c_j \delta_{a_j/c_j}$ determines $D(B)$ for every $B \geq 0$. Conversely, the complete curve $B \mapsto D(B)$ determines μ on $(0, \infty)$. Consequently, μ is a minimal exact unlabeled summary of the complete budget–distortion curve for this representation class.

Put $r_j = a_j/c_j$. The coordinatewise optimum has a water level $\tau > 0$:

$$b_j(\tau) = \frac{1}{2}\log_2(r_j/\tau)_+. \tag{G.3}$$

Writing the result in terms of μ gives two integral identities,

$$B(\tau) = \frac{1}{2\log 2}\int \log(r/\tau)_+ \,\mu(\mathrm{d}r), \qquad D(B(\tau)) = \int \min\{r, \tau\}\,\mu(\mathrm{d}r). \tag{G.4}$$

These identities prove sufficiency. They also make the inverse result possible. At differentiability points the envelope theorem gives

$$D'(B) = -2\log 2\,\tau(B). \tag{G.5}$$

Thus the curve recovers the water level, hence $B(\tau)$. With $s = \log\tau$,

$$\frac{\mathrm{d}}{\mathrm{d}s}B(e^s) = -\frac{1}{2\log 2}\mu((e^s, \infty)) \tag{G.6}$$

at every continuity point. The right-continuous tail, including its jumps, determines μ.

Minimality now has a precise meaning. If another summary T determines the complete curve, then $T(x) = T(y)$ implies $D_x = D_y$, which by Equation (G.6) implies $\mu_x = \mu_y$. Therefore μ is a function of every sufficient summary. It is also strictly coarser than the full unlabeled table η: it retains only the cost-weighted distribution of energy per unit difficulty.

Corollary G.3 (Joint-tail exponent)**.** *If $a_j \asymp j^{-\alpha}$ with $\alpha > 1$ and $c_j \asymp j^{\gamma}$ with $\gamma > -1$, then*

$$D(B) \asymp B^{-(\alpha-1)/(1+\gamma)}.$$

Indeed, the active threshold m obeys $B \asymp m^{1+\gamma}$, while the represented distortion and omitted tail are both $\asymp m^{1-\alpha}$.

When the ingredients can be measured. Controlled interventions provide a route to estimating the ingredients: counts estimate mode frequency, held-out ablations estimate predictive energy, and deliberate rounding perturbations estimate numerical amplification. The identification theorem specifies how these quantities determine the resource curve; extracting them from passive natural-language observations is an open estimation problem.

H PROOFS FOR THE FINITE-BIT AUTOREGRESSIVE MINIMAX LAW

This appendix proves the arbitrary-codebook entropy lemma, the statistical upper and lower bounds, and the power-law corollary in that order.

Lemma H.1 (Arbitrary joint finite-state representations; formal version of Lemma F.1)**.** *Let Θ_j be independent and uniform on $[-\Delta_j, \Delta_j]$, let $p_j > 0$, and put $q_j = p_j \Delta_j^2$. For $B \in \mathbb{N}_0$, every encoder–decoder pair with at most 2^B reconstruction vectors satisfies*

$$\sum_j p_j \mathbb{E}(\Theta_j - \widehat{\Theta}_j)^2 \geq \frac{2}{\pi e} \inf_{\substack{b_j \geq 0 \\ \sum_j b_j \leq B}} \sum_j q_j 2^{-2b_j}. \tag{H.1}$$

The reconstruction vectors may couple all coordinates.

Proof. First restrict to a finite coordinate set A. Write

$$d_j \coloneqq p_j \mathbb{E}(\Theta_j - \widehat{\Theta}_j)^2.$$

For any $S \subseteq A$, data processing and the codebook cardinality imply

$$B \log 2 \geq I(\Theta_S; \widehat{\Theta}). \tag{H.2}$$

Independence gives $h(\Theta_S) = \sum_{j \in S} \log(2\Delta_j)$. Moreover,

$$\begin{aligned} h(\Theta_S \mid \widehat{\Theta}) &= h(\Theta_S - \widehat{\Theta}_S \mid \widehat{\Theta}) \\ &\leq h(\Theta_S - \widehat{\Theta}_S) \\ &\leq \sum_{j \in S} h(\Theta_j - \widehat{\Theta}_j) \\ &\leq \frac{1}{2} \sum_{j \in S} \log\left(2\pi e \, \frac{d_j}{p_j}\right), \end{aligned}$$

where the last line is the Gaussian maximum-entropy inequality at fixed second moment. To see it directly, compare the error density to the centered Gaussian of variance d_j/p_j: nonnegativity of relative entropy gives $h(X) \leq \frac{1}{2} \log(2\pi d_j/p_j) + \mathbb{E}X^2/(2d_j/p_j)$. If an entropy is $-\infty$, the same bound follows by approximation. Zero mean-square error on a nondegenerate uniform coordinate would require infinite information, so it cannot occur for a finite codebook. All information and entropy expressions here use natural logarithms. Therefore

$$I(\Theta_S; \widehat{\Theta}) \geq \frac{1}{2} \sum_{j \in S} \log \frac{2q_j}{\pi e \, d_j}. \tag{H.3}$$

Choose $S = \{j : d_j < 2q_j/(\pi e)\}$ and define

$$b_j = \frac{1}{2 \log 2} \left[\log \frac{2q_j}{\pi e \, d_j}\right]_+.$$

Equations (H.2)–(H.3) give $\sum_j b_j \leq B$. On S,

$$d_j = \frac{2}{\pi e} q_j 2^{-2b_j},$$

while off S, $b_j = 0$ and $d_j \geq (2/(\pi e)) q_j$. Hence

$$\sum_{j \in A} d_j \geq \frac{2}{\pi e} \inf_{\substack{b_j \geq 0 \\ \sum_{j \in A} b_j \leq B}} \sum_{j \in A} q_j 2^{-2b_j}.$$

For countably many coordinates, let $A_m \uparrow \mathbb{N}$ and call the finite variational value $D_{A_m}(B)$. Clearly $D_{A_m}(B) \leq D_q(B)$. Extending an optimizer on A_m by zeros gives

$$D_q(B) \leq D_{A_m}(B) + \sum_{j \notin A_m} q_j.$$

Since $\sum_j q_j < \infty$, $D_{A_m}(B) \uparrow D_q(B)$. Monotone convergence of the distortion completes the proof. Randomized encoders do not improve the Bayes risk: condition on the encoder seed and average the deterministic bound. □

Theorem H.2 (Finite-bit autoregressive minimax law; formal version of Theorem F.2)**.** *Fix $R < \infty$. For every $n \in \mathbb{N}$, $B \in \mathbb{N}_0$, and source family specified by $(p_j, \Delta_j)_{j\geq 1}$ as in Section E.1, there are constants $0 < c_R \leq C_R < \infty$ such that*

$$c_R\{S_q(n) + D_q(B)\} \leq \mathfrak{E}^*(n, B) \leq C_R\{S_q(n) + D_q(B)\}, \tag{H.4}$$

where

$$S_q(n) \coloneqq \sum_j \min\{q_j, n^{-1}\}, \qquad D_q(B) \coloneqq \inf_{\substack{b_j \geq 0 \\ \sum_j b_j \leq B}} \sum_j q_j 2^{-2b_j}. \tag{H.5}$$

The lower bound holds for arbitrary 2^B-state joint encoders and arbitrary fixed shared decoding computation.

Proof. We proceed in four steps.

Step 1: loss curvature. The Bernoulli log-partition function $A(t) = \log(1 + e^t)$ satisfies

$$A''(t) = \sigma(t)(1 - \sigma(t)).$$

On $[-R, R]$, this derivative lies between two positive constants. The excess cross-entropy of logit $\widehat{\theta}$ is the Bregman divergence of A, or equivalently the Kullback–Leibler (KL) divergence, which we denote by KL. Taylor's theorem therefore yields

$$c_R(\theta - \widehat{\theta})^2 \leq \mathrm{KL}(\mathrm{Ber}(\sigma(\theta)) \,\|\, \mathrm{Ber}(\sigma(\widehat{\theta}))) \leq C_R(\theta - \widehat{\theta})^2. \tag{H.6}$$

Clipping a decoded logit to $[-R, R]$ also cannot increase cross-entropy: the derivative with respect to the decoded logit z is $\sigma(z) - \sigma(\theta)$, which has the sign of $z - \theta$. Clipping preserves the decoder's cardinality. Thus both minimax bounds can be proved using bounded decoded logits and weighted squared error.

Step 2: statistical upper bound. Let $N_j \sim \mathrm{Binomial}(n, p_j)$ be the training count of mode j. If $N_j \geq 1$, project the empirical Bernoulli mean onto $[\sigma(-\Delta_j), \sigma(\Delta_j)]$ and apply the logit map. Projection cannot increase probability error, and the logit is Lipschitz on $[\sigma(-R), \sigma(R)]$. Thus its conditional mean-square error is at most C_R/N_j. The zero estimator has error at most Δ_j^2. Choose between them using the observed count, yielding

$$\mathbb{E}[(\widehat{\theta}_j - \theta_j)^2 \mid N_j] \leq C'_R \min\{\Delta_j^2, (N_j + 1)^{-1}\}. \tag{H.7}$$

The identity

$$\mathbb{E}\frac{1}{N_j + 1} = \frac{1 - (1 - p_j)^{n+1}}{(n + 1)p_j} \leq \frac{1}{(n + 1)p_j} \tag{H.8}$$

follows by integrating the binomial generating function. Therefore

$$\begin{aligned} p_j\mathbb{E}(\widehat{\theta}_j - \theta_j)^2 &\leq C'_R \min\left\{p_j\Delta_j^2,\, p_j\mathbb{E}\frac{1}{N_j + 1}\right\} \\ &\leq C''_R \min\{q_j, n^{-1}\}. \end{aligned}$$

Summing proves the $S_q(n)$ upper bound.

Now select an allocation within an arbitrarily small additive error of $D_q(B)$, round each positive b_j down to an integer, and uniformly quantize the clipped estimate on $[-\Delta_j, \Delta_j]$. Flooring uses at most B bits and enlarges 2^{-2b_j} by at most four. By the squared triangle inequality and Equation (H.6),

$$\mathfrak{E}^*(n, B) \leq C_R\{S_q(n) + D_q(B)\}. \tag{H.9}$$

Step 3: statistical lower bound. Fix a finite coordinate set A. On coordinate $j \in A$, use the two parameters $\pm\delta_j$, where

$$\delta_j = c'_R \min\{\Delta_j, (np_j)^{-1/2}\}$$

and $c'_R > 0$ is sufficiently small. Neighboring hypercube vertices differ only when $J = j$, and bounded Bernoulli curvature gives

$$\mathrm{KL}(P_\theta^n \,\|\, P_{\theta^{(j)}}^n) \leq C_R np_j\delta_j^2 \leq C_R(c'_R)^2.$$

Here the design distribution is independent of the hypercube signs. The one-block KL is therefore exactly p_j times the changed coordinate's Bernoulli KL, and independence of blocks multiplies it by n. Choose c'_R so that this product KL is at most $1/8$. For completeness, fix all signs except j and test the remaining sign by the sign of $\widehat{\theta}_j$. A wrong sign incurs squared error at least δ_j^2. The sum of the two testing errors is at least $1 - \mathrm{TV}(P_+, P_-) \geq 3/4$, since $\mathrm{TV}(P_+, P_-) \leq \sqrt{\mathrm{KL}(P_+ \| P_-)/2}$. This inequality follows by partitioning at the positive part of the density difference, applying the log-sum inequality, and using $\mathrm{kl}(a, b) \geq 2(a-b)^2$; the latter follows from its second derivative in a, $1/[a(1-a)] \geq 4$. Averaging over the uniform signs and summing the coordinate losses yields

$$\inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E} \sum_{j \in A} p_j (\widehat{\theta}_j - \theta_j)^2 \geq c_R \sum_{j \in A} \min\{q_j, n^{-1}\}. \tag{H.10}$$

The resulting constant is independent of the size of A. Since the full risk dominates each finite-coordinate risk, taking the supremum over increasing finite sets gives the $S_q(n)$ lower bound; no interchange of an infinite supremum and a prior integral is needed.

Step 4: representation lower bound. Give the encoder θ in addition to the data. This enlarges the encoder class and can only reduce its minimax risk. Under the product-uniform prior on the parameter box, Lemma F.1 gives

$$\inf_{|\mathrm{range}(\widehat{\theta})| \leq 2^B} \sup_{\theta} \mathbb{E} \sum_{j} p_j (\widehat{\theta}_j - \theta_j)^2 \geq c D_q(B). \tag{H.11}$$

The original constrained risk is at least both Equations (H.10) and (H.11). Since $\max\{x, y\} \geq (x+y)/2$, Equation (H.6) proves the matching lower bound in Equation (F.2). □

Corollary H.3 (Power-law energies; formal version of Corollary F.3)**.** *Suppose that $n, B \geq 1$ and, for some $\chi > 0$ and constants $0 < c_- \leq c_+ < \infty$,*

$$c_- j^{-(1+\chi)} \leq q_j \leq c_+ j^{-(1+\chi)} \qquad (j \geq 1).$$

Then

$$\mathfrak{E}^*(n, B) \asymp n^{-\chi/(1+\chi)} + B^{-\chi}. \tag{H.12}$$

The data–bit crossover occurs at $B \asymp n^{1/(1+\chi)}$.

Proof. The comparison assumption does not require the actual sequence q_j to be monotone. Put $\bar{q}_j = j^{-(1+\chi)}$, so that $c_- \bar{q}_j \leq q_j \leq c_+ \bar{q}_j$. Taking infima of the coordinatewise distortion inequalities gives

$$c_- D_{\bar{q}}(B) \leq D_q(B) \leq c_+ D_{\bar{q}}(B).$$

Similarly, $\min\{cq, t\}$ lies between $\min\{c, 1\} \min\{q, t\}$ and $\max\{c, 1\} \min\{q, t\}$. It suffices to prove both rates for $\bar{q}$. Let $m_n = \lfloor n^{1/(1+\chi)} \rfloor$. Then

$$S_{\bar{q}}(n) \asymp \frac{m_n}{n} + \sum_{j > m_n} j^{-(1+\chi)} \asymp n^{-\chi/(1+\chi)}.$$

For $D_{\bar{q}}(B)$, the water-filling solution gives

$$b_j = \frac{1}{2} [\log_2(\bar{q}_j / \tau)]_+.$$

The comparison profile is monotone, so its active set is a prefix $[m]$. For $m \geq 1$, $(m+1)^{-(1+\chi)} \leq \tau \leq m^{-(1+\chi)}$. Writing $v = \tau^{-1/(1+\chi)} \in [m, m+1]$,

$$B = \frac{1+\chi}{2 \ln 2} \{m \ln v - \ln(m!)\} = \Theta(m)$$

as $m \to \infty$, by Stirling's formula and $0 \leq m \ln(v/m) \leq 1$. Active distortion is $m\tau \asymp m^{-\chi}$, and the inactive tail has the same order. Thus $D_q(B) \asymp B^{-\chi}$. The two terms cross when $B^{-\chi} \asymp n^{-\chi/(1+\chi)}$, namely $B \asymp n^{1/(1+\chi)}$. □

I Proof of Adaptation with a Known Ordering

This appendix verifies that one rounded schedule works uniformly over the specified exponent interval.

Theorem I.1 (Adaptation with a known ordering; formal version of Theorem 5.1)**.** *Fix $0 < \chi_{\min} \le \chi_{\max} < \infty$, and suppose the canonical mode order is known. Assume $q_j \asymp j^{-(1+\chi)}$ with common comparison constants and unknown $\chi \in [\chi_{\min}, \chi_{\max}]$. For every $n, B \in \mathbb{N}$, one clipped likelihood estimator and one deterministic integer bit schedule, both independent of χ, satisfy simultaneously throughout this interval*

$$\sup_{\theta_j \in [-\Delta_j, \Delta_j]} \mathbb{E}_\theta \left[\mathcal{L}_\theta(\widehat{f}) - \mathcal{L}_\theta(\theta) \right] \lesssim n^{-\chi/(1+\chi)} + B^{-\chi}. \tag{I.1}$$

Proof. The estimator in the statistical upper bound of Theorem F.2 uses only N_j, Δ_j, and the observed labels, not χ. It contributes $S_q(n) \asymp n^{-\chi/(1+\chi)}$.

Choose $A > \chi_{\max}/2$. Before rounding, the schedule in Equation (5.2) uses

$$A \sum_{j \le m} \log_2(m/j) = A \log_2(m^m/m!) = \Theta(m).$$

Flooring cannot increase the budget. It also preserves a linear lower bound: every $j \le m 2^{-2/A}$ receives at least two unrounded bits and therefore at least one rounded bit. Hence the rounded total is $\Theta(m)$, and the maximal feasible m satisfies $m \asymp B$.

Flooring enlarges each nonzero distortion by at most four. Consequently, uniformly over $\chi \le \chi_{\max}$,

$$\begin{aligned} \sum_{j \le m} q_j 2^{-2b_j} &\lesssim m^{-2A} \sum_{j \le m} j^{2A-1-\chi} \\ &\lesssim m^{-\chi}, \end{aligned}$$

because $2A - \chi \ge 2A - \chi_{\max} > 0$: integral comparison bounds the power sum by $Cm^{2A-\chi}/(2A-\chi)$, with one constant throughout the interval. Likewise the tail constants are uniform because $\chi \ge \chi_{\min} > 0$. For the finitely many budgets before the linear bound on the rounded schedule applies, enlarge the same constant; $B \in \mathbb{N}$ is positive. The omitted tail is $\sum_{j>m} q_j \asymp m^{-\chi}$. Since $m \asymp B$, the representation contribution is $O(B^{-\chi})$, proving Equation (5.1). □

J Proofs for Joint Difficulty

This appendix proves the same-marginal separation and then the positive joint-summary theorem.

Theorem J.1 (Marginals do not identify scaling; formal version of Theorem G.1)**.** *For every $\chi, \gamma > 0$, there are two instances of Equation (G.1) with identical energy multisets and identical cost multisets at every dyadic level, but*

$$\mathcal{A}_{\text{easy}}(C) \asymp C^{-\chi}, \qquad \mathcal{A}_{\text{hard}}(C) \asymp C^{-\chi/(1+\gamma)}. \tag{J.1}$$

The same two exponents hold when selection is replaced by continuous bits with distortion $q_j 2^{-2b_j}$ and cost $c_j b_j$. Hence no summary that retains only the two marginal multisets identifies the optimal exponent in either representation class.

Proof. Fix $\delta > \gamma\chi$. At each level $\ell \ge 1$, create two groups, each of multiplicity 2^ℓ. Their energies are

$$q_\ell^{\mathrm{H}} = 2^{-(1+\chi)\ell}, \qquad q_\ell^{\mathrm{L}} = 2^{-(1+\chi+\delta)\ell},$$

and the level's cost multiset contains 2^ℓ copies of 1 and 2^ℓ copies of $2^{\gamma\ell}$. The easy instance pairs q_ℓ^{H} with unit cost; the hard instance pairs it with $2^{\gamma\ell}$. The energy and cost multisets therefore agree at every level.

Consider a generic component of multiplicity 2^ℓ, per-item energy $2^{-(1+s)\ell}$, and per-item cost $2^{g\ell}$. Buying the component through level L costs

$$\sum_{\ell \le L} 2^{(1+g)\ell} = \Theta(2^{(1+g)L}),$$

and leaves energy

$$\sum_{\ell > L} 2^{-s\ell} = \Theta(2^{-sL}).$$

Thus its selection exponent is $s/(1+g)$. This prefix is also a matching lower bound: within the component, energy per cost decreases with ℓ, so the fractional relaxation is a threshold policy, and every integral policy is no better.

In the easy instance, the high-energy component has exponent χ, while the low-energy component has exponent $(\chi+\delta)/(1+\gamma) > \chi$. Splitting the budget by fixed fractions therefore gives, and the high-energy relaxation lower-bounds, $\mathcal{A}_{\text{easy}}(C) \asymp C^{-\chi}$. In the hard instance, the high-energy component has exponent $\chi/(1+\gamma)$, while the low-energy unit-cost component has exponent $\chi+\delta$. Hence $\mathcal{A}_{\text{hard}}(C) \asymp C^{-\chi/(1+\gamma)}$.

For continuous bits, the optimality condition on an active coordinate is

$$\frac{q_j 2^{-2b_j}}{c_j} = \tau.$$

For the generic component above, the active levels form a prefix through L. At level ℓ, b_ℓ is proportional to $L-\ell$; therefore

$$\sum_{\ell \le L} 2^{(1+g)\ell} b_\ell = \Theta(2^{(1+g)L}).$$

Both active distortion and the inactive tail are $\Theta(2^{-sL})$. The same exponent $s/(1+g)$ follows, and applying it to the four components completes the proof. □

Theorem J.2 (Joint energy–difficulty characterization; formal version of Theorem G.2)**.** *For a finite collection of modes, η is a maximal invariant under permutations of mode labels: it retains all information except the arbitrary names of the modes. For a finite or countable collection, consider the separable problem Equation (E.1) and assume*

$$\sum_j a_j < \infty, \qquad \sum_{j: a_j/c_j > t} c_j < \infty \quad \textit{for every } t > 0.$$

Then the spectrum $\mu = \sum_j c_j \delta_{a_j/c_j}$ determines $D(B)$ for every $B \ge 0$. Conversely, the complete curve $B \mapsto D(B)$ determines μ on $(0,\infty)$. Consequently, μ is a minimal exact unlabeled summary of the complete budget–distortion curve for this representation class.

Proof. For a finite collection, equality of $\eta = \sum_j \delta_{(v_j,a_j,c_j,\kappa_j)}$ is equality of the multisets of tuples, including multiplicities. A bijection between equal multisets is a permutation of mode labels, and every permutation preserves η. Thus η is a maximal invariant.

For the separable allocation, put $r_j = a_j/c_j$. The Lagrangian is

$$\sum_j c_j r_j 2^{-2b_j} + \lambda \left(\sum_j c_j b_j - B \right).$$

The coordinatewise Karush–Kuhn–Tucker equations give

$$b_j(\tau) = \frac{1}{2} \log_2 (r_j/\tau)_+, \qquad \tau = \frac{\lambda}{2 \log 2}.$$

Substitution yields

$$B(\tau) = \frac{1}{2 \log 2} \sum_j c_j \log (r_j/\tau)_+,$$

and

$$D(B(\tau)) = \sum_j c_j \min\{r_j, \tau\}.$$

Here is a direct countable-dimensional justification. Set $A = \sum_j a_j < \infty$. For every $\tau > 0$,

$$\sum_{r_j > \tau} c_j \le A/\tau, \qquad B(\tau) \le \frac{A}{2(\ln 2)\tau},$$

where the second inequality uses $[\ln x]_+ \le x$. On any compact interval of positive water levels, the summands defining $B(\tau)$ have a summable dominating bound proportional to a_j. Hence B is continuous. It is strictly decreasing whenever it is positive, tends to zero at the upper end of its support (possibly infinity), and tends to infinity as $\tau \downarrow 0$, since even one positive coordinate does. Thus every $B > 0$ has a unique water level. The scalar convex inequality

$$a_j 2^{-2b_j} + \lambda c_j b_j \ge a_j 2^{-2b_j(\tau)} + \lambda c_j b_j(\tau)$$

can be summed directly: all distortion and feasible cost sums are finite. With $\lambda = 2(\ln 2)\tau$ and $B(\tau) = B$, it proves optimality over every feasible countable allocation. Layer integration now gives Equation (G.4), proving sufficiency of μ.

To recover the spectrum, let $0 < B_1 < B_2$ and put $\lambda_i = 2(\ln 2)\tau(B_i)$. Applying optimality with each of the two Lagrange multipliers to the other's solution gives

$$-\lambda_1 \le \frac{D(B_2) - D(B_1)}{B_2 - B_1} \le -\lambda_2.$$

The inverse $\tau(B)$ is continuous on $(0, \infty)$. Letting the two budgets coalesce therefore proves, in the countable system itself,

$$D'(B) = -\lambda = -2 \log 2\, \tau(B)$$

at every $B > 0$. Hence the complete curve determines $\tau(B)$ and its inverse $B(\tau)$, with the zero branch above a finite support endpoint. For $s = \ln \tau$, difference quotients have a summable dominating bound: only $r_j > e^{s-h}$ contribute, whose total cost is finite. Thus the right derivative is

$$\frac{\mathrm{d}_+}{\mathrm{d}s} B(e^s) = -\frac{1}{2 \log 2} \sum_{j : r_j > e^s} c_j = -\frac{1}{2 \log 2} \mu((e^s, \infty))$$

and the left derivative replaces $r_j > e^s$ by $r_j \ge e^s$. Their difference recovers every atom's weight; the tail recovers the measure on every positive interval. The endpoint $D(0) = A$ follows by dominated convergence from $B \downarrow 0$. This argument never differentiates a limit of finite-dimensional value functions.

Finally, suppose a summary T determines the complete curve. If $T(x) = T(y)$, then $D_x(B) = D_y(B)$ for every B, so the inversion above gives $\mu_x = \mu_y$. Therefore μ is a function of T, proving minimality in the stated sense. □

Proof of Corollary G.3. Put $\bar{a}_j = j^{-\alpha}$ and $\bar{c}_j = j^{\gamma}$, with $a_- \bar{a}_j \le a_j \le a_+ \bar{a}_j$ and $c_- \bar{c}_j \le c_j \le c_+ \bar{c}_j$. Feasible-set inclusion gives

$$a_- D_{\bar{a},\bar{c}}(B/c_-) \le D_{a,c}(B) \le a_+ D_{\bar{a},\bar{c}}(B/c_+).$$

It suffices to use the monotone ratio $\bar{a}_j/\bar{c}_j = j^{-(\alpha+\gamma)}$. Since $\alpha > 1$ and $\gamma > -1$, its exponent is positive. If m modes are active, put $v = \tau^{-1/(\alpha+\gamma)} \in [m, m+1]$. The budget is

$$B = \frac{\alpha + \gamma}{2 \ln 2} \sum_{j \le m} j^{\gamma} \ln(v/j) \asymp m^{1+\gamma}.$$

For the lower bound restrict to $m/4 \le j \le m/2$; for the upper bound use integral comparison and $\int_0^1 x^{\gamma} \ln(1/x)\, \mathrm{d}x = (1+\gamma)^{-2}$. Replacing m by v in the logarithm adds only $O(m^{\gamma})$. The inactive tail is $\asymp m^{1-\alpha}$. The active distortion is $\tau \sum_{j \le m} c_j$, with $\tau \asymp m^{-(\alpha+\gamma)}$, and has the same order. Eliminating m gives

$$D(B) \asymp B^{-(\alpha-1)/(1+\gamma)}.$$

□

K Proof of the Finite-Precision Realization

This appendix verifies the exact matrix map, the floating-point perturbation, and the lower bound for the stated separable representation.

We use the candidate/query encoding and projections of Section 5 and Equation (5.4). The supplied logit table is $\vartheta_j = w_j^{\mathrm{mode}}$ for $j \le M$, extended by zero outside the represented prefix; θ_j remains the source parameter.

Theorem K.1 (Finite-precision realization; formal version of Theorem 5.2)**.** *Fix $K, M \ge 1$, $R < \infty$, $\beta \ge 0$, and the construction above.*

1. *In exact arithmetic, if $|w_j^{\mathrm{mode}}| \le R$ for every $j \in [M]$, then a query whose matching candidate has any mode $j \ge 1$ satisfies*
$$|z - \vartheta_j| \le 2R(K-1)e^{-\beta}. \tag{K.1}$$

2. *Set $w_j^{\mathrm{mode}} = \theta_j$ for $j \in [M]$ and uniformly quantize it with $b_j \in \mathbb{N}_0$ bits on $[-\Delta_j, \Delta_j]$, using zero when $b_j = 0$. Assume stable max-shifted softmax with exactly representable shifted scores $0, -\beta$ and an exponent range containing their exponentials. Every intermediate result is finite, and every executed addition, multiplication, and division obeys*
$$\mathrm{fl}(a \circ b) = (a \circ b)(1+\delta), \qquad |\delta| \le u, \quad \circ \in \{+, \times, /\},$$
with exact zero results preserved. Exponentials obey $\mathrm{fl}(e^x) = e^x(1+\delta)$, $|\delta| \le u$. Sum both the softmax denominator and the K candidate-weighted scalar contributions by balanced pairwise accumulation. Each candidate contributes its computed attention weight times its stored scalar logit, using zero for modes outside $[M]$. For the resulting predictor f_{fp}, if $(\lceil \log_2 K \rceil + 4)u \le 1/4$, then the held-out excess cross-entropy obeys
$$\mathcal{L}_\theta(f_{\mathrm{fp}}) - \mathcal{L}_\infty \le C_R \left\{ (K-1)^2 e^{-2\beta} + u^2(1+\log K)^2 + \sum_{j=1}^{M} q_j 2^{-2b_j} + \sum_{j>M} q_j \right\}. \tag{K.2}$$

3. *Within the separable scalar-table representation in which coordinate j has at most 2^{b_j} values, the bit-dependent term in Equation (5.6) is tight up to constants.*

Here and below, C_R depends only on the logit bound R.

Proof. For the score matrix $U = W_Q^\top W_K / \sqrt{d_h}$ of the stated construction,
$$x_q^\top U x_t = \begin{cases} \beta, & t = t_\star, \\ 0, & t \ne t_\star. \end{cases}$$
Softmax therefore gives Equation (5.4). Since $\sum_{t \ne t_\star} \alpha_t = 1 - \alpha_\star$ and $|\vartheta_j| \le R$,
$$\begin{aligned} |z - \vartheta_{j_\star}| &= \Big| \sum_{t \ne t_\star} \alpha_t (\vartheta_{j_t} - \vartheta_{j_\star}) \Big| \\ &\le 2R(1 - \alpha_\star) \\ &= 2R \frac{K-1}{e^\beta + K - 1} \\ &\le 2R(K-1)e^{-\beta}. \end{aligned}$$
For $b_j \ge 1$, nearest-point quantization to 2^{b_j} equally spaced levels including the endpoints of $[-\Delta_j, \Delta_j]$ has error at most $\Delta_j/(2^{b_j} - 1) \le 2\Delta_j 2^{-b_j}$. For $b_j = 0$, use zero, whose error is at most Δ_j. Thus both cases give $|\widetilde{w}_j^{\mathrm{mode}} - w_j^{\mathrm{mode}}| \le 2\Delta_j 2^{-b_j}$ for $j \in [M]$. Write $\widetilde{\vartheta}_j = \widetilde{w}_j^{\mathrm{mode}}$ on $[M]$ and $\widetilde{\vartheta}_j = 0$ otherwise. Distractor quantization contributes at most $2R(1 - \alpha_\star)$, so it is absorbed by the routing term.

It remains to justify the arithmetic term. Stable max shifting replaces the scores $s_t \in \{0, \beta\}$ by $\bar{s}_t = s_t - \beta \in \{-\beta, 0\}$ without changing the exact softmax. Let $y_t = e^{\bar{s}_t}$ and $h = \lceil \log_2 K \rceil$. By assumption, the computed unnormalized weight is

$$\widehat{y}_t = y_t(1 + \delta_t), \qquad |\delta_t| \le u.$$

For any finite summands $v_1, \dots, v_K$, a balanced binary tree has at most h additions on each leaf-to-root path. Expanding the computed sum through this tree, using the stated error model at every addition, gives

$$\mathrm{fl}\Big(\sum_{t=1}^{K} v_t\Big) = \sum_{t=1}^{K} v_t(1 + \eta_t), \qquad |\eta_t| \le \gamma_h \coloneqq \frac{hu}{1 - hu}.$$

Indeed, each leaf coefficient is a product of at most h factors $1 + \delta$. Its deviation from one is at most $(1 + u)^h - 1 \le \sum_{k \ge 1} (hu)^k = \gamma_h$. Thus the absolute summation error is at most $\gamma_h \sum_t |v_t|$, including mixed signs and cancellation; for nonnegative summands this is a relative bound. For $K = 1$, no addition is needed and $\gamma_0 = 0$. Applying the positive bound to the computed exponentials gives

$$\widehat{S} = \left(\sum_t \widehat{y}_t\right)(1 + \delta_S), \qquad |\delta_S| \le \gamma_h.$$

Including the final division, the relative-error factors can be bounded explicitly. Put $\bar{\delta} = \sum_t \alpha_t \delta_t$, so $|\bar{\delta}| \le u$, and let $|\delta_{D,t}| \le u$ be the division error. Then

$$\frac{\widehat{\alpha}_t}{\alpha_t} = \frac{(1 + \delta_t)(1 + \delta_{D,t})}{(1 + \bar{\delta})(1 + \delta_S)}.$$

When $(h + 4)u \le 1/4$, we have $u \le 1/16$ and $\gamma_h \le 4hu/3 \le 1/3$. The denominator is therefore at least $(1 - u)(1 - \gamma_h) \ge 5/8$. Subtracting numerator and denominator yields

$$\left|\frac{\widehat{\alpha}_t}{\alpha_t} - 1\right| \le \frac{3u + u^2 + (1 + u)\gamma_h}{(1 - u)(1 - \gamma_h)} \le \frac{2(h + 4)u}{5/8} \le 8(h + 4)u.$$

Consequently

$$\widehat{\alpha}_t = \alpha_t(1 + \varepsilon_t), \qquad |\varepsilon_t| \le 8(h + 4)u.$$

Summing against α_t proves Equation (C.2).

For the output, evaluate the K products $\widehat{v}_t = \mathrm{fl}(\widetilde{\vartheta}_{j_t} \widehat{\alpha}_t)$ and their balanced pairwise sum $\widehat{z}$. This uses the same depth h, even when candidate modes repeat. The stored logits lie in $[-R, R]$, with zero used outside $[M]$. Positivity of the computed weights and the preceding bound give $\sum_t \widehat{\alpha}_t \le 1 + 8(h + 4)u \le 3$. The product and signed-sum bounds now imply

$$\begin{aligned}\left|\widehat{z} - \sum_t \widetilde{\vartheta}_{j_t} \widehat{\alpha}_t\right| &\le \{u + (1 + u)\gamma_h\} \sum_t |\widetilde{\vartheta}_{j_t}| \widehat{\alpha}_t \\ &\le 4R(h + 1)u.\end{aligned}$$

Here $u + (1 + u)\gamma_h = (h + 1)u/(1 - hu)$. Replacing the computed weights by the exact ones adds at most $R\|\widehat{\alpha} - \alpha\|_1 \le 8R(h + 4)u$, whence

$$\left|\widehat{z} - \sum_t \widetilde{\vartheta}_{j_t} \alpha_t\right| \le 36R(h + 1)u \le C_R u(1 + \log K).$$

This proves Equation (C.3) for arbitrary signs, zero coordinates, and repeated candidates. For every matching mode, represented or not, combining the contributions gives

$$|\widehat{z} - \theta_{j_\star}| \le 4R(K - 1)e^{-\beta} + 2\Delta_{j_\star} 2^{-b_{j_\star}} \mathbf{1}\{j_\star \le M\} + |\theta_{j_\star}| \mathbf{1}\{j_\star > M\} + C_R u(1 + \log K).$$

The two mode-specific terms are mutually exclusive. The squared sum is therefore bounded by a constant times routing error, arithmetic error, and their squared mode-specific sum. Averaging over the selected coordinate adds $\sum_{j \le M} q_j 2^{-2b_j} + \sum_{j > M} q_j$. An unrepresented matching mode can still attend to represented distractors; that contribution is already bounded by the routing term above, rather than being assumed zero. Finally $A''(z) \le 1/4$ for every real logit, so the upper KL curvature

bound applies even if floating-point error moves $\widehat{z}$ slightly outside the original box. This proves Equation (5.6).

For the converse, restrict to the separable scalar table. Coordinate j has at most 2^{b_j} reconstruction values. Under the same product-uniform prior on the coordinate logits, the one-dimensional specialization of Lemma F.1 gives Bayes weighted squared error at least $cq_j 2^{-2b_j}$ on each coordinate. Expectations add under this one common prior, so the sum lower-bounds worst-case scalar-table distortion; it is not a pointwise assertion about every parameter. Apply the lower curvature bound to the clipped table decoder to obtain the stated tightness within that representation. The construction stores one $d \times d$ score matrix, one $d \times d$ value matrix, and one d-coordinate head, hence $N = 2d^2 + d$. □

L Proof of the Transformer Bit-Complexity Lemma

This appendix counts the dense forward–backward graph stated in Section C.3. Its sequence length, hidden width, feed-forward width, head count, and depth are $\ell, d, m_{\mathrm{ff}}, H, L$, respectively. The precision indices specify the operand, accumulator, and stored-activation formats whose costs are counted.

Lemma L.1 (Transformer forward–backward bit complexity; formal version of Lemma C.1)**.** *For the graph above, with positive dimensions and conventional dense matrix multiplication,*

$$\mathrm{BitOps}_{\mathrm{fb}} \asymp L\Big[\big(\ell d^2 + \ell d m_{\mathrm{ff}} + \ell^2 d\big)\mathsf{M}(b_{\mathrm{mm}}) + H\ell^2\mathsf{S}(b_{\mathrm{sm}})\Big]. \tag{L.1}$$

$$\mathrm{Mem}_{\mathrm{fb}} \asymp b_w L(d^2 + d m_{\mathrm{ff}}) + b_a L\ell(d + m_{\mathrm{ff}}) + b_{\mathrm{st}} LH\ell^2 + \mathrm{Mem}_{\mathrm{hp}}\,. \tag{L.2}$$

The memory $\mathrm{Mem}_{\mathrm{fb}}$ is measured in bits. The values b_w, b_a, and b_{st} count aggregate bits per logical weight, activation, and saved-attention slot at peak memory; they include any same-shaped gradient or backward-workspace buffer materialized by this eager implementation. The term $\mathrm{Mem}_{\mathrm{hp}}$ contains master weights, optimizer state, and all other explicitly retained high-precision data. A streaming, recomputing attention implementation replaces the quadratic attention state saved for backward by linear online-softmax state plus temporary tile workspace, while leaving the dense attention arithmetic order unchanged.

Proof. Suppressing constant head reshapes, the four query, key, value, and output projections perform $4\ell d^2$ multiply–accumulates per layer. Let Q_{mat}, K_{mat}, and V_{mat} denote the query, key, and value matrices, and let P denote the matrix of attention probabilities. The query–key score product $Q_{\mathrm{mat}}K_{\mathrm{mat}}^{\top}$ and probability–value product PV_{mat} each perform $\ell^2 d$, and the two feed-forward maps perform $2\ell d m_{\mathrm{ff}}$. Thus the forward matrix count is

$$F_{\mathrm{mat}} = 4\ell d^2 + 2\ell^2 d + 2\ell d m_{\mathrm{ff}}. \tag{L.3}$$

For any product $Y = AB$, reverse mode forms $\nabla_A = \nabla_Y B^{\top}$ and $\nabla_B = A^{\top}\nabla_Y$. Each has the same scalar multiply–accumulate count as the forward product. Consequently the matrix part of one forward–backward pass is exactly three times Equation (L.3), up to additions, transposes, and casts.

There are $H\ell$ softmax rows of length ℓ. Stable max shifting, exponentiation, reduction, division, and the reverse-mode row calculation together cost $\Theta(H\ell^2\mathsf{S}(b_{\mathrm{sm}}))$ by the definition of S. Given an incoming row gradient v, the softmax Jacobian is applied by

$$p \odot \big(v - \langle p, v\rangle \mathbf{1}\big),$$

which uses a constant number of length-ℓ passes. No explicit $\ell \times \ell$ Jacobian is formed. Multiplying by L and absorbing fixed projection constants proves Equation (C.5). Normalization, pointwise activation, dropout, and casts have only $O(L\ell(d + m_{\mathrm{ff}}) + LH\ell^2)$ scalar entries and are the stated lower-order terms for this computation.

For memory, the dense projection and feed-forward weights contain $\Theta(L(d^2 + d m_{\mathrm{ff}}))$ scalars. Saved residual, projection, and feed-forward activations contain $\Theta(L\ell(d + m_{\mathrm{ff}}))$ scalars. Eager reverse mode also saves $LH\ell^2$ attention probabilities. Multiplying each collection by its aggregate bits per logical slot (including same-shaped peak gradient or workspace buffers) and adding separately retained high-precision state gives Equation (C.6). A tiled online softmax stores row normalization statistics and tiles rather than the complete probability matrix, so its persistent attention state is linear in $H\ell$; recomputing tiles during backward does not change the dense arithmetic order. These counts also give the two-sided bounds because each named tensor and matrix product is present in the specified graph. □

M Proof of the Compute–Precision Pareto Law

The proof distinguishes the precision required to match a target error from the exact precision chosen by a finite-budget optimizer.

Theorem M.1 (Compute–precision Pareto law; formal version of Theorem C.2)**.** *Assume Equations (C.7) and (C.8), $\kappa_{\mathrm{mm}}(M) \asymp M^{r}$, and $\kappa_{\mathrm{sm}}(M) \asymp M^{r_{\mathrm{sm}}}$, with all displayed relations two-sided and $r, r_{\mathrm{sm}} \geq 0$. For the specified dense graph, $H \leq d$ implies $G \gtrsim S$ and hence $g \geq s$. Item 3 concerns general operation counts satisfying the same two-sided relations; items 1, 2, and 4 apply to both settings.*

1. *The smallest matrix precision that keeps matrix rounding at the model error is*

$$b_{\mathrm{match}}(M) = \left(r + \frac{\chi}{2}\right) \log_2 M + O(1). \tag{M.1}$$

2. *If the matrix term dominates the bracket in Equation (C.8) at matched precisions—in particular in the case $g \geq s$ when the remaining precision factors do not reverse that dominance—then*

$$\inf_{M,b,b_{\mathrm{sm}}:\,\mathrm{cost}\leq C} \mathcal{E} \asymp C^{-\chi/(1+\chi+g)} (\log C)^{\chi\nu/(1+\chi+g)}, \tag{M.2}$$

   up to other declared slowly varying factors.

3. *For general operation counts, if $s > g$, b_{sm} is fixed, and the softmax term dominates computation, then for budgets in the pre-floor range*

$$C^{1/(1+\chi+s)} \lesssim 2^{2b_{\mathrm{sm}}/(\chi+2r_{\mathrm{sm}})},$$

$$\inf_{M,b:\,\mathrm{cost}\leq C} \mathcal{E} \asymp C^{-\chi/(1+\chi+s)}. \tag{M.3}$$

   Reducing matrix precision cannot improve this leading exponent.

4. *At fixed matrix precision b, with softmax precision optimized, the model–rounding crossover and the infimum error over model scales are*

$$M_b \asymp 2^{2b/(\chi+2r)}, \qquad \mathcal{E}_{\mathrm{floor}}(b) \asymp 2^{-2\chi b/(\chi+2r)}. \tag{M.4}$$

   A fixed b_{sm}-bit softmax similarly supports the scaling phase only while

$$M \lesssim 2^{2b_{\mathrm{sm}}/(\chi+2r_{\mathrm{sm}})}. \tag{M.5}$$

The displayed two-sided risk and arithmetic assumptions define the class over which the infima and matching bounds are taken.

Proof. The matrix rounding term is at most a constant multiple of the model term exactly when

$$M^{2r} 2^{-2b} \lesssim M^{-\chi}.$$

Taking base-two logarithms proves Equation (C.9); the two-sided risk assumption gives necessity for matching the model-error scale. For the dense graph, positive head dimensions sum to d, so

$$S(M) \asymp LH\ell^2 \leq Ld\ell^2 \lesssim G(M).$$

Hence $g \geq s$. The general-count branch uses Equations (C.7) and (C.8) with independent nonnegative exponents g, s.

Suppose first that the matrix term dominates the bracket in Equation (C.8). Choose both numerical errors below a constant multiple of $M^{-\chi}$. By Equation (C.9),

$$C \asymp M^{1+\chi+g} (\log M)^{\nu}.$$

Let $\zeta = 1 + \chi + g$. Substitution into $M^{\zeta}(\log M)^{\nu}$ gives cost within constant factors of C at

$$M \asymp C^{1/\zeta} (\log C)^{-\nu/\zeta}.$$

Substituting into $M^{-\chi}$ proves the upper bound in Equation (C.10). For the converse, consider any feasible configuration with risk at most ε. The risk relation gives

$$M \gtrsim \varepsilon^{-1/\chi}, \qquad 2^{-2b} \lesssim \varepsilon, \qquad b \gtrsim \log(1/\varepsilon),$$

using $r \geq 0$. The matrix contribution alone forces

$$C \gtrsim \varepsilon^{-\zeta/\chi}[\log(1/\varepsilon)]^{\nu}.$$

Inverting proves the lower rate for every configuration. For $\nu = 0$, the logarithmic factor equals one and the same argument applies.

Now use general operation counts with $s > g$, fixed b_{sm}, and dominating softmax work before its numerical floor. Matching the matrix error still costs only $b = \Theta(\log M)$, but the leading cost is

$$C \asymp M^{1+\chi+s} b_{\mathrm{sm}}^{\nu_{\mathrm{sm}}} \asymp M^{1+\chi+s}.$$

Within the range where $M^{2r_{\mathrm{sm}}} 2^{-2b_{\mathrm{sm}}} \lesssim M^{-\chi}$, substitution proves Equation (C.11). Conversely, risk at most ε requires $M \gtrsim \varepsilon^{-1/\chi}$, and the softmax cost is at least a constant times $M^{1+\chi+s}$. Thus $C \gtrsim \varepsilon^{-(1+\chi+s)/\chi}$. Since the leading term contains no b, changing matrix precision cannot change this exponent. Equivalently, substituting $M \asymp C^{1/(1+\chi+s)}$ gives exactly the pre-floor budget condition in the theorem statement.

Finally, at fixed b, the model and matrix-rounding terms cross when

$$M^{-\chi} \asymp M^{2r} 2^{-2b}, \qquad M^{\chi+2r} \asymp 2^{2b}.$$

This gives M_b in Equation (C.12). Choose softmax precision so its error is at most the other terms. If $M \leq M_b$, the model term is at least a constant times $M_b^{-\chi}$; if $M \geq M_b$, the matrix-rounding term is, since $r \geq 0$. At M_b both have this order. When $r = 0$, the comparison expression $M^{-\chi} + 2^{-2b}$ decreases to 2^{-2b}; its infimum has the displayed order and M_b is the crossover scale. Replacing r, b by $r_{\mathrm{sm}}, b_{\mathrm{sm}}$ proves Equation (C.13). The fixed-share ratios in Equations (C.14) and (C.15) follow by dividing the baseline cost or storage by its quantized counterpart. □

N Proof of the Compute Phases

This appendix derives each phase only from the two-sided resource assumptions in Section 6.

Theorem N.1 (Precision-aware compute phases; formal version of Theorem 6.1)**.** *Let $\chi > 0$ and $f, \nu \geq 0$. Optimize over configurations using $n(M) \asymp M^{1+\chi}$ blocks whose total excess risk and cost satisfy, uniformly over $M, b \geq 1$,*

$$\mathcal{E}(M, b) \asymp M^{-\chi} + \kappa(M)^2 2^{-2b}, \qquad \mathrm{cost}(M, b) \asymp M^{f+1+\chi} b^{\nu}.$$

Let $\mathcal{R}(C) - \mathcal{L}_\infty = \inf_{\mathrm{cost}(M,b) \leq C} \mathcal{E}(M, b)$. Then:

1. *If $1 \lesssim \kappa(M) \lesssim M^{\rho}$, then the least model-error-matching precision is $b = \Theta(\log M)$, and*

$$\mathcal{R}(C) - \mathcal{L}_\infty \asymp C^{-\chi/(f+1+\chi)} (\log C)^{\chi\nu/(f+1+\chi)} \tag{N.1}$$

   up to other slowly varying factors specified in the assumptions.

2. *If $\kappa(M) = \exp(\Theta(M^{\lambda}))$, $\lambda > 0$, then $b = \Theta(M^{\lambda})$, and the loss exponent is $\chi/(f+1+\chi+\lambda\nu)$.*

3. *If $q_j \asymp j^{-(1+\chi)}$, adaptive per-mode precision has error $\Theta(B^{-\chi})$, whereas a uniform-width prefix has optimal error*

$$\Theta((B/\log B)^{-\chi}). \tag{N.2}$$

Each converse uses the representation, operation set, storage layout, and bit-operation cost specified above.

Proof. Matching numerical error to approximation error requires

$$\kappa(M)^2 2^{-2b} \lesssim M^{-\chi}.$$

Taking base-two logarithms gives Equation (6.3). The two-sided amplification assumption gives the matching necessity.

If $1 \lesssim \kappa(M) \lesssim M^{\rho}$, then $b = \Theta(\log M)$. Multiplying per-block work, number of blocks, and bit-operation cost yields

$$C \asymp M^{f} M^{1+\chi} (\log M)^{\nu} = M^{\zeta} (\log M)^{\nu}, \qquad \zeta = f + 1 + \chi.$$

Substitution into the preceding cost relation gives the budget-matched scale

$$M \asymp C^{1/\zeta} (\log C)^{-\nu/\zeta}.$$

Substitution into $M^{-\chi}$ proves the upper bound in Equation (6.1). Conversely, risk at most ε and $\kappa(M) \gtrsim 1$ require $M \gtrsim \varepsilon^{-1/\chi}$ and $b \gtrsim \log(1/\varepsilon)$. The necessary block count and per-block work give

$$C \gtrsim \varepsilon^{-(f+1+\chi)/\chi} [\log(1/\varepsilon)]^{\nu}.$$

Inverting proves the lower bound, including $\nu = 0$.

If $\kappa(M) = \exp(\Theta(M^{\lambda}))$ for $\lambda > 0$, then Equation (6.3) gives $b = \Theta(M^{\lambda})$, so

$$C \asymp M^{f+1+\chi+\lambda\nu}.$$

This gives the claimed upper rate. Conversely, risk at most ε forces $M \gtrsim \varepsilon^{-1/\chi}$ and $b \gtrsim M^{\lambda} + \log(1/\varepsilon)$, whence

$$C \gtrsim M^{f+1+\chi+\lambda\nu} \gtrsim \varepsilon^{-(f+1+\chi+\lambda\nu)/\chi}.$$

This proves the exponent in both directions.

For adaptive precision, set $a_j = q_j$ and $c_j = 1$ in Corollary G.3; then $D(B) \asymp B^{-\chi}$. A uniform-width prefix of length M and width b uses $B = Mb$ bits. Its omitted tail is $\Theta(M^{-\chi})$, and its represented-coordinate quantization is $\Theta(2^{-2b})$ because $\sum_{j \le M} q_j = \Theta(1)$. Choose $M \asymp B/\log B$ with a sufficiently small constant and $b = \lfloor B/M \rfloor$ for the upper bound in Equation (6.2). For the converse, set $T_B = (B/\log_2 B)^{-\chi}$ and fix $L > 2/\chi$. If $M \le LB/\log_2 B$, the tail is at least $cL^{-\chi} T_B$. Otherwise $b \le B/M < (\log_2 B)/L$, so $2^{-2b} > B^{-2/L} \gtrsim T_B$ for sufficiently large B. Every uniform-width prefix therefore pays at least a constant times T_B, including zero-width prefixes. □

O Experimental Details

This appendix describes the 1,024-mode synthetic study. Its five-seed percentile intervals apply to this study; the other experiments use the BCa procedures and seed or item units specified in their respective captions.

Source families. Each run uses 1024 modes and bounded logits. For each $\chi \in \{0.5, 1, 2\}$, two different frequency–amplitude decompositions produce the same power $q_j \asymp j^{-(1+\chi)}$. Training blocks are independent regenerations. Held-out cross-entropy is evaluated exactly from the population probabilities, so no test-set Monte Carlo error is mixed into the reported training-seed variability.

Grid and methods. The grid contains four values of n, five values of B, three exponents, two decompositions, four allocation methods, and five seeds, for $4 \cdot 5 \cdot 3 \cdot 2 \cdot 4 \cdot 5 = 2400$ configuration–seed pairs. The known-energy method solves the finite energy sequence's water-filling problem. The universal method uses Equation (5.2). Uniform width assigns a common precision to a prefix, and shuffled allocation breaks the energy–precision pairing. Allocations never spend precision beyond the statistical resolution of a mode.

Uncertainty. For each of the 480 configurations, we report the mean over five seeds. The plotted interval is the 2.5 and 97.5 percentile envelope from 5000 bootstrap resamples of the seed mean, using bootstrap seed zero and expanding the interval when necessary to contain the observed mean.

Reproducibility details. The raw rows, summaries, deterministic analytic curves, plotting sources, and experiment code accompany the paper. The experiments use local CPUs, one common exponent specification across configurations, and analytic arithmetic cost as the theoretical resource.

P Mixed Precision in Practice

The resource variables in Section C.3 deliberately separate compute precision from storage precision. In the asymptotic Pareto theorem, b is an effective accuracy parameter with unit roundoff $\Theta(2^{-b})$, and the modeled format width is proportional to b. For a concrete floating-point format, total width and significand precision must instead be distinguished: the numerical horizon uses the actual unit roundoff, not the total number of stored bits. This distinction is necessary in real mixed-precision systems:

- A softmax may compute its maximum, exponential, reduction, and division in 32-bit floating point (FP32) even when the two surrounding attention matrix products use lower precision. NVIDIA Transformer Engine, for example, documents an 8-bit floating-point (FP8) attention path with FP8 matrix multiplications and FP32 softmax (NVIDIA, 2026).
- A matrix kernel may accumulate products in a precision wider than its input and output formats. Accumulation precision therefore belongs in the arithmetic model; it does not imply that the complete attention probability tensor is stored at that precision.
- The softmax output can be cast back to a lower-precision format before the value product. Conversely, an eager implementation may retain the probability tensor for backward. These are distinct choices represented by softmax effective accuracy and b_{st}, respectively.
- Mixed-precision training may retain full-precision master weights and optimizer state even when forward weights, activations, and gradients use a narrower format (Micikevicius et al., 2018). Such residual state is included in $\mathrm{Mem}_{\mathrm{hp}}$, not hidden inside an effective average bit width.
- Tiled online softmax can avoid materializing the quadratic attention matrix by saving normalization statistics and recomputing attention tiles in backward (Dao et al., 2022; NVIDIA, 2026). This changes the storage model and memory traffic, while dense attention still performs a quadratic number of score and value arithmetic operations.

These observations explain why Equations (C.14) and (C.15) are operation and storage ratios rather than throughput predictions. Kernel fusion, memory traffic, device-specific integer or floating-point units, communication, and parallel scheduling can make the same bit-operation count run at different wall-clock speeds. Implementations can therefore differ in speed even when these analytic counts agree.

Q Extended Experimental Results

This appendix examines robustness to spectral shape and ordering, numerical formats, and pretrained-model scale, and provides detailed slope estimates and reference losses. Captions specify uncertainty for each stochastic quantity.

Q.1 Candidate-Count Replication and Exponent Grid

The paired $K = 8$ versus $K = 32$ comparison in Figure 10 tests how candidate count affects the finite-sample construction and trained learner. The profile-by-budget comparison in Figure 5 tests exponent agreement. Table 4b gives unbinned numerical slope estimates. The two-stage study crosses four profiles and prefixes, three sample and state budgets, $K = 8, 32$, and eight paired seeds. Each sample budget restarts the learner, whereas state budgets reuse the fitted weights. Repeated settings at the minimum $n = 64$ are aggregated within seed before bootstrapping. Display bins depend only on G_{μ_M}.

Q.2 Routing, Interventions, and Profile Robustness

The routing measurements and learned/known/frozen comparison in Figure 3(c,f) relate retrieval quality to prediction error. The decomposition and interventions in Tables 1c and 1d isolate the error sources. Figure 6 tests allocation when the spectrum departs from an exact power law.

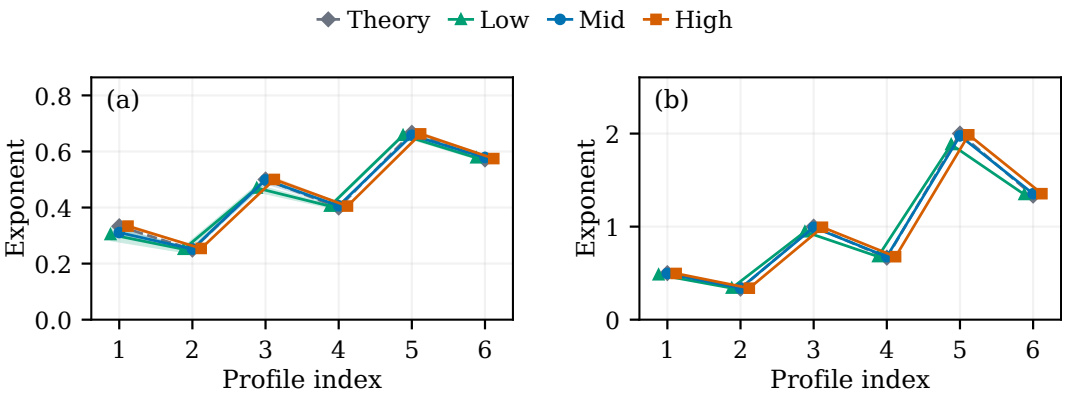


Figure 5: **Six-profile exponent agreement.** (a) Data-limited and (b) memory-limited fitted exponents for profile indices $(0.5, 0), (0.5, 0.5), (1, 0), (1, 0.5), (2, 0), (2, 0.5)$. Green, blue, and orange denote low, middle, and high fixed resource ratios; gray diamonds give the theoretical exponent. Bands are 95% BCa intervals over sixteen paired slopes: eight seeds for each of two frequency–amplitude decompositions (constant and alternating logit radius).

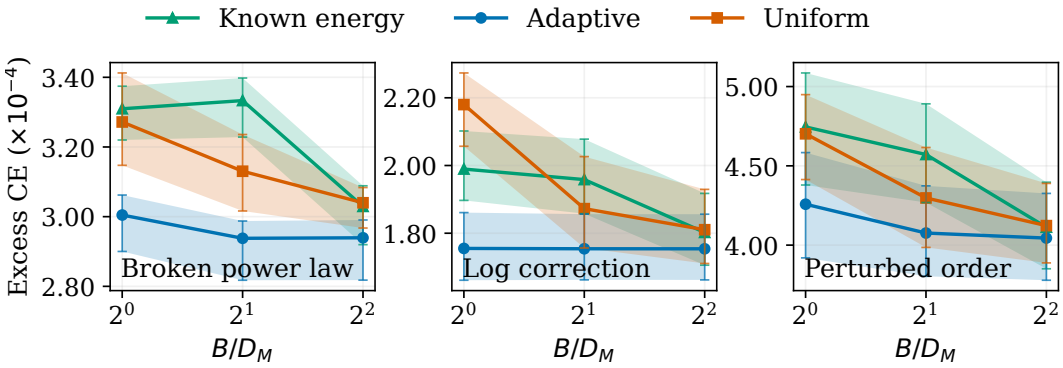


Figure 6: **Allocation beyond exact power laws.** Left, center, and right show a broken power law, logarithmic correction, and perturbed ordering, respectively, at $(\chi, \gamma, M) = (1, 0.5, 256)$. Each panel compares known-energy, adaptive, and uniform allocation. Known-energy uses the true q_i; empirical allocation uses the count-corrected scores in Section S. Bands are 95% BCa intervals over eight paired seeds.

Table 3: **Routing and precision contributions to prediction loss.** At $(\chi, \gamma, M, K) = (1, 0.5, 64, 32)$, data ratio two and unit bit ratio: augmented loss is $L_{\mathrm{lr},q} + M^{-\chi}$; Route+ and Format+ are the positive contrasts R_+ and F_+ defined in Appendix S. Lower is better. Construct. denotes the finite-sample construction. Each $\pm$ is the larger deviation to the 95% BCa endpoints.

Method	Aug. loss	Route+	Format+
Two-stage	0.0191 $\pm$0.0002	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000
Construct.	0.0170 $\pm$0.0002	0.0002 $\pm$0.0000	0.0002 $\pm$0.0001
Known route	0.0192 $\pm$0.0002	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000
Frozen route	0.0605 $\pm$0.0008	0.0339 $\pm$0.0017	0.0000 $\pm$0.0000
No precond.	0.0230 $\pm$0.0003	0.0003 $\pm$0.0000	0.0000 $\pm$0.0000
Uniform bits	0.0191 $\pm$0.0004	0.0000 $\pm$0.0000	0.0000 $\pm$0.0000

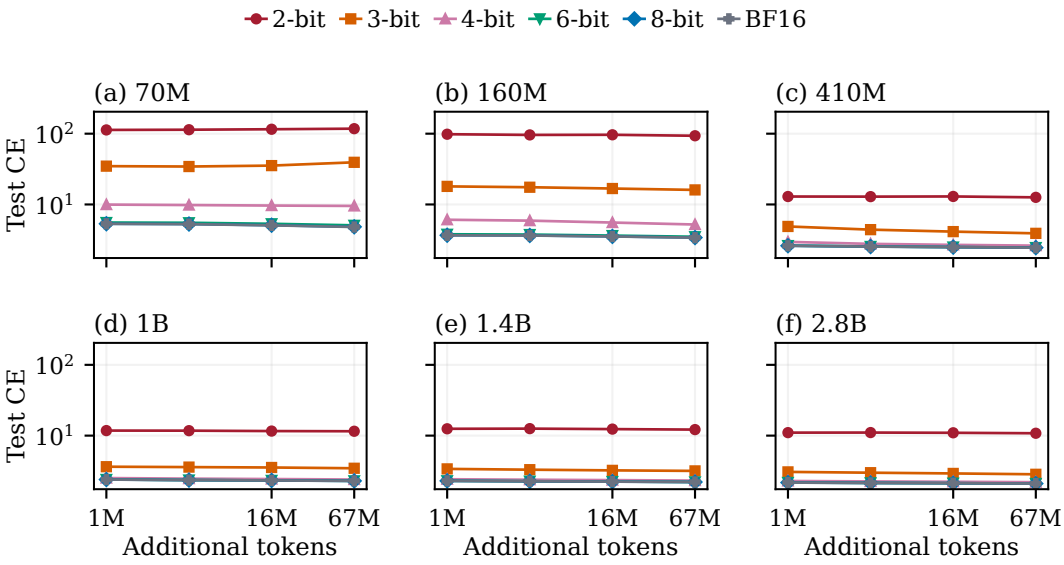


Figure 7: **Complete additional-data and deployment-format scan.** (a–f) Pythia 70M, 160M, 410M, 1B, 1.4B, and 2.8B test CE after four nested additional-token budgets. Curves show complete group-128 2/3/4/6/8-bit deployments and BF16; the shared logarithmic CE axis compares absolute loss and keeps low-bit collapse visible. Bands are 95% BCa intervals over eight paired training seeds evaluated on the same 96 test windows.

Q.3 Stored Precision and Pretrained Transformers

At 70M and 160M, the 2-bit calibration CE reaches 114.67 and 97.78 nats, respectively. These losses exceed the pre-specified $[0, 20]$-nat prediction range, and the fitting procedure yields no converged, in-range solutions when calibration includes 2-bit weights. At 410M–2.8B, all four response functions yield such solutions. A sensitivity analysis fits 4/8-bit and BF16 separately at every scale and predicts 6-bit loss and the largest data budget. Tables 5 to 8 report both calibration choices; Section S defines the response functions and evaluation sets.

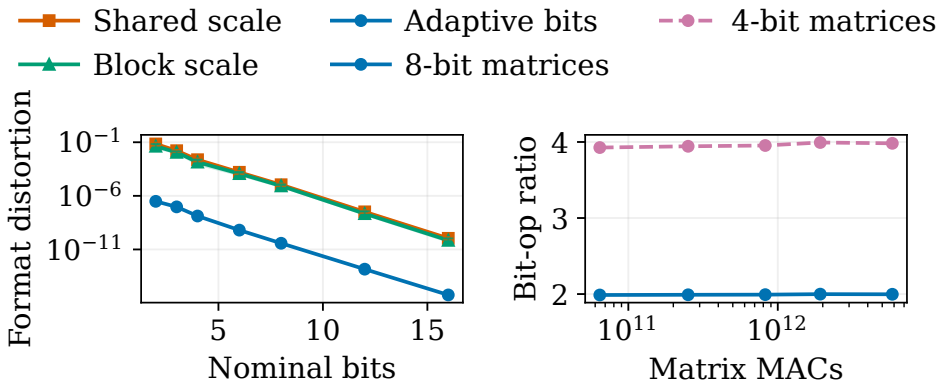


Figure 8: **Stored precision and retained high-precision work.** Left: format distortion versus nominal bits for shared-scale, block-scale, and adaptive-bit representations of the causal source. Right: analytic bit-operation ratios for 8-bit and 4-bit matrix paths with FP32 softmax retained. The source-format curves use paired 95% BCa intervals; the right-hand curves evaluate the analytic operation-count formula.

Predicting unseen data and storage budgets. With 4/8-bit and BF16 calibration, the additive model lowers prediction error at the largest data budget for every scale (Table 5). For 6-bit interpolation, it raises error at 70M/160M and lowers it at the four larger scales (Table 6). Each paired 95%

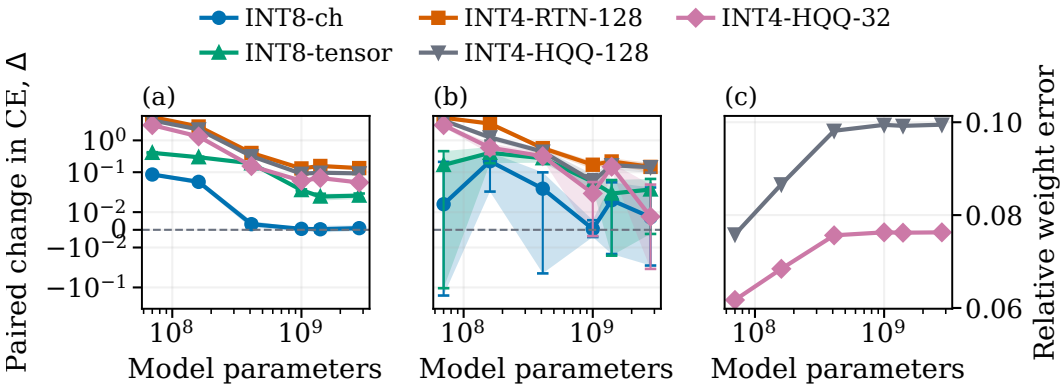


Figure 9: **Finite precision across six Pythia scales.** Left and center: paired cross-entropy changes on WikiText-103 (96 paired windows) and LAMBADA (512 paired examples), respectively. Right: parameter-weighted relative Frobenius error for the two HQQ formats, averaging $\|\widehat{W} - W\|_F / \|W\|_F$ over layers with weights proportional to their parameter counts. Blue circles, green triangles, orange squares, gray down-triangles, and magenta diamonds denote INT8-channel, INT8-tensor, INT4-RTN-g128, INT4-HQQ-g128, and INT4-HQQ-g32, respectively.

Table 4: **Reference losses and spectral slopes.** Left: Pythia BF16 CE over 96 WikiText-103 windows or 512 LAMBADA items; bold/underlined marks the lowest/second-lowest value per dataset. Right: slope against the spectral predictor (reference one), from eight unbinned seed fits. Each ± is the larger deviation to the 95% BCa endpoints.

(a) Pythia BF16 baseline; lower is better.

Model	WikiText	LAMBADA
70M	6.201 ±0.047	6.781 ±0.323
160M	4.388 ±0.056	4.965 ±0.339
410M	2.971 ±0.059	2.283 ±0.216
1B	2.580 ±0.059	1.504 ±0.179
1.4B	<u>2.548</u> ±0.061	<u>1.470</u> ±0.186
2.8B	**2.390** ±0.060	**1.346** ±0.170

(b) Construction and trained-learner slopes.

Profile	Construct.	Trained
$(0.5, 0)$	0.971 ±0.040	0.955 ±0.033
$(0.5, 0.5)$	0.933 ±0.017	0.855 ±0.028
$(1, 0)$	0.969 ±0.029	0.859 ±0.018
$(1, 0.5)$	0.952 ±0.023	0.879 ±0.026

BCa interval for these additive-minus-data differences excludes zero. An interaction further improves extrapolation at 160M, 410M and 1.4B; the corresponding intervals at the other three scales include zero. The benefit of modeling both resources therefore depends on scale and on which resource is extrapolated. Tables 7 and 8 give the predictions when calibration also includes 2-bit weights.

R Experiment and Reproducibility Details

We retain the scientific configuration and seed- or item-level measurements for every reported result. This appendix specifies the sampling scheme, numerical formats, software environment, and aggregation procedure; the accompanying code instantiates the complete grids.

Software and execution framework. The local central-processing-unit (CPU) runs used Python 3.11.15 with NumPy 2.0.2, SciPy 1.13.1, Pandas 2.3.3, PyArrow 21.0.0, and Matplotlib 3.9.4. The graphics-processing-unit (GPU) runs used Python 3.12.3, PyTorch 2.13.0 with CUDA 13.0, TorchAO 0.18.0, Transformers 5.16.1, Datasets 5.0.1, and Hugging Face Hub 1.29.0. The `finitebit_exp` 0.1.0 package defines the scientific grids in Python. PyTorch automatic differentiation supplies the Stage A and Stage V gradients in the mini-batch implementation. The weighted-average FP64 reference implements the same explicit scalar gradient on its theorem-specified active set. Independent

Table 5: **Predicting the largest data budget from 4/8-bit and BF16 calibration.** Root-mean-square error (RMSE, nats; lower is better) over 4/6/8-bit and BF16 test CE at 2^{26} additional tokens. Each response function in Section S is fitted on validation CE at the first three data budgets. Eight paired training seeds and 96 fixed test windows are used; 5,000 seed-level BCa resamples refit each function. Smaller $\pm$ values are rounded-up symmetric envelopes of the 95% intervals. Bold/underline mark the lowest/second-lowest displayed RMSE per row.

Model	Data only	Storage only	Additive	Interaction
70M	2.060 ± 0.061	1.293 ± 0.023	$\underline{1.151} \pm 0.080$	$\mathbf{1.149} \pm 0.087$
160M	0.805 ± 0.016	0.619 ± 0.034	$\underline{0.425} \pm 0.037$	$\mathbf{0.392} \pm 0.036$
410M	0.086 ± 0.004	0.156 ± 0.005	$\underline{0.058} \pm 0.006$	$\mathbf{0.054} \pm 0.006$
1B	$\underline{0.061} \pm 0.003$	0.096 ± 0.002	$\mathbf{0.041} \pm 0.003$	$\mathbf{0.041} \pm 0.003$
1.4B	$\underline{0.057} \pm 0.003$	0.085 ± 0.001	$\mathbf{0.033} \pm 0.004$	$\mathbf{0.033} \pm 0.004$
2.8B	$\underline{0.055} \pm 0.003$	0.074 ± 0.002	$\mathbf{0.034} \pm 0.004$	$\mathbf{0.034} \pm 0.004$

Table 6: **Predicting 6-bit loss from 4/8-bit and BF16 calibration.** RMSE (nats; lower is better) over the first three data budgets, holding out 6-bit weights from calibration. Response functions, eight paired training seeds, 96 fixed test windows, refitted 95% BCa envelopes, and ranking conventions match Table 5.

Model	Data only	Storage only	Additive	Interaction
70M	$\mathbf{1.304} \pm 0.015$	1.845 ± 0.021	$\underline{1.842} \pm 0.023$	$\underline{1.842} \pm 0.023$
160M	$\mathbf{0.649} \pm 0.016$	0.732 ± 0.017	$\underline{0.731} \pm 0.017$	$\underline{0.731} \pm 0.017$
410M	$\underline{0.093} \pm 0.003$	0.094 ± 0.003	$\mathbf{0.081} \pm 0.003$	$\mathbf{0.081} \pm 0.003$
1B	$\underline{0.054} \pm 0.002$	0.055 ± 0.002	$\mathbf{0.046} \pm 0.002$	$\mathbf{0.046} \pm 0.002$
1.4B	0.053 ± 0.002	$\underline{0.047} \pm 0.001$	$\mathbf{0.041} \pm 0.002$	$\mathbf{0.041} \pm 0.002$
2.8B	0.051 ± 0.002	$\underline{0.043} \pm 0.002$	$\mathbf{0.039} \pm 0.002$	$\mathbf{0.039} \pm 0.002$

Table 7: **Predicting the largest data budget with 2-bit weights included in calibration.** RMSE (nats; lower is better) over all six deployment formats at 2^{26} additional tokens, fitted using 2/4/8-bit and BF16 validation CE at the first three budgets. This bounded-response comparison covers 410M–2.8B. The 70M/160M 2-bit calibration losses exceed its $[0, 20]$-nat range; Figure 7 shows their measured quantization losses. Eight paired seeds, 96 test windows, 95% BCa envelopes and ranking conventions match Table 5.

Model	Data only	Storage only	Additive	Interaction
410M	3.776 ± 0.044	1.354 ± 0.008	$\underline{1.328} \pm 0.030$	$\mathbf{1.327} \pm 0.036$
1B	3.426 ± 0.032	1.088 ± 0.010	$\underline{1.050} \pm 0.019$	$\mathbf{1.045} \pm 0.022$
1.4B	3.724 ± 0.031	1.094 ± 0.011	$\underline{1.063} \pm 0.020$	$\mathbf{1.060} \pm 0.028$
2.8B	3.255 ± 0.024	0.797 ± 0.005	$\underline{0.772} \pm 0.006$	$\mathbf{0.771} \pm 0.007$

Table 8: **Predicting 3/6-bit loss with 2-bit weights included in calibration.** RMSE (nats; lower is better) at the first three data budgets, holding out 3/6-bit weights from 2/4/8-bit/BF16 calibration. This comparison covers 410M–2.8B, with the calibration range and scale coverage described in Table 7. Eight paired training seeds, 96 test windows, 95% BCa envelopes and ranking conventions match Table 5.

Model	Data only	Storage only	Additive	Interaction
410M	1.976 ± 0.016	1.573 ± 0.032	$\mathbf{1.568} \pm 0.033$	$\underline{1.569} \pm 0.033$
1B	1.870 ± 0.008	$\underline{1.512} \pm 0.014$	$\mathbf{1.511} \pm 0.014$	$\mathbf{1.511} \pm 0.014$
1.4B	2.152 ± 0.014	$\underline{1.585} \pm 0.012$	$\mathbf{1.584} \pm 0.012$	$\mathbf{1.584} \pm 0.012$
2.8B	1.874 ± 0.003	$\underline{1.145} \pm 0.003$	$\mathbf{1.144} \pm 0.003$	$\mathbf{1.144} \pm 0.003$

worker processes run the two-stage training jobs. Pythia models are loaded once per model–format bundle and evaluated in inference mode. CPU experiments ran on an Apple M5 system; GPU experiments used an NVIDIA RTX PRO 6000 Blackwell Server Edition. Raw seed- and item-level measurements are stored in Parquet files.

Pythia items and operation count. Tokenization is performed once per model tokenizer on the raw WikiText-103 test text, after which contiguous 2,049-token windows provide 2,048 next-token losses per item. All numerical formats receive identical item indices. The analytic operation count includes the query, key, value, and output projections; the two attention matrix products; and both feed-forward products.

Reproducible result generation. Empirical curves are generated from seed- or item-level records using the uncertainty procedure specified in each caption.

The data–memory study crosses six energy–dimension profiles, seven model cutoffs from 16 to 1,024, and three normalized data and state budgets. It evaluates population excess CE over each finite source, including coordinates predicted with zero logit. The mini-batch two-stage study additionally uses the analytic tail proxy described in Section S; the fixed-source weighted-SGD study measures its omitted-coordinate loss directly. The pretrained-model comparisons use identical items for every format, including 96 paired WikiText-103 windows per model–format configuration.

R.1 Additional coupling, robustness, and pretrained-model experiments

The additional experiments comprise a dyadic coupling test with 16 paired seeds, a two-stage study with eight paired seeds, robustness tests for three departures from exact power laws, and six Pythia sizes under six deployment formats on two datasets. The two-stage learner differentiates the Bernoulli loss through the attention-weighted network output and updates the stored value parameters; the direct finite-sample parameter construction is included as a separate likelihood-based comparator.

Six-scale weight-only post-training quantization. All Pythia models use the fixed `step143000` revision. WikiText-103 uses 96 common non-overlapping 2,048-token windows, and LAMBADA uses 512 common examples with last-token cross-entropy. The formats are bfloat16 (BF16), 8-bit integer (INT8) with one tensor scale, INT8 with one scale per output channel, 4-bit integer (INT4) round-to-nearest with groups of 128 weights, and half-quadratic quantization (HQQ) for INT4 groups of 128 or 32 weights. Every reported difference is paired with the BF16 result on the same item. Eager attention receives BF16 scores and returns FP32 softmax probabilities in the measured software stack. The evaluation microbatch is one for both datasets.

The item-level records retain model, format, dataset, item index, cross-entropy, represented weight error, quantized-parameter fraction, and output-logit divergence. Appendix Q reports the detailed figures and tables.

S Joint Resource Evaluation

Fixed-source comparisons. We vary data and value-state budgets while holding the generating distribution fixed. The same-marginal experiment samples the infinite dyadic source at $\chi \in \{1, 2\}$, $\gamma = 1$, with a common normalization for both pairings. A sparse clipped Bernoulli estimator uses unbounded storage, isolating the data exponent. Unobserved coordinates contribute their actual probability mass times the zero-prediction excess cross-entropy. The seven sample budgets are $2^8, 2^{10}, \ldots, 2^{20}$; the primary slope uses the last four budgets for every seed and pairing.

The theorem's learning algorithm. For $(\chi, \gamma) = (0.5, 0), (0.5, 0.5), (1, 0), (1, 0.5)$, the source contains 256 energy groups, each with $d_j = \lceil j^\gamma \rceil$ scalar coordinates. This source size is distinct from the number of sampled regeneration blocks. We compare $K = 8, 32$ candidates with eight independent seeds. At each $n \in \{2^8, 2^{10}, \ldots, 2^{16}\}$, we restart the projected, preconditioned SGD algorithm of Theorem D.5 from zero, using its active set, step sizes and weighted average. We implement the numerical reference in FP64. Calibration uses $S_A = \lceil 4K\lceil \log(eK) + \sqrt{n} \rceil \rceil$ independent blocks. Each fitted state is encoded at $B_V = 8, 16, 32, 64, 128, 256$ bits using the theorem's midpoint scalar

format. Inactive fitted coordinates are zero before encoding; positive-width midpoint decoding also acts on those zero entries. An oracle-coordinate clipped estimator uses the same labeled blocks with the selected coordinate revealed. We report both value and routing bits, calibration and prediction samples, and the losses on 65,536 shared independent test contexts. Oracle and learned routes, before and after quantization, define signed counterfactual differences on these identical contexts.

Additional training and deployed storage. Starting from the Pythia 70M, 160M, 410M, 1B, 1.4B, and 2.8B checkpoints at step143000, each of eight seeds follows one continued-training trajectory with checkpoints after $2^{20}, 2^{22}, 2^{24}, 2^{26}$ additional WikiText-103 tokens. Training uses 2048-token contexts, an effective batch of 32,768 tokens, FP32 parameters and optimizer states, and BF16 computation. AdamW uses learning rate 10^{-5}, $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping at 1, and an eight-step warm-up followed by a constant learning rate. Deployment uses BF16 or group-128 symmetric 2/3/4/6/8-bit round-to-nearest linear weights. Integer payloads, scales, padding, retained BF16 parameters and metadata are all counted. Evaluation uses 96 fixed test windows and 96 fixed validation windows.

Prediction on held-out resource configurations. We fit data-only, storage-only, additive and interacting response models on the first three training budgets at 2/4/8-bit and BF16. The 3/6-bit formats test storage interpolation, while the largest training budget tests data extrapolation. Fits use validation losses; test-window losses remain untouched until evaluation. Comparisons use paired training-seed uncertainty; the common initial pretrained checkpoint is evaluated once and its uncertainty uses windows. Let $t = T/2^{20}$ denote additional tokens in units of the smallest budget, and let s be deployment-package bytes divided by the BF16 package size. Writing $x = t^{-\alpha}$ and $y = s^{-\beta} - 1$, the four fitted CE models are

$$\underbrace{c + ax}_{\text{data only}}, \qquad \underbrace{c + dy}_{\text{storage only}}, \qquad \underbrace{c + ax + dy}_{\text{additive}}, \qquad \underbrace{c + ax + dy + hxy}_{\text{interaction}}.$$

We fit each model scale separately by least squares on validation CE, with $c, a, d \in [0, 20]$, $\alpha, \beta \in [0.01, 4]$, and $h \in [-20, 20]$. The fitting procedure requires convergence and calibration predictions in the pre-specified range $[0, 20]$ nats; held-out predictions are checked against the same range. After observing the large 2-bit losses at 70M and 160M, we also analyze 4/8-bit and BF16 calibration, with 6-bit interpolation and 4/6/8-bit/BF16 extrapolation. This sensitivity analysis uses the same response functions and parameter bounds. Root-mean-square error (RMSE) compares predicted and measured test CE across the held-out resource configurations, after averaging the 96 windows within each configuration.

Route-adjusted and complete-source losses. Let $L_{\mathrm{lr},q}, L_{\mathrm{lr},u}, L_{\mathrm{or},u}$ denote finite-source excess CE for learned-route quantized, learned-route unquantized, and known-route unquantized predictions. All subtract the same source Bayes loss. The route-adjusted analysis defines $E = L_{\mathrm{lr},q} + M^{-\chi}$, $R_+ = (L_{\mathrm{lr},u} - L_{\mathrm{or},u})_+$, and $F_+ = (L_{\mathrm{lr},q} - L_{\mathrm{lr},u})_+$. Its spectrum comparison uses $E - R_+$; its component sum is an upper envelope of E. Here M is the energy-group cutoff and $M^{-\chi}$ is an analytic tail proxy. Our fixed-source study instead evaluates the generating distribution without adding that proxy and retains all four counterfactual losses. The oracle loss splits over actual held-out targets inside and outside A_n; adding the signed differences $L_{\mathrm{lr},u} - L_{\mathrm{or},u}$ and $L_{\mathrm{lr},q} - L_{\mathrm{lr},u}$ gives exactly the complete-source excess loss $L_{\mathrm{lr},q}$. The tail term is the actual zero-prediction KL on omitted targets. This identity holds per test context, before any averaging, by Proposition T.2.

Empirical and known-energy allocation. The robustness study's adaptive allocator uses the count-corrected score

$$s_i = \left[p_i \widehat{\theta}_i^2 - \frac{p_i}{\max\{1, N_i\}} \right]_+,$$

where N_i is the observed count and $\widehat{\theta}_i$ the bounded Bernoulli maximum-likelihood estimate. Both this rule and the known-energy reference water-fill $\sum_i s_i 4^{-b_i}$, floor widths, and assign remaining bits by marginal gains; the reference substitutes $s_i = q_i$. The empirical subtraction prioritizes resolved coordinates, whereas the known-energy reference minimizes the true-energy distortion objective. These distinct objectives allow different finite-sample CE even at the same state budget. The unknown-exponent schedule in Theorem 5.1 is a separate, deterministic construction.

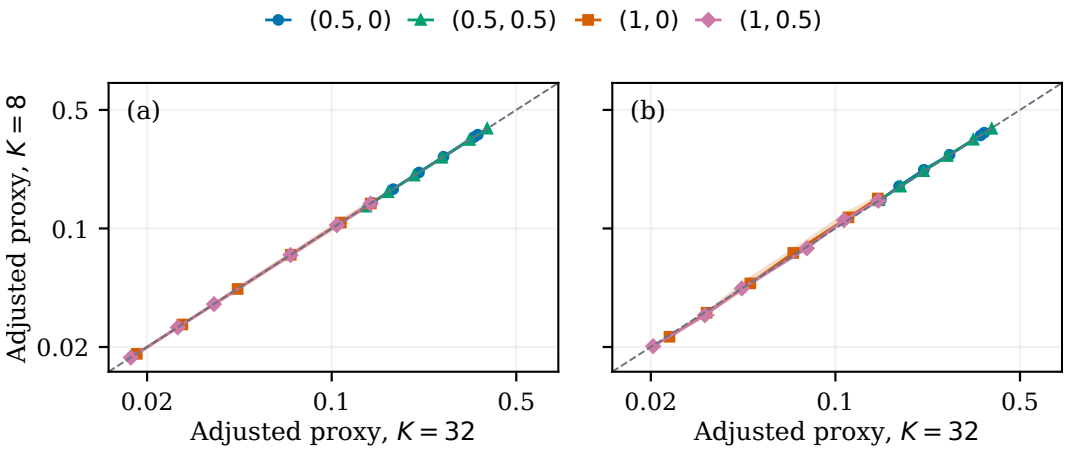


Figure 10: **Candidate-count replication under route-adjusted loss.** (a) Direct finite-sample construction; (b) mini-batch two-stage implementation. Horizontal and vertical axes show $E - R_+$ at $K = 32$ and $K = 8$, respectively, using identical spectral bins. Legends give (χ, γ); the dashed identity line is deterministic. Intervals on both axes are 95% BCa intervals over eight paired seeds.

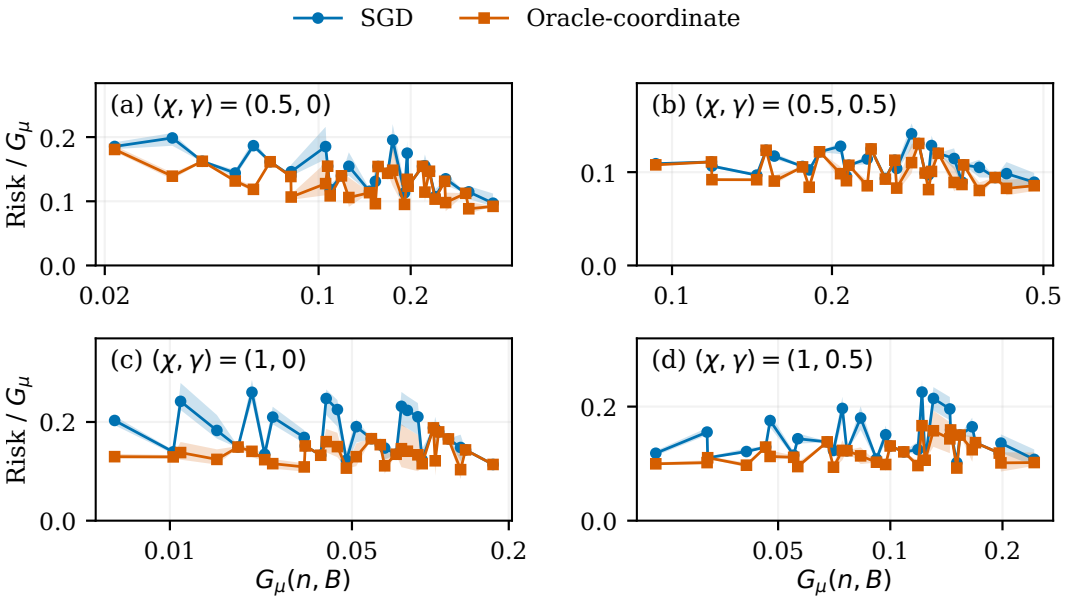


Figure 11: **Unbinned risk-to-spectrum ratios on fixed sources.** Panels (a)–(d) use the four profiles of Figure 14, at $K = 32$. The horizontal axis is G_μ; the vertical axis is measured excess CE divided by G_μ. Each point represents a distinct sample–state budget pair; lines connect budget pairs ordered by G_μ, rather than a fitted response. Blue/orange denote weighted SGD and the oracle-coordinate estimator. Bands retain the raw 95% BCa intervals over eight paired seeds.

T Total Resources and Fixed-Source Evaluation

The following consequence accounts jointly for the value-learning and routing stages. Write $q_{(r)}$ for the decreasing coordinate energies and ξ for their decay exponent.

Proposition T.1 (Total sample and learned-state requirements)**.** *Fix an integer $K \geq 1$, $R > 0$, and $0 < \eta \leq 1$. Use the iid-candidate source, public frequencies and calibration supervision of Theorem D.5, with unit midpoint formats. Suppose $cr^{-(1+\xi)} \leq q_{(r)} \leq Cr^{-(1+\xi)}$, where $\xi > 0$ and $0 < c \leq C < \infty$ are fixed. Let $C_L \geq 1$ be a constant in its four-term risk upper bound. Set*

$$C_D = 2 + C/\xi, \quad a = \max\{1, (1+\xi)/(2\ln 2)\}, \quad C_B = C(1 + 1/\xi)(2a)^\xi.$$

Let $C_A(\eta)$ be an upper-bound constant in the routing hitting-time bound of Proposition D.4 for thresholds at most $1/16$. For $0 < \varepsilon \le 1$, choose

$$h = \min\{1/16, \varepsilon/(4C_L)\},$$
$$n = \max\{1, \lceil (4C_L C_D/\varepsilon)^{(1+\xi)/\xi} \rceil\},$$
$$B_V = \max\{1, \lceil (4C_L C_B/\varepsilon)^{1/\xi} \rceil\}.$$

For $K \ge 2$, take

$$S_A = \lceil C_A(\eta) K\{h^{-1/2} + \ln(eK)\} \rceil,$$
$$B_A = K\{\lceil \log_2(S_A + 1) \rceil + \lceil \log_2(K + 1) \rceil\};$$

for $K = 1$, take $S_A = B_A = 0$. If deployment arithmetic has mean squared logit error at most $\varepsilon/(4C_L)$, the weighted-SGD learner has expected excess risk at most ε, uses $N = n + S_A$ blocks and retains at most $B = B_V + B_A$ learned bits. At fixed source and routing constants,

$$N = O(\varepsilon^{-(1+\xi)/\xi}), \quad B = O(\varepsilon^{-1/\xi}), \qquad S_A/n \to 0, \quad B_A/B_V \to 0.$$

For $\chi > 0$, $\gamma \ge 0$, block energies $a_j \asymp j^{-(1+\chi)}$, dimensions $d_j \asymp j^{\gamma}$ and equal coordinate energies $q_{j,\ell} = a_j/d_j$ within each block, the coordinate exponent is $\xi = \chi/(1+\gamma)$.

Proof. For any integer $m \ge 1$, the energy envelope and integral comparison give

$$\Phi_\mu(n^{-1}) \le m/n + C \sum_{r>m} r^{-1-\xi} \le m/n + (C/\xi) m^{-\xi}.$$

Choose $m = \lceil n^{1/(1+\xi)} \rceil$. Since $n^{1/(1+\xi)} \le m \le 2n^{1/(1+\xi)}$, this is at most $C_D n^{-\xi/(1+\xi)}$, including $n = 1$.

For a storage budget $b \ge 2a$, put $m = \lfloor b/a \rfloor$ and allocate $b_r = (1+\xi)\log_2(m/r)/2$ for $1 \le r \le m$, with zero widths thereafter. The inequality

$$\sum_{r=1}^{m} \ln(m/r) \le \int_0^m \ln(m/x)\, dx = m$$

shows that total width is at most $(1+\xi)m/(2\ln 2) \le am \le b$. Each retained distortion term is at most $Cm^{-1-\xi}$, and the tail is at most $(C/\xi)m^{-\xi}$. Since $m \ge b/(2a)$,

$$D_q(b) \le C(1 + 1/\xi) m^{-\xi} \le C_B b^{-\xi}.$$

For $1 \le b < 2a$, zero allocation has distortion at most $\sum_r q_{(r)} \le C(1 + 1/\xi) \le C_B b^{-\xi}$, proving the same bound for all $b \ge 1$. Summability controls every countable tail. The theorem's scalar implementation floors public continuous widths; this cannot exceed the bit budget and multiplies each distortion term by at most four. Its constant is already included in C_L. A public allocation needs no additional learned index code.

The ceilings in n, B_V can only decrease their nonincreasing error functionals, so each is at most $\varepsilon/(4C_L)$. For $K \ge 2$, the positive threshold $h \le 1/16$ lies in the routing theorem's range. The prescribed S_A exceeds its hitting-time bound. Margin increases with row visits, and counts couple monotonically as calibration blocks are added; leakage therefore remains below $h \le \varepsilon/(4C_L)$. For $K = 1$, leakage is zero. The arithmetic premise bounds the fourth term. Substitution in Equation (D.5) now gives risk at most $C_L \cdot 4\varepsilon/(4C_L) = \varepsilon$.

The independent stages consume distinct blocks, totaling $n + S_A$. Each routing row is reconstructed from its count in $\{0, \dots, S_A\}$ and its destination or unseen symbol. The displayed integer-width fields encode all these possibilities. Concatenating their B_A-bit code with the value code gives at most $2^{B_A + B_V}$ retained states. Numerical reconstruction is charged to the arithmetic term.

For sufficiently small ε, $h = \varepsilon/(4C_L)$. The explicit resource choices then yield

$$S_A/n = O_K(\varepsilon^{1/2+1/\xi} + \varepsilon^{1+1/\xi}),$$
$$B_A/B_V = O_K(\varepsilon^{1/\xi} \ln(1/\varepsilon)).$$

Both tend to zero since $\xi > 0$. Adding the routing costs gives the two total-resource orders.

Finally, every coordinate in block j has energy comparable to $j^{-(1+\chi+\gamma)}$, uniformly over its $d_j \asymp j^\gamma$ coordinates. Upper and lower threshold comparisons therefore give

$$\#\{q_i \geq t\} \asymp \sum_{j \leq c_0 t^{-1/(1+\chi+\gamma)}} j^\gamma \asymp t^{-(1+\gamma)/(1+\chi+\gamma)},$$

with separate positive c_0 for the upper and lower comparisons. Inversion proves $q_{(r)} \asymp r^{-(1+\chi+\gamma)/(1+\gamma)}$ and the asserted ξ. The argument allows bounded nonmonotonicity in the original block labels. □

Proposition T.2 (Fixed-source loss and paired counterfactuals)**.** *Let p be a fixed probability distribution on a finite or countable coordinate set, $|\theta_i| \leq \Delta_i \leq R$, $q_i = p_i \Delta_i^2$, and $d(\theta, z) = D_{\mathrm{KL}}(\mathrm{Ber}(\sigma(\theta)) \| \mathrm{Ber}(\sigma(z)))$. Conditional on a fitted table $|w_i| \leq R$ supported on S, a fresh oracle-routed example has excess risk*

$$L(w, \mathrm{oracle}) = \sum_{i \in S} p_i d(\theta_i, w_i) + \sum_{i \notin S} p_i d(\theta_i, 0),$$
$$0 \leq \sum_{i \notin S} p_i d(\theta_i, 0) \leq \tfrac{1}{8} \sum_{i \notin S} q_i.$$

If all $|\theta_i| = a$, the omitted term equals $p(S^c) d(a, 0)$, also for a data-dependent S.

On the same full candidate context, let L_{rv} be excess loss with route $r \in \{0, 1\}$ (oracle, learned) and value state $v \in \{0, 1\}$ (unquantized, quantized), with all four predicted logits in $[-R, R]$. Pointwise and under every common empirical or population average,

$$\begin{aligned} L_{11} &= L_{00} + (L_{10} - L_{00}) + (L_{11} - L_{10}) \\ &= L_{00} + (L_{01} - L_{00}) + (L_{11} - L_{01}). \end{aligned}$$

The two format contrasts differ by $L_{11} - L_{10} - L_{01} + L_{00}$; individual contrasts may be negative. If two predictions in $[-R, R]$ agree outside an event A, their expected CE differs by at most $2R \Pr(A)$. If A is that some candidate lies in T, then $\Pr(A) \leq K p(T)$.

Proof. For target coordinate i, conditional expected CE minus Bayes CE is exactly $d(\theta_i, z)$. Conditional on training, a fresh target retains law p, giving $\sum_i p_i d(\theta_i, w_i)$. Nonnegative terms permit countable summation by monotone convergence. Splitting at S, where the fitted table is zero on its complement, proves the oracle identity.

Put $A(z) = \ln(1 + e^z)$. Direct substitution gives $d(\theta, z) = A(z) - A(\theta) - A'(\theta)(z - \theta)$. Since $0 < A''(z) = \sigma(z)(1 - \sigma(z)) \leq 1/4$, Taylor's integral remainder implies $0 \leq d(\theta, 0) \leq \theta^2/8$. Multiply by p_i, use $\theta_i^2 \leq \Delta_i^2$, and sum to obtain the tail bound. Exchanging Bernoulli outcomes gives $d(-a, 0) = d(a, 0)$, proving the constant-magnitude formula conditional on each fitted support.

For learned routing, $\sum_s a_{Qs} w_{C_s}$ can be nonzero even when the target is outside S, because a distractor may lie in S. The four losses therefore evaluate the full context with a common target law. Cancelling intermediate terms proves both identities pointwise; linearity preserves them under averaging. Subtracting the two format contrasts gives their stated interaction. For a concrete sign example, the one-bit midpoint format on $[-1, 1]$ sends $w = 0.1$ to 0.5. At $\theta = 0.5$, the contrast is $-d(0.5, 0.1) < 0$; at $\theta = 0.1$, it is $d(0.1, 0.5) > 0$.

Finally, per-label CE $\ell_y(z) = A(z) - yz$ has derivative $\sigma(z) - y$ of absolute value at most one. Hence $|\ell_y(z) - \ell_y(z')| \leq |z - z'| \leq 2R$. Multiplying by the indicator of A and averaging proves the approximation bound, also after subtracting the common Bayes risk. The union bound gives $\Pr(A) \leq \sum_{s=1}^K \Pr(C_s \in T) = K p(T)$; it needs only the candidate marginals, not independence. □

U Predicting Held-Out Resource Configurations

We test whether the coordinate-energy spectrum predicts how a trained predictor responds to additional data or retained state. Six fixed sources contain 256 energy groups each, with 32 candidate positions and logit radius 0.8. Four use the power profiles of the main experiments; two

modify the $(\chi, \gamma) = (1, 0.5)$ profile with a broken-power or logarithmic correction. The source distribution remains fixed across all sample and storage budgets. For each source we use 16 independent training seeds, five sample budgets $n = 2^8, 2^{10}, \ldots, 2^{16}$, and six value-state budgets $B = 8, 16, 32, 64, 128, 256$. Weighted-average projected SGD starts at zero separately for each n and uses the active set and step sizes of the learning theorem. A direct clipped estimator uses the same observations with their relevant training coordinates revealed; both fitted vectors deploy learned routing. Integer midpoint quantization is applied to the fitted vector for each budget.

Independent validation and test sets each contain 65,536 contexts per seed, shared across budgets and methods. We fit response coefficients on validation losses at the first four sample budgets and first five storage budgets. The largest sample budget, largest storage budget, and their intersection define separate held-out tests. We compare the full spectrum, a rank-power approximation with the same total energy and coordinate count, and free one- and two-resource power laws. Prediction error is measured on test loss and normalized by the source's zero-predictor population excess CE. Resource decisions compare retaining the current resources with increasing either resource fourfold. Gains retain their sign; regret is measured relative to the best of these three observed actions. All uncertainty uses paired seed-level BCa intervals with 5,000 resamples.

Source geometry and reproducible observations. Writing $j = 1, \ldots, 256$ for an energy group, set $d_j = \lceil j^\gamma \rceil$, $R_i = 0.8$, and raw group weights $a_j = j^{-(1+\chi)}$. Broken-power weights multiply a_j by $(j/65)^{-0.75}$ for $j \geq 65$; log-corrected weights multiply it by $\log_2(j+2)^{-1.25}$. Group probabilities are normalized weights, distributed equally among their coordinates. The logit sign alternates within each group. Query address q always selects candidate position q. Both this mapping and the coordinate distribution are independent of seed and budget. The four power profiles, denoted P1–P4 in the figures, are $(\chi, \gamma) = (0.5, 0), (0.5, 0.5), (1, 0), (1, 0.5)$; Br and Lg use the two modified $(1, 0.5)$ profiles. Calibration uses $S_A = \lceil 4K(\log(eK) + \sqrt{n}) \rceil$ blocks and routing step size 0.5. Seed-level records retain n, S_A, B_V, B_A, actually used integer bits, fitted vectors, and all four counterfactual losses. Train, validation, and test random streams are distinct; the latter two are shared across the resource grid. Training seeds are 1000–1015.

Response functions and held-out comparisons. For the full spectrum, the response is $a\Phi_\mu(n^{-1}) + b\Phi_\mu(\tau_B)$, $a, b \geq 0$. For the rank-power comparison, least squares regresses log sorted energy on log rank over every coordinate. The resulting power profile is rescaled to the original total energy and uses the same two-coefficient response. The free joint response is

$$c + a(n/256)^{-\alpha} + b(B/8)^{-\beta}, \qquad c, a, b \geq 0, \quad \alpha, \beta \in [0.01, 4].$$

Data-only and memory-only responses omit the other term. Nonlinear fits use six fixed exponent starts for the joint response and four for each one-axis response, selecting the converged solution with smallest validation error. All 960 fits converge. The 20 calibration configurations exclude the largest value of both resources. Test errors average separately over five data extrapolation points, four memory extrapolation points, and their one-point intersection. Each seed is fitted and evaluated before statistical aggregation.

The free joint response gives smaller mean single-axis prediction error than the two-coefficient spectrum response for weighted SGD on every source. At the joint held-out corner, the spectrum's normalized RMSE ranges from 0.0025 to 0.0204, versus 0.0321 to 0.0960 for the free joint power law. The complete and rank-power spectra have similar single-axis errors; at the broken-power joint corner their errors are 0.0028 and 0.0115. The full curve preserves the bend that a single exponent smooths away. These tests resolve two useful properties: joint-resource extrapolation and selection of the resource to increase. They measure a fitted learner response; the minimax theorem supplies the spectral shape up to constants.

Choosing between equal-factor expansions. At each of 16 starting budget pairs, the available actions are to stay, quadruple samples, or quadruple retained value-state bits. The uncalibrated spectrum selects the lowest G_μ, resolving ties in favor of staying. Its calibrated version substitutes the validation-fitted coefficients. The signed gain is current minus selected test loss. Regret subtracts the best measured loss among all three actions, including staying. This definition retains loss increases and allows quantization to improve a realized fit. We first average the 16 decisions within each seed; seed-level paired differences then determine the intervals. For weighted SGD, mean uncalibrated

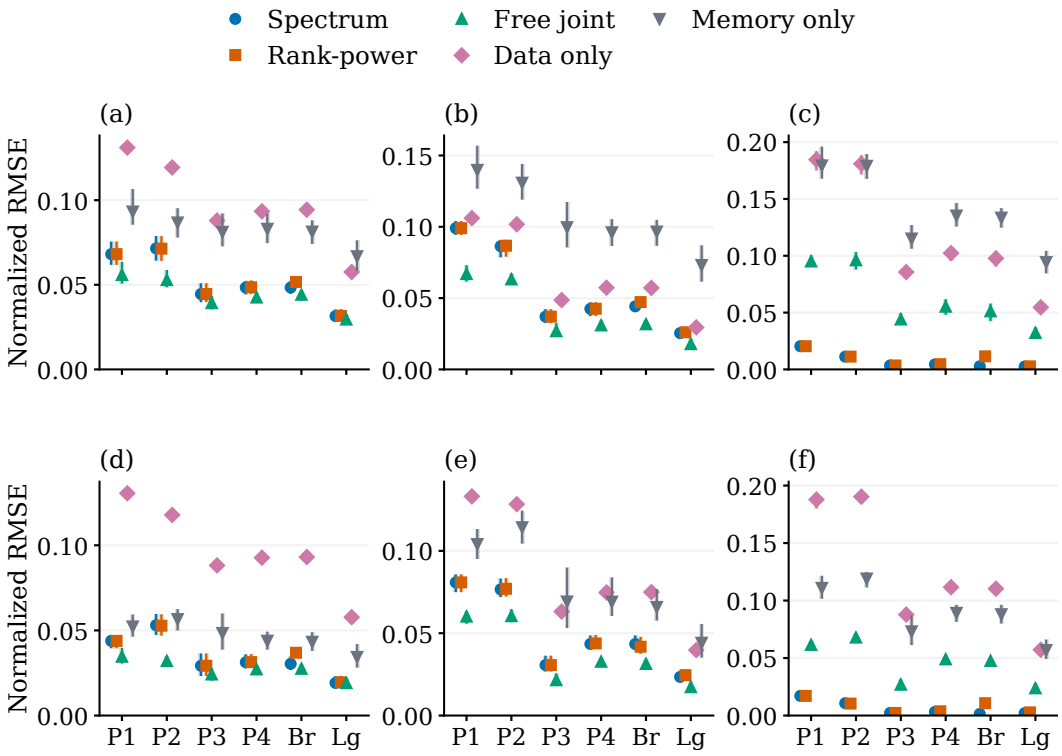


Figure 12: **All response families and both learners.** Top: weighted-average SGD; bottom: direct clipped estimation. (a,d) Data extrapolation; (b,e) memory extrapolation; (c,f) simultaneous extrapolation. Each point is mean seed-level RMSE divided by zero-predictor population excess CE. P1–P4, Br, Lg denote the six fixed sources defined above. Intervals are 95% BCa over 16 seeds. Both empirical losses and predictions use the full source and the same test contexts.

regret ranges from 1.11 to 1.45 in units of 10^{-3} nats, below both fixed-action policies on every source. Calibration further reduces mean regret for this learner; the direct estimator's comparison is reported separately rather than pooled with SGD. Tables 11 and 12 give both signed gains and regrets for all four policies and both learners.

Table 9: **Held-out prediction for weighted-average SGD.** RMSE in 10^{-3} nats (lower is better). D/M/J denote data, memory, and joint extrapolation. Each value averages 16 independently fitted seed-level RMSEs; smaller $\pm$ is a rounded-up symmetric envelope of the raw 95% BCa interval. Best/second displayed values are bold/underlined. Sources are reported separately because their risk surfaces differ.

Source / test	Spectrum	Rank-power	Free joint	Data only	Memory only
P1/D	$\underline{5.04} \pm 0.55$	$\underline{5.04} \pm 0.55$	$\mathbf{4.13} \pm 0.58$	9.70 ± 0.14	6.91 ± 0.99
P1/M	$\underline{7.33} \pm 0.38$	$\underline{7.33} \pm 0.38$	$\mathbf{4.95} \pm 0.46$	7.85 ± 0.23	10.36 ± 1.26
P1/J	$\mathbf{1.51} \pm 0.10$	$\mathbf{1.51} \pm 0.10$	$\underline{7.03} \pm 0.40$	13.67 ± 0.71	13.29 ± 1.23
P2/D	5.29 ± 0.55	$\underline{5.28} \pm 0.55$	$\mathbf{3.90} \pm 0.45$	8.83 ± 0.18	6.43 ± 0.66
P2/M	$\underline{6.39} \pm 0.57$	6.42 ± 0.57	$\mathbf{4.67} \pm 0.31$	7.53 ± 0.23	9.69 ± 0.98
P2/J	$\mathbf{0.83} \pm 0.11$	$\mathbf{0.83} \pm 0.12$	$\underline{7.11} \pm 0.61$	13.39 ± 0.70	13.26 ± 0.85
P3/D	$\underline{3.30} \pm 0.47$	$\underline{3.30} \pm 0.47$	$\mathbf{2.90} \pm 0.31$	6.51 ± 0.03	6.02 ± 0.80
P3/M	$\underline{2.74} \pm 0.39$	$\underline{2.74} \pm 0.39$	$\mathbf{1.98} \pm 0.43$	3.60 ± 0.19	7.38 ± 1.31
P3/J	$\mathbf{0.26} \pm 0.07$	$\mathbf{0.26} \pm 0.07$	$\underline{3.25} \pm 0.44$	6.34 ± 0.34	8.53 ± 0.88
P4/D	$\underline{3.58} \pm 0.30$	3.59 ± 0.30	$\mathbf{3.14} \pm 0.24$	6.92 ± 0.07	6.15 ± 0.66
P4/M	$\underline{3.14} \pm 0.38$	3.15 ± 0.38	$\mathbf{2.28} \pm 0.27$	4.24 ± 0.14	7.12 ± 0.72
P4/J	$\mathbf{0.33} \pm 0.18$	$\underline{0.34} \pm 0.18$	4.08 ± 0.53	7.58 ± 0.51	10.03 ± 0.82
Br/D	$\underline{3.58} \pm 0.30$	3.83 ± 0.26	$\mathbf{3.26} \pm 0.25$	6.98 ± 0.13	6.03 ± 0.54
Br/M	$\underline{3.27} \pm 0.30$	3.49 ± 0.30	$\mathbf{2.34} \pm 0.25$	4.23 ± 0.13	7.16 ± 0.74
Br/J	$\mathbf{0.21} \pm 0.08$	$\underline{0.85} \pm 0.12$	3.78 ± 0.64	7.23 ± 0.54	9.89 ± 0.66
Lg/D	$\underline{2.33} \pm 0.30$	2.34 ± 0.30	$\mathbf{2.18} \pm 0.24$	4.26 ± 0.05	4.95 ± 0.70
Lg/M	$\underline{1.88} \pm 0.28$	1.93 ± 0.28	$\mathbf{1.31} \pm 0.29$	2.18 ± 0.14	5.42 ± 1.03
Lg/J	$\mathbf{0.19} \pm 0.07$	$\underline{0.20} \pm 0.09$	2.37 ± 0.35	4.05 ± 0.34	6.99 ± 0.74

Table 10: **Held-out prediction for direct clipped estimation.** RMSE in 10^{-3} nats (lower is better). D/M/J denote data, memory, and joint extrapolation. Each value averages 16 independently fitted seed-level RMSEs; smaller $\pm$ is a rounded-up symmetric envelope of the raw 95% BCa interval. Best/second displayed values are bold/underlined. Sources are reported separately because their risk surfaces differ.

Source / test	Spectrum	Rank-power	Free joint	Data only	Memory only
P1/D	$\underline{3.25} \pm 0.32$	$\underline{3.25} \pm 0.32$	$\mathbf{2.56} \pm 0.39$	9.66 ± 0.08	3.88 ± 0.51
P1/M	$\underline{5.99} \pm 0.45$	$\underline{5.99} \pm 0.45$	$\mathbf{4.44} \pm 0.32$	9.85 ± 0.25	7.71 ± 0.68
P1/J	$\mathbf{1.27} \pm 0.09$	$\mathbf{1.27} \pm 0.09$	$\underline{4.55} \pm 0.23$	13.90 ± 0.57	8.22 ± 0.78
P2/D	3.93 ± 0.49	$\underline{3.91} \pm 0.49$	$\mathbf{2.36} \pm 0.24$	8.72 ± 0.04	4.20 ± 0.50
P2/M	$\underline{5.67} \pm 0.49$	5.70 ± 0.48	$\mathbf{4.47} \pm 0.31$	9.50 ± 0.20	8.47 ± 0.75
P2/J	$\underline{0.80} \pm 0.11$	$\mathbf{0.77} \pm 0.10$	5.02 ± 0.37	14.09 ± 0.39	8.82 ± 0.56
P3/D	$\underline{2.17} \pm 0.53$	$\underline{2.17} \pm 0.53$	$\mathbf{1.78} \pm 0.29$	6.53 ± 0.03	3.59 ± 0.85
P3/M	$\underline{2.27} \pm 0.43$	$\underline{2.27} \pm 0.43$	$\mathbf{1.59} \pm 0.30$	4.68 ± 0.16	5.13 ± 1.52
P3/J	$\mathbf{0.17} \pm 0.08$	$\mathbf{0.17} \pm 0.08$	$\underline{1.98} \pm 0.39$	6.50 ± 0.20	5.42 ± 1.03
P4/D	$\underline{2.31} \pm 0.35$	2.32 ± 0.35	$\mathbf{2.00} \pm 0.18$	6.86 ± 0.06	3.25 ± 0.40
P4/M	$\underline{3.22} \pm 0.38$	3.24 ± 0.39	$\mathbf{2.42} \pm 0.21$	5.53 ± 0.17	5.12 ± 1.09
P4/J	$\mathbf{0.24} \pm 0.08$	$\underline{0.27} \pm 0.08$	3.61 ± 0.28	8.26 ± 0.21	6.61 ± 0.58
Br/D	$\underline{2.25} \pm 0.31$	2.73 ± 0.23	$\mathbf{2.03} \pm 0.22$	6.89 ± 0.03	3.21 ± 0.42
Br/M	3.22 ± 0.38	$\underline{3.09} \pm 0.45$	$\mathbf{2.32} \pm 0.27$	5.55 ± 0.19	4.88 ± 0.81
Br/J	$\mathbf{0.10} \pm 0.07$	$\underline{0.80} \pm 0.08$	3.51 ± 0.22	8.17 ± 0.30	6.53 ± 0.61
Lg/D	$\underline{1.42} \pm 0.30$	1.45 ± 0.28	$\mathbf{1.41} \pm 0.25$	4.28 ± 0.03	2.55 ± 0.56
Lg/M	$\underline{1.74} \pm 0.22$	1.82 ± 0.23	$\mathbf{1.27} \pm 0.28$	2.94 ± 0.17	3.27 ± 0.85
Lg/J	$\mathbf{0.16} \pm 0.12$	$\underline{0.20} \pm 0.14$	1.75 ± 0.31	4.24 ± 0.21	4.22 ± 0.67

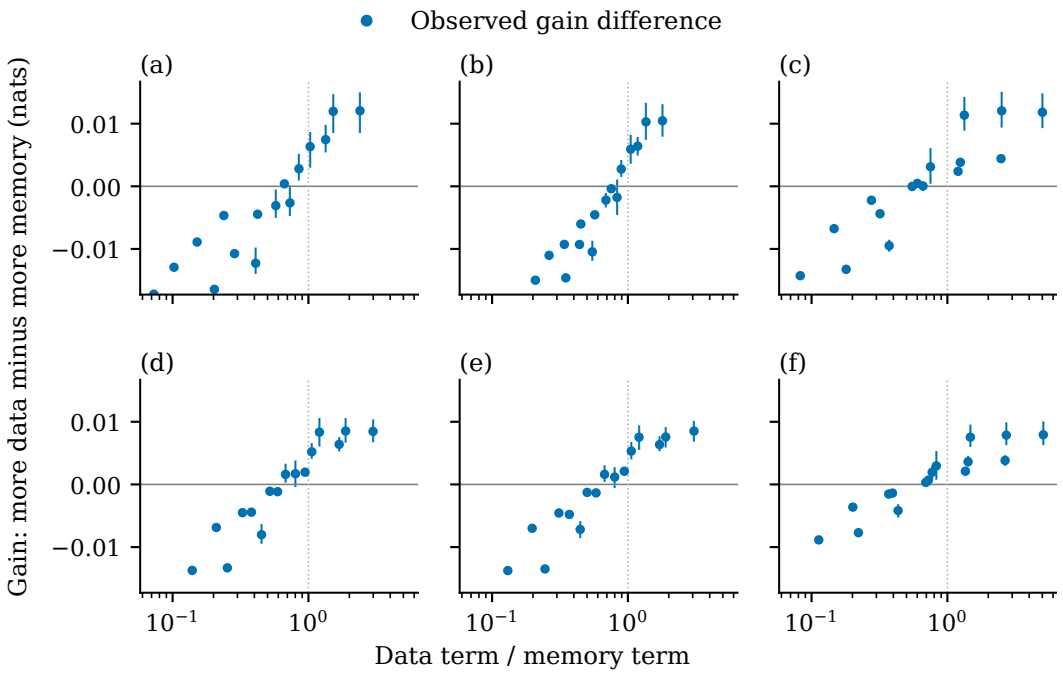


Figure 13: **Which resource improves prediction more?** (a–f) P1, P2, P3, P4, Br, Lg. Each point is a starting budget pair. Horizontal axis: the spectrum's data term divided by its memory term. Vertical axis: test-CE reduction from fourfold data expansion minus that from fourfold memory expansion. Positive values favor more data; negative values favor more memory. Error bars are 95% BCa over 16 paired SGD seeds. Reference lines at zero gain difference and unit spectral ratio are deterministic.

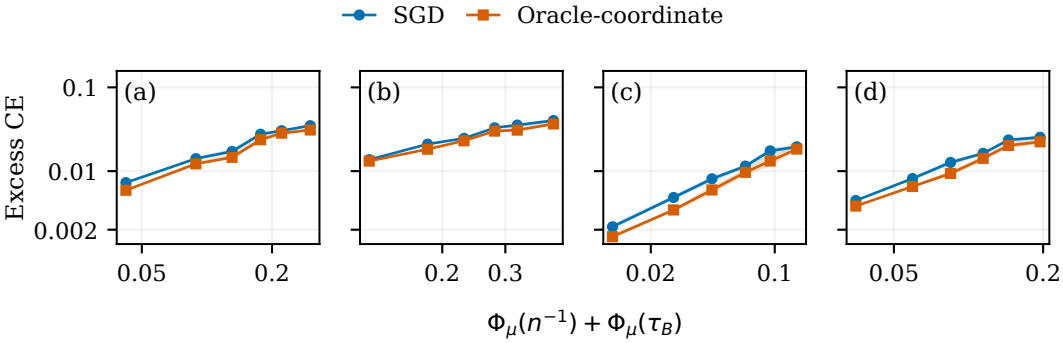


Figure 14: **The four-profile data–memory collapse.** (a–d) P1–P4, each with 256 energy groups and $K = 32$. Excess CE versus $G_\mu(n, B)$ compares weighted-average SGD (blue circles) with oracle-coordinate clipped estimation (orange squares). Within each seed, 30 budget pairs are averaged into six bins chosen from G_μ. Bands are 95% BCa over the eight paired seeds of this experiment. This binned view complements the unbinned held-out prediction tests.

Table 11: **Resource choices for weighted-average SGD.** G is signed test-CE gain (higher is better); R is regret against the best measured action (lower is better), both in 10^{-3} nats. Each seed averages 16 starting configurations. The smaller $\pm$ encloses the 95% paired BCa interval over 16 seeds. Spectrum and calibrated spectrum may also choose no expansion. Best/second displayed values are bold/underlined.

Source / metric	Spectrum	Calibrated	Always data	Always memory
P1/G	$\underline{7.71} \pm 0.59$	$\mathbf{8.15} \pm 0.77$	2.93 ± 0.66	6.21 ± 0.46
P1/R	$\underline{1.32} \pm 0.49$	$\mathbf{0.87} \pm 0.24$	6.09 ± 0.27	2.82 ± 0.70
P2/G	$\underline{7.07} \pm 0.52$	$\mathbf{7.32} \pm 0.80$	2.73 ± 0.63	5.77 ± 0.40
P2/R	$\underline{1.11} \pm 0.26$	$\mathbf{0.86} \pm 0.32$	5.45 ± 0.28	2.41 ± 0.54
P3/G	$\underline{5.17} \pm 0.46$	$\mathbf{6.42} \pm 0.81$	3.34 ± 0.85	3.40 ± 0.29
P3/R	$\underline{1.45} \pm 0.37$	$\mathbf{0.20} \pm 0.20$	3.27 ± 0.15	3.21 ± 0.74
P4/G	$\underline{5.22} \pm 0.40$	$\mathbf{6.22} \pm 0.56$	3.08 ± 0.56	3.76 ± 0.26
P4/R	$\underline{1.30} \pm 0.30$	$\mathbf{0.30} \pm 0.22$	3.44 ± 0.20	2.76 ± 0.46
Br/G	$\underline{5.13} \pm 0.24$	$\mathbf{5.99} \pm 0.47$	2.90 ± 0.41	3.72 ± 0.23
Br/R	$\underline{1.19} \pm 0.24$	$\mathbf{0.33} \pm 0.17$	3.42 ± 0.17	2.60 ± 0.39
Lg/G	$\underline{3.13} \pm 0.37$	$\mathbf{4.21} \pm 0.52$	2.63 ± 0.62	1.91 ± 0.24
Lg/R	$\underline{1.25} \pm 0.29$	$\mathbf{0.17} \pm 0.10$	1.76 ± 0.16	2.48 ± 0.55

Table 12: **Resource choices for direct clipped estimation.** G is signed test-CE gain (higher is better); R is regret against the best measured action (lower is better), both in 10^{-3} nats. Each seed averages 16 starting configurations. The smaller $\pm$ encloses the 95% paired BCa interval over 16 seeds. Spectrum and calibrated spectrum may also choose no expansion. Best/second displayed values are bold/underlined.

Source / metric	Spectrum	Calibrated	Always data	Always memory
P1/G	$\mathbf{9.15} \pm 0.47$	$\underline{8.54} \pm 0.43$	1.57 ± 0.55	8.29 ± 0.32
P1/R	$\mathbf{0.28} \pm 0.17$	$\underline{0.89} \pm 0.16$	7.86 ± 0.36	1.14 ± 0.30
P2/G	$\mathbf{8.73} \pm 0.42$	$\underline{7.97} \pm 0.45$	2.10 ± 0.57	7.65 ± 0.26
P2/R	$\mathbf{0.37} \pm 0.14$	$\underline{1.14} \pm 0.21$	7.00 ± 0.30	1.45 ± 0.39
P3/G	$\underline{5.82} \pm 0.23$	$\mathbf{6.16} \pm 0.56$	2.09 ± 0.72	4.71 ± 0.30
P3/R	$\underline{0.58} \pm 0.39$	$\mathbf{0.24} \pm 0.09$	4.31 ± 0.26	1.69 ± 0.83
P4/G	$\mathbf{5.97} \pm 0.36$	$\underline{5.69} \pm 0.60$	1.47 ± 0.47	5.02 ± 0.22
P4/R	$\mathbf{0.28} \pm 0.15$	$\underline{0.55} \pm 0.12$	4.78 ± 0.21	1.22 ± 0.50
Br/G	$\mathbf{5.85} \pm 0.27$	$\underline{5.64} \pm 0.45$	1.39 ± 0.45	5.06 ± 0.23
Br/R	$\mathbf{0.30} \pm 0.12$	$\underline{0.52} \pm 0.13$	4.77 ± 0.23	1.10 ± 0.39
Lg/G	$\underline{3.45} \pm 0.34$	$\mathbf{3.61} \pm 0.51$	1.18 ± 0.52	2.82 ± 0.26
Lg/R	$\underline{0.37} \pm 0.18$	$\mathbf{0.20} \pm 0.08$	2.63 ± 0.18	0.99 ± 0.44

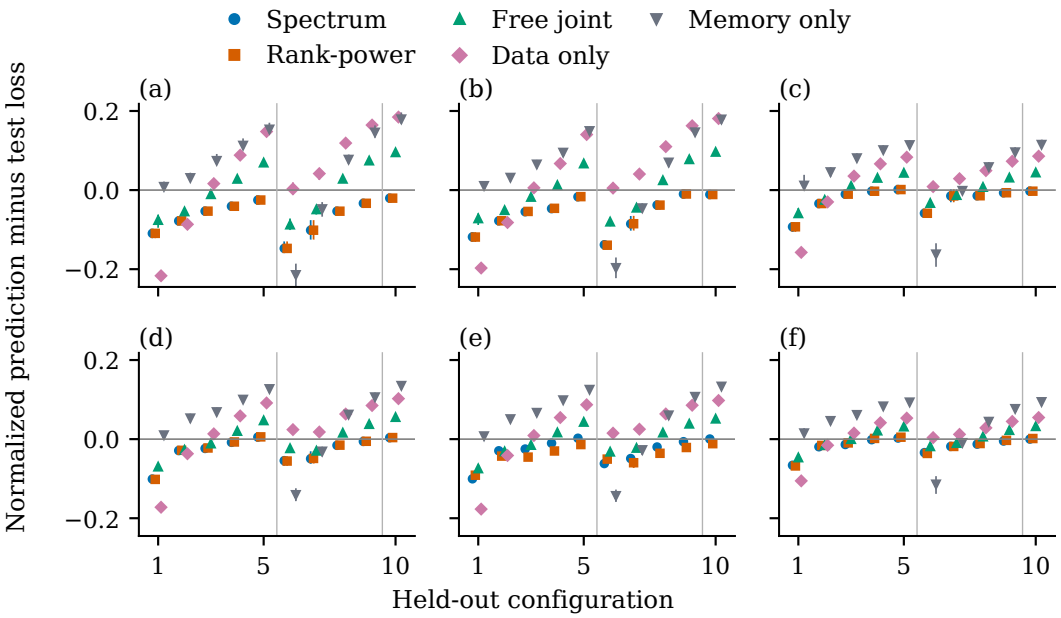


Figure 15: **Unbinned prediction residuals for weighted-average SGD.** (a–f) P1–P4, Br, Lg. Each point is a specific held-out budget pair, not a bin: 1–5 increase memory at maximum data; 6–9 increase data at maximum memory; 10 is the joint maximum. Residuals are predicted minus test excess CE, divided by zero-predictor population excess CE. Positive values overpredict risk. Error bars are 95% paired BCa over 16 seeds, computed after seed-specific fitting.

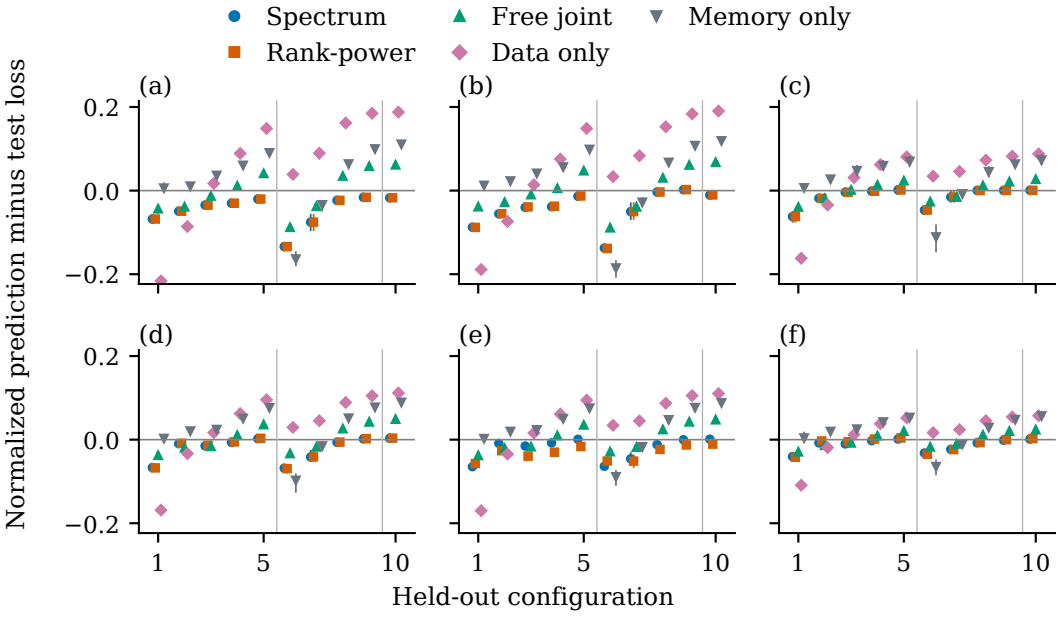


Figure 16: **Unbinned prediction residuals for direct clipped estimation.** Panels, configuration ordering, normalization and uncertainty match Figure 15. All five response families use the same validation-only calibration as the SGD comparison.

U.1 A Fixed-Penalty Comparison for Pythia

Can a format-specific offset explain how deployed loss changes with further training? We supplement the six-scale analysis with

$$\widehat{L}(T, f) = \widehat{L}_{\mathrm{BF16}}(T) + \widehat{\Delta}_f, \qquad \widehat{L}_{\mathrm{BF16}}(T) = c + a(T/2^{20})^{-\alpha}.$$

Fit $c, a \geq 0$, $\alpha \in [0.01, 4]$, to BF16 validation CE at the first three token budgets, using four fixed exponent starts. For 4-bit and 8-bit, $\widehat{\Delta}_f$ is the average validation CE difference from contemporary BF16 over these budgets; $\widehat{\Delta}_{\mathrm{BF16}} = 0$. Interpolate the unseen 6-bit penalty between 4-bit and 8-bit linearly in log deployment-package bytes, including retained parameters and metadata. Neither the largest token budget nor any 6-bit loss enters this fit.

This supplementary comparison was chosen after the original scan. It reuses the higher-bit analysis's calibration and holdouts, with no model reevaluation. The eight training seeds are averaged per configuration before fitting, as in that analysis. All 5,000 paired seed resamples repeat the complete fit; the same resampling indices give paired RMSE differences against each existing response family. The fixed penalty predicts 70M data extrapolation more accurately than the additive response; the additive response is more accurate for the other five scales and for every 6-bit interpolation. Each of these paired BCa differences excludes zero. Tables 13 and 14 report all five response families, without pooling model scales.

Table 13: **Fixed format penalties versus joint responses: largest-budget extrapolation.** RMSE in nats (lower is better). All five responses use the same 4/8-bit/BF16 validation calibration. The fixed-penalty model fits the BF16 data response and interpolates 6-bit penalties in log bytes. The $\pm$ values enclose the raw 95% BCa intervals from 5,000 full refits over eight paired training seeds. Best/second displayed values are bold/underlined; model scales are not pooled.

Model	Data only	Storage only	Additive	Interaction	Fixed penalty
70M	2.060 ± 0.061	1.293 ± 0.023	1.151 ± 0.080	$\underline{1.149} \pm 0.087$	$\mathbf{1.054} \pm 0.050$
160M	0.805 ± 0.016	0.619 ± 0.034	$\underline{0.425} \pm 0.038$	$\mathbf{0.392} \pm 0.036$	0.577 ± 0.046
410M	0.086 ± 0.004	0.156 ± 0.006	$\underline{0.058} \pm 0.006$	$\mathbf{0.054} \pm 0.007$	0.090 ± 0.005
1B	0.061 ± 0.003	0.096 ± 0.002	$\mathbf{0.041} \pm 0.003$	$\mathbf{0.041} \pm 0.004$	$\underline{0.050} \pm 0.004$
1.4B	0.057 ± 0.003	0.085 ± 0.002	$\mathbf{0.033} \pm 0.004$	$\mathbf{0.033} \pm 0.004$	$\underline{0.046} \pm 0.004$
2.8B	0.055 ± 0.003	0.074 ± 0.002	$\mathbf{0.034} \pm 0.004$	$\mathbf{0.034} \pm 0.004$	$\underline{0.046} \pm 0.004$

Table 14: **Fixed format penalties versus joint responses: 6-bit interpolation.** RMSE in nats (lower is better). All five responses use the same 4/8-bit/BF16 validation calibration. The fixed-penalty model fits the BF16 data response and interpolates 6-bit penalties in log bytes. The $\pm$ values enclose the raw 95% BCa intervals from 5,000 full refits over eight paired training seeds. Best/second displayed values are bold/underlined; model scales are not pooled.

Model	Data only	Storage only	Additive	Interaction	Fixed penalty
70M	$\mathbf{1.304} \pm 0.016$	1.845 ± 0.021	$\underline{1.842} \pm 0.023$	$\underline{1.842} \pm 0.023$	1.927 ± 0.025
160M	$\mathbf{0.649} \pm 0.016$	0.732 ± 0.016	$\underline{0.731} \pm 0.017$	$\underline{0.731} \pm 0.017$	0.922 ± 0.020
410M	$\underline{0.093} \pm 0.003$	0.094 ± 0.003	$\mathbf{0.081} \pm 0.003$	$\mathbf{0.081} \pm 0.003$	0.118 ± 0.004
1B	$\underline{0.054} \pm 0.002$	0.055 ± 0.002	$\mathbf{0.046} \pm 0.002$	$\mathbf{0.046} \pm 0.002$	0.066 ± 0.002
1.4B	0.053 ± 0.002	$\underline{0.047} \pm 0.002$	$\mathbf{0.041} \pm 0.002$	$\mathbf{0.041} \pm 0.002$	0.065 ± 0.002
2.8B	0.051 ± 0.002	$\underline{0.043} \pm 0.002$	$\mathbf{0.039} \pm 0.002$	$\mathbf{0.039} \pm 0.002$	0.062 ± 0.002

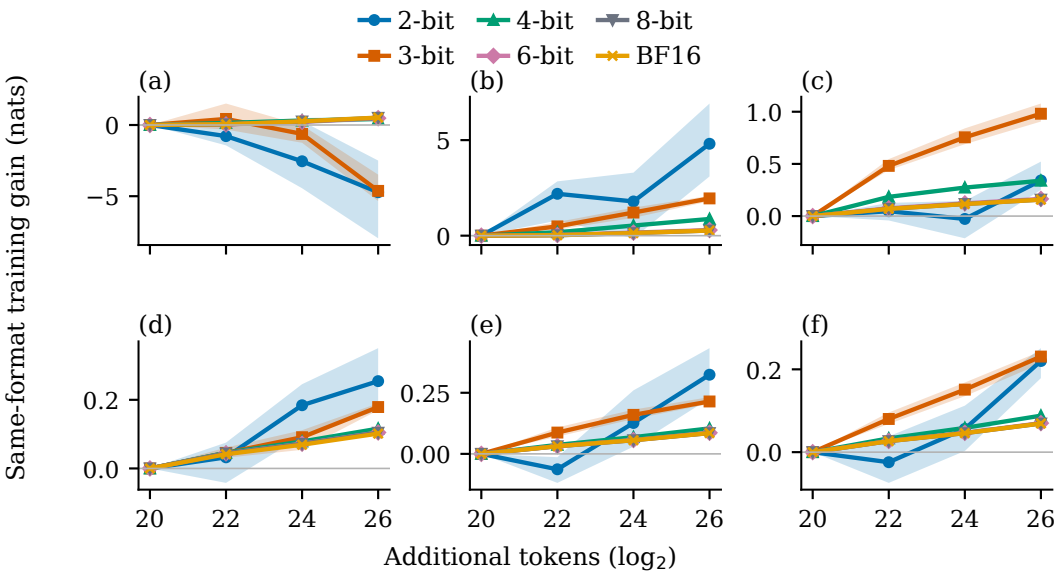


Figure 17: **What additional training changes within each deployment format.** (a–f) Pythia 70M, 160M, 410M, 1B, 1.4B, and 2.8B. Vertical axis subtracts CE at each budget from CE at the first continued-training budget, within the same format and seed. Positive values are improvements; negative values retain measured increases. The horizontal axis is log-base-two additional tokens. Bands are 95% BCa over eight paired training seeds with the same 96 test windows. The first point is identically zero by definition.

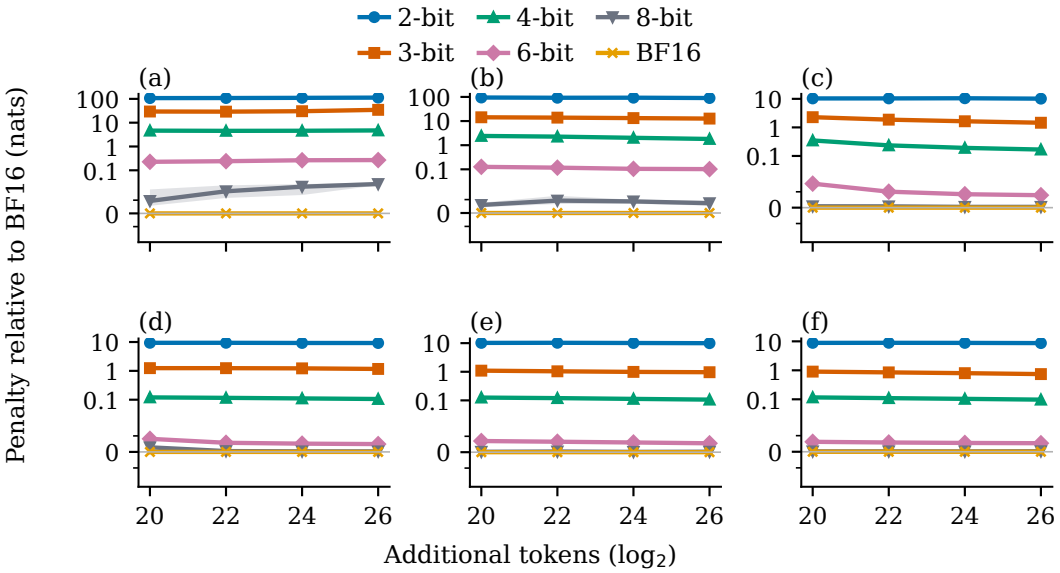


Figure 18: **Quantization penalties relative to the contemporaneous BF16 model.** (a–f) The same six scales as Figure 17. CE differences compare each quantized format with BF16 at the same training budget and seed, rather than with the initial checkpoint. The symmetric-log vertical axis is linear within ± 0.02 nats and logarithmic outside; each panel retains its full measured range. Bands are paired 95% BCa intervals over eight training seeds. BF16 is the zero reference.