Efficient Ranking of Mathematical Reasoning Chains via Calibrated Interpolation Sort

Rohan Dalal [1]

Abstract

Ranking mathematical reasoning chains by quality (for best-of-n decoding, reward model training, or verification) requires expensive pairwise oracle queries. Standard noisy sort needs $\Theta(n \log n)$ such queries per batch. We show that mathematical reasoning chains have a property we call *score commensurability*: when chains are clustered by embedding similarity, quality becomes approximately linear in the latent space, and a small calibration set of k preference pairs suffices to learn a quality direction $\widehat{w}$. Feeding projected scores to interpolation sort then ranks n chains in $O(n \log \log n)$ key comparisons with zero additional oracle queries. On PRM800K (37k human-graded reasoning chains across 900 MATH problems), embedding-based clustering reduces pairwise noise from $\hat{p} \approx 0.31$ to 0.11, enabling calibrated methods to achieve $\tau_b < 0.12$ with $k = 5000$ pairs. A key structural finding: mathematical topic labels do not ensure commensurability. "intermediate algebra" spans polynomial factoring, functional equations, and inequalities with no shared quality direction, but k-means clustering on embeddings does. The calibrated direction $\widehat{w}$ amortizes: once learned, each new batch is ranked with zero additional oracle queries.

1. Introduction

Evaluating mathematical reasoning is expensive. Whether selecting the best chain-of-thought solution via best-of-n decoding, training process reward models, or verifying multi-step proofs, the core task is ranking a batch of n reasoning chains by quality. Each pairwise quality judgment (human or reward model) is a costly *oracle query*, and noisy comparison sort requires $\Theta(n \log n)$ of them per batch. Can the structure of mathematical reasoning reduce this cost?

We argue yes. Mathematical solution quality has a natural scalar structure: a chain with 9/10 correct steps is reliably better than one with 7/10, unlike open-ended text ranked by "helpfulness." The empirical question is whether sentence embeddings preserve this structure, i.e., whether quality is approximately linear in embedding space. We find that it is, but only within semantically similar clusters: when chains addressing similar problems are pooled together, a single direction w^* in embedding space separates good from bad solutions (held-out error $\hat{p} = 0.11$).

This connects to *interpolation sort* (Mehlhorn & Tsakalidis, 1990), which exploits distributional knowledge to sort in $O(n \log \log n)$ key comparisons. Our two-phase approach: (1) learn $\widehat{w} \approx w^*$ from k noisy preference pairs (expensive oracle queries, paid once), then (2) sort each batch by projected scores using interpolation sort ($O(n \log \log n)$ cheap key comparisons, zero oracle queries).

Contributions. (1) We formalize *calibrated interpolation sort* and prove that $k = O(d/\epsilon^2)$ preference pairs suffice for Kendall-τ error $\leq \epsilon/\pi$. (2) On PRM800K (Lightman et al., 2023), we discover that score commensurability (the condition enabling calibrated sort) depends on embedding-based clustering, not mathematical topic labels. (3) We show this is a structural property of mathematical problems: "intermediate algebra" is not semantically coherent, but k-means clusters on sentence embeddings are. (4) We verify on UltraFeedback (Cui et al., 2023) that the same principle holds for general LLM outputs, confirming generality.

2. Problem Setup and Theory

Latent space model. We have n items with representations $z_1, \ldots, z_n \in \mathbb{R}^d$ drawn i.i.d. from $\mathcal{D}$; the base case is $\mathcal{D} = \mathcal{N}(0, I_d)$. Quality is $q(z) = {w^*}^\top z$ for unknown unit vector w^*. We observe k noisy oracle queries: each returns the correct preference with probability $1-p$ ($\mathrm{BSC}(p), p < 1/2$). In practice, oracle labels may be deterministic (derived from ground-truth scores) rather than stochastic; the BSC model then captures the effective noise induced by projecting a non-linear quality function onto a single direction w.

Definition 2.1 (Score commensurability)**.** A pool of n items with embeddings z_i and ground-truth qualities q_i is *p-commensurable* if there exists a unit vector $w \in \mathbb{R}^d$ such that $\frac{1}{\binom{n}{2}} \sum_{i<j} \mathbf{1}\big[\mathrm{sign}(w^\top z_i - w^\top z_j) \neq \mathrm{sign}(q_i - q_j)\big] \leq p$

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Rohan Dalal <rdalal30@gatech.edu>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Algorithm 1 Calibrated Interpolation Sort

Phase 1: Calibration (run once, k oracle queries)
Sample k pairs (z_a, z_b) from embedding pool
Query oracle for pairwise preferences $y_{ab} \in \{-1, +1\}$
Compute differences $\delta_{ab} = z_a - z_b$
Fit logistic regression: $\widehat{w} = \arg\max_w \sum_j \log \sigma(y_j\, w^\top \delta_j)$
Phase 2: Sorting (run per batch, 0 oracle queries)
Input: new batch $z_1, \ldots, z_n$; calibrated $\widehat{w}$
Project: $s_i \leftarrow \widehat{w}^\top z_i$ for each item
Estimate $\hat{\mu}, \hat{\sigma}$ from $\{s_i\}$
Compute predicted ranks: $r_i \leftarrow \Phi((s_i - \hat{\mu})/\hat{\sigma})$
Return interpolation sort of items by s_i using $\{r_i\}$

where $p < 1/2$. Operationally, $\hat{p}$ (held-out logistic error on pairwise labels) estimates the commensurability parameter p. Note that p in Definition 2.1 measures structural mismatch (how well a linear direction explains quality), not injected label noise; the BSC(p) model in Proposition 2.2 treats these symmetrically since both manifest as pairwise disagreement.

Cost model. We distinguish *oracle queries* $\mathcal{O}$ (expensive pairwise preference judgments, e.g., human or reward model) from *key comparisons* $\mathcal{C}$ (cheap scalar comparisons on floats within interpolation sort, $\sim 0.45n$ total, verified from $n = 10^2$ to 10^5).

Method	Oracle queries	Key comp. / batch
Noisy comp. sort	$\Theta(n \log n)$/batch	0
BTL / Copeland	k/batch	$O(n \log n)$
RankNet	k once	$O(n \log n)$
Calibrated sort (ours)	k once	$O(n \log \log n)$

Notation. We report ranking quality as Kendall-τ_b *distance*: $\tau_b = (1 - r)/2$ where r is the Kendall-τ_b correlation, so $\tau_b \in [0, 0.5]$ with lower meaning better agreement. When there are no ties, τ_b equals the fraction of misordered pairs.

Proposition 2.2 (Calibration bound)**.** *Let $z_i \overset{iid}{\sim} \mathcal{N}(0, I_d)$. Given k pairs (z_a, z_b) sampled uniformly at random, with labels $y_{ab} \in \{-1, +1\}$ generated by applying* BSC(p) *noise to* $\mathrm{sign}({w^*}^\top (z_a - z_b))$*, the logistic regression MLE $\widehat{w}$ on pairwise differences $\delta_{ab} = z_a - z_b$ satisfies $\mathbb{E}[\angle(\widehat{w}, w^*)] = O\Big(\sqrt{d/(k(1-2p)^2)}\Big)$.*

Proof. Each difference $\delta_{ab} \sim \mathcal{N}(0, 2I_d)$. The true conditional is $\Pr[y_{ab} = +1 \mid \delta_{ab}] = (1-p)\,\mathbf{1}[{w^*}^\top \delta_{ab} > 0] + p\,\mathbf{1}[{w^*}^\top \delta_{ab} \le 0]$. Fitting a logistic model $\Pr[y = +1 \mid \delta] = \sigma(w^\top \delta)$ is misspecified, but the MLE direction $\widehat{w}/\|\widehat{w}\|$ is consistent for w^* because BSC noise preserves the symmetry $\Pr[y \mid \delta] = \Pr[-y \mid -\delta]$.

Consistency of direction. Write $w = \alpha w^* + w_\perp$ where $w_\perp \perp w^*$. By Theorem 5.14 of van der Vaart (1998), the MLE under a misspecified model converges to the KL-minimizing parameter. Since $w_\perp^\top \delta$ is independent of ${w^*}^\top \delta$ (orthogonal Gaussian projections) and labels depend only on $\mathrm{sign}({w^*}^\top \delta)$, the component $w_\perp^\top \delta$ adds noise without improving likelihood, so the KL-minimizer has $w_\perp = 0$, i.e., $w^*_{\text{logistic}} \propto w^*$.

Sandwich variance. For misspecified M-estimators, the asymptotic covariance is $A^{-1}BA^{-1}$ (Theorem 5.23, van der Vaart 1998), where $A = \mathbb{E}\big[\sigma(w^\top\delta)\,\sigma(-w^\top\delta)\ \ \delta\delta^\top\big]$ and $B = \mathbb{E}\big[\nabla \log \sigma(y\, w^\top \delta)\ \ \nabla \log \sigma(y\, w^\top \delta)^\top\big]$. By rotational symmetry of $\delta \sim \mathcal{N}(0, 2I_d)$, both decompose as $\alpha_\parallel w^* {w^*}^\top + \alpha_\perp (I_d - w^* {w^*}^\top)$ for scalars depending on p.

Angular error. The angle $\angle(\widehat{w}, w^*)$ depends only on $\widehat{w}_\perp$. In the $d-1$ orthogonal directions, each pair contributes Fisher information $O((1-2p)^2/d)$; the $(1-2p)^2$ factor arises because BSC(p) noise attenuates the gradient signal. With k pairs, total perpendicular information is $O(k(1-2p)^2/d)$, giving $\mathrm{Var}[\widehat{w}_\perp] = O(d/(k(1-2p)^2))$. Since $\angle(\widehat{w}, w^*) \approx \|\widehat{w}_\perp\|/\|\widehat{w}_\parallel\|$ for small angles and $\widehat{w}_\parallel$ converges to a constant, Jensen's inequality yields $\mathbb{E}[\angle(\widehat{w}, w^*)] = O\Big(\sqrt{d/(k(1-2p)^2)}\Big)$. □

Proposition 2.3 (Sorting degradation)**.** *Let $z_1, \ldots, z_n \overset{iid}{\sim} \mathcal{N}(0, I_d)$ and let $\widehat{w}$ be any fixed unit vector with $\theta = \angle(\widehat{w}, w^*)$. Then $\mathbb{E}[\tau_b] = \theta/\pi$, where the expectation is over the batch.*

Proof. For any pair (z_i, z_j), let $\delta = z_i - z_j \sim \mathcal{N}(0, 2I_d)$. The pair is misordered iff $\mathrm{sign}({w^*}^\top \delta) \neq \mathrm{sign}(\widehat{w}^\top \delta)$. Define $U = {w^*}^\top \delta$ and $V = \widehat{w}^\top \delta$. Since δ is Gaussian, (U, V) is bivariate normal with $\mathbb{E}[U] = \mathbb{E}[V] = 0$, $\mathrm{Var}[U] = 2$, $\mathrm{Var}[V] = 2$, and $\mathrm{Cov}(U, V) = 2\,{w^*}^\top \widehat{w} = 2\cos\theta$, giving correlation $\rho = \cos\theta$. By Sheppard's formula, $\Pr[\mathrm{sign}(U) \neq \mathrm{sign}(V)] = \arccos(\rho)/\pi = \theta/\pi$. Since $\tau_b = \binom{n}{2}^{-1} \sum_{i<j} \mathbf{1}[\text{misordered}]$, linearity of expectation gives $\mathbb{E}[\tau_b] = \theta/\pi$. □

Corollary 2.4. *With $k = O(d/\epsilon^2)$ calibration queries, calibrated interpolation sort achieves $\mathbb{E}[\tau_b] \le \epsilon/\pi$ using k oracle queries plus $O(n \log \log n)$ key comparisons, versus $\Theta(n \log n/(1 - H(p)))$ oracle queries for noisy comparison sort (Braverman & Mossel, 2008).*

Plug-in estimation. The $O(n \log \log n)$ bound assumes known μ, σ for the Gaussian CDF. In practice, we estimate $\hat{\mu}, \hat{\sigma}$ from the n projected scores. For $n \ge 100$, the estimation error is $O(1/\sqrt{n})$, which shifts predicted ranks by $O(1)$ positions on average, negligible relative to bucket size.

3. Synthetic Validation

We validate the theory on controlled settings before applying it to mathematical reasoning chains.

Synthetic Gaussians. We sweep $d \in \{8, 32, 64, 256\}$, $p \in \{0.05, \ldots, 0.4\}$, k from 10 to 5000 (30 seeds, $n = 1000$). Angular error decays as $O(1/\sqrt{k})$ across all conditions. At $d = 64$, $p = 0.1$: $k = 500$ yields $\cos(\angle(\widehat{w}, w^*)) > 0.91$ ($\tau_b \approx 0.13$); $k = 1000$ yields > 0.96 ($\tau_b < 0.09$). Interpolation sort uses $\sim 0.45n$ key comparisons (constant across n), versus $\sim \log_2 n \cdot n$ for merge sort.

Mixtures of Gaussians. With aligned quality directions across clusters, τ_b matches the Gaussian baseline regardless of cluster count or separation. With angular spread ≥ 0.5 rad, τ_b jumps from 0.09 to 0.29: the bottleneck is quality direction heterogeneity, not non-Gaussianity. This directly foreshadows our PRM800K finding: mixing problems with different quality directions destroys commensurability.

β-VAE latent spaces. β-VAEs (Higgins et al., 2017) on MNIST ($d = 32$, $\beta \in \{0.1, \ldots, 4.0\}$, 15 seeds) show remarkable robustness: despite $12\times$ variation in Anderson-Darling statistics across β, all curves converge to $\tau_b \approx 0.035$ at $k = 3162$. Interpolation sort tolerates moderate distributional mismatch as long as no bucket receives more than $O(1)$ elements.

4. Ranking Mathematical Reasoning Chains

Data. PRM800K (Lightman et al., 2023): 37,464 reasoning chains across 900 MATH competition problems, with step-level human correctness labels. Chain quality = fraction of correct steps (continuous, 84 unique values, mean 0.86). Unlike subjective "helpfulness" scores, this metric has a clear ordinal structure; the empirical question is whether sentence embeddings preserve it linearly, which we measure via $\hat{p}$. Problem types span intermediate algebra (12,194), geometry (5,807), algebra (4,996), precalculus (4,895), number theory (4,041), counting & probability (3,591), and prealgebra (1,958). Embeddings: `all-MiniLM-L6-v2` ($d = 384$, $d_{\text{eff}} = 163$ via PCA at 95% variance).

Three pooling conditions. (A) *Mixed*: $n = 1000$ chains across all problem types, the naïve approach. (B) *Stratified*: $n = 1000$ from the largest k-means cluster ($K = 50$) on embeddings.[1] (C) *Single-type*: $n = 1000$ from "intermediate algebra" only, i.e., topic-based filtering. We compare three calibration methods (linear/logistic, kernel/RBF, MLP $384 \to 128 \to 64 \to 1$) and a RankNet baseline (identical MLP architecture, pairwise cross-entropy, standard sort at inference) against greedy pairwise sort, Copeland counting (wins − losses), and Bradley-Terry MLE (Bradley & Terry, 1952). All use the same k uniformly random oracle pairs (20 seeds, $k \in \{500, \ldots, 5000\}$). Oracle labels are generated deterministically from ground-truth quality scores (with random tie-breaking); no BSC noise is injected. The reported $\hat{p}$ thus reflects pure structural mismatch, the fraction of pairs whose true ordering disagrees with the learned linear direction, not label noise.

[1] With $K = 50$ and 900 prompts, the largest cluster contains $\sim$828 chains. When cluster size $< n$, we use all available chains (effectively $n = 828$). Sensitivity to K is analyzed below.

Table 1. Kendall-τ_b on PRM800K reasoning chains ($d = 384$, $d_{\text{eff}} = 163$, 20 seeds; SE < 0.003). Stratified pooling enables calibrated methods to overtake greedy. Single-type pooling (intermediate algebra) is worse than stratified despite being mathematically "homogeneous."

	Method	$k=500$	$k=1$k	$k=2$k	$k=5$k	$\hat{p}$
Mixed	Greedy	0.387	0.318	0.222	0.142	–
	Copeland	0.336	0.278	0.221	0.155	–
	BTL	0.389	0.298	0.193	**0.093**	–
	Linear	0.421	0.420	0.383	0.342	0.34
	Kernel	0.427	0.411	0.381	0.348	0.34
	MLP	0.408	0.387	0.351	0.318	0.31
Strat.	Greedy	0.375	0.303	0.222	0.286	–
	Copeland	0.326	0.264	0.212	0.152	–
	BTL	0.372	0.274	0.178	**0.096**	–
	RankNet	0.248	0.206	0.157	0.116	0.11
	Linear	0.262	0.208	0.158	0.128	0.13
	Kernel	0.274	0.234	0.202	0.182	0.18
	MLP	0.247	0.206	0.156	0.115	0.11
Single type	Greedy	0.382	0.316	0.237	0.198	–
	Copeland	0.338	0.276	0.219	0.154	–
	BTL	0.391	0.298	0.192	**0.090**	–
	Linear	0.377	0.351	0.328	0.295	0.30
	Kernel	0.383	0.361	0.334	0.306	0.31
	MLP	0.361	0.337	0.303	0.266	0.27

Results (Table 1). On mixed pools, BTL MLE dominates ($\tau_b = 0.093$ at $k = 5000$), consistent with the high $\hat{p} \approx 0.31$–0.34: no single quality direction serves diverse math problems. On stratified pools, $\hat{p}$ drops to 0.11–0.18 and calibrated methods become competitive: MLP reaches $\tau_b = 0.115$ ($\hat{p} = 0.11$), overtaking greedy at $k \approx 500$ (Figure 2). RankNet (identical MLP architecture with standard sort at inference) achieves $\tau_b \approx 0.116$, matching calibrated MLP, confirming that the ranking quality comes from the learned scorer, while interpolation sort contributes an additional $O(n \log n) \to O(n \log \log n)$ reduction in key comparisons. Greedy pairwise sort constructs a tournament graph from k random oracle queries and returns a topological ordering. At small k the graph is sparse and cycles are rare, but as k/n grows, the graph densifies and inconsistencies accumulate: even with deterministic oracle labels, ties and near-ties in ground-truth quality create contradictory edges (e.g., $a > b$, $b > c$, $c > a$ when all three have similar quality scores), and the topological sort must break these cycles arbitrarily. This explains the non-monotonicity in the stratified condition

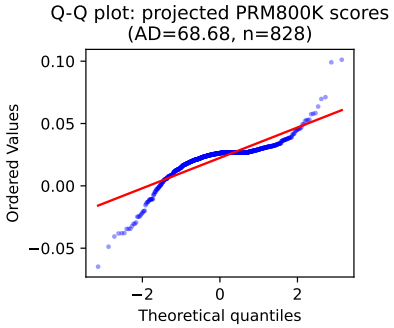

Figure 1. Q-Q plot of projected PRM800K scores against a normal distribution ($n = 828$). Tail deviations (excess kurtosis 8.8, skewness -1.4) are visible but do not prevent interpolation sort from converging.

($0.303 \to 0.286$ at $k = 5000$), and the dramatic degradation on UltraFeedback ($0.234 \to 0.463$; Section 5) where the denser quality score distribution amplifies cycle formation. Concretely, ranking $n = 1000$ chains requires $k = 500$ oracle queries (once) plus $\sim$450 key comparisons per batch, versus $\sim$10,000 oracle queries per batch for merge sort.

Why topic labels fail. Filtering to "intermediate algebra" ($\hat{p} \approx 0.27$–0.30) is worse than embedding-based stratification ($\hat{p} \approx 0.11$–0.18). This is a structural finding about mathematical problem spaces: the topic label "intermediate algebra" spans polynomial factoring, functional equations, and inequalities, problems with fundamentally different solution strategies and quality signatures. Embedding-based clustering groups problems by actual solution structure (e.g., chains that manipulate similar algebraic objects), producing pools where a single quality direction is meaningful. BTL still achieves $\tau_b = 0.090$ in the single-type condition because it estimates item-level abilities without requiring a global quality direction; it is robust to high $\hat{p}$ when k/n is large enough.

Robustness to non-Gaussianity. Proposition 2.3 assumes Gaussian projections. On PRM800K's largest embedding cluster, projected scores $s_i = \widehat{w}^\top z_i$ yield Anderson-Darling statistic 68.7 (5% critical value: 0.783), with excess kurtosis 8.8 and skewness -1.4, clearly non-Gaussian (Figure 1). Despite this, interpolation sort converges as predicted: the β-VAE experiments (Section 3) show $12\times$ variation in Anderson-Darling statistics with identical τ_b convergence. The key requirement is that no bucket receives more than $O(1)$ elements, which holds when the projected score density f satisfies $\sup_x f(x)/\phi_{\hat{\mu},\hat{\sigma}}(x) \leq C$ for some constant C, where $\phi_{\hat{\mu},\hat{\sigma}}$ is the fitted Gaussian density used for bucket assignment.

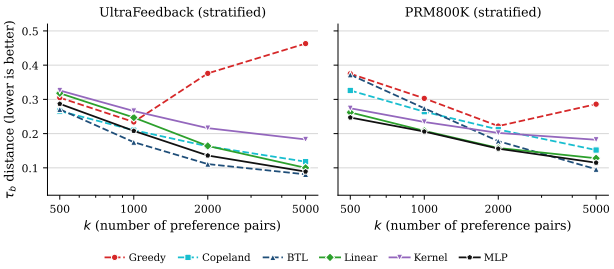

Figure 2. τ_b vs. k on stratified pools. Greedy (red, dashed) degrades at large k; BTL and calibrated methods improve monotonically. MLP (black, solid) matches BTL at $k = 5000$.

Table 2. Kendall-τ_b on UltraFeedback ($d = 384$, 20 seeds; SE < 0.003). The same commensurability pattern holds: stratified pooling enables calibrated methods.

	Method	$k=500$	$k=1$k	$k=2$k	$k=5$k	$\hat{p}$
Mixed	Greedy	0.386	0.327	0.247	0.209	–
	Copeland	0.337	0.279	0.224	0.161	–
	BTL	0.387	0.299	0.198	**0.106**	–
	Linear	0.447	0.447	0.408	0.375	0.38
	Kernel	0.452	0.442	0.412	0.380	0.38
	MLP	0.428	0.406	0.380	0.335	0.34
Strat.	Greedy	0.305	0.234	0.376	0.463	–
	Copeland	0.266	0.210	0.163	0.118	–
	BTL	0.271	0.175	0.111	**0.081**	–
	RankNet	0.288	0.207	0.136	0.089	0.09
	Linear	0.318	0.247	0.164	0.100	0.10
	Kernel	0.326	0.266	0.216	0.183	0.19
	MLP	0.287	0.208	0.136	0.089	0.09

5. Comparison with General LLM Outputs

To test whether the commensurability principle is specific to mathematical reasoning or general, we repeat the experiment on UltraFeedback (Cui et al., 2023) ($\sim$64k prompts, 4 completions each, GPT-4 helpfulness scores; $d = 384$).

Results (Table 2). The same pattern holds: mixed pools have high $\hat{p}$ (0.34–0.38) and BTL dominates; stratified pools reduce $\hat{p}$ to 0.09–0.10 and calibrated MLP matches BTL. Greedy degrades dramatically ($0.234 \to 0.463$). The crossover where calibrated methods overtake greedy occurs at $k \approx 1000$, later than PRM800K's $k \approx 500$ (Table 3), consistent with PRM800K's lower d_{eff} (163 vs. 235) requiring fewer calibration pairs.

Table 3. Dataset comparison (stratified pools). PRM800K's lower d_{eff} and earlier crossover are consistent with fewer effective parameters to learn.

	UltraFeedback	PRM800K
d_{eff} (PCA 95%)	235	163
Best calibrated τ_b	0.089 (MLP)	0.115 (MLP)
Best overall τ_b	0.081 (BTL)	0.096 (BTL)
Greedy-cal. crossover	$k \approx 1000$	$k \approx 500$

6. Related Work

Sorting with distributional knowledge. Interpolation sort achieves $O(n \log\log n)$ or $O(n)$ expected key comparisons when the input distribution is known (Mehlhorn & Tsakalidis, 1990). Recent work on learning-augmented algorithms (Mitzenmacher & Vassilvitskii, 2022) studies sorting and selection when a machine-learned predictor provides approximate item values; our setting is analogous, with the calibrated direction $\widehat{w}$ serving as the predictor and interpolation sort exploiting its accuracy. We connect these classical algorithms to modern generative latent spaces.

Noisy comparison sort and ranking. The lower bound for sorting from $\mathrm{BSC}(p)$ oracle queries is $\Theta(n \log n/(1 - H(p)))$ (Braverman & Mossel, 2008). Item-level methods (Copeland counting, Bradley-Terry MLE, spectral ranking (Shah & Wainwright, 2018)) are minimax-optimal for several ranking objectives. Active methods such as "Just Sort It!" (Maystre & Grossglauser, 2017) use adaptive comparison scheduling under BT models; EZ-Sort (Park et al., 2025) similarly replaces obvious comparisons with cheap CLIP-based predictions. Our approach circumvents the $\Omega(n \log n)$ lower bound by exploiting latent structure, trading adaptive oracle queries for a one-time calibration step.

RLHF and preference learning. Bradley-Terry models underlie most RLHF reward learning (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022). Our calibration is a feature-based Bradley-Terry model restricted to linear functions over the latent space, exploiting embedding structure for sample efficiency. Preference Ranking Optimization (Song et al., 2024) and listwise extensions (Cao et al., 2007) generalize to richer feature maps and Plackett-Luce likelihoods, collapsing repeated pairwise costs via amortized scorers; our contribution adds distributional sorting for key comparison savings. We include a RankNet baseline (pairwise neural ranker with identical architecture) to isolate the contribution of interpolation sort from the learned scoring function.

Process reward models. PRM800K (Lightman et al., 2023) demonstrated that step-level human labels improve mathematical reasoning verification. We provide a complementary angle: given such labels, how efficiently can we rank reasoning chains? The stratified pooling result offers a practical recipe for best-of-n decoding: cluster problems by embedding similarity, calibrate within each cluster using $\sim$500 oracle pairs once, then sort each new batch with zero oracle queries and $\sim$450 key comparisons.

7. Discussion

Why mathematical reasoning is well-suited. Mathematical solution quality has natural ordinal structure: step correctness is binary, and chain quality aggregates these signals.

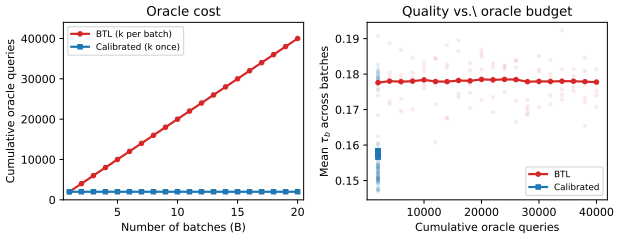

Figure 3. Amortization on PRM800K stratified pool ($k = 2000$, $n = 1000$, 5 seeds). Left: cumulative oracle queries grow linearly for BTL but remain constant for calibrated sort. Right: calibrated sort ($\tau_b = 0.158$) matches or beats BTL (0.178) at fixed oracle cost.

Empirically, this structure manifests as lower effective dimensionality ($d_{\text{eff}} = 163$ vs. 235 for UltraFeedback) and an earlier calibration crossover ($k \approx 500$ vs. 1000). By Proposition 2.2, calibration sample complexity scales as $O(d)$; since PRM800K concentrates 95% of variance in $d_{\text{eff}} = 163$ dimensions (vs. 235 for UltraFeedback), the effective calibration problem has fewer parameters, consistent with the earlier crossover. Whether this reflects mathematical content concentrating quality in fewer latent dimensions, or simply the quality metric being more structured, is an open question.

Embedding structure of mathematical problems. Our most actionable finding for the mathematical AI community is that topic labels are poor proxies for problem similarity. "Intermediate algebra" spans problems with no shared solution structure. K-means clustering on sentence embeddings, even from a general-purpose encoder, groups problems by actual solution strategy (e.g., chains involving similar algebraic manipulations), achieving $\hat{p} = 0.11$ versus 0.30 for topic filtering. This has implications beyond sorting: any method that conditions on problem type (curriculum learning, difficulty estimation, retrieval-augmented generation) may benefit from embedding-based grouping over metadata labels.

When to use calibrated sort vs. BTL. BTL MLE achieves the lowest single-batch τ_b in every condition, so why use calibrated sort? The answer is amortization: BTL requires k oracle queries per batch, while calibrated sort pays k once and ranks each new batch with zero oracle queries. For iterative RLHF with $B \gg 1$ batches under a stable quality criterion, total cost is $k + B \cdot O(n \log\log n)$ vs. $Bk + B \cdot O(n \log n)$. We note that the $O(n \log\log n)$ vs. $O(n \log n)$ key comparison savings is asymptotically meaningful but practically modest for typical n (e.g., $\sim$450 vs. $\sim$10,000 key comparisons at $n = 1000$); the dominant practical benefit is eliminating per-batch oracle queries entirely. Figure 3 confirms this empirically on PRM800K: calibrated sort achieves $\tau_b = 0.158$ (vs. BTL's 0.178) while using $20\times$ fewer oracle queries after $B = 20$ batches.

Practical recipe for best-of-n decoding. For practitioners ranking mathematical reasoning chains: *(1) Embed* all candidate chains with a general-purpose sentence encoder (e.g., all-MiniLM-L6-v2). *(2) Cluster* problems by embedding similarity ($K = 50$ k-means on prompt-level means); verify the largest cluster exceeds your batch size n. *(3) Calibrate once* per cluster: collect $k \approx 500$–2000 pairwise preference labels from the oracle and fit logistic regression on embedding differences (Algorithm 1, Phase 1). *(4) Sort each new batch* by projecting onto $\widehat{w}$ and running interpolation sort: zero oracle queries, $\sim 0.45n$ key comparisons (Phase 2). At $k = 2000$, this achieves $\tau_b < 0.16$ on PRM800K and < 0.14 on UltraFeedback, competitive with BTL at $20\times$ lower cumulative oracle cost after 20 batches.

Best-of-n selection quality. Ranking quality (τ_b) measures global ordering accuracy, but best-of-n decoding only requires identifying the top item. On PRM800K's stratified pool ($n = 828$, 20 seeds), calibrated MLP achieves zero top-1 regret (perfectly selects a quality-1.0 chain) for $k \geq 1000$, even though τ_b is still 0.21. BTL achieves the same at $k \geq 2000$. This confirms that moderate τ_b is sufficient for practical best-of-n pipelines: the top of the ranking converges well before the full ordering does.

Limitations. Our theory assumes linear quality; the MLP's modest improvement ($\hat{p} = 0.09$–0.11 vs. linear's 0.10–0.13) suggests near-linearity holds approximately, but extending the calibration bound to learned feature maps is open. We fix $K = 50$ clusters; the key requirement is that cluster size exceeds n for adequate sampling. Figure 4 shows the commensurability–pool-size tradeoff: $K = 50$ achieves the best MLP $\tau_b = 0.157$ and lowest $\hat{p} = 0.16$; fewer clusters ($K = 10$) yield larger but less homogeneous pools ($\hat{p} = 0.27$, $\tau_b = 0.27$), while more clusters ($K = 200$) shrink the pool below n, limiting statistical power. We report oracle query and key comparison counts but not end-to-end wall-clock time; in practice, embedding computation and MLP training dominate runtime over the sorting step itself, and a full wall-clock comparison would depend heavily on the oracle implementation (human vs. reward model). Finally, our experiments use a general-purpose encoder (`all-MiniLM-L6-v2`); math-specific encoders (e.g., Llemma (Azerbayev et al., 2023)) may yield lower $\hat{p}$ by concentrating quality-relevant variation in fewer dimensions, and investigating this sensitivity is a natural next step.

Impact Statement

This work develops oracle-efficient sorting methods for ranking mathematical reasoning chains. The primary impact is reducing annotation cost for evaluating model-generated solutions, which could make mathematical AI evaluation more accessible. We do not foresee negative societal impacts beyond those inherent to RLHF-trained models generally.

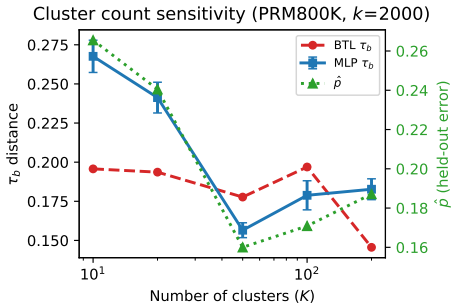

Figure 4. Cluster count sensitivity on PRM800K ($k = 2000$, 20 seeds). $K = 50$ minimizes both $\hat{p}$ and τ_b; too few clusters yield heterogeneous pools, too many shrink pool size below n.

References

Z. Azerbayev, H. Schoelkopf, K. Paster, et al. Llemma: An open language model for mathematics. *arXiv:2310.10631*, 2023.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

M. Braverman and E. Mossel. Noisy sorting without resampling. In *SODA*, pp. 268–276, 2008.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, pp. 4299–4307, 2017.

G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. UltraFeedback: Boosting language models with scaled AI feedback. *arXiv:2310.01377*, 2023.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. *arXiv:2305.20050*, 2023.

Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pp. 129–136, 2007.

M. Mitzenmacher and S. Vassilvitskii. Algorithms with predictions. In *Beyond the Worst-Case Analysis of Algorithms*, pp. 646–662. Cambridge University Press, 2022.

L. Maystre and M. Grossglauser. Just sort it! A simple and effective approach to active preference learning. In *ICML*, pp. 2344–2353, 2017.

K. Mehlhorn and A. Tsakalidis. Data structures. In *Handbook of Theoretical Computer Science*, Vol. A, pp. 301–342, 1990.

Y. Park, H. Chung, and I. Jang. EZ-Sort: Efficient pairwise comparison via zero-shot CLIP-based pre-ordering and human-in-the-loop sorting. In *CIKM*, 2025.

L. Ouyang, J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

F. Song, B. Yu, M. Li, H. Yu, F. Huang, Y. Li, and H. Wang. Preference ranking optimization for human alignment. In *AAAI*, pp. 18990–18998, 2024.

N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *JMLR*, 18(199):1–38, 2018.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.