CERTIFYING FACT-SPECIFIC INFORMATION SURPLUS IN NOISY TRANSFORMER PARAMETER GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

How is information about a learned fact distributed across the parameter groups of a transformer? We study this question using small transformers trained on synthetic data. Within each experimental repetition, models are trained on all possible assignments of a fixed set of binary facts, with shared initialization and data-order randomness across assignments. We model observations of saved parameter groups by adding independent Gaussian noise, which specifies measurement precision. We derive a computable lower bound on joint information surplus: the mutual information about an individual fact in a pair of groups observed together, minus the sum of the mutual informations in their separate observations. The method combines entropies of Gaussian distributions matched to the means and covariances of the observed states with analytic Rényi divergence corrections. We establish a sufficient condition, expressed through covariances between parameter groups, for detecting positive surplus when differences between trained states are small relative to observation noise. Evaluation using Gram matrices of parameter states and rigorous interval arithmetic certifies positivity without Monte Carlo integration. In a confirmatory experiment with a protocol fixed in advance, the method establishes small positive lower bounds across independent training repetitions, whereas the tested pairwise entropy and transport bounds do not certify positivity on the same data. Thus, for the studied source and noise level, jointly observing embeddings and the output group comprising final normalization and the answer head provides more information about an individual fact than the sum available from their separate observations. This result shows that characterizing distributed factual information at a specified observation precision requires accounting for dependencies between parameter groups.

1 INTRODUCTION

Language models can answer factual questions using information acquired during training and retained in their parameters (Petroni et al., 2019; Roberts et al., 2020). Prior work typically evaluates this *parametric knowledge* through behavioral recall or downstream use. We instead ask whether information about a controlled training association is present in the dependence between two predefined parameter groups. Specifically, can their joint observation reveal more about the association than the sum of their separate observations?

We formalize this question using a controlled finite distribution over training-label assignments and an explicit Gaussian observation channel. A small causal transformer learns three symbolic query–label associations. Within each replicate, we train one model on each of the eight label assignments while reusing the same initialization and presentation order. The resulting eight states define a finite source distribution, allowing us to condition on one target association while averaging over the other two. We observe two parameter groups from the same sampled state through independent Gaussian noise. The noise limits which parameter differences can be resolved and is not applied during the model's forward computation. We define the *fact-specific information surplus* as the mutual information available from the joint observation minus the sum available from the two observations separately.

Certifying the sign of this surplus is difficult because approximation errors in the separate mixture-entropy terms need not cancel and may exceed the surplus itself. We derive a lower bound $B_F \leq \Delta_F$ using moment-matched Gaussian entropies and analytic Rényi corrections. For a fixed finite source

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.

at low signal-to-noise ratio (SNR) t, the gap satisfies $0 \leq \Delta_F - B_F = O(t^3)$, while the leading interaction can be of order t^2. All required quantities can be evaluated from Gram matrices indexed by source states, so their matrix dimension depends on the number of states rather than the number of parameters. Rational interval arithmetic then verifies the sign.

After selecting the embedding/output-group pair, the third query's binary target, noise scale $\rho = 8$, and threshold on development data, we evaluate the configuration unchanged on 20 new training replicates. All 20 yield positive certificates, and 19 exceed the prespecified 10^{-4}-bit threshold. Neither the tested pairwise-entropy bound nor the transport/Fano bound yields a positive certificate for any of the same channels. These results provide a finite-source example of positive information surplus under the specified observation channel.

2 Related work

Factual probing and closed-book question answering assess associations available from model parameters (Petroni et al., 2019; Roberts et al., 2020). Controlled capacity studies quantify retained factual information in bits (Allen-Zhu & Li, 2025). Recall, flexible use, and editing address different properties: learning an association need not support its reversal (Berglund et al., 2024), and causal localization need not identify the best editing site (Hase et al., 2023). We ask how information about a specified association is distributed across observations of parameter groups. Appendix F develops these connections.

Information retained in learned weights has also been studied through generalization and invariance (Xu & Raginsky, 2017; Achille & Soatto, 2018; Achille et al., 2019). Fan & Mysore (2026) study individual and joint synaptic information in Hebbian networks, providing the closest precedent for measuring information carried jointly by multiple parameter subsets. Our setting differs in using an enumerated ensemble of trained transformer states and an explicit observation channel.

The signed difference between joint and separate information is negative co-information. In a nonnegative bivariate partial information decomposition (PID), it equals synergy minus redundancy (Williams & Beer, 2010; Bertschinger et al., 2014). Our task is to certify that this difference is positive for a given observation channel.

Gaussian-mixture entropy can be bounded using variational, pairwise, transport, and Taylor constructions (Hershey & Olsen, 2007; Huber et al., 2008; Kolchinsky & Tracey, 2018; Nguyen, 2013; Nielsen & Sun, 2020; Joudeh & Škorić, 2025). Our derivation uses Gaussian-product identities (Kitagawa, 2020) to evaluate the bound, Rényi monotonicity (van Erven & Harremoës, 2014) to control entropy deficits, and low-SNR Gaussian-channel expansions (Prelov & Verdú, 2004; Guo et al., 2005; Mahuas et al., 2023) to characterize its asymptotic accuracy. This combination yields a computable lower bound with third-order error for fixed finite sources while preserving the leading second-order interaction. Outward-rounded interval arithmetic then certifies its sign for the studied transformer states.

3 Problem formulation

Training facts, model state, and a target fact. We use *training fact* to denote the query–label association for one fixed symbolic query. Each of the three symbolic queries has two possible labels: x_1 and x_3 use $\{1, 2\}$, whereas x_2 uses $\{5, 6\}$. For $u = (u_1, u_2, u_3) \in \{1, 2\}^3$, the dataset is

$$D(u) = \{(x_1, u_1), (x_2, 4 + u_2), (x_3, u_3)\}.$$

For example, $D(1, 1, 1)$ assigns answers $(1, 5, 1)$, while $D(1, 1, 2)$ assigns $(1, 5, 2)$: only the third answer changes. Each input contains only the symbolic query; neither its assigned label nor u is provided to the model. The labels define a controlled association task; the model is trained from scratch, without natural-language pretraining.

Fix the task and training specification $\mathcal{C}$ and a realization r of initialization and data-order randomness. One replicate trains all eight datasets with that same randomness:

$$\theta_r(u) = \mathcal{A}(D(u); r), \qquad m_i(u) = \pi_i \theta_r(u) \in \mathbb{R}^{d_i}. \tag{1}$$

The state $\theta_r(u)$ contains learned parameters; π_i selects a predefined group of coordinates. Let $U \sim \mathrm{Unif}(\{1,2\}^3)$ select a row of this state table and let $F = f(U)$ be a balanced binary target. The confirmatory target is $F = U_3 - 1$, distinct from the model's eight-class output: $H(U) = 3$ bits and $H(F) = 1$ bit. All information quantities are conditioned on fixed $(r, \mathcal{C})$, suppressed below. An independent draw of r repeats the entire eight-model construction; these replicates are the statistical units.

Information retained and information observed. Under the uniform source distribution, $I(F; m_i(U))$ measures information about the target in the unnoised parameter group. We instead specify an observation resolution:

$$\bar{m}_i = \tfrac{1}{8}\sum_u m_i(u), \qquad S_i^2 = \tfrac{1}{8}\sum_u \|m_i(u) - \bar{m}_i\|_2^2, \tag{2}$$

$$Y_i = m_i(U) + N_i, \qquad N_i \sim \mathcal{N}(0, \rho S_i^2 I_{d_i}), \quad \rho > 0. \tag{3}$$

For a nonconstant group, $S_i > 0$; a constant group has a constant observation. The noises are mutually independent and independent of U. In the joint observation, both parameter groups correspond to the same sampled assignment U. Each scale is computed across all eight states and retained in every target class. The construction is $F \leftarrow U \rightarrow D(U) \rightarrow \theta_r(U) \rightarrow (Y_i, Y_j)$. In particular, data processing gives $I(F; Y_i) \le I(F; m_i(U))$. Neither quantity measures whether the network's forward computation retrieves the fact or which parameter intervention changes its answer (Geva et al., 2023; Hase et al., 2023).

The noise models an analyst-chosen observation resolution, limiting distinctions from arbitrarily small parameter differences. Rescaling a group also rescales its noise, preserving information; a general nonorthogonal reparameterization with fresh isotropic noise need not. Expected noise energy is $d_i \rho S_i^2$, so equal ρ is not equal physical precision across dimensions. The observation channel is defined from the complete labeled state table, but each observation contains only Y_i, Y_j, or their pair—not the realized U or F; seeds and provenance are not observation features. For a nonconstant group, let φ_Σ be a centered Gaussian density. Then

$$p_i(y) = \tfrac{1}{8}\sum_u \varphi_{\rho S_i^2 I}(y - m_i(u)),$$

$$p_i(y \mid f) = \tfrac{1}{4}\sum_{u: f(u)=f} \varphi_{\rho S_i^2 I}(y - m_i(u)). \tag{4}$$

The joint mixtures use paired means $(m_i(u), m_j(u))$. Figure 1 separates construction of the table from observation of one random row.

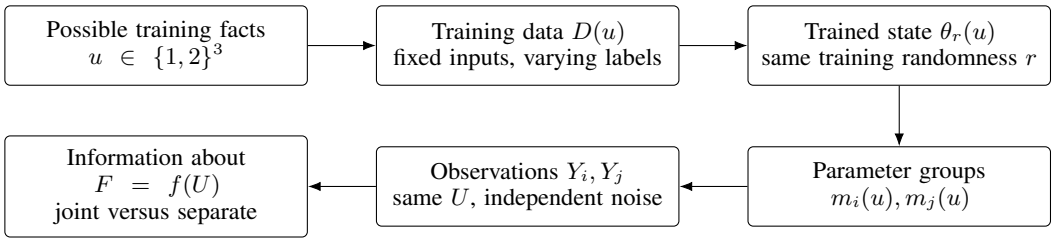

Figure 1: The information experiment. Enumerating the eight assignments gives the analyst a table of trained states. The observer receives noisy parameter vectors for one random assignment. The question is how much those vectors jointly and separately reveal about the selected fact. A new independent choice of training randomness repeats the whole construction.

Joint information about one fact. Mutual information, $I(F;Y) = H(F) - H(F \mid Y)$, is the reduction in uncertainty about F: zero bits for an uninformative observation, one bit if it determines the fact. Define the signed fact-specific surplus, in bits, by

$$\Delta_F = I(F; Y_i, Y_j) - I(F; Y_i) - I(F; Y_j), \tag{5}$$

$$= I(Y_i; Y_j \mid F) - I(Y_i; Y_j). \tag{6}$$

The second equality follows from the chain rule. Positivity means stronger average dependence within fact classes than after mixing them. For example, let K be a fair sign independent of the fair bit F. The signals K and $K(1-2F)$ reveal nothing about F separately, but their product determines it. Independent Gaussian noise preserves equal marginal distributions and distinguishable conditional joint laws. This key example illustrates the information property, without assuming a training mechanism.

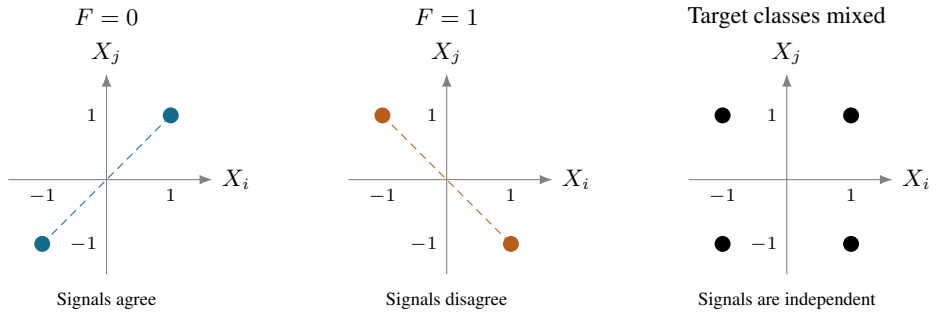

Figure 2: The key example: $X_i = K$, $X_j = K(1-2F)$, with an independent fair sign K and fair bit F. Dots show the signal means, with equal weight in each panel; observation noise is omitted. Both target classes have the same marginal distributions, but opposite joint dependence. Mixing the classes makes the signals independent. Independent Gaussian noise preserves the identical marginal distributions and unconditional independence, while the conditional joint laws remain distinguishable at every finite noise variance.

The sign of Δ_F depends on both the target variable and the observation channel. Conditioning on the complete source state U leaves only independent observation noise, whereas conditioning on F leaves the remaining coordinates of U random and can therefore preserve dependence between the two parameter groups. If each exact group determines F, its noiseless surplus is $-H(F)$; positive noisy surplus does not imply positive surplus in exact weights. The finite source already makes the information finite. Our signed quantity is negative co-information; in a nonnegative bivariate PID it equals synergy minus redundancy, without choosing a PID definition (Williams & Beer, 2010; Bertschinger et al., 2014).

Certification problem. Given the paired state table, source law, target labels, and ρ, the certification task is to compute a verified $b > 0$ with $b \le \Delta_F$, or otherwise return an inconclusive result. We construct an analytical bound $B_F \le \Delta_F$, then verify its value numerically. A nonpositive bound is inconclusive; a positive floating-point estimate alone is not a certificate.

4 A CERTIFICATE FOR JOINT INFORMATION SURPLUS

We express each mixture entropy as the entropy of its moment-matched Gaussian minus a nonnegative deficit, then bound only the deficits that enter the surplus with a negative sign.

4.1 MOMENT MATCHING AND ENTROPY DEFICITS

For either one group or the joint observation,

$$I(F;Y) = h(Y) - \sum_f p_f h(Y \mid F = f), \qquad p_f = \tfrac{1}{2}. \tag{7}$$

Here h is differential entropy in bits. Equation (4) therefore reduces the problem to entropies of finite Gaussian mixtures. For a nonconstant group, divide its observations by the known noise standard deviation $\sqrt{\rho S_i^2}$. This invertible change of units, called *noise whitening*, gives noise covariance I_d and preserves mutual information. For a pair, apply the scaling separately to each group. After this scaling and centering the means of the distribution under consideration, write any required mixture as

$$P = \frac{1}{n}\sum_{a=1}^{n} \mathcal{N}(\mu_a, I_d), \qquad \sum_a \mu_a = 0, \qquad V = \frac{1}{n}\sum_a \mu_a \mu_a^{\mathsf{T}}. \tag{8}$$

Here n is the number of components in the distribution being evaluated: $n = 8$ unconditionally and $n = 4$ within a target class in the main experiment. The *moment-matched Gaussian* $Q = \mathcal{N}(0, I_d + V)$ has the same mean and covariance as P. Since $\log q$ is quadratic,

$$h(P) = h(Q) - d_P, \qquad d_P = D_{\mathrm{KL}}(P\|Q) \geq 0. \tag{9}$$

The divergences in (9)–(10) use base-two logarithms. The deficit has the computable upper bound

$$d_P \leq R_P := D_2(P\|Q) = \log_2 \int \frac{p(y)^2}{q(y)}\,\mathrm{d}y. \tag{10}$$

This is Rényi monotonicity (van Erven & Harremoës, 2014); equivalently, apply Jensen's inequality to $\mathbb{E}_P \log_2(p/q)$. The order-two divergence is useful here because expanding the squared mixture density leaves a finite sum of Gaussian-product integrals. These can be evaluated analytically, unlike the original entropy integral (Section 6). Consequently, each mixture entropy satisfies $h(Q) - R_P \leq h(P) \leq h(Q)$.

Each conditional entropy may be translated separately without changing its value; the replicate-level observation-noise scale remains fixed.

4.2 A COMPUTABLE LOWER BOUND

For group i, let V_i and $V_{i|f}$ be the whitened source covariances, with each conditional covariance centered within its own class. The Gaussian contribution to its information is

$$g_i = \frac{1}{2}\log_2 \det(I + V_i) - \frac{1}{2}\sum_f p_f \log_2 \det(I + V_{i|f}). \tag{11}$$

Define g_{ij} in the same way using the joint covariances, and put $\Delta_G = g_{ij} - g_i - g_j$. This is the surplus obtained from the moment-matched Gaussian entropy terms. Write P_i, P_j, P_{ij} for the whitened unconditional mixtures and $P_{i|f}, P_{j|f}, P_{ij|f}$ for their class-conditional counterparts. Each R_P is computed against that mixture's own moment-matched Gaussian as in (10).

Proposition 1 (Lower bound on fact-specific surplus)**.** *Let U be uniform on a finite set and $F = f(U)$ a balanced binary target. Observe two nonconstant vector-valued functions of U through additive Gaussian noises that are independent of each other and of U, with fixed positive isotropic covariance in each group, as in* (3)*. Then*

$$\boxed{B_F = \Delta_G - R_{P_{ij}} - \sum_f p_f R_{P_{i|f}} - \sum_f p_f R_{P_{j|f}} \leq \Delta_F.} \tag{12}$$

In particular, $B_F > 0$ implies $\Delta_F > 0$.

Proof. The common pure-noise entropy cancels in (7), so (9) gives

$$I(F; Y_i) = g_i - d_{P_i} + \sum_f p_f d_{P_{i|f}}. \tag{13}$$

Substitution of this identity for the two separate observations and their pair into (5) yields

$$\begin{aligned}\Delta_F = \Delta_G &- d_{P_{ij}} + d_{P_i} + d_{P_j} \\ &+ \sum_f p_f d_{P_{ij|f}} - \sum_f p_f d_{P_{i|f}} - \sum_f p_f d_{P_{j|f}}.\end{aligned} \tag{14}$$

Discard the nonnegative terms with positive signs and replace each subtracted deficit by its upper bound R_P from (10). □

Equivalently, $L_{ij} = g_{ij} - R_{P_{ij}}$ lower-bounds joint information, while $U_i = g_i + \sum_f p_f R_{P_{i|f}}$ and U_j upper-bound the separate information. The certificate is $B_F = L_{ij} - U_i - U_j$: positive surplus requires a joint lower bound to exceed the sum of two separate upper bounds.

If either group is constant, $\Delta_F = 0$. For nonconstant groups, the next result identifies when the bound detects a positive interaction.

5 When the certificate detects an interaction

For a nonconstant group, normalize its variation and whiten its observation:

$$X_i = \frac{m_i(U) - \bar{m}_i}{S_i}, \qquad \widetilde{Y}_i = \frac{Y_i - \bar{m}_i}{\sqrt{\rho} S_i} = \sqrt{t}\, X_i + Z_i, \qquad t = \rho^{-1}. \tag{15}$$

Here Z_i is standard Gaussian, $\mathbb{E}\|X_i\|_2^2 = 1$, and X_i is fixed as t varies. These invertible changes of units preserve mutual information; low SNR means large observation noise. Write $C_i = \mathrm{Cov}(X_i)$, $C_{ij} = \mathrm{Cov}((X_i, X_j))$, and $D_{ij} = \mathrm{Cov}(X_i, X_j)$ for the joint covariance's cross block, with analogous conditional notation. The whitened covariances above are $V_i(t) = tC_i$ and $V_{i|f}(t) = tC_{i|f}$. Define

$$\Gamma_F = \mathbb{E}_F \|\mathrm{Cov}(X_i, X_j \mid F)\|_F^2 - \|\mathrm{Cov}(X_i, X_j)\|_F^2. \tag{16}$$

Here $\|\cdot\|_F$ is the Frobenius norm. The coefficient compares squared cross-covariance within fact classes with that after mixing them. For the key example, the conditional cross-covariances are $+1$ and -1, the unconditional one is zero, and $\Gamma_F = 1$.

Lemma 1 (Order of the entropy penalty)**.** *For a fixed finite signal distribution observed through independent standard real Gaussian noise, the order-two penalty against its moment-matched Gaussian satisfies $R_P(t) = O(t^3)$ as $t \to 0^+$. The statement applies to each marginal and class-conditional mixture.*

Proposition 2 (Low-SNR accuracy of the bound)**.** *For a fixed finite source, a balanced binary target F, and independent standard Gaussian noises in the two observations,*

$$\Delta_F(t) = \frac{t^2 \Gamma_F}{2 \ln 2} + O(t^3), \qquad B_F(t) = \frac{t^2 \Gamma_F}{2 \ln 2} + O(t^3). \tag{17}$$

Moreover, the approximation gap is nonnegative and satisfies

$$0 \le \Delta_F(t) - B_F(t) = O(t^3).$$

Consequently, if $\Gamma_F > 0$, then $B_F(t) > 0$ for every sufficiently small positive t.

Why the orders differ. For Lemma 1, set $s = \sqrt{t}$ and compare the mixture and moment-matched Gaussian density ratios with a standard Gaussian density ϕ. After centering, their first derivatives with respect to s at zero vanish and their second derivatives agree. Their difference is therefore $O(s^3)$ in $L^2(\phi)$. The matched density ratio is bounded below for small s, so the corresponding chi-square divergence is $O(s^6) = O(t^3)$. Appendix C gives the domination argument and the conversion to the Rényi penalty.

For the Gaussian part, expand the natural logarithm for any fixed covariance C:

$$\ln \det(I + tC) = t \operatorname{tr} C - \frac{t^2}{2} \operatorname{tr}(C^2) + O(t^3).$$

The first-order traces cancel between a joint covariance and its diagonal blocks. Using the cross block D_{ij} gives

$$\operatorname{tr} C_{ij}^2 = \operatorname{tr} \begin{pmatrix} C_i & D_{ij} \\ D_{ij}^{\mathsf{T}} & C_j \end{pmatrix}^2 = \operatorname{tr} C_i^2 + \operatorname{tr} C_j^2 + 2\|D_{ij}\|_F^2.$$

Subtracting the corresponding conditional expansion yields the coefficient $\Gamma_F/(2 \ln 2)$ in Δ_G. Every deficit in (14) lies between zero and its $O(t^3)$ Rényi penalty, so both the true surplus and its lower bound share that coefficient, and their gap is $O(t^3)$. Gaussian-channel low-SNR expansions are classical (Prelov & Verdú, 2004; Guo et al., 2005); the property needed here is the third-order error control for this computable penalty.

The constants depend on the fixed signals. The proposition gives no finite-noise cutoff: positivity at $\rho = 8$, or $t = 1/8$, requires the numerical check below. The stronger-signal parity example in Section 8 shows why this qualification matters.

6 COMPUTING AND VERIFYING THE CERTIFICATE

Let the rows of M be the centered, whitened means in (8). Define

$$G = MM^{\mathsf{T}}, \qquad K_G = G(I_n + 2G/n)^{-1}, \qquad \delta_{ab}^2 = G_{aa} + G_{bb} - 2G_{ab}.$$

Completing the square reduces the penalty integral to

$$\int \frac{p^2}{q} = \frac{\det(I_n + G/n)}{\sqrt{\det(I_n + 2G/n)}} \frac{1}{n^2} \sum_{a,b=1}^{n} \exp\left[-\frac{\delta_{ab}^2}{4} + \frac{(K_G)_{aa} + 2(K_G)_{ab} + (K_G)_{bb}}{4}\right]. \tag{18}$$

Also, $\det(I_d + V) = \det(I_n + G/n)$. These expressions follow from standard Gaussian-product and determinant identities (Kitagawa, 2020); Appendix B derives the form used here. Paired source ordering gives $G_{ij} = G_i + G_j$. For a target class, let $\mathcal{I}_f$ index its n_f source states and let $\mathbf{1} \in \mathbb{R}^{n_f}$ be the all-ones vector. For $g \in \{i, j\}$, the conditional Gram matrix is

$$G_{g|f} = J_f G_g[\mathcal{I}_f, \mathcal{I}_f] J_f, \qquad J_f = I_{n_f} - \frac{\mathbf{1}\mathbf{1}^{\mathsf{T}}}{n_f}. \tag{19}$$

This selects and recenters each class while retaining the global noise scale. Thus, the transformer experiment requires only 8×8 and 4×4 matrices. Extracting them costs $O(dn^2)$; subsequent matrix operations cost $O(n^3)$, with exact arithmetic also depending on input bit length.

FP32 states are exact dyadic inputs. Their unwhitened Gram matrices, divided by the rational noise variance ρS_i^2, give rational entries even when the whitened vectors themselves are irrational. Determinants and inverses are evaluated exactly; scalar exponential and logarithm enclosures give an outward interval for (12). Appendix E specifies the checks and arithmetic.

Definition 1 (Numerical certificate of positive surplus)**.** For exact inputs satisfying Proposition 1, a certificate is an outward-verified enclosure $B_F \in [b_{\mathrm{lo}}, b_{\mathrm{hi}}]$ with rational endpoints and $b_{\mathrm{lo}} > 0$, together with its exact channel specification. The sign decision is sound because

$$\Delta_F \geq B_F \geq b_{\mathrm{lo}} > 0.$$

The upper endpoint encloses B_F, not Δ_F. An interval containing zero is inconclusive; a genuinely nonpositive B_F cannot become positive through increased numerical precision. The certificate applies to the specified finite channel; numerical enclosure and uncertainty across training replicates are separate questions.

7 EXPERIMENTAL DESIGN

The source in Section 3 gives eight datasets per replicate. Each model is a one-block causal transformer with hidden width 32, four attention heads, feed-forward width 128, and 14,784 parameters. It is trained from scratch on 384 presentations for 24 AdamW updates. The observed groups are token and positional embeddings E (1,760 coordinates), and final normalization plus the answer head H (320 coordinates). Appendix H gives the full partition, tokenization, and training specification. Alternate-format queries yield 68/480 correct answers at initialization and 473/480 at step 24 (14.17% and 98.54%). The increase in accuracy provides descriptive evidence that the associations were learned; the 480 evaluations are not independent samples (Appendix I).

Development used 22 existing replicates and screened 15,708 dependent target/pair/noise combinations in binary64. The E/H pair, target $F = U_3 - 1$, step 24, $\rho = 8$, and threshold $\tau = 10^{-4}$ bit were fixed before drawing randomness for 20 new replicates. Within each replicate, the same initialization and presentation-order seeds are reused for all eight assignments. All 20 pairs were allocated once and retained, without optional stopping or replacement. For training randomness R, drawn through independent uniform 32-bit initialization and presentation-order seeds, define

$$Z_R = \mathbf{1}\{b_{\mathrm{lo}}(R) \geq \tau\}, \qquad p_\tau = \Pr_R[b_{\mathrm{lo}}(R) \geq \tau], \qquad \tau = 10^{-4} \text{ bit}. \tag{20}$$

The primary criterion requires the lower endpoint of a two-sided exact 95% Clopper–Pearson interval for p_τ to exceed 0.5, requiring at least 15 successes out of 20. This is a probability of *certifying* the threshold over training randomness, rather than an estimate of true surplus. No numerical interval overlapped the threshold.

Comparisons and controls. Every method receives the same source, target, states, and noise scale. Corrected Kolchinsky–Tracey (KT) bounds (Kolchinsky & Tracey, 2018) use pairwise distances to bound mixture entropies. The transport/Fano composite tightens single-group upper bounds by pairing opposite-class components and joint lower bounds by classification overlap. A separate development comparison uses the proved Taylor entropy bound of Joudeh & Škorić (2025) at degrees 2, 4, 8, and 10. Appendix D specifies all formulas and their conversion to surplus bounds. Controls include the key/parity example, redundant noisy copies, constant channels, and the replacement

$$P(Y_i, Y_j \mid F) \longmapsto P(Y_i \mid F)P(Y_j \mid F). \tag{21}$$

This preserves each group's information and removes within-class coupling, forcing $\Delta_F = -I(Y_i; Y_j) \leq 0$ by (6).

8 Results

Positive certificates across new training replicates. At the prespecified setting, all 20 new replicates yield positive certificates and 19 exceed 10^{-4} bit. The exact two-sided 95% Clopper–Pearson interval for p_τ is $[0.75126724, 0.99873491]$, meeting the primary criterion. Table 1 compares the bounds; Appendix I reports all 20 outcomes, including the below-threshold replicate 12.

Table 1: Comparison on 20 independent experimental replicates at the prespecified E/H, third-coordinate, step-24, $\rho = 8$ setting. The middle columns count replicates in which a method certifies the stated inequality for Δ_F. A count of zero means no successful certifications, not zero mutual information. The last two columns give the minimum and maximum verified lower endpoints across replicates, in bits; they are not confidence limits or estimates of Δ_F.

	Replicates certified (of 20)		Lower endpoint (bits)	
Method	$\Delta_F > 0$	$\Delta_F \geq 10^{-4}$	Minimum	Maximum
Proposed bound	20	19	0.00006223	0.00047478
Pairwise entropy (KT)	0	0	−0.284158	−0.277968
Transport + Fano	0	0	−0.0748955	−0.0592105

The minimum verified endpoint gives, for each of the 20 fixed channels,

$$\boxed{I(F; Y_E, Y_H) \;\geq\; I(F; Y_E) + I(F; Y_H) + 6.2 \times 10^{-5}.} \tag{22}$$

The constant rounds down the smallest endpoint; 19 channels support the stronger 10^{-4}-bit margin. The mean signed lower endpoint is 0.0003170552 bit. These bounds establish a small positive increment in information from joint observation, without estimating its true size.

Why separate entropy bounds lose the sign. KT and transport/Fano give negative bounds on all 20 channels. For fixed centered signals x_a with uniform source weights, covariance C, and unit noise, the KT entropy offset under scaling by $\sqrt{t}$ is

$$A_\alpha(t) = -\frac{1}{n} \sum_a \log_2 \left(\frac{1}{n} \sum_b e^{-\alpha t \|x_a - x_b\|^2} \right).$$

The mean pairwise squared distance is $2 \operatorname{tr} C$, giving $A_\alpha(t) = 2\alpha t \operatorname{tr} C / \ln 2 + O(t^2)$, whereas the true offset is $t \operatorname{tr} C/(2 \ln 2) + O(t^2)$. The lower choice $\alpha = 1/8$ and upper choice $\alpha = 1/2$ therefore leave first-order slack, which can mask the second-order surplus. The proposed bound preserves that term with third-order error. Transport tightens the single-group bounds here but still leaves a negative surplus bound.

Noise dependence and the penalty budget. In the development example, 6.674×10^{-4} bit of Gaussian interaction minus 3.302×10^{-4} in penalties leaves 3.372×10^{-4} bit. Figure 3 shows the penalty falling below the interaction at $\rho = 8$, then vanishing faster as noise increases. Here $\Gamma_F \approx 0.06725450$, giving the limiting rescaled coefficient 0.04851386. Appendix H.3 retains the component table and individual margin analysis.

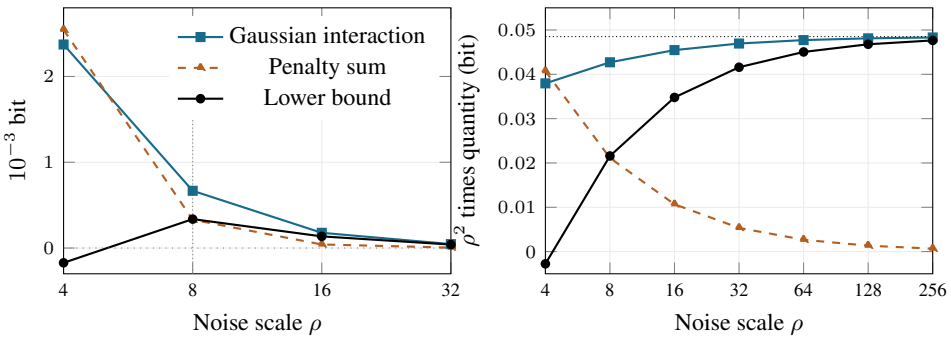

Figure 3: Development example, E/H, third-coordinate target, step 24. Left: the positive margin appears when the penalty falls below the Gaussian interaction; the vertical line marks $\rho = 8$. Right: rescaling by ρ^2 exposes convergence to the leading coefficient (horizontal dotted line), while the rescaled penalty tends to zero. Curves join saved binary64 screening values, not independently certified intervals at every point. The $\rho = 8$ value is separately certified in Table 4.

Controls and a reversal of the bound comparison. On the development E/H channel, the tested Taylor degrees give -0.17692721 bit, versus 0.00033719 for the proposed bound. This comparison uses degrees up to 10 and the component-density ceiling specified in Appendix D. The ordering reverses for synthetic parity at amplitude 2 with fixed unit noise: Taylor gives 0.08428519 bit, while our bound gives -0.84306302, despite positive true surplus. Thus the low-SNR advantage is not uniform dominance (Table 5). These amplitude changes use fixed unit noise, unlike rescaling neural states and recomputing S_i. On a calibration replicate, the conditional-product control preserves both single-group endpoints and gives $B_F \approx -0.00045801$ bit, consistent with its nonpositive true surplus. It tests the role of within-class coupling in the certificate, not a causal retrieval mechanism.

9 Discussion and Conclusion

Under the specified finite source and Gaussian observation channel, the embedding and output groups exhibit a positive fact-specific information surplus. The low-SNR analysis relates this effect to class-conditional cross-covariance, and interval arithmetic certifies its sign in all 20 confirmatory replicates.

The interpretation depends on the source, target, parameterization, and observation law. Exact parameter access or a different metric can change the sign. The small certified margins neither estimate true surplus nor establish a recall benefit. Replication varies initialization and data order for one small task, architecture, group pair, and checkpoint; it does not establish that the observed geometry is specific to transformer learning rather than other label-dependent encodings. The parity reversal also shows that the certificate need not detect positive surplus outside its low-SNR regime.

A larger-model application requires an explicit source law and its ensemble of trained states; one pretrained checkpoint does not supply that ensemble. Gram reduction removes ambient dimension from the matrix inversions, but source enumeration and exact arithmetic remain costs. Within these limits, (22) gives a certified finite-source example of positive information surplus after training.

REPRODUCIBILITY

Appendices B–E provide the proofs and numerical verification procedure. Appendix H specifies training and development selection; Appendix I reports every confirmatory outcome. Code and exact finite-channel inputs for reproducing the certificate calculations are provided in the supplementary materials.

ETHICS STATEMENT

The experiment uses synthetic labels and contains no human participants or personal data.

AI USE STATEMENT

Generative AI tools assisted in formulating and checking mathematical claims, designing experiments, implementing analyses, searching literature, interpreting results, drafting text, and translation. The authors are responsible for the final claims and submitted material.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL https://www.jmlr.org/papers/v19/17-646.html.

Alessandro Achille, Giovanni Paolini, and Stefano Soatto. Where is the information in a deep neural network? arXiv:1905.12213, 2019. URL https://arxiv.org/abs/1905.12213.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/26d3c9a66836ded8f34a944f2bfe868e-Abstract-Conference.html.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5178b2f2d7c44aa390c0777dc77b3f0c-Abstract-Conference.html.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161. URL https://www.mdpi.com/1099-4300/16/4/2161.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Xinhao Fan and Shreesh P. Mysore. Quantifying information stored in synaptic connections rather than in firing activities of neural networks. arXiv:2411.17692v2, 2026. URL https://arxiv.org/abs/2411.17692v2.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

Dongning Guo, Shlomo Shamai (Shitz), and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072. URL https://arxiv.org/abs/cs/0412108.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1772–1791. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.153. URL https://aclanthology.org/2021.eacl-main.153/.

John R. Hershey and Peder A. Olsen. Approximating the Kullback–Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. IV–317–IV–320, 2007. doi: 10.1109/ICASSP.2007.366913. URL https://doi.org/10.1109/ICASSP.2007.366913.

Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. On entropy approximation for Gaussian mixture random vectors. Author manuscript, 2008. URL https://publikationen.bibliothek.kit.edu/1000034848/2581578.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28/.

Basheer Joudeh and Boris Škorić. Numerical evaluation of Gaussian mixture entropy. arXiv:2503.14047v2, 2025. URL https://arxiv.org/abs/2503.14047v2.

Genshiro Kitagawa. Pearson chi-squared-divergence approach to Gaussian mixture reduction and its application to Gaussian-sum filter and smoother. arXiv preprint arXiv:2001.00727, 2020. URL https://arxiv.org/abs/2001.00727. Technical report, Meiji University MIMS-RPB, Statistics and Data Science Series SDS-12.

Artemy Kolchinsky and Brendan D. Tracey. Estimating mixture entropy with pairwise distances. arXiv:1706.02419v4, 2018. URL https://arxiv.org/abs/1706.02419v4. Corrected version of the 2017 Entropy article.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5131–5157. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.256. URL https://aclanthology.org/2025.acl-long.256/.

Gabriel Mahuas, Olivier Marre, Thierry Mora, and Ulisse Ferrari. Small-correlation expansion to quantify information in noisy sensory systems. *Physical Review E*, 108:024406, 2023. doi: 10.1103/PhysRevE.108.024406. URL https://www.phys.ens.psl.eu/~tmora/Physique/MahuasMarre23.pdf.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013. doi: 10.1214/12-AOS1065. URL https://arxiv.org/abs/1109.3250.

Frank Nielsen and Ke Sun. On the chain rule optimal transport distance. arXiv:1812.08113v3, 2020. URL https://arxiv.org/abs/1812.08113v3.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Viacheslav V. Prelov and Sergio Verdú. Second-order asymptotics of mutual information. *IEEE Transactions on Information Theory*, 50(8):1567–1580, 2004. doi: 10.1109/TIT.2004.831784.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437/.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500. URL https://arxiv.org/abs/1206.2459.

Boshi Wang and Huan Sun. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/00a0ebcad584c59dbc439c2af8793638-Abstract-Conference.html.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Knowledge mechanisms in large language models: A survey and perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7097–7135. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.416. URL https://aclanthology.org/2024.findings-emnlp.416/.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. arXiv:1004.2515, 2010. URL https://arxiv.org/abs/1004.2515.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/ad71c82b22f4f65b9398f76d8be4c615-Abstract.html.

Ruoxi Xu, Yunjie Ji, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Ben He, Yingfei Sun, Xiangang Li, and Le Sun. Memorizing is not enough: Deep knowledge injection through reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28682–28693. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1392. URL https://aclanthology.org/2025.acl-long.1392/.

A Interpreting the source and observation model

Figures 1 and 2 distinguish the source construction from the relation that carries information. The following scaling distinction connects the synthetic and neural examples.

The synthetic signals $aK, aK(1-2F)$ at fixed unit noise have $\Gamma_F = a^4$. This amplitude change differs from multiplying a raw neural parameter group by a nonzero scalar and recomputing its scale: (2) then changes the noise by the same factor, preserving the normalized channel. The key bound is positive at tested amplitudes $a = 1/4, 1/2, 1$ and negative at $a = 2$; the true surplus is positive at every finite nonzero amplitude and noise variance.

B Derivation of the Gram integral

Let $\Sigma_Q = I + V$, q be the density of $\mathcal{N}(0, \Sigma_Q)$, and ϕ the standard Gaussian density. For two mixture means, put $T = 2I - \Sigma_Q^{-1}$. Since $\Sigma_Q \succeq I$, $T \succ 0$. Completing the square gives

$$\int \frac{\phi(y-\mu_a)\phi(y-\mu_b)}{q(y)}\,\mathrm{d}y = \det(\Sigma_Q)^{1/2}\det(T)^{-1/2} \times \exp\big[\tfrac{1}{2}(\mu_a+\mu_b)^{\mathsf{T}}T^{-1}(\mu_a+\mu_b) - \tfrac{1}{2}(\|\mu_a\|^2+\|\mu_b\|^2)\big]. \quad (23)$$

The inverse identity

$$T^{-1} = \tfrac{1}{2}[I + (I+2V)^{-1}]$$

turns the exponent into

$$-\frac{\|\mu_a-\mu_b\|^2}{4} + \frac{(\mu_a+\mu_b)^{\mathsf{T}}(I+2V)^{-1}(\mu_a+\mu_b)}{4}.$$

The determinant prefactor is $\det(I+V)/\sqrt{\det(I+2V)}$. Finally,

$$\det(I + M^{\mathsf{T}}M/n) = \det(I + MM^{\mathsf{T}}/n), \qquad M(I + 2M^{\mathsf{T}}M/n)^{-1}M^{\mathsf{T}} = G(I+2G/n)^{-1}.$$

Averaging (23) over all n^2 pairs proves (18). Rank deficiency of G causes no singularity: the two regularized matrices are positive definite.

C Finite-source order argument

We give details for one centered finite signal distribution, allowing fixed weights $w_a \geq 0$, $\sum_a w_a = 1$. Let its values be ξ_a, with $\sum_a w_a \xi_a = 0$. This formulation covers all marginal and class-conditional mixtures used below. Set $s = \sqrt{t}$, $C = \sum_a w_a \xi_a \xi_a^{\mathsf{T}}$, and define the ratios of its noisy density and moment-matched Gaussian density to ϕ:

$$p_s(x)/\phi(x) = \sum_a w_a \exp\Big(sx^{\mathsf{T}}\xi_a - \tfrac{s^2}{2}\|\xi_a\|^2\Big), \quad (24)$$

$$q_s(x)/\phi(x) = \det(I+s^2C)^{-1/2}\exp\big(\tfrac{1}{2}x^{\mathsf{T}}[I-(I+s^2C)^{-1}]x\big). \quad (25)$$

At $s = 0$, both ratios are one. Centering makes their first derivatives zero, and their second derivatives both equal $x^{\mathsf{T}}Cx - \operatorname{tr} C$. Here C is fixed; the covariance of the scaled signal is tC.

To justify a third-order Taylor remainder in $L^2(\phi)$, first note that finite support bounds $\max_a \|\xi_a\|$. Up to order three, derivatives of (24) are bounded for $|s| \leq s_0$ by a polynomial in $\|x\|$ times $\exp(s_0 \max_a \|\xi_a\|\,\|x\|)$. Its square is integrable against ϕ. For (25), choose s_0 sufficiently small that $s_0^2\|C\|_{\mathrm{op}} < 1/4$, unless $C = 0$, in which case the two densities are equal. The exponent is then at most $\|x\|^2/8$. Matrix inverses and their first three derivatives are bounded on this compact interval of s. Consequently, the derivatives of the Gaussian ratio are bounded by $c_0(1+\|x\|^m)\exp(\|x\|^2/8)$ for finite c_0, m. The square of this bound times $\phi(x)$ is integrable. Taylor's theorem therefore yields

$$\left\|\frac{p_s - q_s}{\phi}\right\|_{L^2(\phi)} = O(s^3).$$

Moreover, $C \succeq 0$ implies $q_s/\phi \geq \det(I + s^2 C)^{-1/2}$, which is bounded away from zero for $|s| \leq s_0$. Hence

$$\chi^2(P_s \| Q_s) = \int \phi(x) \, \frac{[(p_s(x) - q_s(x))/\phi(x)]^2}{q_s(x)/\phi(x)} \, \mathrm{d}x$$
$$\leq c_1 \left\| \frac{p_s - q_s}{\phi} \right\|^2_{L^2(\phi)} = O(s^6) = O(t^3).$$

Since $R_P = \log_2(1 + \chi^2(P\|Q))$, this proves Lemma 1. Because only finitely many marginal and conditional mixtures are considered, the neighborhood can be chosen uniformly. Class-dependent centering is legitimate because each entropy and divergence is translation-invariant.

For Proposition 2, define the Gaussian entropy offset $a(C, t) = \frac{1}{2 \ln 2} \ln \det(I + tC)$. The fixed covariances C_i, C_j, C_{ij} are those defined after (15). Block trace cancellation gives

$$a(C_{ij}, t) - a(C_i, t) - a(C_j, t) = -\frac{t^2}{2 \ln 2} \|\mathrm{Cov}(X_i, X_j)\|_F^2 + O(t^3).$$

Subtract the averaged conditional identity to obtain $\Delta_G(t) = t^2 \Gamma_F / (2 \ln 2) + O(t^3)$. Every deficit in (14) is bounded between zero and an $O(t^3)$ penalty. Both Δ_F and B_F therefore have this leading coefficient. More explicitly,

$$\Delta_F - B_F = (R_{P_{ij}} - d_{P_{ij}}) + d_{P_i} + d_{P_j} + \sum_f p_f d_{P_{ij|f}}$$
$$+ \sum_f p_f (R_{P_{i|f}} - d_{P_{i|f}}) + \sum_f p_f (R_{P_{j|f}} - d_{P_{j|f}}).$$

Each term is nonnegative and $O(t^3)$, proving $0 \leq \Delta_F - B_F = O(t^3)$. If $\Gamma_F > 0$, division by t^2 proves eventual positivity. If $\Gamma_F = 0$, this argument gives no sign; higher-order terms matter.

D Details of the comparison bounds

D.1 Entropy bounds and the KT construction

All methods use the same source weights, target, saved states, and noise normalization. Entropy bounds $L_h(P) \leq h(P) \leq U_h(P)$ give

$$\begin{aligned} L_I(P) &= \max\{0, L_h(P) - \mathbb{E}_F U_h(P_{|F})\}, \\ U_I(P) &= \min\{H(F), U_h(P) - \mathbb{E}_F L_h(P_{|F})\}, \\ L_\Delta &= L_I(P_{ij}) - U_I(P_i) - U_I(P_j). \end{aligned} \tag{26}$$

We report this signed lower bound, including negative values. The methods bound the same information quantity; they use different entropy or single-group information bounds.

Pairwise entropy (KT). For equally weighted unit-covariance components, the corrected Kolchinsky–Tracey bounds (Kolchinsky & Tracey, 2018) use

$$A_\alpha(P) = -\frac{1}{n} \sum_a \log_2 \left(\frac{1}{n} \sum_b e^{-\alpha \delta_{ab}^2} \right).$$

The offset is in bits; the exponents $\delta_{ab}^2/8$ and $\delta_{ab}^2/2$ are Gaussian Bhattacharyya and KL distances in natural-log units. Adding $A_{1/8}$ or $A_{1/2}$ to the component noise entropy gives, respectively, a lower or upper entropy bound. We apply both to each unconditional and conditional mixture in (26).

D.2 Finite transport coupling and the joint lower bound

Let $\mathcal{I}_0, \mathcal{I}_1$ be the two equally sized source classes, each of size m. For any bijection $\pi : \mathcal{I}_0 \to \mathcal{I}_1$, generate a uniform pair index independently of F, then choose the member of that pair with label

F. This generates the original conditional mixtures. Revealing the index can only increase the information, so

$$I(F;Y_i) \le U_{\mathrm{tr}}(P_i) := \min_{\pi} \frac{1}{m} \sum_{a\in\mathcal{I}_0} \min\left\{1, \frac{1}{2}\log_2\left(1+\frac{\delta^2_{a,\pi(a)}}{4}\right)\right\}. \tag{27}$$

For a pair of means at squared distance δ^2, the centered signal covariance has rank one and nonzero eigenvalue $\delta^2/4$. Gaussian maximum entropy therefore gives the logarithmic upper bound in (27); the additional ceiling of one bit is $H(F)$. We enumerate $m! = 24$ pairings when $m = 4$, with outward arithmetic; a feasible pairing alone is already a valid upper bound.

For the joint observation, its Bayes error has the upper bound

$$e_* \le \beta := \min\left\{\frac{1}{2}, \frac{1}{2m}\sum_{a\in\mathcal{I}_0}\sum_{b\in\mathcal{I}_1} e^{-\delta^2_{ab}/8}\right\}.$$

This follows by bounding the binary Bayes overlap with the Bhattacharyya overlap and expanding an upper bound over component pairs. Binary Fano and monotonicity of binary entropy $h_{\mathrm{b}}(x) = -x\log_2 x - (1-x)\log_2(1-x)$ on $[0, 1/2]$, with $0\log_2 0 = 0$, give $I(F;Y_i,Y_j) \ge 1 - h_{\mathrm{b}}(\beta)$. The transport/Fano composite reported in Table 1 uses

$$L_{ij} = \max\{L_I^{\mathrm{KT}}(P_{ij}), 1-h_{\mathrm{b}}(\beta)\}, \qquad U_i = \min\{U_I^{\mathrm{KT}}(P_i), U_{\mathrm{tr}}(P_i)\},$$

and the analogous U_j, then reports $L_{ij} - U_i - U_j$.

D.3 Taylor-bound calculation

We use the May 2025 v2 lower-bound theorem of Joudeh & Škorić (2025). It uses an upper bound on the density $m_0 \ge \sup_y p(y)$. For integer degree $C \ge 2$, its monotone Taylor expression in nats is

$$T_C(P) = -\ln m_0 + \sum_{k=1}^{C-1} \frac{1}{k}\mathbb{E}_P(1 - p(Y)/m_0)^k \le h_{\mathrm{nat}}(P). \tag{28}$$

We use $m_0 = (2\pi)^{-k_P/2}$ in the affine signal span of dimension k_P, a valid ceiling for a mixture of unit-covariance components. If that span is smaller than the ambient dimension d,

$$h(P_d) = h(P_{k_P}) + (d-k_P)h(\mathcal{N}(0,1)), \qquad I(F;Y_d) = I(F;Y_{k_P}).$$

The mutual-information identity uses the full unconditional signal span. Class-specific spans are used only for individual conditional entropies; their removed noise entropy is restored before (26) is applied. This removes irrelevant pure-noise coordinates from the entropy-bound comparison.

For mixture weights w_i, let

$$A_a = \frac{\int p(y)^a \,\mathrm{d}y}{m_0^{a-1}} = a^{-k_P/2} \sum_{k_1+\cdots+k_n=a} \frac{a!}{\prod_i k_i!}\prod_i w_i^{k_i} \exp\left[-\frac{1}{2a}\sum_{i<j} k_i k_j \delta^2_{ij}\right]. \tag{29}$$

Expansion of (28) then gives the noise-entropy offset

$$T_C(P) - h_{\mathrm{nat}}(\mathcal{N}(0, I_{k_P})) = H_{C-1} - \frac{k_P}{2} + \sum_{a=2}^{C} \frac{(-1)^{a+1}}{a-1}\binom{C-1}{a-1} A_a,$$

where $H_k = \sum_{j=1}^{k} 1/j$. We take the maximum of this offset and zero, divide by $\ln 2$, and restore the noise entropy. The upper entropy offset is the smaller of $\frac{1}{2}\log_2\det(I+V)$ and the discrete component-label entropy. Coincident means may be merged by summing their weights. These steps give the entropy bounds used in (26); the polynomial-fit approximation is not substituted for a one-sided bound.

The plateau on the development example has a finite-degree explanation. Every expectation in (28) is at most one, so

$$T_C(P) - h_{\mathrm{nat}}(\mathcal{N}(0, I_{k_P})) \le H_{C-1} - k_P/2.$$

At $C = 10$ and unconditional rank $k_P = 7$, this ceiling is $-1691/2520$ nats. The noise-entropy lower bound therefore controls the unconditional mixture. The saved within-class evaluations at rank three also have negative Taylor offsets at the tested degrees. Their noise floors control the upper bounds on single-group mutual information, which sum to about 0.17692721 bit, while the joint lower stays zero. Higher degrees and other valid density ceilings are outside this particular comparison. The amplitude-2 reversal shows that the Taylor method remains useful in another regime.

The Taylor-bound comparison took 2.09 seconds for its fixed set of channels and degrees, including 7,161 mixture-composition terms. This timing and the confirmatory interval-evaluation time describe different workloads and should not be interpreted as a speed ratio between methods.

E NUMERICAL VERIFICATION

E.1 VERIFICATION PROCEDURE

The verification procedure consists of four steps:

1. **Specify the channel.** Supply the two exact, unwhitened centered Gram matrices $\widehat{G}_i, \widehat{G}_j$ in the same source order, the uniform source law, the balanced target labels, the two positive source variances, and the rational noise scale ρ. Input consistency requires symmetry, positive semidefiniteness, centering, and the identity between each source variance and its mean Gram diagonal.
2. **Form the required mixtures.** Divide each Gram matrix by its noise variance, $G_i = \widehat{G}_i/(\rho S_i^2)$, add the resulting matrices for the joint observation, and select and recenter each target-class submatrix. Retain the global noise scales when forming conditional distributions.
3. **Enclose the bound.** Evaluate the regularized determinants and inverses exactly. Use scalar enclosures in (18) to bound each R_P, and combine them with the Gaussian entropy terms according to (12), reversing interval endpoints for negative coefficients. Return rational endpoints $[b_{\mathrm{lo}}, b_{\mathrm{hi}}]$.
4. **Decide positivity.** Certify positive surplus if $b_{\mathrm{lo}} > 0$. Otherwise, this evaluation provides no positive certificate.

E.2 EXACT INPUTS AND RATIONAL MATRIX OPERATIONS

Saved FP32 coordinates are interpreted as exact dyadic inputs. The implementation forms centered Gram entries, determinants, and inverses with rational arithmetic, and checks the inverse identities exactly. Although whitening individual coordinates may involve square roots, its Gram entries are obtained by dividing the original inner products by ρS_i^2; at the prespecified rational noise scale these remain rational. The square root in (18) is handled as a factor of one half after taking a logarithm. If e_{ab} denotes its rational exponent, then

$$R_P = \log_2 \det(I_n + G/n) - \tfrac{1}{2}\log_2 \det(I_n + 2G/n) + \log_2\left(\frac{1}{n^2}\sum_{a,b} e^{e_{ab}}\right).$$

Thus only scalar exponential and logarithm evaluations need enclosures.

E.3 SCALAR ENCLOSURES AND INTERVAL PROPAGATION

The exact-arithmetic procedure uses exact elimination for the regularized Gram determinants and inverses, checking each inverse by multiplication. Scalar arguments are first enclosed on an outward dyadic grid; monotonicity then reduces their evaluation to rational endpoints.

For $0 \le x \le 1$, put $T_k(x) = \sum_{j=0}^{k}(-x)^j/j!$. The alternating-series remainder gives, for $N \ge 1$,

$$0 < T_{2N+1}(x) \le e^{-x} \le T_{2N}(x).$$

For a larger $x \ge 0$, write $x = 2^k y$ with $0 \le y \le 1$, enclose e^{-y}, and square the nonnegative interval k times with outward rounding. Positive exponential arguments use the reciprocal interval: if $0 < \ell \le e^{-x} \le u$, then $1/u \le e^x \le 1/\ell$.

For a positive rational logarithm argument, write $x = 2^k m$, $1 \le m < 2$, and put $z = (m-1)/(m+1)$. The positive series gives

$$\ln m = 2 \sum_{j=0}^{N-1} \frac{z^{2j+1}}{2j+1} + \varepsilon_N, \qquad 0 \le \varepsilon_N \le \frac{2z^{2N+1}}{(2N+1)(1-z^2)}.$$

The tail bound follows by replacing every remaining denominator by $2N+1$ and summing a geometric series. The same formula at $z = 1/3$ encloses $\ln 2$. For intervals $\ln m \in [\ell_m, u_m]$ and $\ln 2 \in [\ell_2, u_2]$, with $\ell_2 > 0$,

$$\log_2 x \in \left[k + \frac{\ell_m}{u_2},\ k + \frac{u_m}{\ell_2} \right].$$

When $m = 1$, the answer is the exact integer k.

The saved verifier uses degrees 40 and 41 for the exponential brackets and 36 terms for each logarithm series. Exponential arguments and logarithm arguments after power-of-two scaling are first enclosed on a grid of spacing 2^{-64}; the scalar routines use outward grids of spacing 2^{-160} to control denominator growth. These are computational settings, whereas the remainder inequalities above establish validity. In summing intervals, a negative coefficient reverses the endpoint order. The known inequality $R_P \ge 0$ also permits intersecting its computed enclosure with $[0, \infty)$. These steps propagate a rigorous enclosure through every term of (12).

E.4 Diagnostic checks and control channels

The key/parity channel has known positive interaction. Independent noisy copies of one target sign satisfy $\Delta_F = -I(Y_i; Y_j) \le 0$, and constant channels cannot have positive surplus. Independent one- and two-dimensional integrations also check the Gram expression. The dependence control preserves both conditional marginal laws and replaces their coupling by (21). Equation (6) then gives $\Delta_F = -I(Y_i; Y_j) \le 0$. This tests whether the certificate depends on within-class coupling while leaving both single-group information problems unchanged. The product uses all $4 \times 4 = 16$ source pairs per class, hence 32 joint components; a fixed permutation of four donors does not produce this law.

An independent implementation passed 13 constructed test cases, 117 comparisons between direct and Gram-matrix entropy calculations, and 12 exact polynomial moment identities. Three successive direct density quadratures agreed with the analytic formula to 3.49×10^{-14}. The quadratures are diagnostics, not certified integration bounds. The rigorous finite-channel statements use the rational enclosures. In the confirmatory results the interval widths are below 2×10^{-18} bit and do not intersect the success threshold.

For the parity control at fixed unit observation-noise variance, take independent fair signs A, B, signals $X_i = aA, X_j = aB$, and binary target $F = (1 - AB)/2$. Each conditional distribution of either group alone is identical across the two fact values, so its true information about F is zero. The unconditional cross-covariance is zero and the conditional cross-covariances are $\pm a^2$, giving $\Gamma_F = a^4$. The interval lower endpoints are:

Amplitude	B_F lower endpoint (bit)
$1/4$	0.00248874
$1/2$	0.02777026
1	0.10370406
2	-0.84306302

These displayed decimals are rounded summaries. Independent noisy copies, constant channels, and the conditional-product construction give no positive certificates. In the parity conditional-product control the target-dependent joint law disappears completely, so the true surplus is zero.

Inputs and scope of the verification. The input checks in Appendix E.1 establish a consistent finite channel. Connecting it to the intended trained states additionally requires extracting the predefined coordinate groups from each saved state in the specified source order and checking their inner products and source variances against the Gram inputs. This validates the experimental inputs; the interval calculation certifies the information bound on those inputs.

The tables report rounded summaries of the exact interval calculations. The noise-dependence curve reports exploratory floating-point calculations; only the selected setting was separately verified by interval arithmetic. Answer accuracy is a descriptive learning diagnostic and is not used in the prespecified certification criterion.

F Broader context of the knowledge question

Scope of the knowledge question. The literature uses several operational meanings of knowledge: producing a correct answer, retaining a training association, using it in a new context, or changing it through a parameter intervention. Wang et al. (2024) organizes knowledge mechanisms around utilization and evolution, while Li et al. (2025) reviews knowledge boundaries and their identification. Our experiment concerns information retained in parameter groups and accessible under a specified observation law.

Recall and generalization of factual associations. Factual probing and closed-book question answering established ways to assess answers available from model parameters (Petroni et al., 2019; Roberts et al., 2020; Jiang et al., 2020; Heinzerling & Inui, 2021). Later controlled studies expose limits of that behavioral view. Berglund et al. (2024) shows that learning an association in one direction need not support its reversal. Wang & Sun (2026) investigates this failure through a conceptual-binding hypothesis and corresponding model designs. Xu et al. (2025) distinguishes memorization, retrieval, reasoning, and association in a synthetic knowledge-injection testbed. These results motivate separating information retained about an association from its availability to a particular computation.

Controlled knowledge capacity. Allen-Zhu & Li (2025) estimates factual knowledge capacity in bits using controlled datasets and studies its dependence on model size, training, architecture, and precision. This provides a close methodological precedent for specifying the source of facts rather than treating benchmark accuracy as an intrinsic memory size. The estimands differ: capacity studies ask how much knowledge a whole model can retain, whereas we fix a three-bit source and ask about one bit in two parameter groups at a declared observation resolution.

Retrieval mechanisms and localization. Feed-forward memories and knowledge-neuron studies investigate how factual predictions depend on internal components (Geva et al., 2021; Dai et al., 2022). Geva et al. (2023) traces factual recall through subject enrichment, relation propagation, and attribute extraction, highlighting the roles of both MLP sublayers and attention. This is a question about the model's forward computation. Our analysis instead compares complete saved parameter groups: information recoverable from those groups need not be causally used by the network when it answers a query.

Knowledge editing and its relation to localization. ROME edits individual factual associations (Meng et al., 2022); MEMIT extends parameter editing to many associations across multiple layers (Meng et al., 2023). However, Hase et al. (2023) finds that localization by causal tracing does not generally identify the best layer for an edit. Thus edit success, causal localization, and information in a parameter observation measure different properties. Editing asks how an intervention changes behavior; our information experiment asks what an observation reveals about the training fact.

G Literal symbolic queries

The prefixes below are fixed across all source assignments. Alphanumeric strings are object and group identifiers; `F0`, `F1`, and `F2` specify field order. They are names and format markers, not program code. The actual inputs use symbolic token IDs.

Table 2: The three fixed training queries and their two possible target answers. The displayed prefixes spell out the symbolic inputs in readable form. Each query has two permitted answers across datasets; the model has eight output classes in every dataset.

Query	Readable training prefix	Choice 1	Choice 2
1	`F1 shop 338B city C01 group 6A56 answer:`	1	2
2	`F2 group E21D shop 0B04 city C02 answer:`	5	6
3	`F0 city C00 group 3356 shop 7312 answer:`	1	2

H Experimental specification and supplementary comparisons

H.1 Inputs and training

We use the datasets $D(u)$ defined in Section 3. Their label generator uses fixed city offsets $(0, 4, 0)$ for the three queries:

$$a_g(u) = \big(c(\mathrm{city}_g) + u_g\big) \bmod 8, \qquad u_g \in \{1, 2\}. \tag{30}$$

Thus the rule is exactly $(u_1, 4 + u_2, u_3)$; it adds no further source randomness. Table 2 records the literal inputs and answers.

Table 3: Disjoint parameter groups. The union is the complete model state; the confirmatory observation uses E and H.

Group	Coordinates	Contents
E	1,760	Token and positional embeddings
A	4,288	First LayerNorm, QKV projection, attention output and biases
M	8,416	Second LayerNorm, both MLP matrices and biases
H	320	Final LayerNorm and bias-free answer head

Each of the three symbolic queries is associated with one fixed training object. An input has 14 tokens: a format token, city, group and object fields in a fixed format-dependent order, and an answer marker. The vocabulary contains 41 symbols. The fixed label rule and binary source offsets determine labels by (30); Table 2 gives the fixed prefixes and their possible answers. Changing an offset changes the corresponding label while leaving the input and its presentation count unchanged.

Each model receives 128 presentations of each group, for 384 in total. A shuffled permutation of the three groups is drawn on each pass. The loss is cross-entropy on the eight answer logits at the last input position; the target answer is absent from the input. Stored diagnostic queries use the same objects in another field order.

The width-32 model has one pre-normalized causal transformer block with four query/key/value attention heads and a GELU feed-forward layer of width 128. It uses learned token and positional embeddings, LayerNorm with $\epsilon = 10^{-5}$, an untied bias-free eight-class output head, and no dropout. Embedding and linear weights are initialized from $\mathcal{N}(0, 0.02^2)$; linear biases start at zero.

Training uses 24 AdamW updates, effective batch size 16, and microbatch size 8. A one-update warmup reaches learning rate 10^{-3}, followed by cosine decay to 10^{-4} at update 24. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient-norm clipping at 1. Computation is FP32 on one CPU thread, with the mathematical attention backend, deterministic operations, and no autocast, fused AdamW, or `foreach` AdamW. Checkpoints are saved at steps $0, 1, 2, 4, 8, 16, 24$.

For each confirmatory replicate, initialization and presentation-order seeds were drawn independently and uniformly from 32-bit integers using the operating-system random source, with replacement. All 20 pairs were allocated once; repeated values would have been retained. The same pair was reset across all eight assignments within its replicate.

H.2 Development selection and saved outcomes

The 22 development replicates comprise two calibration replicates and 20 archived training repetitions. The final screen crossed seven nonempty parity masks of the three source bits, six parameter-group

pairs, and 17 noise levels $\rho = 2^k$, $k = -8, \ldots, 8$, at step 24. Three masks select individual coordinates, giving $22 \times 7 \times 6 \times 17 = 15{,}708$ dependent combinations.

Of these binary64 lower bounds, 4,912 were positive, including 1,320 individual-coordinate cases. The previously evaluated transport, pairwise-entropy, and moment/kernel screens found no positive lower bounds on that grid. These exploratory outcomes are not individual interval certificates. The selected E/H third-coordinate setting at $\rho = 8$ was subsequently interval-certified on all 22 development replicates. No pair, target, noise level, checkpoint, or threshold was changed using the new outcomes.

H.3 Penalty budget

Table 4 separates the Gaussian interaction from the three subtracted penalties. Its descriptive examples include the development channel, replicate 1, replicate 2 (largest lower endpoint), and replicate 12 (below threshold).

Table 4: Certificate components at $\rho = 8$, in 10^{-3} bit. $\overline{R}_{i|F} = \mathbb{E}_F R_{P_{i|F}}$. Values are rounded from saved rigorous intervals; the displayed rounded terms are not themselves an interval calculation.

Replicate	Δ_G	$R_{P_{EH}}$	$\overline{R}_{E\|F}$	$\overline{R}_{H\|F}$	B_F
Development	0.6673833	0.2709291	0.0211634	0.0380997	0.3371911
1	0.5753374	0.2921109	0.0191382	0.0295299	0.2345584
2	0.8439374	0.3092966	0.0194648	0.0403920	0.4747840
12	0.4307701	0.3118183	0.0183955	0.0383312	0.0622251

In the development example, an interaction of 6.674×10^{-4} bit minus penalties totaling 3.302×10^{-4} leaves 3.372×10^{-4} bit. The joint unconditional penalty is the largest subtraction. Replicate 12 has a smaller interaction and therefore less remaining margin, explaining its threshold failure.

H.4 Taylor comparison

These development and synthetic comparisons are separate from the prespecified 20-replicate experiment. The affine-span reduction, entropy floors, valid density ceiling, and upper-bound construction are specified in Section 7 and Appendix D. The amplitude-2 reversal prevents a claim of uniform dominance.

Table 5: Separate development and synthetic comparisons, in bits. Taylor entries use the degree-10 one-sided bound. The neural Taylor value is also unchanged at degrees 2, 4, and 8.

Channel	Taylor bound	Proposed bound
Development example, E/H, $\rho = 8$	-0.17692721	0.00033719
Parity, $a = 1/2$, unit noise	-0.09530575	0.02777026
Parity, $a = 1$, unit noise	-0.12754946	0.10370406
Parity, $a = 2$, unit noise	0.08428519	-0.84306302

H.5 Completed computation and numerical reporting

All 160 confirmatory models and 1,120 checkpoints completed. The series had a reported total runtime of 113.52 seconds on one CPU thread. The listed stage timings were 84.30 seconds for training and validation, 0.97 seconds for geometry extraction, and 26.83 seconds for interval evaluation, totaling 112.10 seconds. The difference from the reported total is 1.42 seconds. The runtime used PyTorch 2.11.0 with CPU FP32 operations; checkpoint files occupied approximately 219 MB. These measured stages reflect the small finite task and are not a per-method timing benchmark.

The smallest exact rational lower endpoint is attained by replicate 12; its decimal expansion begins 0.000062225073447203 bit. It exceeds the rounded-down 6.2×10^{-5}-bit common margin in (22), while failing the prespecified 10^{-4}-bit threshold. The mean signed lower endpoint is 0.0003170552

bit. These are bounds for the declared observation channels, not estimates of true surplus or recall benefit.

I REPLICATION OUTCOMES AND DESCRIPTIVE LEARNING CHECK

Table 6 reports all 20 outcomes. Decisions use the exact saved endpoints, not the rounded display. The confidence interval concerns the probability that a new independent training repetition crosses the fixed threshold. It does not treat the 160 models or 1,120 checkpoints as independent observations, and it is not an interval for the mean true Δ_F.

Table 6: All replication outcomes at E/H, $F = U_3 - 1$, step 24, $\rho = 8$. The primary threshold is 10^{-4} bit.

Replicate	Lower endpoint (bit)	Threshold met
1	0.0002345584	Yes
2	0.0004747840	Yes
3	0.0002878647	Yes
4	0.0002488294	Yes
5	0.0003085870	Yes
6	0.0003231705	Yes
7	0.0003454835	Yes
8	0.0003097979	Yes
9	0.0003567508	Yes
10	0.0003824253	Yes
11	0.0002269977	Yes
12	0.0000622251	No
13	0.0002176049	Yes
14	0.0002498214	Yes
15	0.0003841542	Yes
16	0.0004160903	Yes
17	0.0004121887	Yes
18	0.0003006373	Yes
19	0.0003873572	Yes
20	0.0004117757	Yes

Table 7: Descriptive accuracy of saved alternate-format queries. Each of 160 models contributes three predictions at each checkpoint. Queries and checkpoints are dependent; no confidence interval is assigned to these pooled counts.

Update	0	1	2	4	8	16	24
Correct / 480	68	236	256	274	370	458	473
Accuracy (%)	14.17	49.17	53.33	57.08	77.08	95.42	98.54