RECOVERABLE BUT UNREADABLE: EXACT ACCESSIBILITY LAWS IN NOISY LATENT PARITY CODES AND BEYOND

Anonymous authors
Paper under double-blind review

ABSTRACT

A representation can contain nearly all information about a latent factor while exposing none of it to a simple probe. We distinguish recoverability, the Bayes signal in the full representation, from accessibility, the posterior signal visible through a restricted readout interface, and separate both from finite-sample learnability. We study this gap in noisy latent parity codes. Under the uniform prior, the first order at which a target becomes readable is exactly the minimum number of observed parities whose xor synthesizes an active Walsh component of the target. This witness complexity governs both coordinate-local readout and low-order interaction readout. For balanced binary targets, every sub-threshold local view is independent of the target, giving zero mutual information and chance Bayes accuracy. For random dense mixing below channel capacity, the full latent vector is decodable with exponentially small error, while each coordinate remains unreadable to every sub-threshold local or low-order interface. In the BSC, first witnesses can be exponentially weak, so reveal need not imply practical learnability. We also use the family as an exact probing benchmark and show, including through a continuous learned bridge, that accessibility-aware objectives can reduce low-order exposure while preserving strong global utility across both exact and learned settings.

1 INTRODUCTION

When should we say that a latent factor is "in" a representation? In practice, this phrase often conflates two claims. One is that the factor is recoverable from the full representation by a Bayes-optimal decoder. The other is that the factor is exposed to the restricted interface used by a simple probe, such as a local, linear, or low-degree readout (Hewitt & Manning, 2019; Voita & Titov, 2020; Belinkov, 2022). These claims need not agree. We separate three notions. Let Z denote the latent state, Y the observed representation, and $T = T(Z)$ a target factor. We distinguish:

- **recoverability**: posterior signal about T available in the full representation Y;
- **accessibility**: posterior signal about T exposed through a specified probe interface;
- **learnability**: whether finite data and computation can exploit the accessible signal.

Recoverability and accessibility are population-level quantities; finite-sample behavior is analyzed through witness amplitudes and trained probes.

Our main setting is a noisy latent parity code. A binary matrix $A \in \mathbb{F}_2^{m\times k}$ maps k latent bits to m observed parity coordinates, which are then passed through independent informative binary-input channels. For a latent parity indexed by $q \in \mathbb{F}_2^k$, define the witness complexity

$$r_q(A) := \min\{|v| : v \in \mathbb{F}_2^m,\ A^\top v = q\},$$

where $|v|$ is the Hamming weight of v and all linear algebra is over $\mathbb{F}_2$. Thus $r_q(A)$ is the fewest observed parities whose xor synthesizes the latent parity q. For a general target T, the relevant threshold is the minimum $r_q(A)$ over the active Walsh components of T.

This witness complexity gives an exact answer to the question: *at what order does a target first become readable?* Under the uniform prior on Z, the first readable order is the same for two

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.

interfaces: coordinate-local probes that read at most r observed coordinates, and score-monomial interaction probes of order at most r. For balanced binary targets, the sub-threshold statement is stronger than zero correlation: every sub-threshold local view is independent of the target, has zero mutual information, and gives Bayes accuracy exactly $1/2$.

These laws yield a sharp separation between global information and simple readout. For random dense A below channel capacity, the full latent vector is decodable from Y with exponentially small block error, so each coordinate is nearly fully recoverable from the complete representation. At the same time, with high probability, every coordinate remains unreadable from every sub-threshold local interface and, in the BSC case, from every sub-threshold low-degree polynomial interface. Moreover, the first nonzero witness can have exponentially small correlation, so first reveal need not imply easy finite-sample exploitation.

The parity family also provides a controlled probing and encoder-design benchmark. Since accessibility can be computed directly, encoder objectives can be compared without probe-training confounds. We show that accessibility-aware objectives suppress low-order exposure while preserving global utility in parity-code designs, in learned parity encoders after hardening, and in a continuous synthetic representation experiment beyond parity. Our contributions are:

- We define *accessibility* as interface-relative posterior signal and characterize it through orthogonal projection of the posterior mean.
- For noisy latent parity codes, we prove that the first readable order of any target equals its minimum active witness weight; for balanced binary targets, every sub-threshold local view is independent of the target.
- For random dense mixing below capacity, we establish near-complete global recoverability with exact sub-threshold local and low-order unreadability; in the BSC, first witnesses can be exponentially weak.
- We develop an exact probing and design benchmark and show that accessibility-aware objectives suppress low-order exposure in both parity and continuous learned settings.

Related work. Probing work asks whether information is *present* in a representation or merely *extractable* by a chosen decoder (Alain & Bengio, 2017; Pimentel et al., 2020a;b); control tasks, MDL, and probe-complexity analyses instead emphasize finite-sample fitting and inductive bias (Hewitt & Liang, 2019; Torroba Hennigen et al., 2020; Pimentel & Cotterell, 2021; Cao et al., 2021; White et al., 2021; Immer et al., 2022). Accessibility measures the population posterior signal exposed by an interface before probe optimization. Our analysis relates recoverability to correlation ratio and explained variance (Rényi, 1959), and uses Boolean Fourier methods (O'Donnell, 2014) together with coding-theoretic notions of syndrome witnesses, coset leaders, random-code thresholds, and below-capacity decoding (MacWilliams & Sloane, 1977; Gallager, 1968; Richardson & Urbanke, 2008). The contribution is a representation-learning interpretation in which one witness structure governs local and low-order accessibility while separating global utility, readout exposure, and learnability. This provides an analyzable model of distributed and superposed representations (Hinton et al., 1986; Elhage et al., 2022; Bricken et al., 2023) without probe-training confounds.

2 Recoverability, accessibility, and probe interfaces

Throughout, for any nonconstant target T, write $T^\circ = T - \mathbb{E}[T]$ and $\widetilde{T} = T^\circ / \|T^\circ\|_{L_2}$. Quantitative recoverability and accessibility are computed for $\widetilde{T}$. Structural reveal-order statements use T°, since only Fourier support matters.

Definition 1 (Recoverability and accessibility)**.** For a centered unit-variance target T and representation Y, define

$$\mathrm{Rec}(T;Y) := \|\mathbb{E}[T \mid Y]\|_{L_2(P_Y)}.$$

If $V \subseteq L_2(P_Y)$ is a closed subspace, define

$$\mathrm{Acc}_V(T;Y) := \sup_{\substack{f \in V \\ \|f\|_{L_2(P_Y)} \le 1}} \mathbb{E}[f(Y)T].$$

Recoverability measures the total Bayes signal available in the full representation. Accessibility measures how much of that signal survives when we restrict attention to a probe interface V.

Proposition 1 (Posterior projection principle)**.** *For every closed $V \subseteq L_2(P_Y)$,*

$$\mathrm{Acc}_V(T;Y) = \|\Pi_V \mathbb{E}[T \mid Y]\|_{L_2(P_Y)} \le \mathrm{Rec}(T;Y),$$

where Π_V is the orthogonal projection onto V. In particular,

$$\mathrm{Rec}(T;Y)^2 = \mathrm{Var}(\mathbb{E}[T \mid Y]) = 1 - \mathbb{E}[(T - \mathbb{E}[T \mid Y])^2].$$

So accessibility is not a fitted-probe score. It is the size of the posterior mean after projection onto the allowed interface.

Two interfaces used in this paper. For a coordinate subset $O \subseteq [m]$, define

$$\mathrm{Acc}_O(T) := \left\| \mathbb{E}[\widetilde{T} \mid Y_O] \right\|_{L_2}, \qquad \mathrm{Loc}_r(T) := \sup_{|O| \le r} \mathrm{Acc}_O(T).$$

Thus $\mathrm{Loc}_r(T)$ asks whether any measurable probe reading at most r coordinates can extract centered signal about T. For interaction probes, we define score coordinates in Section 3 and use the subspace

$$V_r^{\mathrm{int}} := \mathrm{span}\{S_v : v \in \mathbb{F}_2^m,\ |v| \le r\}, \qquad \mathrm{Int}_r(T) := \mathrm{Acc}_{V_r^{\mathrm{int}}}(\widetilde{T}; Y).$$

In the BSC case this is exactly the space of degree-$\le r$ multilinear polynomials in the raw observations. For a binary target B with alphabet $\mathcal{B}$, define $\mathrm{BayesAcc}(B \mid \mathscr{H}) := \sup_{g:\, g\ \mathscr{H}\text{-measurable},\, g\ \mathcal{B}\text{-valued}} \mathbb{P}(g = B)$.

Proposition 2 (Balanced binary Bayes bridge)**.** *Let $B \in \{\pm 1\}$ be balanced and let $\mathscr{H}$ be any sigma-field. Then*

$$\mathrm{BayesAcc}(B \mid \mathscr{H}) = \tfrac{1}{2}\big(1 + \mathbb{E}|\mathbb{E}[B \mid \mathscr{H}]|\big).$$

For a coordinate interface O,

$$\tfrac{1}{2}\mathrm{Acc}_O(B)^2 \le \mathrm{BayesAcc}(B \mid Y_O) - \tfrac{1}{2} \le \tfrac{1}{2}\mathrm{Acc}_O(B).$$

Thus zero local accessibility gives Bayes accuracy $1/2$, and small accessibility forces near-chance accuracy.

The parity model below is one exact instance of a more general phenomenon. Recoverability can be high when the full representation identifies the latent state, while accessibility through an interface family can be low when the induced class-conditional marginals are indistinguishable on that family.

Proposition 3 (Global recovery with interface indistinguishability)**.** *Let H be a finite latent state and let $B = B(H) \in \{\pm 1\}$ satisfy $\mathbb{P}(B = +1) = \mathbb{P}(B = -1) = 1/2$. Let $Y \sim P_h$ conditional on $H = h$. Let $\mathcal{I}$ be any family of interfaces, and define $P_{+,I} := \mathcal{L}(Y_I \mid B = +1)$ and $P_{-,I} := \mathcal{L}(Y_I \mid B = -1)$. If a full decoder $\widehat{H}(Y)$ satisfies $\mathbb{P}(\widehat{H}(Y) \ne H) \le \delta$, then $\mathrm{Rec}(B;Y) \ge 1 - 2\delta$. If additionally $\sup_{I \in \mathcal{I}} \mathrm{TV}(P_{+,I}, P_{-,I}) \le \varepsilon$, then*

$$\sup_{I \in \mathcal{I}} \mathrm{BayesAcc}(B \mid Y_I) \le \tfrac{1}{2} + \tfrac{\varepsilon}{2}.$$

Proof. Let $\widehat{B}(Y) := B(\widehat{H}(Y))$. Then $\mathbb{P}(\widehat{B} \ne B) \le \delta$, hence $\mathbb{E}[\widehat{B}B] \ge 1 - 2\delta$. Since $\|\widehat{B}\|_{L_2} = 1$ and B is centered with unit variance,

$$\mathrm{Rec}(B;Y) = \sup_{\|f\|_{L_2(P_Y)} \le 1} \mathbb{E}[f(Y)B] \ge \mathbb{E}[\widehat{B}(Y)B] \ge 1 - 2\delta.$$

For a balanced binary classification problem,

$$\mathrm{BayesAcc}(B \mid Y_I) = \frac{1}{2} + \frac{1}{2}\mathrm{TV}(P_{+,I}, P_{-,I}).$$

Taking the supremum over $I \in \mathcal{I}$ gives the claim. □

The exact parity laws below identify when the interface indistinguishability term is exactly zero for coordinate-local and low-order interaction interfaces, and when the first nonzero witness appears. Thus the role of the parity model is not only to exhibit a separation, but to make the separation exactly computable. Global recovery will be controlled by the usual decoding problem for the full noisy codeword, whereas accessibility will be controlled by which latent Walsh components can be synthesized by a small visible witness. This lets us compare full posterior information, interface-level posterior information, and finite-sample probe behavior within the same model.

3 NOISY LATENT PARITY CODES

Let $Z \in \{\pm 1\}^k$ have prior π, uniform unless stated otherwise, and let $A \in \mathbb{F}_2^{m\times k}$ have rows $a_1, \dots, a_m$. For $q \in \mathbb{F}_2^k$, write $\chi_q(Z) := \prod_{\ell=1}^k Z_\ell^{q_\ell}$. The observation model is $U_j = \chi_{a_j}(Z)$; conditional on Z, the outputs $Y_1, \dots, Y_m$ are independent with $Y_j \sim W(\cdot \mid U_j)$, $j = 1, \dots, m$. The channel output space is $\mathcal{Y}$.

Assumption 1 (Channel regularity)**.** The channel laws $W(\cdot \mid +1)$ and $W(\cdot \mid -1)$ are dominated by a common sigma-finite measure λ. Write $w(y \mid u)$ for the corresponding densities or masses. The score identities and exact reveal laws below only use this dominated binary-input channel structure together with informativeness. Binary-input symmetry in the standard coding-theoretic sense is needed only for the random-coding capacity statement in Section 5. We assume W is informative.

Define $\bar{w}(y) := (w(y \mid +1) + w(y \mid -1))/2$, let $\bar{W}$ have density or mass $\bar{w}$, and set

$$s(y) := \frac{w(y \mid +1) - w(y \mid -1)}{w(y \mid +1) + w(y \mid -1)}$$

on the set where the denominator is positive, and arbitrarily elsewhere. Then $w(y \mid u) = \bar{w}(y)(1 + u\, s(y))$ λ-a.e. Write $S_j := s(Y_j)$ and $\kappa := \mathbb{E}_{\bar{W}}[s(Y)^2]$; informative means $\kappa > 0$.

3.1 WHY THE SCORE IS THE RIGHT LOCAL COORDINATE

A convenient fact about this model is that the local posterior depends on Y_O only through the score vector S_O.

Theorem 1 (Coordinatewise score sufficiency)**.** *For every $O \subseteq [m]$ and every square-integrable target $T(Z)$,*

$$\mathbb{E}[T \mid Y_O] = \mathbb{E}[T \mid S_O] \qquad \textit{a.s.}$$

Thus all latent-predictive information in a coordinate-local interface is preserved by the score vector.

Under the uniform prior, write $\widehat{T}(q) := \mathbb{E}[T^\circ(Z)\chi_q(Z)]$ and $\mathrm{Spec}(T) := \{q \in \mathbb{F}_2^k \setminus \{0\} : \widehat{T}(q) \neq 0\}$. For $v \in \mathbb{F}_2^m$, let $S_v := \prod_{j=1}^m S_j^{v_j}$ and $|v| = \mathrm{wt}(v)$.

Theorem 2 (Interaction identity)**.** *Assume the uniform prior on Z. For every nonconstant target T and every $v \in \mathbb{F}_2^m$,*

$$\mathbb{E}[S_v(Y)\, T^\circ(Z)] = \kappa^{|v|}\, \widehat{T}(A^\top v).$$

In particular, for the parity target $T_q = \chi_q(Z)$ with $q \neq 0$,

$$\mathbb{E}[S_v(Y)\, T_q(Z)] = \kappa^{|v|}\, \mathbf{1}\{A^\top v = q\}.$$

This identity is the algebraic backbone of the paper. It says that a score monomial can correlate with a target only if its syndrome $A^\top v$ lands on an active Walsh component of that target.

For the BSC(ρ), $\rho < 1/2$, with reliability $\mu = 1 - 2\rho$, we have $s(y) = \mu y$ on $\{\pm 1\}$, so V_r^{int} is exactly the space of degree-$\le r$ multilinear polynomials in the raw observations.

3.2 A SMALL EXAMPLE TO KEEP IN MIND

Example 1 (Globally present, still unreadable from one coordinate)**.** Take $k = 3$, $m = 4$, and rows

$$a_1 = e_1 + e_2, \qquad a_2 = e_1 + e_3, \qquad a_3 = e_2 + e_3, \qquad a_4 = e_1 + e_2 + e_3.$$

Then A has rank 3, so the clean parity vector already determines the full latent vector. For the target $Z_1 = \chi_{e_1}(Z)$,

$$e_1 = a_3 + a_4,$$

so Z_1 can be synthesized from two observed parities. No single row equals e_1, so one coordinate is not enough. The witness threshold is therefore 2. The consequences are immediate. Every single-coordinate interface Y_j is completely uninformative about Z_1 under the uniform prior:

$$Z_1 \perp Y_j, \qquad \mathrm{BayesAcc}(Z_1 \mid Y_j) = \tfrac{1}{2}, \qquad j = 1, \dots, 4.$$

But the two-coordinate interface (Y_3, Y_4) already carries nontrivial posterior signal about Z_1, and in the BSC the first nonzero raw witness appears at degree 2.

4 EXACT REVEAL LAWS

For $q \in \mathbb{F}_2^k$, define the witness complexity $r_q(A) := \min\{|v| : A^\top v = q\}$, with value $+\infty$ if no witness exists. For a nonconstant target T, define $r_T(A) := \min_{q \in \mathrm{Spec}(T)} r_q(A)$, with the convention $\min \emptyset = +\infty$. The quantity $r_q(A)$ is the fewest observed parities needed to produce latent parity q. The quantity $r_T(A)$ is the first order at which any active component of T can be exposed.

Theorem 3 (Unified exact reveal)**.** *Assume the uniform prior on Z. For every nonconstant target T,*

$$\min\{r : \mathrm{Loc}_r(T) > 0\} = \min\{r : \mathrm{Int}_r(T) > 0\} = r_T(A).$$

Equivalently, if $r < r_T(A)$, then every measurable probe reading at most r coordinates has zero population correlation with T°, and every score-monomial interaction of order at most r also has zero population correlation with T°. If $r \geq r_T(A)$, then both the local and interaction interfaces contain nontrivial posterior-readout signal. For parity targets, this reduces to $r_q(A)$.

Thus local observation and low interaction order have the same exact reveal threshold.

Corollary 1 (Balanced binary exact local independence law)**.** *Let $B(Z) \in \{\pm 1\}$ be any balanced binary target under the uniform prior and define*

$$r_B(A) := \min_{q \in \mathrm{Spec}(B)} r_q(A).$$

Then

$$\min\{r : \mathrm{Loc}_r(B) > 0\} = r_B(A).$$

Moreover, if $O \subseteq [m]$ and $|O| < r_B(A)$, then

$$B \perp Y_O, \qquad \mathrm{I}(B; Y_O) = 0, \qquad \mathrm{BayesAcc}(B \mid Y_O) = \tfrac{1}{2}.$$

Thus below threshold there is not merely no linear signal, but no readable local signal of any kind.

Corollary 2 (No finite-sample local generalization below threshold)**.** *Assume the uniform prior and let $B(Z) \in \{\pm 1\}$ be balanced. Fix $r < r_B(A)$. Let $\mathcal{A}$ be any possibly randomized training procedure which, given any number n of labeled samples and any side information about A and the channel, outputs a subset $O \subseteq [m]$ with $|O| \leq r$ and a measurable classifier $g : \mathcal{Y}^O \to \{\pm 1\}$. For an independent test sample,*

$$\mathbb{P}\big(g(Y_O) = B\big) = \tfrac{1}{2},$$

where the probability is over the training data, algorithmic randomness, and the independent test sample. Thus below the reveal threshold, no amount of probe-training data can produce a coordinate-local classifier with nontrivial test accuracy, even when the visible subset and classifier are chosen adaptively from the training set.

The exact zero-leakage statements above use the uniform prior. This assumption makes Walsh components exactly orthogonal and turns sub-threshold unreadability into exact independence. Under biased product priors, exact zeros generally become quantitative indistinguishability bounds.

Suppose instead that the latent bits are independent with $\mathbb{E}Z_\ell = \nu_\ell$ and $|\nu_\ell| \leq \bar{\nu} < 1$. For a coordinate target Z_i and observed subset O, define

$$\delta_i(O) := \min\{|t| : t \in \mathrm{span}\{a_j : j \in O\},\ t_i = 1\},$$

with value $+\infty$ if no such t exists. If $P_{Y_O}^{(i,+)}$ and $P_{Y_O}^{(i,-)}$ are the laws of Y_O conditional on $Z_i = +1$ and $Z_i = -1$, then

$$\mathrm{TV}(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}) \leq 2^{|O|/2} \bar{\nu}^{\,\delta_i(O)-1},$$

with the right-hand side interpreted as 0 when $\delta_i(O) = +\infty$. Thus, under biased product priors, exact zeros soften into quantitative indistinguishability whenever any visible combination involving Z_i must involve many other latent bits. The uniform-prior exact law is recovered when no visible combination equals e_i. A full statement and proof are in Section F.

5 RANDOM DENSE MIXING: GLOBAL RECOVERY WITHOUT SIMPLE READOUT

The exact reveal law becomes especially striking for random dense mixing. Let $A \in \mathbb{F}_2^{m\times k}$ have i.i.d. Bernoulli$(1/2)$ entries and rate $R = k/m \in (0,1)$. Write $\alpha_\star(R) := H_2^{-1}(R)$, where $H_2(x) := -x\log_2 x - (1-x)\log_2(1-x)$ and H_2^{-1} is the inverse branch on $[0, 1/2]$.

Theorem 4 (Random dense witness threshold)**.** *For every fixed $\varepsilon > 0$, with high probability over random dense A, simultaneously for all coordinates $i \in [k]$,*

$$(\alpha_\star(R) - \varepsilon)m \le r_{e_i}(A) \le (\alpha_\star(R) + \varepsilon)m.$$

The same statement holds for any polynomial-size family of nonzero syndromes fixed independently of A.

Thus the first coordinate witness typically has weight linear in m.

We also use the classical fact that random linear-code ensembles below channel capacity achieve exponentially small ensemble-average block error under maximum-likelihood decoding (Richardson & Urbanke, 2008). Since the average error already decays as $e^{-\Omega(m)}$, a standard Markov argument upgrades this to a high-probability statement over the random code draw; see Section D.

Theorem 5 (Recoverable, yet unreadable below threshold)**.** *Assume random dense mixing over an informative binary-input symmetric channel W satisfying Assumption 1, with rate $R < C(W)$, where $C(W)$ is its capacity. Then with high probability over A:*

(i) the full latent vector is decodable from Y with block error $e^{-\Omega(m)}$;

(ii) every coordinate is nearly fully recoverable,

$$\mathrm{Rec}(Z_i; Y) \ge 1 - e^{-\Omega(m)} \qquad \forall i \in [k];$$

(iii) for every fixed $\varepsilon > 0$, every coordinate i, and every subset $O \subseteq [m]$ with

$$|O| < (\alpha_\star(R) - \varepsilon)m,$$

one has

$$Z_i \perp Y_O, \qquad \mathrm{BayesAcc}(Z_i \mid Y_O) = \tfrac{1}{2};$$

(iv) for every fixed $\varepsilon > 0$ and every i,

$$\mathrm{Int}_r(Z_i) = 0 \qquad \textit{for all } r < (\alpha_\star(R) - \varepsilon)m.$$

In the BSC, this includes every degree-r multilinear polynomial in the raw observations.

Thus the complete observation nearly reveals each coordinate, while every sub-threshold local interface reveals nothing. The reveal threshold is a population accessibility threshold. Below it, the specified interface contains no posterior signal. At the first reveal order, however, the available signal can be very small. The next corollary quantifies this sample barrier for direct correlation-based exploitation of a specified BSC witness.

Corollary 3 (Sub-threshold impossibility and first-witness sample barrier)**.** *Assume the channel is $\mathrm{BSC}(\rho)$, $\rho < 1/2$, with reliability $\mu = 1 - 2\rho$. For a parity target T_q, every raw multilinear monomial of degree below $r_q(A)$ has zero population correlation with T_q. At degree $r_q(A)$, every first witness has correlation exactly $\mu^{r_q(A)}$.*

Consequently, any procedure that attempts to exploit a specified first witness by direct correlation estimation from labeled samples requires

$$n = \Omega\Big(\mu^{-2r_q(A)}\Big)$$

samples to achieve constant signal-to-noise ratio. Under random dense mixing below capacity, $r_{e_i}(A) = \Theta(m)$ with high probability, so the first witness correlation is exponentially small and direct witness exploitation requires $e^{\Omega(m)}$ samples.

Table 1: Representative fixed and refined encoders. Dense and access-aware models both stay unreadable up to order 2; systematic and global-only encoders expose every coordinate immediately.

Model	Block acc.	Mean Rec	$\mathrm{LocalBayes}_{\leq 2}$	$\mathrm{DegCorr}_{\leq 2}$	Mean r_i	N_1	$N_{\leq 2}$
Systematic	0.8397	0.9524	0.9500	0.9000	1.0	10	10
Dense reference	0.8636	0.9077	**0.5000**	**0.0000**	3.2	**0**	**0**
Global-only	0.7560	0.9447	0.9500	0.9000	1.0	10	10
Access-aware	0.7309	0.8402	**0.5000**	**0.0000**	3.7	**0**	**0**

Table 2: Aggregate refinement statistics over 20 seeds. Global-only refinement remains exposed, while accessibility-aware refinement raises witness thresholds and suppresses low-order readout.

Mode	Block acc.	Mean Rec	$\mathrm{LB}_{\leq 2}$	$\mathrm{DC}_{\leq 2}$	$\bar{r}$	N_1	$N_{\leq 2}$
Glob, 0.00	0.696 ± 0.016	0.942 ± 0.001	0.950	0.900	1.0	10	10
Glob, 0.02	0.694 ± 0.024	0.942 ± 0.003	0.948 ± 0.002	0.896 ± 0.005	1.1 ± 0.1	9.5	10
Glob, 0.05	0.707 ± 0.044	0.940 ± 0.003	0.948 ± 0.002	0.896 ± 0.005	1.1 ± 0.1	9.5	10
Aware, 0.00	0.731 ± 0.009	0.841 ± 0.022	0.574 ± 0.035	0.149 ± 0.069	3.45 ± 0.33	0.75	1.75
Aware, 0.02	0.732 ± 0.013	0.832 ± 0.013	0.522 ± 0.023	0.045 ± 0.045	3.70 ± 0.19	0.5	0.5
Aware, 0.05	0.717 ± 0.010	0.831 ± 0.043	0.597 ± 0.035	0.194 ± 0.069	3.55 ± 0.36	1.25	2.25

6 Exact benchmark experiments and exposure design

The experiments serve two purposes. First, within the parity family, they illustrate the exact benchmark and test whether different encoder objectives lead to different exposure geometries inside the same analyzable model class. Second, through a continuous learned bridge beyond parity, they ask whether the same qualitative distinction between global recoverability and local accessibility remains visible in a neural representation setting. Code is provided in the Supplementary.

6.1 Evaluation protocol and metrics

Unless otherwise stated, all parity benchmark experiments use $k = 10$, $m = 16$, and a $\mathrm{BSC}(0.05)$, so the raw reliability is $\mu = 1 - 2\rho = 0.9$. All order-≤ 2 readability quantities are computed by exhaustive enumeration over all 2^{10} latent states and all coordinate subsets or monomials of order at most 2; reveal-order checks use complete witness search. Full-state block decoding and mean recoverability are estimated by Monte Carlo with exact posterior computations on the finite latent space. Aggregate refinement statistics are computed over 20 independent binary-refinement restarts per setting, and the learned-bridge sweeps use 20 independent seeds per λ_{wit}.

For a parity encoder A with coordinate targets $Z_1, \ldots, Z_k$, we report

$$\mathrm{LocalBayes}_{\leq 2}(A) := \frac{1}{k}\sum_{i=1}^{k} \max_{|O|\leq 2} \mathrm{BayesAcc}(Z_i \mid Y_O), \quad \mathrm{DegCorr}_{\leq 2}(A) := \frac{1}{k}\sum_{i=1}^{k} \max_{|v|\leq 2} \big|\mathbb{E}[X_v Z_i]\big|,$$

where $X_v = \prod_j Y_j^{v_j}$, together with $N_1(A) := \sum_i \mathbf{1}\{r_i(A) = 1\}$, $N_{\leq 2}(A) := \sum_i \mathbf{1}\{r_i(A) \leq 2\}$, and $r_i(A) := r_{e_i}(A)$.

For coordinate targets under the uniform prior, $\mathrm{DegCorr}_{\leq 2}(A) = 0$ if and only if every degree-≤ 2 monomial witness vanishes. We also check theorem consistency on every coordinate target: whether the first nontrivial local Bayes reveal order matches $r_i(A)$, whether the first nontrivial degree reveal order matches $r_i(A)$, and whether all below-threshold local Bayes gaps and degree correlations are numerically zero.

6.2 Encoders and exposure-design objectives

We use two fixed references. **Systematic encoder:** the first k rows are the standard basis, so every coordinate has witness threshold 1. **Screened dense reference:** a full-rank random dense draw retained only if $N_{\leq 2}(A) = 0$, so every coordinate is unreadable from all size-2 local interfaces and every degree-2 monomial witness vanishes.

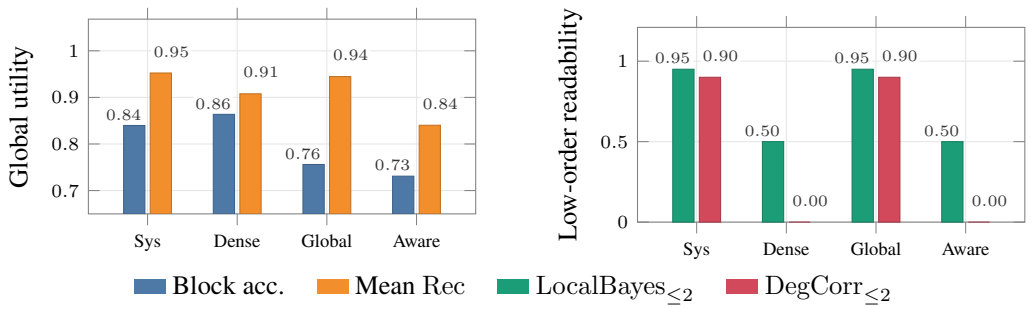

Figure 1: Exact benchmark and design results. Left: global utility. Right: low-order readability. The dense reference and the accessibility-aware encoder preserve strong global performance while remaining unreadable up to order 2. The systematic and global-only encoders are quickly exposed.

Table 3: Theorem-consistency checks for the representative models. First local and degree reveal orders match the witness threshold; below-threshold violations are numerically zero.

Model	Local matches	Degree matches	Max local below-thr. gap	Max degree below-thr. corr.
Systematic	10/10	10/10	0	0
Dense ref.	10/10	10/10	9.99×10^{-16}	0
Global-only	10/10	10/10	0	0
Access-aware	10/10	10/10	0	0

The main design experiment keeps the parity family fixed and changes only the encoder objective. We use a two-stage pipeline: a soft initialization proposes a parity matrix, then exact binary local refinement optimizes the hardened matrix. Every final claim is made on the resulting binary encoder. The *global-only* mode optimizes global code quality, namely minimum and average induced distance, row diversity, and full rank, and does not penalize low-order exposure. The *accessibility-aware* mode augments the same score with explicit penalties on $N_1(A)$ and $N_{\leq 2}(A)$, so low-order witness structure becomes a design target. We sweep an initialization-bias hyperparameter $\eta \in \{0, 0.02, 0.05\}$ with 20 independent binary-refinement restarts per setting. For each mode we report aggregate statistics by setting and one representative encoder from the full mode-level sweep. The representative global-only encoder is chosen for high block accuracy, while the representative access-aware encoder is chosen among runs satisfying $N_{\leq 2}(A) = 0$. Full selection details are given in Section G.

Tables 1 and 2 summarize the results, Figure 1 visualizes the split, and Table 3 reports theorem-consistency checks. Global-only refinement preserves strong recoverability but almost always exposes coordinates at order 1 or 2. Accessibility-aware refinement pushes the encoder into a different part of the design space: low-order readability drops sharply, the mean reveal threshold rises from about 1 to about 3.5–3.7, and unreadable-up-to-order-2 solutions appear reliably. At $\eta = 0.02$, the mean $N_{\leq 2} = 0.5$ across 20 runs implies only ten low-order coordinate leaks over the entire sweep, so at least half of the runs are fully unreadable up to order 2. None of the global-only sweeps produce such a run.

6.3 Finite-sample probes mirror the exact accessibility picture

The exact accessibility calculations predict the finite-sample behavior of trained probes surprisingly well; the resulting learning curves are shown in Figure 2. On exposed encoders, simple probes quickly exploit the available signal. The systematic encoder is solved almost perfectly by linear probes already at $n = 128$, while degree-2 probes rise from 0.7051 ± 0.0200 at $n = 128$ to 0.9498 ± 0.0003 at $n = 8192$. The global-only refined encoder behaves similarly, with degree-2 probe accuracy rising from 0.7866 ± 0.0100 to 0.9550 ± 0.0010. Unreadable-up-to-order-2 encoders look completely different. For the access-aware encoder, linear probes stay at chance throughout the sweep, while degree-2 probe accuracy is only 0.5113 ± 0.0028 at $n = 128$, 0.5043 ± 0.0016 at $n = 2048$, and 0.5034 ± 0.0034 at $n = 8192$. The dense reference likewise remains in the 0.50–0.515 range.

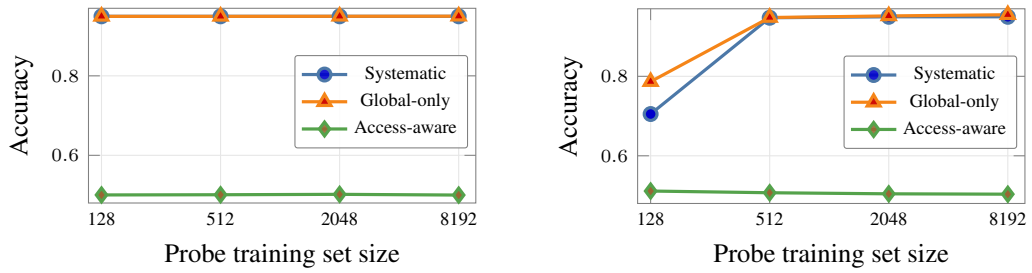

Figure 2: Finite-sample probes track exact accessibility. Left: linear probes. Right: degree-2 probes. Exposed encoders are solved quickly, while order-2-unreadable encoders stay near chance.

The curves are consistent with the population accessibility predictions: when the specified interface exposes signal, simple probes learn it quickly; when the interface has zero population accessibility, the same probes remain near chance. For coordinate-local interfaces, Corollary 2 gives the corresponding finite-sample no-generalization statement below threshold.

6.4 A LEARNED BRIDGE: WITNESS-AWARE END-TO-END TRAINING

To connect the benchmark to learned representations, we train parity encoders end-to-end through a continuous relaxation followed by hardening, using a small MLP decoder for full latent reconstruction; details are in Section G.3. This experiment asks whether accessibility-aware training can drive a learned encoder into the same unreadable regime. The answer is yes. Without witness control, the learned baseline finds distributed but still leaky encoders. With witness-aware regularization, one representative unreadable run at $\lambda_{\text{wit}} = 1.0$ reaches block accuracy 0.8619 and mean recoverability 0.8966, with $\text{LocalBayes}_{\leq 2} = 0.5000$, $\text{DegCorr}_{\leq 2} = 0$, and $\min_i r_i(A) = 3$. Thus every coordinate is unreadable from every size-2 local interface and every degree-2 witness vanishes exactly; trained linear and degree-2 probes stay at chance up to the largest scanned sample size.

6.5 A CONTINUOUS BRIDGE BEYOND THE PARITY FAMILY

We test the distinction in a continuous neural benchmark with a finite latent image world, noisy 12-dimensional representations, and exact Bayes evaluation of one- and two-coordinate local interfaces; see Section H. Across 20 seeds per variant, utility-only reshaping leaves local readability nearly unchanged, whereas accessibility-aware reshaping suppresses low-order exposure while retaining strong global signal. On the protected compositional target, full Bayes accuracy changes from 0.9390 to 0.8944, while best one-coordinate and two-coordinate Bayes accuracies drop from 0.6963 to 0.5814 and from 0.7417 to 0.6367, respectively; quadratic probe accuracy drops from 0.8708 to 0.7032. These results suggest that accessibility-aware shaping can reduce low-order local readability outside the exact parity setting as well.

7 CONCLUSION

This paper formalizes a basic distinction: *recoverability is not readout*. We capture this gap through *accessibility*, the posterior signal exposed by a chosen interface, and characterize it exactly in noisy latent parity codes. Under the uniform prior, minimum witness weight gives the first readable order for both coordinate-local and low-order interaction readout; for balanced binary targets, every sub-threshold local view is independent of the target. For random dense mixing below channel capacity, the full latent vector is decodable with exponentially small error while each coordinate remains unreadable to every sub-threshold local interface; in the BSC, first witnesses may be exponentially weak, so reveal need not imply learnability. The family also provides a controlled probing and encoder-design testbed: accessibility-aware objectives suppress low-order exposure while preserving utility. While the exact laws rely on parity closure and uniform-prior orthogonality, the parity family gives a fully analyzable setting where recoverability, accessibility, and finite-sample witness strength separate without probe-training confounds. The continuous experiment suggests that the same design perspective can also shape local readability beyond the exact parity model.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and code debugging. All AI-assisted text and code were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

Potential benefits. This work clarifies the distinction between recoverability, accessibility, and learnability in representation analysis. A likely positive impact is improved methodological rigor in probing and interpretability research: probe outcomes need not be conflated with the total information present in a representation. Our exact benchmark may also help evaluate probing methods and develop accessibility-aware representation objectives.

Potential risks. The same ideas could be used to construct representations that retain global information while hiding it from simple probes. In beneficial settings, this may support privacy or reduce unintended leakage. In harmful settings, it could make models harder to audit, interpret, or monitor with standard tools.

Limitations and scope. Our sharpest results hold in a highly structured parity-based model and should not be overgeneralized to all neural representations. The continuous learned bridge is qualitative rather than theorem-level evidence beyond parity.

Mitigation and responsible use. We therefore present accessibility-aware shaping as a neutral design axis rather than an unconditional objective. In applications involving safety, fairness, or accountability, reduced simple readability should be weighed against the need for external auditability and transparency.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl. ICLR 2017 Workshop Track.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, October 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/. Published October 4, 2023.

Steven Cao, Victor Sanh, and Alexander Rush. Low-complexity probing via finding subnetworks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 960–966, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.74. URL https://aclanthology.org/2021.naacl-main.74/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. doi: 10.48550/arXiv.2209.10652. URL https://arxiv.org/abs/2209.10652.

Robert G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968. ISBN 9780471290483.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. Distributed representations. In David E. Rumelhart and James L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pp. 77–109. MIT Press, Cambridge, MA, 1986. doi: 10.7551/mitpress/5236.003.0006.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1839–1851, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.129. URL https://aclanthology.org/2022.acl-long.129/.

F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1977. ISBN 9780444851932.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. ISBN 9781107038325. doi: 10.1017/CBO9781139814782.

Tiago Pimentel and Ryan Cotterell. A bayesian framework for information-theoretic probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2869–2887, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.229. URL https://aclanthology.org/2021.emnlp-main.229/.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3138–3153, Online, 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.254. URL https://aclanthology.org/2020.emnlp-main.254/.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3–4):441–451, 1959. doi: 10.1007/BF02024507.

Thomas J. Richardson and Rüdiger Urbanke. *Modern Coding Theory*. Cambridge University Press, 2008. ISBN 9780521852296.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 197–216, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.15. URL https://aclanthology.org/2020.emnlp-main.15/.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 132–138, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.12. URL https://aclanthology.org/2021.naacl-main.12/.

A Proofs for posterior geometry

Proof of Proposition 1. Let $m_T(Y) := \mathbb{E}[T \mid Y]$. For every $f \in L_2(P_Y)$,

$$\mathbb{E}[f(Y)T] = \mathbb{E}[f(Y)m_T(Y)] = \langle f, m_T \rangle_{L_2(P_Y)}.$$

Hence

$$\mathrm{Acc}_V(T;Y) = \sup_{\substack{f \in V \\ \|f\| \le 1}} \langle f, m_T \rangle = \sup_{\substack{f \in V \\ \|f\| \le 1}} \langle f, \Pi_V m_T \rangle = \|\Pi_V m_T\|.$$

The inequality $\mathrm{Acc}_V(T;Y) \le \mathrm{Rec}(T;Y)$ follows because orthogonal projection is contractive. Finally,

$$\mathbb{E}[(T - m_T)^2] = \mathbb{E}[T^2] - \mathbb{E}[m_T^2] = 1 - \mathrm{Rec}(T;Y)^2,$$

since T is centered with unit variance. □

Proof of Proposition 2. Let $m := \mathbb{E}[B \mid \mathscr{H}]$. For any $\mathscr{H}$-measurable $g \in \{\pm 1\}$,

$$\mathbb{P}(g = B) = \frac{1}{2}(1 + \mathbb{E}[gB]) = \frac{1}{2}(1 + \mathbb{E}[g\, m]).$$

This is maximized by $g = \mathrm{sign}(m)$, giving

$$\mathrm{BayesAcc}(B \mid \mathscr{H}) = \frac{1}{2}(1 + \mathbb{E}|m|).$$

For a coordinate subset O, $m = \mathbb{E}[B \mid Y_O]$ and $\mathrm{Acc}_O(B) = \|m\|_{L_2(P_Y)}$. Since $\mathbb{E}[m] = 0$ and $|m| \le 1$ a.s.,

$$\mathbb{E}[m^2] \le \mathbb{E}|m| \le \sqrt{\mathbb{E}[m^2]}.$$

Multiplying by $1/2$ yields the claim. □

B Posterior expansions and score sufficiency

Proof of Theorem 1. For $O \subseteq [m]$ and $z \in \{\pm 1\}^k$,

$$\mathbb{P}(Z = z \mid Y_O = y_O) \propto \pi(z) \prod_{j \in O} w(y_j \mid \chi_{a_j}(z)).$$

Using $w(y \mid u) = \bar{w}(y)(1 + u\, s(y))$,

$$\mathbb{P}(Z = z \mid Y_O = y_O) \propto \pi(z) \prod_{j \in O} \big(1 + \chi_{a_j}(z) s(y_j)\big),$$

since the common factor $\prod_{j \in O} \bar{w}(y_j)$ cancels. Thus the posterior depends on y_O only through $S_O = (s(y_j))_{j \in O}$, so $\mathbb{E}[T \mid Y_O] = \mathbb{E}[T \mid S_O]$ a.s. □

Theorem 6 (Likelihood expansion)**.** *For every $z \in \{\pm 1\}^k$ and almost every $y \in \mathcal{Y}^m$,*

$$p(y \mid z) = \bar{w}^{\otimes m}(y) \sum_{v \in \mathbb{F}_2^m} S_v(y)\, \chi_{A^\top v}(z).$$

Proof. Conditioned on $Z = z$, the outputs are independent and

$$p(y \mid z) = \prod_{j=1}^m w(y_j \mid \chi_{a_j}(z)) = \prod_{j=1}^m \bar{w}(y_j)\big(1 + \chi_{a_j}(z) s(y_j)\big).$$

Expanding the product over all subsets $v \in \mathbb{F}_2^m$ yields

$$p(y \mid z) = \bar{w}^{\otimes m}(y) \sum_{v \in \mathbb{F}_2^m} \prod_{j=1}^m s(y_j)^{v_j} \prod_{j=1}^m \chi_{a_j}(z)^{v_j}.$$

Since $\prod_{j=1}^m \chi_{a_j}(z)^{v_j} = \chi_{A^\top v}(z)$, the formula follows. □

Lemma 1 (Single-coordinate score moment)**.** *For any $u \in \{\pm 1\}$, if $Y \sim W(\cdot \mid u)$ and $S = s(Y)$, then*

$$\mathbb{E}[S] = 0 \text{ under } \bar{W}, \qquad \mathbb{E}[S \mid U = u] = \kappa u.$$

Proof. Using $w(y \mid u) = \bar{w}(y)(1 + u\, s(y))$,

$$\mathbb{E}[S \mid U = u] = \int s(y)\bar{w}(y)(1 + u\, s(y))\, d\lambda(y) = u \int s(y)^2 \bar{w}(y)\, d\lambda(y) = u\kappa.$$

Also $\int s(y)\bar{w}(y)\, d\lambda(y) = 0$. □

Proof of Theorem 2. Conditioned on Z, the outputs are independent and $U_j = \chi_{a_j}(Z)$. By Lemma 1,

$$\mathbb{E}[S_v \mid Z] = \prod_{j:v_j=1} \mathbb{E}[S_j \mid U_j] = \prod_{j:v_j=1} \kappa\, \chi_{a_j}(Z) = \kappa^{|v|}\chi_{A^\top v}(Z).$$

Hence

$$\mathbb{E}[S_v T^\circ] = \mathbb{E}[\mathbb{E}[S_v \mid Z]\, T^\circ(Z)] = \kappa^{|v|}\mathbb{E}[\chi_{A^\top v}(Z) T^\circ(Z)] = \kappa^{|v|}\widehat{T}(A^\top v).$$

For parity targets, Walsh orthogonality gives the indicator formula. □

Proposition 4 (BSC polynomial-interface equivalence)**.** *Assume the channel is* $\mathrm{BSC}(\rho)$*, $\rho < 1/2$, with reliability $\mu = 1 - 2\rho$. Then*

$$s(y) = \mu y, \qquad y \in \{\pm 1\},$$

and consequently

$$V_r^{\mathrm{int}} = \mathrm{span}\Big\{\prod_{j=1}^m Y_j^{v_j} : |v| \le r\Big\},$$

the space of degree-at-most-r multilinear polynomials in the raw observations.

Proof. For the BSC, $w(y \mid u) = (1 + \mu u y)/2$, so $\bar{w}(y) = 1/2$ and $s(y) = \mu y$. Hence

$$S_v(y) = \prod_{j=1}^m (\mu y_j)^{v_j} = \mu^{|v|} \prod_{j=1}^m y_j^{v_j}.$$

Thus score monomials and raw multilinear monomials coincide up to scaling. □

C Exact reveal laws

Lemma 2 (Functions of a coordinate subset)**.** *Assume the uniform prior. For $O \subseteq [m]$, let*

$$\mathcal{V}_O := \mathrm{span}\{a_j : j \in O\} \subseteq \mathbb{F}_2^k.$$

Then for every square-integrable target T,

$$\mathbb{E}[T \mid U_O] = \mathbb{E}[T] + \sum_{q \in \mathcal{V}_O \setminus \{0\}} \widehat{T}(q)\chi_q(Z).$$

In particular, $\mathbb{E}[T \mid U_O]$ is nonconstant iff $\mathrm{Spec}(T) \cap \mathcal{V}_O \neq \emptyset$.

Proof. Under the uniform prior, $\{\chi_q : q \in \mathbb{F}_2^k\}$ is an orthonormal basis of $L_2(Z)$. For $b \in \mathbb{F}_2^{|O|}$,

$$\prod_{j \in O} U_j^{b_j} = \chi_{A_O^\top b}(Z),$$

so the nonconstant functions of U_O are exactly the span of $\{\chi_q : q \in \mathcal{V}_O \setminus \{0\}\}$. Orthogonal projection of the Walsh expansion of T onto this span yields the formula. □

Lemma 3 (Local criterion)**.** *Assume the uniform prior. For a nonconstant target T and subset $O \subseteq [m]$,*

$$\mathrm{Acc}_O(T) > 0 \iff \mathrm{Spec}(T) \cap \mathcal{V}_O \neq \emptyset.$$

Proof. If $\mathrm{Spec}(T) \cap \mathcal{V}_O = \emptyset$, then by Lemma 2 the conditional expectation $\mathbb{E}[T \mid U_O]$ is constant. Since $T - U_O - Y_O$ is a Markov chain,

$$\mathbb{E}[T \mid Y_O] = \mathbb{E}[\mathbb{E}[T \mid U_O] \mid Y_O]$$

is also constant, hence $\mathrm{Acc}_O(T) = 0$.

Conversely, if $q \in \mathrm{Spec}(T) \cap \mathcal{V}_O$, choose $v \in \mathbb{F}_2^m$ supported inside O with $A^\top v = q$. Then S_v is measurable with respect to Y_O, and by Theorem 2

$$\mathbb{E}[S_v T^\circ] = \kappa^{|v|} \widehat{T}(q) \neq 0.$$

Thus some measurable function of Y_O correlates with T°, so $\mathrm{Acc}_O(T) > 0$. □

Proof of Theorem 3. For local interfaces,

$$\mathrm{Loc}_r(T) > 0 \iff \exists O,\ |O| \le r : \mathrm{Acc}_O(T) > 0 \overset{Lemma\ 3}{\iff} \exists O,\ |O| \le r : \mathrm{Spec}(T) \cap \mathcal{V}_O \neq \emptyset.$$

The last condition is equivalent to the existence of $q \in \mathrm{Spec}(T)$ and $v \in \mathbb{F}_2^m$ with $|v| \le r$ and $A^\top v = q$, that is, $r_T(A) \le r$.

For interaction interfaces, by Theorem 2,

$$\mathrm{Int}_r(T) > 0 \iff \exists v,\ |v| \le r : \widehat{T}(A^\top v) \neq 0 \iff \exists q \in \mathrm{Spec}(T) : r_q(A) \le r \iff r_T(A) \le r.$$

This proves the equality of reveal orders. □

Lemma 4 (Balanced binary zero-posterior criterion)**.** *Let $B \in \{\pm 1\}$ be balanced and let $\mathscr{H}$ be a sigma-field. Then*

$$\mathbb{E}[B \mid \mathscr{H}] = 0 \text{ a.s.} \iff B \perp \mathscr{H} \iff \mathrm{I}(B; \mathscr{H}) = 0.$$

Proof. Since B is balanced,

$$\mathbb{P}(B = 1 \mid \mathscr{H}) = \frac{1 + \mathbb{E}[B \mid \mathscr{H}]}{2}.$$

Thus $\mathbb{E}[B \mid \mathscr{H}] = 0$ a.s. iff $\mathbb{P}(B = 1 \mid \mathscr{H}) = 1/2$ a.s., which is equivalent to independence. Independence is equivalent to zero mutual information. □

Proof of Corollary 1. By Theorem 3,

$$\min\{r : \mathrm{Loc}_r(B) > 0\} = r_B(A).$$

If $|O| < r_B(A)$, then $\mathrm{Acc}_O(B) = 0$, that is, $\mathbb{E}[B \mid Y_O] = 0$ a.s. By Lemma 4, $B \perp Y_O$ and $\mathrm{I}(B; Y_O) = 0$. Finally, Proposition 2 gives $\mathrm{BayesAcc}(B \mid Y_O) = 1/2$. □

Proof of Corollary 2. Condition on the training data, the algorithmic randomness, and the selected subset O. The selected subset satisfies $|O| \le r < r_B(A)$. By Corollary 1, for a fresh independent test sample one has $B \perp Y_O$ and B is balanced. Therefore, for the fixed classifier g returned by the training procedure,

$$\mathbb{P}(g(Y_O) = B \mid \text{training data},\ \text{randomness}) = \tfrac{1}{2}.$$

Averaging over the training data and algorithmic randomness gives the claim. □

D Random dense thresholds and sample barriers

Proof of Theorem 4. Fix $i \in [k]$.

Lower bound. Let $0 < \alpha < \alpha_\star(R)$. For each nonzero $v \in \mathbb{F}_2^m$, the random vector $A^\top v$ is uniform on $\mathbb{F}_2^k$, so

$$\mathbb{P}(A^\top v = e_i) = 2^{-k}.$$

Hence by a union bound,

$$\mathbb{P}(r_{e_i}(A) \le \alpha m) \le \sum_{t=1}^{\lfloor \alpha m \rfloor} \binom{m}{t} 2^{-k} \le 2^{m(H_2(\alpha) - R + o(1))}.$$

Since $H_2(\alpha) < R$, this is exponentially small. Union over all i proves the lower bound.

Upper bound. Fix $\varepsilon > 0$ and choose $\beta \in (\alpha_\star(R), \min\{\alpha_\star(R) + \varepsilon, 1/2\})$. Set $t = \lceil \beta m \rceil$ and define

$$X_i := \sum_{v:|v|=t} \mathbf{1}\{A^\top v = e_i\}.$$

Then

$$\mathbb{E}[X_i] = \binom{m}{t} 2^{-k} = 2^{m(H_2(\beta)-R+o(1))} = 2^{\Omega(m)}.$$

The indicators are pairwise independent: if $v \neq v'$, then v, v' are linearly independent over $\mathbb{F}_2$, so $(A^\top v, A^\top v')$ is uniform on $\mathbb{F}_2^{2k}$. Thus $\mathrm{Var}(X_i) \leq \mathbb{E}[X_i]$, and Chebyshev gives

$$\mathbb{P}(X_i = 0) \leq \frac{1}{\mathbb{E}[X_i]} = 2^{-\Omega(m)}.$$

Hence with high probability a witness of weight at most $(\alpha_\star(R) + \varepsilon)m$ exists for e_i. Union over i completes the proof.

□

Lemma 5 (Average-to-high-probability transfer)**.** *Let $E(A) \in [0, 1]$ be any nonnegative performance metric of a random code draw A such that $\mathbb{E}[E(A)] \leq e^{-cm}$. Then for every $0 < c' < c$,*

$$\mathbb{P}\big(E(A) > e^{-c'm}\big) \leq e^{-(c-c')m}.$$

Proof. By Markov's inequality,

$$\mathbb{P}\big(E(A) > e^{-c'm}\big) \leq e^{c'm}\mathbb{E}[E(A)] \leq e^{-(c-c')m}.$$

□

Lemma 6 (Coordinate recoverability from block decoding)**.** *If a decoder $\widehat{Z}(Y) \in \{\pm 1\}^k$ satisfies*

$$\mathbb{P}(\widehat{Z}(Y) \neq Z) \leq \delta,$$

then for every coordinate i,

$$\mathrm{Rec}(Z_i; Y) \geq 1 - 2\delta.$$

Proof. Let $\widehat{T}_i(Y) := \widehat{Z}_i(Y) \in \{\pm 1\}$. Then

$$\mathbb{P}(\widehat{T}_i \neq Z_i) \leq \delta, \qquad \mathbb{E}[\widehat{T}_i Z_i] \geq 1 - 2\delta.$$

Since $\left\|\widehat{T}_i\right\|_{L_2} = 1$,

$$\mathrm{Rec}(Z_i; Y) = \sup_{\|f\| \leq 1} \mathbb{E}[f(Y)Z_i] \geq \mathbb{E}[\widehat{T}_i Z_i] \geq 1 - 2\delta.$$

□

Proof of Theorem 5. Part (i) is the classical random-linear-code achievability theorem below channel capacity under ML decoding (Gallager, 1968; Richardson & Urbanke, 2008), combined with Lemma 5 to pass from exponentially small ensemble-average block error to exponentially small block error with high probability over the random code draw. Full rank of random dense A holds with high probability, so the latent-to-code mapping is injective on the same event.

Part (ii) follows from (i) and Lemma 6.

For (iii), by Theorem 4, with high probability

$$r_{e_i}(A) \geq (\alpha_\star(R) - \varepsilon)m \qquad \forall i.$$

Fix such an A and O with $|O| < (\alpha_\star(R) - \varepsilon)m$. Then no v supported on O can satisfy $A^\top v = e_i$, so $e_i \notin \mathcal{V}_O$. Since Z_i is balanced binary and $\mathrm{Spec}(Z_i) = \{e_i\}$, Lemma 3 gives $\mathrm{Acc}_O(Z_i) = 0$. By Lemma 4, $Z_i \perp Y_O$, and by Proposition 2, $\mathrm{BayesAcc}(Z_i \mid Y_O) = 1/2$.

For (iv), the same high-probability event implies there is no v with $|v| < (\alpha_\star(R) - \varepsilon)m$ and $A^\top v = e_i$. By Theorem 2, every score monomial of degree below threshold has zero correlation with Z_i, hence $\mathrm{Int}_r(Z_i) = 0$ there. In the BSC, score monomials coincide with multilinear raw polynomials up to scaling by Proposition 4. □

Proposition 5 (BSC raw-monomial witness law)**.** *Assume the channel is* $\mathrm{BSC}(\rho)$*,* $\rho < 1/2$*, with reliability* $\mu = 1 - 2\rho$*. For* $v \in \mathbb{F}_2^m$*, define*

$$X_v := \prod_{j=1}^{m} Y_j^{v_j}.$$

Under the uniform prior, for every nonconstant target T*,*

$$\mathbb{E}[X_v\, T^\circ] = \mu^{|v|}\widehat{T}(A^\top v).$$

In particular, for a parity target T_q*,*

$$\mathbb{E}[X_v\, T_q] = \mu^{|v|}\mathbf{1}\{A^\top v = q\}.$$

Hence if $r_q(A) < \infty$*, the first nonzero polynomial witness at reveal order has correlation* $\mu^{r_q(A)}$*.*

Proof. Conditioned on Z, one has $Y_j = \eta_j U_j$ with $\mathbb{E}[\eta_j] = \mu$. Therefore

$$\mathbb{E}[X_v \mid Z] = \prod_{j:v_j=1} \mathbb{E}[Y_j \mid Z] = \prod_{j:v_j=1} \mu\, U_j = \mu^{|v|}\chi_{A^\top v}(Z).$$

Hence

$$\mathbb{E}[X_v T^\circ] = \mu^{|v|}\mathbb{E}[\chi_{A^\top v}(Z)T^\circ(Z)] = \mu^{|v|}\widehat{T}(A^\top v).$$

□

Proof of Corollary 3. By Proposition 5,

$$\mathbb{E}[X_v T_q] = \mu^{|v|}\mathbf{1}\{A^\top v = q\}.$$

If $|v| < r_q(A)$, then no such v can satisfy $A^\top v = q$, so the correlation is zero. If $|v| = r_q(A)$ and $A^\top v = q$, then the correlation is $\mu^{r_q(A)}$.

For a specified first witness, direct correlation estimation reduces to estimating a bounded mean of magnitude $\mu^{r_q(A)}$. The empirical average of n i.i.d. labeled samples has standard deviation $\Theta(n^{-1/2})$, so constant signal-to-noise ratio requires

$$\sqrt{n}\,\mu^{r_q(A)} = \Omega(1),$$

equivalently

$$n = \Omega(\mu^{-2r_q(A)}).$$

The random dense consequence follows from Theorem 4. □

E FINITE-BASIS ACCESSIBILITY FORMULA

Proposition 6 (Finite-basis accessibility formula)**.** *Let* T *be a centered unit-variance target and let* $V = \mathrm{span}\{\phi_1, \ldots, \phi_d\} \subseteq L_2(P_Y)$*, where each* $\phi_a \in L_2(P_Y)$*. Define the Gram matrix and target-correlation vector*

$$G_{ab} := \mathbb{E}[\phi_a(Y)\phi_b(Y)], \qquad b_a := \mathbb{E}[\phi_a(Y)T].$$

Then

$$\mathrm{Acc}_V(T;Y)^2 = b^\top G^\dagger b,$$

where $G^\dagger$ *is the Moore–Penrose pseudoinverse.*

Proof. Any $f \in V$ can be written as $f = \sum_{a=1}^{d} c_a\phi_a$, so

$$\mathbb{E}[fT] = c^\top b, \qquad \|f\|^2_{L_2(P_Y)} = c^\top G c.$$

If $c \in \ker(G)$ then $\sum_a c_a\phi_a = 0$ in L_2, hence $c^\top b = 0$. Therefore

$$\mathrm{Acc}_V(T;Y) = \sup_{c^\top G c \le 1} c^\top b.$$

This is the support function of the ellipsoid induced by G, whose value is $\sqrt{b^\top G^\dagger b}$. □

F GROUPED INTERFACES AND BIASED PRIORS

F.1 GROUPED INTERFACES

Let $\mathcal{G} = \{G_1, \dots, G_M\}$ be a partition of $[m]$. For $I \subseteq [M]$, write

$$G(I) := \bigcup_{t \in I} G_t, \qquad \mathrm{Loc}_{r,\mathcal{G}}(T) := \sup_{|I| \le r} \left\| \mathbb{E}[T \mid Y_{G(I)}] \right\|_{L_2}.$$

Define

$$r_q^{\mathcal{G}}(A) := \min\Big\{ |I| : \exists v \in \mathbb{F}_2^m,\ \mathrm{supp}(v) \subseteq G(I),\ A^\top v = q \Big\}, \qquad r_T^{\mathcal{G}}(A) := \min_{q \in \mathrm{Spec}(T)} r_q^{\mathcal{G}}(A).$$

Proposition 7 (Grouped reveal law)**.** *Under the uniform prior:*

(i) for every nonconstant target T,

$$\min\{r : \mathrm{Loc}_{r,\mathcal{G}}(T) > 0\} = r_T^{\mathcal{G}}(A);$$

(ii) grouped accessibility is invariant under arbitrary measurable bijections with measurable inverse applied separately within each visible group;

(iii) all coordinate factors are individually s-group-accessible iff there exists a left inverse $L \in \mathbb{F}_2^{k \times m}$ with $LA = I_k$ such that each row of L intersects at most s groups.

Proof. For (i), replace O by $G(I)$ in Lemma 3. This yields

$$\mathrm{Loc}_{r,\mathcal{G}}(T) > 0 \iff \exists I,\ |I| \le r,\ \exists q \in \mathrm{Spec}(T),\ \exists v : \mathrm{supp}(v) \subseteq G(I),\ A^\top v = q,$$

which is exactly $r \ge r_T^{\mathcal{G}}(A)$.

For (ii), measurable bijections with measurable inverse preserve the sigma-field generated by the coordinates in each group and hence preserve every grouped interface sigma-field.

For (iii), apply (i) to the coordinate targets $Z_i = \chi_{e_i}$. The existence of an s-group witness for every e_i is equivalent to the existence of a left inverse whose ith row is such a witness. □

F.2 BIASED-PRIOR ROBUSTNESS

Under a biased product prior with

$$\mathbb{E} Z_\ell = \nu_\ell, \qquad |\nu_\ell| \le \bar{\nu} < 1,$$

exact invisibility softens into quantitative indistinguishability on interfaces whose visible row span has no low-weight witness for the target coordinate.

For a coordinate i, let

$$T_i := \frac{Z_i - \nu_i}{\sqrt{1 - \nu_i^2}}, \qquad \delta_i(O) := \min\{|t| : t \in \mathrm{span}\{a_j : j \in O\},\ t_i = 1\},$$

with the convention $\delta_i(O) = +\infty$ if the set is empty. Let $P_{Y_O}^{(i,\pm)}$ denote the law of Y_O conditional on $Z_i = \pm 1$, and define bounded-probe leakage

$$\mathrm{Leak}_O^{\infty}(T_i) := \sup_{f : \mathcal{Y}^O \to [-1,1]} \left| \mathbb{E}[f(Y_O)\, T_i] \right|.$$

Theorem 7 (Biased-prior local indistinguishability)**.** *For every i and $O \subseteq [m]$,*

$$\mathrm{TV}\Big(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}\Big) \le \left(\sum_{\substack{b \in \mathbb{F}_2^{|O|} \\ (A_O^\top b)_i = 1}} \prod_{\ell \ne i : (A_O^\top b)_\ell = 1} \nu_\ell^2 \right)^{1/2} \le 2^{|O|/2} \bar{\nu}^{\,\delta_i(O) - 1},$$

and consequently

$$\mathrm{Leak}_O^{\infty}(T_i) \le \sqrt{1 - \nu_i^2}\, \mathrm{TV}\Big(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}\Big).$$

Proof. Let $n = |O|$, let $P_{U_O}^{(i,\pm)}$ be the law of U_O conditional on $Z_i = \pm 1$, and let

$$\Delta := P_{U_O}^{(i,+)} - P_{U_O}^{(i,-)}.$$

Since Y_O is produced from U_O by an independent channel,

$$\mathrm{TV}(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}) \le \mathrm{TV}(P_{U_O}^{(i,+)}, P_{U_O}^{(i,-)}).$$

For $b \in \mathbb{F}_2^n$, write $\chi_b(u) = \prod_{j=1}^n u_j^{b_j}$ and $t = A_O^\top b$. Then $\chi_b(U_O) = \chi_t(Z)$, so under the product prior,

$$\mathbb{E}[\chi_b(U_O) \mid Z_i = \xi] = \xi^{t_i} \prod_{\ell \neq i : t_\ell = 1} \nu_\ell.$$

Therefore the Fourier coefficient of Δ at b is

$$\widehat{\Delta}(b) = \begin{cases} 2\prod_{\ell \neq i : t_\ell = 1} \nu_\ell, & t_i = 1, \\ 0, & t_i = 0. \end{cases}$$

Using Cauchy–Schwarz and Parseval,

$$\mathrm{TV}(P_{U_O}^{(i,+)}, P_{U_O}^{(i,-)}) = \frac{1}{2} \sum_u |\Delta(u)| \le \frac{1}{2} \sqrt{\sum_b \widehat{\Delta}(b)^2},$$

hence

$$\mathrm{TV}(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}) \le \left(\sum_{\substack{b \in \mathbb{F}_2^n \\ (A_O^\top b)_i = 1}} \prod_{\ell \neq i : (A_O^\top b)_\ell = 1} \nu_\ell^2 \right)^{1/2}.$$

If $(A_O^\top b)_i = 1$, then $t = A_O^\top b$ has weight at least $\delta_i(O)$ and therefore contributes at most $\bar{\nu}^{2(\delta_i(O)-1)}$. Since there are at most 2^n terms,

$$\mathrm{TV}(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}) \le 2^{n/2} \bar{\nu}^{\,\delta_i(O)-1}.$$

Finally, for any $f : \mathcal{Y}^O \to [-1, 1]$,

$$\mathbb{E}[f(Y_O) T_i] = \frac{\sqrt{1-\nu_i^2}}{2} \left(\mathbb{E}[f(Y_O) \mid Z_i = +1] - \mathbb{E}[f(Y_O) \mid Z_i = -1] \right),$$

so

$$|\mathbb{E}[f(Y_O) T_i]| \le \sqrt{1-\nu_i^2}\, \mathrm{TV}(P_{Y_O}^{(i,+)}, P_{Y_O}^{(i,-)}).$$

Taking the supremum over $\|f\|_\infty \le 1$ proves the claim. □

G Extended experimental details

G.1 Refinement objectives, restarts, and representative selection

All parity design experiments use $k = 10$, $m = 16$, and $\mathrm{BSC}(0.05)$. Order-≤ 2 readability summaries are computed by exhaustive enumeration over all latent states and all subsets or monomials of order at most 2; reveal-order checks use complete witness search. Experiments used one NVIDIA V100 32GB GPU, Intel Xeon Gold 6248 CPU, and 256GB RAM.

The *global-only* refinement objective optimizes global code quality only: minimum induced distance, average induced distance, row diversity, and full rank. The *accessibility-aware* refinement objective augments the same score with explicit penalties on $N_1(A)$ and $N_{\le 2}(A)$. We sweep the weak initialization-bias scalar $\eta \in \{0, 0.02, 0.05\}$ and perform 20 independent binary-refinement restarts per setting.

Representative models are selected using the following fixed rules:

Table 4: Representative learned-bridge models. A task-only learned encoder can discover a distributed but still leaky code. Witness-aware learning can push the encoder into the unreadable-up-to-order-2 regime.

Model	Block acc.	Mean Rec	$\mathrm{LB}_{\le 2}$	$\mathrm{DC}_{\le 2}$	$\bar{r}$	$r_{\min}$	W1 rows	Dup. rows
Dense baseline	0.793	0.839	0.586	0.171	3.4	1	1	0
No witness control	0.842	0.841	0.586	0.171	3.6	1	1	0
Witness-aware	**0.862**	**0.897**	**0.500**	**0.000**	3.6	**3**	**0**	0

Table 5: Illustrative learned runs from the 20-seed sweeps, with two seeds shown per setting. Witness-aware training can produce hardened encoders with exact order-2 unreadability.

λ_{wit}	Seed	block acc.	Mean Rec	$\mathrm{LB}_{\le 2}$	$\mathrm{DC}_{\le 2}$	$\bar{r}$	W1 rows	Dup. rows
0.0	0	0.842	0.841	0.586	0.171	3.6	1	0
0.0	1	0.818	0.886	0.626	0.252	3.1	1	0
0.3	0	0.835	0.884	0.590	0.180	3.2	2	1
0.3	1	0.864	0.920	0.676	0.351	2.6	3	0
1.0	0	**0.862**	0.897	**0.500**	**0.000**	**3.6**	**0**	0
1.0	1	0.866	0.919	0.712	0.423	2.6	2	0

- **dense reference:** a high-utility dense full-rank draw satisfying $N_{\le 2}(A) = 0$;
- **global-only representative:** a high-utility run from the global-only sweep;
- **access-aware representative:** a high-utility run from the access-aware sweep subject to $N_{\le 2}(A) = 0$.

Aggregate statistics over all runs are reported in the main text.

G.2 Additional theorem-facing exact checks

For each representative model in the main text, we scan all coordinate targets and compute:

- the exact witness threshold $r_i(A)$;
- the first subset size at which local Bayes accuracy exceeds $1/2$;
- the first degree at which a nonzero raw monomial correlation appears.

These are the quantities summarized in Table 3. Full reveal curves can be generated from the same exhaustive enumeration; the selected models exhibit the predicted jump at the witness-predicted order.

G.3 Additional learned bridge

To connect the exact benchmark to learned representations while staying inside the parity family, we train the parity encoder end-to-end for global latent reconstruction and compare a *no-witness-control* objective against a *witness-aware* objective.

Setup. We fix $k = 10$, $m = 16$, and $\mathrm{BSC}(0.05)$, so $\mu = 0.9$. The encoder is a binary parity matrix $A \in \mathbb{F}_2^{16\times 10}$ learned through a continuous relaxation with temperature annealing followed by a straight-through hardening phase. The decoder is a two-hidden-layer MLP with width 128 and output dimension 10, trained to reconstruct the full latent vector Z from the noisy observation Y. Training uses Adam with learning rate 2×10^{-3}, batch size 512, 2600 total steps, a soft phase of 1900 steps, and temperature annealing from 2.0 to 0.5. For each $\lambda_{\mathrm{wit}} \in \{0, 0.3, 1.0\}$ we train 20 independent seeds.

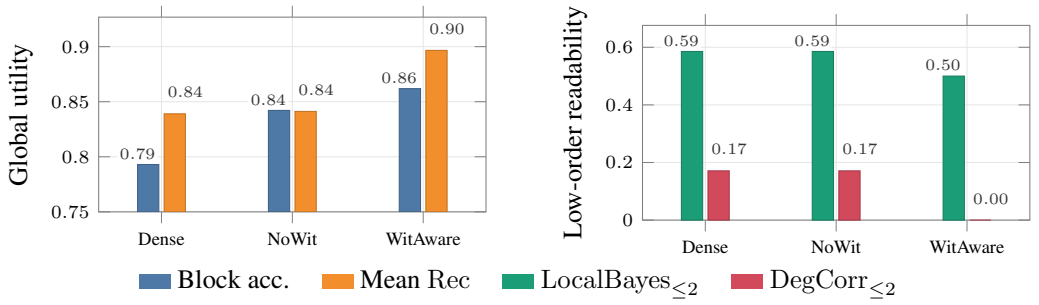

Figure 3: Learned bridge inside the exact parity family. Witness-aware learning preserves strong global performance while removing all size-2 local and degree-2 readout signal.

For encoder probabilities $P_{j\ell} \in (0,1)$, the witness regularizer is a soft penalty against coordinate witnesses of weights 1 and 2. Writing

$$s_{j,i}^{(1)} = P_{j,i} \prod_{t \neq i} (1 - P_{j,t})$$

for the soft event that row j equals e_i, and

$$q_t^{(j,\ell)} = P_{j,t}(1 - P_{\ell,t}) + (1 - P_{j,t})P_{\ell,t}, \qquad s_{j\ell,i}^{(2)} = q_i^{(j,\ell)} \prod_{t \neq i} (1 - q_t^{(j,\ell)})$$

for the soft event that row j xor row ℓ equals e_i, we define

$$\mathcal{R}_{\text{wit}}^{(\leq 2)} = \frac{1}{k} \sum_{i=1}^{k} \left[\max_j s_{j,i}^{(1)} + \max_{j<\ell} s_{j\ell,i}^{(2)} \right].$$

We also use

$$\mathcal{R}_{\text{bin}} = \frac{1}{mk} \sum_{j,\ell} P_{j\ell}(1 - P_{j\ell}), \quad \mathcal{R}_{\text{bal}} = \frac{1}{m} \sum_{j} \left(\frac{1}{k} \sum_{\ell} P_{j\ell} - \frac{1}{2} \right)^2,$$

$$\mathcal{R}_{\text{zero}} = \frac{1}{m} \sum_{j} \prod_{\ell} (1 - P_{j\ell}), \quad \mathcal{R}_{\text{dup}} = \frac{2}{m(m-1)} \sum_{j<\ell} \prod_{t} [P_{jt}P_{\ell t} + (1 - P_{jt})(1 - P_{\ell t})].$$

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{wit}} \mathcal{R}_{\text{wit}}^{(\leq 2)} + \lambda_{\text{bin}} \mathcal{R}_{\text{bin}} + \lambda_{\text{bal}} \mathcal{R}_{\text{bal}} + \lambda_{\text{zero}} \mathcal{R}_{\text{zero}} + \lambda_{\text{dup}} \mathcal{R}_{\text{dup}},$$

with

$$\lambda_{\text{bin}} = 0.02, \qquad \lambda_{\text{bal}} = 0.05, \qquad \lambda_{\text{zero}} = 0.15, \qquad \lambda_{\text{dup}} = 0.08.$$

Here "W1 rows" counts hardened rows of Hamming weight one, and "Dup. rows" counts duplicate hardened rows. Representative learned-bridge results are summarized in Table 4. To illustrate variability across seeds, Table 5 reports example runs from each sweep. A compact visual summary is given in Figure 3.

H CONTINUOUS LEARNED BRIDGE BEYOND THE PARITY FAMILY

We report the continuous representation experiment used in the main text to test whether the recoverability–accessibility distinction remains visible outside the exact parity family. The experiment provides a controlled neural setting in which global Bayes readability, local Bayes readability, and finite-sample probe behavior can be compared directly.

H.1 SETUP

We construct a procedural grayscale image generator with five latent factors,

$$(\text{shape}, \text{scale}, \text{orient}, x, y) \in [2] \times [3] \times [8] \times [6] \times [6],$$

yielding

$$2 \times 3 \times 8 \times 6 \times 6 = 1728$$

latent states. Each state is rendered deterministically as a 48×48 image containing either an arrow or a scalene triangle, with discrete scale, orientation, and position.

From these states we define three balanced binary protected targets:

$$\text{is_triangle}, \qquad \text{left_half}, \qquad \text{triangle_xor_left} = \text{is_triangle} \oplus \text{left_half},$$

where left_half indicates whether the horizontal position lies in the left half of the six-point x grid. All three targets are exactly balanced.

The encoder is a convolutional network

$$f_\theta : \mathbb{R}^{1\times 48\times 48} \to \mathbb{R}^{12},$$

followed by a trainable orthogonal mixer

$$M_\phi \in \mathbb{R}^{12\times 12}.$$

For an image x_s corresponding to latent state s, the clean representation mean is

$$z_s = \text{Norm}\big(M_\phi f_\theta(x_s)\big) \in \mathbb{R}^{12},$$

where every representation is rescaled to fixed radius

$$\|z_s\|_2 = r, \qquad r = 2.0.$$

The observed representation is then

$$Y = z_s + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_{12}), \qquad \sigma = 0.25.$$

The fixed-power constraint prevents trivial improvements in readability by simply inflating representation magnitude relative to the observation noise.

The local interfaces in this experiment are coordinate-local by design: one-coordinate interfaces read a single representation dimension, and two-coordinate interfaces read a pair of dimensions. Since accessibility in the main paper is interface-relative, the learned coordinate system is itself part of the exposure geometry. The orthogonal mixer therefore provides a direct mechanism for changing local exposure without necessarily destroying global information.

A shared MLP decoder predicts all five latent factors from the noisy representation:

$$(\widehat{\text{shape}}, \widehat{\text{scale}}, \widehat{\text{orient}}, \widehat{x}, \widehat{y}).$$

The utility loss is the sum of five cross-entropies. We first train a base encoder–decoder pair on noisy latent reconstruction using a $90/10$ train/validation split of the 1728 states:

$$1555 \text{ train states}, \qquad 173 \text{ validation states}.$$

The base model is trained for 70 epochs with Adam and learning rate 10^{-3}. At the best base epoch, noisy utility under the identity mixer is

$$\text{shape} = 0.9768, \quad \text{scale} = 0.9814, \quad \text{orient} = 0.1639, \quad x = 0.7396, \quad y = 0.8091, \quad \text{joint} = 0.0878.$$

The final bridge uses four variants:

(i) **raw_identity:** pretrained encoder with identity mixer;

(ii) **utility_only_mixer:** freeze the encoder and train the orthogonal mixer plus decoder for utility only;

(iii) **size1_aware_mixer:** utility plus a one-coordinate local exposure penalty;

(iv) **access_aware_mixer:** utility plus both one-coordinate and two-coordinate local exposure penalties.

In all improved bridge variants, the encoder is frozen. The variant stage uses 60 epochs, Adam, and learning rate 8×10^{-4}.

The protected family is

$$\mathcal{P} = \{\text{is_triangle},\ \text{left_half},\ \text{triangle_xor_left}\}.$$

For a protected target $B_t \in \{0, 1\}$ and coordinate j, define

$$\text{Score}_{t,j}^{(1)} = |\mathbb{E}[Y_j \mid B_t = 1] - \mathbb{E}[Y_j \mid B_t = 0]| + 0.25\left|\mathbb{E}[Y_j^2 \mid B_t = 1] - \mathbb{E}[Y_j^2 \mid B_t = 0]\right|.$$

For a coordinate pair (i, j), let

$$\psi_{ij}(Y) = (Y_i,\ Y_j,\ Y_iY_j,\ Y_i^2,\ Y_j^2),$$

and define

$$\text{Score}_{t,ij}^{(2)} = \|\mathbb{E}[\psi_{ij}(Y) \mid B_t = 1] - \mathbb{E}[\psi_{ij}(Y) \mid B_t = 0]\|_2.$$

These are moment-based surrogates for local distinguishability. Let $\text{TopKMean}(\cdot, k)$ be the mean of the largest k entries. To concentrate optimization pressure on the strongest leaks, we use top-k aggregation:

$$\text{Hard}_1 = \frac{1}{|\mathcal{P}|}\sum_{t\in\mathcal{P}} \text{TopKMean}_j\big(\text{Score}_{t,j}^{(1)},\ k_1\big), \qquad k_1 = 4,$$

$$\text{Hard}_2 = \frac{1}{|\mathcal{P}|}\sum_{t\in\mathcal{P}} \text{TopKMean}_{i<j}\big(\text{Score}_{t,ij}^{(2)},\ k_2\big), \qquad k_2 = 8.$$

The final losses are

$$\mathcal{L}_{\text{utility}}, \qquad \mathcal{L}_{\text{utility}} + \lambda_1\text{Hard}_1, \qquad \mathcal{L}_{\text{utility}} + \lambda_1\text{Hard}_1 + \lambda_2\text{Hard}_2,$$

with

$$\lambda_1 = 3.0, \qquad \lambda_2 = 5.0.$$

Because the latent state space is finite and the representation noise is Gaussian, Bayes readability remains numerically tractable under the known generative model. For each latent state s, we compute the representation mean $z_s \in \mathbb{R}^{12}$. For a coordinate subset $O \subseteq [12]$ and target $B \in \{0, 1\}$, the local class-conditional law is

$$p(y_O \mid B = b) = \frac{1}{|S_b|}\sum_{s:B(s)=b} \mathcal{N}\big(y_O; z_{s,O}, \sigma^2 I\big),$$

where

$$S_b = \{s : B(s) = b\}.$$

We report

- **full Bayes accuracy:** $\text{BayesAcc}(B \mid Y_{[12]})$;
- **best1 Bayes accuracy:** $\max_{|O|=1} \text{BayesAcc}(B \mid Y_O)$;
- **best2 Bayes accuracy:** $\max_{|O|=2} \text{BayesAcc}(B \mid Y_O)$.

We also report the posterior-signal proxy

$$\text{RecProxy}(B; Y_O) = \sqrt{\mathbb{E}[m(Y_O)^2]}, \qquad m(Y_O) = 2\mathbb{P}(B = 1 \mid Y_O) - 1.$$

All Bayes quantities are estimated by Monte Carlo using 8192 samples per target and variant.

The most informative result is the protected compositional target triangle_xor_left; see Tables 7 and 8. The results support three observations.

First, *utility-only shaping does not reduce accessibility for free*. Relative to the identity reference, utility-only mixing leaves local Bayes readability almost unchanged:

$$0.6963 \to 0.7036 \text{ for best1}, \qquad 0.7417 \to 0.7405 \text{ for best2}.$$

Table 6: Continuous bridge: mean Bayes readability and utility over the three protected targets. All values are averages over 20 variant-training seeds.

Variant	Full acc.	Best1 acc.	Best2 acc.	Utility joint	Quad probe@8192	MLP probe@8192
raw_identity	0.9593	0.7334	0.8059	0.0859	0.9299	0.9542
utility_only_mixer	0.9597	0.7380	0.8089	0.1254	0.9299	0.9536
size1_aware_mixer	0.9555	0.7227	0.7953	0.1170	0.9248	0.9507
access_aware_mixer	0.9269	0.6267	0.7277	0.1019	0.8390	0.9204

Table 7: Continuous bridge: Bayes readability of the protected compositional target $\mathrm{triangle_xor_left}$. The "proxy" columns report $\mathrm{RecProxy}(B; Y_O)$. All values are averages over 20 variant-training seeds.

Variant	Full acc.	Best1 acc.	Best2 acc.	Full proxy	Best1 proxy	Best2 proxy
raw_identity	0.9390	0.6963	0.7417	0.9071	0.4568	0.5453
utility_only_mixer	0.9401	0.7036	0.7405	0.9074	0.4690	0.5451
size1_aware_mixer	0.9336	0.6812	0.7360	0.9024	0.4134	0.5537
access_aware_mixer	0.8944	0.5814	0.6367	0.8321	0.1942	0.3265

Second, *one-coordinate suppression alone is not enough*. The size-1-aware variant only mildly changes local readability:

$$\text{best1} : 0.6963 \to 0.6812, \qquad \text{best2} : 0.7417 \to 0.7360.$$

Third, *access-aware shaping substantially suppresses low-order local readout*. Relative to the identity reference, the protected compositional target changes as follows:

$$\text{full acc.} : 0.9390 \to 0.8944,$$

$$\text{best1 acc.} : 0.6963 \to 0.5814, \qquad \text{best2 acc.} : 0.7417 \to 0.6367,$$

while the posterior-signal proxy drops more sharply on local interfaces:

$$\text{full proxy} : 0.9071 \to 0.8321,$$

$$\text{best1 proxy} : 0.4568 \to 0.1942, \qquad \text{best2 proxy} : 0.5453 \to 0.3265.$$

Thus local posterior signal falls much faster than full posterior signal.

To separate accessibility from learnability, we also train three probes on noisy representation samples:

- **linear probe:** logistic regression on the raw 12-dimensional representation;
- **quadratic probe:** logistic regression on all linear terms and all pairwise monomials;
- **MLP probe:** a two-hidden-layer probe.

Training sizes are

$$n \in \{128, 512, 2048, 8192\},$$

with a test set of 4096 samples.

H.2 Main results

Table 6 reports the mean Bayes-readability summary over

$$\{\mathrm{is_triangle}, \mathrm{left_half}, \mathrm{triangle_xor_left}\}.$$

Utility-only reshaping leaves local readability nearly unchanged. By contrast, access-aware shaping reduces both one-coordinate and two-coordinate readability, while retaining strong global signal and nontrivial latent-factor utility.

The primitive protected targets show the same pattern with different strengths. For $\mathrm{is_triangle}$, access-aware shaping strongly lowers both best1 and best2 Bayes readability:

$$\text{full} = 0.9857 \to 0.9373, \qquad \text{best1} = 0.7228 \to 0.5790, \qquad \text{best2} = 0.8174 \to 0.7025.$$

Table 8: Continuous bridge: probe accuracy on $\mathrm{triangle_xor_left}$ at training size $n = 8192$. Values are averaged over 20 representation seeds and 3 probe seeds per representation seed.

Variant	Linear	Quadratic	MLP
raw_identity	0.5307	0.8708	0.9310
utility_only_mixer	0.5272	0.8714	0.9277
size1_aware_mixer	0.5290	0.8617	0.9244
access_aware_mixer	0.5308	0.7032	0.8821

For $\mathrm{left_half}$, the effect is weaker:

$$\text{full} = 0.9531 \to 0.9490, \qquad \text{best1} = 0.7811 \to 0.7197, \qquad \text{best2} = 0.8586 \to 0.8440.$$

This likely reflects the fact that $\mathrm{left_half}$ is tightly tied to the utility-relevant latent factor x, so reducing its local exposure is structurally harder without paying more utility cost.

The continuous bridge supports three limited but useful conclusions:

(i) the recoverability–accessibility gap is not purely parity-specific;

(ii) accessibility-aware objectives can reshape one-coordinate and two-coordinate local exposure outside the exact parity family;

(iii) low-order probe learnability tracks these exposure changes, while stronger probes still recover more of the globally available signal.

These results are empirical, but they support the use of accessibility-aware objectives as a design tool beyond the exact parity setting.