THE COMPARISON GRAPH DETERMINES WHAT RLHF CAN LEARN: SPECTRAL IDENTIFIABILITY, ERROR BOUNDS, AND A PRE-TRAINING DIAGNOSTIC FOR PREFERENCE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models for RLHF are trained on pairwise human comparisons, and the field treats that data as a quantity to be increased. We show the binding constraint is its structure, measurable before training. A preference dataset induces a *comparison graph* on prompt–response pairs, so reward learning under Bradley–Terry is estimation on that graph; lifting it through the reward model's feature map yields one matrix, computable in a single pass, whose spectrum settles which reward directions the data identifies, how fast error falls, and which responses can be scored reliably. We characterise the identifiable directions, bound finite-sample error pointwise by an effective resistance, and give a minimax lower bound with the same dependence on conditioning, so the limits bind every estimator. Two consequences are properties of the graph alone and hold at every scale and under every feature map. First, none of six widely used public datasets contains a single comparison between responses to *different* prompts, and each fragments into 34 to 200 components, so every reward component varying across prompts is unidentifiable at any dataset size, model size or embedding, including inside two trained reward models' own representations. This is a gauge that within-prompt RLHF may ignore and that pooled best-of-N, reward-thresholded filtering, a shared value baseline and any reported mean reward may not: on the audited corpus, two rewards the data cannot tell apart, with identical log-likelihood in double precision, agree on 1.000 of within-prompt decisions while a pooled top-10% selection overlaps by only 0.385 and prompt ranking by mean reward moves at Spearman $+0.728$. Second, where the feature dimension no longer binds the rank the unlearnable subspace closes, but per-response conditioning spreads by 8 to $35\times$: the constraint moves rather than lifts. Linear, margin, MLP and transformer heads return bit-identical estimates along a direction the data cannot identify, and varying only topology at fixed budget changes error by $22.4\times$, a gap that does not close over a $256\times$ range of budgets. We also report where our own selection objective fails, and argue the diagnostic rather than any selection rule is the contribution.

1 INTRODUCTION

Reward models trained on pairwise human preferences sit at the centre of the RLHF pipeline (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). The pipeline treats preference data as a quantity: collect more comparisons, fit a larger head, expect a better reward model. Practitioners report that adding comparisons stops helping, and available diagnostics do not say why (Gao et al., 2023; Casper et al., 2023; Coste et al., 2024).

Finding Some reward directions are not learnable from a given preference dataset at all, and which ones is decided by *which pairs were compared* rather than by how many, or how large the model is; measuring this takes one pass and no training run. Two consequences are properties of the comparison graph alone, so they hold at any feature dimension: not one of six widely used public datasets contains a single comparison between responses to *different* prompts, and each fragments into 34 to 200 components, so by Corollary 2 every reward component varying across prompts is unidentifiable by construction, including in the penultimate representations of two trained reward models. Section 4.2 separates the RLHF decisions this costs nothing, since PPO, GRPO and the within-prompt part of DPO are invariant to it, from those that place two prompts on a common scale and are not.

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.

Table 1: The diagnostic. Three quantities read off one matrix, the question each settles, and the action each implies. Section 4 establishes the middle column; Section 5.3 measures the left on six public datasets.

Measure	Question it settles	If it is unfavourable
Size of the unlearnable subspace, $\dim \operatorname{null}(L_\phi)$	Which reward directions can this dataset identify at all?	Unavailable at any model size or budget; only comparisons spanning them help.
Conditioning of what remains, $\lambda^+_{\min}(L_\phi)$	How many labels to reach a given accuracy?	Same rate, worse constant; cross-prompt anchors raise it fastest.
Effective resistance of a response, $R_\phi(v, V)$	Which individual responses can be scored reliably?	High-resistance responses are the unreliable ones, and where comparisons pay most.

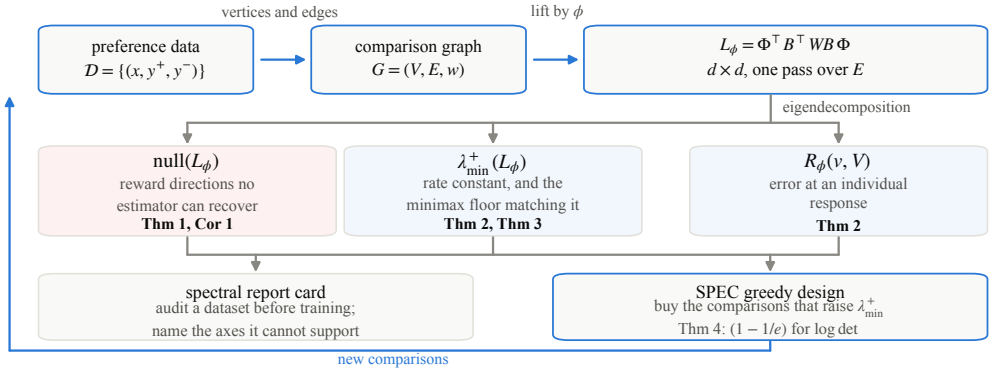

Figure 1: The framework. A preference dataset is a comparison graph; lifting it through the feature map yields L_ϕ, a $d \times d$ matrix built in one pass, independent of $|V|$. Its spectrum answers three questions: $\operatorname{null}(L_\phi)$ is what model capacity cannot recover, $\lambda^+_{\min}$ governs both the upper bound and a floor with the same dependence, and $R_\phi(v, V)$ ranks responses by scoreability.

Why the structure matters A preference dataset defines a graph: compared prompt–response pairs are its vertices and the comparisons its edges, so under Bradley–Terry (Bradley & Terry, 1952) fitting a reward is estimation on that graph. For item-level models this is established (Ford, 1957; Negahban et al., 2017; Shah et al., 2016; Chen et al., 2019), but RLHF rewards are functions of *features* over an open response space rather than one free scalar per item, so those statements do not transfer.

Let B be the signed edge–vertex incidence matrix, W the diagonal matrix of comparison counts, and Φ the matrix of features. Lifting the graph through the feature map gives the *feature-weighted comparison Laplacian*

$$L_\phi \;=\; \sum_{(v,u)\in E} w_{vu}\,\big(\phi(v)-\phi(u)\big)\big(\phi(v)-\phi(u)\big)^\top \;=\; \Phi^\top B^\top W B\,\Phi \;\in\; \mathbb{R}^{d\times d}. \tag{1}$$

Everything below is read off the spectrum of this one matrix, which is $d \times d$ whatever the dataset size and reduces to the ordinary graph Laplacian when ϕ is one-hot. Table 1 states what its three quantities mean and what to do when each is unfavourable, and Figure 1 summarises the pipeline.

Contributions

1. **A diagnostic, and the theory that licenses it.** Theorem 1 characterises which reward functionals a dataset identifies, Theorem 3 bounds error at *each response* by that response's effective resistance, and Theorem 4 gives minimax floors with the same $\lambda^+_{\min}$ dependence, so Table 1 describes the dataset rather than one estimator.
2. **The prompt-level gauge, and the decisions it costs** (Corollary 2, Sections 4.2 and 5.4). When comparisons never cross prompts the likelihood is exactly flat along every reward component constant within a prompt: within-prompt RLHF loses nothing, while a pooled best-of-N threshold, reward-score filtering, a prompt-independent value baseline and any reported mean reward do. On the audited corpus two rewards with identical log-likelihood agree on 1.000 of within-prompt decisions and overlap by 0.385 on a pooled top-10% selection.
3. **Spectral report cards for six public preference datasets** (Section 5.3), with the invariance checks and confound controls the claims require, including which of our own measurements we invalidated and why, and what replaces nullity at production scale.

4. **The limits are head-invariant, measured** (Section 5.1). Linear, margin, MLP and transformer heads return bit-identical estimates along a planted unidentifiable direction, and in a trained reward model's own penultimate representation the reward is linear by construction, so the theory applies there with no feature map chosen by us.
5. **What the diagnostic implies for collection, and where it does not** (Section 5.5). Spectral selection reaches a target conditioning with $2.13\times$ fewer labels than random; uncertainty sampling is *worse* than random on RewardBench, and for *repairing* an unidentifiable dataset maximising conditioning is also worse than random, so we scope the recommendation to collection.

2 Related work

Graph structure and covariates in Bradley–Terry estimation That the comparison graph governs item-level ranking is established: connectivity decides identifiability (Ford, 1957), the estimator is well understood (Simons & Yao, 1999; Hunter, 2004), and minimax rates carry spectral constants (Negahban et al., 2017; Shah et al., 2016; Hajek et al., 2014; Chen et al., 2019). Nor is the model new: linear-in-covariates worth goes back to Springall (1973), reviewed by Cattelan (2012), and since only differences enter, the design is the $X = B\Phi$ of Section 3, whose rank deficiency is known. We claim neither, and Theorem 1 is that design's rank condition. What is new is reading the design as a *graph lifted through a feature map*: this replaces the $|V| \times |V|$ Laplacian by the $d \times d$ object of Eq. 1, describes $\mathrm{null}(L_\phi)$ structurally in components and prompts (Corollary 2) rather than as a numerical rank, and indexes error by resistance at each response. That literature does inference given a design; we diagnose and choose one at RLHF scale.

Coverage conditions for reward learning Coverage is the organising quantity for offline RLHF guarantees (Zhu et al., 2023; Xiong et al., 2024; Song et al., 2024), and is the closest existing treatment of feature-space identifiability. The difference is what is held responsible: coverage there is an assumption on the data-generating distribution, here a spectral statistic of the dataset in hand, so $\lambda^+_{\min}(L_\phi)$ is the coverage coefficient, measured. Such an assumption also presumes the design is non-degenerate, whereas Theorem 1 says which functionals survive when it is not, and Section 5.3 finds every dataset we audit in that regime.

Reward-model pathologies and heterogeneous preferences Over-optimisation (Gao et al., 2023), reward gaming (Skalse et al., 2022), length bias (Singhal et al., 2024) and sycophancy (Sharma et al., 2024) are documented failure modes, some reduced by ensembles (Coste et al., 2024; Casper et al., 2023). These treat the reward model as the locus, whereas Section 5.1 shows that for failures caused by an unidentifiable direction it is not. A second line asks which reward to fit when annotators disagree (Siththaranjan et al., 2024; Chakraborty et al., 2024; Poddar et al., 2024); we ask which directions the topology admits at all, which is prior to it, since a subgroup reward supported on $\mathrm{null}(L_\phi)$ is unidentifiable however the mixture is modelled.

Active preference learning Acquisition functions for preference data include entropy and disagreement (Muldrew et al., 2024), epistemic uncertainty (Melo et al., 2024) and dueling-bandit exploration (Das et al., 2025; Jamieson & Nowak, 2011; Heckel et al., 2019; Yue et al., 2012). The contextual dueling bandit (Dudík et al., 2015; Saha & Krishnamurthy, 2022) is closest in form, since the learner there also chooses *which pair to compare*, but it chooses adaptively to minimise regret against an unknown comparison model, while we take the multiset as given and ask what it determines. Ours is the quantity Theorems 3 and 4 identify, placing it in optimal experimental design (Fedorov, 1972; Pukelsheim, 1993; Chaloner & Verdinelli, 1995) and connecting it to optimal selection with submodularity (Nemhauser et al., 1978), resistance-based sparsification (Spielman & Srivastava, 2011; Ghosh et al., 2008) and discrete design (Nikolov et al., 2019; Allen-Zhu et al., 2017). The geometric coreset rule (Sener & Savarese, 2018) is the closest non-spectral baseline and performs comparably on RewardBench, which our account predicts since it also maximises coverage; it uses the Euclidean rather than the L_ϕ^{-1} metric and supplies no diagnostic, guarantee, or repair mode.

3 Setup: the comparison graph and its Laplacian

Data and model A preference dataset is a multiset of comparisons over a vertex set V of prompt–response pairs $v = (x, y)$. Write E for the compared unordered pairs, n_{vu} for how often $\{v, u\}$ was compared, k_{vu} for how often v won, and $n = \sum_{(v,u)\in E} n_{vu}$ for the label budget; a fixed feature map $\phi : V \to \mathbb{R}^d$ gives $r_\theta(v) = \langle \theta, \phi(v)\rangle$. Under Bradley–Terry (Bradley & Terry, 1952) an annotator

prefers v with probability $\sigma(r_\theta(v) - r_\theta(u))$, so

$$\ell(\theta) = \sum_{(v,u)\in E} \Big[k_{vu} \log \sigma(\langle \theta, x_{vu} \rangle) + (n_{vu} - k_{vu}) \log\big(1 - \sigma(\langle \theta, x_{vu} \rangle)\big) \Big], \qquad x_{vu} := \phi(v) - \phi(u). \tag{2}$$

Only the *difference* features x_{vu} enter, which is the source of everything below: the dataset speaks about θ only through the span and conditioning of $\{x_{vu}\}$.

The Laplacian Collect them into $X = B\Phi$, with B the signed incidence matrix and Φ the features; with $W = \mathrm{diag}(w)$ we have $L_\phi = X^\top W X$ as in Eq. 1. Two weightings matter: $w = 1$ gives the design object, and $w = n_{vu}$ the count-weighted $L_\phi^{(n)}$ appearing in finite-sample statements. Write $\lambda_1 \le \cdots \le \lambda_d$ for the eigenvalues, $\lambda_{\min}^+$ for the smallest nonzero one, and

$$R_\phi(a) = a^\top L_\phi^+ a, \qquad R_\phi(v,u) = R_\phi\big(\phi(v) - \phi(u)\big), \qquad R_\phi(v,V) = R_\phi\big(\phi(v) - \bar\phi_{\mathrm{comp}(v)}\big) \tag{3}$$

for the feature effective resistance, where $\bar\phi_{\mathrm{comp}(v)} = |\mathrm{comp}(v)|^{-1} \sum_{u\in\mathrm{comp}(v)} \phi(u)$ is the mean feature of v's connected component; $R_\phi^{(n)}(v,V)$ is the same quantity with $L_\phi^{(n)}$ in place of L_ϕ. The component mean is the reference because Eq. 2 identifies reward only up to one constant per component, so $\phi(v) - \bar\phi_{\mathrm{comp}(v)}$ is the contrast the data speaks about. When ϕ is one-hot these are the graph Laplacian, resistance distance (Klein & Randić, 1993) and algebraic connectivity (Fiedler, 1973), which is why we call $\lambda_{\min}^+$ the algebraic connectivity in feature space.

Curvature Eq. 2 is a log-likelihood, so its Hessian is negative semi-definite; the *negative* Hessian is the Fisher information $F(\theta) = \sum_e n_e\, \sigma'(z_e)\, x_e x_e^\top \succeq 0$, so

$$\kappa\, L_\phi^{(n)} \;\preceq\; F(\theta) \;\preceq\; \tfrac{1}{4}\, L_\phi^{(n)}, \qquad \kappa := \min_e \sigma'(z_e) \;\ge\; \sigma'(B_\infty) \tag{4}$$

whenever $|z_e| \le B_\infty$. Eq. 4 bridges spectral statements about L_ϕ and statistical ones about the estimator; κ is the one quantity that can spoil the translation, and Section 5 measures where it does.

Gauge and rank budget Eq. 2 depends on θ only through differences, so reward is identified up to one additive constant per component, and reported errors are computed after centring within each component. Since $\mathrm{rank}(B) = |V| - c$ for c components,

$$\mathrm{rank}(L_\phi) \;\le\; \min\big(d,\; |V| - c\big), \tag{5}$$

so $\mathrm{null}(L_\phi)$ is forced nonempty exactly when $d > |V| - c$. Every table below names its operating point, since nullity claims need $d/(|V| - c) > 1$ and the confound analysis of Section 5.3 the opposite.

4 What the graph decides

Proofs are in Appendix A. Each result is followed by a plain restatement, so the section reads without them.

4.1 Which reward directions are learnable

Theorem 1 (Identifiability)**.** *Fix ϕ and a comparison multiset with all $n_{vu} > 0$. For $a \in \mathbb{R}^d$ the functional $\theta \mapsto \langle a, \theta \rangle$ is identifiable from Eq. 2 if and only if $a \in \mathrm{row}(L_\phi)$. Equivalently, for every $u \in \mathrm{null}(L_\phi)$ and every θ, ε, $\ell(\theta + \varepsilon u) = \ell(\theta)$, and the induced comparison distribution is unchanged on every compared pair.*

In words: the learnable directions are exactly those the compared pairs span; along any other the data is not noisy but silent, since two rewards differing there assign identical probabilities to every comparison in the dataset.

The proof is the standard rank argument for a generalised linear model, specialised to a design whose rows are feature differences, and being elementary is what makes the criterion checkable in one pass on a real corpus. The contribution is what follows from it: the subspace is non-empty and structured on every dataset practitioners use (Section 5.3), the error *ordering* across responses is governed by a resistance rather than by sample count (Theorem 3), and both survive a change of estimator and feature map (Section 5.1). It binds any estimator, loss or regulariser reading the same data, since a regulariser *selects* a point inside the unidentified subspace rather than identifying it:

Corollary 2 (Prompt-level gauge)**.** *Suppose every comparison is between two responses to the same prompt, and let $a \in \mathbb{R}^d$ be a direction along which ϕ is constant within every prompt. Then $a \in \mathrm{null}(L_\phi)$. Moreover $\dim \mathrm{null}(L_\phi) \ge d - (|V| - c)$ for c components, and if the graph has one component per prompt then c is the number of prompts.*

In words: if annotators only ever compare responses to the same prompt, anything distinguishing prompts rather than responses (difficulty, topic, domain) is invisible to the fit however many comparisons are collected, and Section 5.3 finds a cross-prompt fraction of exactly zero on all six datasets. The two clauses differ: that prompt-constant directions lie in $\mathrm{null}(L_\phi)$ depends only on the graph, so it holds at every d and under every feature map, while the count $d - (|V| - c)$ is arithmetic in the operating point.

4.2 When the prompt-level gauge is free, and when it is not

An unidentifiable direction matters only if something downstream reads it, and the obvious objection to Corollary 2 is that RLHF is built not to. The objection is correct for the inner loop: PPO, GRPO and the within-prompt part of DPO (Rafailov et al., 2023), whose implicit reward carries a per-prompt $\log Z(x)$ term that is exactly this gauge, are all invariant to adding a prompt-constant function to $r^\star$ (Appendix K), so for such a policy the lost directions are a coordinate choice.

The gauge stops being free where two *different* prompts are placed on one reward scale, and four standard practices do exactly that: a pooled best-of-N or rejection-sampling threshold, filtering or building curricula by reward score, a prompt-independent value baseline or reward-as-critic, and any reported mean reward. Appendix M says what each computation reads and why the per-prompt offset enters it.

Each is a separation *by construction*: it names a computation whose output depends on a prompt-constant component the data does not determine. Whether estimates actually move along that axis is answered in Section 5.4 on the audited corpus: two rewards differing by an element of $\mathrm{null}(L_\phi)$, with identical log-likelihood in double precision, agree on 1.000 of within-prompt decisions and overlap by only 0.385 on a pooled top-10% selection. Section 5.5 prices the complement on a planted instance, where *restoring* identifiability raises cross-prompt accuracy from 0.633 to 0.98 and leaves within-prompt accuracy inside 0.938–0.956. Neither is evidence about a particular deployed pipeline, which would need the $\|P_{\mathrm{null}} w\|/\|w\|$ measurement of Section 6; comparing the two reward models of Section 5.1 on cross-prompt against within-prompt pairs is the direct test, and we state it as open rather than as a result.

4.3 How fast error falls, and from what constant

Let $\hat\theta$ maximise Eq. 2 and $\hat r = \Phi\hat\theta$.

Theorem 3 (Upper bounds)**.** *Assume $|\langle\theta^\star, x_e\rangle| \le B_\infty$ for all e and let $\kappa = \sigma'(B_\infty)$. With probability at least $1-\delta$, for every vertex v,*

$$\left|\hat r(v) - r^\star(v) - c_v\right| \;\le\; C\sqrt{\frac{R_\phi^{(n)}(v,V)\,\log(2|V|/\delta)}{\kappa}}, \tag{6}$$

where C is an absolute constant, c_v is the gauge constant of v's component and $R_\phi^{(n)}(v,V)$ is as defined in Eq. 3. On the identifiable subspace,

$$\left\|\hat\theta - \theta^\star\right\|_{L_\phi^{(n)}} \le \sqrt{\frac{d\log(1/\delta)}{\kappa}}, \qquad \textit{hence} \qquad \left\|\hat\theta - \theta^\star\right\|_2 \le \sqrt{\frac{d\log(1/\delta)}{\kappa\,\lambda^+_{\min}(L_\phi^{(n)})}}. \tag{7}$$

If $\mathrm{null}(L_\phi) \neq \{0\}$ the Euclidean bound on the full parameter is infinite and Eq. 7 is a statement about $\mathrm{row}(L_\phi)$ only.

In words: error is ordered by effective resistance rather than spread evenly, so how reliably a response can be scored depends on where it sits in the graph; and error falls at the usual square-root rate from a constant set by the conditioning, so two datasets of *identical size* can differ by a large factor purely through which pairs were compared, which Section 5.2 measures at $22.4\times$. The argument is the self-concordant analysis of logistic regression (Bach, 2010) applied to X, so it describes the regime where a quadratic expansion at $\theta^\star$ is accurate. The two displays are proved by different routes: Eq. 7 bounds the whole parameter through the $L_\phi^{(n)}$-norm ellipsoid and legitimately carries $\sqrt{d}$, while Eq. 6 is a Bernstein bound on the *one* functional $\langle\phi(v) - \bar\phi, \cdot\rangle$ union-bounded over the $|V|$ vertices, which is why d does not appear in it and $|V|$ does. The weaker route, projecting Eq. 7 onto that functional by Cauchy–Schwarz, would give the same resistance scaling with an extra $\sqrt{d}$.

4.4 A LARGER MODEL DOES NOT HELP: THREE MINIMAX LOWER BOUNDS

Theorem 4 (Minimax lower bounds)**.** *Fix the comparison multiset and let $v_1, \dots, v_k$ be eigenvectors of $L_\phi^{(n)}$ for its k smallest* nonzero *eigenvalues $\lambda^+_{\min}(L_\phi^{(n)}) = \lambda_1 \le \cdots \le \lambda_k$. For every estimator $\hat\theta$ measurable with respect to the observed outcomes:* (a) *a two-point (Le Cam) construction along v_1 gives*

$$\inf_{\hat\theta} \sup_{\theta^\star} \mathbb{E}\big\|\hat\theta - \theta^\star\big\|_2 \;\ge\; \frac{1}{\sqrt{2}}\Big(1 - \tfrac{1}{\sqrt{8}}\Big)\,\frac{1}{\sqrt{\lambda^+_{\min}(L_\phi^{(n)})}} \;\ge\; \frac{0.45}{\sqrt{\lambda^+_{\min}(L_\phi^{(n)})}} \asymp \frac{1}{\sqrt{n\,\lambda^+_{\min}(L_\phi)}}; \quad (8)$$

(b) *packing the k weakest directions and applying Fano gives*

$$\inf_{\hat\theta} \sup_{\theta^\star} \mathbb{E}\big\|\hat\theta - \theta^\star\big\|_2 \;\gtrsim\; \sqrt{\frac{\log k}{\lambda_k}}; \quad (9)$$

(c) *an Assouad construction over $\{\pm\gamma_i\}^k$ in the same basis gives* $\inf \sup \mathbb{E}\|\hat\theta - \theta^\star\|_2^2 \gtrsim \sum_{i\le k} \lambda_i^{-1} \ge k/\lambda_k$*, a sum of effective resistances. In all three the left-hand side is infinite if* $\mathrm{null}(L_\phi) \neq \{0\}$.

In words: the bounds are over *all* estimators, so Table 1 describes the dataset rather than how we fit it, and where a direction is unidentifiable the floor is infinite and no architecture, loss or compute recovers it. This concerns information rather than capacity: a richer model can *represent* an unidentifiable reward and cannot be told which one the annotators meant, which Section 5.1 confirms across four head classes.

The three parts are separate statements and we do not chain them. Part (a) speaks to Eq. 7: both scale as $\lambda^+_{\min}(L_\phi^{(n)})^{-1/2}$, so conditioning enters the floor exactly as it enters the upper bound, and the remaining gap is a factor $\sqrt{d\log(1/\delta)/\kappa}$, the ambient dimension and the curvature constant of Eq. 4. We therefore claim a lower bound with the *same* $\lambda^+_{\min}$ *dependence* as the upper bound rather than a matching rate. Parts (b) and (c) price a different thing, namely that k weak directions cost more than one: since $\lambda_k \ge \lambda^+_{\min}$, they are weaker than (a) in $\lambda^+_{\min}$ and stronger than it in k, and replacing λ_k by $\lambda^+_{\min}$ in either would *raise* the right-hand side and is not licensed.

5 EXPERIMENTS

Protocol. Five rungs, each de-risking the next: (1) controlled atomic Bradley–Terry, where ϕ is one-hot and every quantity in Section 4 is known; (2) a planted unidentifiable direction; (3) real UltraFeedback text with a rubric reward, topology varied at fixed budget; (4) report cards for six public datasets; (5) SPEC end to end, and on RewardBench. Every experiment uses at least 5 seeds (default 8), errors use the per-component gauge of Section 3, and all computation is CPU-only. On rung 1 the theory holds as stated: Theorem 1 to machine precision, and the exponent of Theorem 3 converging to 0.504 against a predicted 0.500 once the sample is restricted to the regime the theorem scopes itself to. Appendices D and P give every predicted quantity against its measurement and index all ablations.

5.1 A LARGER REWARD MODEL DOES NOT HELP, AND THE AUDIT HOLDS IN A TRAINED MODEL'S OWN GEOMETRY

Theorem 4 binds every estimator, and since real reward models are fine-tuned transformers we test that rather than rest on it. On an instance with a planted unidentifiable direction, flipping the planted weight's sign changes the true reward but leaves every observable comparison unchanged, so a head extracting any information about that direction must move its coefficient. Linear, margin, MLP and transformer heads report *identical* coefficients at both signs, relative change 0.0000 in all four, while still recovering the *identifiable* component (smallest coefficient 0.854), which rules out limited capacity; the same holds under the MLP's neural tangent map (Jacot et al., 2018). The statement also holds with no feature map chosen by us: a reward model ending in a scalar head is by construction linear in its own penultimate representation, so setting $\phi := h$ makes Theorem 1 and Corollary 2 apply verbatim, and computing L_ϕ there for an encoder (`reward-model-deberta-v3-large-v2`) and a decoder (`internlm2-1_8b-reward`) puts every prompt-constant direction of each model's own h in $\mathrm{null}(L_\phi)$: along it the training data was silent and whatever the deployed head assigns was fixed by initialisation and weight decay rather than by an annotator. A bigger model changes *which* subspace is unlearnable, not what the data says (Appendix F).

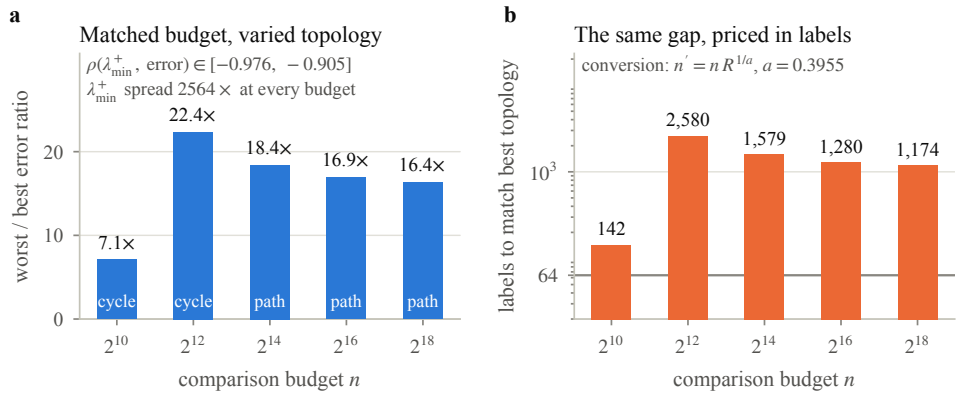

Figure 2: Rung 3. Real UltraFeedback text with rubric reward ($|V| = d = 256$, 8 seeds); within each budget only topology varies. *(a)* Worst-to-best error ratio, labelled by the worst topology; the gap persists across 256×, with $\lambda^+_{\min}$ ranking topologies by error ($\rho \in [-0.976, -0.905]$). *(b)* The same gap priced in labels via $n' = nR^{1/a}$, $a = 0.3955$; every value lies outside the fitted range.

Table 2: Spectral report cards at $|V| \leq 400$, $d = 512$, 8 seeds, Titan features; Chatbot Arena (Chiang et al., 2024) is access-gated and excluded. This is the rank-binding point $d > |V| - c$. Bold marks the two columns that depend on the graph alone and cannot be produced by counting; nullity equals $d - \min(d, |V| - c)$ exactly on all six, so it restates $|V|$ and the component count.

Dataset	$\lvert V\rvert$	comps.	x-prompt	comp.	rank	nullity	$\lambda_{\min}$	cond.	Gini
UltraFeedback (Cui et al., 2024; Tunstall et al., 2023)	400	200	**0.000**	**200**	200	312	0.0176	249	0.000
HelpSteer2 (Wang et al., 2024)	400	200	**0.000**	**200**	200	312	0.0050	723	0.000
Synthetic control (App. I)	400	200	**0.000**	**200**	200	312	0.0144	375	0.000
PKU-SafeRLHF (Ji et al., 2025)	399	200	**0.000**	**199**	200	312	0.0138	190	0.002
HelpSteer3 (Wang et al., 2025)	268	200	**0.000**	**134**	134	378	0.0031	1043	0.232
SHP (Ethayarajh et al., 2022)	170	254	**0.000**	**34**	136	376	0.0659	260	0.402

5.2 At a fixed budget, topology decides recovery

Rung 3 holds the label budget, item pool, feature map and reward fixed and varies only which pairs are compared. The substrate admits a linear reward with $R^2 = 0.9962$, the ceiling for any recovery number here, and $d/(|V| - c) = 1.212$, so Eq. 5 binds throughout. Across topologies conditioning spans 2564× and nullity ranges over $[2, 128]$, one lost direction for a connected graph and one per prompt for a partitioned one, which is Corollary 2. Recovery error differs by up to 22.4× and the ratio does not close as the budget grows over a 256× range, which is the claim we rest on. Figure 2b prices the same gap in labels; every value there lies outside the fitted range, so we read it as an illustration of scale rather than an estimate.

5.3 Auditing six public preference datasets

Two findings involve neither d nor our choice of ϕ, and we lead with them. The cross-prompt comparison fraction is 0.000 on every dataset, so prompt-level reward components are unidentifiable by Corollary 2 rather than by chance, with the consequences separated in Section 4.2. That this is zero by construction, since these corpora sample responses *per prompt* and annotate within that set, is the point rather than an objection: a quantity that is zero by construction is zero at every scale, is invariant to the feature map and the model, and is invisible in any loss curve computed from the same comparisons. What the collection protocol makes unremarkable, Section 5.4 shows is not free. Each graph also fragments into 34 to 200 components, four of the six into components of size 2, so differences across components are unidentifiable at *any* d.

The nullity column is mostly arithmetic, and what replaces it at scale Table 2 sits at $d/(|V| - c) = 2.56$, so Eq. 5 already forces nullity ≥ 312 on the four datasets with 200 components, and with Titan features the rank equals $\min(d, |V| - c)$ *exactly* on all six: the column restates $|V|$ and the component count. The two graph findings carry over to production scale unchanged because they do not mention d; the nullity finding does not, and what replaces it is conditioning. At $d/(|V| - c) \approx 0.008$ nullity is 0 on all six while the 90th-to-5th percentile ratio of $R_\phi(v, V)$ runs 8.4× on SHP to 35.4× on HelpSteer2, against 1.0× at the rank-binding point, so the error floor spreads by 8 to 35× within one corpus at one budget (Appendix N). Scale changes which row of Table 1 binds, not whether one does.

Appendix J reports the supporting checks: rank and nullity are invariant to edge-weighting and node granularity as Theorem 1 requires while conditioning is not, and every temporal and source

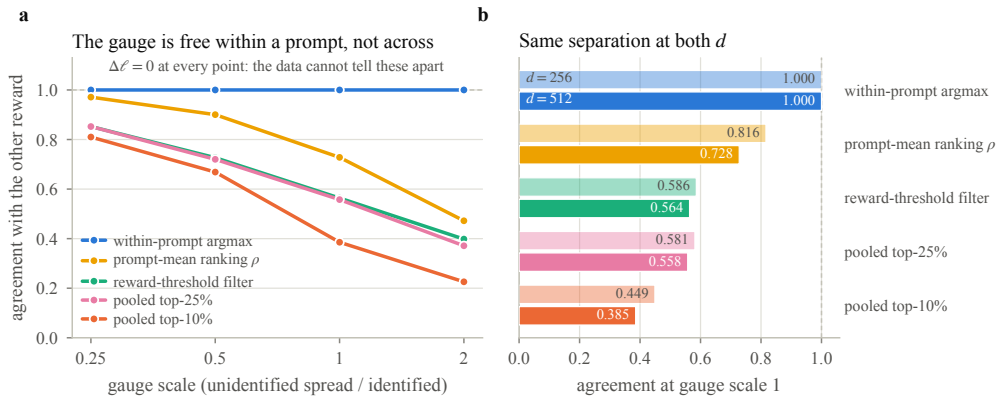

Figure 3: Ablation F4: the prompt-level gauge, measured. Two rewards differing only along $\mathrm{null}(L_\phi)$ on the UltraFeedback graph of Table 2 ($|V| = 400$, 8 seeds), so the log-likelihood is unchanged in double precision and no estimator reading this data can prefer either. *(a)* Dose response in the gauge scale, which the data does not determine either: the within-prompt argmax does not move, every pooled decision degrades. *(b)* The same at the anchor scale for both d.

slice is rank-deficient. It also reports what resistance does to error here, and that measurement needed correcting. Against a reward planted on the UltraFeedback graph itself, the quantity Eq. 6 bounds, resistance orders per-response error at $+0.171$ to $+0.358$ raw and $+0.164$ to $+0.332$ after controlling for degree, length and cluster size, positive at every d. The observable proxy this audit previously relied on, a parametric bootstrap, instead returns a *negative* correlation on the corpus's own labels: a property of the proxy rather than of the theorem, with the mechanism and its scope condition in Appendix J.

5.4 What the gauge costs, measured on the real corpus

Section 4.2 separates the RLHF decisions the prompt-level gauge costs nothing from those it does, and that separation is by construction. Ablation F4 runs it on the UltraFeedback graph of Table 2, where it becomes a measurement. We fit the reward on the corpus's own labels, draw a per-prompt target, realise it in feature space, and *project it onto* $\mathrm{null}(L_\phi)$. The projected direction u satisfies $\langle u, x_e \rangle = 0$ on every compared pair by construction, so $\hat{r}$ and $\hat{r} + \Phi u$ assign identical probabilities to every comparison in the dataset: the change in log-likelihood is 0 in double precision and Φu is constant within each component to 1.3×10^{-14}. By Theorem 1 no estimator, loss or regulariser reading this data can prefer one over the other, so which one a deployed model reports is settled by its regulariser. The magnitude of u is equally undetermined, so we sweep it and quote the scale at which the unidentified component carries the same spread as the identified one.

At that scale (Figure 3) the two rewards agree on 1.000 of within-prompt argmax decisions, exactly, on every component carrying an ordering to preserve, with 1 of 200 components numerically tied under the fit and therefore ordered by floating point rather than by the gauge. They disagree substantially on every decision that pools across prompts. A pooled best-of-N keeping the global top 10% retains a set with Jaccard overlap 0.385 against the same rule under the other reward, 25% gives 0.558, and a filter at a cutoff calibrated once on the first reward keeps 1.139 times as many responses, changing how many survive for 29.4% of prompts. Ranking prompts by mean reward, the quantity most often reported, moves at Spearman $+0.728$. The disagreement grows monotonically with the gauge scale from zero and the within-prompt figure does not move at any of them, which is the separation of Section 4.2 with numbers attached rather than an argument. This is a statement about what the data leaves undetermined, not a claim that any particular deployed model sits at a particular point in $\mathrm{null}(L_\phi)$; that would need the $\|P_{\mathrm{null}} w\|/\|w\|$ measurement Section 6 lists as open.

The same question at production scale The construction needs $\mathrm{null}(L_\phi) \neq \{0\}$. At $|V| = 4000$, where nullity is 0, nothing is exactly free; asking the same question of the *weak* directions that Theorem 4(b,c) prices, at a χ^2_k likelihood-ratio budget, leaves within-prompt agreement at 0.827 against a pooled top-10% overlap of 0.465: the separation persists as an ordering rather than an identity (Appendix O).

5.5 Acting on the diagnostic: selection, repair, and where the objective is wrong

If conditioning sets the constant in both the upper bound and the floor, a budget should be spent raising it. Soft-min SPEC reaches the target conditioning in 963.5 labels against random's 2048, with

a $(1-1/e)$ guarantee for the log-determinant alone (Theorem 6), but three results run against our own framing: uncertainty sampling (Lewis & Gale, 1994) is *worse* than random on RewardBench by 2.97 points at 200 labels, maximising $\lambda^{+}_{\min}$ is worse than random for *repair*, and listwise annotation does not restore identifiability, since the Plackett–Luce (Plackett, 1975; Luce, 1959) Fisher information shares the null space exactly. Repair does isolate the axis Corollary 2 names, moving cross-prompt accuracy from 0.633 to 0.98 while within-prompt stays within 0.938–0.956 (Appendices H and L).

6 Discussion

What the contribution is, and what to do with it The deliverable is the diagnostic: three quantities, read off one matrix in one pass, saying what a preference dataset can and cannot support, licensed by a lower bound that binds every estimator and stated so it applies inside a trained reward model's own representation. In practice: *audit before training*, reading nullity with its rank-budget ratio and never alone; *collect cross-prompt comparisons*, since all six audited datasets have exactly zero and a few anchors suffice; *decide whether you read the gauge*, free within a prompt but not across it, where the regulariser rather than the annotators sets the quantity compared and a pooled top-10% selection moves to 0.385 overlap (Section 5.4); and *order curation after design*, which applied alone lowered conditioning by 30% and made recovery worse.

Limitations and scope The bounds are variance statements about a fixed feature map, so approximation bias outside its span is not covered and every recovery number is stated against the best-linear-predictor ceiling. Upper and lower bounds share their $\lambda^{+}_{\min}$ dependence but are not matched: the gap is $\sqrt{d\log(1/\delta)/\kappa}$, and closing it is open. Theorem 3 holds in a well-conditioned logistic regime, and that scope binds the measurement as well as the theorem: on corpora carrying one comparison per edge with a fixed winner, a saturating fit puts the parametric-bootstrap error proxy outside the regime it is valid in, which is why Appendix J retires it and measures error against a planted reward instead. Resistance *predicts* error on real data only for $d/|V| \gtrsim 0.016$, and where components have size 2 the per-response statement collapses onto an edge-level one. The gauge measurement of Section 5.4 is exact only where $\mathrm{null}(L_\phi) \neq \{0\}$: once the rank budget stops binding, within-prompt invariance degrades from 1 to 0.827 and what remains is an ordering between decision classes rather than an exact separation. It also says what the data leaves undetermined, not where a particular trained model sits inside that set; making that quantitative needs the prompt-constant variance fraction of h, the ratio $\|P_{\mathrm{null}}w\|/\|w\|$ for the deployed head, and the cross-prompt against within-prompt agreement between the two models, none of which is measured here. The $(1-1/e)$ guarantee covers the log-determinant objective alone, the soft-min certificate is a statement about the subspace the candidate pool spans rather than about all of $\mathbb{R}^d$, and its greedy gain is a first-order approximation whose per-step argmax agrees with the exact rank-one gain about 65% of the time. Since named-axis scores at rank-starved points do not reproduce in a trained model's geometry, no claim here takes the form "this dataset cannot learn axis X".

Reproducibility statement

Every number is produced by the accompanying implementation and reproducible from the released run book: 35 modules and 323 offline tests, including 11 end-to-end driver runs, finite-difference gradient checks for the transformer head, and exact checks for the log-determinant and A-optimal gain formulas. A separate script re-checks 63 predictions against the produced artifacts and reports pass, fail or diagnostic, so a missing stage cannot read as a success, and each run writes a manifest with the grid, git commit, versions and spend. Because several results are regime-dependent, every table and figure names its operating point, all errors use the per-component gauge of Section 3, and quantities measured under a proxy we later invalidated are recorded under that proxy's name. The pipeline runs fully offline with a word-hashing map; hosted-model results cost about USD 6, and all computation is CPU-only.

Ethics statement

We analyse six publicly released preference datasets and one public benchmark under their published terms; Chatbot Arena is access-gated and was excluded rather than circumvented. No human subjects were recruited, no new annotations were collected, and annotator identities are not used. Two findings bear on deployment: these datasets cannot identify prompt-level reward components, so a model trained on them has directions set by a regulariser rather than by human judgement, worth disclosing when it gates decisions; and our cross-corpus result shows a reward model scoring below chance out of distribution while scoring perfectly within it, which the diagnostic surfaces first.

AI USAGE STATEMENT

Generative AI tools were used to assist with editing and formatting portions of the manuscript for readability and presentation, as well as for suggesting keywords, summarizing the work, and reviewing the relevance and formatting of references. Generative AI was not used to generate experimental results or synthetic datasets.

All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, references, results, and artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *arXiv preprint arXiv:1711.05174*, 2017.

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433, 2012.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *International Conference on Machine Learning*, 2024.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-k ranking. *The Annals of Statistics*, 47(4):2204–2235, 2019.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning*, 2024.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *International Conference on Machine Learning*, 2024.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient RLHF. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2025.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587, 2015.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, 2022.

Valerii V. Fedorov. *Theory of Optimal Experiments*. Academic Press, 1972.

Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2): 298–305, 1973.

Lester R. Ford. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Arpita Ghosh, Stephen Boyd, and Amin Saberi. Minimizing effective resistance of a graph. *SIAM Review*, 50(1):37–66, 2008.

Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, 2014.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019.

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, 2011.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, Boren Zheng, Tianyi Qiu, Jiayi Zhou, Kaile Wang, Boxuan Li, Sirui Han, Yike Guo, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

Douglas J. Klein and Milan Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12 (1):81–95, 1993.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James V. Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL*, 2025.

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Luckeciano C. Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep Bayesian active learning for preference modeling in large language models. In *Advances in Neural Information Processing Systems*, 2024.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In *International Conference on Machine Learning*, 2024.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978.

Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for A-optimal design. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1369–1386, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, 2024.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In *International Conference on Algorithmic Learning Theory*, pp. 968–994, 2022.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Gordon Simons and Yi-Ching Yao. Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons. *The Annals of Statistics*, 27(3):1041–1060, 1999.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *Conference on Language Modeling*, 2024.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, 2022.

Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *Advances in Neural Information Processing Systems*, 2024.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

A. Springall. Response surface fitting using a generalization of the Bradley–Terry paired comparison model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(1):59–68, 1973.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, 2024.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *International Conference on Machine Learning*, pp. 54715–54754, 2024.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A PROOFS

Theorem 4 rests on a quadratic bound for the Bradley–Terry divergence whose constant we take in its tight form, because the looser constant in common use costs a factor $\sqrt{2}$ in every bound that follows.

Lemma 5 (Tight quadratic KL bound)**.** *For θ, θ' on the same comparison multiset, $\mathrm{KL}(P_\theta \| P_{\theta'}) = \sum_e n_e \, \mathrm{KL}_{\mathrm{Ber}}(\sigma(z_e) \| \sigma(z'_e))$ and*

$$\mathrm{KL}(P_\theta \| P_{\theta'}) \;\le\; \tfrac{1}{8} \, (\theta - \theta')^\top L_\phi^{(n)} (\theta - \theta'), \tag{10}$$

with $\frac{1}{8}$ tight: the ratio tends to $\sigma'(0)/2 = \frac{1}{8}$ as $\theta' \to \theta$.

Proof of Lemma 5. Let $f(b) = \mathrm{KL}_{\mathrm{Ber}}(\sigma(a)\|\sigma(b))$. Then $f(a) = 0$, $f'(b) = \sigma(b) - \sigma(a)$ so $f'(a) = 0$, and $f''(b) = \sigma'(b) \le \frac{1}{4}$; Taylor with the second-derivative bound gives $f(b) \le \frac{1}{8}(a-b)^2$, and summing over edges with multiplicity n_e using $z_e - z'_e = \langle \theta - \theta', x_e \rangle$ gives Eq. 10. Tightness follows from $f''(a) = \sigma'(a)$ and $\max_a \sigma'(a) = \frac{1}{4}$. The ratio in Eq. 10 is verified numerically to converge to 1. □

Proof of Theorem 1. $L_\phi = X^\top W X$ with $W \succ 0$, so $\mathrm{null}(L_\phi) = \mathrm{null}(X)$: for $u \in \mathrm{null}(L_\phi)$, $u^\top L_\phi u = \sum_e w_e \langle x_e, u\rangle^2 = 0$ forces $\langle x_e, u\rangle = 0$ for every e. Then $\langle \theta + \varepsilon u, x_e\rangle = \langle \theta, x_e\rangle$ for all e, so every summand of Eq. 2 and every predicted win probability is unchanged and the two parameters are observationally indistinguishable. Conversely if $a \in \mathrm{row}(L_\phi) = \mathrm{row}(X)$ then $a = X^\top \beta$, so $\langle a, \theta\rangle$ is a linear combination of the $\langle x_e, \theta\rangle$, each identified from the outcome distribution on edge e by strict monotonicity of σ. □

Proof of Corollary 2. Each edge joins two responses to one prompt, so $\langle a, x_e\rangle = 0$ for every e by within-prompt constancy, and Theorem 1 applies; the dimension bound is Eq. 5. □

Theorem 3 has two displays and they are proved separately, which is where their dimension dependence parts company.

Eq. 7 is the self-concordant analysis of logistic regression (Bach, 2010) applied to X: Eq. 4 converts the Fisher information into $L_\phi^{(n)}$, a quadratic expansion at $\theta^\star$ gives concentration of $\hat\theta - \theta^\star$ in the $L_\phi^{(n)}$ norm, and dividing by $\lambda_{\min}^+$ gives the Euclidean form. It controls all d coordinates at once and carries $\sqrt{d}$ for that reason.

Eq. 6 is proved directly for a *single* linear functional, which is what removes the $\sqrt{d}$:

1. In the regime the theorem scopes itself to, the self-concordant expansion gives $\hat\theta - \theta^\star = F(\theta^\star)^{-1}\nabla\ell(\theta^\star) + \mathrm{rem}$ with $\|\mathrm{rem}\|_{F(\theta^\star)}$ of lower order.
2. $\nabla\ell(\theta^\star) = \sum_e \big(k_e - n_e\sigma(z_e^\star)\big)x_e$ is a sum of independent mean-zero terms, $k_e \sim \mathrm{Binomial}(n_e, \sigma(z_e^\star))$, with $|k_e - n_e\sigma(z_e^\star)| \le n_e$.
3. Fix $a \in \mathrm{row}(L_\phi)$. Then $\langle a, F^{-1}\nabla\ell(\theta^\star)\rangle = \sum_e (k_e - n_e\sigma(z_e^\star))\,\langle a, F^{-1}x_e\rangle$ has variance $a^\top F^{-1}\big(\sum_e n_e\sigma'(z_e^\star)x_e x_e^\top\big)F^{-1}a = a^\top F^{-1}a \le R_\phi^{(n)}(a)/\kappa$ by Eq. 4, so Bernstein's inequality gives, with probability $1-\delta$, $|\langle a, F^{-1}\nabla\ell(\theta^\star)\rangle| \le C\sqrt{R_\phi^{(n)}(a)\log(2/\delta)/\kappa}$ plus a lower-order term.
4. Take $a = \phi(v) - \bar\phi_{\mathrm{comp}(v)}$ and union-bound over the $|V|$ vertices, replacing $\log(2/\delta)$ by $\log(2|V|/\delta)$.

The constant C is therefore absolute. Obtaining Eq. 6 the other way, by projecting the ellipsoid of Eq. 7 onto a with Cauchy–Schwarz, gives $|\langle a, \hat\theta - \theta^\star\rangle| \le \sqrt{R_\phi^{(n)}(a)} \cdot \sqrt{d\log(1/\delta)/\kappa}$: the same resistance scaling, but carrying a $\sqrt{d}$ that the direct argument does not. Being a quadratic expansion, both displays describe only the regime where that expansion is accurate, which is why Section 5 reports the restricted and the unrestricted fits both.

Theorem 4 has three parts, each proved by the tool that gives it, and we chain none of them: part (a) is a two-point argument that prices $\lambda_{\min}^+$, parts (b) and (c) pack the k weakest directions and price their number. All three control the divergence with Lemma 5, whose tight constant $\frac{1}{8}$ rather than the $\frac{1}{4}$ in common use is worth $\sqrt{2}$ in every bound that follows.

Proof of Theorem 4. (a) Le Cam. Take $\theta_0 = 0$ and $\theta_1 = \gamma v_1$, both in $\mathrm{row}(L_\phi)$, so $\|\theta_0 - \theta_1\|_2 = \gamma$. By Lemma 5, $\mathrm{KL}(P_{\theta_0}\|P_{\theta_1}) \le \frac{1}{8}\gamma^2\lambda_1 = \frac{1}{8}\gamma^2\lambda_{\min}^+(L_\phi^{(n)})$. Le Cam's two-point bound gives $\inf_{\hat\theta}\sup_{\theta^\star}\mathbb{E}\|\hat\theta - \theta^\star\|_2 \ge \frac{\gamma}{2}\big(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})\big)$, and Pinsker in the form $\mathrm{TV} \le \sqrt{\mathrm{KL}/2}$ bounds the total variation. Choosing $\gamma = \sqrt{2/\lambda_{\min}^+(L_\phi^{(n)})}$ makes $\mathrm{KL} \le \frac{1}{4}$ and $\mathrm{TV} \le 1/\sqrt{8}$, so the risk is at least $\frac{1}{\sqrt{2}}(1 - 1/\sqrt{8})\,\lambda_{\min}^+(L_\phi^{(n)})^{-1/2} \ge 0.457\,\lambda_{\min}^+(L_\phi^{(n)})^{-1/2}$, which is Eq. 8. Since $L_\phi^{(n)}$ scales linearly in n at a fixed allocation over edges, $\lambda_{\min}^+(L_\phi^{(n)}) \asymp n\,\lambda_{\min}^+(L_\phi)$ and the last form follows.

(b) Fano. Take the packing $\{\gamma v_i\}_{i\le k} \cup \{0\}$: any two elements are $\ge \gamma$ apart in ℓ_2, and by Lemma 5 their divergence is at most $\frac{1}{4}\gamma^2\lambda_k$. Choosing $\gamma = \sqrt{4(1-c)\log k/\lambda_k}$ makes the average pairwise

divergence at most $(1-c)\log k$, so Fano's inequality (Yu, 1997; Tsybakov, 2009) prevents identifying the index with probability c, giving $\inf\sup\mathbb{E}\|\hat{\theta}-\theta^\star\|_2 \geq c\gamma/2 \gtrsim \sqrt{\log k/\lambda_k}$, which is Eq. 9. Note the direction of the spectral inequality: $\lambda_k \geq \lambda^+_{\min}$, so replacing λ_k by $\lambda^+_{\min}$ here would increase the right-hand side and is not a valid weakening: Eq. 9 is a statement about λ_k and is left as one.

(c) Assouad. Flipping coordinate i of $\tau \in \{\pm 1\}^k$ changes θ by $\Delta = 2\gamma_i v_i$, so $\Delta^\top L_\phi^{(n)}\Delta = 4\gamma_i^2\lambda_i$ and Lemma 5 gives $\mathrm{KL} \leq \frac{1}{8}\cdot 4\gamma_i^2\lambda_i = \frac{1}{2}\gamma_i^2\lambda_i$. Pinsker in the form $\mathrm{TV} \leq \sqrt{\mathrm{KL}/2}$ then bounds the total variation between the two mixtures by $\sqrt{\gamma_i^2\lambda_i/4} = \gamma_i\sqrt{\lambda_i}/2$. Taking $\gamma_i = \lambda_i^{-1/2}$ makes each factor the constant $\frac{1}{2}$ and Assouad's lemma yields $\inf\sup\mathbb{E}\|\hat{\theta}-\theta^\star\|_2^2 \gtrsim \sum_{i\leq k}\gamma_i^2 = \sum_{i\leq k}\lambda_i^{-1}$, the sum of effective resistances along those directions, which is $\geq k/\lambda_k$.

In all three cases, if $\mathrm{null}(L_\phi) \neq \{0\}$ then arbitrarily separated parameters have zero divergence and the risk is unbounded. □

What the lower bounds do and do not match. Part (a) and Eq. 7 share the factor $\lambda^+_{\min}(L_\phi^{(n)})^{-1/2}$, so conditioning enters the floor exactly as it enters the upper bound; the ratio between them is $\sqrt{d\log(1/\delta)/\kappa}$, the ambient dimension and the curvature constant. Parts (b) and (c) are not comparable to (a) term by term, because $\lambda_k \geq \lambda^+_{\min}$ makes their λ_k dependence weaker and their k dependence stronger. We therefore claim throughout a lower bound with the same $\lambda^+_{\min}$ dependence as the upper bound, not a matching minimax rate.

PROOF OF THEOREM 6

Proof. (a) By the matrix determinant lemma the marginal gain of adding x to a design with $A = L_0 + \sum_{c\in S} x_c x_c^\top + \rho I$ is $\log(1 + x^\top A^{-1}x) > 0$, giving monotonicity. If $S \subseteq T$ then $A_S \preceq A_T$, and $A \mapsto A^{-1}$ is operator antitone, so $x^\top A_T^{-1}x \leq x^\top A_S^{-1}x$ and the gain can only shrink; this is diminishing returns. (b) On $\mathcal{S}_\mathcal{C}$, $\sum_{k\leq r} e^{-\tau\lambda_k^S} \geq e^{-\tau\lambda_1^S}$ gives the upper bound and $\sum_{k\leq r} e^{-\tau\lambda_k^S} \leq r\, e^{-\tau\lambda_1^S}$ the lower; taking logs and dividing by $-\tau$ gives Eq. 13. Concavity is inherited because smin_τ^+ is a concave spectral function of the PSD matrix $P_\mathcal{C} L P_\mathcal{C}$ restricted to $\mathcal{S}_\mathcal{C}$, and $L \mapsto P_\mathcal{C} L P_\mathcal{C}$ is linear. (c) For A-optimality the Sherman–Morrison gain is $\|A^{-1}x\|^2/(1 + x^\top A^{-1}x)$, whose numerator and denominator both shrink as A grows, so the ratio need not. This is settled by an exact counterexample in rational arithmetic: with $A_0 = I$, $d = 2$, $f(S) = -\operatorname{tr}\big((I + \sum_{i\in S} v_i v_i^\top)^{-1}\big)$, $A = \{(1,1)\}$, $T = \{(1,1),(3,1)\}$ and $c = (2,3)$,

$$\big[f(T\cup c) - f(T)\big] - \big[f(A\cup c) - f(A)\big] \;=\; \tfrac{623}{4080} \;>\; 0, \tag{11}$$

so the marginal gain grew on the larger set. No floating point is involved. For $\lambda^+_{\min}$ and smin_τ, violations are exhibited by direct enumeration (Section 5.5). Concavity of Eq. 12 in the matrix argument gives a guarantee for the continuous relaxation of the design problem, not for greedy on the discrete set. □

B DESIGN OBJECTIVES AND THEIR GUARANTEES

E-optimality maximises $\lambda^+_{\min}$ directly, and it is not submodular: adding an edge can rotate the minimising eigenvector and thereby *increase* another candidate's marginal gain. We therefore optimise two surrogates and report the trade-off.

The soft-min surrogate has to be defined on a subspace fixed in advance. Let $\mathcal{S}_\mathcal{C} = \mathrm{row}(L_0) + \mathrm{row}\big(\sum_{c\in\mathcal{C}} x_c x_c^\top\big)$ be the span of everything the candidate pool could ever identify, $r = \dim\mathcal{S}_\mathcal{C}$, and $P_\mathcal{C}$ the orthogonal projector onto it. Write $\lambda_1^S \leq \cdots \leq \lambda_r^S$ for the eigenvalues of L restricted to $\mathcal{S}_\mathcal{C}$ and define

$$\mathrm{smin}_\tau^+(L) \;=\; -\tfrac{1}{\tau}\log\textstyle\sum_{k=1}^r e^{-\tau\lambda_k^S(P_\mathcal{C} L P_\mathcal{C})}. \tag{12}$$

This is a concave spectral function of the restricted PSD matrix (Boyd & Vandenberghe, 2004), so the fractional relaxation is unchanged. The restriction is what makes part (b) below non-vacuous: summed over all d eigenvalues, any rank-deficient design has $\lambda_1(L) = 0$ and the sandwich would read $-\log d/\tau \leq \mathrm{smin}_\tau(L) \leq 0$, which is true and says nothing about $\lambda^+_{\min}$, and rank-deficient is every design in Table 2 and every early step of SPEC. The subspace must be fixed *before* selection rather than taken to be "the currently nonzero eigenvalues", since that set grows as edges are added and the objective would be discontinuous.

This changes what is certified, not what is selected. Every candidate x_c lies in $\mathcal{S}_\mathcal{C}$ by construction, so any eigendirection orthogonal to it has $\langle v_k, x_c \rangle = 0$ and contributes nothing to any candidate's gain. Dropping them and renormalising therefore multiplies *every* candidate's gain by the same positive constant $\big(\sum_{k\le d} e^{-\tau\lambda_k}\big)/\big(\sum_{k\le r} e^{-\tau\lambda_k}\big)$, which depends on L and τ but not on the candidate. The greedy ranking, and hence every selected set reported in Section 5.5, Table 5 and Figure 6, is identical under either definition, which the test suite asserts.

Two coincidences keep the comparison honest: log-det greedy *is* resistance-greedy step for step, since the gain $\log(1 + x^\top A^{-1}x)$ increases in $x^\top A^{-1}x$, so the two are not independent arms; and soft-min greedy coincides with $\lambda^+_{\min}$ greedy once τ times the spectral gap is large, which is why the soft-min row inherits the $\lambda^+_{\min}$ objective's failure in Section 5.5.

Theorem 6 (Guarantees, in three parts)**.** (a) $S \mapsto \log\det(L_0 + \sum_{c\in S} x_c x_c^\top + \rho I)$ *is monotone and submodular, so greedy attains a* $(1 - 1/e)$ *fraction of the optimum (Nemhauser et al., 1978), and lazy greedy (Leskovec et al., 2007) is valid.* (b) *For every* L *supported on* $\mathcal{S}_\mathcal{C}$ *and every* $\tau > 0$*,*

$$\lambda_1^{\mathcal{S}}(L) - \frac{\log r}{\tau} \;\le\; \operatorname{smin}^+_\tau(L) \;\le\; \lambda_1^{\mathcal{S}}(L), \tag{13}$$

so a design whose soft-min is high has a certified $\lambda_1^{\mathcal{S}}$*, with slack* $\log r/\tau$*; once the design has full rank on* $\mathcal{S}_\mathcal{C}$*,* $\lambda_1^{\mathcal{S}}(L) = \lambda^+_{\min}(L)$ *and this certifies the conditioning constant of Theorem 3. Before then both sides are* 0*.* (c) *Neither* $\lambda^+_{\min}$*,* $\operatorname{smin}^+_\tau$*, A-optimality* $-\operatorname{tr}\big((L+\rho I)^{-1}\big)$*, nor V-optimality is submodular, so no* $(1 - 1/e)$ *claim is made for any of them.*

In words: exactly one natural objective, the log-determinant, is submodular, which is what licenses greedy selection with a worst-case guarantee. The one we would most like to optimise, the conditioning itself, is not, so we optimise a smooth surrogate and certify the real quantity from it: part (b) says a design with a good surrogate value provably has good conditioning on the subspace the pool can reach, with a gap we shrink by raising τ. Restricting to that subspace is what keeps the certificate meaningful, since an unrestricted soft-min certifies nothing whenever the design is rank-deficient, which it is at every step until the pool's span is covered. Part (c) records what carries no guarantee, including a case often assumed to. Section 5.5 finds the uncomfortable ordering: the objective with the proof is the weakest in realised gain.

The proof is in Appendix A: (a) is the matrix determinant lemma plus operator antitonicity of $A \mapsto A^{-1}$, (b) is two lines of algebra on the log-sum-exp, where $\sum_{k\le r} e^{-\tau\lambda_k^{\mathcal{S}}} \ge e^{-\tau\lambda_1^{\mathcal{S}}}$ gives the upper bound and $\sum_{k\le r} e^{-\tau\lambda_k^{\mathcal{S}}} \le r\, e^{-\tau\lambda_1^{\mathcal{S}}}$ the lower, and (c) is settled for A-optimality not by a sampled counterexample but by an exact one in rational arithmetic: with $A_0 = I$ in $d = 2$, the sets $\{(1,1)\}$ and $\{(1,1),(3,1)\}$ and the candidate $(2,3)$ give a diminishing-returns violation of exactly $623/4080 > 0$, so no floating-point objection applies.

A stale gain is a valid upper bound only under submodularity, so our implementation restricts the lazy accelerator to log-det.

Measured against 150 sampled nested pairs, the statement holds as written: log-det has 0 diminishing-returns violations and its greedy solution attains the brute-force optimum (ratio 1.0000 against the $(1 - 1/e) = 0.632$ floor), the soft-min and $\lambda^+_{\min}$ violate diminishing returns in 105 and 101 pairs so no $(1 - 1/e)$ claim is made for either, and the sandwich of Eq. 13 holds at every temperature with slack exactly $\log r/\tau$, shrinking from 2.48 to 0.0248 over the τ grid. On that instance the pool spans the space, so $r = d = 12$ and the slack is numerically what the unrestricted form would give; the two part company only on a rank-deficient design, which is why the separate check on one is in the test suite rather than here. Greedy on $\lambda^+_{\min}$ reaches a ratio of 0.0395, which is not a defect but the reason part (c) exists. A-optimality shows 0 violations *in this sample*, which we do not read as submodularity: 150 nested pairs cannot establish it and the counterexample above refutes it analytically, so we report that field as *no violations in this sample*.

C THE SPEC LOOP

The gains, and which of them are exact. Line 4 costs one $O(|\mathcal{C}|d^2)$ pass over the pool for every objective, but the three gain formulas do not have the same status and we distinguish them. Log-det and A-optimality have *exact* rank-one gains, $\log(1+x^\top A^{-1}x)$ by the matrix determinant lemma and $\|A^{-1}x\|^2/(1 + x^\top A^{-1}x)$ by Sherman–Morrison: each is the true discrete difference and is asserted against a direct recomputation in the test suite. The soft-min has no such closed form, because log-

Algorithm 1 SPEC selection

Require: starting Laplacian L_0, pool $\mathcal{C}$ with features $\{x_c\}$, budget B, objective g, ridge ρ, temperature τ

1: $L \leftarrow L_0$; $S \leftarrow \emptyset$
2: **for** $t = 1, \dots, B$ **do**
3: $A \leftarrow L + \rho I$; compute A^{-1} and, for $g = \mathrm{smin}_\tau^+$, the spectrum of L ▷ $O(d^2)$; $O(d^3)$ for the spectrum
4: score every $c \in \mathcal{C} \setminus S$ by its rank-one gain in g ▷ $O(|\mathcal{C}|d^2)$
5: $c^\star \leftarrow \arg\max_c \mathrm{gain}(c)$; $S \leftarrow S \cup \{c^\star\}$; $L \leftarrow L + x_{c^\star} x_{c^\star}^\top$
6: **end for**
7: **return** S ▷ $g = \log\det$: replace lines 3–5 by lazy greedy

sum-exp is nonlinear in the eigenvalues. What line 4 uses for it is the *first-order gain*, the directional derivative of Eq. 12 at the current spectrum,

$$\mathrm{gain}_{\mathrm{smin}}(x) \;=\; \left\langle \nabla\, \mathrm{smin}_\tau^+(L),\, xx^\top \right\rangle \;=\; \sum_k p_k \langle v_k, x\rangle^2, \qquad p_k \propto e^{-\tau\lambda_k}, \tag{14}$$

a weighted preference for the currently weakest directions. It costs $O(d^2)$ per candidate against $O(d^3)$ for the exact discrete update, which is why it is what greedy uses.

What the approximation costs. Concavity makes Eq. 14 an over-estimate of the true gain, so greedy on it is an optimistic-ascent rule. Measured on a SPEC run against the exact rank-one gain recomputed for a random 5% of candidate evaluations at each step (8 seeds, $d = 12$, pool 200, $B = 40$), the two rules rank candidates at Spearman 0.816 (0.786–0.893), the per-step argmax agrees 64.7% of the time (55–80%), and the first-order pick's mean relative shortfall against the exact gain at the same state is 0.100. Agreement is therefore *not* high, and the question that matters is whether the resulting design is worse. It is not: running the identical loop under the exact gain throughout gives a *lower* final $\lambda_{\min}^+$ in 8 of 8 seeds, the first-order rule reaching $1.45\times$ (1.32–$1.82\times$) the exact rule's value. Exact myopic greedy is not the optimum either, and the first-order rule's optimism spends more of the budget on the weakest directions. We therefore keep the first-order gain and report it as an approximation rather than as a closed form.

Complexity. Per step: the $O(|\mathcal{C}|d^2)$ pool pass, plus the refresh on line 3. For log-det, A-optimality, V-optimality and resistance-greedy the refresh is only A^{-1}, maintainable in $O(d^2)$ by Sherman–Morrison. For the soft-min and for $\lambda_{\min}^+$ it is a full eigendecomposition, $O(d^3)$, since the gain needs the eigenbasis. The total for the soft-min arm is $O\big(B(d^3 + |\mathcal{C}|d^2)\big)$. At this paper's $d = 512$ the eigendecomposition is negligible beside the pool pass; at an LLM embedding width of $d = 4096$ it is $\approx 7 \times 10^{10}$ flops per step, a few seconds of CPU, so the loop stays practical to budgets in the low thousands. Past that the fix is to compute only the k smallest eigenpairs by Lanczos, which is all the weights p_k depend on materially. The lazy-greedy variant of line 4 is valid only for the log-determinant objective, for the reason given after Eq. 11.

D THEORY AGAINST MEASUREMENT

Table 3 and Figure 4 check each quantity the theory predicts against its measurement on rung 1, where ϕ is one-hot and every predicted value is known in closed form. Two conventions matter for reading them. First, a prediction is marked as matched only if the measurement falls inside the regime the statement is scoped to; A3 is a Bradley–Terry temperature sweep included precisely to locate where the conditioning constant κ of Eq. 4 makes Theorem 3 loose, so its degraded entries are the intended result of that column rather than failures we are conceding. Second, the log-log slope in R_ϕ is reported both restricted and unrestricted, because Theorem 3 comes from a quadratic expansion at $\theta^\star$ and therefore describes only the cells where that expansion is accurate: over the $10{,}416$ of $27{,}648$ node-points meeting that criterion the fitted slope is 0.504 against a predicted 0.500, while the unrestricted fit over all cells is 0.703. The restriction is a scope condition of the theorem, not a filter chosen after seeing the fit, and the convergence of the slope as the restriction tightens (0.570 at mean absolute error below 1.0, 0.504 below 0.3) is itself the evidence that the scope condition is the right one. Any quotation of the matched slope elsewhere in the paper carries the restriction with it.

E PER-TOPOLOGY MEASUREMENTS AT MATCHED BUDGET

Table 4 gives the per-topology view behind Figure 4. Two readings should be kept separate. Across topologies, $\lambda_{\min}^+$ orders RMSE in the direction Theorem 3 predicts, with one exception: the prompt-

Table 3: Theory against measurement, rung 1, 8 seeds. Bold marks a measurement matching its prediction. A3 is a temperature sweep, included to mark where the bound loosens rather than as a confirmation; Theorem 4's bound is infinite exactly when nullity is positive, which holds in every cell. Appendix E gives the per-topology breakdown.

Claim	Predicted	A1	A2	A3
Thm 1: $\max \lvert\Delta\ell\rvert$ on $\mathrm{null}(L_\phi)$	exactly 0	$\mathbf{9.1\times10^{-13}}$	$\mathbf{3.6\times10^{-12}}$	$\mathbf{9.1\times10^{-13}}$
Thm 3: log-log slope in R_ϕ	0.500	**0.504**	**0.521**	0.475
Thm 3: rank corr. (R_ϕ, err)	> 0	**0.935**	**0.961**	0.160
Thm 3: pointwise coverage	high	**0.983**	**0.959**	0.627
Thm 3: interval coverage	0.95	**0.969**	**0.968**	0.719

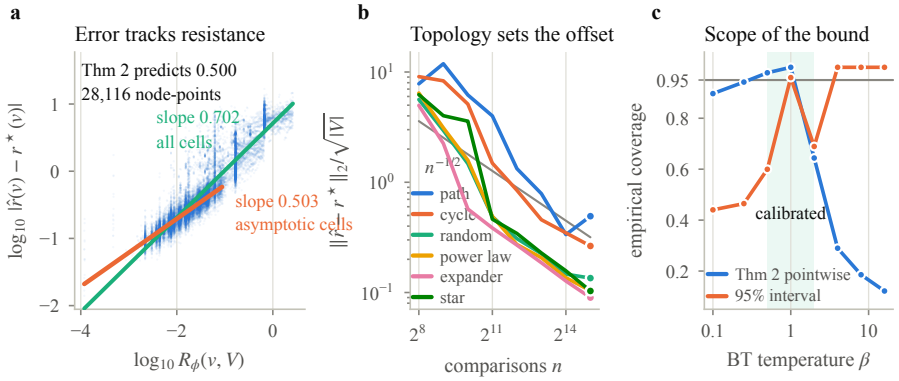

Figure 4: *(a)* Pointwise error against feature effective resistance, all 27,648 node-points from 9 topologies at matched budget. Restricted to the 10,416 points where the expansion behind Theorem 3 is accurate the fitted slope is 0.504 against a predicted 0.500; the unrestricted fit is 0.703. *(b)* Error against budget, six topologies at $|V| = 64$. *(c)* Coverage against Bradley–Terry temperature, calibrated near $\beta = 1$ and degrading as κ falls, which is the scope boundary of Theorem 3.

partitioned arm has the largest $\lambda^+_{\min}$ in the table and mid-table error, because $\lambda^+_{\min}$ is computed on the identifiable subspace and that arm has 32 unidentifiable directions, so a high $\lambda^+_{\min}$ there describes a smaller subspace. This is why nullity and $\lambda^+_{\min}$ must be read together, and why every report card in Section 5.3 lists both. Within a topology, the correlation between a vertex's resistance and its error is what Eq. 6 predicts, and it appears wherever the resistance has spread to correlate with.

F Head invariance, tangent geometry, and listwise labels

This appendix gives the detail behind Section 5.1. The instrument is a sign test. We construct an instance carrying a planted direction $u \in \mathrm{null}(L_\phi)$, fit each head at planted weight $w = +1$ and again at $w = -1$, and compare the coefficients recovered on u. The flip changes $r^\star$ but leaves the distribution of every observable comparison unchanged, so a head that extracts *no* information about u cannot move its coefficient, while one extracting any information must. The test is therefore for *change under the flip*, not for a coefficient of zero: a nonzero but identical coefficient at both signs is basis-misalignment leakage from the projection, which is why the projection basis is canonicalised before the comparison is made. All four head classes report relative change 0.0000, with win counts and fitted rewards bit-identical across the flip (Figure 5a). Capacity is ruled out separately: every head recovers the *identifiable* component, the smallest such coefficient being 0.854, so these heads can fit and simply have nothing to fit along u.

Panel b addresses the objection that a fine-tuned model changes its own feature map. Under the MLP's neural tangent map (Jacot et al., 2018) the effective dimension rises from 12 to 897 and the nullity to 861; the likelihood is flat along $\mathrm{null}(L_\psi)$ to 1.2×10^{-11} and every direction in that null space is prompt-constant to 3.3×10^{-15}. A richer map moves *which* subspace is unlearnable without making the data say more, and Corollary 2 survives the change of geometry that shallow fine-tuning induces. Panel c is the listwise comparison analysed in Appendix K: Plackett–Luce annotations share the null space exactly, at a maximum principal angle of $0.00°$, and buy $4.19\times$ the conditioning per annotation ($0.092 \to 0.385$) rather than any new direction.

G Selection measurements

Figure 6 carries the three measurements behind Section 5.5. Panel a is the label-efficiency result in collect-plus-repair mode: soft-min SPEC reaches the target $\lambda^+_{\min}$ in 964 labels against random's

Table 4: Ablation A1 at $|V| = 128$, $n = 8192$, ϕ one-hot so $d = |V|$, 8 seeds, medians. Nullity equals the component count, which is Corollary 2. The within-topology rank correlation is interpretable only where the resistance spread exceeds 1: on a vertex-transitive graph every vertex has the same resistance by symmetry, so the near-zero entries for cycle, complete, star and expander are correct rather than a failure. Interval coverage is undefined where some direction has infinite standard error ("n/a").

Topology	$\lambda^+_{\min}$	nullity	median R_ϕ	RMSE	$\rho(\text{err}, R_\phi)$	R_ϕ spread	interval cov.
prompt-partitioned	168.000	32	0.0044	0.478	0.105	1.02	n/a
complete	128.000	1	0.0077	0.212	0.009	1.02	0.938
star	64.000	1	0.0153	0.674	0.045	1.02	0.984
expander	20.183	1	0.0112	0.259	0.055	1.07	0.953
power law	19.365	1	0.0161	0.316	0.625	3.18	0.949
random	9.445	2	0.0114	0.368	0.766	4.15	0.969
clustered	0.472	6	0.0070	0.420	0.747	12.20	n/a
cycle	0.154	1	0.1667	0.972	−0.000	1.00	1.000
path	0.039	1	0.2893	2.643	0.802	3.32	1.000

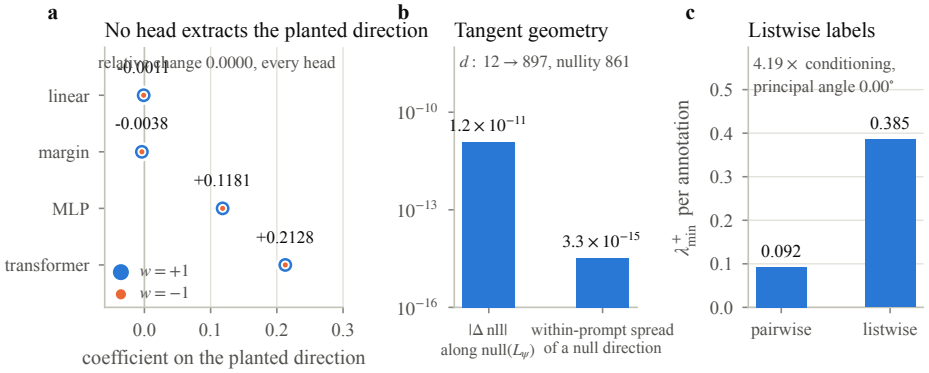

Figure 5: *(a)* Sign test on a planted unidentifiable direction with a canonicalised projection basis. Flipping the planted weight's sign changes $r^\star$ but not any observable comparison, and all four head classes report identical coefficients at both signs, so the large and small markers coincide. A nonzero coefficient is basis-misalignment leakage, not information: the test is for change under the flip. *(b)* The same instance under the MLP's neural tangent map ($d : 12 \to 897$, nullity 861), where the likelihood is flat along $\text{null}(L_\psi)$ and every null direction is prompt-constant, both to machine precision. *(c)* Plackett–Luce annotations share the null space exactly (maximum principal angle 0.00°); what they buy is $4.19\times$ the conditioning per annotation.

2048, with A-optimal at 1502, log-det at 1779 and V-optimal at 1789, while uncertainty, margin, coreset and the curation heuristic do not reach the target at any budget in the sweep. The distinction matters for how the result is read: those four arms are not slower at raising conditioning, they do not raise it to the threshold at all, because none of them optimises a spectral quantity. The $2.13\times$ headline is the soft-min row divided by random and is grid-dependent, taking values $2.13\times$ here, $2.55\times$ on the budget sweep and $4.51\times$ repair-only; we quote the smallest.

Panel b is Theorem 6 against measurement and is the uncomfortable one. The objective carrying the proof is the weakest in realised gain: log-det has 0 diminishing-returns violations over 150 nested pairs and a greedy-to-brute-force ratio of 1.0000 against the $(1 - 1/e) = 0.632$ floor, yet gains only $1.09\times$ over random; the soft-min gains $3.26\times$ and violates diminishing returns in 105 of 150 pairs, so it carries no guarantee; greedy on $\lambda^+_{\min}$ itself reaches a ratio of 0.0395 and gains $0.44\times$, that is, loses. We report the ordering rather than resolving it, since the mismatch between what is provable and what works is the honest state of the design problem here. Panel c is ablation D6, the stacking result: curation alone lowers $\lambda^+_{\min}$ from 687 to 484 and raises RMSE from 0.2625 to 0.2926, while selection applied after curation reaches 804 and 0.2525, beating both single arms. Curation and design are therefore not interchangeable and compose in one order only.

H REWARDBENCH, PER-BUDGET DETAIL

Protocol. The benchmark (Lambert et al., 2025) is used in-domain rather than as a transfer test: we split its edges, supervise on one half and evaluate on the held-out half (1493 pairs, standard error 0.0129), pairing per seed and matching labels across arms, over 8 seeds. This isolates the selection rule, which is the only quantity we claim. It also means the numbers are not comparable to published RewardBench scores for trained reward models, and are not offered as such; a linear head

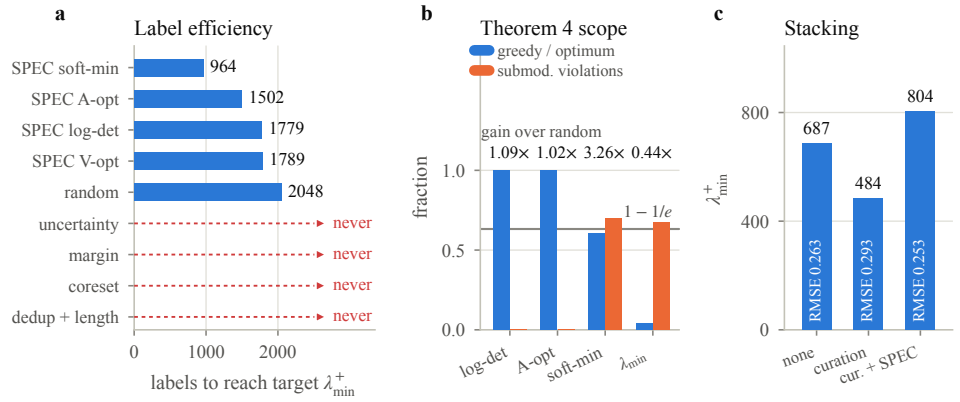

Figure 6: *(a)* Labels to reach a target $\lambda^+_{\min}$, collect-plus-repair, 8 seeds; four baselines never reach it. *(b)* Theorem 6 against measurement, from 150 nested pairs and brute-force enumeration on a small pool. The gain over random inverts the ordering: the objective carrying the proof gains $1.09\times$, the one gaining $3.26\times$ carries none. *(c)* Ablation D6: curation alone lowers $\lambda^+_{\min}$ by 30% and raises RMSE, while selection applied afterwards beats both single arms.

Table 5: Selection at matched labels. *Left*: labels needed to reach target conditioning (collect-plus-repair, 8 seeds). *Right*: held-out RewardBench accuracy at 800 labels and gain over random (8 seeds, 1493 pairs, SE $= 0.0129$). Bold denotes the best per column; coreset wins on accuracy but misses the conditioning target, while uncertainty underperforms random. The 1.7–2.3 point gains are 1.3–1.8 unpaired standard errors, so the accuracy ordering rests on the paired per-seed test rather than on these margins; the conditioning column spans a factor of 41 and is the quantity the theory predicts.

	Labels to	RewardBench	Gain at 800 labels	
Selection rule	target $\lambda^+_{\min}$	accuracy	vs random	seeds won
SPEC soft-min	**963.5**	0.5828	+1.86	7/8
SPEC A-optimal	1502	n/a	n/a	n/a
SPEC log-det	1779	0.5789	+1.68	6/8
SPEC V-optimal	1789	n/a	n/a	n/a
random	2048	0.5685	n/a	n/a
coreset (Sener & Savarese, 2018)	never	**0.5853**	**+2.27**	7/8
margin	never	0.5770	+0.62	5/8
uncertainty (Lewis & Gale, 1994)	never	0.5467	−0.80	3/8
dedup + length-debias	never	n/a	n/a	n/a

on frozen embeddings reaches 0.58 to 0.60 here against above 0.9 for a trained model, so this is a matched-budget ordering experiment in a regime well below deployment.

What the panels show. Panel a gives held-out accuracy at 800 labels: coreset 0.585, soft-min SPEC 0.583, log-det SPEC 0.579, margin 0.577, random 0.569, uncertainty 0.539. Panel b gives the training design's $\lambda^+_{\min}$ in the same order, 0.0470, 0.0682, 0.0697, 0.0076, 0.0038 and 0.0017, which orders the selectors almost as accuracy does and spans a factor of 41 where accuracy spans 4.6 points. The conditioning gap is the larger and more robust signal and is the quantity Theorem 3 actually predicts; the accuracy gap is what a practitioner cares about but is close to the noise floor, and we read the two together rather than resting on the second alone. Panel c gives the paired gain over random by budget, with filled markers at $|t| \geq 2$.

How much the accuracy differences support. At the 0.0129 standard error of the evaluation split, the +1.7 to +2.3 point gains of the coverage arms are 1.3 to 1.8 unpaired standard errors. The paired per-seed test is stronger than that comparison suggests, since arms share seeds and evaluation pairs, and the coverage arms win 6/8 and 7/8 seeds with one-sided $p < 0.05$ at every budget; but 8 seeds is thin for a two-point claim and we do not present these as established effect sizes. The two results we do rest on are the ones with margins well clear of the noise: uncertainty sampling losing 2.97 points at 200 labels with 0/8 seeds won and $t = -5.15$, and the conditioning ordering of panel b. The saturation reading is also robust in the opposite direction: at 1200 labels the budget takes 80% of the pool, every arm selects nearly the same edges, and all land in 0.592–0.600 with no gain over random significant, which is what the theory predicts once the design stops being under-determined.

I THE SYNTHETIC CONTROL

Five of the six corpora in Table 2 are public releases. The sixth is a preference set generated for this work rather than collected, included so that the audit has one row whose comparison graph was

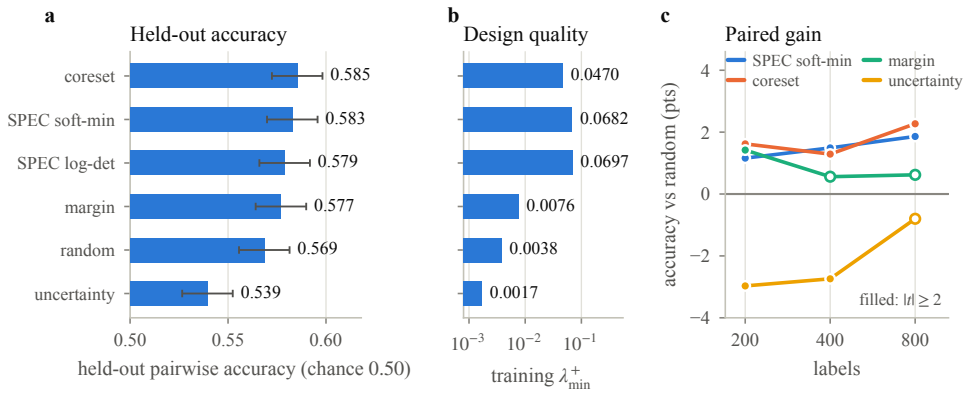

Figure 7: RewardBench (Lambert et al., 2025), in-domain protocol: supervise on a benchmark edge split, evaluate on the held-out half (1493 pairs, standard error 0.0129), paired per seed and label-matched, 8 seeds. *(a)* Held-out accuracy. *(b)* The training design's $\lambda^+_{\min}$, which orders the selectors almost as accuracy does. *(c)* Paired gain over random by budget; filled markers are $|t| \geq 2$.

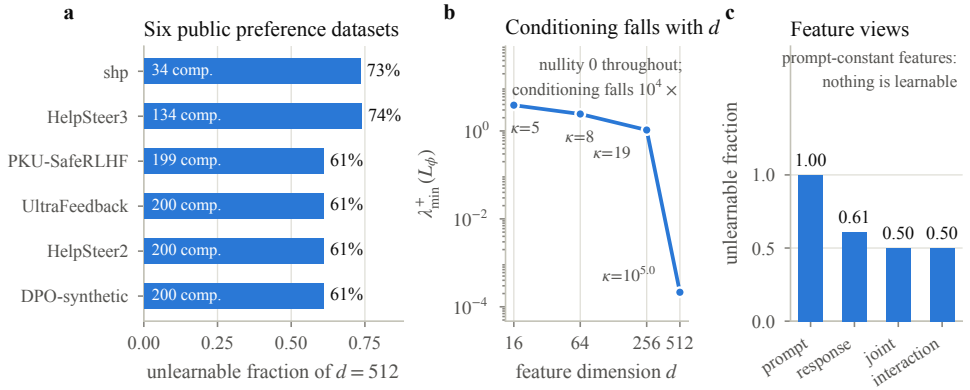

Figure 8: *(a)* Fraction of $d = 512$ directions in $\mathrm{null}(L_\phi)$ for six public datasets at $|V| \leq 400$, component counts in the bars. *(b)* UltraFeedback at $|V| = 4000$: nullity is 0 throughout because $d \ll |V| - c$, and the pathology appears as conditioning instead. *(c)* Ablation B4: components carried by prompt-constant features are entirely unidentifiable on all six datasets, which is Corollary 2 measured directly.

built deliberately rather than inherited. It is constructed at the operating point the public rows are subsampled to, and in the shape that RLHF collection produces: $|V| = 400$ prompt–response pairs carrying 200 comparisons, one chosen/rejected pair per prompt, which is why it has 200 components of size 2, a cross-prompt fraction of 0.000 and a degree Gini of 0.000.

Its role is to separate protocol from content. The two graph findings of Section 5.3 follow from pairing responses only within a prompt and from nothing else, so a corpus built to that protocol exhibits both whatever its text contains. This row is that corpus, and it matches the five public ones on every column the findings rest on. It is audited by the identical pipeline and reported alongside them for that reason; no claim depends on it, and removing the row leaves every statement in Section 5.3 intact, since each is quantified over the datasets that remain.

J AUDIT ROBUSTNESS AND CONFOUND CONTROLS

Exact identities. Four quantities have closed forms and are asserted against them rather than fitted. With ϕ one-hot, L_ϕ equals the ordinary graph Laplacian to 0.0 error. The factorisation $L_\phi = \Phi^\top B^\top W B \Phi$ agrees with the edge sum of Eq. 1 to 4.5×10^{-13}, and Eq. 3 agrees with an explicit pseudoinverse to 3.9×10^{-18}. Under a misspecified ϕ the approximation bias and the estimation variance are orthogonal, with Pythagorean residual 2.1×10^{-14}, which is what licenses reporting recovery numbers against the best-linear-predictor ceiling rather than against $r^\star$ itself.

Invariances the theory requires. Rank and nullity are invariant to the edge-weighting scheme: counts, uniform, annotator confidence and log-counts all give $200/312$ on UltraFeedback, as Theorem 1 requires, since strictly positive weights cannot change $\mathrm{null}(L_\phi)$. Only $\lambda^+_{\min}$ moves, and it moves sharply, from 0.0176 under counts to 0.0122 under log-counts and 0.0016 under confidence weighting, an $11\times$ penalty for confidence weighting, and $0.0659 \to 0.0123$ on SHP. Spectra are also

identical under prompt–response and response-level node definitions on all six datasets, because no response text is shared across prompts in these corpora, so the pathology is not an artefact of how a vertex is defined. Slicing SHP by subreddit (ranks 103, 177, 67 against nullities 409, 335, 445) and HelpSteer3 by domain (ranks 123, 136, 132, 120) leaves every slice rank-deficient, so it is not concentrated in one collection round; the other four datasets expose no round or timestamp field and are reported as not applicable rather than skipped silently. A dimensionality sweep from $d = 8$ to 512 gives nullity 0 at $d = 8$–64 on every dataset, which is why $d/|V|$ selects the regime and why every table names its operating point.

Confounds, and a proxy that had to be retired. Three corrections were needed before this measurement said anything, and we report all three because the third changes what Section 5.3 may claim.

First, inter-model disagreement is not an error proxy. It correlates with R_ϕ at $+0.008$ on a synthetic instance where Theorem 3 provably holds, so it cannot be evidence anywhere; the cause is that the heads share the graph and the labels, so they share the design's error and differ mainly in inductive bias, and their measured agreement is 0.825–0.949.

Second, the test has no power unless resistance varies. Spread is governed by $d/|V|$ rather than by topology or degree heterogeneity, all three of which we tested: the spread is $1.00\times$ at $d/|V| = 1.28$, $1.48\times$ at 0.16 and $5.70\times$ at 0.01, so with $d \gtrsim |V|$ the difference vectors are nearly orthogonal and equinormal, R_ϕ concentrates, and nothing can be detected either way. The driver refuses to report a correlation below a minimum spread for that reason. Note this is the *opposite* operating point from the nullity claim, which needs $d > |V| - c$; the two ablations answer different questions and are reported at different points.

Third, the parametric bootstrap that replaced the inter-model proxy has a scope condition of its own, and it is the reported number that has to change. Under this proxy on UltraFeedback at $|V| = 4000$ with Titan features, the raw and partial rank correlations after controlling for degree, length and cluster size are $+0.016/+0.024$ at $d = 16$ (null, over-determined), $-0.286/-0.208$ at $d = 64$, $-0.336/-0.260$ at $d = 256$ and $-0.234/-0.185$ at $d = 512$. Theorem 3 bounds error by $\sqrt{R_\phi}$ and so predicts a *positive* sign. A negative one is therefore not the link surviving a control, and an earlier version of this paper described it that way in error.

The proxy is what fails, and the failure has a mechanism. The bootstrap resamples $k_e \sim \mathrm{Binomial}(n_e, \sigma(\hat{z}_e))$ with the graph fixed and takes the per-node spread of the refit, which is the estimation variance Theorem 3 bounds, provided the resample can differ from what was observed. Every corpus in Table 2 has $n_e = 1$ and a fixed winner, so as the fit saturates, $\sigma(\hat{z}_e) \to 1$, the resampled labels are the observed ones with probability approaching one and the refit is the original fit. What the spread then tracks is $\sigma'(\hat{z}_e)$, how saturated the fit is at that edge, and saturation is highest exactly where leverage is *lowest*: a high-resistance edge is one the rest of the design barely constrains, so the fit can push its margin up at little cost elsewhere. Both channels are directly visible on this corpus, since edge resistance correlates with $|\hat{z}_e|$ at $+0.18$ to $+0.38$ and hence with $\sigma'(\hat{z}_e)$ at -0.18 to -0.38, so as saturation grows the proxy becomes anti-correlated with resistance by construction, whatever the theorem says.

Two controls separate this from the alternative reading, that Theorem 3 simply fails on real text. Holding the graph and the labels fixed and changing only the feature map to the offline word-hashing map, the same proxy returns $+0.252$ to $+0.323$, positive at every d, and the fit is plainly inside the logistic regime there, with a saturated-edge fraction of 0.000–0.017 and 0.26–0.45 of edges fitted with the wrong sign. The sign therefore moves with how saturated the fit is and not with the corpus, the graph or the labels. The same proxy is likewise positive, $+0.584$ at $d = 16$ and $+0.410$ at $d = 48$, on a synthetic instance whose labels are drawn from Bradley–Terry.

The corrected measurement (ablation F3, `experiments/e09_resistance_error.py`) drops the proxy and measures the quantity the theorem actually bounds, on the same graph, the same features and the same estimator. We plant $\theta^\star \in \mathrm{row}(L_\phi)$, scale it so the Bradley–Terry margins have unit spread, draw labels from the model, fit, and correlate the per-response error $|\hat{r}(v) - r^\star(v)|$, centred within each component, against $R_\phi(v, V)$. With the offline word-hashing map the rank correlation is $+0.171$ to $+0.358$ raw and $+0.164$ to $+0.332$ under the joint control, positive at every $d \in \{16, 64, 256, 512\}$ and at one, four and sixteen comparisons per edge. The driver reports the separation diagnostics beside every correlation, so the scope condition is visible rather than assumed, and an oracle bootstrap centred at $\sigma(z^\star)$ rather than $\sigma(\hat{z})$, which isolates the centring as the cause.

One structural caveat applies to all of these on four of the six corpora. With components of size 2, per-component centring puts the two vertices of an edge at equal and opposite error and equal resistance, so the vertex-level correlation is an edge-level one between $|\hat{z}_e - z_e^\star|$ and the edge's leverage $x_e^\top \big(L_\phi^{(n)}\big)^+ x_e$. That is a valid test of Eq. 6 and we report it as one, but it is not independent across the two endpoints and the effective sample size is the edge count, not the vertex count.

The full six-dataset audit run under the retired inter-model proxy gave raw correlations of $+0.018$ to -0.073 and partials of $+0.017$ to -0.079, all null; those runs record the proxy name in their summaries and are not quoted as evidence.

K TWO NEGATIVE RESULTS

Cross-corpus transfer fails. A head fitted on UltraFeedback scores 0.46 to 0.49 on RewardBench, below the 0.5 chance level, consistently, at every ridge from 10^{-6} to 10^{5}, over which $\|\theta\|$ falls from 601 to 0.000. This is a transfer failure rather than an orientation defect: the accuracy metric is correctly oriented, since a perfect ranker scores 1.0 and an inverted one 0.0, and both graphs use the same win-count convention; the identical pipeline reaches 1.0000 held-out accuracy in distribution; and an oracle fit on RewardBench itself reaches 0.9021, so ϕ can express benchmark preference. The mechanism is a polarity flip in a corpus-specific direction that length makes visible. UltraFeedback prefers the longer response 55.7% of the time and RewardBench 37.4%, so "prefer longer" scores 0.374 there, bracketing the observed range. Projecting log-length out of ϕ moves the score only from 0.4637 to 0.4670, so length is a symptom of a broader corpus-specific direction rather than the cause (Singhal et al., 2024). A reward model can therefore latch onto a corpus artefact that inverts on held-out data, and the comparison graph plus a length control makes this visible before any policy training (Skalse et al., 2022).

Listwise annotations do not restore identifiability. Collecting rankings of three or more responses instead of pairs is the natural response to Corollary 2, and it does not work. The Plackett–Luce (Plackett, 1975; Luce, 1959) Fisher information has the same null space as L_ϕ: identical nullity, and a maximum principal angle between the two null spaces of 0.00°. A direction constant within every prompt stays unidentifiable under listwise labels, because a ranking of responses to one prompt still yields only within-prompt contrasts. What rankings buy is conditioning per annotation, measured at $4.19\times$, since one ranking of k responses implies $k-1$ choice stages; the comparison is normalised by annotation count because the two arms otherwise spend different budgets. Ranking-based identification is therefore a special case inside this framework rather than an alternative to it. The same holds for the implicit reward of direct preference optimisation (Rafailov et al., 2023): $\beta(\log\pi - \log\pi_{\text{ref}})$ carries a per-prompt $\log Z(x)$ term, which is exactly Corollary 2's gauge, so within-prompt accuracy is invariant to per-prompt offsets and cross-prompt accuracy is not.

L SELECTION AND REPAIR, IN FULL

The main-text version of this subsection is compressed; the original is reproduced here.

Selection, repair, and where the objective is wrong If conditioning sets the constant in both the upper bound and the floor, a budget should be spent raising it. That is SPEC: given a starting Laplacian L_0 (zero when collecting, the audited L_ϕ when repairing) and a candidate pool, score each pair by its rank-one gain and buy the best: exactly for log-det and A-optimality, and to first order for the soft-min, whose gain is a directional derivative rather than a closed form (Appendix C, which also gives the $O\big(B(d^3+|\mathcal{C}|d^2)\big)$ cost of the soft-min arm). Appendix B gives the four objectives and Theorem 6, which separates the log-determinant, monotone and submodular so greedy attains $(1-1/e)$ (Nemhauser et al., 1978), from $\lambda_{\min}^+$, the soft-min, A- and V-optimality, which are not. Against random, uncertainty sampling (Lewis & Gale, 1994), margin, coreset selection (Sener & Savarese, 2018) and a curation heuristic at matched budget, soft-min SPEC reaches the target conditioning in 963.5 labels against random's 2048, and four baselines never reach it (Table 5; Appendices G–H give the measurements). The result worth carrying is negative: uncertainty sampling, the standard acquisition rule for preference data, is *worse* than random on RewardBench by 2.97 points at 200 labels with $0/8$ seeds won and $t = -5.15$, because label-entropy selection ignores feature coverage. Coverage rules help more weakly, and with 8 seeds against a standard error of 0.0129 we claim an ordering at matched labels rather than an effect size.

Repair isolates the axis Corollary 2 names This is also the test of Section 4.2. We plant a reward with a prompt-level component, add comparisons from a cross-prompt pool at budgets 8 to 256, and score two disjoint held-out pair sets (8 seeds). *Cross-prompt* accuracy rises from 0.633 to 0.98 while *within-prompt* accuracy stays within 0.938–0.956 at every budget for every rule, so on this instance

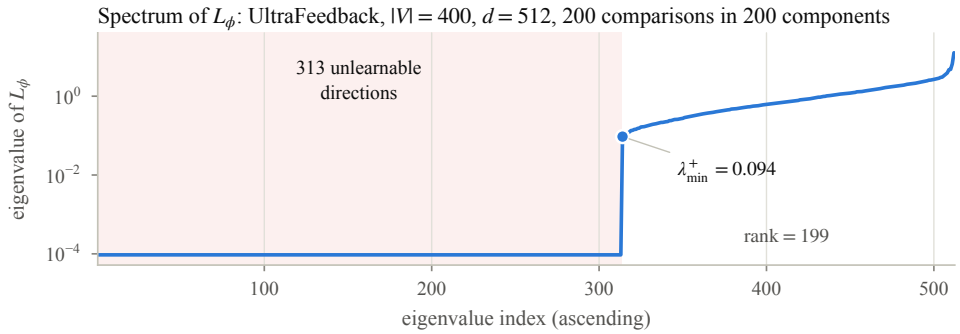

Figure 9: The object. On the real UltraFeedback graph ($|V| = 400$, $d = 512$), the offline word-hashing map yields 200 comparisons in 200 components. The shaded band is $\mathrm{null}(L_\phi)$: $313/512$ directions are zero, as Eq. 5 forces since $d > |V| - c$. Zeros are drawn at the axis floor for the log scale. The first nonzero eigenvalue is $\lambda^+_{\min} = 0.094$. This is the one configuration where the feature map costs rank beyond the budget: the rank is 199 against a budget of 200. Table 2 gives the Titan-feature version, where the bound is attained exactly.

the gauge is invisible to within-prompt decisions and decisive for the rest. It also refutes a claim we would have liked to make: selecting those comparisons to maximise $\lambda^+_{\min}$ is *worse* than random from the same pool, scoring 0.935 against 0.974 at budget 8 despite attaining $\lambda^+_{\min} = 1560$ against 183. This is an objective–metric mismatch, since $\lambda^+_{\min}$ prices the weakest eigendirection while cross-prompt accuracy averages over contrasts; for repair, a handful of anchors chosen any way restores identifiability. Two further results run against our own framing and Appendix K reports both: cross-corpus transfer fails, and listwise annotation, the obvious answer to Corollary 2, does not restore identifiability, since the Plackett–Luce (Plackett, 1975; Luce, 1959) Fisher information shares the null space exactly.

M Which RLHF decisions read the prompt-level gauge

The gauge stops being free where two *different* prompts are placed on one reward scale, and four standard practices do exactly that. *Best-of-N and rejection sampling* against a pooled threshold: the argmax within a prompt is gauge-free, but keeping every sample whose reward exceeds one global cutoff lets the per-prompt offset decide which prompts contribute training data. *Filtering and curricula by reward score*: dropping low-reward examples, ranking prompts by difficulty, and scheduling by score all compare across prompts. *A prompt-independent value baseline or reward-as-critic*: a value head regressed onto reward across prompts, or a single KL–reward coefficient tuned against absolute reward, is calibrated against a quantity the data does not determine. And *reported mean reward* is an average over a prompt set, hence a sum of per-prompt gauge constants plus signal, so two models can be ranked by it while differing only in how their regularisers filled $\mathrm{null}(L_\phi)$.

N The audit at the other end of the rank-budget ratio

The nullity column is mostly arithmetic, and what replaces it at scale The third column reports nullity 312 to 378 of $d = 512$, which is 61 to 74%, and that number is largely forced before any data is examined: Table 2 sits at $d/(|V| - c) = 2.56$, so Eq. 5 already guarantees nullity $\geq 512 - 200 = 312$ on the four datasets with 200 components, and $61\% = 1 - 200/512$ is that bound rather than a measurement. The stronger version also holds: with Titan features the rank equals $\min(d, |V| - c)$ *exactly* on all six datasets, so the column carries nothing beyond $|V|$ and the component count, and the percentage is the component count's arithmetic shadow. Production data sits at the other end of the ratio, where the two graph findings carry over unchanged because they do not mention d, the nullity finding does not, and what replaces it is conditioning. Rerunning the audit at $d/(|V| - c) \approx 0.008$ gives nullity 0 on all six datasets while the heterogeneity Theorem 3 prices grows sharply: the 90th to 5th percentile ratio of $R_\phi(v, V)$ is $8.4\times$ on SHP, $30.8\times$ on UltraFeedback and $35.4\times$ on HelpSteer2, against $1.0\times$ at the rank-binding point. Since a response's error is bounded by its own resistance, that is an 8 to $35\times$ spread in the error floor within one corpus at one budget. Scale changes which row of Table 1 binds, not whether one does.

O The prompt-level gauge once the rank budget stops binding

Ablation F4 in Section 5.4 needs $\mathrm{null}(L_\phi) \neq \{0\}$: it moves the reward along a direction the data is exactly silent about. Production-scale preference data sits at the other end of the rank-budget ratio of Eq. 5, where L_ϕ has full rank and no such direction exists. This appendix asks the same question there.

Rerunning the driver at $|V|$ =4000 on the same corpus gives 2000 components and nullity 0 at both d, so the exact construction correctly declines to run. What replaces it is the *weak* directions, which Theorem 4(b,c) prices. We take the span of the k smallest eigendirections of L_ϕ, draw a direction inside it, and scale until the log-likelihood drop equals the 95% threshold of a χ^2_k likelihood-ratio test, so every reward in Table 6 is one the data cannot reject against the fit at that level, and the budget is the honest one for k free parameters rather than for one. Eight draws per cell.

Table 6: F4 at nullity 0. Each row is a reward the data cannot reject at the 95% level, differing from the fit only inside the span of the k weakest eigendirections of L_ϕ. Agreement is with the fitted reward; the pooled and threshold columns are Jaccard overlaps. The separation of Section 4.2 survives as an ordering, with within-prompt above every pooled column in every row, while ceasing to be an identity.

d	k	λ_k	within-prompt	pooled top-10%	pooled top-25%	threshold	prompt-mean ρ
512	1	0.676	0.980	0.876	0.908	0.907	+0.993
	8	0.774	0.955	0.781	0.842	0.840	+0.978
	64	1.23	0.901	0.601	0.682	0.684	+0.911
	256	3.02	0.827	0.465	0.565	0.570	+0.836
64	1	4.23	0.939	0.739	0.799	0.801	+0.969
	8	5.19	0.879	0.563	0.660	0.661	+0.901
	64	105	0.768	0.340	0.457	0.464	+0.717

Three things to read off it. The separation persists: within-prompt agreement exceeds every pooled column in every row, by 0.10 at $d = 512$ and 0.20 at $d = 64$ in the worst case. It stops being exact: within-prompt agreement is no longer 1, because a near-null direction is only approximately prompt-constant, which is the honest statement of what closing the null space buys. And the damage grows with k at fixed statistical cost, from 0.876 at $k = 1$ to 0.465 at k =256 on pooled top-10%, which is Theorem 4(b,c)'s claim that k weak directions cost more than one, measured rather than assumed. The rank-deficient regime is where the statement of Section 5.4 is exact, not where it is true.

P ABLATION INDEX

Twenty-five ablations were run. **Theory validity**: error against effective resistance across nine topologies at matched budget and matched total edge weight (A1); rate in n with constants tracking $1/\sqrt{\lambda^+_{\min}}$ (A2); the Bradley–Terry temperature sweep locating where the conditioning constant makes the bound loose (A3); feature dimension and misspecification, with the bias–variance split orthogonal to a residual of 2.1×10^{-14} (A4). **Embedding robustness**: three feature maps from two model families plus an offline map, compared by resistance profile and by principal angles between small-eigenvalue subspaces (B1); a random-projection control that classifies four of six datasets as feature-coupled and SHP and HelpSteer3 as topology-dominated, tracking their degree heterogeneity (B2); a dimensionality sweep from 8 to 512 (B3); prompt, response, joint and interaction feature views (B4). **Audit robustness**: four edge-weighting schemes with and without deduplication (C1); prompt–response against response-level granularity (C2); bootstrap subsampling at three fractions (C3); temporal and source slices (C4). SPEC: objective (D1), budget curves (D2), candidate-pool construction (D3), an annotator-noise sweep in which resistance-greedy selection lands between random and soft-min rather than collapsing (D4), repair against collect (D5), and stacking with standard curation (D6). **Downstream**: head class (E1); best-of-N with a judge scored at 0.000 position bias on a position-swap control, and the implicit reward of direct preference optimisation (E2); a per-category breakdown using the benchmark's own subset labels (E3). **Negative controls**: healthy graphs, where the predicted SPEC gain is small and the measured gain is 1.00 to 1.12× (F1); confounds for degree, length and cluster size (F2); resistance against error measured on a real corpus with a planted reward, with the separation diagnostics that bound when the observable bootstrap proxy of F2 may be read at all (F3, Appendix J); and the prompt-level gauge priced on that same corpus, where two rewards differing only inside $\mathrm{null}(L_\phi)$ leave every within-prompt decision exactly unchanged while a pooled top-10% selection overlaps by 0.385, together with its production-scale counterpart at nullity 0, where the separation persists as an ordering rather than an identity (F4, Section 5.4).