IDENTIFYING REWARD FACTORS FOR SPARSE PREFERENCE ELICITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized reward models represent each user's reward as a weighted combination of shared factors. Predictive accuracy alone does not give these factors a stable meaning: an invertible recombination of the factors, offset by a corresponding change in the weights, preserves every prediction while changing the meaning of each weight. We study which factors a population's comparison data identify and how many comparisons are needed to recover a new user's sparse composition once the dictionary is fixed. When each reward is a positive scale times a simplex composition, a population with a specialist for every factor generates a reward cone whose extreme rays are exactly the factors. The dictionary and user compositions are then identifiable up to permutation within the anchored model class. Without anchors, the extreme rays correspond to extreme population mixtures and may exclude sparser newcomers. For a fixed dictionary, the simplex constraint makes the sign of a realizable group contrast test the user's total weight on that group. The decision is independent of the user's scale, although its reliability depends on that scale. CS-Elicit recovers an s-sparse support by testing $O(s \log K)$ tree nodes, using $O\big(s \log K \cdot b^{-2} \log(\log K/(\Delta_\delta))\big)$ noisy comparisons at answer bias $b > 0$. This count matches a counting lower bound up to logarithmic factors, and the weight estimation rate is independent of K. Across language and control tasks, extraction improves dictionary stability across independent data draws. On PersonalLLM, extracting factors from LoRe utilities raises the cross-draw dictionary correlation from 0.42 to 0.63. In the E4 scaling experiment, CS-Elicit reaches 93% held-out accuracy for 1-sparse users with an extracted $K = 10$ dictionary in 10.1 questions, compared with 26.1 for mutual information selection.

1 INTRODUCTION

Personalized alignment aims to make language models responsive to users with different priorities (Sorensen et al., 2024; Kirk et al., 2024). Personalized reward models represent these differences with K shared reward factors and weights specific to each user, allowing adaptation from a small number of comparisons (Poddar et al., 2024; Chen et al., 2025; Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Lee et al., 2026). Accurate preference prediction does not determine the factor coordinates. An invertible change of basis, offset by a corresponding change in the weights, preserves all predictions while changing the meaning of each coordinate. This ambiguity matters when the factors guide elicitation, explain predictions, or let users edit their preferences.

We ask which reward factors a population's comparisons identify and how many questions reveal the factors relevant to a new user. We focus on users whose rewards depend on a small subset of the shared factors. For example, a user may value brevity and caution while placing no weight on the remaining factors. Such sparsity can reduce the cost of elicitation, but sparsity depends on the coordinate system. The factor basis must therefore be identified before sparse preferences can have a consistent meaning across equivalent reward models.

We model each user's reward as a positive scale times a simplex weighted combination of centered, unit norm, linearly independent factors. Exact comparison kernels determine a cone of population rewards. If the population contains a specialist for every factor, the cone's extreme rays recover the dictionary up to relabeling within the anchored model class. This result applies the separability

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.

principle of nonnegative factorization to comparison behavior (Donoho & Stodden, 2003; Arora et al., 2012). Without anchors, a cone with exactly K extreme rays yields a dictionary formed by the rewards of extreme users. If every extreme user has a dense composition, this dictionary cannot represent any strictly sparse newcomer. Under the same condition of K extreme rays, specialists that assign at least $1 - \nu$ weight to their respective factors yield canonical factors within 4ν in L^2 when $\nu < 1/2$.

For a fixed dictionary, simplex normalization enables group tests whose decisions are independent of user scale. A comparison with factor contrast $\gamma(\mathbf{1}_G - c\mathbf{1})$ has preference probability above one half exactly when the user's total weight on G exceeds c. Personal scale affects only the reliability of the answer. A computable margin extends this certificate to general contrasts. CS-Elicit traverses a dyadic partition and finds an s-factor support with $O(s \log K)$ group tests, each based on repeated noisy comparisons. At a fixed positive answer bias, the resulting comparison count matches a counting lower bound up to logarithmic factors. Additional probes estimate the active weights at a rate independent of K. When the rewards of extreme users are known exactly, recovery also holds with impure anchors if the minimum active weight is large enough relative to the impurity and probe error. The analysis does not bound extraction error from finite samples.

Our experiments test the stability of extracted factors, compare acquisition rules with a shared dictionary and estimator, evaluate complete personalization pipelines, and measure how the required number of questions changes with dictionary size. Extraction improves cross-draw dictionary correlation on both language and control tasks, although its effect varies across other perturbations. CS-Elicit gives the highest accuracy for highly sparse users when the response pool contains informative probes, and its advantage shrinks as probe margins weaken.

Contributions. (i) We show how population anchors identify reward coordinates and why extraction without anchors can exclude sparse newcomers (Theorem 1; Proposition 1). On PersonalLLM, extracting factors from LoRe utilities raises the cross-draw dictionary correlation from 0.42 to 0.63.

(ii) We introduce CS-Elicit, which recovers an s-factor support with $O(s \log K)$ group tests, or $O(s \log K \cdot b^{-2} \log(\log K/(\Delta_- \delta)))$ noisy comparisons at answer bias b. It estimates the active weights at a rate independent of K, and its comparison count matches the counting lower bound up to logarithmic factors (Theorem 2).

(iii) Assuming that the rewards of extreme users are known exactly, we bound how anchor impurity changes a new user's coordinates and probe margins (Theorem 3). After 16 answers, our extraction and CS-Elicit pipeline reaches accuracies of 0.635, 0.896, and 0.923 on UF-P-4, Gym, and Kitchen. These values exceed the best pooled or personalized baseline by 0.043, 0.027, and 0.066, respectively.

2 Related Work

Personalized reward models. Models based on latent users (Poddar et al., 2024; Kim & Kim, 2026), prototypes (Chen et al., 2025), and reward factors (Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Lee et al., 2026) share information across a population to adapt from a few comparisons. Other approaches use principal components of preference embeddings (Luo et al., 2025) or train interpretable attributes and sparse experts (Li et al., 2025; Wang et al., 2026). Personalization, however, does not consistently outperform pooled models (Rezk et al., 2025). Theoretical work on heterogeneous preferences studies welfare, aggregation, and regret (Chakraborty et al., 2024; Park et al., 2024; Zhong et al., 2024); Kang (2026) characterize the span of user heads required for constant online regret. We study when the factor basis within a reward subspace is identified, so that sparse preferences have consistent coordinates for elicitation and editing.

Identifiability of nonnegative factorizations. Separability identifies factors from pure rows (Donoho & Stodden, 2003). This principle supports topic models (Arora et al., 2012; 2013), endmember extraction (Winter, 1999; Nascimento & Dias, 2005), and algorithms based on projections or conical hulls (Gillis & Vavasis, 2013; Kumar et al., 2013). Minimum volume methods replace separability with sufficient scattering on a normalized affine slice (Craig, 1994; Huang et al., 2013; Lin et al., 2015; Fu et al., 2018). Because user scales are unknown, the observed rewards in our setting have a conic geometry. Applying minimum volume identification would require choosing a

slice and imposing a scattering condition on its loadings. We instead use anchors to recover extreme rays constructively (Theorem 1) and derive explicit bounds for anchor impurity (Proposition 3). These results complement robustness guarantees for additive noise (Gillis & Vavasis, 2013).

Preference elicitation. Active preference learning chooses comparisons using posterior volume or experimental design criteria (Sadigh et al., 2017; Bıyık et al., 2020; Mukherjee et al., 2024; Qian et al., 2015; Ge et al., 2024). Methods for language models include binary search over reward weights (Oh et al., 2025), uncertainty based selection over user coefficients (Shenfeld et al., 2025), and best arm identification with logistic bandits (Pădurean et al., 2026). Our method first recovers the sparse support and then estimates its weights. Sequential thresholding of individual coordinates requires $\Omega(K)$ tests (Malloy & Nowak, 2014), while group tests can depend logarithmically on K (Dorfman, 1943; Hwang, 1972; Aldridge et al., 2019). The simplex constraint lets us implement group tests with comparisons whose sign is invariant to user scale. We prove upper and lower bounds for recovering the support. One bit compressed sensing also recovers sparse vectors from dense measurements (Boufounos & Baraniuk, 2008; Plan & Vershynin, 2013). Table 4 in Appendix E summarizes these connections.

3 Identification of Compositional Reward Models

Let $\mathcal{T}$ denote the item space, μ a reference distribution on $\mathcal{T}$, and I the population index set. For a finite population, $I = \{1, \dots, N\}$. We represent the reward of user $i \in I$ as

$$r_i(\tau) = w_i^\top f(\tau) = \beta_i \, a_i^\top f(\tau), \quad \beta_i = \|w_i\|_1 > 0, \quad a_i = w_i/\beta_i \in \Delta_{K-1},$$

Here $\Delta_{K-1} := \{a \in \mathbb{R}^K_{\geq 0} : \mathbf{1}^\top a = 1\}$, and $f = (f_1, \dots, f_K)^\top$ is a shared dictionary in $L^2(\mu)$ with $K \geq 2$. The vector w_i is the user's loading, β_i is the user's comparison scale, and a_i is the user's factor composition. The scale is unknown and may differ across users.

Assumption 1 (Normalized reward model)**.** *The reward factors are centered, have unit norm, and are linearly independent in $L^2(\mu)$: $\mathbb{E}_{\tau\sim\mu}[f_k(\tau)] = 0$ and $\|f_k\|_{L^2(\mu)} = 1$ for $k = 1, \dots, K$. Every loading w_i is nonnegative, and the loadings span the coefficient space $\mathbb{R}^K$.*

For user i, the observed comparisons are $\mathcal{D}_i = \{(\tau_{it}^+, \tau_{it}^-, y_{it})\}_{t=1}^{n_i}$, where $y_{it} = 1$ indicates that the first item is preferred. The Bradley–Terry observation model is $\mathbb{P}(y_{it} = 1 \mid i, \tau_{it}^+, \tau_{it}^-) = \sigma\big(w_i^\top [f(\tau_{it}^+) - f(\tau_{it}^-)]\big)$, $\sigma(z) = \frac{1}{1+e^{-z}}$. Answers are conditionally independent. Given all previous queries and answers, the current answer follows this model; asking the same pair again produces a fresh, independent answer.

3.1 Identification from comparison behavior

Define user i's comparison kernel as $P_i(\tau, \tau') = \sigma(r_i(\tau) - r_i(\tau'))$. Let $\mathcal{M}$ contain the parameterizations $\theta = (f, \{w_i\}_{i\in I})$ that satisfy Assumption 1, and let $\mathsf{P}(\theta) = (P_i)_{i\in I}$ denote their indexed kernels. A dictionary is identified within a subclass of $\mathcal{M}$ if $\theta \mapsto \mathsf{P}(\theta)$ is injective on that subclass up to a common permutation of factors.

Definition 1 (Population anchors)**.** *The anchor class is $\mathcal{M}^\star = \{\theta \in \mathcal{M} : e_k \in \mathrm{cl}\{a_i : i \in I\}$ for every $k \in [K]\}$. For a finite population, this condition requires a user i_k with $a_{i_k} = e_k$ for every factor k, and hence $r_{i_k} = \beta_{i_k} f_k$. We call i_k an anchor, or a specialist for factor k. A population is ν-anchored if, for every k, some user satisfies $a_{ik} \geq 1 - \nu$.*

Theorem 1 (Identification of the factor dictionary from comparison behavior)**.** *Let $\theta = (f, \{w_i\}_{i\in I}) \in \mathcal{M}$, $\mathcal{D} = \mathrm{cl\,cone}\{r_i : i \in I\} \subseteq L^2(\mu)$, $\mathcal{M}^\star$ be the anchor class in Definition 1, and $A = \mathrm{cl}\{a_i : i \in I\} \subseteq \Delta_{K-1}$.*

(i) Anchored populations. *If $\theta \in \mathcal{M}^\star$, then $\mathcal{D} = \mathrm{cone}\{f_1, \dots, f_K\}$. Any $\widetilde{\theta} \in \mathcal{M}^\star$ with $\mathsf{P}(\widetilde{\theta}) = \mathsf{P}(\theta)$ satisfies $\widetilde{f} = Pf$ and $\widetilde{w}_i = Pw_i$ for some permutation matrix P. Therefore, the factors and compositions are identified within $\mathcal{M}^\star$ up to relabeling, and each scale β_i is identified.*

(ii) Unanchored populations. *A parameterization in $\mathcal{M}^\star$ with the same kernels exists if and only if $\mathrm{conv}\, A$ has exactly K extreme points $\rho_1, \dots, \rho_K$. Its factors are $g_j = \rho_j^\top f / \|\rho_j^\top f\|_{L^2(\mu)}$, and it is unique within $\mathcal{M}^\star$ up to relabeling. If $\mathrm{conv}\, A$ has more than K extreme points, no anchored parameterization with K factors reproduces the kernels.*

(iii) Approximate anchors. *Suppose $\mathcal{D}$ has exactly K extreme rays and $\theta, \widetilde{\theta} \in \mathcal{M}$ have the same choice kernels. If each parameterization is ν-anchored in its own coordinates with $\nu < 1/2$, then the canonical factors in (ii) can be labeled so that $\|g_k - f_k\|_{L^2(\mu)} \le 4\nu$ for every k. It follows that $d_{\mathrm{perm}}(f, \widetilde{f}) := \min_\pi \max_k \|f_k - \widetilde{f}_{\pi(k)}\|_{L^2(\mu)} \le 8\nu$. For $\nu \le 1/20$, Proposition 3(b) gives the corresponding bounds on composition and scale.*

Theorem 1 characterizes the extreme rays of the observable cone $\mathcal{D}$. They are the reward factors for an anchored population and the rewards of the most extreme users for an unanchored population. Part (ii) shows that comparisons reveal whether an anchored explanation exists, which occurs when $\mathcal{D}$ has K extreme rays. The comparisons alone do not establish that this explanation generated the data. Section 4 shows that the resulting dictionary excludes new users who are sparser than the extreme users. Part (iii) bounds dictionary error in $L^2(\mu)$. Assumption 3 adds a separate pointwise error bound for adaptively selected comparisons.

3.2 Reward dictionary extraction from finite samples

Algorithm 2 in Appendix B.1 fits user rewards by ridge-logistic regression in a common centered feature space containing the factors. It normalizes nonzero rewards, merges near-duplicate directions, and screens for extreme rays using nonnegative projection residuals under a common $L^2(\mu)$ inner product. It then selects K independent directions and refits nonnegative user loadings. With exact rewards and anchors, screening recovers the factor directions by Theorem 1(i). A recovery guarantee from finite samples remains open, so Theorem 3 analyzes the version with exact rewards.

The inputs can be recorded comparisons or comparison logits from a trained personalized reward model. With exact access, centering and integrating the logits recovers user rewards (Lemma 2), connecting extraction to existing personalized models (Poddar et al., 2024; Bose et al., 2025; Shenfeld et al., 2025). Misspecified features and estimation bias can distort this geometry before extraction.

4 Sparse Elicitation of a New User's Composition

Once the dictionary is fixed, we recover the sparse support by asking whether each group G contains an active factor.

A new user has scale $\beta > 0$ and composition $a^\star \in \Delta_{K-1}$. Sparsity depends on the factor coordinates. For a simplicial observed cone, Proposition 1 characterizes the newcomers that remain representable after canonical extreme ray extraction. Proposition 2 in Appendix D shows that support, margin, and scale are invariant across anchored parameterizations.

Proposition 1 (Coverage of sparse newcomers)**.** *Under Assumption 1, suppose* $\mathrm{conv}\, A$ *has exactly K extreme points $\rho_1, \dots, \rho_K$, with extracted factors g_j as in Theorem 1(ii). A new composition $a^\star$ with support S is representable by nonnegative loadings on g if and only if*

$$a^\star \in \mathrm{cone}\{\rho_j : \mathrm{supp}\, \rho_j \subseteq S\}.$$

Thus no newcomer with fewer active factors than every ρ_j is representable. All simplex compositions are representable if and only if the population is anchored.

A dictionary formed from extreme users therefore cannot represent a newcomer who is sparser than every one of them. Without this coverage, the extracted coordinates may fail to form a nonnegative composition, so they cannot support the interpretation of group mass used below.

4.1 Group probes with decisions independent of scale

To recover sparse support, we seek comparisons that distinguish zero group mass from mass at least Δ_-.

Assumption 2 (Sparse composition with margin and a lower bound on scale)**.** *For an integer $1 \le s \le K$, the new user's composition $a^\star \in \Delta_{K-1}$ satisfies $\|a^\star\|_0 \le s$. Let $S := \mathrm{supp}(a^\star)$. The margin $\Delta := \min_{k \in S} a^\star_k$ satisfies $\Delta \ge \Delta_- > 0$, and the scale satisfies $\beta \ge \beta_- > 0$.*

The bounds Δ_- and β_- specify detection resolution and a minimum response scale; the actual values remain unknown. For $x = f(\tau) - f(\tau')$, the following margin certifies separation between $m_G = 0$ and $m_G \ge \Delta_-$, where $m_G = \sum_{k \in G} a_k$.

Definition 2 (Valid group probe and its margin). *For a nonempty proper group $\emptyset \neq G \subsetneq [K]$, a margin level $\Delta_- > 0$, and a contrast vector $x \in \mathbb{R}^K$, define*

$$\vartheta(x; G, \Delta_-) = \min\Big\{ -\max_{k \notin G} x_k, \;\; \Delta_- \min_{k \in G} x_k + (1 - \Delta_-) \min_{k \notin G} x_k \Big\}. \tag{1}$$

The contrast x is a valid probe for (G, Δ_-) if $\vartheta(x; G, \Delta_-) > 0$. The first term certifies that G is absent from the support, and the second certifies that G is present under the least favorable admissible composition. The margin is 1-Lipschitz in $\|\cdot\|_\infty$.

Lemma 1 (Sign test independent of scale). *Let $a \in \Delta_{K-1}$ have nonzero entries at least Δ_-, write $m_G = \sum_{k \in G} a_k$, let $\beta > 0$, and let x be valid for (G, Δ_-) with margin $\vartheta > 0$. Then*

$$m_G = 0 \implies \langle a, x \rangle \leq -\vartheta, \qquad m_G \geq \Delta_- \implies \langle a, x \rangle \geq \vartheta.$$

Consequently, for $p = \sigma(\beta \langle a, x \rangle)$, $|p - \frac{1}{2}| \geq \sigma(\beta\vartheta) - \frac{1}{2}$.

A valid probe acts as a noisy OR query: $p > 1/2$ exactly when G intersects the support. The unknown positive scale does not change this decision boundary, but it does affect confidence. Lemma 1 certifies one contrast. Traversing a dyadic tree requires enough certified pairs at every node and accurate contrast predictions from the frozen dictionary. The next assumption states these two requirements.

Assumption 3 (Probe realizability). *There are constants $\vartheta_0 > \eta_{\mathrm{dict}} \geq 0$ such that the following holds for every nonroot group G in the dyadic partition tree over $[K]$. The response pool contains r pairs, determined only by G and the pool, whose frozen dictionary contrasts $\hat{x} = \hat{f}(\tau) - \hat{f}(\tau')$ satisfy $\vartheta(\hat{x}; G, \Delta_-) \geq \vartheta_0$. Every pair selected by the algorithm has true dictionary contrast $x = f(\tau) - f(\tau')$ satisfying $\|x - \hat{x}\|_\infty \leq \eta_{\mathrm{dict}}$. The r pairs may repeat or be shared across nodes, and each question receives a fresh answer. In the canonical instance,*

$$\hat{x} = \gamma_x(\mathbf{1}_G - c\mathbf{1}) + e, \qquad \gamma_x \geq \gamma, \qquad \|e\|_\infty \leq \eta_{\mathrm{pool}}, \qquad c = \Delta_-/2,$$

so one may take $\vartheta_0 = \gamma\Delta_-/2 - \eta_{\mathrm{pool}}$.

Here η_{pool} measures how closely an available pair realizes the canonical target in the frozen coordinates. The quantity η_{dict} measures the error between the frozen and true contrasts for that pair. The canonical margin in the true coordinates is therefore at least $\gamma\Delta_-/2 - \eta_{\mathrm{pool}} - \eta_{\mathrm{dict}} > 0$. The choice $c = \Delta_-/2$ maximizes this margin (Lemma 4).

Support recovery uses only signs and requires no upper bound on β. Estimating magnitudes requires curvature in the likelihood, which vanishes as the logits saturate.

Assumption 4 (Operating range). *There is a known $R < \infty$ such that every frozen contrast $\hat{x}_t$ used to estimate magnitudes satisfies $\beta|\langle a^\star, \hat{x}_t \rangle| \leq R$. This condition gives the uniform slope bound $c_R := \sigma'(R) > 0$. If only a range R_0 for the true contrasts is known and $\beta \leq \beta_+$, then $R = R_0 + \beta_+\eta_{\mathrm{dict}}$.*

4.2 CS-Elicit and its guarantees

CS-Elicit follows positive groups down a dyadic tree and then estimates weights on the recovered support.

Theorem 2 (Composition identification by questioning). *Let Assumptions 1 and 2 hold, with $S = \operatorname{supp} a^\star$ and $s = |S|$.*

(i) Support identification. *Under Assumption 3, set $b := \sigma(\beta_-(\vartheta_0 - \eta_{\mathrm{dict}})) - \frac{1}{2} > 0$. Algorithm 1 returns $\hat{S} = S$ with probability at least $1 - \delta$ using at most*

$$Q_{\mathrm{supp}} \;\leq\; 2s\lceil \log_2 K \rceil \left\lceil \frac{\log(2\bar{T}/\delta)}{2b^2} \right\rceil \qquad \textit{questions, where} \qquad \bar{T} = 2\lceil \log_2 K \rceil \lceil 1/\Delta_- \rceil.$$

Equivalently, the algorithm performs at most $2s\lceil \log_2 K \rceil$ tree node tests, each using a batch of r comparisons. Thus $Q_{\mathrm{supp}} = O\big(s \log K \cdot b^{-2} \log(\log K/(\Delta_-\delta))\big)$. ***(ii) Weight estimation.*** *Suppose Assumption 4 also holds and $\hat{S} = S$. Compute $\hat{w}$ and $\hat{a}$ as in Algorithm 1. With $n = sr'$ (exactly r' probes for each $k \in S$), the following bound holds with probability at least $1 - \delta$:*

$$\|\hat{a} - a^\star\|_1 \leq \frac{52 s^{3/2}}{\beta_- c_R \gamma} \sqrt{\frac{\log(2s/\delta)}{n}} + \frac{16 s \eta_{\mathrm{dict}}}{c_R \gamma}, \qquad c_R = \sigma'(R).$$

Algorithm 1: CS-Elicit

Input: frozen dictionary $\hat{f}$; response pool $\mathcal{P}$; support resolution Δ_-; repetition count $r = \left\lceil \frac{\log(2\bar{T}/\delta)}{2b^2} \right\rceil$; magnitude parameters (γ, r', R)

Initialize $\mathcal{A} \leftarrow \{[K]\}$ and $\hat{S} \leftarrow \emptyset$;
while $\mathcal{A} \neq \emptyset$ **do**
  Pop $G \in \mathcal{A}$ and split it into halves G_1, G_2;
  for $j \in \{1, 2\}$ **do**
    Ask r questions on the pairs of $\mathcal{P}$ with the largest $\vartheta(\hat{x}; G_j, \Delta_-)$, where $\hat{x} = \hat{f}(\tau) - \hat{f}(\tau')$, chosen as a function of $(G_j, \mathcal{P})$ only; let $\bar{y}_j$ be the mean answer;
    if $\bar{y}_j > 1/2$ **then**
      Add G_j to $\mathcal{A}$ if $|G_j| > 1$, otherwise add its element to $\hat{S}$;
if $\hat{S} = \emptyset$ **then**
  return $\perp$
For each $k \in \hat{S}$, ask exactly r' fresh questions on pairs of $\mathcal{P}$, chosen as a function of $(k, \hat{S}, \mathcal{P})$ only, with $\|\hat{x}_t - \gamma(e_k - \mathbf{1}/(2|\hat{S}|))\|_\infty \le \gamma/(4|\hat{S}|)$; collect the $|\hat{S}|r'$ contrasts and answers in $\mathcal{D}_{\mathrm{mag}}$;
Set $\hat{w} \leftarrow \arg\min_{w \ge 0,\ \mathrm{supp}\, w \subseteq \hat{S},\ |\langle w, \hat{x}_t \rangle| \le R} \sum_t [\log(1 + e^{\langle w, \hat{x}_t \rangle}) - y_t \langle w, \hat{x}_t \rangle]$ for $t \in \mathcal{D}_{\mathrm{mag}}$;
If $\mathbf{1}^\top \hat{w} > 0$, set $\hat{a} \leftarrow \hat{w}/\mathbf{1}^\top \hat{w}$; otherwise choose any $k_0 \in \hat{S}$ and set $\hat{a} \leftarrow e_{k_0}$;
return $\hat{S}$, $\hat{a}$

(iii) Lower bounds. *If $s\Delta_- \le 1$, any adaptive randomized strategy that recovers every admissible support with probability at least $2/3$ requires*

$$\textit{(a)} \quad Q \ge \frac{2s}{3\log 2} \log \frac{K}{s} - 2, \qquad \textit{(b)} \quad Q \ge \frac{\frac{2}{3}\log(K - s + 1) - \log 2}{\kappa^2},$$

where $\kappa^2 = (\beta_- \bar{\gamma} d_s)^2/(4c_R)$, $d_1 = 1$, and $d_s = \Delta_-$ for $s \ge 2$. Bound (b) additionally requires every queried contrast to satisfy $\|x\|_\infty \le \bar{\gamma}$ and $\beta_- |\langle a, x \rangle| \le R$ for every composition in the hard family of Appendix C.

4.3 Elicitation after extraction from exact rewards

Theorem 2 uses the true factor coordinates. Theorem 3 covers the approximately sparse compositions with a signed tail that arise when the dictionary is built from the exact rewards of impure anchors. The result analyzes an idealized version of Algorithm 2 with exact rewards. Define the loading distortion and composition drift as $\varepsilon_1(\nu) = \frac{4\nu(1-\nu)}{1-2\nu}$, $\zeta(\nu) = \frac{2\,\varepsilon_1(\nu)}{1-\varepsilon_1(\nu)}$, so that $\varepsilon_1(\nu) \le 4.3\,\nu$ and $\zeta(\nu) \le 11\,\nu$ for $\nu \le 1/20$.

Theorem 3 (Elicitation after extraction from exact rewards)**.** *Under Assumptions 1 and 2, let $\rho_1, \dots, \rho_K \in \Delta_{K-1}$ be user compositions with $(\rho_k)_k \ge 1 - \nu$ and $\nu \le 1/20$. Let $\hat{g}_k = \rho_k^\top f / \|\rho_k^\top f\|$ be their normalized rewards, which are known exactly. Run CS-Elicit with threshold $c = \Delta/2$ on a dictionary $\tilde{f}$. Suppose each tested group G has r_ν pairs determined by G and the pool. Their contrasts under $\tilde{f}$ are $\gamma(\mathbf{1}_G - c\mathbf{1}) + e$ with $\|e\|_\infty \le \eta_{\mathrm{pool}}$, and these contrasts differ from the corresponding contrasts under $\hat{g}$ by at most η_{est} in $\|\cdot\|_\infty$. If, with $\eta = \eta_{\mathrm{pool}} + \eta_{\mathrm{est}}$,*

$$\gamma\big(\Delta/2 - \zeta(\nu)\big) > \big(1 + \zeta(\nu)\big)\,\eta, \tag{2}$$

then CS-Elicit recovers S with probability at least $1 - \delta$ using at most $2s\lceil \log_2 K \rceil\, r_\nu$ questions, with b' and r_ν given in equation 6 of Appendix D.

For a ν-anchored population whose hull $\mathrm{conv}\, A$ has exactly K extreme points, Algorithm 2 returns such ρ_k from exact rewards (Lemma 5). In this case, $\tilde{f} = \hat{g}$ and $\eta_{\mathrm{est}} = 0$. The theorem does not cover utility estimation or extreme ray selection from finite samples: it does not bound η_{est} or show that the implemented screening selects ν-pure users (Appendix D.4). Condition equation 2 follows

from $\Delta > 22\nu + 2(1 + 11\nu)\eta/\gamma$. The extreme users selected in our PersonalLLM extraction are much less pure than the requirement $\nu \leq 1/20$.

5 EXPERIMENTS

Our experiments address four questions. **(E1)** We measure whether dictionaries learned from population comparisons remain stable across optimizer seeds, user splits, added interior users, and independent data draws. **(E2)** At a fixed query budget, we compare the accuracy and computation of CS-Elicit with random, mutual information, and volume removal acquisition while holding the dictionary and composition estimator fixed. **(E3)** We compare complete sparse personalization pipelines with existing personalized reward models. **(E4)** We measure the questions required to reach a fixed prediction accuracy and how this count changes with dictionary size.

5.1 SETUP

Data and users. The language suite includes PersonalLLM (Zollo et al., 2025), with factors given by the standardized scores of ten published reward models; UltraFeedback (Cui et al., 2023), with four GPT-4 aspect ratings as factors; and UF-P-4 and TL;DR (Stiennon et al., 2020), which provide fixed labels without factor scores. We represent each response x by the final-token hidden state $h(x)$ from Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2026), then reduce it to 512 dimensions with PCA fitted on training prompts. The control suite includes Gymnasium (de Lazcano et al., 2024) and Franka Kitchen (Fu et al., 2020); their factors are normalized observation coordinates or moving object joints. When factors are known, users are semi-synthetic s-sparse or Dirichlet compositions $a \in \Delta_{K-1}$ that answer through the Bradley–Terry channel $\sigma(\beta\, a^\top (f(x) - f(x')))$. We set $\beta = 5$ for language elicitation, $\beta = 1$ for the stability study in Appendix F.2, and calibrate β to a population Bayes accuracy of 0.80 on control tasks. In the complete pipeline experiment, PersonalLLM users are the Dirichlet mixtures of Bose et al. (2025) with deterministic labels, and the four GPT-4 aspects define the UF-P-4 users. TL;DR is the only dataset with human labels linked to persistent annotators. Table 5 in Appendix F.1 describes user construction, data splits, label counts, and item overlap.

Methods. We use two BT references: a pooled model fitted on seen users without target-user labels, and an independent model fitted on all labeled comparisons from the target user. The personalized models are LoRe (Bose et al., 2025), VPL (Poddar et al., 2024), SPL (Kim & Kim, 2026), PAL (Chen et al., 2025), and Sparse-LoRe, a gated LoRe variant that separates sparsity from extraction (Appendix F.1). We compare random acquisition, mutual information (MI), volume removal (VR) (Sadigh et al., 2017), and CS-Elicit. CS-Elicit asks each dyadic node once because the repetition schedule in Theorem 2(i) exceeds the experimental budgets.

Metrics. We report held-out preference accuracy, negative log-likelihood, and wall-clock time. To compare dictionaries, we pair their factors with a Hungarian assignment and average the Pearson correlations between paired factor scores on held-out items. Appendix F gives the settings, hyperparameters, and compute for each experiment.

5.2 RESULTS

Stability of extracted dictionaries. Theorem 1 predicts that extraction is invariant to optimization and to the addition of interior users. Table 1 evaluates LoRe, Sparse-LoRe, Identified LoRe (factors extracted from LoRe utilities), and Extreme-ray (factors extracted from individual user utilities) across optimizer seeds, user splits, added dense users, and independent data draws. Appendix F.2 gives the protocol. On PersonalLLM, Identified LoRe raises the cross-draw correlation from 0.42 to 0.63. Direct extreme-ray extraction has the highest cross-draw correlation (0.83) and lowers split-induced prediction disagreement to 3.0%, compared with 9–10% for the LoRe variants. The methods are generally stable on Control, where Identified LoRe gives the highest recovery. PersonalLLM recovery remains high even when the population has no factor structure, so stability alone does not verify the generating factors. The corresponding elicitation results therefore describe the extracted coordinates.

Efficient elicitation for unseen users. Table 2 isolates query selection: every rule uses the ground-truth dictionary, the same composition estimator, and the same users, candidate pool, answers, and

Table 1: Dictionary stability. Score correlations on held-out items (higher is better), and half-split prediction disagreement (Disagr.). *Random* is the noise floor of Gaussian dictionaries.

Dataset	Method	Recovery	Seed	Split	Dense	Draw	Disagr. (%)
PersonalLLM	LoRe	0.36 ± 0.14	0.55 ± 0.16	0.79 ± 0.13	0.54 ± 0.17	0.42 ± 0.10	8.8 ± 2.2
	Sparse-LoRe	0.14 ± 0.02	0.26 ± 0.02	$\mathbf{0.97 \pm 0.00}$	0.27 ± 0.04	0.27 ± 0.03	9.3 ± 1.8
	Identified LoRe	$\mathbf{0.62 \pm 0.12}$	0.69 ± 0.15	0.77 ± 0.06	0.71 ± 0.15	0.63 ± 0.14	10.1 ± 1.8
	Extreme-ray	0.52 ± 0.14	$\mathbf{1.00 \pm 0.00}$	0.88 ± 0.11	$\mathbf{0.99 \pm 0.04}$	$\mathbf{0.83 \pm 0.12}$	$\mathbf{3.0 \pm 4.2}$
	Random	0.09 ± 0.04	0.47 ± 0.02	0.47 ± 0.02	0.47 ± 0.02	0.47 ± 0.02	48.7 ± 5.6
Control	LoRe	0.85 ± 0.13	0.93 ± 0.10	0.86 ± 0.13	0.92 ± 0.10	0.84 ± 0.15	$\mathbf{1.7 \pm 1.2}$
	Sparse-LoRe	0.85 ± 0.12	0.90 ± 0.09	$\mathbf{0.91 \pm 0.08}$	0.90 ± 0.09	0.89 ± 0.11	3.0 ± 2.0
	Identified LoRe	$\mathbf{0.87 \pm 0.11}$	0.92 ± 0.08	0.88 ± 0.11	0.93 ± 0.07	$\mathbf{0.93 \pm 0.08}$	3.9 ± 2.9
	Extreme-ray	0.80 ± 0.15	$\mathbf{1.00 \pm 0.00}$	0.85 ± 0.12	$\mathbf{0.94 \pm 0.11}$	0.88 ± 0.11	3.8 ± 3.2
	Random	0.44 ± 0.21	0.44 ± 0.14	0.44 ± 0.14	0.44 ± 0.14	0.44 ± 0.14	30.8 ± 9.3

Table 2: Held-out deterministic preference accuracy (%) of 1-sparse unseen users after $Q = 8$ answers on the ground-truth dictionary, mean $\pm$ standard deviation across seeds.

Dataset	Pooled BT	Random	MI	VR	CS-Elicit	Indep. BT, all pairs
PersonalLLM	77.5 $\pm$1.4	83.9 $\pm$0.5	87.3 $\pm$0.3	83.4 $\pm$1.9	**93.0** $\pm$0.8	96.7 $\pm$0.4
UltraFeedback	90.5 $\pm$0.6	96.4 $\pm$0.5	99.9 $\pm$0.0	90.7 $\pm$0.4	**100.0** $\pm$0.0	100.0 $\pm$0.0
MountainCar (K=2)	91.6 $\pm$0.7	95.4 $\pm$1.2	97.6 $\pm$0.6	93.4 $\pm$1.0	**99.3** $\pm$0.5	98.0 $\pm$0.5
CartPole (K=4)	61.6 $\pm$1.5	85.9 $\pm$1.7	91.7 $\pm$1.3	85.5 $\pm$3.3	**94.3** $\pm$1.7	97.5 $\pm$0.3
Acrobot (K=6)	63.1 $\pm$2.8	77.7 $\pm$2.9	84.5 $\pm$2.8	73.1 $\pm$2.2	**97.3** $\pm$1.6	95.2 $\pm$1.8
LunarLander (K=8)	58.7 $\pm$1.9	78.9 $\pm$1.7	84.7 $\pm$1.4	77.2 $\pm$3.0	**89.0** $\pm$3.6	96.4 $\pm$0.6
Kitchen (K=14)	55.6 $\pm$0.5	63.3 $\pm$1.5	63.4 $\pm$2.1	62.4 $\pm$1.8	**64.4** $\pm$1.8	77.3 $\pm$2.6

test pairs. After eight answers, CS-Elicit has the highest mean accuracy for 1-sparse users on all seven datasets. Its largest gain over MI is 12.8 points on Acrobot, and it reaches 100% accuracy on UltraFeedback. CS-Elicit also matches or exceeds the independent BT reference on three datasets while using eight labels instead of 456–512. Appendix F.3 studies less sparse users and weak probes. Appendix F.6 reports exact support recovery and runs the repetition schedule from Theorem 2(i) when the pool contains enough distinct valid pairs; this condition holds only for MountainCar and UltraFeedback at $s = 1$. At the reference accuracy target, CS-Elicit takes 3.0–26.6 ms per user, which is 6.3–10.6$\times$ less time than MI (Appendix F.4).

End-to-end comparison. Table 3 compares our extracted-dictionary pipeline after 16 CS-Elicit answers with pooled BT and four personalized reward models. The personalized baselines use their native adaptation procedures with 16 random answers, while independent BT uses all available labels. Our pipeline has the highest accuracy on UF-P-4, Gym, and Kitchen; its largest gain over the best pooled or personalized baseline is 0.066 on Kitchen. It is within 0.007 of the best method on PersonalLLM and TL;DR. NLL is less consistent because the pipeline is overconfident on PersonalLLM and UF-P-4. UF-P-4 contains only four users: the models fit these users on training prompts and evaluate them on held-out prompts. This dataset therefore measures adaptation to known users on new prompts. When the dictionary and estimator are fixed, CS-Elicit is best only on UF-P-4, while MI matches or exceeds it on the control tasks (Appendix F.7). This comparison shows that the frozen dictionary and estimator account for most of the complete pipeline's gain.

Scaling of the question count. Figure 1(a) reports the answers needed to reach 93% accuracy for 1-sparse users as K grows, with all rules using the same candidate pairs and composition estimator (Appendix F.9). Up to the ground-truth dimension, CS-Elicit stays close to one tree pass, $2\lceil \log_2 K \rceil$, and uses the fewest questions except on Acrobot. On PersonalLLM, increasing K from 2 to 10 raises the CS-Elicit count from 1.6 to 10.1 questions. Over the same range, MI rises from 2.6 to 26.1 and random selection from 7.7 to 52.6.

Panel (b) marks the range of this advantage. As the support grows, node margins shrink and strong probes become rare. At $s = 3$, no node has $\beta\vartheta \geq 1$ on three of the four tasks, and only 6% do on MountainCar, even when most nodes remain formally certified. The predicted logarithmic

Table 3: End-to-end personalization of unseen users after $Q = 16$ answers: held-out accuracy (Acc.) and negative log-likelihood (NLL).

	PersonalLLM		TL;DR		UF-P-4		Gym (4 tasks)		Kitchen	
Method	Acc.	NLL	Acc.	NLL	Acc.	NLL	Acc.	NLL	Acc.	NLL
Pooled BT	.848±.122	.366±.210	.654±.029	.621±.025	.592±.063	.671±.025	.775±.102	.572±.086	.614±.089	.680±.026
LoRe	**.932**±.073	.820±.844	.675±.070	.684±.179	.581±.074	1.07±.184	.814±.086	.719±.291	.857±.075	.759±.288
VPL	.927±.117	**.209**±.331	.705±.048	**.564**±.060	.588±.061	.671±.038	.864±.074	.500±.095	.789±.127	.551±.171
SPL	.927±.119	.211±.332	**.706**±.047	**.564**±.060	.587±.063	**.670**±.037	.857±.077	.504±.095	.799±.138	.544±.194
PAL	.923±.106	.264±.155	.689±.037	.584±.034	.590±.070	.673±.054	.869±.070	**.489**±.082	.852±.078	.519±.093
Ours	.926±.116	.376±.131	.699±.045	**.564**±.052	**.635**±.051	.927±.092	**.896**±.078	.493±.205	**.923**±.130	**.503**±.577
Indep. BT	.966±.043	.124±.168	.685±.041	.605±.072	.639±.065	.629±.060	.954±.030	.416±.050	.958±.032	.332±.053

dependence therefore appears only when the candidate pool contains sufficiently strong group tests. Panel (c) evaluates the support returned by the one-answer-per-node walk; it does not evaluate the repeated tests required for exact recovery by Theorem 2(i). At $K = K_{\text{true}}$, the walk recovers the exact support for 27–79% of users on the four control tasks, and the rate is 62% on PersonalLLM at $K = 10$. At $K = 4K_{\text{true}}$, the range falls to 0–31%. Reaching an accuracy target with fewer questions therefore does not imply exact support recovery.

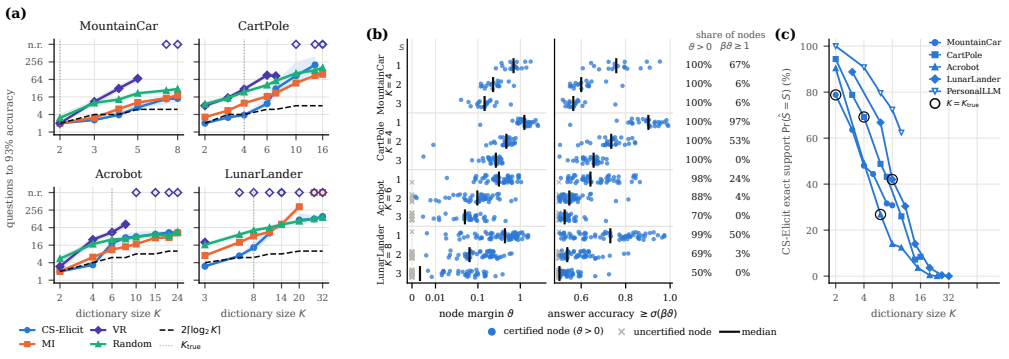


Figure 1: (a) Questions needed to reach 93% unseen-user prediction accuracy as the dictionary grows ($s = 1$). Open markers: target not reached. (b) Node margins of the same tasks at $s = 1, 2, 3$ and the per-answer accuracy they guarantee. (c) Probability that CS-Elicit's walk returns exactly the true support; mean over seeds, rings mark $K = K_{\text{true}}$.

6 Conclusion

Population comparisons identify a shared reward dictionary when each factor has an anchor: the factors are the extreme rays of the population reward cone and are identifiable up to permutation within the anchored model class. Once the dictionary is fixed, simplex weights allow realizable group contrasts to test group mass independently of user scale. CS-Elicit then recovers sparse supports with $O(s \log K)$ tree-node tests and, at fixed answer bias, a number of comparisons that is logarithmic in K.

The experiments show where these results apply in practice. On PersonalLLM, extracting factors from LoRe utilities raises the mean cross-draw dictionary correlation from 0.42 to 0.63. In the E4 scaling experiment with an extracted $K = 10$ dictionary, CS-Elicit reaches 93% accuracy for semi-synthetic 1-sparse users in 10.1 questions on average, compared with 26.1 for mutual information. The complete pipeline's gains depend on dictionary quality and on whether the candidate pool contains informative probes.

AI USE STATEMENT

Large language models were used to improve the wording of the text and to check the exposition of the proofs. The problem formulation, the theorems and their proofs, the algorithms, the experiments, and the analysis of the results are the authors' own work, and the authors take full responsibility for the content.

REFERENCES

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: an information theory perspective. *Foundations and Trends® in Communications and Information Theory*, 15(3-4): 196–392, 2019.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning*, pp. 280–288. PMLR, 2013.

André Barreto, Vincent Dumoulin, Yiran Mao, Mark Rowland, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing individual human preferences with reward features. *Advances in Neural Information Processing Systems*, 38:82119–82154, 2026.

Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning. *arXiv preprint arXiv:2005.02575*, 2020.

Avinandan Bose et al. LoRe: Personalizing LLMs via low-rank reward modeling. *arXiv preprint arXiv:2504.14439*, 2025.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21. IEEE, 2008.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations*, 2025.

Maurice D Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):542–552, 1994.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Rodrigo de Lazcano, Kallinteris Andreas, Jun Jet Tai, Seungjae Ryan Lee, and Jordan Terry. Gymnasium robotics, 2024. URL `http://github.com/Farama-Foundation/Gymnasium-Robotics`.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.

Robert Dorfman. The detection of defective members of large populations. *The Annals of mathematical statistics*, 14(4):436–440, 1943.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *arXiv preprint arXiv:1803.01257*, 2018.

Luise Ge, Brendan Juba, and Yevgeniy Vorobeychik. Learning linear utility functions from pairwise comparison queries. *arXiv preprint arXiv:2405.02612*, 2024.

Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2013.

Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2013.

Frank K Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67(339):605–608, 1972.

Enoch Hyunwook Kang. Personalized alignment revisited: The necessity and sufficiency of user diversity. *arXiv preprint arXiv:2605.09119*, 2026.

Gihoon Kim and Euntai Kim. Swap-guided preference learning for personalized reinforcement learning from human feedback, 2026. URL `https://arxiv.org/abs/2603.12595`.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *International Conference on Machine Learning*, pp. 231–239. PMLR, 2013.

Gyuseok Lee, Wonbin Kweon, Zhenrui Yue, SeongKu Kang, Jiawei Han, and Dong Wang. Uncertainty-aware variational reward factorization via probabilistic preference bases for llm personalization. *arXiv preprint arXiv:2604.00997*, 2026.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. Prefpalette: Personalized preference modeling with latent attributes. *arXiv preprint arXiv:2507.13541*, 2025.

Chia-Hsiang Lin, Wing-Kin Ma, Wei-Chiang Li, Chong-Yung Chi, and ArulMurugan Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *International Conference on Learning Representations*, volume 2026, pp. 133805–133838, 2026.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. Rethinking diverse human preference learning through principal component analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19857–19870, 2025.

Matthew L Malloy and Robert D Nowak. Sequential testing for sparse recovery. *IEEE Transactions on Information Theory*, 60(12):7862–7873, 2014.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

José MP Nascimento and José MB Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

Minhyeon Oh, Seungjoon Lee, and Jungseul Ok. Comparison-based active preference learning for multi-dimensional personalization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 33145–33166, 2025.

Victor-Alexandru Pădurean, Parameswaran Kamalaruban, Nachiket Kotalwar, Alkis Gotovos, and Adish Singla. Inference-time personalized alignment with a few user preference queries. *Advances in Neural Information Processing Systems*, 38:85125–85156, 2026.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Li Qian, Jinyang Gao, and HV Jagadish. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment*, 8(11):1322–1333, 2015.

Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment. *arXiv preprint arXiv:2512.23067*, 2025.

Dorsa Sadigh, Anca Dragan, Shankar Sastry, and Sanjit Seshia. Active preference-based learning of reward functions. 2017.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*, 2025.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Yifan Wang, Jinyi Mu, Mayank Jobanputra, Yu Wang, Ji-Ung Lee, Soyoung Oh, Isabel Valera, and Vera Demberg. Sparse mixture-of-experts reward models learn interpretable and specialized experts for personalized preference modeling. *arXiv preprint arXiv:2606.04284*, 2026.

Michael E Winter. N-findr: An algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *Imaging spectrometry V*, volume 3753, pp. 266–275. SPIE, 1999.

Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

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

Thomas Zollo, Andrew Siah, Naimeng Ye, Li Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, volume 2025, pp. 66949–66971, 2025.

A Behavioral and Conic Preliminaries

Lemma 2 (What comparison behavior reveals)**.** *Two reward parameterizations induce the same comparison kernel for user i if and only if their rewards differ by a constant almost surely. Centering sets this constant to zero, and*

$$r_i(\tau) = \mathbb{E}_{\tau' \sim \mu}\big[\sigma^{-1}(P_i(\tau, \tau'))\big] \quad \textit{for almost every } \tau.$$

The map $w \mapsto (\beta, a) = (\|w\|_1, w/\|w\|_1)$ is also a bijection from the nonzero, nonnegative loadings to $(0, \infty) \times \Delta_{K-1}$.

Proof. Because σ is strictly increasing, two kernels are equal if and only if all reward differences are equal. By Fubini's theorem, a function whose pairwise differences vanish almost surely is almost surely constant, and centering removes this constant. Integrating $\sigma^{-1}(P_i(\tau, \tau')) = r_i(\tau) - r_i(\tau')$ over τ' yields the displayed identity. For $w \geq 0$, summing $w = \beta a$ over its coordinates gives $\beta = \|w\|_1$ and then $a = w/\|w\|_1$. □

Lemma 3 (Reduction to conic transformations)**.** *Under Assumption 1, any two behaviorally equivalent, normalized parameterizations with K factors satisfy $\tilde{f} = Mf$ and $\tilde{w}_i = M^{-\top} w_i$ for an invertible matrix M. The matrix $M^{-\top}$ maps every observed loading into the nonnegative orthant. Equivalent parameterizations have the same unordered dictionary exactly when M is a permutation matrix.*

Proof. Lemma 2 implies that the centered rewards are equal. The spanning condition in Assumption 1 then implies that both dictionaries span the same K dimensional function space, so $\tilde{f} = Mf$ for some invertible M. Matching coefficients gives $\tilde{w}_i = M^{-\top} w_i$, and nonnegativity gives the conic constraint. Unit norms rule out positive diagonal rescaling, while nonnegativity rules out sign changes. Only permutations preserve the unordered dictionary. □

B Proof of Factor Extraction

Proof of Theorem 1. Identify $\operatorname{span}\{f_k\}$ with $\mathbb{R}^K$ through the map $c \mapsto c^\top f$. Under this identification, $\mathcal{D}$ corresponds to $\operatorname{cl}\operatorname{cone}\{a_i : i \in I\} = \operatorname{cone}(A)$. The inclusion $\subseteq$ holds because $\operatorname{cone}(A)$ is closed, since A is compact and bounded away from the origin. The reverse inclusion holds because every point of A is a limit of compositions and $\operatorname{cl}\operatorname{cone}\{a_i\}$ is a closed convex cone. Since $A \subseteq \Delta_{K-1}$, the slice $\{x \in \operatorname{cone}(A) : \mathbf{1}^\top x = 1\}$ equals $\operatorname{conv} A$, so $f_k \in \mathcal{D}$ if and only if $e_k \in \operatorname{conv} A$. If $e_k \in \operatorname{conv} A$, then e_k is an extreme point of both Δ_{K-1} and $\operatorname{conv} A$. It therefore belongs to A, because every extreme point of the convex hull of a compact set belongs to that set. The converse is immediate. When $e_k \in A$, its ray is extreme in $\mathbb{R}^K_{\geq 0}$ and remains extreme in the subcone $\operatorname{cone}(A)$, so its image is an extreme ray of $\mathcal{D}$. For an anchored population, $\operatorname{cone}(A)$ contains every e_k and lies within $\mathbb{R}^K_{\geq 0}$. It therefore equals the orthant, and $\mathcal{D} = \operatorname{cone}\{f_1, \ldots, f_K\}$ has the factor rays as its extreme rays. Any behaviorally equivalent anchored parameterization generates the same cone $\mathcal{D}$, so its factors span the same rays. Unit norm fixes the representative of each ray, and Lemmas 2 and 3 recover the loadings, compositions, and scales up to permutation.

For (ii), let $A = \operatorname{cl}\{a_i\} \subseteq \Delta_{K-1}$ and $\mathcal{C} = \operatorname{cl}\operatorname{cone}(A)$. The extreme rays of $\mathcal{C}$ correspond to the extreme points of its unit mass slice $\overline{\operatorname{conv}}(A)$. If there are exactly K such rays, their unit mass representatives $\rho_1, \ldots, \rho_K$ are linearly independent because the compositions span $\mathbb{R}^K$ and lie in $\operatorname{cone}\{\rho_j\}$. They also belong to A, since the extreme points of the convex hull of a compact set belong to that set. Define $g_j = (\rho_j^\top f)/\|\rho_j^\top f\|$. Every reward has a unique nonnegative expansion in the g_j, which gives an equivalent anchored parameterization whose factors are the normalized rewards of extreme users. Any other parameterization in $\mathcal{M}^\star$ with the same kernels generates the same observable cone. This cone is simplicial with extreme rays spanned by the g_j, so the argument for (i) shows that its dictionary equals g up to relabeling. If a pure point e_k is absent from A, this dictionary differs from the generating dictionary. If the number of extreme rays differs from K, no anchored representation with K factors can generate the same observable cone, because such a representation would make the cone simplicial with exactly K extreme rays.

For (iii), Lemma 2 shows that the two parameterizations have the same rewards. Thus $\mathcal{D}$ and its K extreme rays are the same functions under both parameterizations. Let $\hat{g}_1, \ldots, \hat{g}_K$ be the unit

Algorithm 2: Reward factor discovery with nonnegative weight refitting

Input: per-user comparisons $\mathcal{D}_i = \{(\tau_{it}^+, \tau_{it}^-, y_{it})\}_{t=1}^{n_i}$ for users $i \in [N]$; feature map ϕ and ridge parameter λ_v; dictionary size K; numerical screening tolerance ϵ_0.

1. Fit centered $\widehat{r}_i$ by equation 3; retain their original scales.
2. Form $q_i = \widehat{r}_i / \|\widehat{r}_i\|_{L^2(\mu)}$ for nonzero estimates; flag near-zero estimates and merge duplicate directions (approximately at tolerance in finite data).
3. Screen candidate directions by their nonnegative projection residual onto the other distinct directions, retaining those whose residual exceeds ϵ_0; in the exact case a direction is extreme iff $q_i = \sum_{j \neq i} c_j q_j$, $c_j \geq 0$, is infeasible.
4. Select K screened independent directions using nonnegative projection residuals; if this is not possible, report a discovery failure.
5. Set $\widehat{f}_k$ to their centered unit-norm representatives.
6. For each user, fit $\widehat{w}_i \in \arg\min_{w \in \mathbb{R}^K_{\geq 0}} \|\widehat{r}_i - w^\top \widehat{f}\|^2_{L^2(\mu)}$ and set $\widehat{\beta}_i = \mathbf{1}^\top \widehat{w}_i$, $\widehat{a}_i = \widehat{w}_i / \widehat{\beta}_i$.

Output: $\widehat{f}$, $\{\widehat{w}_i\}$, and discovery/refit residuals.

norm representatives of these rays. In the coordinates of θ, Lemma 5 labels the extreme points of $\operatorname{conv} A$ so that $(\rho_k)_k \geq 1 - \nu$ and $\hat{g}_k = \rho_k^\top f / \|\rho_k^\top f\|$. Proposition 3(a) then gives $\|\hat{g}_k - f_k\| \leq 4\nu$. The same argument in the coordinates of $\widetilde{\theta}$ gives $\|\hat{g}_{\pi(k)} - \widetilde{f}_k\| \leq 4\nu$ for some permutation π. The triangle inequality yields $d_{\mathrm{perm}}(f, \widetilde{f}) \leq 8\nu$, and Proposition 3(b) transfers each composition to the $\hat{g}$ coordinates and back. □

Remark 1 (Identification is relative to the anchor class)**.** *For any $\theta \in \mathcal{M}$ and any linearly independent $v_1, \ldots, v_K \in \mathbb{R}^K$ with $\operatorname{cone}\{v_1^\top f, \ldots, v_K^\top f\} \supseteq \mathcal{D}$, expanding each r_i in the normalized functions $v_j^\top f / \|v_j^\top f\|$ gives a parameterization in $\mathcal{M}$ with the same kernels, so within $\mathcal{M}$ no dictionary is identified, anchored or not; the anchored parameterizations are exactly those whose cone equals $\mathcal{D}$, and Theorem 1(i) identifies the dictionary among them. If $\theta \notin \mathcal{M}^\star$, this freedom moves the factors themselves. Some $e_k \notin A$, hence $e_k \notin \operatorname{cone}(A)$ by the proof of part (i), and separation of the point e_k from the closed convex cone $\operatorname{cone}(A)$ gives $h \in \mathbb{R}^K$ with $h_k = h^\top e_k < 0 \leq h^\top a_i$ for every i. For $0 < t < 1/|h_k|$ let $N_t = I + t\, e_k h^\top$, which is invertible with $\det N_t = 1 + t h_k > 0$ and inverse transpose $N_t^{-\top} = I - \frac{t}{1+th_k}\, h\, e_k^\top$. The loadings $N_t w_i = w_i + t\,(h^\top w_i)\, e_k$ are nonnegative, they span $\mathbb{R}^K$, and $w_i^\top f = (N_t w_i)^\top N_t^{-\top} f$, so the functions $\widetilde{f}_j^{(t)} = f_j - \frac{t h_j}{1+th_k} f_k$, normalized to unit norm with the loadings rescaled accordingly, form a parameterization in $\mathcal{M}$ with the same kernels. Some $j \neq k$ has $h_j \neq 0$, because $h = h_k e_k$ would force $a_{ik} = 0$ for every i against the spanning condition, and for that j the factor $\widetilde{f}_j^{(t)}$ is a combination of f_j and f_k with two nonzero coefficients, hence a multiple of no f_l. An unanchored population therefore admits a continuum of behaviorally equivalent dictionaries, none of which is a relabeling of f, and the anchored representation of part (ii), when it exists, is another such dictionary.*

B.1 Reward dictionary extraction from finite samples

We first fit each user's reward in a common feature space. Let $\phi : \mathcal{T} \to \mathbb{R}^d$ be a fixed feature map with $\mathbb{E}_\mu[\phi] = 0$, and assume that the true factors lie in its span. Define $\delta\phi_{it} = \phi(\tau_{it}^+) - \phi(\tau_{it}^-)$ as the contrast vector of user i's t-th comparison, and $\ell(y, z) = \log(1 + e^z) - yz$ as the logistic loss. We estimate

$$\widehat{v}_i = \arg\min_{v \in \mathbb{R}^d} \left\{ \frac{1}{n_i} \sum_{t=1}^{n_i} \ell(y_{it}, v^\top \delta\phi_{it}) + \frac{\lambda_v}{2} \|v\|_2^2 \right\}, \quad \widehat{r}_i(\tau) = \widehat{v}_i^\top \phi(\tau), \quad \text{with } \lambda_v > 0. \tag{3}$$

All later geometric operations use the same inner product in function space. In feature coordinates, a shared reference sample can approximate $\Sigma_\phi = \mathbb{E}_\mu[\phi(\tau)\phi(\tau)^\top]$. The implementation assumes that the true factors lie in the span of the feature map. Misspecified features, ridge bias from finite

samples, and coordinate changes that alter the penalty can distort the reward geometry before factor extraction.

We then normalize the nonzero fitted rewards, merge approximately duplicate directions, and use nonnegative projection residuals under a common $L^2(\mu)$ inner product to screen for extreme rays. From the screened rays, we select K independent directions as the estimated factors. We then refit each user's nonnegative weights, scale, and composition in this dictionary. With exact rewards and an anchored population, the screening recovers the factor directions by Theorem 1(i).

C Proofs of Sparse Elicitation

Proof of Lemma 1. If $m_G = 0$ then $\langle a, x\rangle \le \max_{k\notin G} x_k \le -\vartheta$. Suppose $m_G \ge \Delta_-$. Positivity of the second term in equation 1 together with $\min_{k\notin G} x_k \le \max_{k\notin G} x_k < 0$ forces $\min_{k\in G} x_k > 0 > \min_{k\notin G} x_k$, so the lower bound $\langle a, x\rangle \ge m_G \min_{k\in G} x_k + (1-m_G)\min_{k\notin G} x_k$ is nondecreasing in m_G and is at least its value at $m_G = \Delta_-$, which is at least ϑ. A user whose nonzero weights are at least Δ_- has $m_G = 0$ or $m_G \ge \Delta_-$, and the statements about p follow because σ is increasing with $\sigma(0) = \frac{1}{2}$ and $\beta > 0$. For the canonical contrast, $-\max_{k\notin G} x_k \ge \gamma_x c - \eta$ and $\Delta_- \min_{k\in G} x_k + (1-\Delta_-)\min_{k\notin G} x_k \ge \Delta_-(\gamma_x(1-c)-\eta) + (1-\Delta_-)(-\gamma_x c - \eta) = \gamma_x(\Delta_- - c) - \eta$. Each term of equation 1 is a minimum or maximum of coordinates of x combined with absolute coefficients summing to one, hence 1-Lipschitz in $\|\cdot\|_\infty$, and so is their minimum. For the converse, if $\max_{k\notin G} x_k \ge 0$ then $a = e_j$ with $j \in \arg\max_{k\notin G} x_k$ has $m_G = 0$ and $\langle a, x\rangle \ge 0$. Otherwise the second term of equation 1 is nonpositive, and $a = \Delta_- e_k + (1-\Delta_-)e_j$ with $k \in \arg\min_{l\in G} x_l$ and $j \in \arg\min_{l\notin G} x_l$ has nonzero weights Δ_- and $1-\Delta_- \ge \Delta_-$, has $m_G = \Delta_-$, and satisfies $\langle a, x\rangle = \Delta_- \min_{l\in G} x_l + (1-\Delta_-)\min_{l\notin G} x_l \le 0$. □

Remark 2. *For the canonical contrast $x = \gamma_x(\mathbf{1}_G - c\mathbf{1}) + e$ with $c \le \Delta_-$ and $\|e\|_\infty \le \eta_{\mathrm{pool}}$, we have $\vartheta(x; G, \Delta_-) \ge \gamma_x \min\{c, \Delta_- - c\} - \eta_{\mathrm{pool}}$. Conversely, if $\Delta_- \le \frac{1}{2}$ and $\vartheta(x; G, \Delta_-) \le 0$, some admissible composition under one of the two hypotheses has the wrong sign. A positive margin therefore characterizes uniform sign separation.*

Proof of Theorem 2(i) and (iii). Let $\bar{T} = 2\lceil \log_2 K\rceil \lceil 1/\Delta_-\rceil$, $b = \sigma(\beta_-(\vartheta_0 - \eta_{\mathrm{dict}})) - \frac{1}{2}$, and $r = \lceil \log(2\bar{T}/\delta)/(2b^2)\rceil$, as in Algorithm 1. Since $\vartheta(\cdot; G, \Delta_-)$ is 1-Lipschitz in $\|\cdot\|_\infty$ by Lemma 1, every pair selected under Assumption 3 has true-coordinate margin at least $\vartheta_0 - \eta_{\mathrm{dict}}$, so Lemma 1 and $\beta \ge \beta_-$ give $|p - \frac{1}{2}| \ge b$ for every test the algorithm performs, with $p > \frac{1}{2}$ exactly when the tested group meets S. Theorem 3 uses instead the bound $\sigma(z) - \frac{1}{2} \ge \frac{1}{6}\min\{z, 1\}$ of Lemma 6, which gives the smaller value $\frac{1}{6}\min\{\beta_-(\vartheta_0 - \eta_{\mathrm{dict}}), 1\} \le b$.

Answer table. By Assumption 3, the r pairs asked at a node G are determined by G and the pool, not by earlier answers. By the conditional independence of answers (Section 3), the transcript therefore has the same law as the following construction: before the interaction, draw independent answers $y_{G,j} \sim \mathrm{Ber}(p_{G,j})$ for every nonroot node G and $j \le r$, where $p_{G,j}$ is the response probability of the j-th pair of G, and let the algorithm read $y_{G,1}, \dots, y_{G,r}$ when it tests G. A pair repeated within or across nodes receives distinct table entries, i.e. fresh answers. If answers cannot be refreshed, as with recorded comparisons, the argument below still holds when the r pairs of each node are distinct pairs with independent answers: Hoeffding's inequality is applied within a node, and the union bound needs no independence across nodes.

Upper bound. Let $\mathcal{G}_S$ be the set of children of tree nodes that meet S; it is determined by S, and it has at most $2s\lceil \log_2 K\rceil \le \bar{T}$ elements. Hoeffding's inequality applied to the table entries of each $G \in \mathcal{G}_S$ and a union bound over $\mathcal{G}_S$ imply that all empirical means for $G \in \mathcal{G}_S$ lie strictly within b of their expectations with probability at least $1-\delta$. On this event Lemma 1 classifies every group in $\mathcal{G}_S$ correctly, and the algorithm tests exactly these groups. At any tree level at most s disjoint groups intersect the support, so at most $2s$ child groups are tested per level and all support singletons are reached. This proves the upper bound. Each test is a batch of r comparisons, and $\log(2\bar{T}/\delta) = \log(4\lceil \log_2 K\rceil \lceil 1/\Delta_-\rceil/\delta)$, which gives the $O(\cdot)$ form; the $O(s \log K)$ count refers to tree-node tests, not to comparisons.

Counting lower bound, (iii)(a). Place a uniform prior on the $\binom{K}{s}$ compositions that assign mass $1/s$ to a size-s support. These compositions are admissible because $s\Delta_- \le 1$. Queries are functions

of past answers and internal randomness, so Q binary answers carry at most $Q \log 2$ nats about the support. Fano's inequality at success probability $2/3$ gives $Q \geq \frac{2}{3 \log 2} \log \binom{K}{s} - 2$, and $\log \binom{K}{s} \geq s \log(K/s)$ gives the stated form.

Information lower bound, (iii)(b). For $s \geq 2$ fix mass $(1 - \Delta_-)/(s-1)$ on coordinates $1, \ldots, s-1$ and place the remaining mass $d_s = \Delta_-$ on one unknown coordinate $j \in \{s, \ldots, K\}$, with scale β_-; these compositions are admissible because $s\Delta_- \leq 1$. For $s = 1$ place all mass $d_1 = 1$ on an unknown coordinate $j \in [K]$, so that the number of locations is again $K - s + 1$.

Let $M = K - s + 1$ and condition on the protocol's internal randomness and on any past transcript. The next query x is then fixed. For two possible locations j, ℓ, their logits differ by at most $2\beta_- \bar{\gamma} d_s$. Since σ is $1/4$-Lipschitz, their Bernoulli parameters p_j, p_ℓ differ by at most $\beta_- \bar{\gamma} d_s / 2$. The operating-range assumption gives $p_\ell(1 - p_\ell) = \sigma'(z_\ell) \geq c_R$, and hence

$$D_{\mathrm{KL}}(\mathrm{Ber}(p_j) \| \mathrm{Ber}(p_\ell)) \leq \frac{(p_j - p_\ell)^2}{p_\ell(1 - p_\ell)} \leq \frac{(\beta_- \bar{\gamma} d_s)^2}{4 c_R} = \kappa^2.$$

The chain rule for relative entropy therefore bounds the divergence between any two length-Q transcript distributions by $Q\kappa^2$, even for adaptive queries. Under the uniform prior on the M locations, convexity of relative entropy in its second argument bounds the mutual information by $Q\kappa^2$. Fano's inequality at error probability at most $1/3$ gives $Q\kappa^2 \geq (2/3) \log M - \log 2$, proving (iii)(b). The claim is vacuous when $M = 1$ or its numerator is nonpositive. □

Remark 3 (Relation to active elicitation)**.** *General active elicitation methods (Sadigh et al., 2017; Mukherjee et al., 2024; Oh et al., 2025) repeatedly optimize over the full weight vector. CS-Elicit ranks candidate pairs by their certified margins, traverses the dyadic tree, and repeats sign tests to gain confidence. Its number of tree node tests matches the $s \log(K/s)$ counting bound in Theorem 2(iii)(a) up to the factor $\log K / \log(K/s)$. Repetition adds a factor of $b^{-2} \log(\log K/(\Delta_- \delta))$ to the number of comparisons, where b depends on β_-. The error in magnitude estimation depends on s rather than K.*

Corollary 1 (Dense coordinate certification)**.** *For even K and resolution $\Delta \leq 2/K$, any protocol that estimates every $a \in \Delta_{K-1}$ to ℓ_∞ error below $\Delta/2$ with probability at least $2/3$ requires at least $K/3 - 2$ binary answers.*

Proof. Pair coordinates and encode one bit per pair by assigning weights $1/K \pm \Delta/2$. An estimate within $\Delta/2$ recovers all $K/2$ bits. The transcript has entropy at most $Q \log 2$, and Fano's inequality gives the result. The task certified here is dense uniform accuracy, which differs from sparse support recovery in that every coordinate must be resolved. □

Lemma 4 (Two sided control and threshold choice for canonical probes)**.** *Let the new user have scale $\beta > 0$ and let a question use a pair whose probe vector is $x = \gamma_x(\mathbf{1}_G - c\,\mathbf{1}_{[K]}) + e_x$ with $\gamma_x \geq \gamma > 0$ and $\|e_x\|_\infty \leq \eta$; write $m_G = \sum_{k \in G} a^\star_k$ and $p = \sigma(\beta \langle a^\star, x \rangle)$, so that $\langle a^\star, x \rangle = \gamma_x(m_G - c) + \langle a^\star, e_x \rangle$ with $|\langle a^\star, e_x \rangle| \leq \eta$ because $\mathbf{1}^\top a^\star = 1$.*

(i) (Operating range.) *If $\beta |\langle a^\star, x \rangle| \leq R$, then $|p - \frac{1}{2}| \geq c_R \beta \big(\gamma |m_G - c| - \eta\big)$ with $c_R = \sigma'(R)$, and $|p - \frac{1}{2}| \leq \frac{1}{4} \beta |\langle a^\star, x \rangle|$ for every probe.*

(ii) (Threshold.) *Over $c \in (0, \Delta)$ the worst-case net signal $\min\{\gamma c,\ \gamma(\Delta - c)\} - \eta$ is maximized at $c = \Delta/2$, where both hypotheses $m_G = 0$ and $m_G \geq \Delta$ attain $\gamma |m_G - c| \geq \gamma \Delta/2$; the amplitude γ is a common factor and leaves the optimal threshold unchanged.*

Proof. The decomposition of $\langle a^\star, x \rangle$ is $\mathbf{1}^\top a^\star = 1$ together with Hölder's inequality for $\|a^\star\|_1 = 1$. (i) σ' is even and decreasing on $[0, \infty)$, so $|\sigma(z) - \frac{1}{2}| = |\int_0^z \sigma'| \geq |z| \sigma'(R)$ for $|z| \leq R$ and $\leq |z| \sigma'(0) = |z|/4$ for every z. (ii) The two branches of the minimum are linear in c with opposite slopes and coincide at $c = \Delta/2$, and both hypotheses attain equality there. □

Proof of Theorem 2(ii). Write $w^\star = \beta a^\star$, so that $w^\star \geq 0$, $\operatorname{supp} w^\star = S$, $\|w^\star\|_1 = \beta$, and, until the dictionary-error paragraph, $\Pr(y_t = 1) = \sigma(\langle w^\star, x_t \rangle)$, where x_t denotes the frozen contrast $\hat{x}_t$ used by the estimator. Given $\hat{S} = S$, the magnitude pairs are determined by S and the pool, exactly r' per $k \in S$, and their answers are fresh; so, conditionally on the support stage, the $n = sr'$ answers

are independent with these means. Let $\mathcal{W} = \{w \in \mathbb{R}^K_{\geq 0} : \operatorname{supp} w \subseteq S,\ |\langle w, x_t\rangle| \leq R \text{ for all } t\}$, a convex set containing $w^\star$ by Assumption 4, which constrains the frozen contrasts, and let $L_n(w) = \frac{1}{n}\sum_{t=1}^n [\log(1 + e^{\langle w, x_t\rangle}) - y_t\langle w, x_t\rangle]$. Vectors supported on S are treated as elements of $\mathbb{R}^s$.

Design. Let $\bar{x}_t = \gamma(e_{k(t)} - c\mathbf{1}_S)$ be the canonical part of x_t on S, so that $x_{t,S} = \bar{x}_t + e_{t,S}$. Since each $k \in S$ receives n/s probes, $\frac{1}{n}\sum_t \bar{x}_t\bar{x}_t^\top = \frac{\gamma^2}{s}\big[I - (2c - sc^2)\mathbf{1}\mathbf{1}^\top\big]$, whose eigenvalues are γ^2/s on $\mathbf{1}^\perp$ and $\gamma^2(1-sc)^2/s \geq \gamma^2/(4s)$ along $\mathbf{1}$, because $sc \leq 1/2$. For a unit vector $u \in \mathbb{R}^s$ the triangle inequality in ℓ_2 over t gives

$$\Big(\tfrac{1}{n}\textstyle\sum_t \langle x_{t,S}, u\rangle^2\Big)^{1/2} \geq \Big(\tfrac{1}{n}\textstyle\sum_t \langle \bar{x}_t, u\rangle^2\Big)^{1/2} - \max_t |\langle e_{t,S}, u\rangle| \geq \tfrac{\gamma}{2\sqrt{s}} - \sqrt{s}\,\eta_{\mathrm{pool}} \geq \tfrac{\gamma}{4\sqrt{s}},$$

using $|\langle e_{t,S}, u\rangle| \leq \|e_t\|_\infty \|u\|_1 \leq \eta_{\mathrm{pool}}\sqrt{s}$ and $\eta_{\mathrm{pool}} \leq \gamma/(4s)$. Hence $\Sigma_n := \frac{1}{n}\sum_t x_{t,S}x_{t,S}^\top \succeq \frac{\gamma^2}{16s} I$.

Strong convexity. On $\mathcal{W}$, $\nabla^2 L_n(w) = \frac{1}{n}\sum_t \sigma'(\langle w, x_t\rangle) x_{t,S}x_{t,S}^\top \succeq c_R\Sigma_n \succeq \lambda I$ with $\lambda = c_R\gamma^2/(16s)$, because $|\langle w, x_t\rangle| \leq R$ and σ' is even and decreasing on $[0, \infty)$.

Gradient at the truth. $g := \nabla L_n(w^\star) = \frac{1}{n}\sum_t (\sigma(\langle w^\star, x_t\rangle) - y_t)x_{t,S}$ has mean zero, and its j-th coordinate is a sum of independent terms whose t-th term ranges over an interval of length $|x_{t,j}|/n$. For the n/s probes of type j, $|x_{t,j}| \leq \gamma(1-c) + \eta_{\mathrm{pool}} \leq 2\gamma$; for the others, $|x_{t,j}| \leq \gamma c + \eta_{\mathrm{pool}} \leq \gamma/s$. Hence $\sum_t |x_{t,j}|^2/n^2 \leq (4\gamma^2/s + \gamma^2/s^2)/n \leq 5\gamma^2/(ns)$, and Hoeffding's inequality with a union bound over $j \in S$ gives, with probability at least $1 - \delta$, $|g_j| \leq \gamma\sqrt{5\log(2s/\delta)/(2ns)}$ for all j, so $\|g\|_2 \leq \gamma\sqrt{5\log(2s/\delta)/(2n)}$.

Estimation error. Since $\hat{w}$ minimizes L_n over the convex set $\mathcal{W} \ni w^\star$, first-order optimality gives $\langle \nabla L_n(\hat{w}), w^\star - \hat{w}\rangle \geq 0$. Adding the two strong-convexity inequalities $L_n(w^\star) \geq L_n(\hat{w}) + \langle \nabla L_n(\hat{w}), w^\star - \hat{w}\rangle + \frac{\lambda}{2}\|w^\star - \hat{w}\|_2^2$ and $L_n(\hat{w}) \geq L_n(w^\star) + \langle g, \hat{w} - w^\star\rangle + \frac{\lambda}{2}\|\hat{w} - w^\star\|_2^2$ yields $\lambda\|\hat{w} - w^\star\|_2^2 \leq -\langle g, \hat{w} - w^\star\rangle \leq \|g\|_2\|\hat{w} - w^\star\|_2$, hence

$$\|\hat{w} - w^\star\|_2 \leq \frac{\|g\|_2}{\lambda} \leq \frac{16s}{c_R\gamma}\sqrt{\frac{5\log(2s/\delta)}{2n}}.$$

Finally $\|\hat{w} - w^\star\|_1 \leq \sqrt{s}\,\|\hat{w} - w^\star\|_2$, and for $\hat{a} = \hat{w}/\|\hat{w}\|_1$ and $a^\star = w^\star/\|w^\star\|_1$,

$$\|\hat{a} - a^\star\|_1 \leq \frac{\|\hat{w} - w^\star\|_1}{\|w^\star\|_1} + \|\hat{w}\|_1\Big|\frac{1}{\|\hat{w}\|_1} - \frac{1}{\|w^\star\|_1}\Big| \leq \frac{2\|\hat{w} - w^\star\|_1}{\beta},$$

which gives the bound with constant $32\sqrt{5/2} < 52$ and $\beta \geq \beta_-$. On the same event, $\hat{w} \neq 0$ whenever the right-hand side is below 2, because then $\|\hat{w} - w^\star\|_1 < \|w^\star\|_1$, so $\hat{a}$ is defined.

Dictionary error. If the contrasts used by the estimator are x_t while the user responds to x_t^{true} with $\|x_t^{\mathrm{true}} - x_t\|_\infty \leq \eta_{\mathrm{dict}}$, written η in this paragraph, the true logit is $\langle w^\star, x_t\rangle + \epsilon_t$ with $|\epsilon_t| \leq \|w^\star\|_1\eta = \beta\eta$. Let $\bar{L}$ be the expectation of L_n under the true responses and $\bar{w}$ its minimizer over $\mathcal{W}$. Then $\bar{L}$ has the same Hessian as L_n and is λ-strongly convex on $\mathcal{W}$, and $\nabla\bar{L}(w^\star) = \frac{1}{n}\sum_t(\sigma(\langle w^\star, x_t\rangle) - \sigma(\langle w^\star, x_t\rangle + \epsilon_t))x_{t,S}$ has coordinates bounded by $\frac{\beta\eta}{4}\big[\frac{1}{s}(\gamma(1-c) + \eta_{\mathrm{pool}}) + \gamma c + \eta_{\mathrm{pool}}\big] \leq \frac{\beta\eta\gamma}{2s}$, so $\|\nabla\bar{L}(w^\star)\|_2 \leq \beta\eta\gamma/(2\sqrt{s})$ and, by the argument above applied to $\bar{L}$, $\|\bar{w} - w^\star\|_2 \leq 8\beta\eta\sqrt{s}/(c_R\gamma)$. The difference $L_n - \bar{L}$ is linear in w with gradient $\xi = \frac{1}{n}\sum_t(\mathbb{E}y_t - y_t)x_{t,S}$, which obeys the same Hoeffding bound as g; the optimality argument applied at $\bar{w}$ gives $\lambda\|\hat{w} - \bar{w}\|_2 \leq \|\xi\|_2$. The triangle inequality and the normalization step then add $16s\eta/(c_R\gamma)$ to the bound. □

D From Factor Identification to Elicitation

This appendix proves Propositions 2 and 1 and Theorem 3. It also gives the quantitative stability result for approximate anchors used in these results. Throughout, $\|\cdot\|_{1\to 1}$ and $\|\cdot\|_{\infty\to\infty}$ denote the operator norms on $\mathbb{R}^{K\times K}$ induced by $\|\cdot\|_1$ and $\|\cdot\|_\infty$. They equal the maximum absolute column sum and maximum absolute row sum, respectively.

D.1 COORDINATE INVARIANCE AND COVERAGE

For $v \in \mathbb{R}^K$ with $\mathbf{1}^\top v > 0$ write $\mathsf{N}(v) = v/(\mathbf{1}^\top v)$. Let $\tilde{f} = Mf$ be a normalized K-factor parameterization behaviorally equivalent to f on the population (Lemma 3). A reward $\beta a^\top f$ with $a \in \Delta_{K-1}$ is rewritten as $\tilde{\beta}\, \phi_M(a)^\top \tilde{f}$ with

$$\phi_M(a) = \mathsf{N}(M^{-\top} a), \qquad \tilde{\beta} = \beta\, \mathbf{1}^\top M^{-\top} a, \tag{4}$$

whenever $\mathbf{1}^\top M^{-\top} a > 0$. The map ϕ_M gives a new user's composition in the alternative coordinates.

Proposition 2 (Coordinate invariance of support, margin, and scale)**.** *Let $\tilde{f} = Mf$ be as above. The following are equivalent:*

(i) M is a permutation matrix;

(ii) for every $k \in [K]$, $\phi_M(e_k)$ is defined and is a vertex of Δ_{K-1};

(iii) for every $a \in \Delta_{K-1}$, $\phi_M(a)$ is defined, lies in Δ_{K-1}, and $|\operatorname{supp} \phi_M(a)| = |\operatorname{supp} a|$.

Under (i), $\phi_M(a) = Ma$ and $\tilde{\beta} = \beta$. The transformation relabels the support while preserving the sparsity s, margin Δ, and scale β. CS-Elicit, when run in the $\tilde{f}$ coordinates with the relabeled tree, performs the same tests with the same probe signal (γ, η). Therefore, for an anchored population, the quantities S, s, Δ, and β of a new user are invariants of that user's comparison kernel across all anchored parameterizations (Theorem 1(i)).

Proof of Proposition 2. Since $\tilde{f} = Mf$, $\beta a^\top f = \beta a^\top M^{-1} \tilde{f} = \beta (M^{-\top} a)^\top \tilde{f} = \beta (\mathbf{1}^\top M^{-\top} a)\, \mathsf{N}(M^{-\top} a)^\top \tilde{f}$, which is equation 4.

(i)⇒(iii). If M is a permutation matrix then $M^{-\top} = M$, $\mathbf{1}^\top M a = 1$, and $\phi_M(a) = Ma$ permutes coordinates, which preserves membership in Δ_{K-1} and support size; $\tilde{\beta} = \beta \mathbf{1}^\top M a = \beta$. The dyadic tree over the $\tilde{f}$ labels is the relabeling of a tree over the f labels, and a pair with $\tilde{f}$-contrast $\tilde{x} = \gamma(\mathbf{1}_{\tilde{G}} - c\mathbf{1}) + \tilde{e}$ has f-contrast $x = M^{-1}\tilde{x} = M^\top \tilde{x} = \gamma(\mathbf{1}_G - c\mathbf{1}) + M^\top \tilde{e}$ with $G = \{k : M_{jk} = 1,\ j \in \tilde{G}\}$, $M^\top \mathbf{1} = \mathbf{1}$, and $\|M^\top \tilde{e}\|_\infty = \|\tilde{e}\|_\infty$; so the same tests are performed with the same signal, and the budget $2s\lceil \log_2 K \rceil r$ of Theorem 2(i), which holds for any labeling, is unchanged.

(iii)⇒(ii). e_k is 1-sparse, so $\phi_M(e_k)$ is a 1-sparse point of Δ_{K-1}, i.e. a vertex.

(ii)⇒(i). $\phi_M(e_k) = \mathsf{N}(M^{-\top} e_k) = e_{j(k)}$ with $\mathbf{1}^\top M^{-\top} e_k > 0$ means that column k of $M^{-\top}$ equals $c_k e_{j(k)}$ with $c_k > 0$. Invertibility forces $k \mapsto j(k)$ to be a bijection, so $M^{-\top} = P \operatorname{diag}(c)$ where P is the permutation matrix with $P_{j(k),k} = 1$. Hence $M = P \operatorname{diag}(c)^{-1}$ and $\tilde{f}_{j(k)} = f_k / c_k$; unit norms give $c_k = 1$, so $M = P$.

The final claim follows from Theorem 1(i): in an anchored population every anchored behaviorally equivalent parameterization has a permutation matrix M. □

Proof of Proposition 1. R is invertible because the population loadings span $\mathbb{R}^K$ and lie in $\operatorname{cone}\{\rho_j\}$ (as in the proof of Theorem 1(ii)). In matrix form $\hat{g} = D^{-1} R^\top f$, so $f = R^{-\top} D \hat{g}$ and $\beta a^{\star\top} f = \beta a^{\star\top} R^{-\top} D \hat{g} = (\beta D R^{-1} a^\star)^\top \hat{g}$; uniqueness of coefficients in a linearly independent system gives $\hat{w} = \beta D R^{-1} a^\star$. Since D is a positive diagonal matrix, $\hat{w} \geq 0$ iff $R^{-1} a^\star \geq 0$ iff $a^\star = Rc$ for some $c \geq 0$, which is the first equivalence in (i).

First note a support-face fact. If $a^\star = \sum_j c_j \rho_j$ with $c \geq 0$, then for $k \notin S$ the identity $0 = a^\star_k = \sum_j c_j (\rho_j)_k$ is a sum of nonnegative terms, so $c_j (\rho_j)_k = 0$ for every j; every j with $c_j > 0$ therefore has $\operatorname{supp} \rho_j \subseteq S$. The converse inclusion is trivial, and the two inclusions give the second equivalence in (i).

(ii) If $\{e_k : k \in S\} \subseteq \{\rho_j\}$ then $F_S = \operatorname{conv}\{e_k : k \in S\} \subseteq \operatorname{cone}\{\rho_j\}$. Conversely, $e_k \in F_S \subseteq \operatorname{cone}\{\rho_j\}$ and the support-face fact applied with $S = \{k\}$ give $e_k \in \operatorname{cone}\{\rho_j : \operatorname{supp} \rho_j \subseteq \{k\}\}$; the only point of Δ_{K-1} supported in $\{k\}$ is e_k, so some $\rho_j = e_k$. Taking $S = [K]$, every $a^\star \in \Delta_{K-1}$

is representable iff $\{e_k\} \subseteq \{\rho_j\}$, i.e. $\{\rho_j\} = \{e_k\}$ since both sets have K elements. Finally, A is compact and $C = \operatorname{conv} A$. If the population is anchored then each $e_k \in A \subseteq C \subseteq \Delta_{K-1}$ is an extreme point of Δ_{K-1} and hence of C, so $\{e_k\} \subseteq \operatorname{ext} C = \{\rho_j\}$. Conversely, the extreme points of the convex hull of a compact set belong to the set, so $\{\rho_j\} = \{e_k\}$ implies $e_k \in A$ for every k, i.e. the population is anchored.

(iii) follows from the support-face fact with $\{\rho_j : \operatorname{supp} \rho_j \subseteq S\} = \emptyset$, whose cone is $\{0\} \not\ni a^\star$.

For the last sentence, if $\hat{a}_l < 0$ then $\sum_{k \in \{l\}} \hat{a}_k < 0$ and $\sum_{k \neq l} \hat{a}_k = 1 - \hat{a}_l > 1$. □

D.2 Quantitative stability under approximate anchors

The transfer bounds use the constants defined in the main text:

$$\varepsilon_1(\nu) = \frac{4\nu(1-\nu)}{1-2\nu}, \qquad \zeta(\nu) = \frac{2\varepsilon_1(\nu)}{1-\varepsilon_1(\nu)}. \tag{5}$$

Lemma 5 (Extreme users of a ν-anchored population are approximate anchors)**.** *Let $C = \operatorname{conv} \operatorname{cl}\{a_i : i \in I\}$, and suppose the population is ν-anchored with $\nu < 1/2$. For each $k \in [K]$, there is an extreme point ρ of C with $\rho_k \geq 1 - \nu$. The points obtained for different k are distinct. If C has exactly K extreme points, they can be labeled so that $(\rho_k)_k \geq 1 - \nu$ for every k, and the anchored representation in Theorem 1(ii) is $\hat{g}_k = \rho_k^\top f / \|\rho_k^\top f\|$.*

Proof. C is compact and convex. The linear functional $a \mapsto a_k$ attains its maximum over C, and the set of maximizers is a nonempty compact convex face of C; by Minkowski's theorem that face is the convex hull of its extreme points, and an extreme point of a face is an extreme point of C. Since some population user has $a_{ik} \geq 1 - \nu$, the maximum is at least $1 - \nu > 1/2$. A point of Δ_{K-1} cannot have two coordinates exceeding $1/2$, so the maximizers obtained for different k are different. If $|\operatorname{ext} C| = K$, the K distinct maximizers exhaust the extreme points, and relabeling ρ_k as the maximizer of the k-th coordinate gives $(\rho_k)_k \geq 1 - \nu$. The last claim is the construction in the proof of Theorem 1(ii). □

Proposition 3 (Transfer of dictionary, composition, and scale under approximate anchors)**.** *Let $\rho_1, \ldots, \rho_K \in \Delta_{K-1}$ satisfy $(\rho_k)_k \geq 1 - \nu$ with $\nu < 1/2$, and write $\rho_k = (1 - \nu_k)e_k + \nu_k q_k$ with $\nu_k = 1 - (\rho_k)_k \in [0, \nu]$ and $q_k \in \Delta_{K-1}$, $(q_k)_k = 0$. Let $R = [\rho_1 \cdots \rho_K]$, $d_k = \|\rho_k^\top f\|_{L^2(\mu)}$, $D = \operatorname{diag}(d_1, \ldots, d_K)$, and $\hat{g}_k = \rho_k^\top f / d_k$.*

(a) R is invertible with $\|R^{-1} - I\|_{1\to 1} \leq 2\nu/(1-2\nu)$; $d_k \in [1 - 2\nu_k, 1]$; the $\hat{g}_k$ are centered, unit norm, and linearly independent; and

$$\|\hat{g}_k - f_k\|_{L^2(\mu)} \leq 4\nu_k \leq 4\nu \qquad (k \in [K]).$$

(b) Every reward $r = \beta a^\top f$ with $\beta > 0$ and $a \in \Delta_{K-1}$ equals $\hat{w}^\top \hat{g}$ with $\hat{w} = \beta D R^{-1} a$. If $\varepsilon_1(\nu) < 1$, then $\hat{\beta} := \mathbf{1}^\top \hat{w} > 0$ and $\hat{a} := \hat{w}/\hat{\beta}$ satisfies

$$\mathbf{1}^\top \hat{a} = 1, \qquad \|\hat{a} - a\|_1 \leq \zeta(\nu), \qquad \|\hat{a}\|_1 \leq 1 + \zeta(\nu), \qquad |\hat{\beta}/\beta - 1| \leq \varepsilon_1(\nu),$$

with ε_1, ζ as in equation 5; for $\nu \leq 1/20$, $\varepsilon_1(\nu) \leq 4.3\nu$ and $\zeta(\nu) \leq 11\nu$. If moreover $a \in \operatorname{cone}\{\rho_1, \ldots, \rho_K\}$, then $\hat{a} \in \Delta_{K-1}$.

Proof. (a) Column k of $E := I - R$ is $e_k - \rho_k = \nu_k(e_k - q_k)$, whose ℓ_1 norm is $2\nu_k$ because $(q_k)_k = 0$ and $\mathbf{1}^\top q_k = 1$. Hence $\|E\|_{1\to 1} = \max_k 2\nu_k \leq 2\nu < 1$, so $R = I - E$ is invertible, $R^{-1} = \sum_{m \geq 0} E^m$, and $\|R^{-1} - I\|_{1\to 1} \leq \sum_{m \geq 1} (2\nu)^m = 2\nu/(1-2\nu)$. Next, $\|q_k^\top f\| \leq \sum_l (q_k)_l \|f_l\| = 1$, so $d_k = \|(1-\nu_k) f_k + \nu_k q_k^\top f\|$ lies in $[(1-\nu_k) - \nu_k, (1-\nu_k) + \nu_k] = [1 - 2\nu_k, 1]$. Centering of $\hat{g}_k$ follows from $\mathbb{E}_\mu f = 0$, unit norm holds by construction, and linear independence follows from $\hat{g} = D^{-1} R^\top f$ with $D^{-1} R^\top$ invertible. Finally, $u_k := \rho_k^\top f$ satisfies $\|u_k - f_k\| = \nu_k \|q_k^\top f - f_k\| \leq 2\nu_k$ and $|1 - d_k| = \big|\|f_k\| - \|u_k\|\big| \leq \|f_k - u_k\|$, so

$$\|\hat{g}_k - f_k\| \leq \|u_k/d_k - u_k\| + \|u_k - f_k\| = |1 - d_k| + \|u_k - f_k\| \leq 4\nu_k.$$

(b) From $\hat{g} = D^{-1}R^\top f$ we get $f = R^{-\top}D\hat{g}$, hence $\beta a^\top f = \beta a^\top R^{-\top}D\hat{g} = (\beta DR^{-1}a)^\top \hat{g}$. Let $v = DR^{-1}a$. Using $\|a\|_1 = 1$, $\|D\|_{1\to1} = \max_k d_k \le 1$, and $\|D - I\|_{1\to1} = \max_k |1 - d_k| \le 2\nu$,

$$\|v-a\|_1 \le \|D(R^{-1}-I)a\|_1 + \|(D-I)a\|_1 \le \|D\|_{1\to1}\|R^{-1}-I\|_{1\to1} + \|D-I\|_{1\to1} \le \frac{2\nu}{1-2\nu} + 2\nu = \varepsilon_1(\nu).$$

Hence $t := \mathbf{1}^\top v - 1$ satisfies $|t| \le \|v - a\|_1 \le \varepsilon_1$, so $\hat{\beta} = \beta(1 + t) \ge \beta(1 - \varepsilon_1) > 0$ and $|\hat{\beta}/\beta - 1| \le \varepsilon_1$. With $\hat{a} = v/(1 + t)$ and $1 + t \ge 1 - \varepsilon_1$,

$$\|\hat{a} - a\|_1 \le \frac{\|v - a\|_1}{1 + t} + \|a\|_1\Big|\frac{1}{1 + t} - 1\Big| \le \frac{\varepsilon_1}{1 - \varepsilon_1} + \frac{\varepsilon_1}{1 - \varepsilon_1} = \zeta(\nu),$$

$\|\hat{a}\|_1 \le \|a\|_1 + \|\hat{a} - a\|_1 \le 1 + \zeta$, and $\mathbf{1}^\top\hat{a} = \mathbf{1}^\top v/(1 + t) = 1$. The numerical constants follow because $\varepsilon_1(\nu)/\nu = 4(1 - \nu)/(1 - 2\nu)$ and $\zeta(\nu)/\varepsilon_1(\nu) = 2/(1 - \varepsilon_1(\nu))$ are increasing in ν, and $\varepsilon_1(1/20) = 19/90$, $\zeta(1/20) = 38/71$, so that $\varepsilon_1(\nu) \le \frac{38}{9}\nu \le 4.3\nu$ and $\zeta(\nu) \le \frac{760}{71}\nu \le 11\nu$ on $(0, 1/20]$. If $a \in \mathrm{cone}\{\rho_j\}$ then $R^{-1}a \ge 0$, so $v \ge 0$ and $\hat{a} \ge 0$.

□

Proposition 3(a) provides the bound used in Theorem 1(iii). Its constants are independent of K and the Gram matrix of f because composition transfer is an identity in loading space, while $L^2(\mu)$ enters only through $d_k \in [1 - 2\nu, 1]$. Part (b) shows that with approximate anchors, every sparse new user remains approximately sparse in the extracted coordinates, although the remaining tail may contain negative entries.

D.3 Robust probes and the full guarantee

Lemma 6 (Group mass probe with a signed tail)**.** *Let a new user's reward be $\hat{\beta}\,\hat{a}^\top\hat{g}$ with $\hat{\beta} > 0$, $\mathbf{1}^\top\hat{a} = 1$, and $\|\hat{a} - a^\circ\|_1 \le \zeta$ for some $a^\circ \in \Delta_{K-1}$. For a pair whose $\hat{g}$-contrast is $\hat{x} = \gamma(\mathbf{1}_G - c\mathbf{1}) + e$ with $\|e\|_\infty \le \eta$, the response probability $p = \sigma\Big(\hat{\beta}\langle\hat{a}, \hat{x}\rangle\Big)$ satisfies, with $\hat{m}_G = \sum_{k\in G}\hat{a}_k$ and $m^\circ_G = \sum_{k\in G} a^\circ_k$,*

$$|\hat{m}_G - m^\circ_G| \le \zeta, \qquad \mathrm{sgn}(p - \tfrac{1}{2}) = \mathrm{sgn}(\hat{m}_G - c) \ \textit{whenever}\ \gamma|\hat{m}_G - c| > (1 + \zeta)\eta,$$

and in that case $|p - \frac{1}{2}| \ge \frac{1}{6}\min\big\{\hat{\beta}\big[\gamma|\hat{m}_G - c| - (1 + \zeta)\eta\big], 1\big\}$.

Proof. Since $\mathbf{1}^\top\hat{a} = 1$, $\langle\hat{a}, \hat{x}\rangle = \gamma(\hat{m}_G - c) + \langle\hat{a}, e\rangle$, and $|\langle\hat{a}, e\rangle| \le \|\hat{a}\|_1\|e\|_\infty \le (1 + \zeta)\eta$ because $\|\hat{a}\|_1 \le \|a^\circ\|_1 + \|\hat{a} - a^\circ\|_1$. The sign claim follows because σ is increasing with $\sigma(0) = 1/2$ and $\hat{\beta} > 0$. The tail bound is $|\sum_{k\in G}(\hat{a}_k - a^\circ_k)| \le \|\hat{a} - a^\circ\|_1$. For the last claim, $|p - \frac{1}{2}| = \sigma(|z|) - \frac{1}{2}$ with $|z| = \hat{\beta}|\langle\hat{a}, \hat{x}\rangle| \ge \hat{\beta}[\gamma|\hat{m}_G - c| - (1 + \zeta)\eta]$, and $\sigma(z) - \frac{1}{2} = \frac{1}{2}\tanh(z/2) \ge \frac{1}{6}\min\{z, 1\}$ for $z \ge 0$: by concavity of $\tanh$ on $[0, \infty)$, $\tanh(z/2) \ge z\tanh(1/2) \ge z/3$ for $z \in [0, 1]$, and $\tanh(z/2) \ge \tanh(1/2) > 1/3$ for $z \ge 1$. □

Theorem 3 uses the following bias bound and repetition count:

$$\begin{aligned} b' &= \tfrac{1}{6}\min\Big\{\beta_-\big(1 - \varepsilon_1(\nu)\big)\Big[\gamma\big(\tfrac{\Delta}{2} - \zeta(\nu)\big) - \big(1 + \zeta(\nu)\big)\eta\Big],\ 1\Big\}, \\ \bar{T} &= 2\lceil\log_2 K\rceil\lceil 1/\Delta\rceil, \quad r_\nu = \Big\lceil\frac{\log(2\bar{T}/\delta)}{2b'^2}\Big\rceil. \end{aligned} \tag{6}$$

Proof of Theorem 3. Step 1 (the new user in the extracted coordinates). Since $\varepsilon_1(1/20) < 1$, Proposition 3(b) applied to $r^\star = \beta a^{\star\top} f$ gives $r^\star = \hat{\beta}\hat{a}^\top\hat{g}$ with $\mathbf{1}^\top\hat{a} = 1$, $\|\hat{a} - a^\star\|_1 \le \zeta := \zeta(\nu)$, and $\hat{\beta} \ge \beta(1 - \varepsilon_1(\nu)) \ge \beta_-(1 - \varepsilon_1(\nu)) > 0$.

Step 2 (every group is well separated from the threshold). Let $G \subseteq [K]$ and $c = \Delta/2$. If $G \cap S = \emptyset$, then $|\hat{m}_G| \le \sum_{k\notin S}|\hat{a}_k| = \sum_{k\notin S}|\hat{a}_k - a^\star_k| \le \zeta$, so $\hat{m}_G - c \le -(\Delta/2 - \zeta) < 0$ by equation 2. If $G \cap S \ne \emptyset$, then $m^\star_G \ge \Delta$ and Lemma 6 gives $\hat{m}_G \ge \Delta - \zeta$, so $\hat{m}_G - c \ge \Delta/2 - \zeta > 0$. In both cases $|\hat{m}_G - c| \ge \Delta/2 - \zeta$. Each pair asked for G has $\tilde{f}$-contrast $\gamma(\mathbf{1}_G - c\mathbf{1}) + e$ and $\hat{g}$-contrast $\hat{x} = \gamma(\mathbf{1}_G - c\mathbf{1}) + e + d$ with $\|d\|_\infty \le \eta_{\mathrm{est}}$, so $\|e + d\|_\infty \le \eta_{\mathrm{pool}} + \eta_{\mathrm{est}} = \eta$; the user responds to

$\hat{x}$ because $r^\star = \hat{\beta}\hat{a}^\top \hat{g}$. By equation 2 and Lemma 6 the probe for G has $p_G > \frac{1}{2}$ iff G meets S, and $|p_G - \frac{1}{2}| \geq b'$.

Step 3 (tests and questions). Let $\mathcal{T}$ be the dyadic tree over $[K]$, of depth $L = \lceil \log_2 K \rceil$ with the singletons as leaves, and let $\mathcal{G}_S$ be the set of children of nodes that meet S. At each level at most s disjoint nodes meet S, so $|\mathcal{G}_S| \leq 2sL$; moreover $s\Delta \leq \sum_{k\in S} a^\star_k \leq 1$ gives $s \leq \lceil 1/\Delta \rceil$, so $|\mathcal{G}_S| \leq \bar{T}$. For each $G \in \mathcal{G}_S$ let $\bar{y}_G$ be the mean of the answers to its r_ν pairs, which are fixed by G and the pool; by the answer-table construction in the proof of Theorem 2(i) these answers are independent Bernoulli variables whose parameters all lie on the side of $\frac{1}{2}$ found in Step 2, at distance at least b' from it; write p_G for their average. By Hoeffding's inequality and a union bound over the fixed set $\mathcal{G}_S$, with probability at least $1 - \bar{T} \cdot 2e^{-2r_\nu b'^2} \geq 1 - \delta$ we have $|\bar{y}_G - p_G| < b'$ for all $G \in \mathcal{G}_S$. On this event every test in $\mathcal{G}_S$ returns the correct verdict, $\operatorname{sgn}(\bar{y}_G - \frac{1}{2}) = \operatorname{sgn}(p_G - \frac{1}{2})$, so CS-Elicit (which tests the two children of every positive node, the root being positive without a test because $\hat{m}_{[K]} = 1$) tests exactly the nodes of $\mathcal{G}_S$, follows every root-to-leaf path to each singleton $\{k\}$ with $k \in S$, and rejects every tested singleton outside S. It outputs S after at most $|\mathcal{G}_S|\, r_\nu \leq 2s\lceil \log_2 K \rceil r_\nu$ questions.

Step 4 (boundary statements). Step 1 uses only $(\rho_k)_k \geq 1 - \nu$, not extremality of the ρ_k. At $\nu = 0$ and $\eta_{\text{est}} = 0$ we have $\varepsilon_1 = \zeta = 0$, equation 2 reads $\gamma\Delta > 2\eta$, and $b' = \frac{1}{6}\min\{\beta_-(\gamma\Delta/2 - \eta), 1\} \leq \sigma(\beta_-(\gamma\Delta/2 - \eta)) - \frac{1}{2} = b$ by Lemma 6, so the count is that of Theorem 2(i) with b replaced by this smaller bound. The converse statements are Proposition 1(ii) and (iii). □

Remark 4 (Probes specified in the true coordinates)**.** *If Assumption 3 is imposed in the true coordinates, contrasts transfer by $\hat{x} = D^{-1}R^\top x$. In the setting of Proposition 3, row k of $D^{-1}R^\top - I$ equals $\big(\frac{1-\nu_k}{d_k} - 1\big)e_k^\top + \frac{\nu_k}{d_k} q_k^\top$, whose absolute sum is at most $\frac{\nu_k}{1-2\nu_k} + \frac{\nu_k}{1-2\nu_k}$ because $d_k \in [1 - 2\nu_k, 1]$. Hence $\|D^{-1}R^\top - I\|_{\infty\to\infty} \leq \varepsilon_\infty(\nu) := 2\nu/(1-2\nu)$ and $\|D^{-1}R^\top\|_{\infty\to\infty} = \max_k 1/d_k \leq 1 + \varepsilon_\infty(\nu)$. A pair with true contrast $x = \gamma(\mathbf{1}_G - c\mathbf{1}) + e$, $c \leq 1/2$, therefore has $\hat{g}$-contrast $\hat{x} = \gamma(\mathbf{1}_G - c\mathbf{1}) + \hat{e}$ with*

$$\|\hat{e}\|_\infty \leq (1 + \varepsilon_\infty(\nu))\,\eta + \varepsilon_\infty(\nu)(1-c)\,\gamma, \qquad \|\hat{x}\|_\infty \leq (1 + \varepsilon_\infty(\nu))\|x\|_\infty.$$

Theorem 3 then holds with η replaced by $(1 + \varepsilon_\infty(\nu))\eta + \varepsilon_\infty(\nu)(1-c)\gamma$ and $\bar{\gamma}$ by $(1 + \varepsilon_\infty(\nu))\bar{\gamma}$; the coupling condition becomes $\gamma\Delta > 2\eta + O(\nu)(\gamma + \eta)$ with these explicit constants.

Remark 5 (General valid probes in the extracted coordinates)**.** *Theorem 3 holds for arbitrary valid probes in the $\hat{g}$ coordinates. If every tested group uses pairs with $\vartheta(\hat{x}; G, \Delta) \geq \vartheta_0$ and $\|\hat{x}\|_\infty \leq \bar{\gamma}$, then equation 2 becomes $\vartheta_0 > \zeta(\nu)\bar{\gamma}$ and b' becomes $\frac{1}{6}\min\{\beta_-(1 - \varepsilon_1(\nu))(\vartheta_0 - \zeta(\nu)\bar{\gamma}), 1\}$, because $\langle \hat{a}, \hat{x}\rangle$ differs from $\langle a^\star, \hat{x}\rangle$ by at most $\|\hat{a} - a^\star\|_1 \|\hat{x}\|_\infty \leq \zeta(\nu)\bar{\gamma}$ and Lemma 1 applies to $a^\star$. The canonical instance, with $\vartheta_0 = \gamma\Delta/2 - \eta$ and $\bar{\gamma} = \gamma + \eta$, is equation 2.*

D.4 Scope of Theorem 3 and the gap for finite samples

Theorem 3 analyzes an idealized extraction. It assumes that (a) the K dictionary directions are the normalized rewards of users with purity $(\rho_k)_k \geq 1 - \nu$, and (b) for the pairs asked by CS-Elicit, the working dictionary predicts the contrasts of these rewards within η_{est}. The selected users need not be extreme: Step 1 of the proof uses only their purity through Proposition 3(b). Given exact rewards from a ν-anchored population whose hull has exactly K extreme points, Algorithm 2 satisfies (a) by Lemma 5 and (b) with $\eta_{\text{est}} = 0$. A known lower bound $\Delta_- \leq \Delta$ may replace Δ throughout, and Remark 4 covers probes specified in the true coordinates.

The implemented procedure estimates rewards from finitely many comparisons. We do not prove that it satisfies (a) or (b) with high probability. Such a guarantee would require the following three components, and our analysis supplies only part of the first.

(i) *Utility estimation.* Suppose the factors lie in the span of ϕ and the parameters and features for each user are bounded. Maximum likelihood bounds for Bradley–Terry models (Zhu et al., 2023) then control $\|\widehat{v}_i - v_i\|$ in the empirical design norm at rate $\sqrt{(d + \log(1/\delta))/n_i}$. By the Cauchy–Schwarz inequality, this result controls the contrast error of each pair through $\|\phi(\tau) - \phi(\tau')\|$ measured in the inverse design norm. Normalization by $\|\widehat{r}_i\|$ adds a relative error of the same order. These bounds can be transferred to our ridge estimator, but they do not yield η_{est}. If the factors lie outside the span of ϕ, a bias remains as n_i grows. Appendix F.8 shows that this bias is large for our language features.

(ii) *Ray selection.* We have no guarantee that screening with tolerance ϵ_0 and greedy residual selection returns users with purity at least $1 - \nu$ when the rewards are noisy. Estimation noise can move interior users outside the hull and create spurious extreme points. Robustness results for successive projection (Gillis & Vavasis, 2013) analyze a different selection rule and a separable model with additive noise. Their constants depend on the conditioning of the factor Gram matrix, whereas the constants in Proposition 3 do not.

(iii) *Composition.* Even if (i) and (ii) hold, η_{est} and the realized ν must satisfy equation 2. At $\nu = 0.05$, this condition already requires $\Delta > 1.07$, which no composition can satisfy. In our PersonalLLM extraction, the selected users have purity as low as 0.37 (Appendix F.8).

A guarantee for the complete pipeline with finite samples therefore remains open. Theorem 3 isolates the error caused by impure anchors, but it does not control estimation or selection error.

E Position Relative to Prior Work

Table 4 records, for each line of work discussed in Section 2, which of the features of this paper it provides. A dash marks a partial answer or a question that does not arise in that line of work.

Table 4: Prior work against the features of this paper. ✓: provided; ✗: absent; –: partial, or the question does not arise in that line of work. The column on the user's scale refers to the Bradley–Terry temperature of each user. Our guarantee in extracted coordinates is partial: it assumes exact extreme-user rewards and does not cover finite-sample extraction (Appendix D.4).

Line of work	Shared factors with user weights	Factor basis identified beyond its span	Unknown per-user scale handled	Representability of sparse new users	New-user elicitation guarantee	Questions logarithmic in K	Scale-free decision rule	Lower bounds for elicitation	Guarantee in extracted coordinates
Latent-user variational reward models (Poddar et al., 2024; Kim & Kim, 2026; Lee et al., 2026)	–	✗	–	✗	✗	✗	✗	✗	✗
Prototype and ideal-point models (Chen et al., 2025)	✓	✗	–	✗	✗	✗	✗	✗	✗
Low-rank and factorized rewards (Bose et al., 2025; Shenfeld et al., 2025; Barreto et al., 2026; Luo et al., 2025)	✓	✗	–	✗	–	✗	✗	✗	✗
Attribute-supervised and mixture-of-experts rewards (Li et al., 2025; Wang et al., 2026)	✓	–	–	✗	✗	✗	✗	✗	✗
Theory of personalized alignment (Kang, 2026; Chakraborty et al., 2024; Park et al., 2024; Zhong et al., 2024)	✓	–	✗	✗	✗	✗	✗	–	✗
Separable and minimum-volume factorization (Donoho & Stodden, 2003; Arora et al., 2012; Gillis & Vavasis, 2013; Fu et al., 2018)	✓	✓	–	✗	✗	✗	✗	✗	✗
Active elicitation of a dense reward (Sadigh et al., 2017; Mukherjee et al., 2024; Ge et al., 2024)	–	✗	–	✗	✓	✗	✗	✓	✗
Active elicitation for language-model users (Oh et al., 2025; Shenfeld et al., 2025; Pădurean et al., 2026)	✓	✗	✗	✗	–	✗	✗	✗	✗
Sparse recovery from binary tests (Hwang, 1972; Aldridge et al., 2019; Malloy & Nowak, 2014; Plan & Vershynin, 2013)	✗	✗	–	✗	–	✓	–	✓	✗
This paper	✓	✓	✓	✓	✓	✓	✓	✓	–

F Detailed Experiments

F.1 Setup

Shared protocol. PersonalLLM uses the standardized scores of ten published reward models as factors, and UltraFeedback uses four GPT-4 aspect ratings. The control tasks use normalized observation coordinates for MountainCar, CartPole, Acrobot, and LunarLander, and moving object joints for Franka Kitchen. UF-P-4 and TL;DR provide recorded labels without factor scores. Semi-synthetic users are s-sparse or Dirichlet compositions $a \in \Delta_{K-1}$ that prefer x to x' with probability $\sigma(\beta\, a^\top (f(x) - f(x')))$. We set $\beta = 5$ for language elicitation, $\beta = 1$ for the stability study, and calibrate β to a population Bayes accuracy of 0.80 on control tasks. Seen users train the shared models, validation users select hyperparameters, and unseen users are reserved for evaluation. We fix all settings on development seeds and report results from fresh seeds.

We represent a response x by the final-token hidden state $h(x)$ of Skywork-Reward-V2-Llama-3.1-8B, reduced to 512 dimensions by PCA fitted on training prompts. A trajectory is represented by the coordinate-wise mean of its normalized observations. The oracle dictionary is the identity in these coordinates; on PersonalLLM, it is the ridge readout of the reward-model scores from $h(x)$. The pooled Bradley–Terry reference is fitted on seen users without labels from the target user. The independent reference is fitted on every labeled comparison from the target user. Sparse-LoRe, defined below, separates sparsity from extraction. CS-Elicit asks each dyadic node once because the repetition schedule in Theorem 2(i) exceeds the experimental budgets.

For semi-synthetic users, accuracy measures agreement with the noiseless preference; for other users, it measures agreement with the recorded label. We also report negative log-likelihood and wall-clock time. To compute score correlation, we pair the factors of two dictionaries by a Hungarian assignment and average their Pearson correlations on held-out items. This measure can be positive for unrelated dictionaries, so we report random dictionaries of the same shape as a noise floor.

Sparse-LoRe retains LoRe's shared basis V, initialization, and training objective, but replaces the loading with

$$a_{ik} = \frac{g_{ik} \exp(w_{ik})}{\sum_{\ell=1}^{K} g_{i\ell} \exp(w_{i\ell})},$$

where $g_{ik} \in [0, 1]$ is a hard-concrete gate trained with an expected-L_0 penalty. The gates select the support size, and validation data from seen users determine the penalty weight.

User construction and split independence. Table 5 describes how each dataset defines users and labels, gives the size of each split, and records item overlap across splits. The control tasks are *fully synthetic*: both compositions and answers are simulated. PersonalLLM and UltraFeedback in E1, E2, and E4 are *semi-synthetic*: responses and factor scores come from the datasets, while sampled user compositions answer through the Bradley–Terry channel. The PersonalLLM and UF-P-4 corpora used for the complete pipeline are also semi-synthetic. Their fixed labels are generated from reward-model or GPT-4 scores. TL;DR is the only dataset with *recorded human labels linked to persistent identities*; its users are the worker IDs from Stiennon et al. (2020).

All methods use the same splits, which are fixed before training. Validation contains no test user or test pair. On the three language corpora used for the complete pipeline, seeds change only the random question streams. In the other experiments, each data seed draws a new population.

End-to-end corpora. PersonalLLM follows the protocol of Bose et al. (2025). We standardize the ten reward-model scores on the training prompts and draw 300 users from $\mathrm{Dirichlet}(0.1 \cdot \mathbf{1}_{10})$. On each prompt, a user compares the highest- and lowest-scoring responses among eight candidates under their mixture. The first 150 users are seen, and the remaining 150 are unseen. The shared models use LoRe's budget of 47 pairs per seen user, or 0.5% of the training prompts. Each test user's pool contains one pair for every training prompt, so users share items and differ only in their labels. The test set contains 200 pairs on random prompts from the dataset's test split. Of the 998 test prompts, 28 also occur in the training split; as a result, 1.8% of test pairs share a prompt with a pair used to fit the shared models. We remove any pool pair that is identical to one of the same user's test pairs.

TL;DR follows LoRe's Reddit protocol. We retain workers who labeled at least 50 comparisons in the training split of Stiennon et al. (2020) and at least 20 in its validation split; 39 workers meet these criteria. A fixed shuffle assigns 19 workers to the seen set and 20 to the unseen set. The shared models use all 46,190 training-split comparisons from seen workers because we remove LoRe's 150-pair cap. For each unseen worker, training-split comparisons form the query pool and validation-split comparisons form the uncapped test set. The splits contain different posts. No test post appears in the pairs used to fit the shared models, and only 3 of the 6,267 test posts occur anywhere on the training side. Workers can share items: a seen worker also labeled 25% of the pool comparisons for unseen workers, and those seen-worker labels enter the shared fit. Test-set sizes range from 29 to 7,278 comparisons per worker, so we first average accuracy within each worker and then across workers.

UF-P-4 is the benchmark from Poddar et al. (2024), rebuilt with the authors' relabeling code. Each of its four users prefers the response with the higher GPT-4 rating for one aspect: helpfulness, honesty, instruction following, or truthfulness. We keep only pairs on which the four users do not all agree, which removes 119,972 of the 243,328 joined rows. We split prompts $90/10$ into training and test sets. Because the dataset has no separate unseen-user population, the same four users are used for fitting and evaluation. For each user, 4,000 training pairs fit the shared models, another 500 training-side pairs form the validation set, the remaining 23,318 form the query pool, and 500 pairs on test prompts form the test set. No test prompt or response appears on the training side, although 98.9% of pairs have labels from more than one user. UF-P-4 therefore measures adaptation to known users on new prompts.

PersonalLLM and TL;DR use four-fold cross-validation over seen users. In each fold, we refit every shared model on the other three folds, elicit the held-out users from their pools, and evaluate them on their own test pairs. PersonalLLM has 15 users per fold; TL;DR evaluates all seen workers across the folds.

Elicitation, stability, and scaling corpora. For the language datasets, we partition prompts into disjoint query, validation, and test sets. In E2, candidate pairs are within-prompt pairs from 1,500 PersonalLLM query prompts and 7,000 UltraFeedback query prompts. The corresponding test sets contain pairs from 200 and 800 test prompts. Seen, validation, and unseen users draw pairs from the same candidate set, so items can recur across users but cannot cross from the candidate set into the test set. For each control cell, we divide 800 trajectories into training, elicitation, calibration, and test blocks and construct every pair within one block. No trajectory in a test pair appears in any other split.

F.2 Stability of extracted dictionaries (E1)

Populations. PersonalLLM uses the ten reward models as factors ($K = 10$), with $\beta = 1$ and 150 seen, 30 validation, and 50 unseen users. Each seen user provides 4,096 within-prompt comparisons for dictionary learning, and each unseen user provides 512 training and 256 test comparisons. User compositions are either hard s-sparse vectors, with weight $1/s$ on s uniformly selected factors for $s \in \{1, 2, 3\}$, or draws from $\mathrm{Dirichlet}(0.1 \cdot \mathbf{1}_{10})$. We also run a uniform control population with $a = \mathbf{1}/K$ for every user. Because this population has no factor structure to recover, we omit it from Table 1. Each population uses three data seeds.

The Control suite uses 800 episodes per task, split by trajectory into training (40%), elicitation (25%), calibration (15%), and test (20%) sets. Each condition has 250 seen, 50 validation, and 50 unseen users. Every user provides 200 training, 256 calibration, and 500 test comparisons. We calibrate β within each cell to a population Bayes accuracy of 0.80. The eleven conditions include MountainCar at $s = 1$, and CartPole, Acrobot, and LunarLander at both $s = 1$ and their main sparsities of 2, 3, and 3. We use random-policy and controller-augmented trajectory banks and draw each condition with five data seeds.

Features. For PersonalLLM, $h(x)$ is the L2-normalized, final attention-masked hidden state of Skywork-Reward-V2-Llama-3.1-8B, which is also the input to its reward head. We project these states onto the leading 512 principal components of the 60,176 training responses; these components retain 96.5% of the variance. A single constant rescales the features so that their mean row norm matches the features used to tune the pipeline's regularization grids. A ridge readout of the ten reward-model scores from these features has a mean training R^2 of 0.75.

Table 5: User construction and splits. Label counts are per user. *Pool*: the labelled pairs a test user's questions are drawn from; the independent-BT reference is fitted on all of them. CV: four-fold cross-validation over seen users. SFF: Stiennon et al. (2020). RM: reward model.

	PersonalLLM, UltraFeedback	PersonalLLM	TL;DR	UF-P-4	Gym, Kitchen
Experiments	E1, E2, E4 (UltraFeedback: E2)	E3	E3	E3	Gym: E1–E4; Kitchen: E2, E3
Regime	semi-synthetic	semi-synthetic	recorded human labels	semi-synthetic (GPT-4 ratings)	fully synthetic
Users	s-sparse or Dirichlet compositions of the 10 RM scores or the 4 aspect ratings; redrawn per seed	$\mathrm{Dirichlet}(0.1)$ mixtures of 10 standardized RM scores (Bose et al., 2025)	SFF worker IDs	one user per GPT-4 aspect (Poddar et al., 2024)	s-sparse compositions of state coordinates or object joints; redrawn per seed
Labels	sampled from BT, $\beta = 5$ ($\beta = 1$ in E1)	best vs. worst of 8 responses, deterministic	the worker's recorded choice	higher rating on the user's aspect, deterministic	sampled from BT, Bayes accuracy 0.80
Inclusion	all users; UltraFeedback ties excluded from accuracy	all users	≥ 50 training and ≥ 20 validation comparisons: 39 workers	pairs the 4 users disagree on; 119,972 of 243,328 rows dropped	all users
Fitting users (labels)	150 (4,096)	150 seen (47)	19 seen workers (84–12,984; 46,190 in total)	the 4 users (4,000)	250 (200)
Validation	30 users	60 seen users, CV	19 seen workers, CV	500 held-out pairs per user	50 users
Test users	50 per seed; 3 seeds (E4: 5)	150 unseen	20 unseen workers	the same 4 users	25 per seed; 5 seeds
Pool / test labels	512 / 256	9,401 / 200	median 734 / 1,079 (ranges 51–5,631 / 29–7,278)	23,318 / 500	456 / 500
Items across splits	query and test prompts disjoint; candidate items shared by all users	dataset's train/test prompt split; 28 of 998 test prompts recur (1.8% of test pairs); all users share the items	SFF training vs. validation posts; no test post in the shared fit; 25% of pool comparisons also labelled by a seen worker	90/10 prompt split; no test prompt or response on the training side; 98.9% of pairs labelled by > 1 user	trajectories split $40/25/15/20\%$; each pair within one block

Dictionaries. Sparse-LoRe uses the vendored fitter, with its gate penalty selected on validation users. LoRe and Identified LoRe share the same objective and search grid. The grid includes the vendored random initialization and a spectral initialization, which sets V to the top-K right singular vectors of the seen users' Bradley–Terry utilities and $W = 0$; dense loadings and Sparse-LoRe gates; L2 penalties on V of 0, $1.6 \cdot 10^{-5}$, $1.6 \cdot 10^{-4}$, and $1.6 \cdot 10^{-3}$ per training pair; and either the vendored number of optimization steps or $10/3$ times that number. LoRe selects the grid point whose basis gives the highest validation accuracy. Identified LoRe instead selects the point whose dictionary, extracted by Algorithm 2, gives the highest validation accuracy. Most cells select random initialization with a nonzero penalty. Extreme-ray applies Algorithm 2 to individual Bradley–Terry utilities, with the utility ridge chosen on validation users.

Perturbations. Each refit reruns the complete pipeline. Extreme-ray and Sparse-LoRe reselect their hyperparameters for every refit, while LoRe and Identified LoRe reuse the settings selected on the reference data. We test four perturbations. (i) We replace the optimizer seed with three fresh seeds. (ii) We split the seen users into two disjoint random halves three times while holding the seed fixed. (iii) We add three draws of 20% more seen users, 30 on PersonalLLM and 50 on Control, with $a = \mathbf{1}/K$. These users answer the comparisons of one seen user under fresh response noise. (iv) We compare the full-data dictionaries for every pair of data seeds within a condition.

Metrics. We compute score correlation on 7,520 held-out test-prompt responses for PersonalLLM and 160 held-out test trajectories for Control. Cross-seed comparisons use the items from the first seed. Each *Random* row uses ten pairs of Gaussian dictionaries with the learned shape in each cell and applies the same scoring procedure, including disagreement. To compute disagreement, we fit each unseen user's composition in both dictionaries using 512 training comparisons and ridge 0.1 for PersonalLLM, or 256 calibration comparisons and ridge 10^{-3} for Control. We then report the fraction of test comparisons for which the two dictionaries predict opposite signs. On Control, the full-data refits reproduce the dictionaries used in the reported elicitation runs.

Methods and per-condition results. Tables 6, 7, and 8 give condition-level results for the four methods in the main stability table. Identified LoRe applies Algorithm 2 to LoRe utilities. This procedure differs from the extraction from Sparse-LoRe utilities used in the complete pipeline experiment. Extreme-ray applies Algorithm 2 to individual user utilities. The uniform PersonalLLM control separates reproducibility from recovery of the generating factors. Although this population has no variation in composition, its recovery correlations are 0.67 for Identified LoRe and 0.59 for Extreme-ray. Stability alone therefore does not establish recovery of the generating reward models.

F.3 Fixed-dictionary elicitation and control pools (E2)

Ground-truth factors. PersonalLLM uses the standardized scores of ten published reward models, and UltraFeedback uses four GPT-4 aspect ratings (Cui et al., 2023). MountainCar, CartPole, Acrobot, and LunarLander use normalized observation coordinates with $K = 2, 4, 6, 8$, respectively. Franka Kitchen uses $K = 14$ moving object joints.

Protocol of Table 2. UltraFeedback provides four aspect ratings—helpfulness, honesty, instruction following, and truthfulness—for $69{,}636$ completions of $17{,}409$ prompts. Franka Kitchen uses the $K = 14$ object joint positions that move in the D4RL demonstrations, measured over 50-step windows as described below. Each response or trajectory is represented directly in its factor coordinates, so the ground-truth dictionary is the identity. Unseen users answer through the Bradley–Terry channel. We set $\beta = 5$ on the language datasets, which gives a Bayes accuracy of 0.84–0.85, and calibrate β to a population Bayes accuracy of 0.80 on the control tasks.

Pooled BT uses the same seen users and comparisons: 150 language users with $4{,}096$ comparisons each, or 250 control users with 200 comparisons each. It receives no labels from the target user. The independent BT reference fits a separate model to all target-user labels, using 512 comparisons on language datasets and 456 on control tasks, with its ridge selected on validation users. It is a full-data reference. Dictionary-based rules can exceed it because the reference estimates all K coordinates from noisy labels, while a recovered sparse support fixes the user's direction. The repetition count in Theorem 2(i) exceeds these budgets, so CS-Elicit asks each dyadic-tree node once and accepts a pair when $\vartheta > 0$. Appendix F.6 runs the theorem's schedule where the candidate pool permits it.

We fix settings on development seeds and report fresh seeds: three seeds with 50 unseen users each for the language datasets, and five seeds with 25 users each for the control tasks. UltraFeedback ratings are integers, so accuracy excludes pairs with an exact tie in the user's true preference. MI and CS-Elicit reach ceiling accuracy on UltraFeedback for 1-sparse and Dirichlet users.

Less sparse users. Table 9 adds 2-sparse and Dirichlet(0.001) language users. Dirichlet(0.001) is the nearly one-hot population used by Bose et al. (2025). For 2-sparse UltraFeedback users, CS-Elicit exceeds MI by 2.6 points and also exceeds the independent BT reference; the median margin remains 0.75 with four factors. On PersonalLLM, MI exceeds CS-Elicit by 5.3 points. The median margin falls to 0.04, only 70% of nodes are certified, and the walk can require $2s\lceil \log_2 K \rceil = 16$ answers. For the PersonalLLM Dirichlet population, the two rules differ by 0.2 points ($p = 0.68$). VR is below MI in every row.

Control users. The control protocol uses synthetic users and answers built on the task's ground-truth coordinates $f(\tau) \in \mathbb{R}^K$. A user is a composition $a \in \Delta_{K-1}$ with exactly s nonzero entries, each equal to $1/s$, so an active factor always carries mass $1/s$ and the node threshold $\Delta_- = 1/s$ leaves no undetermined band. Supports are assigned by walking a random permutation of all $\binom{K}{s}$ supports, drawing a fresh permutation whenever one is exhausted, so every set of s factors is used by equally many users, up to one. Each condition has 250 seen, 50 validation, and 50 unseen users, taken from the same walk, so the three roles have the same mix of supports. A user prefers trajectory τ to τ' with probability $\sigma(\beta\, a^\top (f(\tau) - f(\tau')))$, and the answer is sampled from it. The coefficient β is set per condition by bisection on training pairs so that the population's Bayes accuracy is 0.80, searched over $[10^{-3}, 10^{3}]$. Trajectories are split into training (40%), elicitation (25%), calibration (15%), and test (20%) sets, and every user answers 200 training, 256 calibration, and 500 test pairs, each drawn uniformly from the pairs of one set. The 456 training and calibration pairs are the "all pairs" of the independent BT reference, and a new user's questions are pairs of elicitation trajectories. Features are z-scored with statistics of the training trajectories only. The fixed-dictionary comparison in Table 2 evaluates 25 unseen users per seed from these generated splits; its control rows use $s = 1$.

Table 6: Stability by PersonalLLM population. Mean ± standard deviation across the data seeds of each condition, or across its seed pairs for Draw; columns, methods, and *Random* rows as in Table 1. Best method per column in bold.

Condition	Method	Recovery	Seed	Split	Dense	Draw	Disagr. (%)
Hard, $s = 1$	LoRe	0.35 ± 0.20	0.59 ± 0.25	0.79 ± 0.12	0.54 ± 0.27	0.37 ± 0.07	8.5 ± 1.3
	Sparse-LoRe	0.14 ± 0.02	0.27 ± 0.02	0.97 ± 0.01	0.26 ± 0.05	0.27 ± 0.04	9.2 ± 2.4
	Identified LoRe	0.55 ± 0.11	0.66 ± 0.14	0.73 ± 0.08	0.68 ± 0.11	0.51 ± 0.06	10.5 ± 3.0
	Extreme-ray	$\mathbf{0.60 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.87 \pm 0.11}$	$\mathbf{1.2 \pm 0.1}$
	Random	0.12 ± 0.04		0.47 ± 0.01			49.8 ± 3.4
Hard, $s = 2$	LoRe	0.33 ± 0.09	0.57 ± 0.10	0.78 ± 0.10	0.60 ± 0.20	0.48 ± 0.16	8.6 ± 4.1
	Sparse-LoRe	0.14 ± 0.02	0.26 ± 0.03	$\mathbf{0.97 \pm 0.00}$	0.26 ± 0.03	0.27 ± 0.03	9.3 ± 1.8
	Identified LoRe	$\mathbf{0.66 \pm 0.03}$	0.70 ± 0.16	0.76 ± 0.03	0.74 ± 0.12	$\mathbf{0.70 \pm 0.04}$	10.0 ± 1.5
	Extreme-ray	0.31 ± 0.15	$\mathbf{1.00 \pm 0.00}$	0.74 ± 0.06	$\mathbf{0.99 \pm 0.02}$	0.69 ± 0.10	$\mathbf{7.3 \pm 7.8}$
	Random	0.10 ± 0.04		0.47 ± 0.02			53.0 ± 6.6
Hard, $s = 3$	LoRe	0.28 ± 0.05	0.40 ± 0.09	0.92 ± 0.03	0.40 ± 0.08	0.39 ± 0.07	8.7 ± 2.0
	Sparse-LoRe	0.15 ± 0.03	0.27 ± 0.03	$\mathbf{0.97 \pm 0.00}$	0.28 ± 0.05	0.28 ± 0.04	9.4 ± 2.8
	Identified LoRe	0.56 ± 0.17	0.61 ± 0.20	0.83 ± 0.06	0.61 ± 0.23	0.52 ± 0.13	8.9 ± 1.3
	Extreme-ray	$\mathbf{0.60 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.97 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.96 \pm 0.01}$	$\mathbf{1.6 \pm 0.6}$
	Random	0.06 ± 0.06		0.48 ± 0.05			45.8 ± 6.5
$\mathrm{Dirichlet}(0.1)$	LoRe	0.47 ± 0.17	0.65 ± 0.09	0.67 ± 0.13	0.61 ± 0.06	0.44 ± 0.11	9.6 ± 1.6
	Sparse-LoRe	0.15 ± 0.03	0.26 ± 0.02	$\mathbf{0.97 \pm 0.00}$	0.29 ± 0.03	0.28 ± 0.03	9.1 ± 0.7
	Identified LoRe	$\mathbf{0.72 \pm 0.03}$	0.81 ± 0.03	0.77 ± 0.03	0.83 ± 0.04	0.81 ± 0.02	10.9 ± 1.3
	Extreme-ray	0.57 ± 0.03	$\mathbf{1.00 \pm 0.00}$	0.84 ± 0.04	$\mathbf{0.96 \pm 0.07}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{1.9 \pm 0.3}$
	Random	0.10 ± 0.05		0.48 ± 0.02			46.3 ± 4.8
Uniform†	LoRe	0.46 ± 0.19	0.53 ± 0.18	0.80 ± 0.12	0.57 ± 0.25	0.46 ± 0.10	8.1 ± 2.0
	Sparse-LoRe	0.15 ± 0.03	0.26 ± 0.04	0.97 ± 0.00	0.27 ± 0.04	0.28 ± 0.02	9.2 ± 2.9
	Identified LoRe	$\mathbf{0.67 \pm 0.01}$	0.72 ± 0.13	0.81 ± 0.03	0.76 ± 0.09	0.72 ± 0.02	8.9 ± 1.0
	Extreme-ray	0.59 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.4 \pm 0.1}$
	Random	0.08 ± 0.04		0.46 ± 0.03			54.2 ± 7.5

Table 7: Stability on MountainCar and CartPole. Mean ± standard deviation across the data seeds of each condition, or across its seed pairs for Draw; columns, methods, and *Random* rows as in Table 1. Best method per column in bold.

Condition	Method	Recovery	Seed	Split	Dense	Draw	Disagr. (%)
MountainCar, $s = 1$	LoRe	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.5 \pm 0.3}$
	Sparse-LoRe	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.01	0.99 ± 0.01	0.6 ± 0.2
	Identified LoRe	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.5 \pm 0.3}$
	Extreme-ray	0.99 ± 0.01	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.01	$\mathbf{0.5 \pm 0.7}$
	Random	0.04 ± 0.30		0.22 ± 0.13			44.3 ± 6.2
MountainCar, $s = 1$, mixed	LoRe	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.5 \pm 0.1}$
	Sparse-LoRe	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.98 ± 0.03	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.5 \pm 0.2}$
	Identified LoRe	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.5 \pm 0.1}$
	Extreme-ray	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.6 ± 0.2
	Random	0.12 ± 0.23		0.17 ± 0.12			49.2 ± 8.7
CartPole, $s = 1$	LoRe	0.98 ± 0.03	0.94 ± 0.08	0.93 ± 0.08	0.98 ± 0.01	0.96 ± 0.04	$\mathbf{0.8 \pm 0.5}$
	Sparse-LoRe	$\mathbf{0.99 \pm 0.02}$	0.97 ± 0.02	$\mathbf{0.98 \pm 0.02}$	0.98 ± 0.01	$\mathbf{0.98 \pm 0.03}$	1.5 ± 0.6
	Identified LoRe	$\mathbf{0.99 \pm 0.02}$	0.91 ± 0.12	0.72 ± 0.14	$\mathbf{0.99 \pm 0.02}$	$\mathbf{0.98 \pm 0.02}$	4.7 ± 1.3
	Extreme-ray	0.50 ± 0.06	$\mathbf{1.00 \pm 0.00}$	0.71 ± 0.08	0.73 ± 0.22	0.95 ± 0.04	3.1 ± 0.6
	Random	0.55 ± 0.10		0.43 ± 0.09			25.9 ± 5.4
CartPole, $s = 2$	LoRe	$\mathbf{0.85 \pm 0.08}$	0.92 ± 0.10	0.93 ± 0.08	0.90 ± 0.16	0.79 ± 0.11	1.0 ± 0.4
	Sparse-LoRe	0.75 ± 0.03	0.98 ± 0.01	$\mathbf{0.98 \pm 0.02}$	0.97 ± 0.02	0.98 ± 0.01	1.2 ± 0.4
	Identified LoRe	0.76 ± 0.01	0.97 ± 0.04	0.97 ± 0.06	0.96 ± 0.03	$\mathbf{0.99 \pm 0.01}$	1.3 ± 1.0
	Extreme-ray	0.77 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.03	$\mathbf{0.6 \pm 0.2}$
	Random	0.55 ± 0.06		0.48 ± 0.07			30.3 ± 7.6
CartPole, $s = 2$, mixed	LoRe	0.81 ± 0.04	0.91 ± 0.12	0.92 ± 0.11	0.94 ± 0.12	0.95 ± 0.04	0.9 ± 0.5
	Sparse-LoRe	0.82 ± 0.02	0.98 ± 0.02	0.98 ± 0.02	0.98 ± 0.01	$\mathbf{0.99 \pm 0.01}$	1.3 ± 0.5
	Identified LoRe	0.81 ± 0.02	0.97 ± 0.03	0.98 ± 0.02	0.99 ± 0.02	$\mathbf{0.99 \pm 0.01}$	1.1 ± 0.6
	Extreme-ray	$\mathbf{0.85 \pm 0.02}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.8 \pm 0.4}$
	Random	0.43 ± 0.06		0.48 ± 0.02			30.5 ± 6.4

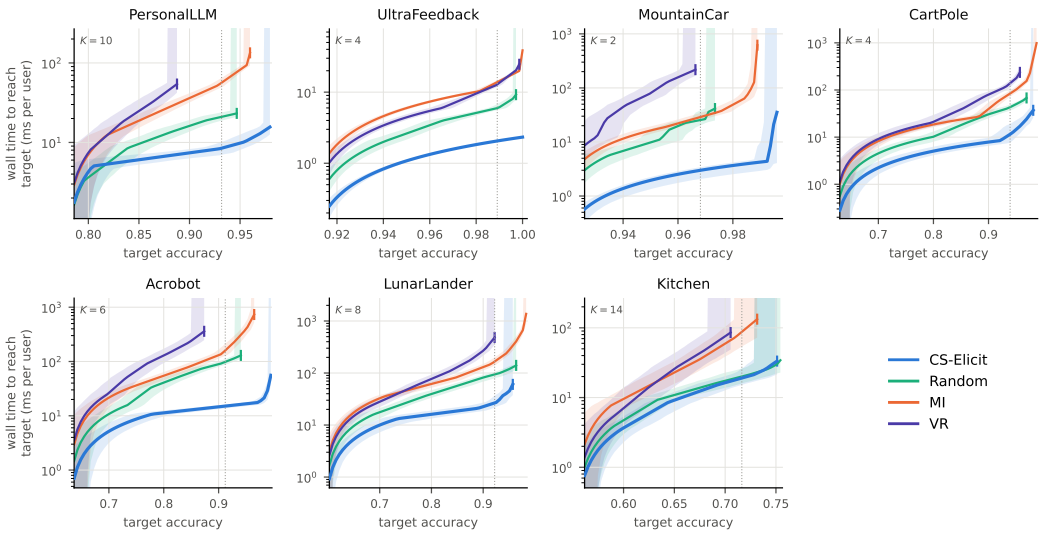


Figure 2: Wall-clock cost of reaching a given prediction accuracy, for the 1-sparse users of Table 2 on the ground-truth dictionary. A tick ends the line where the rule never reaches the target.

Language users, budgets, scales, and seed counts for this comparison are specified alongside that table.

Franka Kitchen. We sample 800 windows of 50 steps from the D4RL `kitchen-mixed` demonstrations (Minari `D4RL/kitchen/mixed-v2`) (Fu et al., 2020). Each window begins at a uniformly selected step of a uniformly selected demonstration. The raw observations contain 21 object joint positions. We remove the 7 joints whose dataset standard deviation is below 1% of the largest because they barely move, leaving $K = 14$ factors. A window's feature is the mean joint position over its 50 steps. The resulting reward is linear in object state, and a 1-sparse user prefers windows in which one selected joint is higher in its range.

For 1-sparse users, bisection reaches the upper limit $\beta = 10^3$ in four of five seeds. The resulting Bayes accuracy is 0.79–0.80 on training pairs and 0.78–0.79 on test pairs. Many joints remain nearly fixed within most windows: for eight of the fourteen joints, 60–84% of training pairs differ by less than 0.05 standard deviations. A 1-sparse user answers these pairs near chance for any β. For 2-sparse users, β ranges from 4.8 to 6.1.

F.4 Wall-clock cost of elicitation

Figure 2 plots the wall time each rule in Table 2 spends per user. A CS-Elicit question costs no more than a random one: at $Q = 8$ it takes 0.9–2.6 ms, against 1.1–4.6 ms for random, because its next pair is read from a probe tree built before the user arrives. MI and VR instead score a fresh subset of 256–384 candidate pairs against 64–96 posterior samples at every step, and cost 3.2–10.3 and 2.3–7.4 ms per question. Because CS-Elicit also needs fewer questions, its time-to-target curve lies below those of the other rules on all five datasets. At the reference target (90% of the accuracy gain), CS-Elicit needs 3.0–26.6 ms per user. This is 2.5–8.5× less than random and 6.3–10.6× less than MI, and VR does not reach the target on three of the five datasets. The gap widens with the budget: after 256 questions on the control tasks, MI has spent 0.8–1.4 s per user and CS-Elicit 35–66 ms. MI's extra time buys accuracy only on LunarLander (98.3% against 95.9%) and, by 0.2 points, on CartPole. These per-user times exclude the one-time construction of the probe tree, which takes 0.5–6.9 s per population and is shared by all later users.

Runs and timing. Figure 2 and Table 10 reuse the runs of Table 2 without rerunning them. These are the 1-sparse users on the ground-truth dictionary: five fresh seeds with 25 unseen users each on every control task ($K = K_{\text{true}}$), and three fresh seeds with 50 unseen users each on PersonalLLM ($\beta = 5$). The time of a user at budget Q is the wall-clock time the rule has spent on that user up to

its Q-th answer, measured inside the rule. It covers choosing each pair, updating the rule's posterior, and refitting the composition at every evaluated budget, so all rules pay the same refits. It excludes the work shared by all users of a population: the dictionary, the pooled reference, and CS-Elicit's probe tree and probe search. Held-out evaluation is also excluded. At each step MI and VR score a fresh random subset of candidate pairs against samples from a particle posterior. On the control tasks this is 256 pairs, 64 samples, and $2{,}048$ particles, the settings frozen on development seeds. On PersonalLLM it is the defaults of 384 pairs, 96 samples, and $4{,}096$ particles. CS-Elicit asks one answer per tree node and then draws random pairs, so after its walk each question costs about as much as a random one. Each seed of a condition ran in one process, with all rules timed sequentially in that process. Up to 35 such processes shared a 48-core Intel Xeon w7-3455. We report the median time over users and the mean accuracy over users at each budget.

Targets and intervals. The time to reach a target accuracy is the first budget at which the mean-accuracy curve reaches it. It is log-interpolated in time and in questions between the two budgets that bracket the crossing. The reference target of Table 10 is 90% of the way from the zero-question accuracy to the highest accuracy that CS-Elicit, random, and MI all reach. VR is excluded from this ceiling because its accuracy falls past 128 questions on three control tasks, which would cap the target low. Intervals come from $1{,}000$ bootstrap resamples of users, with the same resample used for all rules of a dataset. A resample in which a rule never reaches the target counts as an infinite time.

Setup cost. CS-Elicit builds its probe tree once per population, before any new user arrives. On the control tasks we report the logged tree-planning and probe-search stages, median over seeds. These stages cover both dictionaries of each run, the ground-truth and an extracted one, so they overstate the setup for the ground-truth dictionary alone. On PersonalLLM the setup was not logged. We re-timed it on the saved ground-truth dictionaries of the same three seeds, as the tree plan plus the per-node probe pools built before the first user; it took 4.85–4.95 s. The break-even count $N^\star$ is the setup time divided by the per-user time saved against a rule at the reference target. A population with more than $N^\star$ new users therefore spends less total time with CS-Elicit, setup included. The evaluation populations have 25 (control) and 50 (PersonalLLM) unseen users, so against MI the setup is recovered within one evaluation population on MountainCar, CartPole, and Acrobot, but not on LunarLander or PersonalLLM.

F.5 Probe realization in the elicitation experiments

CS-Elicit can only use a node of the dyadic tree if the candidate pool realizes a valid probe for it (Assumption 3), and a single answer to that probe is reliable only if $\sigma(\beta\vartheta)$ is well above $\frac{1}{2}$ (Lemma 1). This section shows both quantities for the cells of Table 2. For every node G we take the pair with the largest margin $\vartheta(x; G, \Delta_-)$ of Definition 2 over the pool, in ground-truth coordinates and at $\Delta_- = 1/s$, which is the pair CS-Elicit asks. A node is certified when $\vartheta > 0$.

Figure 3 shows the selected contrasts for one seed per dataset. A valid probe at $s = 1$ is positive on every factor of the group and negative on every other factor, and this is the pattern of the certified rows. The realized contrasts are not the canonical probe $\gamma(\mathbf{1}_G - c\,\mathbf{1})$ of Assumption 3: their entries vary within and outside the group, and the margin is set by the weakest of them. On UltraFeedback the four aspect ratings are integers and strongly separated pairs are plentiful, so every node has a margin near 0.75 at $s = 2$. On PersonalLLM at $s = 2$ the uncertified nodes are mainly the large groups, where a pair must raise several reward models together while lowering all the others; five of the six root nodes across seeds are uncertified, and the median node has only two pairs with $\vartheta > 0$. On Kitchen, most object joints do not move in a given 50-step window, so the contrasts are zero on them, and no pair is strictly positive on all factors of a group. The nodes counted as certified there have $\vartheta < 10^{-4}$.

Figure 4 gives the margin of every node across seeds, together with the per-answer accuracy $\sigma(\beta\vartheta)$ that Lemma 1 guarantees for it. The sign of CS-Elicit's advantage over MI in Table 2 tracks the second quantity, and the certified share predicts it poorly. CS-Elicit is significantly better than MI in every cell where at least 43% of the nodes have $\beta\vartheta \geq 1$, and it ties or loses where a quarter of the nodes or fewer do (PersonalLLM at $s = 2$ and Dirichlet, Kitchen), although PersonalLLM certifies 63–70% of its nodes in those cells. LunarLander is intermediate: 96% of its nodes are certified but only 43% have $\beta\vartheta \geq 1$, with a median guaranteed accuracy of 0.66. On Kitchen, the calibrated scale $\beta = 10^3$ does not compensate for the vanishing margins. No node reaches $\beta\vartheta \geq 1$, and

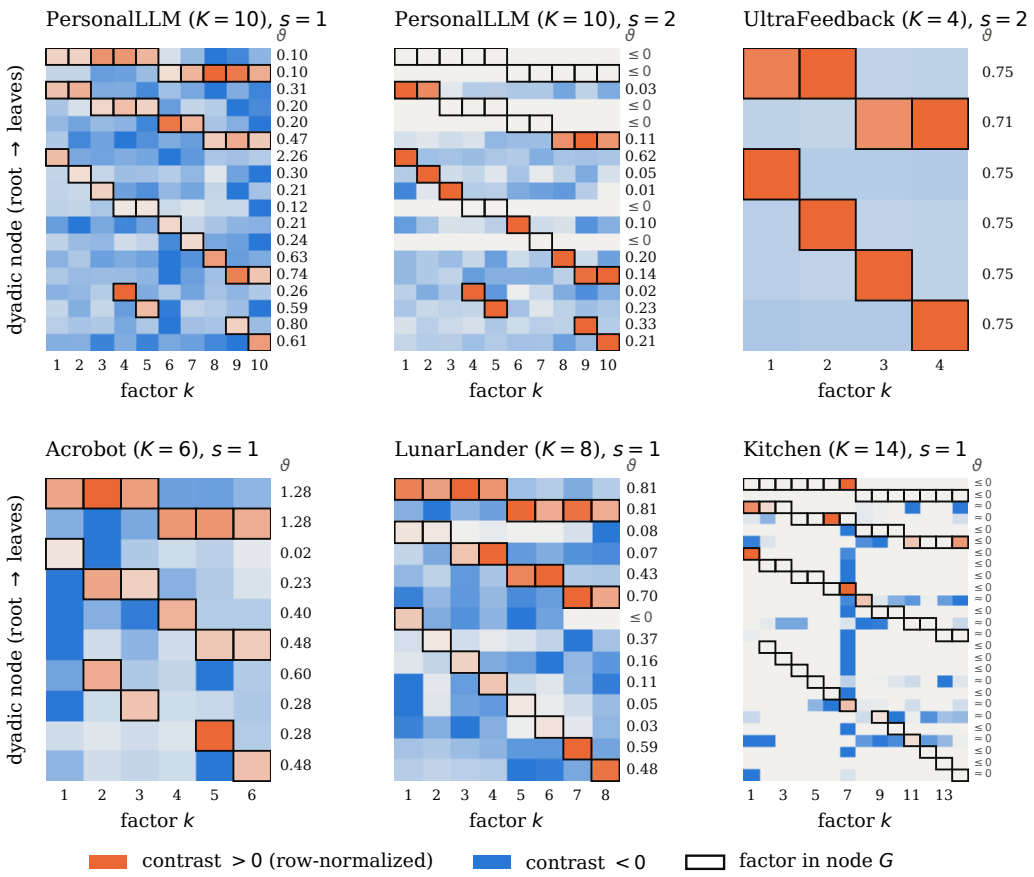


Figure 3: Realized probes. Each row is a node of the dyadic tree of one cell (seed 3), ordered from the root to the leaves, and shows the ground-truth contrast $x = f(\tau) - f(\tau')$ of the pair with the largest margin, normalized by $\max_k |x_k|$. Boxes mark the factors in the node's group G; the column on the right gives the margin $\vartheta(x; G, 1/s)$. At $s = 1$ a valid probe is positive exactly on the boxed factors. Rows that are blank and end in ≤ 0 have no pair with a positive margin.

the guaranteed accuracy is 0.5 at every node, so the 43% certified share in Figure 4 is a numerical artifact and the tree is effectively uncertified. The best margin also overstates what repeated tests can use. Assumption 3 asks for r distinct pairs of margin ϑ_0 at every node, and the ninth-best pair's median margin is 40–51% below the best on PersonalLLM (0.19 against 0.32 at $s = 1$), Acrobot (0.32 against 0.52), and LunarLander (0.12 against 0.25), and zero at PersonalLLM's median node at $s = 2$ and Dirichlet. Only on MountainCar and UltraFeedback at $s = 1$ can every node meet the theorem's requirement with distinct pairs; the share is 14–72% on the other realizable cells and 0% on Kitchen and on PersonalLLM beyond $s = 1$ (Appendix F.6).

F.6 Support recovery and the repetition schedule

Theorem 2(i) guarantees $\hat{S} = S$ when every node of the dyadic tree is tested r times on r distinct pairs of margin at least ϑ_0, whereas the runs of Table 2 ask each node once. This section measures what the one-shot walk recovers, and runs the theorem's procedure wherever the candidate pool allows it. The cells are those of Table 2 and Table 9 with $s \in \{1, 2\}$, on the ground-truth dictionary: the same seeds, users, candidate pools, trees, and $\Delta_- = 1/s$.

Procedures. We simulate the complete walk of Algorithm 1 for each unseen user without an answer budget, 200 times per user with fresh Bradley–Terry answers. A node is tested on its r distinct pairs with the largest margin $\vartheta(x; G, \Delta_-)$ and declared positive when the majority of the r answers is.

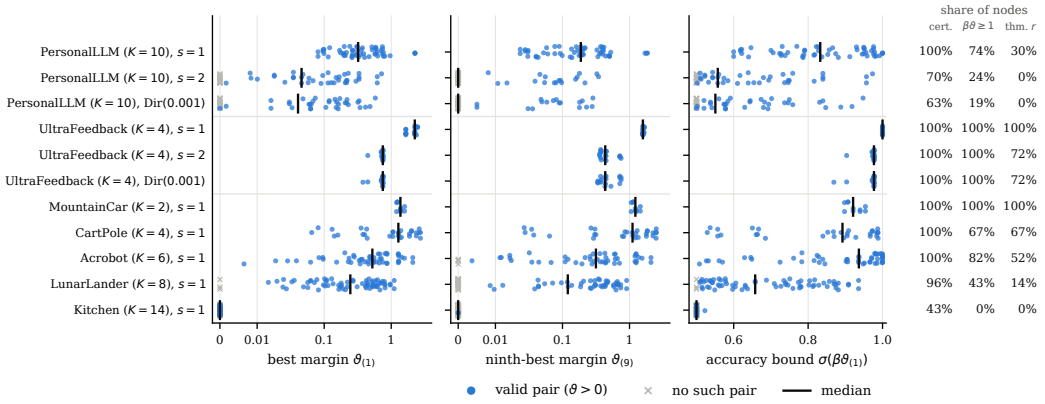


Figure 4: Node margins and the per-answer accuracy they guarantee, for every cell of Table 2 (all nodes of the oracle dictionary's tree over all seeds: three on the language datasets, five on the control tasks). Left: margin $\vartheta_{(1)}$ of the best distinct pair per node, the pair CS-Elicit asks, on a symmetric-log axis; nodes without a valid pair are plotted at 0. Second: margin $\vartheta_{(9)}$ of the ninth-best distinct pair, the weakest pair a node tested nine times would use. Third: the lower bound $\sigma\big(\beta\vartheta_{(1)}\big)$ on the probability that a single answer is correct (Lemma 1), 0.5 for uncertified nodes. Right: share of nodes that are certified (cert.), that have $\beta\vartheta_{(1)} \geq 1$, i.e. a single answer correct with probability at least 0.73, and whose r-th best margin meets the requirement $r \geq \log(2\bar{T}/\delta)/(2b^2)$ of Theorem 2(i) for some r (thm. r; $\delta = 0.05$). Bars mark medians.

The one-shot procedure ($r = 1$) is the rule of Table 2; we also use $r \in \{3, 9, 17\}$. The theorem's schedule sets $r = \lceil \log(2\bar{T}/\delta)/(2b^2) \rceil$ with $\delta = 0.05$, $\beta_- = \beta$, and $\eta_{\text{dict}} = 0$ (the dictionary is exact), and takes ϑ_0 to be the smallest r-th best margin over all nodes of the tree, as Assumption 3 requires. We use the smallest r that satisfies its own requirement; if no r does, the theorem does not apply to the cell. A node without a valid pair is not asked and its factors are left out of $\hat{S}$, as in the runs. For fixed r, a node with fewer than r valid pairs is asked on all of them. Majority votes over more than 64 answers use the normal approximation to the Poisson-binomial count.

Metrics. Exact recovery is $\Pr(\hat{S} = S)$. The false inclusion rate is $\mathbb{E}|\hat{S} \setminus S|/(K - s)$ and the false exclusion rate is $\mathbb{E}|S \setminus \hat{S}|/s$. The depth-wise error is the share of tested nodes at a depth whose decision differs from whether the group intersects S. All are averaged over users and then over seeds.

Applicability. The theorem applies in 2 of the 13 cells, MountainCar and UltraFeedback at $s = 1$, where it needs $r = 15$ and $r = 11$ answers per node and 29 and 44 questions per user, beyond the budget of 8. In every other cell some node of the tree lacks enough distinct valid pairs of the required margin (Table 11). Even if every node could reuse its best pair, the worst node would need 206 answers on PersonalLLM at $s = 1$, and 785–3,688 on CartPole and on Acrobot at $s = 1$. On Kitchen, LunarLander, and Acrobot and PersonalLLM at $s = 2$, some node has no valid pair at all. The best margin also overstates the margin available to r distinct pairs: the median node's ninth-best margin is 0.12 against a best margin of 0.25 on LunarLander at $s = 1$, and 0.19 against 0.32 on PersonalLLM.

Is repetition conservative or essential? Where the theorem applies, repetition is essential on MountainCar, where the one-shot walk recovers 88.3% of supports and the theorem's schedule 100%, and it changes nothing on UltraFeedback, where the one-shot walk is already exact. Outside the theorem's regime, repeating a node on more distinct pairs still helps at $s = 1$ when the pool supplies them: exact recovery rises from 62.0% to 89.7% at nine answers per node on PersonalLLM, from 54.5% to 73.8% on CartPole, and from 38.4% to 52.4% on LunarLander, at 7–9 times the questions. At $s = 2$ repetition does not help on Acrobot, PersonalLLM, and LunarLander, where recovery stays at 2–16%: false exclusions dominate (62–93%), because nodes with no valid pair cannot be asked, and a positive node whose best pairs have margins near zero is missed by any number of votes. The one-shot walk's errors are mostly false exclusions. Repetition reduces the

node error at every depth except the first on PersonalLLM at $s = 2$ (Table 12). A false negative at depth 1 excludes half of the factors, so the one-shot error at the top of the tree (17.9% on PersonalLLM at $s = 1$, 4.9–6.1% on LunarLander and MountainCar) propagates to every factor below it.

Support recovery by rule. Table 13 compares the supports implied by each rule's fitted composition after $Q = 8$ answers. With support $\{k : \hat{a}_k \geq \Delta_-/2\}$, CS-Elicit leads MI at $s = 1$ on every dataset except MountainCar (tied on UltraFeedback), in part because its fit is restricted to the tree's support, which makes its composition sparse. With the threshold-free top-s coordinates, which uses the true s, MI matches or exceeds CS-Elicit on every cell except Acrobot at $s = 1$ and UltraFeedback at $s = 2$, including PersonalLLM at $s = 1$ (85.3% against 73.3%), where CS-Elicit is the more accurate rule. CS-Elicit's accuracy advantage in Table 2 is therefore not an advantage in support identification: a composition fitted on an approximately correct support predicts preferences well even when the support is not exact.

F.7 End-to-end personalization and fitting ablation (E3)

Table 3 compares complete pipelines at $Q = 16$. Random and CS-Elicit use the same gated composition estimator within each frozen linear dictionary. Our pipeline applies Algorithm 2 to Sparse-LoRe utilities to construct its dictionary. VPL and SPL use their native inference procedures.

Why native VPL and SPL inference is retained. The Gym control ablation replaces each model with the top-K right singular vectors of its centered, decoder-linearized seen-user utilities, then fits nonnegative weights for each new user in this frozen dictionary. Under the same random question sessions and gated fit, this change lowers mean accuracy by 0.114 for VPL and 0.088 for SPL (Table 14). The results support using the native baseline procedures. They do not establish that a nonnegative fit can reproduce an arbitrary signed decoder representation.

Validation and timing for the control fitting ablation. Validation selects the gated fit over $\{150, 500\}$ optimization steps, learning rates $\{0.01, 0.05\}$, and L_0 penalties $\{0, 0.1, 1, 10\}$ plus the basis-specific penalty. The simplex comparator uses ridge 10^{-3} without a separate grid search, so the two fits have different tuning budgets. We fix question sessions before refitting. At the native Sparse-LoRe settings, the batched gated implementation exactly reproduces its unseen-user accuracies. Timing includes question selection and one fit at $Q = 16$. We amortize gated fitting over each batch of users, so the reported value is a per-user batch cost rather than the latency of fitting one user alone.

Detailed end-to-end results. Table 3 computes NLL from held-out answers: human labels on TL;DR, rule-generated labels on PersonalLLM and UF-P-4, and simulated Bradley–Terry answers on the control tasks. Language accuracy is agreement with the held-out label. Control accuracy is agreement with the noiseless preference on pairs with a strict preference. The Gym column pools MountainCar, CartPole, Acrobot, and LunarLander with $K = 8$ and 500 test users. Kitchen uses $K = 7$ and 125 test users. Independent BT uses a median of 9,401 labels per user on PersonalLLM, 734 on TL;DR, 23,318 on UF-P-4, and 456 on the control tasks.

For each dataset, validation selects one extracted dictionary by its average accuracy over the four question rules. This procedure does not tune the dictionary specifically for CS-Elicit. Validation also selects the gated composition fit over a simplex fit; the gated fit ties or wins on four of five datasets. PAL uses PAL-A from the authors' code. Validation selects up to 300 new-user epochs, more than the official schedule. Under the official schedule, PAL is less accurate on the control tasks, including 0.638 rather than 0.852 on Kitchen. On UF-P-4, validation selects one prototype, so PAL reduces to a pooled model.

Table 16 fixes the dictionary and estimator and varies only the questions. CS-Elicit has the highest accuracy on UF-P-4, with 0.635 compared with 0.583–0.596. MI is more accurate on Gym, with 0.935 compared with 0.896, and on Kitchen, with 0.932 compared with 0.923; the Kitchen difference is within noise. Random and volume removal are more accurate on PersonalLLM, with 0.935–0.937 compared with 0.926. The four rules differ by at most 0.002 on TL;DR.

Using validation users only, we also searched for a control setting favorable to CS-Elicit across four learned dictionaries, $K \in \{8, \dots, 256\}$, $Q \in \{4, 8, 16, 32\}$, and four metrics. No selected setting places CS-Elicit significantly above all three other rules on test users. The closest cases are Acrobot at $Q = 8$, where CS-Elicit leads on every metric but is significantly better only than volume removal, and Kitchen at $Q = 8$, where it exceeds random and volume removal and ties MI. Table 15 gives results for each Gym task. Independent BT uses 28–1,457 times as many questions and gives the highest accuracy on every dataset except TL;DR. On TL;DR, half of the unseen annotators have at most 734 pairs for fitting a 512-dimensional reward. Every CS-Elicit result uses a budget-truncated walk, so Theorem 2 does not apply to these runs.

F.8 Representation limit of the feature space

The dictionaries in Table 3 are linear maps of a fixed feature space. Their recovery of the generating factors is therefore limited by the best linear readout from those features. We measure this limit on PersonalLLM, where the ten standardized reward-model scores provide known ground-truth factors. The feature map contains 512-dimensional Skywork-Reward-Llama-3.1-8B-v0.2 representations. The *oracle* dictionary is a least-squares readout of the ten scores fitted on training prompts. The *extracted* dictionary uses greedy successive projection on the per-user loadings of a Sparse-LoRe model trained on 100 seen users from a Dirichlet(0.1) population. It selects $K = 10$ users and uses their rewards as factors. We evaluate both dictionaries on 66 test prompts that do not occur in the Sparse-LoRe training data. After matching factors by a Hungarian assignment, we report the held-out R^2 from regressing each ground-truth factor on the dictionary and the Pearson correlation between matched factors (Table 17).

The oracle dictionary recovers the ground truth only partially, with mean $R^2 = 0.67$, matched correlation 0.82, and a minimum factor R^2 of 0.44. The extracted dictionary has mean $R^2 = 0.44$, correlation 0.52, and spans 68% of the oracle readout. The ten users selected as extreme rays have ground-truth purities from 0.37 to 0.996, with mean 0.77. Thus ν can reach 0.63, far above the requirement $\nu \leq 1/20$ in Theorem 3. These representation and purity gaps limit every question rule on this dictionary and help explain why question selection has a smaller effect in the complete pipeline than with the ground-truth dictionary in Table 2.

F.9 Question-count scaling (E4)

Figure 1 uses $s = 1$, five fresh seeds with 50 unseen users per seed, a common target of 93% noiseless-preference accuracy, and log interpolation between evaluated query budgets. The four Gymnasium tasks (de Lazcano et al., 2024) use inverse-quadratic state-region dictionaries with $K = \lceil K_{\text{true}} r \rceil$ for $r \in \{1/3, 2/3, 1, 4/3, 5/3\}$. PersonalLLM uses the first $K \in \{2, 4, 6, 8, 10\}$ reward models and an extracted dictionary. Open markers denote targets that remain unreached; these cases stay censored at the largest evaluated budget. CS-Elicit uses one answer per node in this experiment, so the results do not test the high-confidence repetition schedule in Theorem 2. Panel (c) reads the tree output at the largest evaluated budget, after every walk has finished. For PersonalLLM, a recovered factor counts as the reward model assigned to it by a Hungarian match on score correlation. The mean matched correlation is at least 0.988 for every K.

Table 8: Stability on Acrobot and LunarLander. Mean ± standard deviation across the data seeds of each condition, or across its seed pairs for Draw; columns, methods, and *Random* rows as in Table 1. Best method per column in bold.

Condition	Method	Recovery	Seed	Split	Dense	Draw	Disagr. (%)
Acrobot, $s = 1$	LoRe	0.94 ± 0.08	0.93 ± 0.04	0.86 ± 0.09	0.93 ± 0.07	0.90 ± 0.09	$\mathbf{2.3 \pm 0.6}$
	Sparse-LoRe	0.91 ± 0.04	0.86 ± 0.03	0.82 ± 0.03	0.88 ± 0.07	0.86 ± 0.05	5.0 ± 1.7
	Identified LoRe	$\mathbf{1.00 \pm 0.00}$	0.91 ± 0.09	$\mathbf{0.90 \pm 0.08}$	$\mathbf{0.95 \pm 0.05}$	$\mathbf{0.99 \pm 0.00}$	2.8 ± 0.7
	Extreme-ray	0.86 ± 0.12	$\mathbf{1.00 \pm 0.00}$	0.70 ± 0.12	0.82 ± 0.08	0.79 ± 0.08	10.4 ± 2.8
	Random	0.54 ± 0.07		0.51 ± 0.07			29.1 ± 3.7
Acrobot, $s = 3$	LoRe	0.74 ± 0.08	0.86 ± 0.12	0.79 ± 0.10	0.91 ± 0.09	0.69 ± 0.09	$\mathbf{1.5 \pm 0.7}$
	Sparse-LoRe	0.76 ± 0.09	0.81 ± 0.07	0.82 ± 0.04	0.79 ± 0.07	0.79 ± 0.09	4.8 ± 1.3
	Identified LoRe	$\mathbf{0.79 \pm 0.07}$	0.82 ± 0.07	0.80 ± 0.10	0.93 ± 0.03	0.79 ± 0.06	5.1 ± 3.2
	Extreme-ray	0.78 ± 0.05	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.98 \pm 0.03}$	$\mathbf{0.80 \pm 0.12}$	4.1 ± 1.0
	Random	0.53 ± 0.05		0.50 ± 0.03			24.4 ± 3.9
Acrobot, $s = 3$, mixed	LoRe	0.75 ± 0.14	0.88 ± 0.15	0.76 ± 0.11	0.87 ± 0.14	0.70 ± 0.10	$\mathbf{1.6 \pm 0.4}$
	Sparse-LoRe	0.72 ± 0.09	0.80 ± 0.04	$\mathbf{0.83 \pm 0.05}$	0.80 ± 0.05	0.78 ± 0.09	4.6 ± 0.5
	Identified LoRe	$\mathbf{0.76 \pm 0.05}$	0.91 ± 0.08	0.81 ± 0.04	0.87 ± 0.10	$\mathbf{0.86 \pm 0.06}$	7.0 ± 1.3
	Extreme-ray	0.67 ± 0.06	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.83 \pm 0.06}$	$\mathbf{0.98 \pm 0.03}$	0.81 ± 0.06	5.3 ± 1.8
	Random	0.51 ± 0.04		0.49 ± 0.03			27.2 ± 2.9
LunarLander, $s = 1$	LoRe	0.90 ± 0.14	0.97 ± 0.03	0.88 ± 0.09	0.86 ± 0.06	0.82 ± 0.15	$\mathbf{4.3 \pm 1.5}$
	Sparse-LoRe	0.85 ± 0.05	0.81 ± 0.02	0.86 ± 0.06	0.82 ± 0.06	0.77 ± 0.05	5.4 ± 1.0
	Identified LoRe	$\mathbf{0.92 \pm 0.10}$	0.95 ± 0.07	$\mathbf{0.90 \pm 0.09}$	0.90 ± 0.04	$\mathbf{0.87 \pm 0.09}$	6.5 ± 1.2
	Extreme-ray	0.87 ± 0.15	$\mathbf{1.00 \pm 0.00}$	0.74 ± 0.04	$\mathbf{0.93 \pm 0.07}$	0.77 ± 0.15	7.2 ± 1.2
	Random	0.53 ± 0.03		0.52 ± 0.03			25.7 ± 1.8
LunarLander, $s = 3$	LoRe	0.72 ± 0.05	0.88 ± 0.15	0.76 ± 0.09	0.85 ± 0.11	0.75 ± 0.06	$\mathbf{2.4 \pm 0.5}$
	Sparse-LoRe	0.74 ± 0.06	0.80 ± 0.02	$\mathbf{0.83 \pm 0.02}$	0.80 ± 0.02	0.79 ± 0.06	3.7 ± 0.4
	Identified LoRe	$\mathbf{0.76 \pm 0.02}$	0.85 ± 0.01	0.80 ± 0.02	0.83 ± 0.05	$\mathbf{0.85 \pm 0.03}$	6.8 ± 1.6
	Extreme-ray	0.74 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.81 ± 0.03	$\mathbf{0.98 \pm 0.04}$	0.79 ± 0.03	4.5 ± 0.4
	Random	0.51 ± 0.03		0.53 ± 0.03			26.1 ± 2.5
LunarLander, $s = 3$, mixed	LoRe	0.69 ± 0.04	0.92 ± 0.13	0.69 ± 0.07	0.84 ± 0.10	0.63 ± 0.04	$\mathbf{2.4 \pm 0.5}$
	Sparse-LoRe	0.76 ± 0.06	0.84 ± 0.01	$\mathbf{0.85 \pm 0.04}$	0.85 ± 0.04	0.82 ± 0.03	4.7 ± 0.6
	Identified LoRe	$\mathbf{0.78 \pm 0.02}$	0.86 ± 0.05	0.81 ± 0.03	0.86 ± 0.03	$\mathbf{0.88 \pm 0.04}$	6.2 ± 1.3
	Extreme-ray	0.77 ± 0.02	$\mathbf{1.00 \pm 0.00}$	0.81 ± 0.03	$\mathbf{0.98 \pm 0.04}$	0.80 ± 0.03	4.3 ± 1.0
	Random	0.52 ± 0.03		0.53 ± 0.03			25.6 ± 4.6

Table 9: Full version of Table 2, with 2-sparse and Dirichlet users on the language datasets. Held-out deterministic preference accuracy (%) of unseen users after $Q = 8$ answers, mean ± standard deviation across seeds (three per language setting, five per control task). All four query-selection rules use the *ground-truth dictionary* and the same composition estimator, so the table isolates the choice of questions. Pooled BT uses no labels from the target user. Independent BT, all pairs, is a per-user Bradley–Terry model fitted on every labeled comparison of the target user (512 on the language datasets, 456 on the control tasks), with its ridge chosen on validation users; it is a full-data reference, not an upper bound for rules that use the dictionary. $^{+}$/$^{-}$: CS-Elicit is better/worse than MI by a paired Wilcoxon test over users at $p < 0.01$. Best rule per row in bold. UltraFeedback with s=1 and Dirichlet users is at ceiling for MI and CS-Elicit.

Dataset	Users	Pooled BT	Random	MI	VR	CS-Elicit	Indep. BT, all pairs
PersonalLLM	s=1	77.5 ±1.4	83.9 ±0.5	87.3 ±0.3	83.4 ±1.9	**93.0**$^{+}$ ±0.8	96.7 ±0.4
	s=2	82.7 ±1.8	83.7 ±1.2	**88.6** ±1.3	85.5 ±1.1	83.2^{-} ±1.4	96.1 ±0.4
	Dir(0.001)	74.9 ±3.9	82.1 ±1.2	**86.2** ±0.5	82.1 ±2.7	86.0 ±2.0	96.6 ±0.2
UltraFeedback	s=1	90.5 ±0.6	96.4 ±0.5	99.9 ±0.0	90.7 ±0.4	**100.0**$^{+}$ ±0.0	100.0 ±0.0
	s=2	92.0 ±0.5	93.4 ±0.5	96.6 ±0.4	92.1 ±0.6	**99.2**$^{+}$ ±0.3	97.9 ±0.3
	Dir(0.001)	91.0 ±0.5	96.3 ±0.6	99.7 ±0.1	90.9 ±0.5	**99.9**$^{+}$ ±0.0	99.5 ±0.3
MountainCar (K=2)	s=1	91.6 ±0.7	95.4 ±1.2	97.6 ±0.6	93.4 ±1.0	**99.3**$^{+}$ ±0.5	98.0 ±0.5
CartPole (K=4)	s=1	61.6 ±1.5	85.9 ±1.7	91.7 ±1.3	85.5 ±3.3	**94.3**$^{+}$ ±1.7	97.5 ±0.3
Acrobot (K=6)	s=1	63.1 ±2.8	77.7 ±2.9	84.5 ±2.8	73.1 ±2.2	**97.3**$^{+}$ ±1.6	95.2 ±1.8
LunarLander (K=8)	s=1	58.7 ±1.9	78.9 ±1.7	84.7 ±1.4	77.2 ±3.0	**89.0**$^{+}$ ±3.6	96.4 ±0.6
Kitchen (K=14)	s=1	55.6 ±0.5	63.3 ±1.5	63.4 ±2.1	62.4 ±1.8	**64.4** ±1.8	77.3 ±2.6

Table 10: Per-user wall time (ms) to first reach the reference target accuracy, for 1-sparse users on the ground-truth dictionary. The target is 90% of the way from the zero-question accuracy to the highest accuracy that CS-Elicit, random, and MI all reach. Brackets: 95% bootstrap interval over users (∞: more than 2.5% of resamples never reach the target). Parentheses: interpolated number of questions. n.r.: not reached within the budget. Setup: one-time construction of CS-Elicit's probe tree per population, in seconds. $N^\star$: number of users after which the setup is recovered against MI or random.

Dataset	Target (%)	CS-Elicit	Random	MI	VR	Setup (s)	$N^\star$ vs MI	$N^\star$ vs random
MountainCar	96.8	3.0 [2.0, 3.4] (3 q)	25.5 [20.6, 38.6] (53 q)	28.9 [24.8, 34.1] (6 q)	n.r.	0.47	18	21
CartPole	93.9	11.5 [7.5, 19.2] (7 q)	43.2 [37.2, 49.9] (37 q)	84.2 [71.0, 97.5] (13 q)	139.2 [114.4, 160.6] (42 q)	1.35	18	42
Acrobot	91.2	15.1 [13.7, 16.2] (6 q)	98.5 [86.6, 111.2] (81 q)	159.8 [126.3, 210.2] (20 q)	n.r.	2.52	17	30
LunarLander	92.2	26.6 [21.8, 37.9] (14 q)	94.0 [79.1, 105.0] (50 q)	166.3 [141.9, 198.6] (19 q)	482.9 [344.7, ∞] (126 q)	6.91	49	102
PersonalLLM	93.2	8.5 [7.0, 10.1] (8 q)	21.0 [18.8, 23.9] (44 q)	56.4 [49.3, 64.9] (18 q)	n.r.	4.91	102	392

Table 11: The unbudgeted CS-Elicit walk on the ground-truth dictionary. Cert.: nodes with a valid pair, over all seeds. Pairs: distinct valid pairs of the median node. $\vartheta_{(1)}/\vartheta_{(9)}$: best and ninth-best margin of the median node. r at $\vartheta_{(1)}$: answers per node the theorem would need if every node could reuse its best pair, at the node where this is largest (∞: some node has no valid pair). Exact: $\Pr(\hat{S} = S)$ (%) for r answers per node on r distinct pairs; Thm.: the schedule of Theorem 2(i), n/a where it does not apply. FI, FE: false inclusion and exclusion rates (%), and Q: questions per user, each for $r = 1$ / $r = 9$.

						Exact (%)							
Dataset	s	Cert.	Pairs	$\vartheta_{(1)}/\vartheta_{(9)}$	r at $\vartheta_{(1)}$	r=1	3	9	17	Thm.	FI (%)	FE (%)	Q
PersonalLLM	1	54/54	157	0.32 / 0.19	206	62.0	79.0	89.7	94.8	n/a	0.3 / 0.0	36.1 / 10.2	6.3 / 59.5
	2	38/54	2	0.05 / 0.00	∞	1.7	2.1	2.4	2.4	n/a	0.2 / 0.0	93.0 / 91.3	1.0 / 7.0
UltraFeedback	1	18/18	1,505	2.25 / 1.58	10	100.0	100.0	100.0	100.0	100.0 (r=11)	0.0 / 0.0	0.0 / 0.0	4.0 / 36.0
	2	18/18	84	0.75 / 0.43	13	93.6	99.3	100.0	100.0	n/a	1.3 / 0.0	2.0 / 0.0	5.4 / 48.6
MountainCar	1	10/10	3,373	1.37 / 1.22	12	88.3	97.1	99.9	100.0	100.0 (r=15)	6.1 / 0.0	6.1 / 0.0	2.0 / 18.0
CartPole	1	30/30	2,927	1.28 / 1.11	785	54.5	64.7	73.8	78.1	n/a	8.5 / 5.3	24.4 / 10.3	4.0 / 36.0
	2	30/30	263	0.21 / 0.15	3,688	35.2	46.3	59.0	63.6	n/a	17.9 / 8.8	28.6 / 16.0	5.3 / 48.7
Acrobot	1	50/50	290	0.52 / 0.32	2,701	77.3	81.8	84.3	84.6	n/a	3.6 / 2.1	7.1 / 5.9	5.2 / 46.0
	2	42/50	12	0.10 / 0.01	∞	14.8	14.5	15.1	15.6	n/a	4.5 / 1.7	61.9 / 62.5	4.5 / 28.1
LunarLander	1	67/70	44	0.25 / 0.12	∞	38.4	48.4	52.4	53.5	n/a	3.4 / 1.7	49.1 / 41.3	5.5 / 41.2
	2	34/70	0	0.00 / 0.00	∞	5.0	5.9	6.3	6.6	n/a	0.8 / 0.1	80.0 / 74.0	4.5 / 39.5
Kitchen	1	56/130	0	0.00 / 0.00	∞	0.0	0.0	0.0	0.0	n/a	0.0 / 0.0	100.0 / 100.0	0.0 / 0.0
	2	3/130	0	0.00 / 0.00	∞	0.0	0.0	0.0	0.0	n/a	0.0 / 0.0	100.0 / 100.0	0.0 / 0.0

Table 12: Depth-wise node error (%) of the unbudgeted walk, $r = 1$ / $r = 9$ answers per node: the share of tested nodes at each depth of the dyadic tree whose decision is wrong. –: no node tested at that depth (Kitchen: both depth-1 nodes lack a valid pair).

Dataset	s	Depth 1	Depth 2	Depth 3	Depth 4
PersonalLLM	1	17.9 / 5.0	11.7 / 2.0	5.6 / 1.4	2.2 / 0.0
	2	25.3 / 26.4	16.3 / 3.6	8.6 / 1.4	8.7 / 0.2
UltraFeedback	1	0.0 / 0.0	0.0 / 0.0	–	–
	2	1.1 / 0.0	1.5 / 0.0	–	–
MountainCar	1	6.1 / 0.0	–	–	–
CartPole	1	1.5 / 0.0	24.2 / 13.1	–	–
	2	5.6 / 0.3	25.0 / 14.2	–	–
Acrobot	1	0.2 / 0.0	10.5 / 6.2	6.0 / 3.7	–
	2	16.1 / 15.0	24.7 / 19.2	11.8 / 9.3	–
LunarLander	1	4.9 / 0.0	19.8 / 13.9	24.3 / 16.7	–
	2	11.6 / 1.7	15.1 / 4.3	11.1 / 4.2	–
Kitchen	1	–	–	–	–
	2	–	–	–	–

Table 13: Exact support recovery (%) of each rule's fitted composition after $Q = 8$ answers, from the runs of Table 9. Left: support $\{k : \hat{a}_k \geq \Delta_-/2\}$. Right: the s largest coordinates of $\hat{a}$, which uses the true s. Tree: CS-Elicit's tree output $\hat{S}$ itself, one answer per node as in Table 2.

		$\hat{a}_k \geq \Delta_-/2$				Top-s				
Dataset	s	Random	MI	VR	CS-Elicit	Random	MI	VR	CS-Elicit	Tree
PersonalLLM	1	29.3	24.0	15.3	66.7	43.3	85.3	52.7	73.3	59.3
	2	3.3	8.0	10.7	3.3	14.0	32.0	15.3	16.0	2.0
UltraFeedback	1	63.3	100.0	93.3	100.0	76.7	100.0	99.3	100.0	100.0
	2	20.7	89.3	69.3	92.7	48.7	93.3	86.0	94.0	92.0
MountainCar	1	75.2	99.2	68.8	96.8	75.2	99.2	68.8	96.8	91.2
CartPole	1	54.4	72.8	53.6	78.4	68.0	82.4	68.0	82.4	55.2
	2	33.6	54.4	36.8	45.6	47.2	61.6	48.0	55.2	41.6
Acrobot	1	44.8	68.8	28.8	92.8	57.6	91.2	39.2	95.2	78.4
	2	12.8	35.2	11.2	27.2	29.6	65.6	20.0	39.2	15.2
LunarLander	1	31.2	32.8	11.2	64.8	49.6	78.4	52.8	68.8	38.4
	2	5.6	12.0	8.8	11.2	18.4	46.4	20.8	24.0	5.6
Kitchen	1	18.4	0.0	9.6	19.2	38.4	66.4	36.8	39.2	0.0
	2	0.8	0.0	3.2	3.2	8.0	20.8	9.6	12.8	0.0

Table 14: Native versus frozen-dictionary inference on Gym control at $Q = 16$, with random questions. Mean accuracy over four environments and five seeds (20 cells).

Method	Native	Frozen + simplex	Frozen + gated
VPL	0.8643	0.7241	0.7506
SPL	0.8573	0.7429	0.7689

Table 15: Per-task accuracy for the control row of Table 3 ($Q = 16$, mean $\pm$ standard deviation over the 125 test users of each task; Gym at $K = 8$, Kitchen at $K = 7$). Columns as in Table 3; the four question rules use one extracted dictionary and the shared fit. Best $Q = 16$ accuracy per task in bold.

	End-to-end baselines					Extracted dictionary, shared fit				
Task	Pooled BT	LoRe	VPL	SPL	PAL	Random	MI	VR	CS-Elicit	Ceiling
MountainCar	.823±.072	.824±.077	.877±.060	.861±.061	.870±.068	.895±.066	**.935**±.027	.905±.057	.895±.065	.955±.018
CartPole	.706±.107	.815±.081	.839±.081	.842±.085	.859±.065	.886±.074	**.936**±.040	.901±.077	.886±.097	.954±.018
Acrobot	.827±.074	.835±.077	.893±.060	.883±.064	.903±.051	.915±.056	**.945**±.029	.912±.067	.907±.071	.968±.013
LunarLander	.743±.094	.783±.100	.847±.079	.844±.088	.844±.080	.889±.086	**.926**±.045	.894±.068	.894±.076	.939±.048
Kitchen	.614±.089	.857±.075	.789±.127	.799±.138	.852±.078	.899±.084	**.932**±.068	.897±.088	.923±.130	.958±.032

Table 16: Question rules on the same frozen extracted dictionary as ours in Table 3, all followed by the same gated fit, after $Q = 16$ answers: accuracy and NLL, mean $\pm$ standard deviation over test users. MI and VR are APReL's mutual information and corrected volume removal, choosing questions with their own posterior. Best per dataset in bold.

	Random		MI		VR		CS-Elicit (ours)	
Dataset	Acc.	NLL	Acc.	NLL	Acc.	NLL	Acc.	NLL
PersonalLLM	.935±.092	**.359**±.114	.925±.110	.392±.118	**.937**±.087	.367±.116	.926±.116	.376±.131
TL;DR	**.700**±.054	**.564**±.057	**.700**±.055	**.564**±.058	.698±.060	.565±.057	.699±.045	**.564**±.052
UF-P-4	.596±.067	.926±.097	.592±.065	**.915**±.081	.583±.082	.998±.201	**.635**±.051	.927±.092
Gym (4 tasks)	.896±.072	.475±.096	**.935**±.037	**.438**±.053	.903±.068	.470±.088	.896±.078	.493±.205
Kitchen	.899±.084	.431±.171	**.932**±.068	**.367**±.105	.897±.088	.487±.346	.923±.130	.503±.577

Table 17: Representation limit on PersonalLLM ($K = 10$, 512-dimensional features, held-out test prompts). R^2: held-out fit of each ground-truth factor from the dictionary. Corr.: Pearson correlation of Hungarian-matched factors. Span: R^2 of the oracle readout from the dictionary.

Dictionary	Mean R^2	Min. R^2	Mean corr.	Span
Ground truth	1.00	1.00	1.00	–
Oracle linear readout	0.67	0.44	0.82	1.00
Extracted	0.44	0.22	0.52	0.68