RETRIEVAL CAPACITY AND ROBUSTNESS OF STATIC ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can query slots replace attention heads without sacrificing retrieval capacity or reliability? To the best of our knowledge, we establish the first joint sharp characterization of these two resources for static quadratic-coverage retrieval. For one tokenwise pooling round with h heads, R fixed query slots per head and $K = hR$ reads, stable length-uniform approximation of coverage over D prototypes is possible exactly when $K \geq D$. The converse allows arbitrary finite value width, length-dependent scores and joint nonlinear Lipschitz decoding, without a data-dependent bypass. Reliability obeys a different law: against s hidden whole-head corruptions, uniformly beating the constant baseline over unbounded count imbalance with bounded decoder–value sensitivity requires $h \geq 2s+1$ and $hR \geq (2s+1)D$, with matching constructions under sufficient contrast and noise margin. Exact finite-budget minimax frontiers quantify what read allocation, score contrast, value scale, noise and decoder sensitivity can and cannot compensate for. An exact decoder oracle isolates fixed-source representation limits, while a count-uniform certificate supplies trainable robust margins over an infinite known-prototype multiset class. All 20 construction-matched fits attain the predicted finite-budget frontiers, and 2,592 grouped-fault configurations match the exact formula. After a 100-step teacher prefix, direct-margin continuation recovers certified robustness in 15/15 fresh blocks at $D = 8, 16, 32$. Complementary frozen-Qwen3-4B stress tests probe bounded readout and native-head sensitivity, not transfer of the static theorem. The result separates interchangeable read channels from irreplaceable fault domains.

1 INTRODUCTION

Attention architectures allocate information across heads, query states and value coordinates (Vaswani et al., 2017). Learned pooling queries in Set Transformer (Lee et al., 2019), latent cross-attention in Perceiver (Jaegle et al., 2021), and register tokens (Darcet et al., 2024) make query states an explicit architectural resource. Head pruning and retrieval-head specialization further show that head count alone is an incomplete description of model behavior (Michel et al., 2019; Wu et al., 2025; Xiao et al., 2025). These observations motivate a structural question: *which resources supplied by attention heads can additional query slots replace, and which remain irreducible?*

The missing theoretical distinction. Existing theory establishes Transformer universality through contextual mappings (Yun et al., 2020), relates head count to approximation complexity (Yu et al., 2026), and identifies head-rank bottlenecks (Bhojanapalli et al., 2020; Amsel et al., 2025), while other work separates one-head from two-head attention-only models on graph-based retrieval tasks (Tesfaye et al., 2026). These results do not specify the joint head–slot substitution law studied here: whether equal numbers of pooled reads imply equal clean capacity *and* equal robustness when reads in one head fail together. To the best of our knowledge, we give the first sharp characterization of this distinction for static quadratic-coverage retrieval, coupling read count and fault-domain grouping with finite contrast, noise and decoder-sensitivity budgets.

We analyze one static round with $K = hR$ reads. Scores and values depend only on their token and channel, and a joint Lipschitz decoder receives all input-dependent information through the pooled values. For coverage of D prototypes, the exact stable, length-uniform threshold is $K = D$. Below it, a *common* background hides a prototype from every read simultaneously; at it, Gibbs reads attain

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.

any fixed accuracy with bounded decoder–value sensitivity and logarithmically growing score scale. Thus the capacity threshold counts reads, not heads in isolation.

Reliability breaks this equivalence. With bounded LV, uniformly beating the constant baseline over arbitrarily large count imbalance against s hidden whole-head corruptions requires

$$\boxed{h \geq 2s+1 \qquad \text{and} \qquad hR \geq (2s+1)D.}$$

Matching constructions attain both requirements with sufficient contrast and noise margin (Theorem 3.3). An extra slot can supply a read, but cannot create another fault domain inside the same head. Here independent fault domains mean distinct corruption groups, not statistically independent head activations. Classical redundancy (Chen et al., 2018) explains the $2s+1$ ingredient; the new result is the exact attention-specific error law that also accounts for how types share finite-contrast reads.

Why the result is broad and useful. The lower bounds range over arbitrary tokenwise score and value maps, arbitrary finite value dimensions and joint nonlinear Lipschitz decoders; the length-uniform converse even permits length-dependent scores. They are therefore not artifacts of linear decoding, scalar values or a fixed softmax temperature. The capacity theorem treats continuous token inputs, and the reliability construction covers an infinite multiset class with arbitrary omissions and bounded complete-input count imbalance. This parameterization-independent viewpoint applies to static softmax pooling and fixed-query cross-attention interfaces satisfying the assumptions, not just to one trained network. Exact frontiers make the distinction operational: they identify a resource shortfall, specify an attaining allocation, and quantify the error cost when contrast, noise margin or independent heads are insufficient. State width, metadata and compute remain separately counted resources.

Contributions.

(i) **A sharp head–slot substitution threshold.** We prove the necessary-and-sufficient $K = hR \geq D$ condition for stable length-uniform coverage, with a quantitative sensitivity penalty below threshold (Theorem 3.1).

(ii) **Exact finite-budget capacity–reliability laws.** Matching lower bounds and constructions jointly resolve read allocation, contrast, value scale, noise and whole-head faults, separating interchangeable reads from indispensable fault domains (Theorems 3.2 and 3.3).

(iii) **From resource limits to certified learning.** An exact fixed-encoder decoder oracle and a strictly enlarged count-uniform certificate distinguish representation and readout limitations and provide trainable robust-margin conditions (Section 4). Paired continuation studies test recovery after partial construction supervision; pretrained-model stress tests examine external signals (Section 5).

Learning retains known types, replica assignments and matched aggregation. The Qwen studies are complementary diagnostics, not a transfer of the static minimax guarantees to contextual, multiround Transformers.

2 Static reads and the resource being counted

Let $a_1, \dots, a_D \in [0,1]^d$ be distinct prototypes, $D \geq 2$, and set

$$f_i(x) = \|x - a_i\|_2^2/d, \qquad H(X) = \max_{i\in[D]} \min_{t\in[T]} f_i(x_t), \qquad \gamma = \min_{i\neq r} f_i(a_r) > 0. \tag{1}$$

Coverage is zero on a prototype sequence exactly when no type is missing. A static K-channel model computes

$$A_j(X) = \frac{\sum_t e^{s_j(x_t)} v_j(x_t)}{\sum_t e^{s_j(x_t)}}, \qquad \widehat{H}(X) = G(A_1(X), \dots, A_K(X)). \tag{2}$$

Each score is finite. Scores and values depend only on the token and channel. They may be retuned for each sequence length. Channel values have diameter at most V in their stated norm, and G is L-Lipschitz on the product of the value convex hulls, equipped with $\|z\|_{\mathrm{blk}} = \max_j \|z_j\|_j$. Write $C = LV$. There is no input-dependent bypass, positional score, or sample identifier.

For a query-to-data round with h heads and R fixed query slots, $K = hR$ counts simultaneous evaluations when normalization is over data tokens only. This is not an equality of parameter counts, state widths, FLOPs, or KV-cache sizes. The upper construction requires room for the queries and pooled coordinates. Replacing heads by slots without those resources is outside the claim.

Architectural correspondence and boundary. A one-round cross-attention module with h independent softmax heads and R fixed query slots per head produces exactly hR summaries of the form in Equation (2), after absorbing the query–key maps into s_j and the value projection into v_j. Parameter sharing within a head does not change the number of reads, while a whole-head fault acts jointly on its R summaries. Positional scores, contextualized token maps, query–query interaction, residual bypasses, or later access to token states break this reduction. The results are therefore sharp laws for this pooling interface, not universal laws for an arbitrary Transformer.

3 Sharp laws for capacity and reliability

3.1 How many reads are needed uniformly over length?

Let $\mathcal{E}_T = \sup_{X \in ([0,1]^d)^T} |\widehat{H}(X) - H(X)|$. The first theorem does not bound score magnitude or value width.

Theorem 3.1. *For $1 \le K < D$ and $T \ge D$, every model in Equation (2) satisfies*

$$\mathcal{E}_T \ge \frac{1}{2}\left[\gamma - \frac{C}{2 + \lfloor (T-D)/K \rfloor}\right]_+. \tag{3}$$

For every $\beta > 0$, with D channels there is a quadratic-feature construction with

$$0 \le \widehat{H}(X) - H(X) \le \frac{\log T}{\beta}, \qquad V_\infty \le 1, \qquad L_{\mathrm{blk}} \le 3. \tag{4}$$

Thus, for each fixed $0 < \varepsilon < \gamma/2$, error at most ε at every $T \ge D$ with $\sup_T C_T < \infty$ is achievable if and only if $K \ge D$, when the required state and readout resources are available.

Why a nonlinear decoder does not remove the obstruction. Each channel designates one score-maximizing prototype. When $K < D$, some type is designated by none. Repeat the designated types in a common background, then append either the missing type or an already present type. The target changes by at least γ, while all pooled values move by at most $V/(2 + \lfloor (T-D)/K \rfloor)$. This is one shared input pair, not separate counterexamples for separate heads. Lipschitz continuity then controls the joint decoder.

The upper construction uses scores $-\beta f_i(x)$, shared values $(x, \|x\|_2^2/d)$, and the maximum of D affine reconstructions. Its pooled width is $D(d+1)$, with $O(Dd)$ prototype metadata and an $O(\log D)$-depth ReLU maximum tree. Choosing $\beta = (\log T)/\varepsilon$ keeps decoder sensitivity bounded, but does not keep score precision bounded. Below threshold, fixed accuracy requires $C = \Omega(T/K)$. Appendix A gives the complete proof and head–slot realization.

3.2 At fixed length, contrast and value width matter

To obtain an exact comparator for experiments, specialize to a finite *balanced-omission* source. Assume every prototype has the same nearest-neighbor gap $\Delta > 0$. For each i, define X_i^+ to contain $D-1$ copies of a_i and $n(D-1)$ copies of each other type. Define X_i^- to omit a_i and contain $1 + n(D-1)$ copies of every other type. Both have length $T = (D-1)[1 + n(D-1)]$, with labels 0 and Δ, respectively. This enumerated source has $2D$ indexed inputs. It is not an IID held-out dataset.

Channel j now has score span at most B_j, value diameter at most V_j in the infinity norm, and coordinatewise adversarial noise radius η_j. Its finite value width may be chosen. The decoder is globally L-Lipschitz in the unweighted coordinate infinity norm. Let $\mathcal{E}^*_{\mathrm{res}}$ be the infimum of worst absolute error over all source inputs and their noise boxes, including the choice of encoder and decoder.

Theorem 3.2 (Finite-budget robust frontier)**.** *For integers $D \geq 2$, $n, K \geq 1$, and finite nonnegative budgets, define*

$$g_j(r) = \frac{1}{1 + n(r-1) + n(D-r)e^{-B_j}}, \quad 1 \leq r \leq D,$$
$$t_* = \max_{\substack{r_j \in \{0,\dots,D\} \\ \sum_j r_j = D}} \min_{j: r_j > 0} \{V_j g_j(r_j) - 2\eta_j\}. \tag{5}$$

Then the exact optimum is

$$\boxed{\mathcal{E}^*_{\mathrm{res}} = \tfrac{1}{2}[\Delta - L(t_*)_+]_+.} \tag{6}$$

A matching construction uses D useful value coordinates in total, partitioned into channel widths r_j. Independently prescribed narrower per-channel widths need not attain this optimum.

Mechanism. Assign r_j types to channel j, give those types score B_j, and store a distinct indicator coordinate for each assigned type. A present type contributes at least $V_j g_j(r_j)$, while an absent type contributes zero. Noise removes $2\eta_j$ from their separation. A clipped minimum readout attains Equation (6). Conversely, the score-span constraint limits how many types one channel can separate above any given margin. A capacity-counting argument forces a common type at or below t_* in every channel. This proves optimality even with arbitrary finite value width (Appendix B).

For homogeneous budgets, let $r = \lceil D/K \rceil$ when $K \leq D$, and $r = 1$ otherwise. Equation (6) reduces to

$$\mathcal{E}^*_{\mathrm{res}} = \tfrac{1}{2}[\Delta - L[Vg(r) - 2\eta]_+]_+. \tag{7}$$

For $K \geq D$, $g(1) = 1/[1 + n(D-1)e^{-B}]$. Thus sufficient head–slot coverage does not prevent dilution at fixed contrast as n grows. For $a = (\Delta/L + 2\eta)/V \in (0, 1)$, zero error in this regime is equivalent to $B \geq \max\{0, \log[n(D-1)a/(1-a)]\}$. Uniformly repeating every token instead leaves attention unchanged. At a fixed n, fewer than D channels may attain zero error using wider values. This does not contradict the length-uniform threshold. Crucially, monotonicity in the allowed contrast budget concerns an infimum over encoders and decoders. It does not imply that multiplying the logits of a fixed pretrained model improves its task accuracy.

3.3 When tokens cannot replace heads: correlated read failures

Equal hR does not imply equal reliability if all R reads in one head fail together. Group the static channels into h heads of R reads, with the same per-channel budgets (B, V, η) and global decoder sensitivity L. This is a grouped-read model, not a hardware fault simulator or a parameter-matched shared-K/V implementation. An adversary either erases up to s whole heads and reveals their identities, or replaces up to s whole-head memories by arbitrary vectors without revealing their identities. The decoder is globally L-Lipschitz, separately for each revealed erasure mask.

The result also moves beyond the enumerated source. Assume the prototypes are equidistant with gap Δ. Let $\mathcal{C}_n$ contain every nonempty prototype multiset that either misses a type or has all types present with largest/smallest count ratio at most the integer n. Then H equals Δ or 0, respectively. There is no fixed sequence length or access to a sample identifier.

Theorem 3.3 (Exact grouped-read reliability frontier)**.** *Set $q = s + 1$ for revealed erasures and $q = 2s + 1$ for hidden corruptions. If $h < q$, the minimax robust error on $\mathcal{C}_n$ is $\Delta/2$, regardless of R, contrast or value width. Otherwise, with $r = \lceil qD/(hR) \rceil$ and $g(r) = [1+n(r-1)+n(D-r)e^{-B}]^{-1}$, it is exactly*

$$\boxed{\mathcal{E}^*_{\mathrm{fault}} = \tfrac{1}{2}[\Delta - L[Vg(r) - 2\eta]_+]_+.} \tag{8}$$

When $h \geq q$, the optimum is attained using q indicator copies per type on distinct heads, with qD useful pooled coordinates in total. When $h < q$, the constant predictor $\Delta/2$ attains the optimum.

The lower bound counts how many types can be strongly exposed by each head, then uses one common missing type whose separating heads can be erased or split between two corruption patterns. For the upper bound, balance the copies across reads. Take a maximum over surviving copies for erasures, or a median over the $2s + 1$ copies for corruptions, followed by a clipped minimum across

types. These order statistics preserve the infinity-norm Lipschitz budget. Full proofs and a fault-aware decoder oracle are in Appendix D.

Thus nontrivial accuracy uniformly over count imbalance requires *both* $h \geq q$ and $hR \geq qD$ under bounded LV, even if contrast grows. At these thresholds, $B = \log n + b$ suffices for zero error when $V > 2\eta + \Delta/L$ and b is sufficiently large. For one hidden head corruption, 1×24 and 3×8 have the same read count but fundamentally different guarantees. Replication and median robustness are classical. The claim here is the exact attention-specific frontier and its head–slot obstruction, not a new error-correcting code.

4 An exact diagnostic of representation and decoding

For a fixed encoder, let z_a be the memory of source input a, y_a its label, and Q_a its noise box, with common coordinate radii h_k. The exact best decoder is computable from actual memories, without optimizing another attention layer.

Proposition 4.1 (Best Lipschitz decoder on fixed boxes)**.** *Set* $\delta_{ab} = \max_k [|z_{ak} - z_{bk}| - 2h_k]_+$. *Then*

$$\mathcal{E}^*_{\mathrm{dec}}(A) = \inf_{\mathrm{Lip}_\infty(G) \leq L} \max_a \sup_{z \in Q_a} |G(z) - y_a| = \tfrac{1}{2} \max_{a,b} [|y_a - y_b| - L\delta_{ab}]_+. \tag{9}$$

Necessity follows from the closest points of each pair of boxes. A maximum of distance cones attains the bound through classical scalar Lipschitz extension (McShane, 1934). This gives a computable benchmark for how much error the fixed representation forces (Appendix C).

For any admissible trained predictor on the balanced source,

$$\mathcal{E}(A, G) - \mathcal{E}^*_{\mathrm{res}} = \underbrace{\mathcal{E}^*_{\mathrm{dec}}(A) - \mathcal{E}^*_{\mathrm{res}}}_{\text{representation gap}} + \underbrace{\mathcal{E}(A, G) - \mathcal{E}^*_{\mathrm{dec}}(A)}_{\text{decoding gap}}. \tag{10}$$

Both terms are nonnegative, but neither is automatically an optimization gap. The first can include encoder restrictions. The second can include decoder architecture or a conservative norm constraint, as well as training failure. This distinction prevents interpreting every failed fit as a capacity lower bound.

The diagnostic is stable to measured representation error: if corresponding memories change by at most δ in block infinity norm and labels by at most ξ, with unchanged noise and fault budgets, their decoder oracles differ by at most $L\delta + \xi$ (Appendix I). This transfers a margin only when the discrepancy is controlled; it does not turn a constructed encoder into a guarantee for an independently trained network.

The formula also suggests a diagnostic trainable readout. With fixed training boxes and labels in $[0, \Delta]$, use

$$G_h(z) = \mathrm{clip}_{[0,\Delta]} \max_a \{h_a - L \operatorname{dist}_\infty(z, Q_a)\}. \tag{11}$$

Its heights h_a are learned, rather than set to the analytical optimum. For fixed boxes the objective

$$U(h) = \max\{0, \max_b (y_b - h_b), \max_{a,b} (h_a - L\delta_{ab} - y_b)\} \tag{12}$$

is a convex robust-error upper bound whose minimum equals $\mathcal{E}^*_{\mathrm{dec}}(A)$. The decoder stores the training anchors, so it is a source-adapted test of attainable decoding, not a general-purpose architecture or a held-out-label prediction guarantee.

5 Experiments: from construction to certified continuation

The experiments connect exact resource optima, attainable decoding and robust-margin learning. The first two controls use the complete balanced source with $D = n = 8$, $\Delta = 1/32$, $T = 399$, $B = \log 128$, and $V = L = 1$. Four fixed (K, η) regimes span zero and nonzero resource optima. Each uses five initialization seeds and 2,000 Adam steps. Results use final iterates, not selected checkpoints.

All errors below are divided by Δ, so the constant predictor has error 0.5. Brackets denote per-seed numerical robust-error lower/upper bounds averaged over seeds, not confidence intervals.

Attainment means excess at most 0.01Δ relative to the stated optimum. FP64 evaluation does not constitute outward-rounded formal certification. The construction-matched protocol and frozen-encoder diagnostic are in Appendix E.

5.1 Finite-source decoding controls

Freeze the matching construction of Theorem 3.2 and train a minimum/clip readout with exact noise-box extrema. All 20 final fits attain the resource optimum within the stated tolerance, including numerical zero in the three zero-optimum regimes (Table 1). The nonzero-optimum regime tests attainment of the predicted frontier, not a zero-error claim. This is calibration of a construction-matched family, not a comparison establishing superiority over generic architectures.

Table 1: Calibration of the construction-matched minimum/clip readout, all five seeds in each regime. Errors are divided by Δ, with mean $\pm$ seed SD. Attainment is excess at most 0.01Δ.

K, η	Resource optimum	Trained robust error	Attained
4, 0	0.0000	0.0000 ± 0.0000	5/5
4, 0.03	0.0000	0.0000 ± 0.0000	5/5
4, 0.05	0.3933	0.3940 ± 0.0000	5/5
8, 0.05	0.0000	0.0000 ± 0.0000	5/5

A second control freezes all 20 jointly MSE-trained table encoders and trains three fresh readouts each (60 fits). In the clearest $K = 8, \eta = 0.05$ regime, all five encoders have $\mathcal{E}^*_{\mathrm{dec}}(A) = \mathcal{E}^*_{\mathrm{res}} = 0$. Their learned anchor envelopes approach zero with mean normalized robust upper bound 1.45×10^{-4}, or absolute error 4.53×10^{-6}. The envelope stores source anchors and is not a held-out guarantee. The complete construction-matched protocol and budgets appear in Appendix E.

5.2 Attaining the whole-head reliability frontier

Equation (8), the fixed-memory fault oracle and actual constructed attention agree across all 2,592 deterministic configurations under exhaustive fault masks. The maximum error residual is 6.94×10^{-18}. These checks include both sides of the head-count threshold and verify the implementation rather than statistical generalization or learning.

Table 2 isolates the claimed resource separation. Increasing slots from eight to 24 preserves clean capacity but cannot create independent fault domains inside one head. Conversely, merely using three heads is insufficient when a type's replicas are concentrated in one of them. Reliability comes from distributing the same useful reads across independent groups, not from head count as a proxy for more parameters.

For learned calibration within the attainable regime, fix $K = 24$, $D = 8$, $V = L = 1$, $\eta = 0.05$ and $B = \log 16$. Each type has three indicator copies placed on distinct heads. At each $h \in \{3, 6, 12, 24\}$, five random initializations train only the intercept and slope of the clipped min-of-medians readout for 2,000 projected Adam steps on the 16 balanced examples at $n = 8$. The encoder and replication layout are constructed, not learned. All 20 final fits have zero numerical worst error over every allowed whole-head corruption and continuous noise box.

Each calibrated model is evaluated on 1,000 new histograms at each $n \in \{8, 32, 128\}$, including multiple missing types. All retain zero worst error at $n = 8$, and at larger n with the declared contrast adjustment $B(n) = \log 16 + \log(n/8)$. No test labels or anchors enter calibration. The adjustment receives the imbalance bound and changes the score budget, so this is not fixed-budget length extrapolation. An additional bounded-fault study, with corrupted coordinates in $[0, 1]$, also yields zero source and held-out worst error for all 20 recalibrated median readouts. Complete layouts, concentrated-copy, aggregation and fixed-contrast controls are in Appendices F, G and H. These are construction-guided fault-injection tests, not end-to-end discovery or hardware-error recovery.

5.3 Learning robust margins beyond table imitation

The conditional learning problem. All arms use known prototype types, a fixed replica layout and a min-of-medians readout. The shared-K/V encoder's features, keys, values and queries are trained.

Table 2: The central equal-read comparison for one hidden head corruption. All rows have $K = 24$, $D = 8$, $n = 8$, $B = \log 16$ and $\eta = 0.05$. Errors are exact decoder oracles normalized by Δ. "Separated" places the three copies of every type on distinct heads; "concentrated" keeps a type's copies in one head. The complete grid is in Appendix G.

Layout	Reads K	Clean error	One-fault error
1×24	24	0.000	0.500
3×8, separated	24	0.000	0.000
3×8, concentrated	24	0.000	0.500

After a teacher prefix, direct continuation removes table targets but retains type identities and replica assignments. The question is whether this partially supervised encoder can learn the required robust margins without continuing to match the teacher tables. Appendix Q accounts for the fixed structural priors.

A sufficient neighborhood of the construction. If ϵ_v bounds value error and ρ bounds score-error oscillation, every multiset's memory error is at most $\delta_{\mathrm{tab}} = \epsilon_v + \min\{1, \rho/4\}$. For the matched unit-Lipschitz readout,

$$\delta_{\mathrm{tab}} \leq \frac{g(1) - 2\eta - \Delta}{2} \tag{13}$$

suffices for zero error throughout $\mathcal{C}_n$ with one arbitrary head fault. This is an a posteriori condition, not optimizer convergence. The token-map proof, local continuous-input extension and representability argument appear in Appendix J.

At $D = 8$, full supervision, teacher-then-task and gradient-balanced fading supervision each meet the condition in 30/30 fits. All 120 supervised scale-study fits at $D = 16, 32$ also meet it, using full-teacher or balanced-fade schedules. These use 2,000 updates, $h = 3, 6$, $K = 3D$, key/value width D, hidden width 32 and $B_D = \log 16 + \log((D-1)/7)$, keeping $g(1) = 2/9$ at $n = 8$. Resources grow with D. Full protocols and controls are in Appendix K.

A strictly larger certified representation family. For each assigned replica ℓ, minimize its actual scalar attention output over positive count ratios at most n. Let $A(n)$ be the smallest such minimum and U the largest off-type value in an assigned coordinate. With fixed readout bias b, the direct condition

$$A(n) \geq b + \eta, \qquad U \leq b - \Delta - \eta \tag{14}$$

suffices for zero robust error throughout $\mathcal{C}_n$. Sorting the values reduces each minimum to $D + 1$ count vectors. The region certified by Equation (14) strictly contains that certified by our table-approximation bound (Proposition L.1). Thus proximity to the teacher tables is stronger than robust retrieval requires. This turns the actual worst-count margin into a trainable target, evaluated by elementary linear-fractional optimization rather than by enumerating all multisets.

Recovery beyond the table-approximation region. An initial paired $h = 3$ study resets Adam after 100 teacher updates and optimizes violations of Equation (14) to 2,000 total updates. All 15 final models meet the condition and have zero worst error on 1,000 fresh $n = 8$ counts each. A separately frozen confirmation repeats this in five new blocks at each of $D = 8, 16, 32$ (Table 3). All corresponding prefixes have nonzero source error. Thus the result is recovery, not retention, and all these final direct models lie outside the table-bound sufficient neighborhood. The frozen protocol and paired confirmation are reported in Appendix N.

The paired evidence establishes that continued table imitation is unnecessary in this tested regime. Continued teacher supervision also succeeds, so this is an alternative learning route, not an accuracy gain over the teacher. Direct constraints and finite-source task loss differ in count coverage as well as objective form. The focused protocols and paired controls are in Appendices M and N. Inference uses no teacher or test labels, while retaining known types, replica layout and matched readout. Fresh counts test multiplicity robustness, not unseen-type generalization.

Table 3: Conditional margin learning after a shared 100-update teacher prefix. Each dimension has five fresh paired blocks, with $h = 3$ and $n = 8$. Columns count final models and give their certificate-margin range. All frozen prefixes have nonzero source error. Paired controls are in Appendix N.

D	Source zero	Fresh zero	Direct condition	Margin range
8	5/5	5/5	5/5	$[0.00098, 0.00629]$
16	5/5	5/5	5/5	$[0.00035, 0.00637]$
32	5/5	5/5	5/5	$[0.00193, 0.00851]$

5.4 Bounded pooled-value readout from a frozen pretrained model

The theorem concerns static reads, whereas a pretrained Transformer adds contextual token interactions and residual paths. We therefore use Qwen3-4B only for two mechanism-oriented stress tests. The first asks whether frozen contextual features remain usable through a bounded pooled-value interface after a length shift. The second asks whether redistributing a fixed native intervention budget across heads changes the model's answer likelihood after matching the realized projection norm. These tests examine whether the relevant signals remain visible outside the construction. They do not transfer the minimax guarantees to Qwen.

We use RULER-derived variable tracking (VT) (Hsieh et al., 2024) with ten parsed candidate assignments and five target variables. Frozen Qwen3-4B (Yang et al., 2025) supplies assignment-endpoint features. A lightweight reader learns K softmax pools of fixed ten-dimensional address values. The decoder sees only the pooled vectors, with value diameter at most one and logit Lipschitz constant 20. Its ten independent membership decisions have no K-dependent output cap. The main query encodes the question alone. We train on 80 4K sources, retain 20 for validation, and test on 40 sources at each of 16K and 32K. Three seeds and $K \in \{1, 3, 4, 5, 6, 8\}$ give 54 fits across contextual candidates, a contextual-query ablation and an input-embedding candidate control. All budgets use 68,096 trainable parameters.

Table 4: Bounded pooled-value VT readout at the smallest tested budget, $K = 1$, over 80 test sources and three seeds. Exact and recall use independent membership thresholds. Top-5 is a separate known-cardinality decoder. Both Qwen rows have the same clean outcomes at every tested K.

Candidate / query features	Exact (%)	Recall (%)	Top-5 exact (%)
Qwen / question only	98.75	99.75	100.00
Qwen / contextual query	98.75	99.75	100.00
Input / question only	0.00	34.08	1.25

The question-only reader reaches 98.75% exact recovery (source-clustered 95% bootstrap interval $[96.25, 100]\%$), versus 0% for input candidates. Exact recovery is 100% at 16K and 97.5% at 32K. Removing the pooled memory reduces thresholded recall to zero. Thus contextual representations support accurate, bounded-interface readout beyond the training length without a contextual query bypass. Importantly, $K = 1$ already matches larger budgets: one vector can carry several target coordinates. The result complements the finite-width mechanism but does not establish a five-read threshold. Contextual features, candidate parsing and set decisions differ from the static coverage model. These are not official RULER generation scores. The complete grid, dilution controls and audit details are in Appendix O.

Native head-output intervention. We separately intervene on native Qwen3-4B head outputs during full prefill on fresh RULER-derived direct-retrieval sources. Layer 30, suffix positions and heads 11, 6 and 5 are selected once on disjoint V7 calibration sources. V8 and V9 then freeze that choice and the intervention code without reselection. Each of three rotations per family replaces the same eight token–head vectors, and the two families have balanced aggregate head–position marginals and matched actual BF16 output-projection change. Each length uses 64 new source-paired examples.

The dispersed-minus-concentrated log-probability margin is $+0.01693$ $[0.00794, 0.02683]$ at 16K and $+0.02224$ $[0.01383, 0.03125]$ at 32K after matching the realized projection-output norm within

0.051%. The likelihood effect repeats across lengths, while exact-set accuracy does not improve. Sham and donor controls pass. The complete behavioral outcomes, protocol and integrity checks are in Appendix P.

What the pretrained-model controls establish. The VT controls rule out a reader-only solution and a contextual-query bypass. Norm matching excludes scalar projection magnitude as the sole explanation of the native margin contrast. Head identity, perturbation direction and token position remain coupled, and the native intervention does not improve exact-set accuracy. The pretrained-model evidence is therefore diagnostic: it establishes bounded readout and repeatable margin sensitivity, while the static theorems carry the capacity and reliability claims.

Claim–evidence correspondence. Theorem 3.1 gives the all-length capacity threshold. Table 1 instead tests the finite-source optimum of Theorem 3.2; it is not an all-length threshold experiment. The 2,592 grouped-fault checks and equal-K oracles in Table 2 test Theorem 3.3. Direct-margin continuation establishes conditional certified recovery, whereas Qwen tests provide external diagnostics without a behavioral gain or transfer of the minimax guarantee. None establishes general task-only discovery, fixed-budget all-length extrapolation or a pretrained-model compression method. Claim-bearing controls are in Appendices F–Q; complete exploratory records remain in the reproducibility package.

6 Relation to prior theory

Head count, rank and expressivity. Yu et al. (2026) study approximation complexity with one classification-token read per head and constrained embeddings/readouts. Bhojanapalli et al. (2020); Amsel et al. (2025) analyze head-rank bottlenecks; Tesfaye et al. (2026) give a dimension- and precision-independent one-versus-two-head separation for an attention-only endpoint-selection task. We instead count all fixed-query reads, constrain LV rather than parameter count, and determine exact coverage and grouped-fault errors. The model classes are not nested; Appendix Q details the distinctions.

Queries, set representations and practical heads. Learned seed queries (Lee et al., 2019), latent cross-attention (Jaegle et al., 2021) and registers (Darcet et al., 2024) motivate query states as resources. Their full contextual or iterative architectures are not instances of our static model. Deep Sets (Zaheer et al., 2017) supplies a permutation-invariant modeling framework; subsequent set-encoding lower bounds (Wagstaff et al., 2019; 2022) constrain latent dimension, whereas our obstruction persists with arbitrary finite value width. Head-pruning experiments (Michel et al., 2019), retrieval-head identification (Wu et al., 2025) and head-aware KV allocation (Xiao et al., 2025) concern redundancy and specialization in trained models, not exact head–slot substitution under hidden corruption.

Stability, dilution and redundancy. Attention influence and bounded-logit dispersion have established analyses (Hahn, 2020; Veličković et al., 2025). Our length-uniform lower bound permits length-dependent, unbounded score budgets; the finite-budget frontier quantifies the optimal contrast/noise tradeoff. Attention-operator Lipschitz analyses address input-domain and sequence-length effects (Kim et al., 2021; Castin et al., 2024); our L instead constrains the post-pooling decoder, not the full attention map. Norm-constrained GroupSort networks approximate Lipschitz functions (Anil et al., 2019); our oracle optimizes over all admissible scalar Lipschitz decoders on fixed uncertainty sets, using classical extension (McShane, 1934). DRACO (Chen et al., 2018) supplies a redundancy precedent, while FT-Transformer (Dai et al., 2025) protects attention computations against soft errors. These ingredients do not determine the grouped softmax-retrieval frontier; novelty lies in the joint resource characterization.

Conclusion. Static retrieval has two distinct architectural resources: pooled reads supply coverage capacity, while disjoint heads supply fault domains. Matching thresholds and exact finite-budget frontiers show when query slots are substitutes for heads and when no number of extra slots can repair a head shortage. Decoder diagnostics and count-uniform margins connect this characterization to conditional representation learning. The pretrained-model studies are complementary stress tests, not general Transformer guarantees or evidence of compression gains.

Reproducibility and responsible use

Proofs and claim-bearing protocols follow in the appendix. Tables are generated from retained result records. The source package preserves all executed arms, final checkpoints, formal test inputs and independent replays, including results not summarized in the manuscript. Confirmatory seeds and final iterates are not selected by outcome. The findings do not establish deployment safety or general language-model compression. Generative AI use is disclosed separately below.

AI use declaration

Generative AI assisted the experimental implementation, literature searches, and manuscript revision. The authors are responsible for the final proofs, citations, data, and conclusions.

References

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over Quantity in Attention Layers: When Adding More Heads Hurts. *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/9c537882044c8b5352c363e840872ddb-Abstract-Conference.html.

Cem Anil, James Lucas, and Roger Grosse. Sorting Out Lipschitz Function Approximation. *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 97:291–301, 2019. https://proceedings.mlr.press/v97/anil19a.html.

Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-Rank Bottleneck in Multi-head Attention Models. *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 119:864–873, 2020. https://proceedings.mlr.press/v119/bhojanapalli20a.html.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How Smooth Is Attention? *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, 235:5817–5840, 2024. https://proceedings.mlr.press/v235/castin24a.html.

Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. DRACO: Byzantine-resilient Distributed Training via Redundant Gradients. *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 80:903–912, 2018. https://proceedings.mlr.press/v80/chen18l.html.

Huangliang Dai, Shixun Wu, Jiajun Huang, Zizhe Jian, Yue Zhu, Haiyang Hu, and Zizhong Chen. FT-Transformer: Resilient and Reliable Transformer with End-to-End Fault Tolerant Attention. *arXiv preprint arXiv:2504.02211v2*, 2025. doi: 10.48550/arXiv.2504.02211. https://arxiv.org/abs/2504.02211v2.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. *International Conference on Learning Representations (ICLR)*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.html.

Michael Hahn. Theoretical Limitations of Self-Attention in Neural Sequence Models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306. https://aclanthology.org/2020.tacl-1.11/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the Real Context Size of Your Long-Context Language Models? *First Conference on Language Modeling (COLM)*, 2024. https://arxiv.org/abs/2404.06654.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General Perception with Iterative Attention. *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 139:4651–4664, 2021. https://proceedings.mlr.press/v139/jaegle21a.html.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz Constant of Self-Attention. *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 139:5562–5571, 2021. https://proceedings.mlr.press/v139/kim21i.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 97:3744–3753, 2019. https://proceedings.mlr.press/v97/lee19d.html.

E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934. https://doi.org/10.1090/S0002-9904-1934-05978-0.

Paul Michel, Omer Levy, and Graham Neubig. Are Sixteen Heads Really Better than One? *Advances in Neural Information Processing Systems*, 32, 2019. https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html.

Amanuel Tesfaye, Zeno Kujawa, Rajmohan Rajaraman, and Ravi Sundaram. Two (narrow) heads are better than (an arbitrarily wide) one. *International Conference on Learning Representations (ICLR)*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/221ec998e345bf4a128bf6c48e1aadee-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not Enough (for Sharp Size Generalisation). *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 267:61190–61211, 2025. https://proceedings.mlr.press/v267/velickovic25a.html.

Edward Wagstaff, Fabian Fuchs, Martin Engelcke, Ingmar Posner, and Michael A. Osborne. On the Limitations of Representing Functions on Sets. *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 97:6487–6494, 2019. https://proceedings.mlr.press/v97/wagstaff19a.html.

Edward Wagstaff, Fabian B. Fuchs, Martin Engelcke, Michael A. Osborne, and Ingmar Posner. Universal Approximation of Functions on Sets. *Journal of Machine Learning Research*, 23(151):1–56, 2022. https://jmlr.org/papers/v23/21-0730.html.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. *International Conference on Learning Representations (ICLR)*, 2025. https://openreview.net/forum?id=EytBpUGB1Z.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. *International Conference on Learning Representations (ICLR)*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/5c1ddd2e59df46fd2aa85c833b1b36ed-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. https://arxiv.org/abs/2505.09388.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The Effect of Attention Head Count on Transformer Approximation. *International Conference on Learning Representations (ICLR)*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d72a514fc6948515af3ee69c1059776-Abstract-Conference.html.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are Transformers universal approximators of sequence-to-sequence functions? *International Conference on Learning Representations (ICLR)*, 2020. https://openreview.net/forum?id=ByxRM0Ntvr.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. *Advances in Neural Information Processing Systems*, 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

A PROOF OF THE LENGTH-UNIFORM THRESHOLD

A common-background inequality. If u, v, w lie in a common convex set of diameter at most V, then

$$\|(1-\alpha)u + \alpha v - (1-\alpha')u - \alpha' w\| \le V \max(\alpha, \alpha'), \qquad \alpha, \alpha' \in [0,1]. \tag{15}$$

For $\alpha \ge \alpha'$, write the difference as $\alpha'(v-w) + (\alpha - \alpha')(v-u)$ and apply the triangle inequality. The other ordering is symmetric.

Lower bound in Theorem 3.1. For each channel choose $r_j \in \arg\max_r s_j(a_r)$ and set $\mathcal{B} = \{r_j : j \le K\}$. Since $|\mathcal{B}| \le K < D$, choose $i \notin \mathcal{B}$. Let $m = 1 + \lfloor (T-D)/K \rfloor$. Construct a length-$(T-1)$ background C by including each prototype except a_i once and adding $m-1$ copies of each distinct prototype in $\mathcal{B}$. This uses at most

$$D - 1 + K(m-1) \le T - 1$$

positions. Fill any remaining positions with a prototype other than a_i. For any $b \ne i$, compare $X^+ = (C, a_i)$ and $X^- = (C, a_b)$. The first covers every prototype, so $H(X^+) = 0$. The second omits a_i, so $H(X^-) \ge \gamma$.

For channel j, the background normalizer is at least $me^{s_j(a_{r_j})}$. Either appended token has score at most $s_j(a_{r_j})$, hence attention mass at most $1/(m+1)$. Their pooled reads share the same conditional background average. By Equation (15),

$$\|A_j(X^+) - A_j(X^-)\|_j \le V/(m+1).$$

Taking the block maximum and applying the decoder's Lipschitz bound gives $\gamma \le 2\mathcal{E}_T + LV/(m+1)$. Nonnegativity proves Equation (3). Repeated tokens have identical scores and values, which is where position and context independence enter. The argument applies to every model at each length, including length-dependent score and value maps. □

Upper bound and threshold. For prototype i, use

$$k(x) = v(x) = (x, \|x\|_2^2/d), \qquad q_i = \beta(2a_i/d, -1).$$

The score $q_i^\top k(x)$ differs from $-\beta f_i(x)$ by a token-independent constant, so its softmax weights are $p_{it} \propto e^{-\beta f_i(x_t)}$. If $u_i = (\bar{x}_i, \bar{q}_i)$ denotes the read, set

$$g_i(u_i) = \|a_i\|_2^2/d - 2a_i^\top \bar{x}_i/d + \bar{q}_i, \qquad G(u) = \max_i g_i(u_i).$$

Then $g_i(A_i(X)) = \sum_t p_{it} f_i(x_t)$. Let $z_i = \min_t f_i(x_t)$ and $Z_i = \sum_t e^{-\beta f_i(x_t)}$. The entropy identity and $Z_i \ge e^{-\beta z_i}$ imply

$$\beta \sum_t p_{it} f_i(x_t) = -\sum_t p_{it} \log p_{it} - \log Z_i \le \log T + \beta z_i.$$

A weighted average is at least its minimum. Taking the maximum over i proves Equation (4). Values lie in $[0,1]^{d+1}$. Each affine reconstruction has coefficient ℓ_1 norm $1 + 2\|a_i\|_1/d \le 3$, and taking a maximum preserves this block-max Lipschitz bound.

For $K < D$ and bounded C_T, the lower bound tends to $\gamma/2$. For $K = D$, choose $\beta_T = (\log T)/\varepsilon$ and obtain $C_T \le 3$. Unused channels can be ignored. This proves the exact length-uniform threshold.

To realize the construction using h heads and R slots with $hR \ge D$, assign each prototype to one slot/head pair. Store its query metadata in that slot's head-specific block, use linear projections to select the block, and share the quadratic K/V features. Pool over data tokens only and apply the same maximum decoder. This requires the stated query metadata and $D(d+1)$ active pooled coordinates. It does not hold state width fixed as the head/slot configuration changes. □

B Proof of the finite-budget frontier

Theorem 3.2. Write $w_{ji} = e^{s_j(a_i)}$ and

$$\alpha_{ji} = \frac{w_{ji}}{w_{ji} + n\sum_{\ell \neq i} w_{j\ell}}, \qquad q_j(r) = V_j g_j(r) - 2\eta_j.$$

For any set of r types, choose the one with smallest weight w. The other $r-1$ selected weights are at least w, and every remaining weight is at least we^{-B_j}. Therefore some type in the set has $\alpha_{ji} \leq g_j(r)$.

Let $c_j = \max(\{0\} \cup \{r : q_j(r) > t_*\})$. The preceding argument shows that channel j has at most c_j types satisfying $V_j\alpha_{ji} - 2\eta_j > t_*$. If $\sum_j c_j \geq D$, one could choose integers $r_j \leq c_j$ summing to D. Monotonicity of q_j would make every active group's margin strictly larger than t_*, contradicting Equation (5). Some common type i thus satisfies $V_j\alpha_{ji} - 2\eta_j \leq t_*$ for every channel.

Conditioning the positive read X_i^+ on types other than i gives exactly the negative read X_i^-. The diameter bound implies $\|A_j(X_i^+) - A_j(X_i^-)\|_\infty \leq V_j\alpha_{ji}$, regardless of finite value width. The distance between their product noise boxes is consequently at most $(t_*)_+$. Their labels differ by Δ, so every L-Lipschitz decoder has worst error at least $[\Delta - L(t_*)_+]_+/2$.

For attainment, partition the types into groups S_j of optimal sizes r_j. Give a group member score B_j and every other type score zero. For each $i \in S_j$, allocate one value coordinate $V_j\mathbf{1}\{x = a_i\}$. A missing type has coordinate zero. On a positive source input, every other count divided by a present type's count is at most n. Its coordinate is at least $V_j g_j(r_j)$. This construction uses D useful coordinates across the active channels. No minimal value-dimension claim is made.

If $t_* \leq 0$, the constant $\Delta/2$ attains the bound. Otherwise set

$$u(z) = \min_{j, i \in S_j}(z_{ji} - \eta_j), \qquad \ell = \min(L, \Delta/t_*), \qquad \epsilon = (\Delta - \ell t_*)/2.$$

The function

$$G(z) = \mathrm{clip}_{[\epsilon, \Delta-\epsilon]}\{\Delta/2 + \ell(t_*/2 - u(z))\}$$

is globally L-Lipschitz. Positive boxes have $u \geq t_*$ and negative boxes have $u \leq 0$, yielding error exactly ϵ. This proves both attainability and the formula, including zero budgets. □

Homogeneous budgets. For $K \leq D$, at least one active group has size $r \geq \lceil D/K \rceil$. Since g is nonincreasing in r, its margin bounds the minimum. A balanced partition attains this bound. For $K > D$, use D singleton groups and ignore the rest, giving $r = 1$. These observations prove Equation (7). When $L, V > 0$, $K \geq D$, and $a = (\Delta/L + 2\eta)/V$, zero error is equivalent to $g(1) \geq a$. For $0 < a < 1$, rearrangement gives the contrast threshold in the main text. At $a = 1$ it is approached only as $B \to \infty$, while $a > 1$ leaves a positive floor. If $L = 0$ or $V = 0$, the optimum is $\Delta/2$.

C Proof of the decoder diagnostic

For the product boxes, choosing the closest coordinatewise points gives $\mathrm{dist}_\infty(Q_a, Q_b) = \delta_{ab}$. If a decoder has robust errors e_a, e_b, then

$$|y_a - y_b| \leq e_a + e_b + L\delta_{ab}.$$

This proves the lower bound in Proposition 4.1. Let $e = \frac{1}{2}\max_{a,b}[|y_a - y_b| - L\delta_{ab}]_+$ and define

$$G_e(z) = \max_a\{y_a - e - L\,\mathrm{dist}_\infty(z, Q_a)\}.$$

Distance to a nonempty set is 1-Lipschitz, so G_e is L-Lipschitz. On Q_b, its own term is $y_b - e$, and each term is at most $y_a - e - L\delta_{ab} \leq y_b + e$. Thus its robust error is at most e, proving equality. Clipping to the label interval does not increase error. Intersecting boxes and $L = 0$ are included.

For Equation (12), the untruncated learned envelope on Q_b lies between h_b and $\max_a(h_a - L\delta_{ab})$, so $U(h)$ is an error upper bound. All entries in its maximum are affine in h, hence U is convex. Setting $h_a = y_a - e$ with $e = \mathcal{E}^*_{\mathrm{dec}}(A)$ yields $U(h) \leq e$ by the pairwise inequalities. Proposition 4.1 gives the reverse bound. This proves the minimum claim without asserting convergence of any particular optimizer.

D GROUPED FAILURES: ORACLE, LOWER BOUND AND ATTAINMENT

An independent fixed-memory comparator. Write z_{ah} for all coordinates of head h on source input a. For noise radius η, define

$$d_{ab,h} = [\|z_{ah} - z_{bh}\|_\infty - 2\eta]_+, \qquad d_{ab}^{[q]} = q\text{th largest of } (d_{ab,1}, \dots, d_{ab,h}).$$

Set $d_{ab}^{[q]} = 0$ when $q > h$. On any finite labeled source, the optimal decoder error under the stated faults is

$$\mathcal{E}^*_{\mathrm{dec,fault}}(A) = \tfrac{1}{2} \max_{a,b} [|y_a - y_b| - L d_{ab}^{[q]}]_+. \tag{16}$$

For erasures, a common mask removes at most s head distances, so their smallest surviving maximum is the $(s+1)$st largest. Apply the distance-cone argument of Proposition 4.1 on the retained coordinates separately for each mask. Maximizing over masks and input pairs proves the formula, including $s \geq h$.

For hidden corruptions, each input's admissible memories form a finite union of cylinders: select up to s unrestricted head blocks and retain noise boxes on the others. Two cylinders can match on the union of their corrupted heads, removing at most $2s$ distances. Every set of at most $2s$ heads can be split into two sets of at most s, so the distance between the two unions is exactly $d_{ab}^{[2s+1]}$. If e denotes the right-hand side of Equation (16), the function

$$G(z) = \max_a \{y_a - e - L \operatorname{dist}_\infty(z, \mathcal{U}_a)\}$$

has error at most e on every uncertainty union $\mathcal{U}_a$. Distance to a nonempty set is 1-Lipschitz even for these unbounded sets. Comparing closest admissible memories gives the reverse inequality. Thus the formula is an exact oracle, not just a fault attack lower bound. It applies to arbitrary value maps and head widths.

Lower bound in Theorem 3.3. The balanced-omission pairs of Theorem 3.2 belong to $\mathcal{C}_n$. For each read j, retain the notation $\alpha_{ji} = w_{ji}/(w_{ji} + n\sum_{\ell \neq i} w_{j\ell})$. The score-span argument shows that at most $r-1$ types can satisfy $\alpha_{ji} > g(r)$. Call a head strong for a type if at least one of its reads satisfies this inequality. Across all heads, the number of strong head–type incidences is at most

$$hR(r-1) < qD.$$

Consequently there is one type i with at most $q-1$ strong heads. For every other head, the entire positive/negative memory-block distance is at most $Vg(r)$, regardless of its finite value widths.

For erasures, remove those at most $s = q-1$ strong heads. For corruptions, split them into two sets of at most s and corrupt one set on the positive input and the other on the negative input. The adversary can make both received memories coincide on their union. On all other heads, opposite noise displacements reduce the distance to at most $[Vg(r) - 2\eta]_+$. Comparing the two labels through an L-Lipschitz decoder proves the lower bound. This is also the order-statistic oracle argument above.

If $h < q$, erasures can remove all heads, or the two corruption sets can cover all heads. In either case the two labels become indistinguishable, forcing error at least $\Delta/2$. The constant predictor attains this value. □

Attainment on the entire source class. Assume $h \geq q$. Number types from zero. Assign type i to the q heads $(qi + \ell) \bmod h$, $0 \leq \ell < q$. These heads are distinct. Each head receives at most $\lceil qD/h \rceil$ types. Within each head distribute its types across its R reads as evenly as possible. No read receives more than $\lceil \lceil qD/h \rceil / R \rceil = \lceil qD/(hR) \rceil = r$ types. Empty reads may be ignored. Use score B for assigned types and zero otherwise, with one value $V\mathbf{1}\{x = a_i\}$ for each assigned type. This uses qD useful coordinates. Enforcing standard shared headwise value features can introduce additional unused coordinates, which are not included in that useful-width count.

For a fully covered multiset with counts c_i, every assigned coordinate for type i is

$$\frac{Vc_i e^B}{e^B \sum_{k \in S_j} c_k + \sum_{k \notin S_j} c_k} \geq \frac{V}{1 + n(|S_j| - 1) + n(D - |S_j|)e^{-B}} \geq Vg(r).$$

Only $c_k/c_i \le n$ was used. This applies to all such multisets, not only the balanced witnesses. For any omitted type, all its uncorrupted coordinates are zero, irrespective of other counts or the number of additional missing types.

For each type, aggregate its surviving copies by a maximum under erasures. At least one remains. Under hidden corruptions, aggregate its $2s+1$ copies by a median. At least $s+1$ are honest because the copies lie on distinct heads. Let $u(z)$ be the minimum of these type-level aggregates. Positive inputs satisfy $u \ge Vg(r) - \eta$, and negative inputs satisfy $u \le \eta$. The aggregate is 1-Lipschitz in the infinity norm, separately for each revealed mask. If $m = Vg(r) - 2\eta > 0$, set

$$\lambda = \min(L, \Delta/m), \qquad G(z) = \mathrm{clip}_{[0,\Delta]}\{\Delta/2 + \lambda(Vg(r)/2 - u(z))\}.$$

Its worst error is at most $[\Delta - \lambda m]/2$, equal to the claimed value. If $m \le 0$, use the constant $\Delta/2$. The lower bound gives equality. Zero L or V and zero contrast are included. □

Uniform imbalance and contrast compensation. If $hR < qD$ and $h \ge q$, then $r \ge 2$ and $g(r) \le [1 + n(r-1)]^{-1} \to 0$, even with unbounded contrast. Bounded LV therefore precludes any uniform error strictly below $\Delta/2$. If $hR \ge qD$ and $h \ge q$, all assigned reads can be singletons. For $B = \log n + b$, their smallest positive mass is $g(1) = [1 + (D-1)e^{-b}]^{-1}$, independent of n. Given $a = (\Delta/L + 2\eta)/V \in (0,1)$, choose $b \ge \max\{0, \log((D-1)a/(1-a))\}$ for zero error. The same fixed decoder then works at all n. This uses a supplied count-imbalance bound and increases the score budget. It is not unconditional extrapolation at fixed contrast or a new temperature scaling principle. The fault-free imbalance special case has $q = 1$.

E Core attainment protocols

Construction-matched calibration. The source uses the paired-code prototypes of Section 3 with $D = n = 8$, $\Delta = 1/32$, $T = 399$, $B = \log 128$, and $V = L = 1$. We test $(K, \eta) \in \{(4,0), (4,0.03), (4,0.05), (8,0.05)\}$. For each regime, the analytic encoder from Theorem 3.2 is frozen and the minimum/clip readout $\mathrm{clip}_{[0,\Delta]}\{b - \lambda \min_k(z_k - a_k)\}$ is trained for 2,000 projected Adam steps with seeds $101, 211, 307, 401, 503$. Learning rates are 10^{-2}, 10^{-3} and 10^{-4} over the first half and final two quarters. Gradient norm is capped at five. The final iterate is evaluated, with no checkpoint or seed selection. Exact box-corner extrema provide the robust errors in Table 1. All 20 fits attain the corresponding resource optimum within 0.01Δ.

Frozen-encoder decoding diagnostic. In the $K = 8, \eta = 0.05$ regime, all five jointly MSE-trained table encoders have $\mathcal{E}^*_{\mathrm{dec}}(A) = \mathcal{E}^*_{\mathrm{res}} = 0$. We freeze each encoder and its source anchors, then optimize the convex Lipschitz-envelope objective in Equation (12). The mean normalized robust upper bound is 1.45×10^{-4} (absolute error 4.53×10^{-6}). This diagnostic tests whether the learned representation is decodable by a source-adapted Lipschitz extension. It is not a held-out-label or generic-architecture result.

Evaluation and reproducibility. Training uses FP32 and evaluation uses FP64. The reported numerical bounds are not outward-rounded certificates. Result records and final checkpoints remain in the reproducibility package. Only claim-bearing protocols are summarized here. Two-step smoke runs are excluded from all formal counts.

F Additional optimization and transfer controls

The main text focuses on tests that identify or attain the theoretical frontiers. This appendix records the theorem-directed aggregation and transfer controls; no claim is made that task loss or a generic decoder automatically discovers the construction.

Whole-head intervention and aggregation controls. With $K = 24$, all tested layouts have zero clean decoder-oracle error. Under one hidden head corruption, $h < 3$ has oracle error 0.5Δ, as does the $h = 3$ control that concentrates each type's copies in one head. Separating replicas permits zero error with the matched readout. Keeping contrast fixed at $n = 128$ instead gives mean per-fit worst normalized error one. Replacing the median with a mean gives one under explicit arbitrary

Table 5: One whole-head corruption at fixed $K = 24$, $D = 8$, and $\eta = 0.05$. First two numeric columns are exact balanced-source decoder oracles at $n = 8$, $B = \log 16$. Last three report the mean of five per-fit worst errors on 1,000 new histograms, normalized by Δ. Only two readout coefficients are trained. "Comp." increases contrast by $\log(n/8)$, without retraining. A dash denotes no calibration fit. The starred control places all copies of each type in one head.

h	R	Clean oracle	Fault oracle	New $n = 8$	$n = 128$ fixed	$n = 128$ comp.
1	24	0.000	0.500	–	–	–
2	12	0.000	0.500	–	–	–
3	8	0.000	0.000	0.000	1.000	0.000
3*	8	0.000	0.500	–	–	–
6	4	0.000	0.000	0.000	1.000	0.000
12	2	0.000	0.000	0.000	1.000	0.000
24	1	0.000	0.000	0.000	1.000	0.000

corruptions at $n = 8$ on the same memory. These are checks of the theorem's necessary conditions and chosen readout, not evidence against every alternative trained decoder.

In the separately recalibrated bounded-fault control, median fits have source and held-out worst error zero in 20/20 cases. Mean fits average 0.50047 on both, against constant error 0.5. Their opposite-label aggregate uncertainty intervals overlap, giving scalar decoder oracle 0.5 in every case. Both arms clip memories to $[0, 1]$ and use the same exact bounded-fault objective.

G Reliability experiment: protocol and independent checks

Source and exact attacks. The separate study uses equidistant prototypes $a_i = \sqrt{D/64}\, e_i \in [0, 1]^D$, giving $\|a_i - a_j\|_2^2/D = \Delta = 1/32$ for the tested $D \in \{4, 8\}$. Scores are B on a read's assigned types and zero otherwise. Values are the assigned type indicators, with $V = L = 1$ throughout. Attention is computed from actual integer multiplicities and scores, not from the formula for g. A separate expanded-token computation checks multiplicity compression. No source labels enter the encoder.

Every subset of at most s faulty heads is enumerated, including the empty set. Order-statistic readouts are coordinatewise monotone, so opposite noise-box corners and extreme faulty coordinates give exact robust output extrema. Replacements $\pm 10^6$ lie outside the honest range and realize the limiting protected order statistics. For the mean ablation, they are explicit finite attacks attaining the maximum possible clipped error Δ, not a claim of an unclipped infinite-attack optimizer. The fixed-memory oracle in Equation (16) is computed independently from pairwise head-block distances.

Grid, including degenerate and failing regimes. The full Cartesian grid is $D \in \{4, 8\}$, $h \in \{1, 2, 3, 6\}$, $R \in \{1, 2, 4\}$, $n \in \{1, 8, 128\}$, $B \in \{0, \log 16, \log 128\}$, $\eta \in \{0, 0.05\}$, $s \in \{0, 1, 2\}$, and both fault models. All 2,592 configurations are retained. When $h < q$, the attaining predictor is the constant $\Delta/2$, not a successful retrieval. The maximum observed formula–oracle residual is zero and the formula–attained-error residual is 6.94×10^{-18} in FP64. An additional 48 random dense score/value encoders satisfy the theoretical lower bound. Exhaustive removal of head distances checks the order-statistic oracle against a second implementation. These checks and expanded-token attention have maximum residual 1.12×10^{-16}. None is an outward-rounded certificate.

Matched-resource intervention and fitting. The $K = 24$ study allocates three copies per type and 24 useful coordinates in every row of Table 5. For $h < 3$, some copies necessarily share a head. For $h \geq 3$, the theorem's layout is randomized by type, head and slot permutations. The score table always contains $KD = 192$ entries. This matches grouped read counts and useful coordinates, not physical shared-K/V FLOPs. The additional starred placement control was added after the initial grid check. It assigns type i only to head $i \bmod 3$, cycling its three copies across that head's slots. It uses the same counts and budgets, and its clean/fault oracles are computed from actual memories. No readout is fitted for this control. Only b and λ are learned in $G(z) = \mathrm{clip}_{[0,\Delta]}(b - \lambda u(z))$ with $0 \leq \lambda \leq 1$. The readout uses the min of typewise medians. Training data are the complete

16-example source at $n = 8$. Let u_a^-, u_a^+ be its exact attacked extrema. The convex objective is

$$\max\{0,\ b - \lambda \min_{a:y_a=0} u_a^-,\ \Delta - b + \lambda \max_{a:y_a=\Delta} u_a^+\}.$$

It upper-bounds clipped worst error and uses no resource-oracle values. Seeds $101, 211, 307, 401, 503$ initialize $b \sim \mathcal{U}(0, \Delta)$ and $\lambda \sim \mathcal{U}(0.02, 0.1)$. Adam uses learning rates 0.005, 0.001, 0.0001 over the first half, next quarter and last quarter of 2,000 updates, with standard $(0.9, 0.999)$ moments and 10^{-8} denominator offset. The slope is projected after every update. Final iterates are used, with no seed selection. The analytic encoder and simple two-parameter calibration make this substantially easier than learning a Transformer.

Held-out multiplicities and ablations. For each seed and each $n \in \{8, 32, 128\}$, generate 500 fully covered histograms with independent counts uniform on $\{1, \dots, n\}$, and 500 incomplete histograms. For the latter, draw the number of missing types uniformly from $1, \dots, D-1$, choose their identities uniformly without replacement, and set their counts to zero. No resulting normalized histogram equals a training example. All layouts and contrast ablations share these inputs: 15,000 generated test instances, reused across 100,000 layout/contrast evaluations. They are not 100,000 independent samples. Tests enumerate all allowed head faults and continuous noise extrema, not merely one sampled perturbation. The five seeds vary initialization, layout and test inputs, not five independent large-model training runs. Zero reported errors are FP64 values, not an empirical claim about every possible real-world input.

At larger imbalance, contrast compensation changes only $B(n) = \log 16 + \log(n/8)$. It receives n as an externally specified test-class bound, not an estimate inferred without additional cost. The predictor's learned coefficients, indicators and type assignment remain fixed. The fixed-contrast and mean-aggregation ablations are retained. The latter changes the aggregate without retraining and isolates susceptibility to unbounded replacement, not an optimized mean-model comparison. Replication and the min/median architecture are theory-guided choices, not discoveries from a blind search.

Artifacts. `reliability/run_reliability.py` contains the full NumPy-only CPU experiment. Its smoke mode is separate from the formal run. All grid outputs, 20 final calibration states and histories, layouts, held-out histograms, ablations, source hash and environment versions are retained. The formal run uses Python 3.12.14 and NumPy 2.3.5. This small experiment requires no GPU.

H Bounded-fault aggregation control

This separate 40-fit study retains the original constructed encoder, $D = 8$, $K = 24$, $n = 8$, $B = \log 16$, $\eta = 0.05$ and $h \in \{3, 6, 12, 24\}$. Whole-head replacements have each coordinate in $[0, 1]$; honest coordinates receive bounded noise, then both arms clip memories to $[0, 1]$. Median and mean arms independently train their two affine-readout coefficients for 2,000 updates with the original initialization seeds and schedule. Monotonicity gives exact extrema over all head positions and continuous boxes for both arms. Each fit tests on 1,000 new histograms of the same known types at $n = 8$; layouts and arms share inputs within each seed. This makes 5,000 distinct generated histograms reused for 40,000 evaluations, not 40,000 independent samples. Scalar-interval decoder oracles check whether information has been lost by aggregation itself. The original unbounded mean ablation is retained, not substituted.

I Stability of the decoder diagnostic

Consider two labeled memory collections (z_a, y_a) and $(\widetilde{z}_a, \widetilde{y}_a)$ with the same source indexing, coordinate grouping, noise radii, Lipschitz constant and fault budget. Suppose $\max_a \|z_a - \widetilde{z}_a\|_\infty \le \delta$ and $\max_a |y_a - \widetilde{y}_a| \le \xi$. For the box decoder oracle of Proposition 4.1, or the arbitrary-head-replacement oracle in the reliability proof,

$$\left|\mathcal{E}^*_{\mathrm{dec}}(z, y) - \mathcal{E}^*_{\mathrm{dec}}(\widetilde{z}, \widetilde{y})\right| \le L\delta + \xi. \tag{17}$$

This is a stability consequence of the oracle, not a new extension theorem or an end-to-end learning guarantee.

Proof. Each coordinate difference between two memories changes by at most 2δ. Subtraction of the unchanged noise radius, positive part, coordinate maximum, and a fixed order statistic are all nonexpansive in the maximum norm. Thus each pairwise noise-box distance, including the $(2s+1)$st largest head distance for hidden replacements, changes by at most 2δ. When there are fewer than $2s+1$ heads, both effective distances are zero. The absolute label difference changes by at most 2ξ. Applying the positive part and maximum over pairs in the oracle formula and then dividing by two gives the stated bound. □

An empirically measured encoder discrepancy therefore consumes a quantifiable error budget. A large or unmeasured discrepancy cannot justify transferring the constructed encoder's zero-error claim. The accompanying deterministic perturbation checks test this formula with nonzero oracle errors and a tight two-point example. They are implementation checks, not a statistical generalization result.

J From a learned token map to the robust construction

Let the teacher scores s^*_{hi} expose one assigned prototype per read at contrast B, with three copies per type on distinct heads. Its shared head value map sends prototype a_i to e_i. For a learned shared-K/V encoder on the same prototypes, define

$$\epsilon_v = \max_{h,i} \|v_h(a_i) - e_i\|_\infty, \qquad \rho = \max_{h,r} \left[\max_i (s_{hr}(a_i) - s^*_{hr}(a_i)) - \min_i (s_{hr}(a_i) - s^*_{hr}(a_i))\right].$$

Then, for every nonempty prototype multiset, regardless of its length or multiplicities,

$$\|A(X) - A^*(X)\|_\infty \le \delta_{\mathrm{tab}} := \epsilon_v + \min\{1, \rho/4\}. \tag{18}$$

To see this, first hold the attention distribution fixed and replace the value table, costing at most ϵ_v. Interpolate the score table linearly. The derivative of any teacher-value coordinate's pooled mean is its covariance with the score error. Cauchy–Schwarz and the variance bound for a bounded random variable give an absolute derivative at most $\rho/4$, since the teacher-value range is at most one. Integrating gives the result, also capped by one. Zero-count types are omitted, so the same proof applies to incomplete multisets.

For full coverage with count ratio at most n, each assigned teacher coordinate is at least $g = [1 + n(D-1)e^{-B}]^{-1}$. A missing type instead has all three teacher copies zero. Under one arbitrary head replacement and honest-coordinate noise of radius η, each median remains between honest values. The minimum across type medians therefore separates the two cases by at least $g - 2\eta - 2\delta_{\mathrm{tab}}$. The fixed globally one-Lipschitz readout

$$G(z) = \mathrm{clip}_{[0,\Delta]}\left(\frac{g+\Delta}{2} - \min_i \mathrm{median}^3_{\ell=1}\, z_{i\ell}\right)$$

has zero worst error whenever

$$\delta_{\mathrm{tab}} \le \frac{g - 2\eta - \Delta}{2}. \tag{19}$$

At $D = 8, n = 8, B = \log 16, \eta = .05, \Delta = 1/32$, this threshold is approximately 0.0454861. This is a sufficient condition on the learned token maps, not an optimization convergence guarantee.

Why the shared-K/V class can contain such a representation. Let the augmented frozen token-feature matrix on the D prototypes have row rank D, and key width d_k and value width be at least D. Linear least squares can then fit head keys $k_h(a_i) = e_i$ and value logits $M(2e_i - \mathbf{1})$. Set each assigned query's first D coordinates to $\sqrt{d_k}\, M(2e_j - \mathbf{1})$ and pad unused coordinates with zero. With scaled dot-product keys, sigmoid- bounded scores and sigmoid values, both desired tables are approximated with error at most a constant times e^{-M}. Thus every positive margin in Equation (19) can be met with finite parameters. No parameter-norm restriction is claimed, and this construction uses prototype identities and a preassigned replication pattern. It is supervised finite-type representability, not evidence that task-only gradient descent discovers those identities or assignments.

Nearby continuous tokens. The same argument applies if the score/value bounds hold uniformly on the ℓ_∞ boxes of radius ε around each prototype. For equidistant prototypes of normalized squared separation Δ, perturbing a token by at most ε per coordinate changes its normalized squared distance to any prototype by at most $\xi = 2\sqrt{\Delta}\,\varepsilon + \varepsilon^2$. The minimum over present tokens and maximum over target prototypes are nonexpansive. Thus true continuous coverage differs from the binary type-coverage label by at most ξ. If the token-box version of Equation (19) holds, the fixed readout's continuous coverage error is at most ξ, under the same noise and head-fault model. This is local input transfer, not transfer to new semantic types. Interval propagation can check a sufficient token-map bound numerically without using test labels for fitting.

K Construction-supervised representation learning

All runs use a fixed slope-one min-of-medians readout, three copies of each type on distinct heads, and a shared-K/V encoder whose features, keys, values and queries are trained. Teacher targets specify the assigned type in the scores and prototype indicators in the values. This is structural supervision, not task-label-only discovery.

Table 6: Construction-supervised learning at $D = 8$. Each row contains 30 final fits: two layouts, three independently trained prototype geometries and five initialization seeds. Errors are normalized by Δ.

Training	Zero robust error	Mean robust error	Uniform condition
Full teacher	30/30	0.0000	30/30
Teacher then task	30/30	0.0000	30/30
Gradient-balanced fade	30/30	0.0000	30/30

Teacher-then-task uses 1,000 teacher updates followed by 1,000 robust-task updates. Gradient-balanced fading initializes the teacher coefficient from the ratio of the two full-encoder gradient norms and reduces it linearly to zero during the first half. Every fit uses Adam with learning rates $0.01, 0.002, 0.0004$ over the first half and final two quarters, gradient norm cap five, and the final iterate rather than a selected checkpoint. Seeds are $101, 211, 307, 401, 503$ and geometry seeds are $2026091751, 2026091773, 2026091781$. No test outcome selects a run.

Each model is tested on 1,000 complete or incomplete histograms at $n = 8, 32, 128$. At larger n, the reported compensated condition uses $B(n) = \log 16 + \log(n/8)$ and therefore is not fixed-budget extrapolation. A coordinate-box check at radius 0.005 also leaves all 90 fits within the sufficient condition. The worst sampled continuous error is 0.03515Δ, below the corresponding label-shift bound 0.05737Δ. NumPy replay of saved checkpoints reconstructs the encoder, attention, readout extrema and fault oracle independently, with maximum metric residual 2.23×10^{-15}.

The claim is conditional construction-supervised representability and learning, not general task-only discovery. Full run records remain in the reproducibility package.

L A direct multiplicity certificate for learned tables

This diagnostic applies to the fixed replicated readout, without requiring the learned tables to approximate prescribed indicator tables. For a nonnegative integer s and each type i, fix $q = 2s + 1$ assigned reads on distinct heads. An assigned read ℓ has scores $s_{\ell t}$ and scalar values $v_{\ell t} \in [0, 1]$ on the D prototype types. The value is the coordinate consumed by this type's replica statistic. Scores and value tables may otherwise be arbitrary. Write $i(\ell)$ for its assigned type and $w_{\ell t} = e^{s_{\ell t}} > 0$. The predictor is $\mathrm{clip}_{[0,\Delta]}(b - \min_i \mathrm{med}_{\ell : i(\ell) = i} z_\ell)$, with at most s hidden whole-head corruptions and honest-coordinate noise bounded by η. This readout is globally 1-Lipschitz in the infinity norm. The prototypes remain equidistant.

Proposition L.1 (Learned-table count certificate)**.** *For a positive integer n, define*

$$a_\ell(n) = \min_{c\in[1,n]^D} \frac{\sum_t c_t w_{\ell t} v_{\ell t}}{\sum_t c_t w_{\ell t}}, \qquad A(n) = \min_\ell a_\ell(n), \tag{20}$$

$$U = \max_\ell \max_{t\neq i(\ell)} v_{\ell t}. \tag{21}$$

The worst error on $\mathcal{C}_n$ is at most

$$\max\{\mathrm{clip}_{[0,\Delta]}(b+\eta-A(n)), \mathrm{clip}_{[0,\Delta]}(\Delta-b+\eta+U)\}. \tag{22}$$

Each $a_\ell(n)$ is computed by sorting the D values increasingly and checking the $D+1$ count vectors that give count n to a prefix and count 1 to its complement. The first term in Equation (22) *is the exact worst error over complete multisets under this fixed readout. The incomplete-input term is a sufficient upper bound, not a general exact formula.*

Proof. Positive count vectors with ratio at most n can be scaled into $[1,n]^D$ without changing attention. Conversely the specified vertices are integer multisets, so any minimum attained there is realizable. For one read, let μ be the minimum ratio on the compact count box. Every feasible c satisfies $\sum_t c_t w_t(v_t-\mu)\geq 0$, with equality at a minimizer. Minimizing this linear expression sets $c_t = n$ when $v_t < \mu$ and $c_t = 1$ when $v_t > \mu$. Ties have zero coefficient. Consequently a sorted-prefix vertex also attains μ. This is an elementary linear-fractional optimization fact, not a new general optimization theorem.

On complete inputs every honest selected coordinate is at least $A(n)-\eta$. A median of $2s+1$ copies has at least $s+1$ honest copies, so the minimum of medians is also at least this quantity. The first error bound follows. It is attained by a read and count vertex attaining $A(n)$: perturb its honest coordinate downward by η and corrupt s other replica heads to sufficiently negative values. The target median equals $A(n)-\eta$, and no other median can be smaller than this global bound. Clipping preserves attainment. For $s=0$ the same statement holds with the single copy and no corruption.

On incomplete inputs choose any absent type i. Every honest copy for that type averages only values at most U, regardless of the remaining counts or scores. At least $s+1$ such copies are at most $U+\eta$, so their median, and hence the minimum across types, is at most $U+\eta$. The label is Δ, giving the second bound. This argument also covers arbitrarily imbalanced incomplete multisets. □

In particular, $A(n)\geq b+\eta$ and $U\leq b-\Delta-\eta$ suffice for zero robust error. Unlike the token-map approximation condition, this certificate measures the actual learned tables and does not specify teacher score targets. It still uses prototype identities, a known imbalance budget, replica assignments and a matched readout. It does not establish unrestricted task-only learning or convergence of SGD. Evaluating all qD copies costs $O(qD^2\log D)$ arithmetic operations.

Strictly larger sufficient region than table approximation. For the three-copy teacher with mass $g(1)$ and $b=(g(1)+\Delta)/2$, the earlier condition $\delta_{\mathrm{tab}}\leq (g(1)-2\eta-\Delta)/2$ implies this certificate. Indeed, its multiplicity-uniform memory bound gives $A(n)\geq g(1)-\delta_{\mathrm{tab}}$, while every off-type value is at most $\epsilon_v\leq\delta_{\mathrm{tab}}$. The inclusion can be strict. At the experiment's budgets, $g(1)=2/9$, $\eta=1/20$ and $\Delta=1/32$. Keep the teacher scores but scale its indicator values by $9/10$. Then $A(8)=1/5>b+\eta$ and $U=0<b-\Delta-\eta$, so the direct certificate gives zero error. In contrast, $\delta_{\mathrm{tab}}=1/10>131/2880$, so the table-approximation test does not certify this representation. This does not contradict the earlier sufficient condition or imply that either test is necessary.

M DIRECT-MARGIN CONTINUATION PROTOCOL

We test whether the multiplicity certificate can replace continued regression to teacher tables after a short construction-supervised prefix. The frozen setting is $D=8$, $h=3$, $R=8$, $K=24$, $B=\log 16$, $\eta=0.05$, $\Delta=1/32$ and $n=8$, with the fixed min-of-three-medians readout. For each assigned replica ℓ, the objective penalizes violations of

$$A_\ell(8)\geq b+\eta+\tau, \qquad U_\ell\leq b-\Delta-\eta-\tau,$$

Table 7: Fresh 100-step-prefix confirmation at $n = 8$. Each cell contains five paired blocks. The Direct column counts models satisfying Equation (14). Fresh zero uses 1,000 unseen count vectors per model.

D	Prefix	Continuation	Source zero	Fresh zero	Direct
8	100	Direct margin	5/5	5/5	5/5
8	100	Continued teacher	5/5	5/5	5/5
16	100	Direct margin	5/5	5/5	5/5
16	100	Continued teacher	5/5	5/5	5/5
32	100	Direct margin	5/5	5/5	5/5
32	100	Continued teacher	5/5	5/5	5/5

where $\tau = 0.01\Delta$, $A_\ell(8)$ is computed by the sorted-prefix formula of Proposition L.1, and U_ℓ is the largest off-type value. It uses type identities and declared budgets but no teacher score or value targets.

Fifteen shared prefix states come from 100 teacher updates over three geometries and five initialization seeds. Every prefix has nonzero source robust error and violates Equation (14). Adam is reset and each model continues to 2,000 total updates. All 15 direct-margin continuations meet the direct condition and have zero worst error on 1,000 fresh $n = 8$ count vectors. Their earlier table-approximation bounds remain above the sufficient threshold, so the direct certificate covers representations that the table-neighborhood condition does not. Continued teacher supervision also succeeds in all 15 pairs. Therefore the result establishes an alternative certified continuation route, not an accuracy improvement over the teacher. A finite-source task-loss control reaches zero on the source but only 3/15 fresh-count tests, separating count-class coverage from source fitting.

An independent NumPy implementation replays the 90 formal checkpoints, attention tables and exact noise/fault extrema, and enumerates all 2^8 count-box vertices for the direct floor. Across 497,340 source evaluations, the maximum residual is 3.56×10^{-15}. Numerical neural checks use ordinary FP64 rather than outward-rounded certification. Full run records remain archived. Only the frozen continuation protocol supports the stated claim.

N FRESH-BLOCK CONFIRMATION OF CERTIFIED CONTINUATION

A separately frozen confirmation uses five new paired (initialization, geometry) blocks at each $D \in \{8, 16, 32\}$, with a 100-update teacher prefix and 2,000 total updates. Resources scale as $K = 3D$, key/value width D, and $B_D = \log 16 + \log((D-1)/7)$, so $g(1) = 2/9$ at $n = 8$. All arms copy the same saved prefix and reset Adam. Formal test counts are shared within each paired block, and no formal outcome selects a seed, checkpoint or hyperparameter.

Every one of the 15 frozen prefixes has nonzero source error, and every final direct model lies outside the older table-bound sufficient neighborhood. Hence success is recovery by the direct margins rather than retention of an already feasible state. The equally successful teacher continuation prevents an accuracy-superiority claim. A paired finite-source task-loss control reaches fresh-count zero in 10/15 fits and the direct condition in 0/15, supporting the narrower claim that explicit count-class margins improve certificate attainment in this protocol.

The certificate is for $\mathcal{C}_8$. With fixed B_D, none of the final arms establishes extrapolation to the larger imbalance values $n = 32, 128$. Contrast-compensated checks change the score budget and are reported only in the archived records. A separate NumPy evaluator replays all final models and prefix states using a bisection/count-witness solver rather than the training implementation. Across 1,700 metric comparisons and 534,600 model–source evaluations, its maximum residual is 7.11×10^{-15}. Full run records remain in the reproducibility package. Only the frozen 100-step protocol is used as manuscript evidence.

O FROZEN-QWEN VARIABLE-TRACKING PROTOCOL

Features and access restrictions. We freeze `Qwen/Qwen3-4B` at revision `1cfa9a7208912126459214e8b04321603b3df60c`, using BF16, SDPA and a seed-271828 Gaussian projection with normalized columns to 512 dimensions. Candidates are the last tokens overlapping complete `VAR lhs = rhs` assignments. Parsing uses input text, not reference answers. The main query is a separate model forward on the explicit question alone. The contextual query ablation uses the queried-value state in the complete prompt. The input-candidate control changes only candidate states to input embeddings.

A genuine bounded value interface. After candidate permutation, address i has value $e_i \in \mathbb{R}^{10}$. Learned 64-dimensional normalized keys, query and slot offsets give scores in $[-6, 6]$. For $z_j = \sum_i p_{ji} e_i$, the only decoder input is $Z = (z_1, \ldots, z_K)$ and its logits are

$$\ell_i = 20 \left(K^{-1} \sum_{j=1}^{K} z_{ji} - 1/10 \right).$$

This fixed decoder is 20-Lipschitz in the coordinate infinity norm, and its sigmoid outputs are 5-Lipschitz. Hard set decisions are not Lipschitz. The primary prediction uses $\ell_i \geq 0$, independently for ten candidates. The secondary rule selects five logits for every K, including $K = 1$. No raw-score, query or key bypass enters decoding. One-hot values retain multiple candidate coordinates, so no one-answer-per-read restriction is imposed. All 252 five-of-ten oracle sets pass the five-vector decoder check, and a single legal mixture can also encode five targets.

Training and source coordinates. The three arms use $K \in \{1, 3, 4, 5, 6, 8\}$ and seeds 11, 29 and 47, giving 54 fits. Each fit uses 1,500 AdamW updates, batch size 16, learning rate 10^{-3}, weight decay 10^{-3} and gradient clipping at 5. Binary cross-entropy is applied only after value pooling. The score encoder has 68,096 parameters at every K and the decoder has none. The 4K training ordinals are 300–379, validation is 380–399, and the 16K/32K test ordinals are 460–499 at each length. Ordinals are zero-based. Training randomly permutes candidates each step. Final evaluation uses a deterministic permutation seeded only by task, length and ordinal, never reference labels. No formal-outcome success gate or final-checkpoint selection is used.

Complete outcomes and controls. All six K values give Qwen thresholded exact recovery 98.75%, recall 99.75% and top-5 exact recovery 100%, for both query variants. Input candidates give thresholded exact recovery 0%, recall 34.08–35.50%, and top-5 exact recovery 1.25%. For each configuration there are 80 test-source clusters, each evaluated for three seeds. Intervals use 2,000 bootstrap resamples of source means, conditional on the trained seeds. An all-correct empirical bootstrap interval is degenerate and is not certainty about population accuracy. The thresholded Qwen error occurs on the same 32K source, ordinal 485, for all seeds, with one missing target. Top-5 recovers it.

The zero-memory control has zero thresholded recall. Fixed-feature dilution adds 32 or 256 identical score-zero, value-zero terms to each softmax pool. For question-only Qwen, 32 copies retain 98.75% exact recovery, while 256 give 97.08–98.33% across K, with no monotone read-count advantage. These are not new long-text forwards. In particular, for $K = 1$, dilution uniformly rescales the candidate weights and cannot change top-5 ranking. Across 100 validation/test sources per fit, four conditions and two rules, all 43,200 records are retained and their set metrics independently recomputed.

Audit and interpretation. An independent audit found that an earlier evaluation permutation seed used a full-record hash containing the answer field. The reported evaluation therefore re-scores every unchanged checkpoint with coordinate-only seeds and independently reproduces the set metrics. Training permutations are label-independent, and the retained artifacts include checkpoint hashes and both evaluation records. Contextual candidate features and preprocessing remain outside the static token-local model. Five VT answers do not identify the theorem's prototype count D, so this experiment tests finite-task bounded readout and cross-length transfer rather than the uniform lower bound.

P NATIVE QWEN HEAD-OUTPUT INTERVENTION

Frozen site and independent panels. The model and revision match Appendix O. A disjoint V7 calibration selects zero-indexed layer 29, suffix-token positions and heads 11, 6 and 5. The two reported panels reuse that site and identical intervention logic without head, layer or seed reselection. V8 uses 64 fresh 16K sources and V9 uses 64 fresh 32K sources, with no source overlap with calibration or prior formal panels. The task is RULER-derived direct retrieval rather than an official RULER evaluation.

Matched intervention. For each source we construct three concentrated and three dispersed rotations. Every mask selects eight token–head vectors, and aggregate head–position marginals match between families. Both arms use complete prefill and replace selected native head outputs with donor-derived changes. Scaling is chosen from the realized BF16 output projection, not an FP32 linear proxy. The maximum relative mismatch in the realized projection-output norm is 0.0484% at 16K and 0.0504% at 32K. The pre-declared endpoint is the source-paired difference in answer-versus-alternative log-probability margin. Exact-set generation is reported beside it rather than used to select a configuration.

At 16K, dispersed minus concentrated margin is $+0.016929$ with 95% interval $[0.007944, 0.026826]$; the exact-set difference is -0.5208 percentage points with interval $[-1.5625, 0]$. At 32K, the margin contrast is $+0.022244$ with interval $[0.013834, 0.031248]$, while both arms are correct on all 64 sources. The post-minus-pre normalization contrasts contain zero in both panels, so the data do not identify how much of the remaining effect is attributable to the matching rule.

Controls and scope. Sham logits and generations are identical to clean runs. Input-suffix replacement follows the donor output, and native output projection is independently reproduced. The selected-three-head and all-head replacements also verify that the intervention can alter generation, although they are not the matched primary comparison. Complete code, configurations, source records, all 18 conditions, server and local analyses, and SHA256 receipts are retained for both panels. These checks support a repeatable small margin sensitivity. They do not establish accuracy improvement, a unique causal role for grouping, head compression, or transfer of the static-reader threshold to a Transformer with residual bypass and contextual token interactions.

Q WHAT THE RESOURCE LAWS ESTABLISH AND WHAT LEARNING USES

The resource separation. Table 8 contrasts the objects actually controlled by nearby results. Our lower bound permits arbitrary tokenwise scores and finite value widths but restricts stability and forbids positional or contextual input paths. The upper construction supplies queries and pooled state. Hence $K = hR$ identifies the number of reads, not an equivalence at fixed parameters, FLOPs or KV-cache size. Likewise the fault theorem is for the declared grouped-read model, not automatically for every parameter-tied implementation of multi-head attention.

Breadth and limits of the comparison. The cited head-count, rank, graph-selection and set-encoding results constrain different model classes and tasks; none is claimed as a special case of our theorems. Our qualified priority claim concerns their joint counterpart here: a sharp read-count substitution threshold for quadratic coverage together with exact finite-budget grouped-fault frontiers. The generality is over admissible static score/value maps, finite value widths and Lipschitz decoders, not over arbitrary contextual Transformers or retrieval objectives.

The advance is a sharp characterization. Theorem 3.1 supplies a matching threshold, not only an impossibility example. Theorem 3.2 determines the exact finite-source optimum across read allocations and contrast/noise budgets. Theorem 3.3 extends the minimax characterization to an infinite multiset class and proves that read capacity cannot substitute for independent fault domains. The classical $2s + 1$ threshold alone leaves this contrast-dependent retrieval frontier undetermined. The count criterion further gives a polynomial-time test and learning target over an infinite input class, with a sufficient region strictly larger than that of Equation (13). Indistinguishability, Lipschitz extension and linear-fractional optimization supply established proof ingredients. The new claims are

Table 8: How the present resource laws extend the theoretical picture. These comparisons do not assert that either model class subsumes the other.

Reference	Controlled object or established ingredient	Present characterization
Yu et al. (2026), Sections 3–5	Single-query-per-head approximation and parameter complexity under embedding/readout constraints.	An exact stable-coverage threshold counting all K static reads, with LV as the resource budget rather than parameter count.
Lee et al. (2019), Section 3.2	Attention pooling with multiple learned seed queries.	Sharp conditions under which query slots provide coverage capacity and under which independent heads remain indispensable.
Amsel et al. (2025)	Head-rank limits for nearest-neighbor retrieval, even with many low-rank heads.	Read-count and grouped-fault frontiers at a controlled decoder–value scale.
Tesfaye et al. (2026)	One-versus-two-head separation for attention-only endpoint selection, independent of dimension and precision.	A head–slot substitution law for coverage with joint nonlinear stable decoding and grouped faults.
Wagstaff et al. (2022)	Latent-dimension requirements for universal Deep Sets approximation.	Read and stability requirements for coverage that persist even with arbitrary finite value width.
Veličković et al. (2025), Lemma 2.1	Dispersion as the number of logits grows at bounded logit spread.	A read threshold allowing length-dependent scores and a separate finite-budget minimax law quantifying the exact robust-error cost.
Chen et al. (2018), Theorem 3.1	The classical $2s + 1$ redundancy requirement for Byzantine recovery.	A grouped-attention minimax law integrating read allocation, finite contrast, honest noise and Lipschitz decoding.

the stated resource laws and attention-specific characterization, not general versions of those tools or dominance over every earlier certificate.

The learning question and its conditioning information. Table 9 separates fixed information, trainable parameters and information used only for evaluation. Prototype coordinates are known training inputs. Their identities specify teacher targets or the assigned coordinates used by the certificate objective. Fresh geometries are trained separately, not evaluated by a single unchanged model. A test count vector changes multiplicities or omits types from that model's known vocabulary. This design addresses robust continuation within a supplied representation layout, not learning a latent vocabulary.

Supervised scale reference. Table 10 summarizes the earlier supervised approximation panel, distinct from short-prefix confirmation. Complete construction-supervised resource counts are in Appendix F.

What the existing controls identify. Frozen prefixes distinguish recovery from retaining an already feasible representation. Direct continuation establishes that continued regression to the teacher tables is unnecessary in the tested regime. Equally successful continued-teacher models prevent an accuracy-superiority claim. Task-only source fits and fresh-count tests distinguish finite-source fitting from multiplicity robustness, but the objectives have different coverage: direct constraints address an entire count class, whereas task loss sees the enumerated source. The existing augmentation study trains generic readouts on frozen encoders and does not remove this confound for matched encoder continuation. Isolating that question would require a separate diverse-count task-continuation control. No such new experiment is claimed here.

The retained controls delimit the learned claim. An inference-time teacher is absent, but this does not make training label-only or the architecture prior-free. Removing known types, learning replica assignments or learning a generic readout would change the learning problem and remains unestablished.

Table 9: Structural-prior accounting for the matched continuation study. The teacher prefix and readout layout are shared by all compared arms.

Component	Supplied or optimized during training	Used at inference
Prototype coordinates/types	Known training vocabulary. Type identities index supervision and certificate constraints.	Token coordinates enter the encoder. The learned parameters retain type-specific structure.
Replicas and readout	Type-to-replica assignment, head grouping, min-of-medians rule and bias are fixed, not trained.	The same assignment and aggregation are used.
Encoder	Features, shared keys/values and query vectors are optimized. All branches copy the same saved prefix and reset Adam.	Final parameters only, without teacher targets or test labels.
Continuation objective	Teacher: table targets. Direct: worst-count margins using type identities and declared budgets. Task: robust loss on the finite balanced source.	No continuation loss is evaluated.
Test budgets and inputs	Fresh count vectors are evaluation inputs, not selected training examples.	Fixed-B and externally compensated-B conditions are reported separately. Compensation uses the specified n.

Table 10: Construction-supervised protocols: source robust zero and the uniform sufficient condition, at $n = 8$. Each cell includes all 30 fits (two layouts, three geometries, five seeds). $D = 8$ is the earlier reference. Resources grow as $K = 3D$ and $B_D = \log 16 + \log((D-1)/7)$.

Training	$D = 8$	$D = 16$	$D = 32$
Full teacher	30/30	30/30	30/30
Balanced fade	30/30	30/30	30/30