DECISION-SUFFICIENT PREDICTIVE REPRESENTATIONS FOR SYMMETRIC FIXED-CARDINALITY LOSSES

Anonymous authors
Paper under double-blind review

ABSTRACT

What information about uncertainty must a predictor expose to support optimal decisions? We answer this question for symmetric fixed-cardinality overlap losses, where decisions and outcomes are fixed-size subsets and the loss depends only on their overlap. Using the Johnson decomposition, we derive coefficients directly from the scalar loss profile to identify the active interaction layers, without constructing the full loss matrix. Projection onto these layers preserves every expected-loss comparison. Conversely, every nonzero direction in the active subspace yields two feasible outcome distributions with no common optimal decision. These results establish the minimum universally sufficient affine-linear interface dimension. Existing convex-calibration bounds give the same exact dimension, attained by a squared embedding. For $K = n/2$, complement symmetry eliminates positive even-order layers for a family including odd-m majority thresholds, explaining non-prefix requirements. A layer-wise regret bound separates omitted information, estimation error, and decoding error. Synthetic experiments and frozen-predictor ablations on a balanced, candidate-conditioned Wikipedia task show how loss-specific information yields decision gains, with estimation error and decision margins shaping its value. Comparisons of direct active-target training with full prediction followed by projection distinguish the choice of information from the choice of estimator.

1 INTRODUCTION

Predictive decision systems must determine what information about uncertainty to pass from prediction to optimization. While decision-focused learning aligns prediction with downstream objectives or differentiates through optimization (Donti et al., 2017; Elmachtoub & Grigas, 2022; Wilder et al., 2019; Pogancic et al., 2020; Berthet et al., 2020), the choice of predictive interface determines which distinctions the optimizer can use. Predicting the full outcome distribution can be costly, whereas item-wise marginals may omit dependencies that change the optimal decision. We study this choice for symmetric fixed-cardinality overlap losses, $L(S, e) = \ell(|S \cap e|)$, where a decision selects K items and the outcome contains exactly m items. This formulation captures fixed-cardinality instances of budgeted recommendation and multi-label prediction (Schultheis et al., 2024). It also captures complete-evidence retrieval with a fixed number of supporting passages, where success requires recovering the entire evidence set (Yang et al., 2018; Xiong et al., 2020).

Figure 1 illustrates why information requirements depend on the loss. Exactly two of four sites fail, and two sites can be protected. The figure compares two possible distributions of the failure pair, denoted P and Q, with identical item-wise failure probabilities. Under additive reward, which awards one unit for each protected failure, selecting AD maximizes expected reward under both distributions. Under all-or-nothing reward, which is earned only when both failures are protected, the optimal choices are AB or CD under P, but AC or BD under Q. No single protection decision is optimal under both distributions. A decision rule based only on the shared marginals therefore cannot be optimal under both distributions, even with exact optimization. In contrast, a pairwise interface retains the probabilities of joint failures, allowing the optimizer to distinguish P from Q and select an optimal protection pair under each distribution. Thus, item-wise information suffices for additive reward, whereas all-or-nothing reward requires information about joint failures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

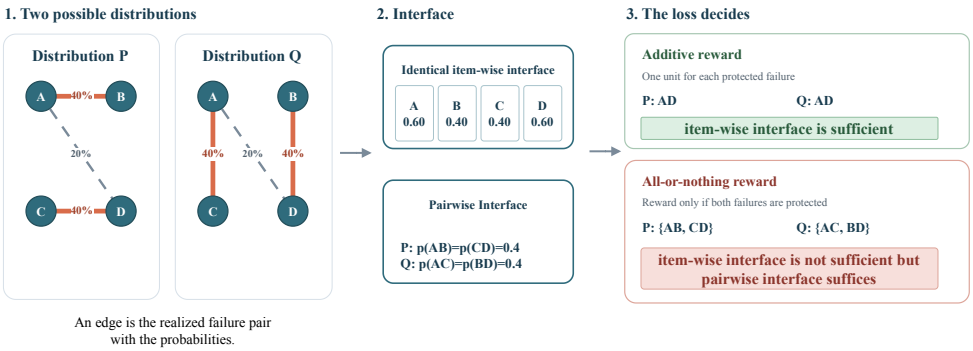


Figure 1: Four-site motivation. The two outcome distributions have identical item-wise marginals. Additive reward selects AD under both distributions; under all-or-nothing reward, no single protection decision is optimal under both. Pairwise failure probabilities distinguish the distributions and suffice for optimal all-or-nothing decisions.

This example raises a general question: for a given overlap loss, which components of the outcome distribution must a predictive interface retain to support optimal decisions? Classical comparisons of statistical information and recent work on task-relevant representations provide general frameworks for studying this question (Blackwell, 1953; Le, 1964; van Rooyen & Williamson, 2014; Sevetlidis, 2026; Bennouna et al., 2026; Ye et al., 2026). Loss embeddings and Johnson spectral theory offer complementary tools for constructing and analyzing such representations (Ramaswamy et al., 2013; Ciliberto et al., 2016; 2020; Delsarte, 1973; Filmus, 2014). For symmetric fixed-cardinality overlap losses, we seek an explicit characterization directly from the scalar loss profile: which components suffice, why omitting any required direction can prevent optimal decisions, and what minimum dimension an affine-linear interface must have.

We characterize this information through active Johnson layers, orthogonal components of increasing interaction order. Our coefficients compute the active subspace directly from the scalar loss profile. Projection preserves all expected-loss comparisons, and every nonzero active direction separates feasible distributions with no common optimal decision. The resulting minimum dimension is therefore both sufficient and necessary. Complement symmetry explains why required orders can skip intermediate layers. A layer-wise error bound then connects this exact characterization to learning: it separates the cost of omitted information from errors in estimating and decoding the retained components. Our experiments examine these quantities alongside sample size and decision margins.

Contributions.

- **Explicit loss-specific active support.** We derive coefficients that identify the active Johnson layers directly from the overlap loss, without constructing the full loss matrix. Projection preserves every expected-loss comparison; complement symmetry yields a family with inactive even orders.
- **Direction-wise necessity and minimum interface dimension.** We establish the decision cost of failing to distinguish each active direction through an exact average-regret identity. This yields the minimum dimension of a universally sufficient affine-linear interface, equal to the active-subspace dimension D_L.
- **Exact calibration dimension and a matching surrogate.** Applying existing convex-calibration bounds to the active-space characterization gives the exact dimension D_L. A standard squared embedding attains this dimension, with an explicit bound transferring surrogate excess risk to decision regret.
- **Decision value and finite-sample learning.** A spectral error bound separates omission, estimation, and decoding. Synthetic and candidate-conditioned Wiki experiments relate higher-order decision gains to sample size, estimation error, and decision margins, and compare direct training with full prediction followed by projection.

2 Related Work

Decision-sufficient representations identify information needed for optimization. Blackwell comparison, Le Cam deficiency, and Bayes sufficiency provide general task-preservation principles (Blackwell, 1953; Le, 1964; van Rooyen & Williamson, 2014; Sevetlidis, 2026). For linear programs, Bennouna et al. (2026) characterize sufficient cost measurements geometrically and construct minimum or least-cost datasets; Ye et al. (2026) develop pointwise sufficiency, hardness results, and learning guarantees. For a finite loss, expected decision costs are Lp, and cost geometry characterizes which measurements preserve their optimizer. We exploit overlap symmetry to compute their decision-relevant dependence on p directly from $\ell(q)$: the interaction orders and dimension follow without enumerating the loss matrix (Table 1).

Table 1: Comparison with closely related frameworks. Matrix-free means avoiding explicit construction of a finite loss matrix: ✓ supplied, − not supplied, N/A not applicable to LP cost queries. CCdim denotes convex calibration dimension.

Work	Representation	Matrix-free	Characterization	Dimension
Bennouna et al. (2026)	LP cost geometry	N/A	Geometric sufficiency	Relevant subspace
Ye et al. (2026)	LP cost queries	N/A	Pointwise sufficiency	Learning bounds in $d^\star$
Ramaswamy series (2013/16)	Finite loss matrix	−	Calibration bounds	Rank/ CCdim bounds
Ciliberto et al. (2016/20)	Implicit embedding	✓	Sufficient construction	Embedding-dependent
This work	Closed form from $\ell(q)$	✓	Direction-wise necessity	Exact affine minimum; exact CCdim

Calibration and structured prediction provide the learning foundations. Ramaswamy et al. (2013) construct calibrated surrogates from low-rank loss factorizations, while Ramaswamy & Agarwal (2016) give general convex calibration dimension bounds. Implicit loss embeddings support consistent structured prediction without enumerating finite loss matrices (Ciliberto et al., 2016; 2020); elicitation and embedding theory further clarify prediction dimension (Frongillo & Kash, 2015; 2021; Finocchiaro et al., 2024). For our loss family, active support yields an exact minimum affine-interface dimension and matching convex calibration dimension.

Johnson theory supplies classical slice decompositions and inclusion ranks (Delsarte, 1973; Gottlieb, 1966; Wilson, 1990; Brouwer & Haemers, 2011; Filmus, 2014). We identify active layers from $\ell(q)$ and construct a necessity witness for every direction in their span. Related F1 and Jaccard analyses study variable-cardinality power sets (Zhang et al., 2020; Zhang, 2026a;b). Our characterization complements decision-focused learning and predict-then-optimize methods (Donti et al., 2017; Elmachtoub & Grigas, 2022; El Balghiti et al., 2019; Wilder et al., 2019; Pogancic et al., 2020; Berthet et al., 2020) by identifying the information passed from prediction to optimization.

3 Loss-Specific Predictive Interfaces

Let $[n] = \{1, \dots, n\}$, with outcome space $\Omega_m = \binom{[n]}{m}$ and decision space $\Omega_K = \binom{[n]}{K}$, where $1 \le m, K < n$. We study losses $L(S, e) = \ell(|S \cap e|)$. At a fixed input context, let $p \in \Delta_{\Omega_m}$ be the conditional outcome distribution, where Δ_{Ω_m} is the probability simplex. For the loss vector $L_S = (L(S, e))_{e \in \Omega_m}$, define expected loss, optimal decisions, and decision regret by

$$\begin{aligned} R_L(S; p) = \langle L_S, p \rangle, \qquad \mathrm{Bayes}_L(p) = \arg\min_{S \in \Omega_K} R_L(S; p), \\ \mathrm{Reg}_L(S; p) = R_L(S; p) - \min_{T \in \Omega_K} R_L(T; p). \end{aligned} \tag{1}$$

Definition 1 (Predictive interface and universal Bayes sufficiency)**.** *A predictive interface is a map $R : \Delta_{\Omega_m} \to \mathcal{Z}$, where $\mathcal{Z}$ is the interface output space. It is universally Bayes-sufficient for L if there exists a single decoder $d : \mathcal{Z} \to \Omega_K$ such that $d(R(p)) \in \mathrm{Bayes}_L(p)$ for every $p \in \Delta_{\Omega_m}$.*

The decoder observes only $R(p)$. Universal sufficiency is a population-level property of the information available to this decoder.[1]

Optimal decisions depend only on expected-loss differences. Define

$$W_L = \operatorname{span}\{L_S - L_T : S, T \in \Omega_K\}, \qquad R_W(p) = P_{W_L} p. \tag{2}$$

Here P_{W_L} is orthogonal projection. Since $L_S - L_T \in W_L$, $R_L(S;p) - R_L(T;p) = \langle L_S - L_T, p\rangle = \langle L_S - L_T, P_{W_L} p\rangle$. Thus R_W preserves all comparisons, including ties, and is universally Bayes-sufficient. Both W_L and optimal decisions are unchanged by $L'(S,e) = aL(S,e) + b(e)$ with $a > 0$.

3.1 Computing the active interface

To compute W_L from the scalar profile, use utility $u = -\ell$ and let $A_u[S,e] = u(|S \cap e|)$, with expected-utility vector $g_p = A_u p$. Define $\mathrm{Bayes}_u(p) = \arg\max_S g_p(S)$ and $\mathrm{Reg}_u(S;p) = \max_T g_p(T) - g_p(S)$; these equal $\mathrm{Bayes}_L(p)$ and $\mathrm{Reg}_L(S;p)$, respectively.

Classical Johnson theory decomposes functions on the fixed-cardinality space Ω_k as $\mathbb{R}^{\Omega_k} = \bigoplus_{j=0}^{\min(k,n-k)} V_j^{(k)}$, with $\dim V_j^{(k)} = \binom{n}{j} - \binom{n}{j-1}$. We use the convention $\binom{a}{b} = 0$ unless a, b are integers with $0 \le b \le a$. The layer $V_0^{(k)}$ consists of constant functions; $V_1^{(k)}$ captures item-wise information, and higher layers capture interaction components not represented at lower orders. A layer is matched when it exists in both the outcome and decision spaces, namely $0 \le j \le d_* := \min(m, n-m, K, n-K)$. On each matched layer,

$$A_u|_{V_j^{(m)}} = \lambda_j(u) T_j, \tag{3}$$

where $T_j : V_j^{(m)} \to V_j^{(K)}$ is a fixed isometry with normalization specified in Appendix B. Unmatched layers map to zero; on each matched nonconstant layer, $\lambda_j(u)$ determines whether it affects decision comparisons.

To compute these coefficients, put $q_{\max} = \min(m, K)$. Attainable overlaps satisfy $\max(0, m + K - n) \le q \le q_{\max}$; extend u arbitrarily to the remaining integers in $\{0, \dots, q_{\max}\}$. Expand $u(q) = \sum_{s=0}^{q_{\max}} \alpha_s \binom{q}{s}$ with $\alpha_s = (\Delta^s u)(0)$, where $\Delta u(q) = u(q+1) - u(q)$ is the forward difference. For $1 \le j \le d_*$,

$$\lambda_j(u) = \sum_{s=j}^{q_{\max}} \alpha_s \lambda_j^{(s)}, \tag{4}$$

$$\lambda_j^{(s)} = \sqrt{\binom{K-j}{s-j}\binom{n-s-j}{K-s}\binom{m-j}{s-j}\binom{n-s-j}{m-s}}.$$

These coefficients are independent of the extension to unattainable overlaps; Appendix B gives the derivation.

The constant layer contributes equally to every decision and cancels from comparisons. Define the active support by

$$J_L = \{1 \le j \le \min(m, n-m) : P_{V_j^{(m)}} W_L \neq \{0\}\}. \tag{5}$$

The next section identifies J_L through the nonzero coefficients, yielding R_W without constructing the full loss matrix.

4 Active Support: Sufficiency and Necessity

4.1 Active-support characterization and sufficiency

Theorem 1 (Active-support characterization and sufficiency)**.** *For every matched nonconstant layer,*

$$j \in J_L \iff \lambda_j(u) \neq 0, \qquad W_L = \bigoplus_{j \in J_L} V_j^{(m)}, \qquad \dim W_L = \sum_{j \in J_L} \left(\binom{n}{j} - \binom{n}{j-1} \right). \tag{6}$$

[1] Appendix D treats randomized interfaces and decoders through their induced distributions over decisions.

Hence the active representation $R_J(p) = P_{\oplus_{j\in J_L} V_j^{(m)}} p$ equals $R_W(p)$. It preserves every expected-loss comparison, including ties, and is universally Bayes-sufficient for L.

Proof. Let $C_K = I - |\Omega_K|^{-1}\mathbf{1}\mathbf{1}^\top$ center decision coordinates. Since $u = -\ell$, $W_L = \mathrm{im}(A_u^\top C_K)$. Centering removes the constant layer. By (3), each remaining matched block is $\lambda_j(u)T_j^\top$, whose image is $V_j^{(m)}$ exactly when $\lambda_j(u) \neq 0$; unmatched blocks vanish. Orthogonality gives the direct sum and dimension formula. Sufficiency follows from (2). Appendix C gives the representation-theoretic details. □

4.2 Direction-wise necessity and decision cost

Every nonzero active direction separates a pair with no common optimal decision. Write $\mathrm{range}(f) = \max_S f(S) - \min_S f(S)$.

Theorem 2 (Active-subspace necessity and exact regret identity)**.** *Fix $0 \neq h \in W_L$, $N = |\Omega_m|$, and $p_0 = \mathbf{1}/N$. For $0 < \epsilon \leq (N\|h\|_\infty)^{-1}$, let $p^\pm = p_0 \pm \epsilon h$. These are feasible distributions, and $\mathrm{Bayes}_u(p^+) \cap \mathrm{Bayes}_u(p^-) = \varnothing$. If an interface has the same output, or the same output distribution, under p^+ and p^-, then every decoder based only on that interface induces one common distribution ρ over decisions and satisfies*

$$\frac{1}{2}\left\{\mathbb{E}_\rho \,\mathrm{Reg}_u(S; p^+) + \mathbb{E}_\rho \,\mathrm{Reg}_u(S; p^-)\right\} = \frac{\epsilon}{2}\,\mathrm{range}(A_u h) > 0. \tag{7}$$

Proof. Since $W_L \subseteq \mathbf{1}^\perp$, the stated bound makes $p^\pm$ nonnegative with unit sum. Put $f = A_u h$. If f were constant, then $\langle L_S - L_T, h\rangle = 0$ for every S, T, forcing $h \in W_L \cap W_L^\perp = \{0\}$. Thus f is nonconstant. Symmetry gives $A_u p_0 = c\mathbf{1}$, so $A_u p^\pm = c\mathbf{1} \pm \epsilon f$. The optimal sets are therefore $\arg\max f$ and $\arg\min f$, which are disjoint even with ties.

For each decision S, the regrets $\epsilon(\max f - f(S))$ and $\epsilon(f(S) - \min f)$ sum to $\epsilon\,\mathrm{range}(f)$. Identical interface output distributions induce a common decision distribution ρ under any fixed decoder. Averaging this identity over ρ proves (7). □

Any interface, including a nonlinear one, that fails to distinguish a constructed pair cannot be universally sufficient.

4.3 Minimum dimension and non-prefix interfaces

Corollary 1 (Minimum affine-linear interface dimension)**.** *Let $D_L = \dim W_L$. Among affine-linear interfaces $R(p) = Mp + c \in \mathbb{R}^d$, the smallest output dimension that can be universally Bayes-sufficient for L is*

$$D_L = \sum_{j\in J_L} \left(\binom{n}{j} - \binom{n}{j-1}\right). \tag{8}$$

Proof. An orthonormal basis matrix B of W_L gives a sufficient D_L-dimensional interface Bp, since $P_{W_L} p = B^\top B p$. If $d < D_L$, then $M|_{W_L}$ has a nonzero kernel vector h. Theorem 2 constructs $p^\pm$ with identical interface outputs and no common optimal decision, ruling out sufficiency. For $D_L = 0$, any fixed decision suffices. □

For $J_L \neq \emptyset$, Theorems 1–2 imply that the prefix $T_r = P_{V_1 \oplus \cdots \oplus V_r}$ is sufficient exactly when $r \geq \max J_L$, but it can contain redundant layers.

Non-prefix structure follows from complement symmetry. If $K = n/2$ and $u(q) + u(m-q) = c$ on attainable overlaps, all positive even-order layers are inactive (Appendix E). Majority thresholds with odd m satisfy this condition. For $u(q) = \mathbf{1}\{q \geq 2\}$ at $n = 6$, $K = m = 3$, the active support is $J_L = \{1, 3\}$; Table 2 compares its dimension with the two prefixes. In contrast, all-or-nothing utility $u(q) = \mathbf{1}\{q = m\}$ has prefix support when $m \leq K$; Appendix B gives its support and dimension, and Appendix E derives the threshold example.

Table 2: Layer identity and interface dimension. The threshold utility is $u(q) = \mathbf{1}\{q \geq 2\}$ at $n = 6$, $m = K = 3$. Sufficiency is with respect to every outcome distribution.

Retained layers	Dimension	Universally sufficient
$V_1 \oplus V_3$ (active)	10	Yes
$V_1 \oplus V_2$	14	No
$V_1 \oplus V_2 \oplus V_3$ (prefix)	19	Yes

5 Calibrated Learning of the Active Interface

Section 4 identifies the information needed for optimal decisions. We now construct a dimension-optimal convex objective for learning it. Calibration means that, at every p, vanishing surrogate excess risk yields vanishing decision regret after decoding. Let $\mathrm{CCdim}(L)$ be the smallest dimension admitting such a convex surrogate and decoder. This measures the dimension of the surrogate prediction space.

Corollary 2 (Exact calibration dimension)**.** *For every symmetric fixed-cardinality overlap loss,*

$$\mathrm{CCdim}(L) = \mathrm{affdim}(L) = \dim W_L = \sum_{j \in J_L} \left(\binom{n}{j} - \binom{n}{j-1} \right). \tag{9}$$

The value is zero for a report-constant loss.

Proof. The loss vectors have affine dimension $\mathrm{affdim}(L) = D_L$. The uniform distribution equalizes all decisions, since each has $\binom{K}{q}\binom{n-K}{m-q}$ outcomes at overlap q. At this full-support equalizer, the two-sided feasible subspace $\{v : \mathbf{1}^\top v = 0,\ v \perp W_L\}$ has dimension $|\Omega_m| - 1 - D_L$. The feasible-subspace bound of Ramaswamy & Agarwal (2016) gives $\mathrm{CCdim}(L) \geq D_L$; their affine construction gives the matching upper bound. Appendix F verifies the conditions for applying these bounds and gives the full argument. □

Squared loss learns the conditional mean $\xi(p)$ of an observed active-coordinate target $\phi(e)$, retaining every expected-loss comparison. With $D = D_L$, an orthonormal basis $w_1, \ldots, w_D$ of W_L, and a reference decision S_0, define

$$\phi(e) = (w_1[e], \ldots, w_D[e]), \quad \xi(p) = \mathbb{E}_p \phi(e), \quad b(S)_r = \langle L_S - L_{S_0}, w_r \rangle. \tag{10}$$

For $\psi(z; e) = \|z - \phi(e)\|_2^2$, write $\mathrm{Reg}_\psi(z; p) = \mathbb{E}_p \psi(z; e) - \inf_{z'} \mathbb{E}_p \psi(z'; e)$ for its conditional excess risk.

Corollary 3 (Dimension-optimal squared surrogate)**.** *Fix an ordering of decisions and let $d(z)$ select the first decision in $\arg\min_S \langle b(S), z \rangle$. Then ψ is convex calibrated and, with $B_L = \max_{S,T} \|b(S) - b(T)\|$,*

$$\mathrm{Reg}_L(d(z); p) \leq B_L \|z - \xi(p)\|_2 = B_L \sqrt{\mathrm{Reg}_\psi(z; p)}. \tag{11}$$

Together with Corollary 2, it is dimension-optimal among convex calibrated Euclidean surrogates.

Proof. Orthonormal projection gives $\langle b(S), \xi(p) \rangle = R_L(S; p) - R_L(S_0; p)$, so $d(\xi(p))$ is Bayes-optimal. Completing the square gives $\mathrm{Reg}_\psi(z; p) = \|z - \xi(p)\|^2$. If $S = d(z)$ and S^* is Bayes-optimal, decoder optimality and Cauchy–Schwarz yield $R_L(S; p) - R_L(S^*; p) \leq \|z - \xi(p)\| \|b(S) - b(S^*)\|$, proving (11) and calibration. Corollary 2 gives dimension optimality; see Appendix G for further details. □

Layer-wise decision error. Let $p_j = P_{V_j^{(m)}} p$ and $\widehat{p}_j = P_{V_j^{(m)}} \widetilde{p}$, where $\widetilde{p}$ is the reconstructed prediction. A decoder within $\varepsilon_{\mathrm{dec}}$ of maximizing $A_u \widetilde{p}$ satisfies

$$\mathrm{Reg}_u(\widehat{S}; p) \leq \sqrt{2 \sum_{j \in J_L} \lambda_j(u)^2 \|p_j - \widehat{p}_j\|_2^2} + \varepsilon_{\mathrm{dec}}. \tag{12}$$

For omitted layers, $\widehat{p}_j = 0$, giving their weighted signal energy. On retained layers, $\lambda_j(u)^2$-weighted squared loss has conditional excess risk equal to the estimation term (Appendix G). The experiments relate these errors to decision margins.

6 Experiments: From Sufficient Information to Decision Value

Given a loss, our workflow identifies its active support, selects a predictor using validation layer errors, and decodes its active coordinates. Exact-distribution ablations test information requirements; synthetic learning studies examine estimation; a controlled Wiki task implements the workflow. We report utility gaps for exact distributions, decision utility for learned predictions, and reference regret for DGP-B dynamics. Primary comparisons enumerate all feasible decisions.

6.1 Which information does each loss require?

Experimental design. We first test the loss-to-interface mapping, holding distributions fixed and changing only the downstream utility. At $n = 6$, $m = K = 3$, each of 20 paired seeds generates 100 distributions by perturbing the uniform distribution along V_1, V_2, V_3. We evaluate additive utility $q/3$, pair reward $q(q-1)/6$, and threshold utility $\mathbf{1}[q \geq 2]$, all scaled to $[0, 1]$. For every distribution, we retain the constant component together with V_1, $V_1 \oplus V_2$, $V_1 \oplus V_3$, or the full vector, and include a constant-only baseline. No predictor is trained. Exact decoding maximizes projected expected utility. We measure the utility gap to the full-information optimum, evaluating each selected decision under the original distribution. Zero gap means that projection preserves optimal decision value.

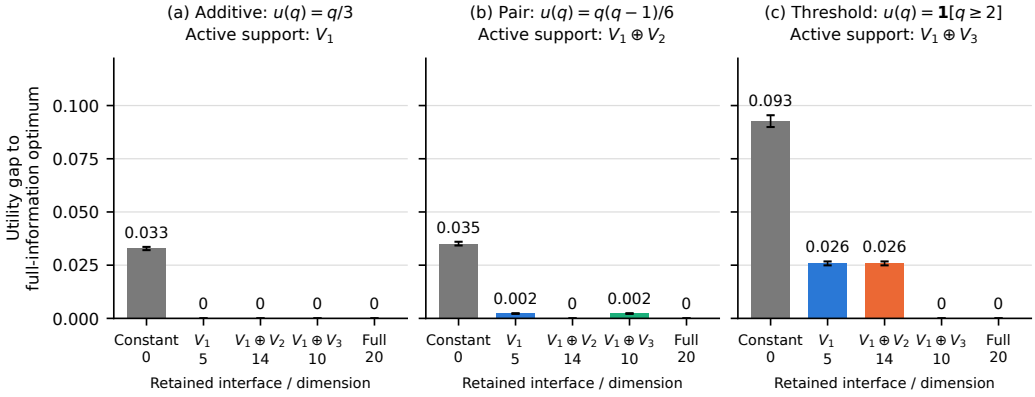


Figure 2: Loss-specific information requirements. Utility gaps on the same outcome distributions under three utilities. The 14-dimensional $V_1 \oplus V_2$ is sufficient for pair reward, whereas the 10-dimensional $V_1 \oplus V_3$ is sufficient for threshold utility.

Changing the loss changes which layers matter. For additive utility, all interfaces retaining V_1 have zero gap (Figure 2a). For pair reward, $V_1 \oplus V_2$ matches full prediction, while both V_1 and $V_1 \oplus V_3$ incur a gap of 0.002. For threshold utility, $V_1 \oplus V_3$ matches full prediction, while both V_1 and $V_1 \oplus V_2$ incur 0.026. Constant-only gaps are 0.033, 0.035, and 0.093, respectively. The required nonconstant layers are therefore $\{1\}$, $\{1, 2\}$, and $\{1, 3\}$ for the three utilities.

Layer identity matters more than dimension alone. The ordering of the two-layer interfaces reverses between pair reward and threshold utility. Adding the inactive layer changes neither decisions nor utility, whereas adding the missing active layer closes the gap. These exact-distribution comparisons isolate information loss; learned interfaces additionally depend on estimation quality. Appendix H.2 gives the construction and complete numerical results.

6.2 When does estimating active information improve decisions?

Experimental design. Three synthetic studies examine when the required information becomes useful through sample size, estimation error, and training dynamics. We first use continuous Top-3 outcomes, formed by selecting the three largest correlated item scores. At $n = 10$, $m = 3$, $K = 5$, we fit full predictors with $5{,}000$ or $50{,}000$ samples and 20 paired seeds. We freeze each predictor at epoch 100 and compare projections retaining different layers. Appendix H.3 gives architectures and supplementary endpoints.

More data can make estimated higher-order information useful. Under threshold utility, retaining V_3 changes success relative to V_1 by -1.073 percentage points (pp) at $5{,}000$ fitting samples and $+0.231$ pp at $50{,}000$ (Figure 3). Seeds favoring active increase from 1 to 15 out of 20. Retaining V_2 improves pair reward at both sizes, while additive utility is unchanged.

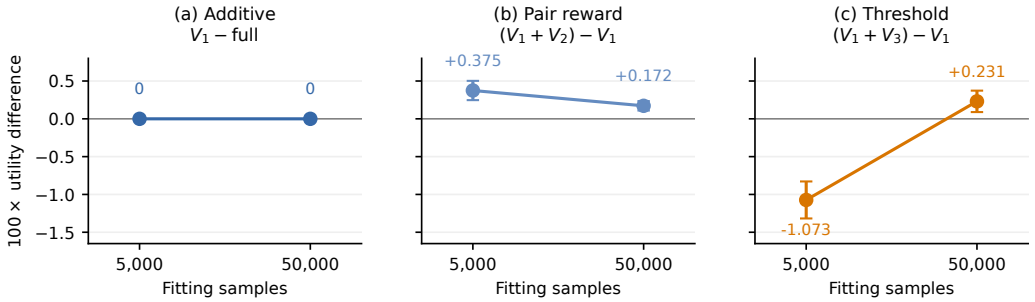


Figure 3: Continuous Top-3: projections of the same frozen predictor at epoch 100. Points show means and 95% paired-seed intervals at two fitting sizes.

Decision margins govern the effect of higher-order information. Within the Top-3 study, a unique V_1 optimum is preserved when its margin exceeds the range of V_3 score contributions. At $50{,}000$ samples, retaining V_3 changes 61.948% of decisions; 89.967% of these changes leave realized success unchanged. Gains are positive in the upper two validation-defined range-to-margin quartiles, but negative in the highest predicted-energy quartile (Appendix H.3).

Estimation error can outweigh the value of retained information. To quantify estimation error against the true conditional distribution, we use a controlled-posterior study at $n = 10$. At its highest V_3 strength, retaining exact V_3 improves threshold utility by 0.049 pp, while the direct active model's squared V_3 error remains 43.578 times the layer energy at $50{,}000$ samples. This setting combines a small exact-information gain with substantial estimation error, corresponding to the terms in Equation (12). Appendix H.4 gives the signal–sample grid.

Training loss and decision quality can evolve differently. To examine these quantities over training, we use the logit-generated DGP-B with independently trained V_1 and active predictors. Continued training lowers V_1 loss while its reference regret rises. Active reference regret falls and remains lower under both solvers, with an exact-decoder gap of 0.010 at epoch 100 (Figure 4).

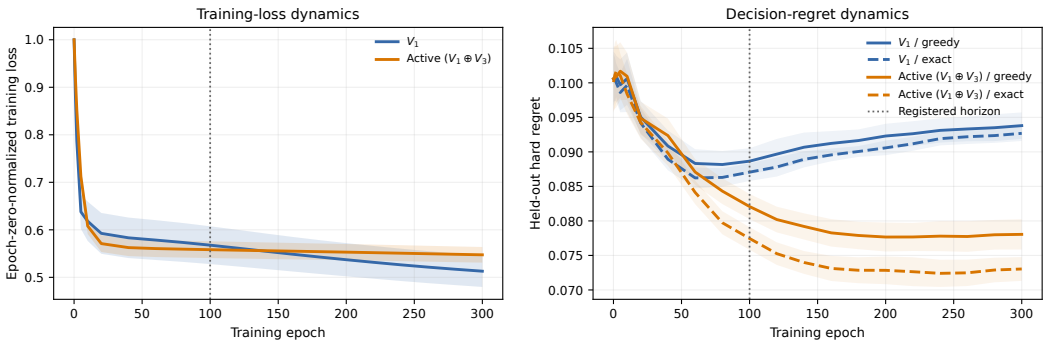


Figure 4: DGP-B, $8{,}000$ fitting samples and 20 paired seeds: within-run normalized loss (left) and reference regret (right), with 95% intervals. The reference also observes the latent shock; dotted lines mark epoch 100.

6.3 How does the workflow apply to real data?

Experimental design. We use a controlled WikiSeeAlso task with $n = 6$, $m = K = 3$. Each candidate set contains three true labels and three negatives with matched label statistics. All six candidates are randomly ordered. Three frozen constructions provide about $28{,}500$ training articles each and share $8{,}726$ test articles. Five objectives share Set MLP32, initialization, minibatch order, and a 300-epoch budget, with ten paired model seeds. Appendix I.2.1 gives checkpoint selection and source-level paired inference.

Loss-specific validation selects the predictor. We minimize $\sum_{j \in J_L} \lambda_j^2 \, \mathrm{MSE}_{\mathrm{val},j}$ over two checkpoints per objective, using observed validation targets. All three losses select CE in all 30 construction–seed groups, as does full-Brier validation. We freeze the predictor and decode 5, 14, or 10 active coordinates for additive, pair, or threshold utility.

Table 3: Wiki threshold success from the same validation-selected predictor within each construction and seed.

Interface	Coordinates	Success (%)	Gain over V_1 (pp)
V_1	5	96.383	0
$V_1 \oplus V_2$	14	96.383	0
$V_1 \oplus V_3$ (active)	10	96.659	+0.276
Full prediction	20	96.659	+0.276

Active coordinates preserve full decisions and improve on V_1. Retaining V_3 gives a significant 0.276 pp threshold gain, positive in all ten seeds (Table 3). Active and full decisions coincide, as do V_1 and $V_1 \oplus V_2$ decisions. Ten active coordinates replace 20 stored probabilities. Additive decisions are unchanged; the pair-reward gain over V_1 is 0.017 pp and is not significant.

Decision margins locate the value of higher-order information. Active and V_1 disagree on 9.396% of cases, with a conditional success gain of 2.935 pp. In the highest validation-defined range-to-margin bin, the gain is 0.763 pp. The highest predicted-V_3-norm bin instead has zero gain: concentrated predictions already give the same decision.

The interface and its estimator are separate choices. The workflow exceeds direct unweighted active training by 1.292 pp, an estimator comparison distinct from the 0.276 pp information effect. Appendix I.2 reports all matched objectives, including negative projection gains for direct active training.

7 Discussion and Implications

The active support specifies the information required by a loss and its minimum affine-linear interface dimension. This characterization guides three design choices: what the predictor exposes, how that information is estimated, and how it is decoded. The loss profile identifies relevant layers and their score-error weights; the estimator determines how accurately they are learned. The controlled Wiki task illustrates this separation: full CE learns useful higher-order information, active projection preserves its decisions, and removing V_3 reduces threshold utility.

Support identification, prediction, and decoding have distinct computational costs. The coefficient formula identifies active layers without constructing the full loss matrix. Bounded active order r gives $D_L = O(n^r)$ interface coordinates; exact enumeration considers $\binom{n}{K}$ decisions. The trained parameter count depends on whether these coordinates are predicted directly or extracted from a full predictor.

Limitations. The characterization assumes fixed outcome and decision cardinalities and symmetric overlap losses. Variable cardinalities and item-dependent losses require structure beyond a single Johnson decomposition.

Reproducibility statement

Complete proofs are provided in Appendices A–G. Appendices H–I specify data construction, models, training and decoding, independent sample sizes, and statistical procedures. The accompanying reproduction guide maps the reported tables and figures to experiment configurations and scripts.

AI use statement

Generative AI tools assisted with proof development and verification, counterexample search, experiment code, figure preparation, and manuscript polishing. The authors independently reviewed and verified all AI-assisted material and take responsibility for the final paper, claims, code, and artifacts.

References

Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. What data enables optimal decisions? an exact characterization for linear optimization. *Advances in Neural Information Processing Systems*, 38:166020–166049, 2026.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable pertubed optimizers. *Advances in neural information processing systems*, 33:9508–9519, 2020.

David Blackwell. Equivalent comparisons of experiments. *The annals of mathematical statistics*, pp. 265–272, 1953.

Andries E Brouwer and Willem H Haemers. Distance-regular graphs. In *Spectra of graphs*, pp. 177–185. Springer, 2011.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. *Advances in neural information processing systems*, 29, 2016.

Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21 (98):1–67, 2020.

Philippe Delsarte. An algebraic approach to the association schemes of coding theory. *Philips Res. Rep. Suppl.*, 10:vi+–97, 1973.

Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. *Advances in neural information processing systems*, 30, 2017.

Othman El Balghiti, Adam N Elmachtoub, Paul Grigas, and Ambuj Tewari. Generalization bounds in the predict-then-optimize framework. *Advances in neural information processing systems*, 32, 2019.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management science*, 68(1): 9–26, 2022.

Yuval Filmus. Orthogonal basis for functions over a slice of the boolean hypercube. *arXiv preprint arXiv:1406.0142*, 2014.

Jessie Finocchiaro, Rafael M Frongillo, and Bo Waggoner. An embedding framework for the design and analysis of consistent polyhedral surrogates. *Journal of Machine Learning Research*, 25(63): 1–60, 2024.

Rafael Frongillo and Ian A Kash. Vector-valued property elicitation. In *Conference on Learning Theory*, pp. 710–727. PMLR, 2015.

Rafael M Frongillo and Ian A Kash. Elicitation complexity of statistical properties. *Biometrika*, 108 (4):857–879, 2021.

Daniel Henry Gottlieb. A certain class of incidence matrices. *Proceedings of the American Mathematical Society*, 17(6):1233–1237, 1966.

L Le. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, pp. 1419–1455, 1964.

Marin Vlastelica Pogancic, Anselm Paulus, Vit Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In *ICLR*, 2020.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016.

Harish G Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. *Advances in Neural Information Processing Systems*, 26, 2013.

Erik Schultheis, Wojciech Kotlowski, Marek Wydmuch, Rohit Babbar, Strom Borman, and Krzysztof Dembczynski. Consistent algorithms for multi-label classification with macro-at-k metrics. In *International Conference on Learning Representations*, volume 2024, pp. 13440–13469, 2024.

Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning. *arXiv preprint arXiv:2606.04045*, 2026.

Brendan van Rooyen and Robert C Williamson. Le cam meets lecun: Deficiency and generic feature learning. *arXiv preprint arXiv:1402.4884*, 2014.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

Richard M Wilson. A diagonal form for the incidence matrices of t-subsets vs. k-subsets. *European Journal of Combinatorics*, 11(6):609–615, 1990.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, et al. Answering complex open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*, 2020.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Yuhan Ye, Saurabh Amin, and Asuman Ozdaglar. Learning decision-sufficient representations for linear optimization. *arXiv preprint arXiv:2603.18551*, 2026.

Mingyuan Zhang. Exact rank and convex calibration dimension lower bounds for the multi-label f1 loss. *arXiv preprint arXiv:2608.08399*, 2026a.

Mingyuan Zhang. Exponential convex calibration dimension for the multi-label jaccard measure. *arXiv preprint arXiv:2608.13549*, 2026b.

Mingyuan Zhang, Harish Guruprasad Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pp. 11246–11255. PMLR, 2020.

Appendices A–G give the representation-theoretic background and complete proofs. Appendix H specifies the synthetic experiments and supplementary comparisons; Appendix I gives the WikiSeeAlso protocols, estimation controls, and diagnostics.

A JOHNSON REPRESENTATION BACKGROUND

We summarize the Johnson decomposition underlying the layer-wise action of overlap utilities. Let $\Omega_k = \binom{[n]}{k}$, and equip $\mathbb{R}^{\Omega_k}$ with the standard Euclidean inner product. Permuting the items induces an action of S_n on this function space. Its orthogonal decomposition is (Filmus, 2014)

$$\mathbb{R}^{\Omega_k} = \bigoplus_{j=0}^{\min(k,n-k)} V_j^{(k)}, \qquad \dim V_j^{(k)} = \binom{n}{j} - \binom{n}{j-1},$$

with $\binom{n}{-1} = 0$. Each $V_j^{(k)}$ is an irreducible representation of type $(n-j, j)$ and occurs exactly once. The layer $V_0^{(k)}$ consists of constant functions, while $V_1^{(k)}$ consists of centered additive item functions.

An operator between two slices is equivariant if it commutes with simultaneous relabeling of the items. Since $|\pi S \cap \pi e| = |S \cap e|$ for every $\pi \in S_n$, the overlap-utility matrix $A_u[S, e] = u(|S \cap e|)$ is equivariant. The irreducible types are distinct, so Schur's lemma implies that A_u maps each source layer only into the matching target layer. Source layers with no matching target layer are annihilated.

On every matched layer,

$$A_u|_{V_j^{(m)}} = \lambda_j(u) T_j, \qquad 0 \le j \le \min(m, n-m, K, n-K),$$

where $T_j : V_j^{(m)} \to V_j^{(K)}$ is an equivariant isometry. The intertwining map is unique up to a scalar. Appendix B constructs T_j through common harmonic lifts, fixes its normalization and sign, and derives the coefficient $\lambda_j(u)$.

B SPECTRAL COEFFICIENT FORMULA: COMPLETE DERIVATION

We derive the spectral coefficients by expanding the utility profile in a binomial basis and computing the action of each inclusion kernel under a common Euclidean normalization.

Utility expansion. Put $q_{\max} = \min(m, K)$ and define $B_{k,s}[X, A] = \mathbf{1}\{A \subseteq X\}$ for $X \in \Omega_k$, $A \in \Omega_s$. We use the ordinary convention $\binom{a}{b} = 0$ unless $a, b \in \mathbb{Z}$ and $0 \le b \le a$. Attainable overlaps satisfy $\max(0, m + K - n) \le q \le q_{\max}$. Extend u arbitrarily to the remaining integers in $\{0, \dots, q_{\max}\}$ and write its Newton expansion

$$u(q) = \sum_{s=0}^{q_{\max}} \alpha_s \binom{q}{s}, \qquad \alpha_s = (\Delta^s u)(0).$$

Here $\Delta u(q) = u(q+1) - u(q)$ is the forward difference. Counting s-sets contained in $S \cap e$ gives

$$A_u = \sum_{s=0}^{q_{\max}} \alpha_s B_{K,s} B_{m,s}^\top. \tag{B.1}$$

Harmonic lifts and normalization. For $1 \le j \le \lfloor n/2 \rfloor$, let $H_j = \ker(B_{j,j-1}^\top)$ and put $H_0 = \mathbb{R}$. For $h \in H_j$ and $j \le k \le n - j$, write $x_k(h) = B_{k,j} h$. These harmonic lifts realize the Johnson layers: $x_k(H_j) = V_j^{(k)}$ (Filmus, 2014). Their Euclidean norms satisfy

$$\|x_k(h)\|^2 = \rho_{k,j} \|h\|^2, \qquad \rho_{k,j} = \binom{n-2j}{k-j}. \tag{B.2}$$

To obtain this identity, let $U_k = B_{k+1,k}$ for $j \le k < n - j$. Counting inclusions and using harmonicity gives $U_k x_k = (k+1-j) x_{k+1}$ and $U_k^\top x_{k+1} = (n-k-j) x_k$. Taking inner products yields

$$(k+1-j)\|x_{k+1}(h)\|^2 = (n-k-j)\|x_k(h)\|^2.$$

Iterating from $x_j(h) = h$ proves (B.2).

For every matched $0 \le j \le d_* := \min(m, n-m, K, n-K)$, define the harmonic-lift-aligned isometry

$$T_j x_m(h) = \sqrt{\frac{\rho_{m,j}}{\rho_{K,j}}}\, x_K(h). \tag{B.3}$$

Equation (B.2) makes the lift injective, so this defines a linear isometry on the entire layer, with its sign fixed by the common harmonic lift.

Layer coefficients. Fix a matched layer $0 \le j \le d_*$. For $j \le s \le k$, incidence composition gives

$$B_{k,s} B_{s,j} = \binom{k-j}{s-j} B_{k,j}. \tag{B.4}$$

For $j \le s \le q_{\max}$, equivariance gives $B_{m,s}^\top x_m(h) = c_{m,s,j} x_s(h)$. Pairing with $x_s(h)$ and applying (B.4) at $k = m$ yields

$$c_{m,s,j} = \binom{m-j}{s-j} \frac{\rho_{m,j}}{\rho_{s,j}} = \binom{n-s-j}{m-s}.$$

Thus

$$B_{m,s}^\top x_m(h) = \binom{n-s-j}{m-s} x_s(h), \tag{B.5}$$

and hence

$$B_{K,s} B_{m,s}^\top x_m(h) = \binom{K-j}{s-j} \binom{n-s-j}{m-s} x_K(h). \tag{B.6}$$

For $s < j$, the intermediate s-slice has no layer of type $(n-j, j)$, so this action is zero. Comparing (B.6) with the normalized isometry in (B.3) gives

$$\lambda_j^{(s)} = \binom{K-j}{s-j} \binom{n-s-j}{m-s} \sqrt{\frac{\rho_{K,j}}{\rho_{m,j}}} \qquad (j \le s \le q_{\max}).$$

Substituting (B.2) and simplifying gives the nonnegative basis coefficient

$$\lambda_j^{(s)} = \sqrt{\binom{K-j}{s-j} \binom{n-s-j}{K-s} \binom{m-j}{s-j} \binom{n-s-j}{m-s}}. \tag{B.7}$$

By (B.1),

$$\boxed{\lambda_j(u) = \sum_{s=j}^{q_{\max}} \alpha_s \lambda_j^{(s)}, \qquad A_u|_{V_j^{(m)}} = \lambda_j(u) T_j.} \tag{B.8}$$

Although each $\lambda_j^{(s)}$ is nonnegative, the signed weights α_s can make $\lambda_j(u)$ negative or zero by cancellation. Changing the extension on unattainable overlaps leaves A_u unchanged and therefore preserves every coefficient under the fixed normalization. Source layers with no matching target layer are annihilated.

AON. For $u(q) = \mathbf{1}\{q = m\}$ with $1 \le m \le K < n$, use the same indicator profile on all $q \in \{0, \dots, m\}$. Its only nonzero Newton coefficient is $\alpha_m = 1$, so

$$\lambda_j^2 = \binom{K-j}{m-j} \binom{n-m-j}{K-m}.$$

This is positive exactly for $0 \le j \le \min(m, n-K)$; after decision centering, $J = \{1, \dots, r\}$ with $r = \min(m, n-K)$. The active and smallest sufficient prefix interfaces coincide, with dimension $\sum_{j=1}^{r} [\binom{n}{j} - \binom{n}{j-1}] = \binom{n}{r} - 1$. If $K < m$, containment is impossible, the utility matrix is zero, and $J = \emptyset$; a zero-dimensional interface suffices.

Exact coefficient checks. Exact overlap matrices and harmonic lifts verify the coefficient formula independently for all 140 interior triples $1 \le m, K < n$ through $n = 8$, across 9,559 utility/triple cases. The checks include restricted-overlap extensions and engineered cancellations, agreeing in every case.

C ACTIVE-SUPPORT CHARACTERIZATION AND SUFFICIENCY

We use the layer action derived in Appendix B to identify the decision-relevant subspace and prove that its projection preserves all Bayes decisions, establishing Theorem 1.

Decision centering. Let $C_K = I - |\Omega_K|^{-1}\mathbf{1}\mathbf{1}^\top$ center decision coordinates. Each row of $C_K L_\ell$ is an average of loss differences, $L_S - |\Omega_K|^{-1}\sum_T L_T = |\Omega_K|^{-1}\sum_T (L_S - L_T)$. Conversely, subtracting two centered rows recovers $L_S - L_T$. Since $A_u = -L_\ell$,

$$W_L = \mathrm{rowspan}(C_K L_\ell) = \mathrm{im}(A_u^\top C_K).$$

The operator C_K removes the constant decision layer $V_0^{(K)}$ and is the identity on every $V_j^{(K)}$ with $j \geq 1$.

Active layers. On each matched nonconstant layer, the adjoint of $A_u|_{V_j^{(m)}} = \lambda_j(u) T_j$ is $A_u^\top|_{V_j^{(K)}} = \lambda_j(u) T_j^\top$. Since T_j is an isometry onto $V_j^{(K)}$, its adjoint maps that layer onto $V_j^{(m)}$. Thus the block image is the entire source layer when $\lambda_j(u) \neq 0$ and zero when $\lambda_j(u) = 0$. Unmatched layers contribute zero. Orthogonality of the Johnson layers therefore gives

$$J_L = \{1 \leq j \leq d_* : \lambda_j(u) \neq 0\}, \qquad W_L = \bigoplus_{j \in J_L} V_j^{(m)}, \qquad d_* = \min(m, n-m, K, n-K).$$

Using the layer dimensions from Appendix A,

$$\dim W_L = \sum_{j \in J_L} \left(\binom{n}{j} - \binom{n}{j-1} \right).$$

Bayes sufficiency. For any $p \in \Delta_{\Omega_m}$, put $z = P_{W_L} p$. Since $L_S - L_T \in W_L$ for every pair of decisions,

$$\langle L_S - L_T, p \rangle = \langle L_S - L_T, z \rangle.$$

All expected-loss comparisons, including ties, are preserved, so $\mathrm{Bayes}_L(p) = \arg\min_{S \in \Omega_K} \langle L_S, z \rangle$. Selecting the first minimizer under a fixed ordering of Ω_K defines a single decoder that is Bayes for every p. Hence $R_J(p) = P_{W_L} p$ is universally Bayes-sufficient. If $J_L = \emptyset$, then $W_L = \{0\}$, all decisions have equal loss, and any fixed decision suffices.

For $L' = aL + \mathbf{1}b^\top$ with $a > 0$, the identity $C_K L' = a C_K L$ gives $W_{L'} = W_L$. Positive scaling and outcome-only shifts also preserve all Bayes decisions.

D ACTIVE-SUBSPACE NECESSITY PROOF DETAILS

We prove Theorem 2 by constructing two distributions along any nonzero active direction. Their optimal decision sets are disjoint, and an interface that cannot distinguish them incurs an exact positive average regret.

Feasible distributions. Let $0 \neq h \in W_L$, set $N = |\Omega_m|$, and let $p_0 = \mathbf{1}/N$. Appendix C gives $W_L \subseteq \mathbf{1}^\perp$, so $p^\pm = p_0 \pm \epsilon h$ have unit mass. For

$$0 < \epsilon \leq (N \|h\|_\infty)^{-1},$$

both distributions are coordinatewise nonnegative, and strict inequality makes them interior to the probability simplex.

Disjoint optimal decisions. Put $f = A_u h$, with $u = -\ell$. If f were constant, then for every pair of decisions S, T,

$$0 = f(S) - f(T) = -\langle L_S - L_T, h \rangle.$$

Thus $h \in W_L^\perp$. Together with $h \in W_L$, this would imply $h = 0$, a contradiction; hence f is nonconstant. By permutation symmetry, every decision has the same expected utility under the uniform distribution, so $A_u p_0 = c\mathbf{1}$ for some scalar c. Therefore

$$A_u p^+ = c\mathbf{1} + \epsilon f, \qquad A_u p^- = c\mathbf{1} - \epsilon f.$$

Thus Bayes decisions for the two distributions are respectively $\arg\max f$ and $\arg\min f$. These sets remain disjoint under ties: if some decision attained both extrema, then $\max f = \min f$, contradicting that f is nonconstant.

Exact regret identity. Suppose the interface has the same output under p^+ and p^-, or the same output distribution if randomized. Any fixed decoder based only on this output, deterministic or randomized, then induces a common decision distribution ρ. For each fixed decision S,

$$\mathrm{Reg}_u(S; p^+) = \epsilon(\max f - f(S)), \qquad \mathrm{Reg}_u(S; p^-) = \epsilon(f(S) - \min f).$$

Their sum is $\epsilon \operatorname{range}(f)$ for every S. Averaging over the common distribution ρ yields

$$\frac{1}{2}\left\{\mathbb{E}_\rho \mathrm{Reg}_u(S; p^+) + \mathbb{E}_\rho \mathrm{Reg}_u(S; p^-)\right\} = \frac{\epsilon}{2}\operatorname{range}(f) > 0.$$

Consequently, no decoder based on this interface can be Bayes-optimal for both distributions.

E Complement symmetry and non-prefix interfaces

Complement symmetry explains how an overlap loss can require a higher-order layer while ignoring an intermediate one.

Complement action. Assume $K = n/2$, so the complement of a decision is again in Ω_K. Define $(Cf)(S) = f(S^c)$. Using the harmonic lifts from Appendix B, let $h \in H_j$ and write $x_K(h)(S) = \sum_{A \subseteq S, |A|=j} h(A)$. For $|B| < j$, summing the harmonicity equations over all $(j-1)$-sets containing B gives $(j - |B|) \sum_{A \supseteq B, |A|=j} h(A) = 0$. Inclusion–exclusion therefore yields

$$x_K(h)(S^c) = \sum_{\substack{B \subseteq S \\ |B| \le j}} (-1)^{|B|} \sum_{\substack{A \supseteq B \\ |A|=j}} h(A) = (-1)^j x_K(h)(S).$$

Only the terms with $|B| = j$ survive. Since these lifts span $V_j^{(K)}$, $C|_{V_j^{(K)}} = (-1)^j I$.

Vanishing even layers. Suppose $u(q) + u(m-q) = c$ for every attainable overlap q. Since $|S^c \cap e| = m - |S \cap e|$, this condition gives $CA_u = c\mathbf{1}\mathbf{1}^\top - A_u$. For $v \in V_j^{(m)}$ in a matched layer with $j \ge 1$, we have $v \perp \mathbf{1}$, so $CA_u v = -A_u v$. On the other hand, $A_u v \in V_j^{(K)}$ gives $CA_u v = (-1)^j A_u v$. Combining the identities yields

$$\left(1 + (-1)^j\right) A_u v = 0.$$

Thus every matched even-order layer with $j \ge 2$ has $\lambda_j(u) = 0$. For odd m, the majority utility $u(q) = \mathbf{1}\{q \ge (m+1)/2\}$ satisfies the condition with $c = 1$. The coefficient formula in Appendix B determines which odd-order layers are active.

The three-item threshold. For $n = 6$, $m = K = 3$, and $u(q) = \mathbf{1}\{q \ge 2\}$, the profile values at $q = 0, \ldots, 3$, the Newton coefficients, and the spectral coefficients are

$$u = (0, 0, 1, 1), \qquad \alpha = (0, 0, 1, -2), \qquad (\lambda_0, \lambda_1, \lambda_2, \lambda_3) = (10, 4, 0, -2).$$

Hence $J_L = \{1, 3\}$: V_2 is inactive while V_3 is active. The dimensions are $\dim V_1 = \dim V_3 = 5$ and $\dim V_2 = 9$. The active interface $V_1 \oplus V_3$ is therefore 10-dimensional and universally Bayes-sufficient. The second-order prefix $V_1 \oplus V_2$ has 14 dimensions but omits the active V_3 layer and is insufficient by Theorem 2. The third-order prefix $V_1 \oplus V_2 \oplus V_3$ is sufficient, but uses 19 dimensions. Corollaries 1 and 2 establish 10 as the minimum dimension for universally Bayes-sufficient affine-linear interfaces and convex calibrated Euclidean surrogates, respectively.

F EXACT CONVEX CALIBRATION DIMENSION

We prove Corollary 2 by matching the affine-dimension upper bound with a lower bound at the uniform distribution. Write $N = |\Omega_m|$, $D_L = \dim W_L$, and $p_0 = \mathbf{1}/N$. A global constant shift makes the loss nonnegative without changing calibration or affine dimension. Merge identical loss rows, which only relabels equivalent decisions and leaves $\mathrm{CCdim}(L)$ and W_L unchanged. By symmetry, all remaining rows are coordinate permutations of one another, so they have the same Euclidean norm and the same mean $c = \langle L_S, p_0 \rangle$. For any such row L_S, the vector $p_\delta = p_0 + \delta(c\mathbf{1} - L_S)$ is a full-support probability distribution for sufficiently small $\delta > 0$. For every other row L_T,

$$\langle L_T - L_S, p_\delta \rangle = \frac{\delta}{2} \|L_T - L_S\|_2^2 > 0.$$

Thus each retained decision is uniquely optimal somewhere, as required in the framework of Ramaswamy & Agarwal (2016).

The affine dimension of the loss rows is $\dim \mathrm{span}\{L_S - L_T\} = D_L$. Transposing our matrix to the decision-column convention of Ramaswamy & Agarwal (2016), their Theorem 12 gives $\mathrm{CCdim}(L) \le D_L$.

For the matching lower bound, let

$$Q_S = \{p \in \Delta_{\Omega_m} : \langle L_S - L_T, p \rangle \le 0 \text{ for all } T\}$$

be the region where S is Bayes-optimal. The uniform distribution p_0 equalizes all decisions because the loss rows have equal sums. Hence every decision-comparison inequality is tight at p_0, while all probability coordinates are strictly positive. A perturbation is feasible in both directions precisely when it preserves total mass and leaves all these comparisons unchanged. By Lemma 15 of Ramaswamy & Agarwal (2016), the resulting feasible subspace is

$$\mathcal{F}_S = \{v : \mathbf{1}^\top v = 0,\ v \perp W_L\}, \qquad \dim \mathcal{F}_S = N - 1 - D_L,$$

where the dimension follows from $W_L \subseteq \mathbf{1}^\perp$. Since p_0 has support size N, their Theorem 16 yields

$$\mathrm{CCdim}(L) \ge N - \dim \mathcal{F}_S - 1 = D_L.$$

Together with the upper bound and Theorem 1, this proves the stated equality. If $D_L = 0$, all loss rows coincide and any fixed decision is optimal, so a zero-dimensional surrogate suffices.

G CALIBRATED SURROGATES AND LAYER-WISE DECISION ERROR

We prove Corollary 3, separate omission and estimation errors in the decision-regret bound, and derive loss-specific training weights. A margin criterion then gives a sufficient condition for preserving the decision when adding a predicted component.

Squared surrogate and calibration. Put $D = \dim W_L$. Let $w_1, \dots, w_D$ be an orthonormal basis of W_L and fix a reference decision S_0. Define

$$\phi(e) = (w_1[e], \dots, w_D[e]), \quad \xi(p) = \mathbb{E}_p \phi(e), \quad b(S)_r = \langle L_S - L_{S_0}, w_r \rangle.$$

Since $L_S - L_{S_0} \in W_L$, orthonormal reconstruction gives

$$\mathbb{E}_p[L(S, e) - L(S_0, e)] = \langle b(S), \xi(p) \rangle.$$

Fix an ordering of decisions and define the decoder

$$d(z) = \text{the first decision in } \arg\min_S \langle b(S), z \rangle.$$

Then $d(\xi(p)) \in \mathrm{Bayes}_L(p)$. Define the convex surrogate $\psi(z; e) = \|z - \phi(e)\|_2^2$. Completing the square gives

$$R_\psi(z; p) = \|z - \xi(p)\|_2^2 + \mathbb{E}_p \|\phi(e) - \xi(p)\|_2^2, \qquad \mathrm{Reg}_\psi(z; p) = \|z - \xi(p)\|_2^2.$$

If $S = d(z)$ and S^* is Bayes-optimal, link optimality gives

$$\mathrm{Reg}_L(S; p) \le \langle \xi(p) - z, b(S) - b(S^*) \rangle \le B_L \|z - \xi(p)\|_2,$$

where $B_L = \max_{S,T} \|b(S) - b(T)\|_2$. Hence

$$\mathrm{Reg}_L(d(z); p) \le B_L \sqrt{\mathrm{Reg}_\psi(z; p)}.$$

At fixed p, the smallest positive decision regret, when present, is strictly positive because the decision set is finite. The bound therefore keeps surrogate excess risk away from zero for non-Bayes decisions, proving convex calibration. Corollary 2 gives dimension optimality. If $D = 0$, all loss rows coincide; the zero surrogate on $\mathbb{R}^0$ with any fixed decision is calibrated.

Layer-wise omission and estimation error. For $u = -\ell$, minimizing expected loss is equivalent to maximizing expected utility. Let $p \in \Delta_{\Omega_m}$, let $\widetilde{p} \in \mathbb{R}^{\Omega_m}$ be a reconstructed prediction, and write $\lambda_j = \lambda_j(u)$. Suppose the decoder returns $\widehat{S}$ satisfying

$$\max_S (A_u \widetilde{p})_S - (A_u \widetilde{p})_{\widehat{S}} \le \varepsilon_{\mathrm{dec}}.$$

Write $p_j = P_{V_j^{(m)}} p$ and $\widehat{p}_j = P_{V_j^{(m)}} \widetilde{p}$. Utility regret is $\mathrm{Reg}_u(S; p) = \max_T (A_u p)_T - (A_u p)_S$. With $S^* \in \arg\max_S (A_u p)_S$ and $f = A_u(p - \widetilde{p})$, the decoder guarantee gives

$$\mathrm{Reg}_u(\widehat{S}; p) \le f(S^*) - f(\widehat{S}) + \varepsilon_{\mathrm{dec}} \le \mathrm{range}(f) + \varepsilon_{\mathrm{dec}}. \tag{G.1}$$

The constant component cancels from the range. With $C_K = I - |\Omega_K|^{-1} \mathbf{1}\mathbf{1}^\top$, the block decomposition gives

$$C_K f = \sum_{j \in J_L} \lambda_j T_j (p_j - \widehat{p}_j), \qquad \|C_K f\|_2^2 = \sum_{j \in J_L} \lambda_j^2 \|p_j - \widehat{p}_j\|_2^2.$$

The second identity uses the isometry of each T_j and orthogonality of its target layers. For distinct decisions S, T, $\|\mathbf{e}_S - \mathbf{e}_T\|_2 = \sqrt{2}$, so $\mathrm{range}(f) \le \sqrt{2} \|C_K f\|_2$. Substitution into (G.1) proves (12).

For a retained layer set U, set $\widehat{p}_j = 0$ on each omitted nonconstant layer $j \notin U$. The squared score error then decomposes as

$$\sum_{j \in J_L} \lambda_j^2 \|p_j - \widehat{p}_j\|_2^2 = \underbrace{\sum_{j \in J_L \cap U} \lambda_j^2 \|p_j - \widehat{p}_j\|_2^2}_{\text{estimation}} + \underbrace{\sum_{j \in J_L \setminus U} \lambda_j^2 \|p_j\|_2^2}_{\text{omission}}. \tag{G.2}$$

Taking expectations over contexts and applying Jensen gives

$$\mathbb{E}\,\mathrm{Reg}_u \le \sqrt{2 \sum_{j \in J_L} \lambda_j^2 \mathbb{E}\|p_j - \widehat{p}_j\|_2^2} + \mathbb{E}\varepsilon_{\mathrm{dec}}.$$

Loss-specific training weights. Let Q_j have orthonormal columns spanning $V_j^{(m)}$, and define the observed target $\phi_j(e) = Q_j^\top \mathbf{e}_e$ and its conditional mean $\xi_j(p) = Q_j^\top p$. For retained coordinates $z = (z_j)_{j \in J_L \cap U}$, consider

$$\psi_{\lambda,U}(z; e) = \sum_{j \in J_L \cap U} \lambda_j(u)^2 \|z_j - \phi_j(e)\|_2^2. \tag{G.3}$$

Completing the square in each layer gives

$$\mathrm{Reg}_{\psi_{\lambda,U}}(z; p) = \sum_{j \in J_L \cap U} \lambda_j(u)^2 \|z_j - \xi_j(p)\|_2^2 = \sum_{j \in J_L \cap U} \lambda_j(u)^2 \|\widehat{p}_j - p_j\|_2^2,$$

where $\widehat{p}_j = Q_j z_j$. Thus weighting by $\lambda_j(u)^2$ makes the conditional surrogate excess risk equal to the retained-layer score error in (G.2). With $U = J_L$, decode $\widetilde{p} = \sum_{j \in J_L} Q_j z_j$ using the guarantee above. Then

$$\mathrm{Reg}_u(\widehat{S}; p) \le \sqrt{2\,\mathrm{Reg}_{\psi_{\lambda,J_L}}(z; p)} + \varepsilon_{\mathrm{dec}}.$$

All weights are positive on J_L, so exact decoding ($\varepsilon_{\mathrm{dec}} = 0$) gives a convex calibrated surrogate in D_L coordinates.

Decision stability and margins. Let S_g be the unique maximizer of base decision scores g, with margin $M = g(S_g) - \max_{T \neq S_g} g(T)$. For an additional score component h,

$$(g+h)(S_g) - (g+h)(T) \geq M - \mathrm{range}(h), \qquad T \neq S_g.$$

Hence $M > \mathrm{range}(h)$ preserves the maximizing decision. For the threshold comparison between V_1 and $V_1 \oplus V_3$ in a frozen predictor, take $g = A_u \widehat{p}_1$ and $h = A_u \widehat{p}_3$. The resulting criterion compares the predicted contribution of V_3 with the existing V_1 decision gap, yielding the prediction-defined diagnostic used in Appendix H.3.

H Synthetic experimental protocols and supplementary results

H.1 Study overview and evaluation metrics

The studies follow the questions in Sections 6.1–6.2: which information a loss requires, when learning it improves decisions, and how training and decoding affect its value. The exact-distribution ablation and the controlled-posterior teacher evaluate Bayes regret under a known outcome distribution. Continuous Top-3 uses held-out realized utility. DGP-A–C use reference regret under $p(e \mid x, \eta)$, which conditions on the observed input and a latent shock. All studies use paired test instances across interfaces. We distinguish the number of actual fitting contexts, N_{fit}, from the pool size before validation splitting, N_{pool}. The DGP-A–C MLPs reserve 20% of that pool for validation; Top-3 and controlled-posterior studies draw validation contexts separately, and ridge uses its entire fitting pool. Utility differences are reported in percentage points (pp); regret tables use a $100\times$ scale. Losses, layer energies, and error bounds retain their original scales.

Table 4: Synthetic study sizes. N_{fit} counts actual fitting contexts. Each study uses 20 seeds; 95% intervals for contrasts use paired seed means.

Study	Seed IDs	N_{fit}	Test/seed
Exact-distribution ablation	1200–1219	–	100 distributions
Continuous Top-3	2000–2019	5,000/50,000	10,000
Controlled posterior	1800–1819	1,000–50,000	10,000
DGP-A–C fixed MLP	1400–1419	2,000/8,000	1,200
Ridge/Brier	1300–1319	100–10,000	1,200
Early-stopped MLP	1400–1419	400/2,000/8,000	1,200
Inactive-target control	1400–1419	8,000	1,200

H.2 Exact-distribution information ablation

At $n = 6$, $m = K = 3$, independently draw unit vectors z_j by normalizing standard Gaussian vectors in V_j, set

$$h = \frac{0.600 z_1 + 0.600 z_2 + z_3}{\|0.600 z_1 + 0.600 z_2 + z_3\|_2}, \qquad p = \frac{\mathbf{1}}{20} + \frac{0.850}{20\|h\|_\infty} h,$$

and project this exact distribution before decoding. The same 100 distributions for each of 20 seeds (1200–1219) are evaluated under $q/3$, $q(q-1)/6$, and $\mathbf{1}[q \geq 2]$. Each projection retains V_0. Target-utility ties within 10^{-10} are broken using expected overlap from that same projection, followed by lexicographic order. The constant interface therefore selects the lexicographically first decision. The reported gap is the full-information optimal expected utility minus the selected decision's expected utility, both evaluated under p. Intervals are computed across the 20 paired seed means.

The active supports are $\{1\}$, $\{1,2\}$, and $\{1,3\}$; the largest score discrepancy between a full distribution and its loss-active projection is below 10^{-12}. For additive utility, retaining V_1 suffices. The sufficient two-layer interface changes from $V_1 \oplus V_2$ for pair reward to $V_1 \oplus V_3$ for threshold utility. Adding the other, inactive layer to V_1 changes neither decisions nor utility.

H.3 Continuous Top-3: sample size and learned information value

Shared learning protocol. The continuous Top-3 and controlled-posterior studies use $n = 10$, $m = 3$, $K = 5$, two-hidden-layer width-128 ReLU MLPs, and one observed outcome per fitting

Table 5: Exact-distribution ablation at $n = 6$, $m = K = 3$. Entries are $100\times$ mean utility gap [95% paired seed interval]. Dimension excludes the constant component, except full prediction, whose 20 ambient coordinates have 19 simplex degrees of freedom.

Interface	Dim.	Additive	Pair	Threshold
Constant	0	3.283 [3.207, 3.359]	3.512 [3.424, 3.600]	9.268 [8.991, 9.545]
V_1	5	0 [0, 0]	0.227 [0.211, 0.244]	2.584 [2.492, 2.676]
$V_1 \oplus V_2$	14	0 [0, 0]	0 [0, 0]	2.584 [2.492, 2.676]
$V_1 \oplus V_3$	10	0 [0, 0]	0.227 [0.211, 0.244]	0 [0, 0]
Full	20	0 [0, 0]	0 [0, 0]	0 [0, 0]

context. All interfaces share the data and training budget. Squared targets represent V_1, $V_1 \oplus V_2$, $V_1 \oplus V_3$, or the full outcome vector. Full prediction estimates a residual around the uniform distribution, without softmax or simplex clipping. Training uses AdamW with learning rate 0.002, weight decay 10^{-4}, and minibatches of 512. Independent validation and test sets contain 5,000 and 10,000 contexts per seed. Epoch 100 is the pre-specified primary endpoint within a fixed 300-epoch training window; epoch 300 and the earliest minimum-validation-loss checkpoint are supplementary endpoints. Exact decoding uses target utility, retained-interface expected overlap within tolerance 10^{-10}, and lexicographic order. Additive, pair, and threshold utilities are $q/3$, $q(q-1)/6$, and $\mathbf{1}[q \geq 2]$. Intervals are paired Student-t intervals across 20 teacher/data/model seeds. Supplementary comparisons receive Holm adjustment.

Continuous Top-3 outcomes. For seeds 2000–2019, sample $x \sim N(0, I_6)$ and generate

$$r = xA + 0.800\, zB^\top + 0.600\, \epsilon, \qquad z \sim N(0, I_3), \quad \epsilon \sim N(0, I_{10}).$$

The latent variables are mutually independent and independent of x. The seed-specific matrix A has independent $N(0, 1/6)$ entries; the rows of B are independent Gaussian vectors normalized to unit length. The three largest scores form the observed outcome. Fitting sizes $N_{\text{fit}} = 5{,}000$ and $50{,}000$ are nested subsets of each seed's pool. The outcome distribution is induced by correlated item scores. Evaluation uses held-out realized utility; N in the Top-3 figures denotes N_{fit}.

Information retained from a frozen predictor. Table 6 compares projections of the same full predictor. Retaining V_3 reduces threshold utility by 1.073 pp at $N_{\text{fit}} = 5{,}000$ and improves it by 0.231 pp at $N_{\text{fit}} = 50{,}000$. For threshold decoding, adding V_2 to V_1 changes neither scores nor decisions, and the active projection preserves full-prediction comparisons. For additive decoding all retained interfaces agree; for pair utility, retaining V_2 improves utility at both fitting sizes.

Table 6: Continuous Top-3 frozen-predictor contrasts at epoch 100. Entries are $100\times$ utility differences [95% paired-seed interval].

Utility	Retained-interface contrast	$N_{\text{fit}} = 5{,}000$	$N_{\text{fit}} = 50{,}000$
Additive	$V_1 - \text{full}$	0 [0, 0]	0 [0, 0]
Pair	$(V_1 \oplus V_2) - V_1$	0.375 [0.247, 0.502]	0.172 [0.110, 0.235]
Threshold	$(V_1 \oplus V_3) - V_1$	−1.073 [−1.318, −0.828]	0.231 [0.089, 0.373]

Independently trained interfaces. Independent active training exceeds independent V_1 by 0.479 pp at $N_{\text{fit}} = 5{,}000$ and 0.374 pp at $N_{\text{fit}} = 50{,}000$. Relative to independently trained $V_1 \oplus V_2$, the differences are -1.242 and $+0.219$ pp (Figure 5). These comparisons change both the fitted backbone and the estimated layer components, whereas projection holds the underlying predictor fixed.

Margins and prediction-defined strata. For a frozen prediction, let the V_1 margin be the difference between its best and second-best target-utility scores, and let the V_3 range be its largest minus smallest score contribution. A unique V_1 optimum is preserved when its margin exceeds that range (Appendix G). When the range reaches the margin, the added component can change the decision; its realized value is measured by the utility difference on changed decisions. For $D = \mathbf{1}\{S_{\text{active}} \neq S_{V_1}\}$,

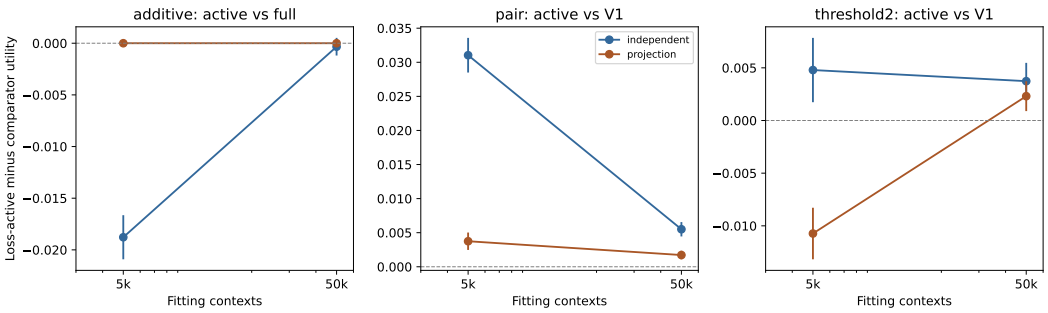


Figure 5: Continuous Top-3 projection and independent-training comparisons at epoch 100. Projection holds the full predictor fixed; independent training changes the supervised target. Intervals use 20 paired seeds.

the utility difference equals $\Pr(D = 1)\mathbb{E}[u(S_{\text{active}}, e) - u(S_{V_1}, e) \mid D = 1]$. Table 7 decomposes this difference. At $N_{\text{fit}} = 50{,}000$, active and V_1 disagree on 61.948% of contexts, but 89.967% of these disagreements have equal realized success. The pooled conditional gain is 0.373 pp, giving an overall gain of 0.231 pp.

Table 7: Threshold decision changes for the same frozen Top-3 predictor at epoch 100. Win/tie/loss compare active utility with V_1 among disagreements; a tie means equal realized success. Disagreement percentages and conditional gains pool source counts across 20 paired seeds; overall gains average seed contrasts.

N_{fit}	Disagree (%)	Win (%)	Tie (%)	Loss (%)	Pooled conditional gain (pp)	Overall (pp)
5,000	73.066	4.914	88.703	6.383	−1.469	−1.073
50,000	61.948	5.203	89.967	4.830	0.373	0.231

Quartiles are defined by validation predictions and frozen before test evaluation. At $N_{\text{fit}} = 50{,}000$, the highest predicted V_3-energy quartile has a gain of -0.686 pp (95% paired-seed interval $[-0.952, -0.420]$). The upper two range-to-margin quartiles instead have gains of 0.481 pp $[0.148, 0.813]$ and 0.442 pp $[0.153, 0.731]$. Figure 6 reports all four quartiles in each partition.

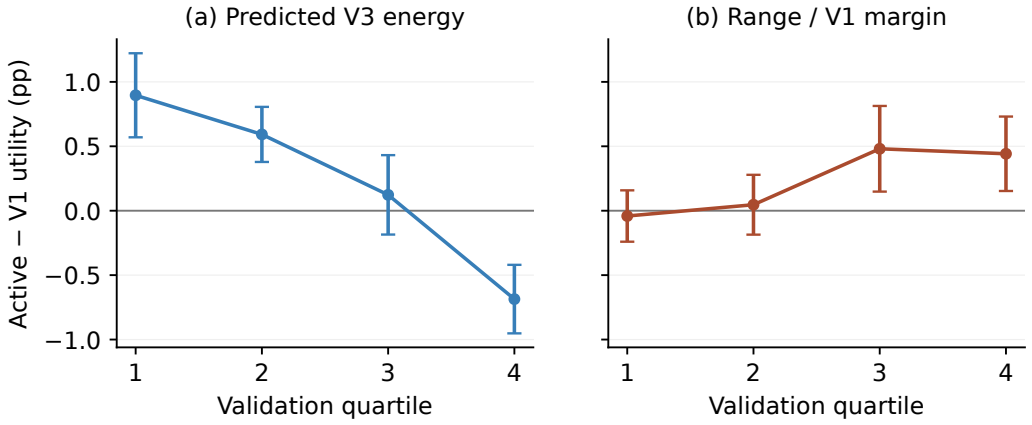


Figure 6: Continuous Top-3: threshold gains from retaining V_3 in the same frozen full predictor, at $n = 10$, $N_{\text{fit}} = 50{,}000$, and epoch 100. Energy and range-to-margin quartiles define separate partitions using validation predictions. Points show means and 95% paired-seed intervals across 20 seeds.

Supplementary training endpoints. At $N_{\text{fit}} = 50{,}000$, the frozen-predictor threshold active-minus-V_1 mean is 0.111 pp at epoch 300, with an interval covering zero. The validation-selected mean is 0.264 pp, with a positive nominal interval but without significance after the supplementary Holm adjustment. The epoch-100 sign change occurs between the two measured fitting sizes.

H.4 Controlled posterior: signal strength and estimation

Using the learning protocol in Appendix H.3, the teacher for seeds 1800–1819 draws $x \sim N(0, I_6)$. Let Q_j have orthonormal columns spanning V_j, and draw each seed-specific $A_j \in \mathbb{R}^{6\times \dim V_j}$ with independent $N(0, 1/6)$ entries. Define

$$b_j = \max_e \sum_r |Q_j[e, r]|, \qquad s_j(x) = \frac{Q_j \tanh(A_j^\top x)}{b_j},$$

and set

$$p(e \mid x) = \frac{1 + 0.200\, s_1(x, e) + 0.200\, s_2(x, e) + 0.500\, \gamma s_3(x, e)}{\binom{n}{3}}, \qquad \gamma \in \{0, 0.250, 1\}.$$

Each $s_j(x)$ lies in V_j and is bounded in absolute value by one. The probabilities are nonnegative and sum to one, and $\gamma = 0$ removes posterior V_3 exactly. The design crosses the three strengths with fitting sizes $1{,}000$, $5{,}000$, $20{,}000$, and $50{,}000$ at $n = 10$. Robustness cells use $n = 16$, $m = 3$, $K = 8$, $\gamma = 1$, and sizes $5{,}000$ and $50{,}000$. Here $N = N_{\text{fit}}$ in the figures, and the conditional distribution given observed x is known exactly.

Exact information value and estimation error. At $\gamma = 1$, the exact-posterior gain from retaining V_3 is 0.049 pp at $n = 10$ and 0.005 pp at $n = 16$, averaged over independent validation contexts.

For the direct active model at $n = 10$, write $p_3 = P_{V_3} p$ and $\widehat{p}_3 = P_{V_3}\widehat{p}$. The seed-averaged ratio $\mathbb{E}\|\widehat{p}_3 - p_3\|^2 / \mathbb{E}\|p_3\|^2$ falls from 1881.863 at $N_{\text{fit}} = 1{,}000$ to 43.578 at $N_{\text{fit}} = 50{,}000$. Direct active-minus-$V_1$ utility at the latter size is -0.138 pp. Between epochs 100 and 300, its training loss falls but validation loss rises and Bayes regret increases from 1.534 to 1.562 on a $100\times$ scale. Figure 7 shows the error-to-energy ratio, and Figure 8 gives the full comparison grid.

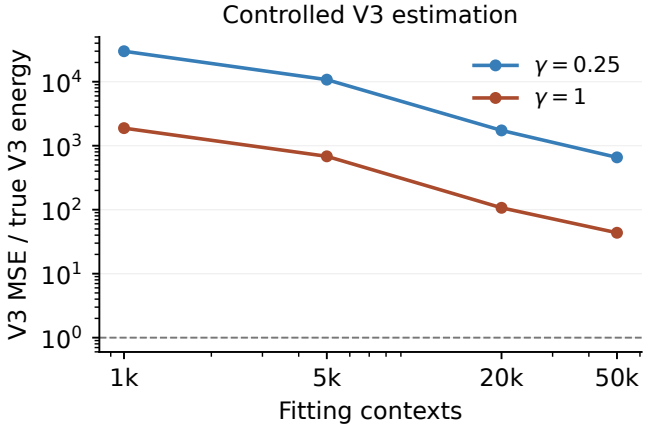


Figure 7: Controlled posterior: V_3 mean squared error divided by true layer energy for independently trained active predictors at $n = 10$ and epoch 100. Curves average 20 paired seeds; the dashed line marks equality of error and energy.

From layer error to decision scores. We apply (G.2) to each frozen interface. The threshold weights $(\lambda_1^2, \lambda_3^2)$ are $(640, 24)$ at $n = 10$ and $(90552, 1008)$ at $n = 16$. Layer errors sum squared orthonormal coordinates and average over test contexts, matching the coefficient normalization. The following error and regret values use the original utility scale. For $n = 10$, $\gamma = 1$, and $N_{\text{fit}} = 50{,}000$, the direct active model's weighted score error is 0.081, yielding an aggregate mean-regret upper bound

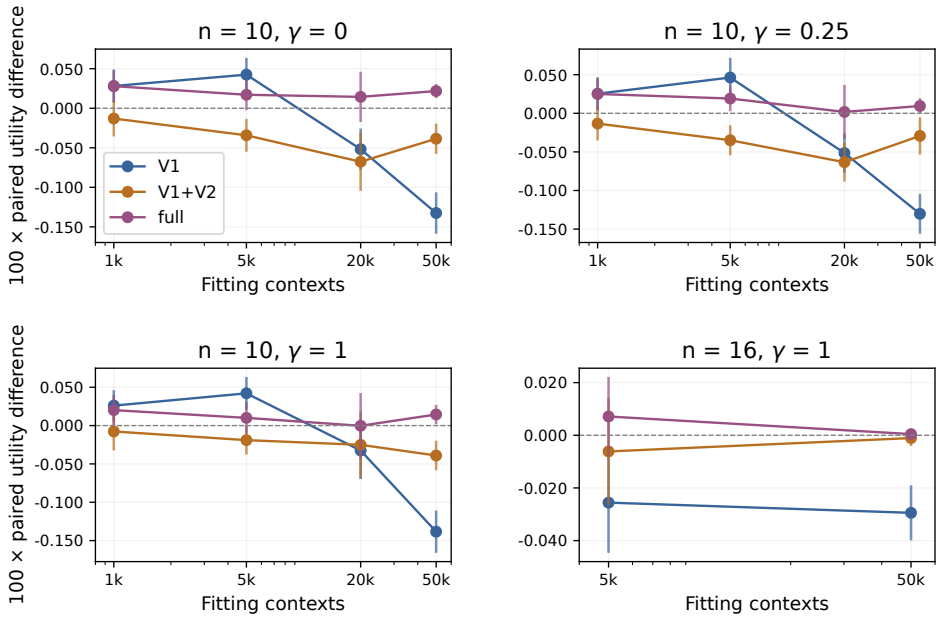


Figure 8: Controlled-posterior signal and sample-size comparisons at epoch 100. Positive utility differences favor independently trained active over the named comparator. Curves show paired-seed means and 95% intervals.

of 0.401; its attained mean regret is 0.015. For the V_1 model, omitted V_3 contributes 3.768×10^{-4} in weighted energy, whereas error on the retained V_1 contributes 0.527. Its bound is 1.027 and attained regret is 0.014. Estimation error dominates the omitted-layer energy in this setting.

H.5 Logit-generated outcomes and training dynamics

DGP-A–C. Independently sample $x \sim N(0, I_6)$ and a latent logit shock $\eta \sim N(0, I_{|\Omega_m|})$ per row. We use $m = 3$, $K = n/2$, even n, and threshold utility $u(q) = \mathbf{1}[q \geq 2]$ throughout the DGP-A–C studies. Let Q_j be an orthonormal basis for V_j. Draw $A_j \in \mathbb{R}^{6 \times \dim V_j}$ with independent $N(0, 0.8^2)$ entries and define

$$Z_j(x) = \tanh(x^\top A_j) Q_j^\top .$$

The learner's input is x; the reference distribution conditions on both x and its unobserved shock η:

$$p(e \mid x, \eta) = \mathrm{softmax}(\tau\{s_1 Z_1(x) + s_2 Z_2(x) + s_3 Z_3(x)\} + \sigma\eta)_e .$$

Labels are sampled from this vector, and the evaluator measures

$$\mathrm{Reg}_{\mathrm{ref}}(\widehat{S}; x, \eta) = \max_{|S|=K} \mathbb{E}_{p(\cdot|x,\eta)} u(|S \cap e|) - \mathbb{E}_{p(\cdot|x,\eta)} u(|\widehat{S} \cap e|).$$

The reported gap includes the value of the unobserved shock, in addition to regret relative to the Bayes decision based on $p(e \mid x) = \mathbb{E}_\eta p(e \mid x, \eta)$. Paired method comparisons use identical test rows and cancel the common oracle baseline. The three logit profiles are

	s_1	s_2	s_3	σ	τ	logit structure
A	0.900	0	0.450	0.050	1	smaller V_3 weight
B	0.350	0	1.350	0.050	1.100	larger V_3 weight
C	0.700	1.400	1	0.200	1	mixed-layer logits.

The weights control logit components; softmax mixes their contributions across posterior layers. The random-feature map is $\Phi(x) = (1, \sqrt{2/64}\cos(x^\top W + b))$, with standard Gaussian W and $b \sim \mathrm{Unif}([0, 2\pi]^{64})$, frozen per seed.

Training and decoding. The DGP-A–C campaign maps each context to 65 input features and trains a two-hidden-layer width-64 ReLU MLP with a linear output head. Training uses the fixed 80% fitting split, full-batch AdamW, learning rate 0.002, weight decay 10^{-4}, and 100 epochs. The primary $n = 10$ trajectory uses $N_{\text{fit}} = 8{,}000$ from a pool of $N_{\text{pool}} = 10{,}000$ and continues to epoch 300 with the same data and initialization. At each checkpoint, all solvers decode the same frozen interface estimate. The generating distribution evaluates each selected decision's reference regret. The ridge and early-stopped MLP studies below examine robustness across fitting sizes and training procedures.

For DGP-A–C, exact decoding enumerates lexicographically ordered decisions and selects the first maximizer. Greedy begins with the empty set and adds the first maximum-gain element; one-swap begins at that greedy decision and repeatedly accepts the first improvement larger than 10^{-12} until none remains.

Within-interface loss and regret trajectories. Losses are divided by each run's initial value, measuring relative training progress. Reference regret in the following trajectory summaries uses the original utility scale, as in Figure 4. In the primary DGP-B cell ($n = 10$, $N_{\text{fit}} = 8{,}000$), normalized V_1 loss falls from 0.568 to 0.513 between epochs 100 and 300, while its greedy regret rises by 0.005. Active-interface greedy regret falls from 0.082 to 0.078 and stabilizes around epochs 180–240. At epoch 300, active greedy regret is 0.016 lower than V_1 greedy regret (95% CI $[0.014, 0.018]$).

Interface gains under alternative solvers. In the same cell, replacing V_1 by $V_1 \oplus V_3$ at epoch 100 reduces regret by 0.007 under greedy decoding (95% CI $[0.005, 0.008]$) and by 0.010 under exact decoding (95% CI $[0.008, 0.011]$). Active-interface greedy also improves on V_1-exact by 0.005 (95% CI $[0.003, 0.007]$).

H.6 Interface and solver comparisons

Table 8 reports epoch-100 reference regret over paired seeds 1400–1419 at $n = 6$, with $N_{\text{pool}} = 2{,}500$ and $N_{\text{fit}} = 2{,}000$. Table 9 extends the cross-solver comparison to $n = 10, 12$.

Table 8: Synthetic MLP reference regret ($100\times$) at epoch 100, $n = 6$, $m = K = 3$, and $N_{\text{fit}} = 2{,}000$. Lower is better; the final column is the paired greedy-minus-exact regret difference.

DGP-B: larger V_3 logit weight				
Interface	greedy	one-swap	exact	greedy − exact
V_1	10.949	10.947	10.947	0.002
$V_1 \oplus V_2$	11.058	10.809	10.809	0.249
$V_1 \oplus V_3$	8.782	8.156	8.063	0.719
$V_1 \oplus V_2 \oplus V_3$	8.721	7.795	7.700	1.020
full	8.661	7.706	7.594	1.067

DGP-C: mixed-layer logits				
Interface	greedy	one-swap	exact	greedy − exact
V_1	12.265	11.920	11.920	0.346
$V_1 \oplus V_2$	12.274	11.170	11.170	1.103
$V_1 \oplus V_3$	10.956	10.359	10.302	0.654
$V_1 \oplus V_2 \oplus V_3$	11.411	9.902	9.753	1.658
full	11.186	9.755	9.525	1.660

On the $100\times$ reference-regret scale, DGP-B's paired greedy-to-exact gains are 0.002 with 95% CI $[-0.133, 0.138]$ under V_1, 0.719 $[0.580, 0.858]$ under $V_1 \oplus V_3$, and 1.067 $[0.890, 1.244]$ under the full interface. The paired difference V_1-exact minus $(V_1 \oplus V_3)$-greedy is 2.165 with 95% CI $[1.945, 2.385]$.

Greedy evaluates partial sets during construction, whereas the active support characterizes comparisons between final K-sets. For threshold utility, all singleton scores are zero, so the first greedy choice follows the tie rule even for V_1. One-swap exchanges elements at fixed cardinality K; exact decoding enumerates all K-sets. Each solver uses only the named reconstructed interface.

The same cross-solver contrast is positive in all six configurations in Table 9:

Table 9: Cross-solver reference-regret contrasts ($100\times$) at epoch 100. Positive values favor the active interface with greedy decoding over V_1 with exact decoding.

n	N_{fit}	DGP	V_1-exact − active-greedy	95% CI
6	2,000	B	2.165	[1.945,2.385]
6	2,000	C	0.964	[0.621,1.306]
10	2,000	B	0.547	[0.383,0.712]
10	8,000	B	0.498	[0.325,0.672]
12	2,000	B	0.232	[0.061,0.403]
12	8,000	B	0.225	[0.058,0.392]

H.7 Ridge/Brier robustness across sample sizes

At $n = 6$, $m = K = 3$, ridge regression fits each target on the same random features using $\widehat{B} = (\Phi^\top\Phi + 0.010I)^{-1}\Phi^\top Y$. Table 10 reports threshold reference regret under exact decoding. The common ridge fitting operator commutes with output projection. Consequently, V_1 and $V_1 \oplus V_2$ give identical threshold decisions, as do active, prefix, and full prediction. The table groups these equivalent interfaces into two columns per DGP, isolating the decision value of the estimated V_3 component under the same fitting operator.

Table 10: Ridge/Brier reference regret ($100\times$) at $n = 6$, $m = K = 3$. The V_1 column also represents $V_1 \oplus V_2$; the active column represents $V_1 \oplus V_3$, prefix, and full. Lower is better.

	DGP-A		DGP-B		DGP-C	
N_{fit}	V_1	Active	V_1	Active	V_1	Active
100	11.701	11.779	11.296	10.548	13.642	13.357
250	10.468	10.636	10.983	9.567	12.790	12.127
500	9.777	9.944	10.858	8.782	12.134	11.297
1,000	8.834	8.953	10.642	7.951	11.266	10.044
2,500	7.391	7.519	10.253	6.652	9.916	8.267
5,000	6.257	6.314	9.888	5.656	8.982	7.111
10,000	5.347	5.280	9.495	4.844	8.214	6.094

Under DGP-A, retaining estimated V_3 increases reference regret through $N_{\text{fit}} = 5{,}000$ and reduces it at 10,000. Under DGP-B and DGP-C it reduces regret throughout the grid. The same sufficient interface thus has different estimation tradeoffs across the three generating distributions.

H.8 Early-stopped MLPs and inactive auxiliary targets

At $n = 6$, $m = K = 3$, the common MLP architecture has two ReLU hidden layers of width 64 and is trained by AdamW with learning rate 0.002, weight decay 10^{-4}, 20% validation, patience 25, and at most 220 epochs. Training is full-batch; the best validation checkpoint is updated only after an improvement larger than 10^{-7}. Interface-specific training seeds add offsets $101, 211, 307, 401, 503$ for $V_1, V_1 + V_2, V_1 + V_3$, prefix, and full, respectively. Each resulting seed controls both the training/validation split and model initialization. Interfaces share the source pool and test set and are fitted separately. The pool sizes $N_{\text{pool}} = 500, 2{,}500, 10{,}000$ yield $N_{\text{fit}} = 400, 2{,}000, 8{,}000$ in Table 11.

In DGP-C at $N_{\text{fit}} = 8{,}000$, prefix training has lower regret than direct active training (Table 11). A separately fitted ablation uses the same DGP and early-stopping procedure, with independent data and initialization seeds. Reference regret ($100\times$) after zeroing V_2 within each frozen predictor is:

trained head	all coordinates	zero V_2	difference
$V_1 + V_2$	8.737	8.737	0
prefix	6.501	6.501	0.

Table 11: Early-stopped MLP reference regret ($100\times$) at $n = 6$, $m = K = 3$. All interfaces use the same backbone architecture and validation rule; their parameters are fitted separately.

DGP	N_{fit}	V_1	$V_1 + V_2$	$V_1 + V_3$	prefix	full
A	400	11.283	11.170	11.240	11.468	11.571
A	2,000	8.899	9.520	9.349	10.089	10.015
A	8,000	6.463	7.450	6.771	7.524	7.640
B	400	11.382	11.282	10.252	10.487	10.583
B	2,000	10.949	10.796	7.906	8.486	8.807
B	8,000	10.191	10.269	5.501	5.940	5.839
C	400	12.516	12.498	12.120	12.252	12.421
C	2,000	10.843	10.988	9.670	9.770	9.749
C	8,000	9.055	8.675	7.027	6.600	6.594

Removing V_2 preserves each frozen predictor's decisions. The independent-head comparison measures the outcome of separate fitting procedures; the frozen-head ablation isolates the inactive layer's zero contribution to threshold decoding.

I WikiSeeAlso protocols, mechanism comparisons, and diagnostics

I.1 Dataset, candidate construction, and independent samples

The WikiSeeAlso task uses $n = 6$, $m = 3$, and $K = 3$. Johnson dimensions are $\dim V_1 = 5$, $\dim V_2 = 9$, and $\dim V_3 = 5$. Additive utility $u_{\text{add}}(q) = q$ activates only V_1; threshold utility $u_{\text{thr}}(q) = \mathbf{1}[q \geq 2]$ activates $V_1 \oplus V_3$. Every frozen prediction is evaluated under both utilities using exact enumeration. Evaluation reports realized threshold success and mean overlap on held-out articles.

LF-WikiSeeAlso-320K supplies article features and related-article labels.[2] Features use 32-dimensional SVD article embeddings and label prototypes fitted on 623,774 training articles outside the validation split. A prototype is the mean embedding of articles bearing that label. Each candidate has 97 input features: the article embedding, the unit-normalized prototype, their elementwise product, and their cosine similarity. For a training article, its own contribution is removed from label support and prototypes. The counts in Table 12 describe interface fitting.

W3 is a candidate-conditioned task. For each article, we sample three true labels without replacement from those with effective training support at least 10. Negative pools cross quarter-octave support bins with within-bin prototype-norm deciles, computed from training labels with support at least 10. Articles are retained if every selected positive's cell contains at least seven nontrue labels. For each positive, we take up to eight negatives in its cell nearest in support and norm, using

$$d(\ell, \ell_+) = \left(\frac{\log c_\ell - \log c_{\ell_+}}{\log(2^{1/4})} \right)^2 + \left(\frac{\nu_\ell - \nu_{\ell_+}}{0.050} \right)^2,$$

where c_ℓ is support and ν_ℓ is the mean prototype's norm. We select the negative whose article–prototype cosine similarity is closest to the positive's. All true labels and previously selected negatives are excluded. Final candidate order is randomized and frozen before interface comparisons. Construction seeds 20260921–20260923 define three versions of this task.

Table 12: Independent source articles in the three Wiki constructions. Each source has one candidate instance within a construction.

Construction	Train	Validation	Test
Wiki 20260921	28,558	3,132	8,938
Wiki 20260922	28,523	3,132	8,956
Wiki 20260923	28,574	3,106	8,959

[2] Extreme Classification Repository.

The Wiki constructions share 8,726 test articles; their union has 9,217 articles, and 4,613 common articles have the same unordered candidate set in all three constructions. Article-blind, validation-selected metadata baselines improve on uniform random by 3.334, 3.216, and 3.086 pp, respectively; the pre-specified shortcut criterion is 5 pp. The baselines use label support, prototype norm, and candidate position. Their results and the permutation diagnostics are provided in the benchmark records.

I.2 Matched objectives and frozen-information comparisons

The matched-objective study uses five objectives, three frozen constructions and ten model seeds (20–29), for 150 fits. Every model is a single-head Set MLP32 with 6,305 parameters and 20 raw outcome scores. A shared two-layer item encoder feeds a shared subset scorer using the subset mean and invariant set context. Within each construction and model seed, all objectives start from identical parameters and use identical minibatch orders. Feature standardization is fitted on training data.

Objectives and reconstructed predictions. Write $r = r_\theta(x) \in \mathbb{R}^{20}$, $t = \mathbf{e}_e$, $b = \mathbf{1}/20$, $P_j = Q_j Q_j^\top$, and $P_{13} = P_1 + P_3$. Here Q_j is an orthonormal Johnson basis. For threshold utility, $\lambda_1^2 = 16$, $\lambda_2^2 = 0$, $\lambda_3^2 = 4$, and $\dim V_1 = \dim V_3 = 5$. Define $c = b + r - \bar{r}\mathbf{1}$ and $s = \mathrm{softmax}(r)$. The single-example training objectives are

$$\begin{aligned}
\psi_{\mathrm{A}}(r;e) &= 2 \sum_{j\in\{1,3\}} \|Q_j^\top (r-t)\|^2, && \text{unweighted active,}\\
\psi_{\mathrm{B}}(r;e) &= \tfrac{1}{5}\left(16\|Q_1^\top (r-t)\|^2 + 4\|Q_3^\top (r-t)\|^2\right), && \text{weighted active,}\\
\psi_{\mathrm{C}}(r;e) &= \tfrac{20}{19}\|c-t\|^2, && \text{centered raw Brier,}\\
\psi_{\mathrm{D}}(r;e) &= \tfrac{20}{19}\|s-t\|^2, && \text{softmax Brier,}\\
\psi_{\mathrm{E}}(r;e) &= -\log s_e / \log 20, && \text{full CE.}
\end{aligned}$$

All objectives have loss one at the uniform prediction. The decoder receives $b + P_{13}r$ for A/B, c for C, and s for D/E. A/B discard the unsupervised V_2 component of their raw outputs. C is a signed unit-mass reconstruction; D/E are probability vectors. Frozen ablations retain the constant component together with V_1, $V_1 \oplus V_2$, $V_1 \oplus V_3$, or the entire reconstructed vector. Each V_1 column evaluates a projection of the row's frozen predictor.

Training, endpoints, and decoding. All fits use AdamW, learning rate 0.002, weight decay 10^{-4}, batch size 512, and 300 epochs on an H200 in deterministic FP32. The primary endpoint is epoch 300. For every objective, the secondary endpoint is the earliest minimum, over all training epochs, of the same validation score error

$$R_{\mathrm{score}}(\widetilde{p}, t) = \tfrac{1}{5}\left(16\|Q_1^\top (\widetilde{p}-t)\|^2 + 4\|Q_3^\top (\widetilde{p}-t)\|^2\right).$$

Test evaluation follows completion of all 150 fits. Exact decoding enumerates all 20 decisions. Within tolerance 10^{-8} of the maximum predicted target utility, it maximizes predicted overlap; remaining ties within 10^{-8} are resolved by lexicographic order of sorted candidate IDs. The same rule applies to every projection and objective. Both utilities and all projections are evaluated on each frozen endpoint.

Paired inference. Results use the common 8,726 source articles. We average the three constructions within each source/model-seed pair, then independently resample articles and the ten model seeds (2,000 bootstrap draws, seed 20260912). Intervals condition on these fixed constructions. The primary contrast is B minus A at epoch 300. Two-sided p values compare its magnitude with the centered bootstrap distribution, using the add-one finite-draw correction. The secondary Holm family consists of B minus A at validation selection and C minus A, D minus C, E minus D, and E minus A at both endpoints, totaling nine comparisons. Frozen-projection intervals describe retained-information effects separately.

Frozen-predictor information comparisons. Table 13 reports the effect of retaining V_3 within each frozen predictor. Full CE has positive active-minus-V_1 seed means at both endpoints for all ten seeds. At epoch 300, its projections disagree in 14.662% of source/construction/seed cases. Among

disagreements, active wins in 8.465%, loses in 4.843%, and ties in realized utility in 86.692%, yielding a conditional gain of 3.621 pp and overall gain of 0.531 pp. The three raw-output objectives instead have negative projection gains in all ten seed means at both endpoints.

Table 13: Matched Wiki objectives: active minus V_1 projection of the same frozen predictor, in threshold-success pp [95% source/model-seed interval].

Objective	Fixed epoch 300	Validation-selected
Unweighted active	-1.528 $[-1.872, -1.220]$	-1.007 $[-1.179, -0.840]$
Weighted active	-1.647 $[-1.888, -1.407]$	-1.195 $[-1.362, -1.042]$
Centered raw Brier	-1.998 $[-2.504, -1.508]$	-0.819 $[-0.983, -0.669]$
Softmax Brier	0.316 $[0.216, 0.416]$	0.016 $[-0.007, 0.040]$
Full CE	0.531 $[0.438, 0.628]$	0.262 $[0.189, 0.340]$

Training-objective comparisons. Tables 14–15 compare separately fitted predictors. The primary weighting effect is 0.096 pp, with interval $[-0.219, 0.426]$ and $p = 0.566$. Five of ten seed means are positive. Its construction-specific effects on the common cohort are 0.359, 0.156, and -0.227 pp. At validation selection, the effect is -0.001 pp, interval $[-0.203, 0.203]$, with $p = 0.993$. Weighting lowers score error at both endpoints, while full CE improves threshold success over unweighted active training by 2.010 pp at epoch 300 and 1.465 pp at validation selection. Score error is computed against observed one-hot targets and includes outcome noise.

Table 14: Matched Wiki objectives: active-interface threshold success (%) and common score error on the same 8,726 test articles. Selected checkpoints minimize validation score error; all reported values are test metrics.

	Threshold success		Score error	
Objective	Fixed 300	Selected	Fixed 300	Selected
Unweighted active	94.417	95.223	0.455	0.430
Weighted active	94.513	95.222	0.439	0.423
Centered raw Brier	93.815	95.313	0.454	0.429
Softmax Brier	95.762	95.832	0.507	0.437
Full CE	96.427	96.688	0.446	0.409

Table 15: Secondary matched-training contrasts. Threshold-success differences are in pp; p values use Holm correction across these nine comparisons. A: unweighted active; B: weighted active; C: centered raw Brier; D: softmax Brier; E: full CE.

Contrast	Endpoint	Difference	95% CI	Holm p
C − A	Fixed	-0.602	$[-1.276, 0.106]$	0.268
D − C	Fixed	1.947	$[1.409, 2.488]$	0.004
E − D	Fixed	0.665	$[0.501, 0.833]$	0.004
E − A	Fixed	2.010	$[1.710, 2.342]$	0.004
B − A	Selected	-0.001	$[-0.203, 0.203]$	0.993
C − A	Selected	0.090	$[-0.144, 0.319]$	0.939
D − C	Selected	0.519	$[0.306, 0.748]$	0.004
E − D	Selected	0.856	$[0.674, 1.040]$	0.004
E − A	Selected	1.465	$[1.220, 1.717]$	0.004

Invariance and permutation checks. For all 150 fits at both endpoints, full/active and $V_1/V_1 \oplus V_2$ threshold decisions coincide, and all projections yield identical additive decisions. Maximum score discrepancies are 1.499×10^{-14} for full/active threshold and 4.463×10^{-14} for additive comparisons. On 32 contexts per fixed-endpoint model, 20 candidate permutations give zero report disagreement and maximum prediction equivariance error 2.639×10^{-5}.

I.2.1 VALIDATION-SELECTED LOSS-SPECIFIC WORKFLOW

Table 3 evaluates the complete workflow with ten new model seeds, 60–69, on the same three frozen constructions and common test articles. The architecture, five objectives, initialization matching, minibatch ordering, and 300-epoch budget are those above, giving 150 additional fits. For each objective, the checkpoint pool contains epoch 300 and the earliest minimum validation threshold score error. All selection criteria use this same two-checkpoint pool per objective.

Loss-specific validation and coordinate export. For additive $q/3$, pair $q(q-1)/6$, and threshold $\mathbf{1}[q \geq 2]$, the coefficient formula gives squared layer weights $(4, 0, 0)$, $(4, 4/9, 0)$, and $(16, 0, 4)$. For each construction and model seed, we select from the pool by minimizing

$$\widehat{R}_L(f) = \sum_{j \in J_L} \lambda_j^2 \frac{1}{N_{\text{val}}} \sum_i \left\| Q_j^\top \left(\widehat{p}_f(x_i) - \mathbf{e}_{e_i} \right) \right\|^2.$$

Ties use the earliest epoch, then a fixed objective ordering. The selected predictor is frozen before test evaluation. The decoder receives only $(Q_j^\top \widehat{p}_f(x))_{j \in J_L}$ and uses precomputed decision operators, with the exact-decoding and tie-breaking rule above. All losses select CE in all 30 construction–seed groups. Full-Brier validation and CE validation within the two archived CE checkpoints select the same models.

Paired information and estimation effects. The primary contrast is threshold active minus V_1 of the same selected predictor. We average constructions within each source/model-seed pair and jointly bootstrap articles and model seeds with 2,000 draws. Table 16 reports the contrasts. New-seed threshold gains are positive in all ten seed means; replaying seeds 20–29 gives 0.262 pp with interval $[0.188, 0.344]$. Both cohorts reuse the same articles. The direct-active comparator minimizes unweighted threshold-active squared loss and uses validation checkpoint selection. Its comparison with the workflow changes the estimator as well as its predictions.

Table 16: Validation-selected Wiki workflow: paired test-utility differences in pp. Intervals resample source articles and model seeds, conditional on the frozen constructions and validation selections.

Utility	Contrast	Difference	95% CI
Threshold	Active $-$ V_1	0.276	$[0.199, 0.356]$
Threshold	Active $-$ $(V_1 \oplus V_2)$	0.276	$[0.199, 0.356]$
Threshold	Active $-$ full	0	$[0, 0]$
Threshold	Workflow $-$ direct active	1.292	$[1.076, 1.518]$
Threshold	Workflow $-$ full-Brier selection	0	$[0, 0]$
Pair	Active $-$ V_1	0.017	$[-0.009, 0.044]$
Additive	Active $-$ V_1	0	$[0, 0]$

Layer estimation and decision relevance. On observed one-hot layer targets, test MSEs for V_1, V_2, V_3 are 0.091, 0.235, and 0.147; their R^2 values relative to the training-mean predictor are 0.634, 0.477, and 0.414. Observed target variations are 0.250, 0.450, and 0.250. These diagnostics measure prediction of observed targets, not the energy of the unobserved conditional posterior. We define ρ as the predicted V_3 threshold-score range divided by the best–second-best V_1 score margin, floored at 10^{-10}. Quartile cutoffs for ρ and $\|P_3 \widehat{p}\|$ use validation predictions only. Conditional-effect intervals resample the gain numerator and membership denominator together. The strata in Table 17 are descriptive prediction-defined comparisons. On disagreement cases, active wins, ties, and loses in 7.717%, 87.502%, and 4.781% of cases. In the highest predicted-norm bin, mean probability outside the modal outcome is below 10^{-4} and the mean V_1 margin is 0.667: high norm accompanies an already stable decision.

Interface size and numerical identities. The exported additive, pair, and threshold interfaces contain 5, 14, and 10 coordinates. The full output stores 20 probabilities with 19 independent degrees of freedom. All networks retain 6,305 parameters and 20 raw outputs; the size reduction concerns the prediction-to-decoder interface. Active and full decisions coincide for all losses; threshold V_1 and $V_1 \oplus V_2$ decisions coincide, and additive decisions agree across all interfaces. The maximum active/full expected-score discrepancy is 1.488×10^{-14}.

Table 17: Threshold decision value in validation-defined prediction strata. Membership and disagreement are percentages; gains and intervals are in pp.

Group	Membership	Disagreement	Gain	95% CI
All cases	100	9.396	0.276	$[0.199, 0.356]$
Active/V_1 disagreement	9.396	100	2.935	$[2.136, 3.757]$
Highest ρ bin	33.545	24.931	0.763	$[0.541, 0.989]$
Highest predicted-V_3-norm bin	15.335	0.010	0	$[0, 0]$

I.3 Alternative architectures and training endpoints

This study varies architecture and training schedule for full-outcome softmax prediction. The set backbone applies a shared two-layer ReLU item encoder, concatenates each outcome subset's mean embedding with the invariant set mean, and applies a shared scalar scorer. The interaction backbone additionally uses symmetric pair messages and subset products. Candidate permutations therefore permute outcome scores consistently.

Set MLP32 uses constant-rate training for 300 epochs; Interaction128 uses cosine-rate training for 600 epochs. Both minimize full-outcome cross-entropy and use five paired seeds, 20–24, on each of the three constructions (30 fits). We evaluate the fixed endpoint and the checkpoint selected by validation cross-entropy. All validation-selected Interaction128 checkpoints are at epoch 5.

Exact decoding enumerates the 20 decisions and maximizes predicted target utility. Decisions within 10^{-8} of the maximum are ordered lexicographically by their sorted candidate IDs. Unlike Appendix I.2, this study uses validation cross-entropy for checkpoint selection and omits overlap-based tie-breaking. The decoder also applies to Appendix I.4; Appendix I.5 examines its tie-rule sensitivity.

Paired intervals use the common 8,726 test sources. We first average the three construction-specific differences within each article/model-seed pair, then independently bootstrap source articles and the five paired model seeds (2,000 draws). Inference conditions on the three fixed constructions. Table 18 reports active-minus-V_1 threshold success. The Set MLP32 validation projection gain has a positive joint bootstrap interval; its paired five-seed t-test gives $p = 0.069$.

Table 18: Wiki frozen-predictor projection: threshold success of $V_1 \oplus V_3$ minus V_1 (pp). Intervals resample common source articles and five paired model seeds after averaging the three fixed constructions. Validation checkpoints minimize validation cross-entropy; fixed endpoints are epoch 300 for Set MLP32 and 600 for Interaction128.

Predictor	Endpoint	Difference	95% CI
Set MLP32	Validation	$+0.124$	$[0.024, 0.241]$
Interaction128	Validation	$+0.286$	$[0.186, 0.390]$
Set MLP32	Fixed 300	$+0.413$	$[0.270, 0.551]$
Interaction128	Fixed 600	-5.955	$[-6.334, -5.556]$

I.4 Encoder refitting and auxiliary-target training

The training-mechanism study uses the width-32 set backbone and a 300-epoch budget. Let $r_\theta(x) \in \mathbb{R}^{20}$ denote raw outcome scores and let $Q_J \in \mathbb{R}^{20\times 10}$ contain an orthonormal basis of $V_1 \oplus V_3$. Full prediction and direct active training expose

$$\widehat{p}_\theta = \mathrm{softmax}(r_\theta), \qquad z_{\mathrm{full}} = Q_J^\top \widehat{p}_\theta, \qquad z_{\mathrm{direct}} = Q_J^\top r_\theta.$$

Full CE trains against the observed outcome. Direct active training minimizes $\|z_{\mathrm{direct}} - Q_J^\top \mathbf{e}_e\|^2/10$, where $\mathbf{e}_e$ is its one-hot vector. Both routes pass the reconstructed active vector $Q_J z$ to the decoder.

The study adds four arms per construction and seed (60 fits): direct active squared loss; a freshly initialized active head fitted on a frozen, validation-selected full-CE encoder; an active head with a zero-weight V_2 auxiliary head; and the matched auxiliary arm with weight one. The latter two arms

share active-head initialization, main-loss weight, and batch order. The zero-weight auxiliary arm and direct active training give identical decisions at both endpoints under both utilities.

All four arms use AdamW, batch size 512, learning rate 0.002, and weight decay 10^{-4}. Their validation endpoint minimizes active projected MSE over the pre-specified checkpoints, with earlier epochs breaking ties; the full-CE comparator minimizes validation cross-entropy. Frozen refitting starts from that comparator's validation-selected encoder. Decoding and paired inference follow Appendix I.3, with bootstrap seed 20260907. Table 19 reports the training contrasts, using the full-CE predictors from that study. Frozen refitting includes this pretraining followed by 300 epochs of head fitting. At the validation endpoint, refitting and full projection share the same encoder; the fixed full predictor instead uses its epoch-300 encoder.

Table 19: Wiki training mechanisms: paired threshold-success differences (pp, 95% source/ model-seed intervals). Validation checkpoints minimize cross-entropy for full prediction and active MSE otherwise.

Contrast	Fixed epoch 300	Validation endpoint
Full projection − direct	$2.034\ [1.518, 2.540]$	$1.349\ [1.008, 1.729]$
Frozen/refit − direct	$0.584\ [-0.072, 1.171]$	$-0.447\ [-0.978, 0.073]$
Frozen/refit − full projection	$-1.451\ [-1.788, -1.082]$	$-1.795\ [-2.178, -1.506]$
V_2 objective on − off	$0.251\ [-0.500, 0.964]$	$0.023\ [-0.364, 0.470]$
Null auxiliary − direct	$0\ [0, 0]$	$0\ [0, 0]$

Full CE followed by projection outperforms direct active training at both endpoints. Refitting an active head does not retain the original full predictor's advantage: even with the same validation-selected encoder, it reduces success by 1.795 pp. The V_2 auxiliary-target intervals include zero at both endpoints. Full projection minus direct active also improves mean additive overlap by 0.012 (CI $[0.004, 0.019]$) at the fixed endpoint and 0.011 ($[0.004, 0.018]$) at validation-selected endpoints. These gains compare separately fitted predictors; within each frozen predictor, all retained interfaces give the same additive decisions.

I.5 Training dynamics and decision stability

Training trajectories. Figure 9 compares direct active training, frozen refitting, and V_2 auxiliary training at 17 pre-specified checkpoints. Test curves use the common articles; validation MSE uses each construction's full validation cohort.

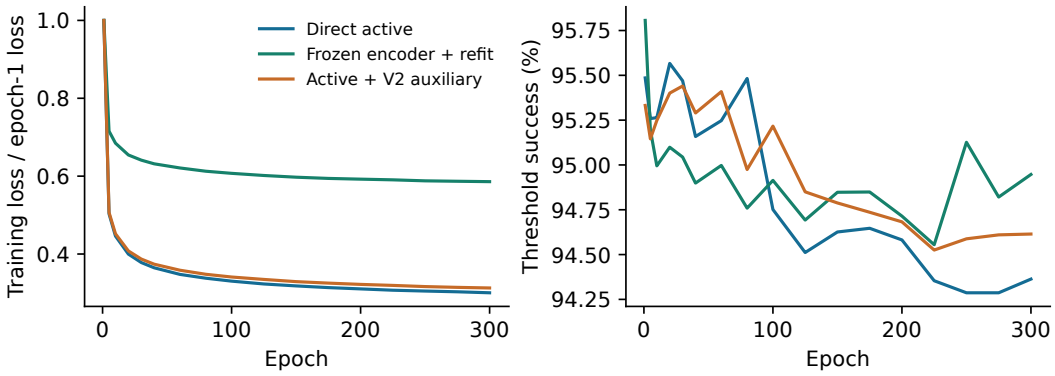


Figure 9: Wiki training dynamics across three fixed constructions and five model seeds, evaluated on $8{,}726$ common test articles. Training losses are normalized within each run to epoch 1, then averaged over constructions and seeds.

Between epochs 20 and 300, direct active training lowers training loss while success falls from 95.567% to 94.363% and validation active MSE rises from 17.668 to 18.127 (all MSEs here are multiplied by 1,000). With V_2 auxiliary training, validation MSE rises from 17.652 to 18.186 and

success falls from 95.401% to 94.615%. Frozen refitting lowers validation MSE from 22.722 to 21.104 and test MSE from 28.735 to 27.663, while success falls from 95.099% to 94.947%.

Saturation and near-tied decisions. At Interaction128's epoch-600 endpoint, 80.429% of predictions put probability greater than $1 - 10^{-6}$ on a single outcome. At the threshold-decoder tolerance 10^{-8}, 99.078% have multiple maximizing decisions, compared with 3.812% at the validation endpoint. On each construction's full test cohort, the average active-minus-V_1 contrast is -5.895 pp under lexicographic tie-breaking; Table 18 reports -5.955 pp on the common 8,726 articles. In the full-cohort contrast, tied cases contribute -5.925 pp and the remainder 0.030 pp. A post-hoc sensitivity analysis re-decodes the same frozen predictions with predicted-overlap tie-breaking within the threshold tolerance. This changes the contrast to 0.184 pp, with at most 10^{-8} reduction in predicted threshold utility. The contrast localizes the decision instability to saturated, nearly tied predictions.

Validation-selected shrinkage. The shrinkage diagnostic decodes $P_1\widehat{p} + \alpha P_3\widehat{p}$ with $\alpha \in \{0, 0.250, 0.500, 0.750, 1\}$ selected solely by validation threshold success (smaller α breaks ties). Table 20 reports the selected rule's gains over V_1. Shrinkage brings the fixed-endpoint Interaction128 contrast close to zero. In the other three settings, its gains are smaller than those from retaining the full active component (Table 18).

Table 20: Validation-selected shrinkage: threshold-success gains over V_1 (pp, 95% intervals).

Predictor	Fixed endpoint	Validation endpoint
Set MLP32	$0.333\ [0.224, 0.461]$	$0.098\ [0.012, 0.205]$
Interaction128	$0.005\ [-0.002, 0.015]$	$0.192\ [0.117, 0.274]$

I.6 Supplementary Wiki cohort: layer predictability and training size

Cohort and cross-fitting. W1 retains the native exact-three-label slice, whereas W3 samples three supported true labels. W1 matches negatives in quarter-octave support bins and prototype-norm strata, with minimum support 10 selected from $\{2, 5, 10\}$ using train/validation counts. It contains 7,929 training, 854 validation, and 707 test articles. Both tasks exclude all true labels from negatives and randomize candidate order.

Three-fold source-level cross-fitting uses five model seeds and width-32 full-Brier predictors (AdamW, batch size 512, learning rate 0.002, weight decay 10^{-4}). Each predictor fits on two folds (5,286 sources) and predicts the held-out fold. Supervised prototypes exclude held-out sources; unsupervised SVD and candidates remain fixed. Endpoints are epoch 60 and the minimum-surrogate-loss checkpoint on the original validation cohort. Observed one-hot V_3 target variation is 0.250; out-of-fold V_3 prediction R^2 is 0.676 at epoch 60 and 0.673 at validation selection.

Training-size comparison. The W1 sweep varies interface-fitting articles for independently trained squared-loss interfaces at epoch 60, fixing the architecture, budget, candidates, and features fitted on the larger training population. Table 21 reports all four sizes: $N_{\text{fit}} = 227$ uses ten source-level subsample draws; other sizes use one draw, with five paired model seeds throughout. Active exceeds $V_1 \oplus V_2$ in all ten draw means at $N_{\text{fit}} = 227$, but its contrasts with both interfaces are negative at the larger fitting sizes. This sweep shows no crossover toward an active advantage as the fitting sample grows.

Table 21: W1 threshold-success differences at epoch 60 (pp, 95% source-bootstrap intervals). Paired effects are averaged over seeds and draws before resampling the 707 test articles.

N_{fit}	Draws	Active $- V_1$	Active $- (V_1 \oplus V_2)$
227	10	$0.057\ [-0.074, 0.215]$	$0.685\ [0.402, 0.976]$
1,000	1	$-0.849\ [-1.584, -0.141]$	$-1.216\ [-1.980, -0.566]$
5,000	1	$-1.047\ [-1.669, -0.481]$	$-1.414\ [-2.065, -0.849]$
7,929	1	$-0.990\ [-1.641, -0.368]$	$-0.736\ [-1.358, -0.113]$