RoPoLL: Robust Panel of LLM Judges

Anish Acharya [1] **Kris W. Pan** [1] **Brian Verkhovsky** [1]

Abstract

The LLM Jury, a *Panel of LLM Evaluators* (PoLL) (Verga et al., 2024) reporting consensus scores, has become a practical alternative to single judge LLM evaluation, yet its statistical behavior remains poorly understood. Formalizing the setup under the Huber contamination model, we show that PoLL incurs unbounded bias under any positive contamination, regardless of jury size, whenever a single judge fails in a biased, LLM-typical way (mode collapse, sycophancy, safety refusal). We frame jury consensus as classical robust mean estimation and propose RoPoLL (**Ro**bust **P**anel **o**f **LL**M-as-Judge), which preserves the PoLL panel and substitutes the aggregation function with a robust mean estimator, instantiated with the geometric median (GM): tuning-free, with the optimal finite-sample breakdown point $1/2$. A finite-sample error bound and an information-theoretic minimax lower bound match on the parametric rate $\sigma\sqrt{d/N}$ and differ on the breakdown floor by a factor of $\sqrt{d}$, a statistical–computational gap that polynomial-time RoPoLL pays relative to the intractable Tukey halfspace median. Across 13 open-weight judges (4 B–675 B), three reward-model benchmarks, and four corruption regimes at rates up to 50%, RoPoLL dominates PoLL on every biased corruption type: by $\approx 19\%$ on cross-dimensional attacks at matched compute, and by orders of magnitude on heavy-tailed Byzantine adversaries. A 3-judge RoPoLL committee at 38 B beats Mistral-Large-3 (675 B) by $1.31\times$ on HelpSteer 2 under 30% `bimodal-random` corruption, an $18\times$ parameter advantage with strictly better accuracy. A Noisy-GT control confirms the premium is paid against *biased* contamination, not benign Gaussian imprecision (where PoLL is statistically optimal).

[1]Amazon Web Services. Correspondence to: Anish Acharya <achanish@amazon.com>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.

1. Introduction

Reliable evaluation is the bottleneck in aligning Large Language Models (LLMs). Human evaluation does not scale to modern alignment cycles, so the field has converged on the *LLM-as-a-Judge* paradigm (Zheng et al., 2023), in which an LLM scores outputs along one or more quality attributes; follow-up work trains open judges to match this behavior (Kim et al., 2024) and standardizes rubric-based protocols (Li et al., 2023; Dubois et al., 2024; Ye et al., 2024). A single judge, however, is a single point of statistical failure. Its backbone's biases (position, verbosity, self-enhancement, sycophancy, refusal artifacts) are well documented (Wang et al., 2023; Panickssery et al., 2024; Saito et al., 2023; Stureborg et al., 2024), propagate uncorrected to every score, and fix the cost-quality profile to that of one model.

A natural remedy is to evaluate by committee. The *Panel of LLM Evaluators* (PoLL) of Verga et al. (2024) ensembles smaller, diverse, cheaper backbones and reports the arithmetic mean as the consensus, matching or exceeding a single large judge. Related multi-model evaluators (peer-rank discussion (Li et al., 2024), multi-agent debate (Chan et al., 2024), and deeper/wider judge networks (Zhang et al., 2024)) vary the panel structure but inherit PoLL's aggregation rule. PoLL is optimal precisely when judge errors are light-tailed and centered on the truth, in which case averaging N judges contracts variance at the parametric rate $1/N$ (Proposition 1), the clean-baseline efficiency shown in Figure 10.

The problem: Byzantine failures, not Gaussian noise. Real judges fail as *biased point masses far from the truth*, not symmetric noise: a parser fallback to all-zeros (*mode collapse*), saturation near the maximum (*sycophancy*), a per-axis-plausible yet jointly anomalous vector (*cross-attribute confusion*), or out-of-scale values (*heavy-tailed hallucination*). These are not rare—parser failure alone reaches 33% on the smallest judge and a $3.4\%/0.6\%$ mean on HelpSteer 3/2 (Figure 2). Exactly this regime is what classical robust statistics (Huber, 1964; Tukey, 1960; Minsker, 2015; Lugosi & Mendelson, 2019) and Byzantine-robust optimization (Blanchard et al., 2017; Yin et al., 2018; El Mhamdi et al., 2018; Acharya et al., 2022; 2025) identify as adversarial for mean-style aggregation: under the Huber model (Assumption 1), the four modes are instan-

Algorithm 1 RoPoLL

Require: Jury scores $\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_N \in \mathbb{R}^d$; tolerance $\epsilon > 0$; stability $\eta > 0$

$\mathbf{z}^{(0)} \leftarrow \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{y}}_i$
for $t = 0, 1, 2, \dots$ **do**
  $w_i^{(t)} \leftarrow 1/\max(\|\mathbf{z}^{(t)} - \hat{\mathbf{y}}_i\|_2, \eta)$ for each i
  $\mathbf{z}^{(t+1)} \leftarrow \sum_i w_i^{(t)} \hat{\mathbf{y}}_i \,/\, \sum_i w_i^{(t)}$
  if $\|\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}\|_2 < \epsilon$ **then**
    break
  end if
end for
return $\hat{\mathbf{y}}_{\mathrm{GM}} \leftarrow \mathbf{z}^{(t+1)}$

tiations of Q_i (`zeros`, `inverted`, `bimodal-random`, `cauchy-far`), and Proposition 2 shows PoLL's bias is unbounded under *any* positive contamination, regardless of N.

Overview of our approach. We propose RoPoLL (**Ro**bust **P**anel **o**f **LL**M-as-Judge), a drop-in replacement for the arithmetic-mean step of PoLL with a robust mean estimator. We instantiate it with the geometric median (GM): alone among the classical candidates it is simultaneously *tuning-free*, *joint-distance preserving* (unlike the coordinate-wise median, which misses the cross-attribute structure of Example 1), and optimal at the $1/2$ breakdown point (§3.1), and it is computed via the modified Weiszfeld iteration (Algorithm 1, §3.2) at $O(Nd\log(1/\epsilon))$ per query.

Contributions.

- **Formalisation.** We give the first formal treatment of LLM jury aggregation as a robust mean-estimation problem (§2): we model the pipeline as a Markov kernel (Definition 2), define the LLM Jury (Definition 3), and characterize judge failures as Byzantine faults under the Huber model (Assumption 1). Proposition 2 shows PoLL admits unbounded bias under this model.
- **Algorithm and theory.** We propose RoPoLL (§3) and prove (§4) a finite-sample upper bound $\|\hat{\mathbf{y}}_{\mathrm{GM}} - \mathbf{y}^\star\|_2 \le C_{\alpha+\beta}\rho$ (Theorem 1), a correlated-jury extension (Lemma 3), and a matching minimax lower bound (Theorem 2) that agrees on the rate $\sigma\sqrt{d/N}$ and exposes a $\sqrt{d}$ statistical–computational gap, the price of GM's tractability over the intractable Tukey median.
- **Large-scale empirical validation.** Across 13 open-weight judges (4 B–675 B) on three benchmarks (§5), RoPoLL beats PoLL by up to three orders of magnitude on heavy-tailed and cross-dimensional attacks, and a 38 B committee beats Mistral-Large-3 (675 B) by $1.31\times$ under 30% `bimodal-random` corruption ($18\times$ fewer parameters); a Noisy-GT control confirms the premium is paid against *biased* contamination, not imprecision.
- **Open release of the judge-output corpus.**[1] We release the full 13-judge $\times$ three-benchmark corpus ($\sim$ 28K scored (judge, sample) cells: raw text, parsed scores, reference labels, rubric, parser, and corruption pipeline; §D.1)—to our knowledge the first standardized corpus of LLM-jury outputs.

A detailed related-work discussion is deferred to Appendix A; full proofs and the synthetic gallery are in Appendices B–C.2.

2. Problem Setup

We evaluate a *system agent* $\mathcal{M} : \mathcal{P} \to \mathcal{R}$ mapping prompts in $\mathcal{P}$ to responses in $\mathcal{R}$; for each instance $x = (p, \mathcal{M}(p)) \in \mathcal{X} \triangleq \mathcal{P} \times \mathcal{R}$ the goal is to estimate a vector of d attribute scores rating how good the response is for the prompt, conditional on the realized x.

Definition 1 (Reward and latent functional)**.** *We specialize throughout to the homogeneous bounded-scalar reward space $\mathcal{Y} = [0, K]^d \subset \mathbb{R}^d$, on which a measurable* latent reward functional $\mathbf{y}^\star : \mathcal{X} \to [0, K]^d$ *gives the canonical, unobservable attribute-wise assessment of r for p.*

Because $\mathbf{y}^\star$ is unobservable, evaluation proceeds through an observable *reference protocol* $\mathcal{A}$, a Markov kernel $\mathcal{X} \rightsquigarrow [0, K]^d$ (Billingsley, 1995; Dudley, 2002; Kallenberg, 2002) yielding the benchmark dataset $\mathcal{D} = \{(x_j, \mathbf{y}_j^{\mathrm{ref}})\}_{j=1}^{M}$ with $\mathbf{y}_j^{\mathrm{ref}} \sim \mathcal{A}(\cdot \mid x_j)$ (noiseless idealization $\mathcal{A}(\cdot \mid x) = \delta_{\mathbf{y}^\star(x)}$).

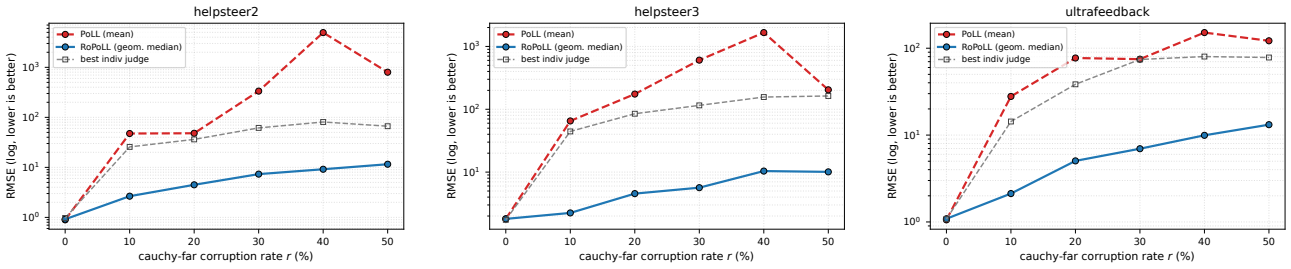
Definition 2 (Rubric and LLM-as-Judge)**.** *A* rubric ρ *fixes the d attributes, the score range $[0, K]$, and the output schema, together with a deterministic encoder* $\mathrm{enc}_\rho : \mathcal{X} \to \mathcal{P}$. *An* LLM judge *is a triplet $f = (\mathcal{M}_f, \rho, \phi)$ with backbone kernel $\mathcal{M}_f : \mathcal{P} \rightsquigarrow \mathcal{R}$ and deterministic parser $\phi : \mathcal{R} \to \mathbb{R}^d$, inducing the pipeline*

$$x \xrightarrow{\mathrm{enc}_\rho} \mathrm{enc}_\rho(x) \xrightarrow{\mathcal{M}_f} T_f \xrightarrow{\phi} \hat{\mathbf{y}}_f(x) = \phi(T_f) \in \mathbb{R}^d. \quad (1)$$

Remark 1 (Parser-induced atoms and operational stochasticity)**.** *The parser ϕ is part of the estimator: it maps malformed outputs, refusals, or missing fields to a fixed fallback such as* $\mathbf{0}$*, so $\hat{\mathbf{y}}_f(x)$ can carry point masses even when T_f is diffuse (itself non-degenerate at temperature* 0 *due to inference-stack non-determinism). This is the formal counterpart of the* mode collapse *failure mode (§1) and is how backbone-level noise becomes parser-level Huber contamination.*

Definition 3 (LLM jury and aggregation)**.** *A* jury *is N judges $\mathcal{J} = \{f_1, \dots, f_N\}$ sharing ρ and ϕ but with distinct backbones, producing score vectors $\{\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_N\} \subset \mathbb{R}^d$ on instance x. An* aggregation function $\mathcal{A} : (\mathbb{R}^d)^N \to \mathbb{R}^d$ *returns a consensus $\hat{\mathbf{y}}_{\mathrm{agg}}$; the objective is to minimize $\|\hat{\mathbf{y}}_{\mathrm{agg}} - \mathbf{y}^\star\|_2$.*

[1] Released Dataset: https://github.com/aws/RoPoLL



(a) HelpSteer 2 *(b)* HelpSteer 3 *(c)* UltraFeedback

Figure 1. **PoLL vs. RoPoLL under heavy-tailed `cauchy-far` corruption.** RMSE vs. per-case rate r (log y) for the MEDIUM jury (N=3), best single judge as gray dashed reference. PoLL diverges by one to three orders of magnitude while RoPoLL stays bounded, instantiating Proposition 2 (coordinate-wise MEDIAN, competitive here, in Figure 11).

The central question is *which $\mathcal{A}$ stays accurate when judges fail in arbitrary, possibly adversarial ways*, which we formalize via the classical contamination model of robust statistics.

Assumption 1 (Huber ϵ-Contamination Model). *Each judge $f_i \in \mathcal{J}$ has a contamination rate $\alpha_i \in [0,1)$, and the conditional law of $\hat{\mathbf{y}}_i$ given $\mathbf{y}^\star$ is the mixture $\hat{\mathbf{y}}_i \sim (1-\alpha_i)P_i + \alpha_i Q_i$, where the* competent *component P_i is unbiased ($\mathbb{E}_{P_i}[\hat{\mathbf{y}}_i] = \mathbf{y}^\star$, finite second moment) and the* corruption *component Q_i is arbitrary on $\mathbb{R}^d$; an indicator $Z_i \sim \text{Bernoulli}(\alpha_i)$ selects which component generates $\hat{\mathbf{y}}_i$.*

Instantiations and empirical grounding. The unrestricted Q_i captures every LLM-judge failure mode in the single-judge bias literature: mode collapse ($\delta_{\mathbf{0}}$), sycophancy ($\delta_{K\mathbf{1}}$, Wang et al., 2023; Stureborg et al., 2024), anti-correlated Byzantine attacks ($\delta_{K\mathbf{1}-\mathbf{y}^\star}$), cross-attribute confusion ($\text{Unif}\{0,K\}^d$, the cross-dimensional mode of Example 1), and heavy-tailed adversaries (Cauchy). These are the four synthetic regimes (`zeros`, `inverted`, `bimodal-random`, `cauchy-far`) evaluated in §5. Such failures are not hypothetical: across our 13-judge grid (Figure 2), naturally-occurring parser-failure rates span 0.59% on HelpSteer 2 to 3.38% on HelpSteer 3 (up to 33% for the smallest judge), so α is dataset-dependent across one to two orders of magnitude and the distribution-free class $\{Q_i\}$ is the right object of study.

Assumption 2 (Conditional Independence)**.** *Conditioned on $\mathbf{y}^\star$, the judge outputs $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N$ are mutually independent.*

Remark 2 (i.i.d. as baseline; correlated extension)**.** *Assumption 2 is the standard i.i.d. baseline of robust statistics. Real LLM juries trained on overlapping corpora violate this: inter-judge correlation $\bar{\gamma} \in [0.3, 0.7]$ is typical (Figure 21 and Section D.2). Lemma 3 (§4.1) extends Theorem 1 to the equicorrelated case and shows that the breakdown structure ($C_{\alpha+\beta}$ and ρ) is unchanged; only the high-probability event weakens, from $1-\exp(-N\beta^2/2)$ (Hoeffding under independence) to $1-1/(\beta^2 N_{\text{eff}})$ (Chebyshev under correlation), with $N_{\text{eff}} = N/(1+(N-1)\bar{\gamma}_W)$.*

Assumption 3 (Sub-Gaussian Competent Noise)**.** *For each judge f_i, the competent component P_i is σ_i^2-sub-Gaussian, i.e. $\mathbb{E}_{P_i}[\exp(\lambda \mathbf{u}^\top(\hat{\mathbf{y}}_i - \mathbf{y}^\star))] \le \exp(\lambda^2\sigma_i^2/2)$ for all $\mathbf{u} \in \mathbb{S}^{d-1}, \lambda \in \mathbb{R}$; σ_i^2 is the per-judge* skill*. This is non-restrictive: any distribution on a bounded set is sub-Gaussian, and scores live in $[0,K]^d$.*

Assumption 4 (Minority Corruption). *The* effective contamination fraction $\alpha \triangleq \frac{1}{N}\sum_{i=1}^N \alpha_i < 1/2$.

The threshold $\alpha < 1/2$ is information-theoretically tight: a corrupted majority can simulate any target law, rendering $\mathbf{y}^\star$ unidentifiable.

Collecting the assumptions above, the complete observation model for a single evaluation instance is:

$$\hat{\mathbf{y}}_i \;=\; (1-Z_i)\,(\mathbf{y}^\star + \boldsymbol{\epsilon}_i) \;+\; Z_i\,\boldsymbol{\eta}_i,\ \forall i = 1,\ldots,N \quad (2)$$

where $Z_i \sim \text{Bernoulli}(\alpha_i)$ are independent latent corruption indicators, $\boldsymbol{\epsilon}_i \sim P_i - \mathbf{y}^\star$ is zero-mean and σ_i^2-sub-Gaussian (Assumption 3), and $\boldsymbol{\eta}_i \sim Q_i$ is the arbitrary corruption noise, independent of $\boldsymbol{\epsilon}_i$ and Z_i. The statistician observes only $\{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N\}$ and has no access to $\{Z_i\}$ or $\{Q_i\}$. The canonical jury aggregator is the arithmetic mean adopted by PoLL (Verga et al., 2024), $\hat{\mathbf{y}}_{\text{mean}} \triangleq \frac{1}{N}\sum_i \hat{\mathbf{y}}_i$, which on clean juries enjoys the parametric variance-reduction rate.

Proposition 1 (Variance Reduction for the Clean Jury). *If $\alpha_i = 0$ for all i (every judge competent), then $\hat{\mathbf{y}}_{\text{mean}}$ is unbiased, and under Assumption 2 with $\boldsymbol{\Sigma}_i \preceq \sigma^2 \mathbf{I}_d$, $\mathbb{E}[\|\hat{\mathbf{y}}_{\text{mean}} - \mathbf{y}^\star\|_2^2 \mid \mathbf{y}^\star] = \frac{1}{N^2}\sum_i \text{tr}(\boldsymbol{\Sigma}_i) \le d\sigma^2/N$ (proof in Appendix B.1).*

Corollary 1 (Effective Jury Size Under Correlation)**.** *If additionally $\text{Cov}(\hat{\mathbf{y}}_i \mid \mathbf{y}^\star) = \boldsymbol{\Sigma}$ and $\text{Cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \mid \mathbf{y}^\star) = \gamma\,\boldsymbol{\Sigma}$ $(i \ne j)$, then the MSE rate is $\text{tr}(\boldsymbol{\Sigma})/N_{\text{eff}}$ with the* effective jury size $N_{\text{eff}} \triangleq N/(1+(N-1)\gamma)$.

The design implication: for any $\gamma > 0$ the effective jury size N_{eff} saturates at $1/\gamma$, so for the moderate $\gamma \in [0.3, 0.7]$ measured across our diverse LLM backbones it saturates

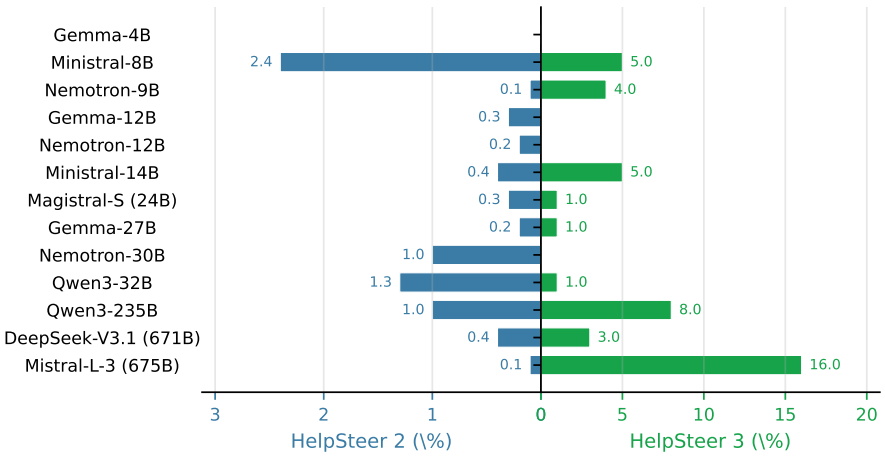


Figure 2. **Naturally-occurring parser-failure rates motivate the contamination model.** Horizontal bars per judge (sorted by parameter count, top = smallest) restricted to the 13-judge pool common to both benchmarks (Claude-Opus/Sonnet/Haiku-4.5 are HS 3-only and excluded here for panel alignment; their HS 3 statistics appear in Table 2). The natural failure regime is *dataset-dependent*: 0.59% mean on HelpSteer 2 and 3.38% mean on HelpSteer 3—with the smallest judge (Gemma-4B) failing on 33% of HS 3 multilingual signed-preference samples (full 16-judge pool). Each parser-failure event is a Dirac mass at the fallback vector $\mathbf{0}$, instantiating $Q = \delta_{\mathbf{0}}$ in Assumption 1 (mode collapse). Naturally-occurring rates already span 0% to 33%, motivating the synthetic sweep $r \in [0\%, 50\%]$ studied in §5, which covers this natural regime and stress-tests beyond.

at $N \approx 2$–3, motivating the three-judge committees of §5 (synthetic validation in §4.1).

The next result shows that the $1/N$ variance-reduction rate of Proposition 1 is irrelevant the moment any contamination is present.

Proposition 2 (Unbounded Bias of PoLL). *Under Assumption 1, suppose each Q_i has finite first moment $\mu_i^Q \triangleq \mathbb{E}_{Q_i}[\hat{\mathbf{y}}_i] \in \mathbb{R}^d$. Then*

$$\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] \;=\; \mathbf{y}^\star \;+\; \frac{1}{N}\sum_{i=1}^{N} \alpha_i \left(\mu_i^Q - \mathbf{y}^\star\right), \qquad (3)$$

and for any $\alpha > 0$ the conditional bias $\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] - \mathbf{y}^\star$ cannot be uniformly bounded under Assumption 1, regardless of N.

Placing a single corrupted mean $\mu_{i_0}^Q$ at distance $\propto N$ makes (3) arbitrarily large, exactly cancelling the $1/N$ averaging, so a larger jury cannot help (proof in Appendix B.2). This is the **central impossibility** motivating RoPoLL: we seek an aggregator that both matches the mean's $O(\sigma\sqrt{d/N})$ rate when clean and stays bounded under arbitrary contamination with $\alpha < 1/2$. The geometric median (§3) achieves both.

3. Robust Panel of LLM Judges

Proposition 2 forces us to abandon the arithmetic mean: under contamination its bias is unbounded over the corruption class regardless of jury size N. RoPoLL is a drop-in replacement for the PoLL aggregation step that keeps the panel but swaps the mean for a robust mean estimator; we now choose that estimator.

3.1. Choosing the Robust Estimator

Three classical robust mean estimators are natural candidates. The *trimmed mean* discards the β-fraction of points farthest from the current estimate and averages the rest; the *coordinate-wise median* (CoMed) takes the univariate median per attribute, $\hat{\mathbf{y}}_{\text{Med}} = \arg\min_{\mathbf{z}} \sum_i \|\mathbf{z} - \hat{\mathbf{y}}_i\|_1$; and the *geometric median* (GM) minimizes the same objective in the joint ℓ_2 norm, $\hat{\mathbf{y}}_{\text{GM}} = \arg\min_{\mathbf{z}} \sum_i \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2$ (Definition 4). The trimmed mean needs the contamination rate: $\beta < \alpha$ readmits the bias of Proposition 2 while $\beta > \alpha$ inflates variance. CoMed is tuning-free but decouples the coordinates, which is exactly the wrong move when corruption is structured across attributes.

Example 1 (Cross-Dimensional Corruption)**.** *Consider a jury evaluating on two attributes with ground truth $\mathbf{y}^\star = (2.5, 2.5)$ on $[0,5]^2$, and a corrupted judge that outputs $\hat{\mathbf{y}}_{\text{corr}} = (0, 5)$. Each coordinate individually lies in the plausible range $[0,5]$, so CoMed treats it as unremarkable per axis; jointly, however, the displacement $\|\hat{\mathbf{y}}_{\text{corr}} - \mathbf{y}^\star\|_2 = \sqrt{12.5} \approx 3.54$ is large, and the geometric median downweights it (Figure 3).*

This generalizes to the whole class of per-coordinate-plausible but jointly anomalous corruptions, whose canonical instance is the `bimodal-random` class (§5): with each corrupted coordinate drawn from $\{0, K\}$, the marginal $\frac{1}{2}(\delta_0 + \delta_K)$ looks plausible while the joint vector sits at a

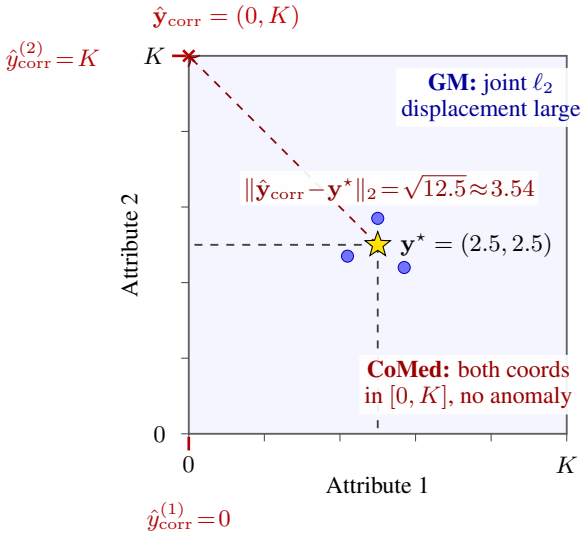


Figure 3. **Cross-dimensional corruption (Example 1).** Three competent judges (blue dots) cluster around the truth $\mathbf{y}^\star = (2.5, 2.5)$ in the score box $[0, K]^2$ with $K = 5$. A corrupted judge outputs $\hat{\mathbf{y}}_{\text{corr}} = (0, K)$: *each coordinate individually* lies in the plausible range $[0, K]$ (red axis ticks), so any *coordinate-wise* estimator sees nothing anomalous on either axis. Jointly, however, the corrupted vector lies at ℓ_2 distance $\sqrt{12.5} \approx 3.54$ from $\mathbf{y}^\star$ (red dashed arrow), and the geometric median's joint-distance objective downweights it. This is the qualitative reason RoPoLL uses GM rather than CoMed; the empirical analogue at scale is the `bimodal-random` sweep of §5.

random hypercube corner, so any coordinate-wise estimator incurs $\Omega(\alpha)$ per-coordinate bias (Huber, 1964, Thm. 5.1) composing in ℓ_2 with a $\sqrt{d}$ factor. The geometric median avoids both failure modes: it is tuning-free, attains the optimal $1/2$ breakdown point, and acts on *joint* ℓ_2 distance, so we instantiate RoPoLL with it (empirical comparisons against CoMed and the trimmed mean are in §5). At the small $N \leq 5$ of LLM panels plain GM is the right default: geometric median-of-means (Lugosi & Mendelson, 2019; Hopkins, 2020) would improve the large-N breakdown floor but degenerates to plain GM at block size 1.

3.2. The Geometric Median: Definition and Properties

Definition 4 (RoPoLL via Geometric Median)**.** *Given jury outputs $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N \in \mathbb{R}^d$, the* RoPoLL *estimate of $\mathbf{y}^\star$ is*

$$\hat{\mathbf{y}}_{\text{GM}} \triangleq \arg\min_{\mathbf{z} \in \mathbb{R}^d} \sum_{i=1}^{N} \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2. \tag{4}$$

Its modern robustness analysis is due to Lopuhaä & Rousseeuw (1991); Small (1990); Vardi & Zhang (2000); Acharya et al. (2022; 2025); we collect the properties we use:

Definition 5 (Finite-Sample Breakdown Point)**.** *For an estimator $T : (\mathbb{R}^d)^N \to \mathbb{R}^d$ and a sample $\hat{\mathbf{y}}_{1:N} \in (\mathbb{R}^d)^N$, the* finite-sample breakdown point *of T at $\hat{\mathbf{y}}_{1:N}$ is the smallest fraction m/N such that there exists a corrupted sample $\hat{\mathbf{y}}'_{1:N}$ differing from $\hat{\mathbf{y}}_{1:N}$ in at most m coordinates for which $\|T(\hat{\mathbf{y}}'_{1:N}) - T(\hat{\mathbf{y}}_{1:N})\|_2$ can be made arbitrarily large (Lopuhaä & Rousseeuw, 1991).*

Proposition 3 (Properties of the Geometric Median)**.** *Let $F(\mathbf{z}) = \sum_{i=1}^{N} \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2$. A minimizer exists ($F$ convex and coercive) and is unique when the $\hat{\mathbf{y}}_i$ are not collinear (F strictly convex); the geometric median is affine equivariant, $\text{GM}(\mathbf{U}\hat{\mathbf{y}}_i + \mathbf{b}) = \mathbf{U}\,\text{GM}(\hat{\mathbf{y}}_i) + \mathbf{b}$ for orthogonal $\mathbf{U}$ and translation $\mathbf{b}$; and its finite-sample breakdown point is $\epsilon^\star = \lceil N/2 \rceil / N \to 1/2$, optimal among translation-equivariant estimators (Lopuhaä & Rousseeuw, 1991).*

(Proof in Appendix B.3.)

Remark 3 (Discrete breakdown threshold at small N)**.** *At $N = 3$ the integer breakdown cutoff is "one corrupted of three": two corrupted judges break the geometric median regardless of α. The sweep to per-cell rate $r = 50\%$ straddles this cutoff (in expectation 1.5 of 3 judges corrupted), which is why the corruption-class dependence in Figure 11 appears: beyond-breakdown cells still average out under symmetric Q_i (`zeros`, `inverted`) but not under biased Q_i (`bimodal-random`, `cauchy-far`).*

The Weiszfeld Iteration. The geometric median has no closed form for $d \geq 2$ (Bajaj, 1988); setting $\nabla F(\mathbf{z}) = \mathbf{0}$ gives the Weiszfeld fixed-point (Eq. 5), a reweighted mean that downweights distant points, with a modified weight $1/\max(\|\mathbf{z} - \hat{\mathbf{y}}_i\|_2, \eta)$ handling the data-point singularity (Weiszfeld, 1937; Vardi & Zhang, 2000), yielding Algorithm 1.

$$\mathbf{z} = \frac{\sum_{i=1}^{N} \hat{\mathbf{y}}_i / \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2}{\sum_{i=1}^{N} 1/\|\mathbf{z} - \hat{\mathbf{y}}_i\|_2}. \tag{5}$$

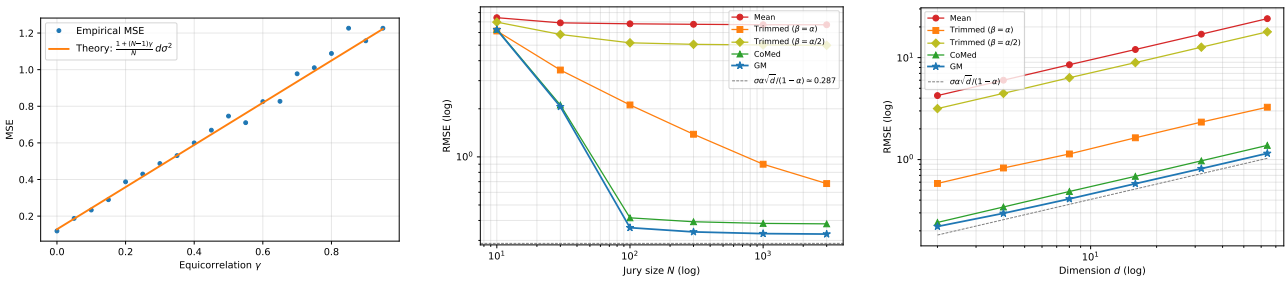
It converges linearly for non-collinear data, so tolerance ϵ costs $O(Nd \log(1/\epsilon))$, microseconds for a typical jury and negligible against per-judge inference (Appendix B.4).

4. Theoretical Guarantees

4.1. Finite-Sample Error Bound

We bound $\|\hat{\mathbf{y}}_{\text{GM}} - \mathbf{y}^\star\|_2$ under our Huber model in two steps: a deterministic geometric lemma about the geometric median (Lemma 1, due to Minsker (2015)) and a probabilistic lemma controlling the sub-Gaussian cluster radius (Lemma 2). Combining them yields Theorem 1. Full proofs are in Appendix B.

Lemma 1 (Geometric Breakdown of GM). (Minsker (2015), Lemma 2.1; building on Lopuhaä & Rousseeuw (1991).) *Let $x_1, \ldots, x_k \in \mathbb{R}^d$ and let x_* be any minimizer of $z \mapsto \sum_{j=1}^{k} \|z - x_j\|_2$ (a geometric median). Fix $\alpha \in (0, 1/2)$, $r > 0$, and $z \in \mathbb{R}^d$. If $|\{j : \|x_j - z\|_2 \leq$*



(a) Corollary 1. *(b)* Theorem 1. *(c)* Lemma 2.

Figure 4. **Theory validation. (a)** Clean-jury MSE matches the closed form $\frac{1+(N-1)\gamma}{N} d\sigma^2$ of Corollary 1, with N_{eff} saturating at $1/\gamma$. **(b)** Under worst-case Huber contamination the geometric median converges to the predicted breakdown floor $\sigma\alpha\sqrt{d}/(1-\alpha)$ (Theorem 1) while the mean plateaus above it (Proposition 2). **(c)** The cluster radius tracks the predicted $\sqrt{d}$ scaling (Lemma 2).

$r\}| \geq (1-\alpha)k$, *then*

$$\left\|x_* - z\right\|_2 \;\leq\; C_\alpha\, r, \qquad C_\alpha \;\triangleq\; \frac{1-\alpha}{\sqrt{1-2\alpha}}. \tag{6}$$

This deterministic bound relates the geometric median to any target z via how concentrated the inputs are around z; the constant C_α is sharp and diverges as $\alpha \to 1/2$, matching the breakdown point of GM (Figure 7).

Lemma 2 (Sub-Gaussian Cluster Radius). *Under Assumptions 1–4, let $\beta \in (0,\, 1/2-\alpha)$ be a slack parameter. With probability at least $1-\exp(-N\beta^2/2)$,*

$$\left|\left\{i \in [N] : \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq \rho\right\}\right| \;\geq\; (1-\alpha-\beta)\,N, \tag{7}$$

where the cluster radius is

$$\rho \;=\; \sigma\left(C_1\sqrt{d} \;+\; \sqrt{\frac{1}{c}\,\log\frac{2(1-\alpha)}{\beta}}\right), \tag{8}$$

and $C_1, c > 0$ are absolute constants (from the sub-Gaussian-norm tail bound, Step 1 of the proof in Appendix B.6).

The slack β trades off cluster radius against contamination threshold: larger β permits smaller ρ but raises the effective threshold from α to $\alpha+\beta$ in the geometric step.

Theorem 1 (RoPoLL Breakdown Bound under Huber Contamination). *Under Assumptions 1–4, fix any slack $\beta \in (0,\, 1/2-\alpha)$. With probability at least $1-\exp(-N\beta^2/2)$,*

$$\left\|\hat{\mathbf{y}}_{\text{GM}} - \mathbf{y}^\star\right\|_2 \;\leq\; \underbrace{\frac{1-\alpha-\beta}{\sqrt{1-2\alpha-2\beta}}}_{C_{\alpha+\beta}} \cdot \underbrace{\sigma\left(C_1\sqrt{d} + \sqrt{\tfrac{1}{c}\log\tfrac{2(1-\alpha)}{\beta}}\right)}_{\rho}. \tag{9}$$

Interpretation. The geometric constant $C_{\alpha+\beta}$ depends only on the contamination rate and diverges as $\alpha+\beta \to 1/2$, while the cluster radius ρ depends only on σ, d and does *not* shrink with N: under arbitrary Q the asymptotic-N floor is set by the cluster radius, distribution-free over $\{Q_i\}$ (Figure 6). Figure 4 validates this on synthetic data: the geometric median converges to the predicted floor $\sigma\alpha\sqrt{d}/(1-\alpha)$ while PoLL stays above it, and the cluster radius scales as $\sqrt{d}$.

Beyond i.i.d. Real juries violate the conditional independence of Assumption 2 ($\bar{\gamma} \in [0.3, 0.7]$, Figure 21 and Section D.2). The next lemma extends the bound to the equicorrelated case: the breakdown structure ($C_{\alpha+\beta}$ and ρ) is unchanged and only the success probability weakens, from exponential in N to polynomial in N_{eff}.

Lemma 3 (RoPoLL under Equicorrelated Juries). *Replace Assumption 2 with the weaker* equicorrelated-indicator *assumption: for the cluster indicators $W_i \triangleq \mathbb{1}\{Z_i = 0,\ \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq \rho_p\}$ of Lemma 2, $\text{Cov}(W_i, W_j) \leq \bar{\gamma}_W \sqrt{\text{Var}(W_i)\text{Var}(W_j)}$ for all $i \neq j$, with $\bar{\gamma}_W \in [0,1]$. Under Assumptions 1, 3, 4 and the equicorrelated-indicator assumption, fix any slack $\beta \in (0, 1/2-\alpha)$. With probability at least*

$$1 - \frac{1}{\beta^2\, N_{\text{eff}}}, \qquad N_{\text{eff}} \;\triangleq\; \frac{N}{1+(N-1)\bar{\gamma}_W}, \tag{10}$$

the bound $\|\hat{\mathbf{y}}_{\text{GM}} - \mathbf{y}^\star\|_2 \leq C_{\alpha+\beta}\,\rho$ of Theorem 1 holds, with $C_{\alpha+\beta}$ and ρ unchanged. At $\bar{\gamma}_W = 0$, the equicorrelated assumption reduces to independence and Theorem 1's exponential bound $\exp(-N\beta^2/2)$ recovers (which is strictly tighter than (10)).

Empirical $\bar{\gamma}_W$. We estimate the indicator correlation $\bar{\gamma}_W$ directly from the co-incidence of cluster events $\{W_i = 1\}$ on our 13-judge panels (Section D.3 and Table 3): $\bar{\gamma}_W \in [0.45, 0.53]$ on HelpSteer 2 and UltraFeedback, in line with the inter-judge score correlations $\bar{\gamma} \in [0.49, 0.71]$ of Figure 21. At N=3 this gives $N_{\text{eff}} \in [1.45, 1.58]$, leaving the

high-probability event non-trivial for the moderate slack β used in practice.

4.2. Minimax Lower Bound

A matching information-theoretic lower bound shows the upper bound is essentially tight.

Theorem 2 (Minimax Lower Bound). *Under the same assumptions as Theorem 1,*

$$\inf_{\hat{\mathbf{y}}} \sup_{F \in \mathcal{F}_{\alpha,\sigma}} \mathbb{E}_F[\|\hat{\mathbf{y}} - \mathbf{y}^\star\|_2] \;\geq\; c\sigma\left(\sqrt{d/N} \;+\; \frac{\alpha}{1-\alpha}\right), \tag{11}$$

where the infimum is over all measurable estimators of the form $\hat{\mathbf{y}}(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N)$, $\mathcal{F}_{\alpha,\sigma}$ is the class of joint distributions consistent with Assumptions 1, 2, 3, and 4, and $c > 0$ is a universal constant.

Comparison. The two bounds match exactly on the parametric rate $\sigma\sqrt{d/N}$ and differ by a $\sqrt{d}$ factor on the breakdown floor ($C_\alpha \sigma\sqrt{d}$ vs. $\sigma\alpha/(1-\alpha)$), a statistical–computational gap: the floor is matched only by the intractable Tukey halfspace median (NP-hard for $d \geq 3$ (Tukey, 1975; Donoho & Gasko, 1992; Johnson & Preparata, 1978; Aloupis, 2006), sub-exponential via smoothed depth (Chen et al., 2018)), whereas the geometric median pays the $\sqrt{d}$ for $O(Nd\log(1/\epsilon))$ tractability. For LLM juries the trade is favorable: $d \leq 5$, so the overhead is $\leq 2.2\times$ and the variance term dominates the floor at $N = 3$. The i.i.d. baseline (identical σ_i, partially relaxed by Lemma 3) leaves per-judge heterogeneity and explicit dependence (Li et al., 2024; Chan et al., 2024; Zhang et al., 2024) to future work.

Proof techniques. Full proofs are in Appendix B. Theorem 1 composes a deterministic breakdown lemma for the geometric median (Lemma 1) with a sub-Gaussian bound on the competent-cluster count (Lemma 2), placing the GM within $C_{\alpha+\beta}\rho$ of $\mathbf{y}^\star$ once a $(1-\alpha-\beta)$ majority concentrates; Lemma 3 keeps this deterministic core and only swaps the Hoeffding step for a Chebyshev bound on $\sum_i W_i$, so ρ and $C_{\alpha+\beta}$ are unchanged and only the success probability weakens; Theorem 2 is a Le Cam two-point argument matching the $\sqrt{d/N}$ variance term (Pinsker) and the $\alpha/(1-\alpha)$ floor (modulus of continuity of the Huber neighborhood) (Figures 6 to 8).

5. Experiments

We evaluate RoPoLL against PoLL (the arithmetic-mean baseline of Verga et al. (2024)) and the coordinate-wise Median on three reward-model benchmarks under a per-case corruption pipeline that exposes the corruption-type dependence predicted by Theorem 1 and Example 1.

5.1. Setup

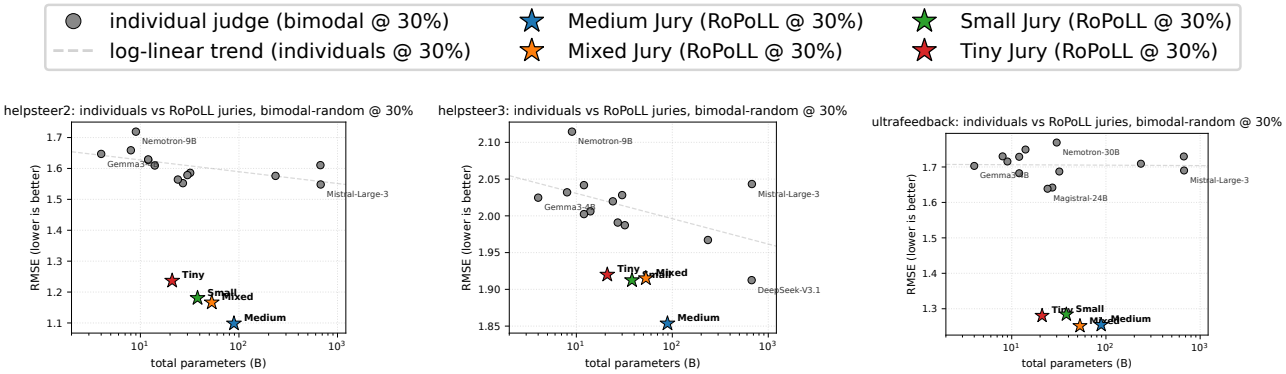
Datasets. We use three reward-model benchmarks with complementary ground truth: **HelpSteer 2** (Wang et al., 2024) (1,000 samples, 0–4 Likert across five human-rated attributes), **HelpSteer 3** (Wang et al., 2025) (100-sample multilingual preference slice, reduced to a signed `overall_preference` scalar on $[-4, 4]$), and **UltraFeedback** (Cui et al., 2024) (1,000 samples, 1–5 across four attributes, GPT-4 reference).

Judges and juries. We score every sample with 13 open-weight judges spanning 4–675 B at temperature 0 under a shared rubric (Mistral-Large-3 675 B and DeepSeek-V3.1 671 B down to Gemma-4B), and curate four three-judge committees trading size against compute: Medium $\approx$ 89 B, Mixed $\approx$ 53 B, Small $\approx$ 38 B, and Tiny $\approx$ 21 B (Table 2 lists the full pool). We fix $N = 3$ as the saturation knee of Corollary 1 for the measured $\gamma \in [0.3, 0.7]$ (corroborated by the ablation in Figure 12), and compare PoLL (mean), the coordinate-wise Median, and RoPoLL (Algorithm 1) on the same score vectors per sample.

Per-case corruption protocol. Holding the jury at three, we corrupt individual (sample, judge) *cells* at per-case rate $r \in \{0\%, 10\%, \ldots, 50\%\}$, the realistic pattern of a judge occasionally emitting a bad score; this range covers and stress-tests the natural failure rates of Figure 2 (0.59% on HS 2 to 33% on HS 3). The four corruption types are `zeros` (**0** parser fallback), `inverted` ($K\mathbf{1} - \mathbf{y}^\star$, worst-case Byzantine), `bimodal-random` (each coordinate set to 0 or K, the cross-dimensional mode of Example 1), and `cauchy-far` ($\mathbf{y}^\star + 10 + 2(s_{\max} - s_{\min})\mathbf{t}$, $\mathbf{t}$ standard Cauchy, a biased heavy-tailed attack). We report RMSE against the reference labels; per-judge calibration breakdowns are in Appendix C.3.

Heavy-Tailed Corruption. The `cauchy-far` attack is the empirical analogue of Proposition 2: each corrupted slot has an unbounded first moment, so one contaminated judge can drag PoLL arbitrarily. Figure 1 confirms this: on the Medium jury PoLL's RMSE exceeds RoPoLL's by one to three orders of magnitude at every $r \geq 10\%$ (peaking at HelpSteer 2, $r = 40\%$: $\approx$ 4,951 vs. $\approx$ 9.2, a $540\times$ ratio). The coordinate-wise Median is competitive here: under heavy-tailed attacks *any* robust aggregator beats the mean, as the theory predicts.

Cross-Dimensional. The `bimodal-random` attack drives each coordinate of a corrupted score to an extremum: plausible per coordinate but jointly anomalous, the failure mode of Example 1. Figure 5 plots RMSE against parameter count for the 13 individual judges and the four committees under identical 30% corruption. On HelpSteer 2 and UltraFeedback all four committees sit below the individual-at-



(a) HelpSteer 2 *(b)* HelpSteer 3 *(c)* UltraFeedback

Figure 5. **Parameter efficiency under `bimodal-random` corruption ($r = 30\%$).** RMSE vs. parameter count (log scale): gray circles are the 13 individual judges (dashed log-linear fit), coloured stars the four RoPoLL juries at their aggregate budget. The committees sit on or below the individual scaling trend; clean and `zeros` counterparts are in Figures 9 and 10.

30% scaling trend; the headline: the SMALL committee at 38 B reaches RMSE 1.18, beating Mistral-Large-3's 1.55 at 675 B (a $1.31\times$ advantage at $18\times$ fewer parameters), and on HelpSteer 3 the MEDIUM committee at 89 B matches DeepSeek-V3.1 (671 B) at RMSE 1.85.

The compute-matched comparison (RoPoLL vs. PoLL on the *same* committee) is even more direct: on SMALL at 30% `bimodal-random`, RoPoLL (RMSE 1.18) beats PoLL (RMSE $\approx$ 1.45) by $\approx 19\%$ at identical inference cost (Figure 11). *Robust aggregation, not the ensemble itself, delivers the win.*

Bounded Mean-Preserving Corruptions. The `zeros` and `inverted` attacks are *mean-preserving* when the corrupted point sits at the scale midpoint—an empirical accident making these the hardest regimes to separate the two methods. Even here (Figure 9, 30% `zeros`) all four committees sit at or below the individual-at-30% line and the gap to PoLL stays small (≤ 0.3 RMSE for MEDIUM): RoPoLL is no worse, and the practitioner cannot choose which regime the next corruption falls into.

Clean-Baseline Parameter Efficiency. Does robustness cost accuracy when there is *no* corruption? Theorem 1 predicts a small premium at $\alpha = 0$, and Figure 10 confirms it: at $r = 0\%$ the MEDIUM/MIXED/SMALL committees sit below the individual scaling line (TINY on-trend), with a clean-case premium of at most $+6.4\%$ RMSE (median $+0.9\%$)—a small fraction of the gains under corruption.

Ablations and Controls. On jury size, RMSE drops sharply from $N = 1$ to $N = 3$ and levels off (a fourth judge falls within the standard-deviation band), confirming the knee predicted by Corollary 1 (Figure 12). On the estimator, the full grid (Figure 11) confirms the corruption-type dependence of Theorem 1: the RoPoLL/PoLL gap is negligible under mean-preserving `zeros`/`inverted` (± 0.3 RMSE), positive at every $r \geq 10\%$ under `bimodal-random`, and one to three orders of magnitude under `cauchy-far`; the coordinate-wise MEDIAN matches RoPoLL except on `bimodal-random` (cross-dimensional structure, Example 1, is invisible to it).

Noisy-GT control. Is the "insurance premium" paid on a phantom—are real judge failures merely imprecise rather than biased? A *Noisy-GT* adversary injecting $\hat{\mathbf{y}}_{\text{noisy}} = \text{clip}(\mathbf{y}^\star + \boldsymbol{\epsilon}, 0, K)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.8^2\mathbf{I})$, in place of the adversarial vectors, makes all three aggregators *improve* as its rate rises (PoLL slightly preferred, since averaging unbiased noise is optimal), confirming that the RoPoLL premium is paid against *biased* contamination, not imprecision.

Practical Recommendation. Use RoPoLL as the default jury aggregator: the clean-case premium is small ($\leq 6\%$) and the real threat is biased contamination, not imprecision. The jury size follows the saturation law $N_{\text{eff}} \to 1/\gamma$ (Corollary 1): for the diverse pools here ($\gamma \approx 0.49$–0.71) the knee is $N \approx 3$, higher for more orthogonal pools. The synthetic gallery and per-model breakdowns are in Appendices C.2 and C.3.

6. Conclusion

We recast LLM-jury aggregation as robust mean estimation: PoLL admits unbounded bias under any positive contamination, whereas RoPoLL (the geometric median of the panel) attains a minimax-tight finite-sample bound and empirically dominates PoLL on biased corruption at a small clean-baseline premium —*robust aggregation, not the ensemble itself, delivers the win*. Open directions include per-judge heterogeneity and explicit dependence.

References

Acharya, A., Hashemi, A., Jain, P., Sanghavi, S., Dhillon, I. S., and Topcu, U. Robust training in high dimensions via block coordinate geometric median descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 11145–11168. PMLR, 2022.

Acharya, A., Sanghavi, S., Dimakis, A., and Dhillon, I. S. Geometric median (GM) matching for robust k-subset selection from noisy data. In *International Conference on Machine Learning*, pp. 372–419. PMLR, 2025.

Aloupis, G. Geometric measures of data depth. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:147–158, 2006.

Bajaj, C. The algebraic degree of geometric optimization problems. *Discrete & Computational Geometry*, 3(2): 177–191, 1988.

Billingsley, P. *Probability and Measure*. John Wiley & Sons, New York, 3 edition, 1995.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2024.

Chen, M., Gao, C., and Ren, Z. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2024.

Donoho, D. L. and Gasko, M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20(4):1803–1827, 1992.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Dudley, R. M. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002.

El Mhamdi, E. M., Guerraoui, R., and Rouault, S. The hidden vulnerability of distributed learning in Byzantium. In *International Conference on Machine Learning*, pp. 3521–3530, 2018.

Esary, J. D., Proschan, F., and Walkup, D. W. Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

Hopkins, S. B. Mean estimation with sub-Gaussian rates in polynomial time. *The Annals of Statistics*, 48(2):1193–1213, 2020.

Huber, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Joag-Dev, K. and Proschan, F. Negative association of random variables, with applications. *The Annals of Statistics*, 11(1):286–295, 1983.

Johnson, D. S. and Preparata, F. P. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

Kallenberg, O. *Foundations of Modern Probability*. Springer, New York, 2 edition, 2002.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., et al. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Li, R., Patel, T., and Du, X. PRD: Peer rank and discussion improve large language model based evaluations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=YVD1QqWRaj.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023.

Lopuhaä, H. P. and Rousseeuw, P. J. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1): 229–248, 1991.

Lugosi, G. and Mendelson, S. Sub-Gaussian mean estimators. *The Annals of Statistics*, 47(2):783–794, 2019.

Massart, P. *Concentration Inequalities and Model Selection: École d'Été de Probabilités de Saint-Flour XXXIII – 2003*. Springer, 2007.

Minsker, S. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Panickssery, A., Bowman, S. R., and Feng, S. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4NJBV6Wp0h.

Pitt, L. D. A gaussian correlation inequality for symmetric convex sets. *The Annals of Probability*, pp. 470–474, 1977.

Rockafellar, R. T. *Convex analysis*, volume 28. Princeton university press, 1997.

Saito, K., Wachi, A., Wataoka, K., and Akimoto, Y. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Small, C. G. A survey of multidimensional medians. *International Statistical Review*, 58(3):263–277, 1990.

Stureborg, R., Alikaniotis, D., and Suhara, Y. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

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

Tukey, J. W. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, pp. 448–485, 1960.

Tukey, J. W. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, volume 2, pp. 523–531, 1975.

Vardi, Y. and Zhang, C.-H. The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., and Lewis, P. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., and Sui, Z. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Egert, G., Zhang, P., Kamalakara, A. S., and Kuchaiev, O. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024.

Wang, Z., Zeng, J., Delalleau, O., Egert, D., Evans, E., Shin, H.-C., Soares, F., Dong, Y., and Kuchaiev, O. HelpSteer3: Human-annotated feedback and edit data to empower inference-time scaling in open-ended general-domain tasks. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25640–25662, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1246. URL https://aclanthology.org/2025.acl-long.1246/.

Weiszfeld, E. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal*, 43:355–386, 1937.

Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Mun, Y., Lee, J., Park, B., Shin, S., Kim, S., et al. FLASK: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*, 2024.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659, 2018.

Zhang, X., Yu, B., Yu, H., Lv, Y., Liu, T., Huang, F., Xu, H., and Li, Y. Wider and deeper LLM networks are fairer LLM evaluators. *arXiv preprint arXiv:2407.13275*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623, 2023.

Appendix

The appendix collects the deferred proofs and supporting material for the body of the paper.

A. Related Work

LLM-as-Judge evaluation and per-judge biases. The LLM-as-Judge paradigm was established by Zheng et al. (2023) (MT-Bench, Chatbot Arena), demonstrating that strong models such as GPT-4 can serve as reliable proxies for human annotators. Subsequent work has extended the paradigm along several axes: open-source judges with fine-grained rubrics (Kim et al., 2024); automated frameworks for instruction-following models (Li et al., 2023; Dubois et al., 2024); and skill-level evaluation (Ye et al., 2024). A parallel literature documents systematic biases of single judges—position, verbosity, self-enhancement, sycophancy, and prompt-format sensitivity (Wang et al., 2023; Panickssery et al., 2024; Saito et al., 2023; Stureborg et al., 2024). These findings motivate the use of diverse judge panels but treat each judge in isolation; no prior work analyzes the *aggregation* step or its failure modes.

Jury and panel evaluation. Verga et al. (2024) introduced the Panel of LLM Evaluators (POLL), our direct predecessor: a diverse committee of smaller backbones aggregated by the arithmetic mean. Their work established the practical value of LLM juries but did not analyze robustness; the mean aggregator is used without justification, and no failure modes are considered. Zhang et al. (2024) studied how panel width and depth affect evaluation fairness, again without robustness guarantees. The key gap across this literature is the absence of any analysis of catastrophic failure modes or formal robustness properties of the aggregation rule. Our Proposition 2 closes this gap: under any positive contamination rate POLL (Verga et al., 2024) admits unbounded bias regardless of N.

Multi-agent debate and structured aggregation. A distinct family of multi-judge methods produces aggregated judgments through structured *interaction* rather than independent scoring. Li et al. (2024) propose peer-rank discussion among judges, in which each judge sees others' scores and updates its own; Chan et al. (2024) propose multi-agent debate, in which judges argue over a verdict before consensus. These methods change the joint distribution of $(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N)$—they introduce dependence by design, breaking Assumption 2—and trade independence for deliberation-driven error reduction. Whether they exhibit the same Byzantine-failure mode as POLL is an open question. The Huber-contamination analysis of this paper does not directly apply to such interactive aggregators, but the corruption-class diagnosis (point masses far from the truth) likely transfers, suggesting robust extensions of debate-based aggregation as a future direction. Majority voting in mathematical reasoning (Cobbe et al., 2021) is a related but coarser ensemble technique on binary correctness; the analogue of Proposition 2 for vote-based aggregation on $\{0, 1\}$ outputs is the standard $\alpha < 1/2$ Byzantine threshold.

Calibration as a complementary paradigm. A separate line of work removes judge bias *at the source* via per-judge calibration on a labeled validation slice (Zheng et al., 2023). Calibration assumes a stationary, recoverable bias and trades worst-case guarantees for average-case efficiency; ROPOLL assumes nothing on the corruption distribution and pays a constant-factor insurance premium to bound the worst case. The two are complementary: ROPOLL can aggregate calibrated scores, and the calibration-RoPoLL composition—together with extensions to heterogeneous, correlated, and dependent juries—is left to future work.

Robust statistics and the geometric median. The Huber contamination model (Huber, 1964) and the breakdown point (Tukey, 1960) are the classical framework for estimation under arbitrary corruption. The geometric median attains the optimal $1/2$ breakdown for any translation-equivariant estimator (Lopuhaä & Rousseeuw, 1991; Small, 1990; Vardi & Zhang, 2000); in high dimensions, Minsker (2015) established sub-Gaussian concentration for the geometric median of means—the result Theorem 1 adapts to contaminated juries—and Lugosi & Mendelson (2019) developed sub-Gaussian mean estimators with optimal dimension dependence. Recent applications to ML pipelines include block-coordinate GM descent for robust training (Acharya et al., 2022) and GM Matching for robust subset selection (Acharya et al., 2025). Our setting differs from this literature on three axes: *(i) low dimension* ($d \in \{4, 5\}$ evaluation attributes, so the $\sqrt{d/N}$ rate is dominated by constants and the $1/(1-2\alpha)$ contamination factor is the load-bearing dependence); *(ii) structured contamination* (Q_i arises from specific LLM failure modes—parser fallback, sycophancy, refusals, cross-attribute confusion—which inform the four empirical corruption types in §5); and *(iii) heterogeneous workers* (per-judge σ_i, α_i vary across the panel, outside the i.i.d. regime that the classical robust-statistics literature targets). Among broader alternatives in the robust-aggregation toolbox, the *half-space (Tukey) median* attains the optimal breakdown $1/2$ in any dimension but is NP-hard to compute and prohibitive at $d \geq 5$ (Small, 1990); *median of means* (Lugosi & Mendelson, 2019) targets heavy-tailed data rather than Huber contamination concentrated in a minority of judges; the geometric median's tuning-free $1/2$ breakdown, joint-distance

Result	One-line statement
Section 2 Problem Setup	
Assumption 1	Huber ϵ-contamination model.
Assumption 2	Conditional independence of judges.
Assumption 3	σ-sub-Gaussian competent noise.
Assumption 4	Minority corruption $\alpha < 1/2$.
Proposition 1	Clean-jury MSE $\mathrm{tr}(\mathbf{\Sigma})/N$.
Corollary 1	Effective jury size $N_{\text{eff}}{=}N/(1{+}(N{-}1)\gamma)$.
Proposition 2	POLL bias unbounded for any $\alpha > 0$.
Section 3 Robust Panel of LLM Judges	
Example 1	Cross-dimensional corruption: per-coord plausible, jointly anomalous.
Definition 4	ROPOLL via geometric median.
Definition 5	Finite-sample breakdown point.
Proposition 3	GM existence, uniqueness, equivariance, breakdown $1/2$.
Algorithm 1	Modified Weiszfeld iteration: $O(Nd\log(1/\epsilon))$.
Section 4 Theoretical Guarantees	
Lemma 1	GM within $C_\alpha r$ if $(1{-}\alpha)$ fraction of points are within r.
Lemma 2	Sub-Gaussian cluster radius $\rho{=}\sigma(C_1\sqrt{d}{+}\sqrt{\log(1/\beta)/c})$.
Theorem 1	ROPOLL error $\leq C_{\alpha+\beta}\rho$ w.p. $1{-}\exp(-N\beta^2/2)$.
Lemma 3	Same bound under equicorrelated juries, w.p. $1{-}1/(\beta^2 N_{\text{eff}})$.
Theorem 2	Minimax lower bound $\Omega(\sigma(\sqrt{d/N}{+}\alpha/(1{-}\alpha)))$.

Table 1. **Roadmap of formal results.** Each row links to the result's full statement (clickable reference); proofs are deferred to the Section B

objective, and $O(Nd\log(1/\epsilon))$ cost make it the right default for the small-d, small-N, heterogeneous-worker, one-shot regime that LLM juries occupy. A systematic empirical comparison against the broader family is left to future work (§6).

Byzantine-robust distributed learning. The connection between robust aggregation and Byzantine fault tolerance has been worked out in distributed optimization: Krum (Blanchard et al., 2017), coordinate-wise median and trimmed mean as gradient aggregators (Yin et al., 2018), and Bulyan (El Mhamdi et al., 2018). This literature targets N from tens to thousands of workers, with adversarial perturbations composed across thousands of training rounds. The LLM-jury setting shares the mathematical structure but differs operationally on three axes: *(a) small N* (juries operate at $N \in \{3,\dots,13\}$ where every judge is materially expensive, requiring tight finite-sample guarantees); *(b) per-sample heterogeneity* (the contamination indicator Z_i is conditional on the prompt-response pair x, not per round); *(c) no iterative learning loop* (LLM-jury aggregation is one-shot at evaluation time, so the per-instance bias bound matters directly rather than its cumulative effect across rounds). These differences explain why our analysis emphasizes finite-sample distribution-free guarantees over the corruption class (Theorem 1); the heterogeneity of the worker pool, judge correlation, and explicit dependence (in debate-based methods) are left to future work and have no direct analogue in the Byzantine distributed-learning literature.

To our knowledge, we are the first to formalize LLM jury aggregation as a robust estimation problem, prove finite-sample contamination guarantees in this setting, and evaluate robustness systematically against both natural and adversarial judge failures at scale.

B. Complete Proofs and Full Theoretical Development

B.1. Proof of Proposition 1 (Variance Reduction)

Proof of Proposition 1. Under $\alpha_i = 0$, Assumption 1 gives $\mathbb{E}[\hat{\mathbf{y}}_i \mid \mathbf{y}^\star] = \mathbf{y}^\star$ and $\mathrm{Cov}(\hat{\mathbf{y}}_i \mid \mathbf{y}^\star) = \mathbf{\Sigma}_i$. Linearity of conditional expectation gives $\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] = \mathbf{y}^\star$, and bilinearity of covariance gives

$$\mathrm{Cov}(\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star) \;=\; \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{Cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \mid \mathbf{y}^\star)\,. \tag{12}$$

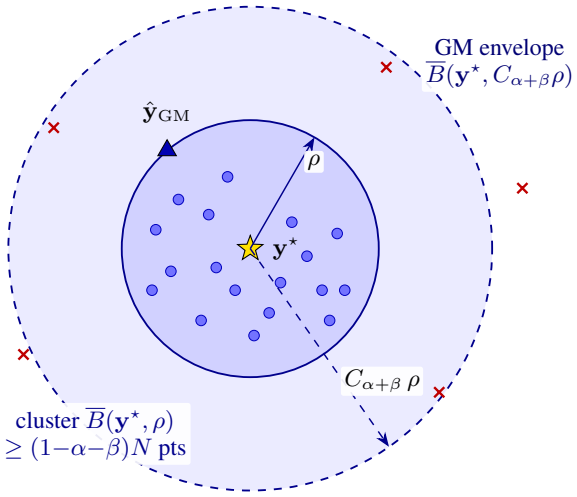


Figure 6. **Geometry of Theorem 1.** Lemma 2 guarantees that at least $(1-\alpha-\beta)N$ judge outputs (blue dots) lie inside the *cluster ball* $\overline{B}(\mathbf{y}^\star, \rho)$ of sub-Gaussian radius ρ (solid disk). Lemma 1 then forces the geometric median $\hat{\mathbf{y}}_{\mathrm{GM}}$ (blue triangle) to lie inside the *GM envelope* $\overline{B}(\mathbf{y}^\star, C_{\alpha+\beta}\rho)$ (dashed disk), *regardless of where the remaining* $(\alpha+\beta)N$ *corrupted points (red* $\times$*) are placed*—this is the distribution-free breakdown property of the geometric median. The two-step composition is exactly Theorem 1.

Independence (Assumption 2) zeroes the cross-covariances, yielding

$$\mathrm{Cov}(\hat{\mathbf{y}}_{\mathrm{mean}} \mid \mathbf{y}^\star) = \frac{1}{N^2}\sum_{i=1}^{N} \mathbf{\Sigma}_i, \quad \mathbb{E}\big[\|\hat{\mathbf{y}}_{\mathrm{mean}} - \mathbf{y}^\star\|_2^2 \,\big|\, \mathbf{y}^\star\big] = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{tr}(\mathbf{\Sigma}_i), \tag{13}$$

the second equality using that the conditional error is centered. The final bound follows from $\mathrm{tr}(\mathbf{\Sigma}_i) \leq d\sigma^2$ whenever $\mathbf{\Sigma}_i \preceq \sigma^2 \mathbf{I}_d$. ■

Proof of Corollary 1. Substituting the equicorrelated structure $\mathrm{Cov}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \mid \mathbf{y}^\star) = \gamma\,\mathbf{\Sigma}$ for $i \neq j$ and $\mathrm{Cov}(\hat{\mathbf{y}}_i \mid \mathbf{y}^\star) = \mathbf{\Sigma}$ into (12) gives $\mathrm{Cov}(\hat{\mathbf{y}}_{\mathrm{mean}} \mid \mathbf{y}^\star) = \frac{1+(N-1)\gamma}{N}\,\mathbf{\Sigma}$; taking traces yields the MSE expression. ■

Remark 4 (Implication for Jury Design)**.** *Proposition 1 and Corollary 1 formalize the classical benefit of a jury: independent, diverse, conditionally unbiased judges reduce estimator variance, with an effective sample-size penalty determined by their pairwise dependence. Proposition 2 shows why this benefit is insufficient under contamination: the arithmetic mean's bias is unbounded over the corruption class (Assumption 1) regardless of* N*, so any variance reduction the jury affords is dominated by an adversarial choice of* $\{Q_i\}$*. A robust aggregation rule is therefore needed to preserve the signal of the competent majority while attenuating contamination bias—this is the role of* RoPoLL *in §3.*

B.2. Proof of Proposition 2 (Unbounded Bias of PoLL)

For convenience we recall the statement: under Assumption 1 and finite first moments $\mu_i^Q \triangleq \mathbb{E}_{Q_i}[\hat{\mathbf{y}}_i]$, the mean's conditional bias is

$$\mathbb{E}[\hat{\mathbf{y}}_{\mathrm{mean}} \mid \mathbf{y}^\star] - \mathbf{y}^\star \;=\; \frac{1}{N}\sum_{i=1}^{N} \alpha_i \Big(\mu_i^Q - \mathbf{y}^\star\Big), \tag{14}$$

and is not uniformly bounded over the corruption class as long as $\alpha > 0$, regardless of N.

Proof of Proposition 2. We prove the two claims in turn: the explicit bias formula (14), and the impossibility of a uniform bound over the corruption class.

Step 1: Per-judge expectation. Fix $i \in [N]$. By Assumption 1, conditional on $\mathbf{y}^\star$ the law of $\hat{\mathbf{y}}_i$ is the mixture $(1-\alpha_i)P_i + \alpha_i Q_i$ with selector $Z_i \sim \mathrm{Bernoulli}(\alpha_i)$. By the law of total expectation,

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{y}}_i \mid \mathbf{y}^\star] &= (1-\alpha_i)\,\mathbb{E}_{P_i}[\hat{\mathbf{y}}_i] + \alpha_i\,\mathbb{E}_{Q_i}[\hat{\mathbf{y}}_i] \\ &= (1-\alpha_i)\,\mathbf{y}^\star + \alpha_i\,\mu_i^Q,\end{aligned}$$

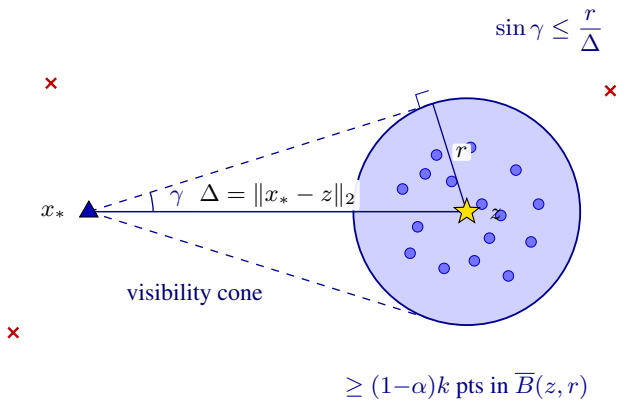


Figure 7. **Geometric core of Lemma 1.** The contradiction hypothesis $\Delta \triangleq \|x_* - z\|_2 > C_\alpha r$ places x_* *outside* the cluster ball $\overline{B}(z,r)$, so the ball subtends a cone of half-angle γ at x_* with $\sin\gamma = r/\Delta$ (right-angle at the tangent point shown). Every cluster point $x_j \in \overline{B}(z,r)$ lies inside this cone, hence makes angle $\gamma_j \le \gamma$ with the central ray $x_* \to z$, so $\cos\gamma_j \ge \sqrt{1-r^2/\Delta^2} = \alpha/(1-\alpha)$ when $\Delta > C_\alpha r$ with $C_\alpha = (1-\alpha)/\sqrt{1-2\alpha}$. Summing this lower bound over the $(1-\alpha)k$ cluster indices forces the directional derivative $DF(x_*; z - x_*)$ to be strictly negative, contradicting the first-order optimality of the geometric median—hence $\Delta \le C_\alpha r$.

where the second equality uses $\mathbb{E}_{P_i}[\hat{\mathbf{y}}_i] = \mathbf{y}^\star$ (competent unbiasedness, Assumption 1) and the finite-first-moment assumption on Q_i to identify $\mathbb{E}_{Q_i}[\hat{\mathbf{y}}_i] = \mu_i^Q$. Rearranging,

$$\mathbb{E}[\hat{\mathbf{y}}_i \mid \mathbf{y}^\star] - \mathbf{y}^\star \;=\; \alpha_i\,(\mu_i^Q - \mathbf{y}^\star). \tag{15}$$

Step 2: Linearity of the mean. By the linearity of expectation applied to $\hat{\mathbf{y}}_{\text{mean}} = N^{-1}\sum_{i=1}^N \hat{\mathbf{y}}_i$,

$$\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] \;=\; \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\hat{\mathbf{y}}_i \mid \mathbf{y}^\star].$$

Substituting (15) yields (14), proving the first claim.

Step 3: Adversarial corruption distribution. Suppose $\alpha = N^{-1}\sum_i \alpha_i > 0$. Then there exists at least one index $i_0 \in [N]$ with $\alpha_{i_0} > 0$. Let $B > 0$ be arbitrary, $\mathbf{e}_1$ be the first standard basis vector, and consider the adversarial choice

$$Q_{i_0} \;=\; \delta_{\mathbf{y}^\star + (NB/\alpha_{i_0})\,\mathbf{e}_1}, \tag{16}$$

the Dirac mass placed at the indicated point; take $\{Q_i\}_{i\neq i_0}$ to be any distributions with $\mu_i^Q = \mathbf{y}^\star$ (e.g., $Q_i = P_i$ itself, which gives zero contribution to the bias). Then $\mu_{i_0}^Q = \mathbf{y}^\star + (NB/\alpha_{i_0})\,\mathbf{e}_1$, and (14) reduces to

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] - \mathbf{y}^\star &= \frac{1}{N}\,\alpha_{i_0}\,(\mu_{i_0}^Q - \mathbf{y}^\star) \\ &= \frac{1}{N}\,\alpha_{i_0}\,\frac{NB}{\alpha_{i_0}}\,\mathbf{e}_1 \;=\; B\,\mathbf{e}_1.\end{aligned}$$

Hence $\left\|\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] - \mathbf{y}^\star\right\|_2 = B$.

Step 4: Conclusion. Since $B > 0$ was arbitrary, no constant $C(\alpha, N, d, \sigma)$ depending only on the model parameters of Assumptions 1–4 can satisfy

$$\sup_{\{Q_i\}} \left\|\mathbb{E}[\hat{\mathbf{y}}_{\text{mean}} \mid \mathbf{y}^\star] - \mathbf{y}^\star\right\|_2 \;\le\; C(\alpha, N, d, \sigma).$$

The bias is therefore unbounded over the corruption class for every fixed N, completing the proof. ■

Remark 5 (Why N does not help)**.** *The construction (16) scales $\mu_{i_0}^Q$ with N: the adversary's per-judge displacement grows linearly with the jury size, exactly cancelling the $1/N$ averaging. This is the formal statement of why variance reduction (Proposition 1) cannot rescue the mean under contamination: variance contracts at rate $1/N$, but bias is preserved by the adversary irrespective of N, and the bias term dominates as long as $\alpha > 0$.*

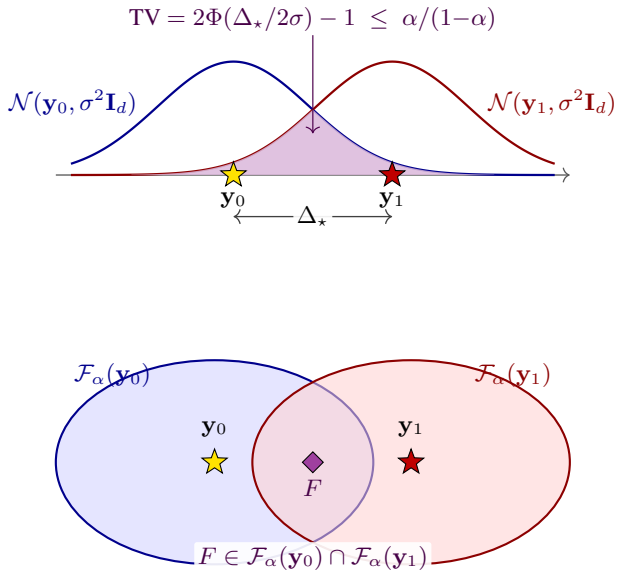


Figure 8. **Modulus of continuity for Theorem 2. Top:** total variation between two equal-covariance Gaussians at separation $\Delta_\star$ is $2\Phi(\Delta_\star/2\sigma) - 1$ (*dimension-free*; depends only on the line through the two centers). The overlap is shaded; when the overlap mass exceeds $\alpha/(1-\alpha)$, the two Huber neighborhoods touch. **Bottom:** the contamination class $\mathcal{F}_\alpha(\mathbf{y}) = \{(1-\alpha)\mathcal{N}(\mathbf{y}, \sigma^2\mathbf{I}_d) + \alpha Q\}$ is depicted as a cloud of distributions around each center; under the threshold above, a single distribution $F \in \mathcal{F}_\alpha(\mathbf{y}_0) \cap \mathcal{F}_\alpha(\mathbf{y}_1)$ is consistent with *both* truths. No estimator can distinguish $\mathbf{y}_0$ from $\mathbf{y}_1$ on observations drawn from F, hence Le Cam's two-point bound forces minimax error $\geq \Delta_\star/4 \geq \frac{\sqrt{2\pi}}{4}\sigma\alpha/(1-\alpha)$, *independent of* N.

B.3. Proof of Proposition 3

Proof of Proposition 3. (i) Existence. Each summand $\mathbf{z} \mapsto \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2$ is the Euclidean norm of an affine function of $\mathbf{z}$, hence continuous and convex (see any standard reference on convex analysis). Sums of continuous convex functions are continuous and convex, so F is continuous and convex. For coercivity, fix any data point $\hat{\mathbf{y}}_1$; by the reverse triangle inequality

$$F(\mathbf{z}) \;\geq\; \|\mathbf{z} - \hat{\mathbf{y}}_1\|_2 \;\geq\; \|\mathbf{z}\|_2 - \|\hat{\mathbf{y}}_1\|_2 \;\to\; \infty \quad \text{as } \|\mathbf{z}\|_2 \to \infty.$$

Since F is continuous and coercive, the sublevel set $\{\mathbf{z} : F(\mathbf{z}) \leq F(\mathbf{0})\}$ is nonempty, closed, and bounded, hence compact in $\mathbb{R}^d$. Weierstrass's theorem then yields a minimizer.

(ii) Uniqueness. The function $\mathbf{z} \mapsto \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2$ is strictly convex on every line not passing through $\hat{\mathbf{y}}_i$ and affine on the line through $\hat{\mathbf{y}}_i$ in the direction of any other point. Suppose the data are not collinear: then for any line $\mathcal{L} \subset \mathbb{R}^d$ there exists at least one $\hat{\mathbf{y}}_i \notin \mathcal{L}$, so the corresponding summand is strictly convex along $\mathcal{L}$. Hence F is strictly convex along every line, hence strictly convex on $\mathbb{R}^d$, and the minimizer is unique (Vardi & Zhang, 2000).

(iii) Affine equivariance. Let $\mathbf{U} \in \mathbb{R}^{d\times d}$ be orthogonal and $\mathbf{b} \in \mathbb{R}^d$. For all $\mathbf{z} \in \mathbb{R}^d$,

$$\|\mathbf{U}\mathbf{z} + \mathbf{b} - (\mathbf{U}\hat{\mathbf{y}}_i + \mathbf{b})\|_2 \;=\; \|\mathbf{U}(\mathbf{z} - \hat{\mathbf{y}}_i)\|_2 \;=\; \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2,$$

where the second equality uses $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_d$. Summing over i, $F^{\mathbf{U},\mathbf{b}}(\mathbf{U}\mathbf{z} + \mathbf{b}) = F(\mathbf{z})$ where $F^{\mathbf{U},\mathbf{b}}$ is the objective on the transformed sample. The map $\mathbf{z} \mapsto \mathbf{U}\mathbf{z} + \mathbf{b}$ is a bijection on $\mathbb{R}^d$, so the two minimizers are related by exactly this transformation.

(iv) Breakdown point. We show that the GM tolerates any corruption of strictly fewer than $\lceil N/2 \rceil$ points and that this threshold is tight.

Sufficiency. Suppose $m < \lceil N/2 \rceil$ points are arbitrarily replaced and denote the corrupted sample $\hat{\mathbf{y}}'_{1:N}$. The competent set $S = \{i : \hat{\mathbf{y}}'_i = \hat{\mathbf{y}}_i\}$ has $|S| = N - m > N/2$, hence $|S| > |S^c|$. Let $\mathbf{z}' = \mathrm{GM}(\hat{\mathbf{y}}'_{1:N})$ be the corrupted GM. By the subgradient optimality condition for the convex objective F',

$$\mathbf{0} \in \partial F'(\mathbf{z}') = \sum_{i:\mathbf{z}' \neq \hat{\mathbf{y}}'_i} \frac{\mathbf{z}' - \hat{\mathbf{y}}'_i}{\|\mathbf{z}' - \hat{\mathbf{y}}'_i\|_2} + (\text{ball terms for ties}).$$

Each unit-vector term has norm 1. If $\|\mathbf{z}'\|_2$ were unbounded as the adversary varies the corrupted points within their m-coordinate budget, then for the competent points $i \in S$ the unit vectors $(\mathbf{z}' - \hat{\mathbf{y}}_i)/\|\mathbf{z}' - \hat{\mathbf{y}}_i\|_2$ would all lie in a small cone (all pointing approximately from the bounded competent cluster toward $\mathbf{z}'$), so their sum has norm at least $|S|(1 - o(1))$. The corrupted contribution has norm at most $|S^c| < |S|$, hence the total subgradient has norm at least $|S| - |S^c| > 0$, contradicting the optimality $\mathbf{0} \in \partial F'(\mathbf{z}')$. Therefore $\|\mathbf{z}'\|_2$ remains bounded, i.e. no m-budget corruption can drive the GM to infinity.

Necessity. With $m = \lceil N/2 \rceil$ corrupted points all placed at a common location $\hat{\mathbf{y}}'_{i_0} = M \cdot \mathbf{e}_1$ for arbitrarily large M, the corrupted set forms a majority (or tie if N is even) and the GM moves to within $O(1)$ of $M\,\mathbf{e}_1$ as $M \to \infty$ (Lopuhaä & Rousseeuw, 1991). Hence the breakdown point is exactly $\epsilon^\star = \lceil N/2 \rceil / N$, which tends to $1/2$ as $N \to \infty$. This is the optimal breakdown for any translation-equivariant estimator (Lopuhaä & Rousseeuw, 1991). ■

B.4. Weiszfeld Iteration: Full Derivation, Convergence, and Cost

For completeness, this subsection gives the full derivation, convergence statement, and cost analysis for the Weiszfeld iteration sketched in §3.2.

Derivation. At a non-data point $\mathbf{z} \neq \hat{\mathbf{y}}_i$ for all i, the gradient of the GM objective $F(\mathbf{z}) = \sum_i \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2$ is

$$\nabla F(\mathbf{z}) = \sum_{i=1}^{N} \frac{\mathbf{z} - \hat{\mathbf{y}}_i}{\|\mathbf{z} - \hat{\mathbf{y}}_i\|_2}. \tag{17}$$

Setting $\nabla F(\mathbf{z}) = \mathbf{0}$ and rearranging gives the fixed-point equation (5) of §3.2,

$$\mathbf{z} \;=\; \frac{\sum_{i=1}^{N} \hat{\mathbf{y}}_i / \|\mathbf{z} - \hat{\mathbf{y}}_i\|_2}{\sum_{i=1}^{N} 1/\|\mathbf{z} - \hat{\mathbf{y}}_i\|_2},$$

which is the Weiszfeld iteration $\mathbf{z} \leftarrow T(\mathbf{z})$. When the current iterate coincides with a data point $\hat{\mathbf{y}}_j$, the denominator $\|\mathbf{z} - \hat{\mathbf{y}}_j\|_2 = 0$ creates a singularity; the modified step of Vardi & Zhang (2000) replaces the weight by

$$w_i^{(t)} = \frac{1}{\max\big(\|\mathbf{z}^{(t)} - \hat{\mathbf{y}}_i\|_2,\ \eta\big)} \tag{18}$$

for a small stability parameter $\eta > 0$, recovering Algorithm 1.

Convergence. Vardi & Zhang (2000) prove that the modified Weiszfeld iteration converges to the unique geometric median at a linear rate whenever the data are not collinear: there exists $\rho \in (0, 1)$ depending on the data configuration with $\|\mathbf{z}^{(t)} - \hat{\mathbf{y}}_{\mathrm{GM}}\|_2 \le \rho^t \|\mathbf{z}^{(0)} - \hat{\mathbf{y}}_{\mathrm{GM}}\|_2$. The number of iterations to reach tolerance ϵ is therefore $O(\log(1/\epsilon))$.

Cost. Each iteration computes N Euclidean distances in $\mathbb{R}^d$ and one weighted average, costing $O(Nd)$ arithmetic operations. With $O(\log(1/\epsilon))$ iterations the total cost is $O(Nd \log(1/\epsilon))$. For typical LLM juries ($N \le 20$, $d \le 5$, $\epsilon = 10^{-8}$) this amounts to a few hundred floating-point operations—microseconds on any modern processor, while a single LLM judge invocation costs seconds of GPU time. The aggregation step is computationally negligible relative to the inference cost of the jury.

B.5. Proof of Lemma 1

For convenience, we recall Lemma 1: let $x_1, \ldots, x_k \in \mathbb{R}^d$ and let x_* be any minimizer of $z \mapsto \sum_{j=1}^{k} \|z - x_j\|_2$; fix $\alpha \in (0, 1/2)$, $r > 0$, $z \in \mathbb{R}^d$. If $|\{j : \|x_j - z\|_2 \le r\}| \ge (1 - \alpha)k$, then $\|x_* - z\|_2 \le C_\alpha\, r$ with $C_\alpha = (1 - \alpha)/\sqrt{1 - 2\alpha}$.

Proof of Lemma 1. We give the proof of Minsker (2015), with the geometric setup made explicit. The argument is by contradiction: assume $\|x_* - z\|_2 > C_\alpha r$ and derive a violation of the optimality of x_*.

For brevity write $\Delta \triangleq \|x_* - z\|_2$ and let $F(y) \triangleq \sum_{j=1}^{k} \|y - x_j\|_2$ denote the geometric-median objective. Since x_* minimizes the convex function F on $\mathbb{R}^d$, the one-sided directional derivative of F at x_* in any direction $v \in \mathbb{R}^d$ is

non-negative (standard convex analysis; see e.g. Rockafellar, 1997, Theorem 23.1). Taking $v \triangleq z - x_*$:

$$DF(x_*; v) \;\triangleq\; \lim_{t \downarrow 0} \frac{F(x_* + tv) - F(x_*)}{t} \;\geq\; 0. \tag{19}$$

Step 1: Compute the directional derivative. The function $y \mapsto \|y - x_j\|_2$ is the Euclidean norm of an affine function; it is differentiable at any $y \neq x_j$ with gradient $(y - x_j)/\|y - x_j\|_2$ (the Fermat–Weber gradient, classical; cf. Weiszfeld, 1937; Vardi & Zhang, 2000). For j with $x_j = x_*$, the directional derivative of $y \mapsto \|y - x_*\|_2$ at x_* in direction v equals $\|v\|_2$ (the Euclidean norm is positively homogeneous, so its right-hand directional derivative at the origin is $\|v\|_2$). Letting $K_* = \{j : x_j = x_*\}$, the total directional derivative decomposes as

$$DF(x_*; v) \;=\; \sum_{j \notin K_*} \frac{\langle x_* - x_j, v \rangle}{\|x_* - x_j\|_2} \;+\; |K_*|\, \|v\|_2.$$

Substituting $v = z - x_*$ and dividing by $\|v\|_2 = \Delta > 0$:

$$\frac{DF(x_*; z - x_*)}{\Delta} \;=\; -\sum_{j \notin K_*} \cos\gamma_j \;+\; |K_*|, \tag{20}$$

where γ_j is the angle at x_* between the rays $x_* \to x_j$ and $x_* \to z$, defined for $j \notin K_*$ by $\cos\gamma_j = \langle x_j - x_*,\, z - x_* \rangle / (\|x_j - x_*\|_2\, \Delta)$.

Step 2: Lower-bound $\cos\gamma_j$ for points near z. Let $J \triangleq \{j : \|x_j - z\|_2 \leq r\}$ denote the indices of points within distance r of z. By hypothesis, $|J| \geq (1 - \alpha)k$.

For $j \in J$, the point x_j lies in the closed ball $\overline{B}(z, r)$. The angle γ_j at x_* between the rays $x_* \to x_j$ and $x_* \to z$ is at most the half-angle subtended by the ball $\overline{B}(z, r)$ as seen from x_*. Since $\|x_* - z\|_2 = \Delta$ and the ball has radius r, elementary geometry gives

$$\sin\gamma_j \;\leq\; \frac{r}{\Delta}, \qquad \cos\gamma_j \;\geq\; \sqrt{1 - \frac{r^2}{\Delta^2}}. \tag{21}$$

By assumption $\Delta > C_\alpha r$, so $r/\Delta < 1/C_\alpha$ and (21) yields

$$\cos\gamma_j \;>\; \sqrt{1 - \frac{1}{C_\alpha^2}} \quad \text{for all } j \in J. \tag{22}$$

For $j \in J^c \setminus K_*$ (points farther than r from z that do not coincide with x_*), we have only the trivial bound $\cos\gamma_j \geq -1$.

Step 3: Combine. A short observation simplifies the algebra: the constant $C_\alpha = (1 - \alpha)/\sqrt{1 - 2\alpha} \geq 1$ for $\alpha \in [0, 1/2)$ (with equality only at $\alpha = 0$). Combined with the contradiction hypothesis $\Delta > C_\alpha r \geq r$, this implies that every $j \in K_*$ (where $x_j = x_*$, so $\|x_j - z\|_2 = \Delta > r$) satisfies $j \in J^c$. Therefore $K_* \subseteq J^c$, and the partition $J^c = (J^c \setminus K_*) \cup K_*$ gives $|J^c \setminus K_*| = |J^c| - |K_*|$.

Substituting into (20) and using the angular bounds from Step 2:

$$\begin{aligned}
\frac{DF(x_*; z - x_*)}{\Delta} \;&=\; -\sum_{j \in J} \cos\gamma_j \;-\; \sum_{j \in J^c \setminus K_*} \cos\gamma_j \;+\; |K_*| \\
&<\; -|J|\, \sqrt{1 - 1/C_\alpha^2} \;+\; |J^c \setminus K_*| \;+\; |K_*| \\
&=\; -|J|\, \sqrt{1 - 1/C_\alpha^2} \;+\; |J^c| \\
&\leq\; -(1 - \alpha)k\, \sqrt{1 - 1/C_\alpha^2} \;+\; \alpha k,
\end{aligned}$$

where the final line uses $|J| \geq (1 - \alpha)k$ and $|J^c| \leq \alpha k$.

We now show that the choice $C_\alpha = (1 - \alpha)/\sqrt{1 - 2\alpha}$ makes this strictly negative. Compute:

$$1 - \frac{1}{C_\alpha^2} \;=\; 1 - \frac{1 - 2\alpha}{(1 - \alpha)^2} \;=\; \frac{(1 - \alpha)^2 - (1 - 2\alpha)}{(1 - \alpha)^2} \;=\; \frac{\alpha^2}{(1 - \alpha)^2}.$$

Hence $\sqrt{1-1/C_\alpha^2} = \alpha/(1-\alpha)$, and:

$$\frac{DF(x_*; z-x_*)}{\Delta} \;<\; -(1-\alpha)k \cdot \frac{\alpha}{1-\alpha} + \alpha k \;=\; -\alpha k + \alpha k \;=\; 0.$$

This contradicts (19), which required $DF(x_*; z-x_*) \geq 0$. Therefore the assumption $\|x_* - z\|_2 > C_\alpha r$ must fail, proving (6). ■

Remark 6 (Sanity checks for C_α)**.** *The constant $C_\alpha = (1-\alpha)/\sqrt{1-2\alpha}$ behaves as expected at the boundary cases:*

- *At $\alpha = 0$: $C_0 = 1$. The lemma reduces to "if all k points lie within r of z, then their geometric median lies within r of z," which is immediate because the geometric median lies in the convex hull of the points, hence in $\overline{B}(z, r)$.*
- *At $\alpha = 1/4$: $C_{1/4} = (3/4)/\sqrt{1/2} = 3/(2\sqrt{2}) \approx 1.061$.*
- *At $\alpha = 0.3$: $C_{0.3} = 0.7/\sqrt{0.4} \approx 1.107$.*
- *As $\alpha \to 1/2$: $C_\alpha \to \infty$. The lemma becomes vacuous, matching the breakdown point of the geometric median: with corrupted majority, no constant bound on $\|x_* - z\|_2$ is possible.*

Remark 7 (Tightness)**.** *The constant $C_\alpha = (1-\alpha)/\sqrt{1-2\alpha}$ is sharp in the sense that the same proof technique cannot give a smaller constant: the inequality $\sin\gamma_j \leq r/\Delta$ in (21) is achieved when x_j lies on the boundary of $\overline{B}(z, r)$ at the tangent point from x_*, and the bound on the directional derivative is tight for that configuration. A matching example: place $(1-\alpha)k$ points on the boundary of $\overline{B}(z, r)$ at the tangent points from a location x_* at distance $C_\alpha r$ from z, and place the remaining αk points at x_* itself. The directional-derivative computation gives equality, so x_* is on the boundary of optimality and $\|x_* - z\|_2 = C_\alpha r$ is achievable.*

Remark 8 (Where this lemma is used)**.** *Lemma 1 is the geometric core of all breakdown-point bounds for the geometric median. It is purely deterministic and contains no probability. We apply it with $z = \mathbf{y}^\star$ and r taken to be a high-probability bound on the radius of the ball containing the majority of the samples; the next subsection (§B.6) provides exactly this bound for sub-Gaussian competent components.*

B.6. Proof of Lemma 2

For convenience, we recall Lemma 2: under Assumptions 1–4, for any slack $\beta \in (0, 1/2-\alpha)$, with probability at least $1 - \exp(-N\beta^2/2)$, at least $(1-\alpha-\beta)N$ of the N judge outputs lie within distance $\rho = \sigma\big(C_1\sqrt{d} + \sqrt{(1/c)\log(2(1-\alpha)/\beta)}\big)$ of $\mathbf{y}^\star$, where $C_1, c > 0$ are absolute constants from the sub-Gaussian-norm tail bound derived in Step 1 below.

Note on heterogeneous parameters. Assumptions 1 and 3 are stated per-judge (α_i and σ_i). Throughout this proof we read α as the global mean contamination $\alpha = (1/N)\sum_i \alpha_i$ from Assumption 4 (which we are entitled to do because Hoeffding in Step 2 only sees $\sum_i \mathbb{E}W_i$; per-judge heterogeneity averages out at this aggregation step), and σ as $\sigma = \max_i \sigma_i$ (the worst-case sub-Gaussian parameter, used in Step 1 to bound every i simultaneously).

The proof is in three stages: (1) control the tail of one competent sample's deviation $\|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2$ at probability p via a covering-net argument; (2) count, via Hoeffding, how many of the N judges fall inside the resulting ball; (3) pick p and a Hoeffding slack u so that the count exceeds $(1-\alpha-\beta)N$ with the claimed probability.

Proof of Lemma 2. Step 1: tail bound for the norm of a single competent sample. For each judge i, write the noise decomposition $\hat{\mathbf{y}}_i = (1-Z_i)(\mathbf{y}^\star + \boldsymbol{\epsilon}_i) + Z_i\,\boldsymbol{\eta}_i$ of Assumption 1, where $Z_i \sim \mathrm{Bern}(\alpha)$ selects competent ($Z_i = 0$) vs. corrupted ($Z_i = 1$). Conditional on $Z_i = 0$, Assumption 3 states that $\boldsymbol{\epsilon}_i \in \mathbb{R}^d$ is σ-sub-Gaussian, i.e. for every $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$\mathbb{E}\big[\exp\big(\langle \boldsymbol{\lambda}, \boldsymbol{\epsilon}_i\rangle\big) \,\big|\, Z_i = 0\big] \;\leq\; \exp\big(\tfrac{1}{2}\sigma^2\|\boldsymbol{\lambda}\|_2^2\big). \tag{23}$$

We now show, from (23) alone,

$$\Pr\big[\|\boldsymbol{\epsilon}_i\|_2 > \sigma(C_1\sqrt{d} + t) \,\big|\, Z_i = 0\big] \;\leq\; \exp(-c\,t^2), \qquad \forall\, t > 0, \tag{24}$$

for absolute constants $C_1, c > 0$.

We prove (24) directly from (23) via a covering-net argument over the unit sphere $\mathbb{S}^{d-1} = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\|_2 = 1\}$. All conditioning is on $\{Z_i = 0\}$; we drop the conditioning bar in this step for readability.

Step 1a: scalar projections are sub-Gaussian. Fix any unit vector $\mathbf{u} \in \mathbb{S}^{d-1}$. Setting $\boldsymbol{\lambda} = \lambda \mathbf{u}$ in (23):

$$\mathbb{E}\big[\exp(\lambda \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle)\big] \;\le\; \exp\big(\tfrac{1}{2}\sigma^2\lambda^2\big), \qquad \forall\, \lambda \in \mathbb{R}. \tag{25}$$

That is, the scalar variable $\langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle$ is σ-sub-Gaussian in $\mathbb{R}$. By Markov's inequality applied to $\exp(\lambda \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle)$:

$$\Pr\big[\langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle > s\big] \;\le\; \exp(-\lambda s + \tfrac{1}{2}\sigma^2\lambda^2), \tag{26}$$

and minimizing the right-hand side over $\lambda > 0$ at $\lambda = s/\sigma^2$ gives the sharp scalar Hoeffding-style bound

$$\Pr\big[\langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle > s\big] \;\le\; \exp\big(-s^2/(2\sigma^2)\big), \qquad \forall\, s > 0. \tag{27}$$

Step 1b: discretize the sphere with a $1/2$-net. Let $\mathcal{N} \subset \mathbb{S}^{d-1}$ be a $1/2$-net of the sphere in Euclidean distance: every $\mathbf{u} \in \mathbb{S}^{d-1}$ is within distance $1/2$ of some $\mathbf{u}' \in \mathcal{N}$. Such a net exists with cardinality

$$|\mathcal{N}| \;\le\; 5^d \tag{28}$$

by a volumetric covering argument (Vershynin, 2018, Cor. 4.2.13: the unit sphere admits an ϵ-net of size $(1 + 2/\epsilon)^d$; take $\epsilon = 1/2$).

Step 1c: net-supremum approximates the true supremum. By definition, $\|\boldsymbol{\epsilon}_i\|_2 = \sup_{\mathbf{u} \in \mathbb{S}^{d-1}} \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle$. Pick the maximizer $\mathbf{u}^*$ and let $\mathbf{u}' \in \mathcal{N}$ satisfy $\|\mathbf{u}^* - \mathbf{u}'\|_2 \le 1/2$. Then

$$\begin{aligned} \langle \mathbf{u}^*, \boldsymbol{\epsilon}_i \rangle \;&=\; \langle \mathbf{u}', \boldsymbol{\epsilon}_i \rangle \;+\; \langle \mathbf{u}^* - \mathbf{u}', \boldsymbol{\epsilon}_i \rangle \\ &\le\; \max_{\mathbf{u} \in \mathcal{N}} \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle \;+\; \tfrac{1}{2}\,\|\boldsymbol{\epsilon}_i\|_2 \end{aligned}$$

where the second line uses Cauchy–Schwarz and $\|\mathbf{u}^* - \mathbf{u}'\|_2 \le 1/2$. Since the left-hand side equals $\|\boldsymbol{\epsilon}_i\|_2$, rearranging gives

$$\|\boldsymbol{\epsilon}_i\|_2 \;\le\; 2 \max_{\mathbf{u} \in \mathcal{N}} \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle. \tag{29}$$

Step 1d: union bound over the net. Combining (29), (27), and (28): for any $r > 0$,

$$\begin{aligned} \Pr\big[\|\boldsymbol{\epsilon}_i\|_2 > 2r\big] \;&\le\; \Pr\Big[\max_{\mathbf{u} \in \mathcal{N}} \langle \mathbf{u}, \boldsymbol{\epsilon}_i \rangle > r\Big] \\ &\le\; |\mathcal{N}| \exp\big(-r^2/(2\sigma^2)\big) \\ &\le\; \exp\big(d \log 5 - r^2/(2\sigma^2)\big). \end{aligned}$$

Substituting $r = \sigma\sqrt{2(d \log 5 + s)}$ for $s > 0$:

$$\Pr\Big[\|\boldsymbol{\epsilon}_i\|_2 > 2\sigma\sqrt{2(d \log 5 + s)}\Big] \;\le\; \exp(-s).$$

Using $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$:

$$2\sigma\sqrt{2(d \log 5 + s)} \;\le\; 2\sigma\sqrt{2d \log 5} + 2\sigma\sqrt{2s} \;=\; C_1\sigma\sqrt{d} + C_2\sigma\sqrt{s},$$

with $C_1 = 2\sqrt{2 \log 5} \le 4$ and $C_2 = 2\sqrt{2}$. Substituting $s = c\,t^2$ with $c = 1/C_2^2 = 1/8$: $C_2\sigma\sqrt{s} = C_2\sigma\sqrt{ct^2} = C_2\sigma\, t/C_2 = \sigma\, t$. Hence for all $t \ge 0$,

$$\Pr\big[\|\boldsymbol{\epsilon}_i\|_2 > C_1\sigma\sqrt{d} + \sigma t\big] \;\le\; \exp(-c\,t^2), \tag{30}$$

which is exactly (24) with the same absolute constants $C_1 = 2\sqrt{2 \log 5} \le 4$ and $c = 1/8$.

Remark on the explicit constants. The covering radius $1/2$, net size 5^d, and resulting prefactor $C_1 = 2\sqrt{2 \log 5}$ are not optimized; sharper chaining bounds (Boucheron et al., 2013, §5.4) reduce C_1 towards 1 at the cost of a more involved proof. For our purposes the order $\sigma(\sqrt{d} + t)$ is what matters, so we proceed with the simpler bound.

Step 1c: solve for the radius at tail probability p. Set $t = \sqrt{(1/c)\log(1/p)}$ in (24); the right-hand side becomes $\exp(-c \cdot (1/c)\log(1/p)) = \exp(\log p) = p$. Defining

$$\rho_p \triangleq \sigma\Big(C_1\sqrt{d} + \sqrt{(1/c)\log(1/p)}\Big), \tag{31}$$

we obtain the per-sample tail bound

$$\Pr\big[\|\boldsymbol{\epsilon}_i\|_2 > \rho_p \,\big|\, Z_i = 0\big] \le p. \tag{32}$$

Step 2: count of judges within ρ_p of $\mathbf{y}^\star$. For each $i \in [N]$ define the indicator

$$W_i \triangleq \mathbb{1}\{Z_i = 0 \text{ and } \|\boldsymbol{\epsilon}_i\|_2 \le \rho_p\}. \tag{33}$$

On $\{W_i = 1\}$ judge i is competent and within distance ρ_p of $\mathbf{y}^\star$ (since $\hat{\mathbf{y}}_i - \mathbf{y}^\star = \boldsymbol{\epsilon}_i$ when $Z_i = 0$); hence

$$\sum_{i=1}^{N} W_i \le \big|\{i \in [N] : \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \le \rho_p\}\big|. \tag{34}$$

The right-hand count is the cluster size we want to lower-bound, so it suffices to lower-bound $\sum W_i$.

Step 2a: marginal mean of W_i. By the tower rule, $\mathbb{E}W_i = \Pr[Z_i = 0]\ \Pr[\|\boldsymbol{\epsilon}_i\|_2 \le \rho_p \mid Z_i = 0]$. Using $\Pr[Z_i = 0] = 1 - \alpha$ from Assumption 1 and (32),

$$\mathbb{E}W_i \ge (1 - \alpha)(1 - p). \tag{35}$$

Step 2b: independence of W_i across i. Each W_i is a measurable function of $(Z_i, \boldsymbol{\epsilon}_i)$ (for $Z_i = 1$, the value of $\boldsymbol{\eta}_i$ does not enter W_i because the indicator forces $Z_i = 0$). By Assumption 2 the tuples $\{(Z_i, \boldsymbol{\epsilon}_i, \boldsymbol{\eta}_i)\}_{i=1}^N$ are mutually independent, hence so are the W_i.

Step 2c: Hoeffding's inequality. Each $W_i \in \{0, 1\} \subseteq [0, 1]$. Hoeffding's inequality (Boucheron et al., 2013, Theorem 2.8) applied to the independent bounded variables W_i states: for any $u > 0$,

$$\Pr\left[\frac{1}{N}\sum_{i=1}^{N} W_i - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}W_i < -u\right] \le \exp(-2Nu^2). \tag{36}$$

Combining (36) with the lower bound (35) on each $\mathbb{E}W_i$:

$$\Pr\left[\sum_{i=1}^{N} W_i < (1 - \alpha)(1 - p)N - uN\right] \le \exp(-2Nu^2). \tag{37}$$

Step 3: choose p and u to expose slack β. We want the lower-bound count $(1-\alpha)(1-p)N - uN$ to be at least $(1-\alpha-\beta)N$:

$$(1 - \alpha)(1 - p) - u \ge 1 - \alpha - \beta \quad \Longleftrightarrow \quad (1 - \alpha)\, p + u \le \beta.$$

Split the slack β equally between the per-sample tail and the Hoeffding deviation by choosing

$$p = \frac{\beta}{2(1 - \alpha)}, \qquad u = \frac{\beta}{2}. \tag{38}$$

Then $(1 - \alpha)p = \beta/2$, so $(1 - \alpha)p + u = \beta$ exactly, verifying the constraint. Substituting $u = \beta/2$ into (37):

$$\Pr\left[\sum_{i=1}^{N} W_i < (1 - \alpha - \beta)N\right] \le \exp\big(-2N(\beta/2)^2\big) = \exp(-N\beta^2/2). \tag{39}$$

Substituting $p = \beta/(2(1 - \alpha))$ into (31), the radius becomes

$$\rho \triangleq \rho_p\big|_{p=\beta/(2(1-\alpha))} = \sigma\left(C_1\sqrt{d} + \sqrt{\tfrac{1}{c}\log\tfrac{2(1-\alpha)}{\beta}}\right),$$

which is exactly (8).

Step 4: assemble the conclusion. On the complementary event of (39), which has probability at least $1 - \exp(-N\beta^2/2)$, the bound $\sum W_i \geq (1 - \alpha - \beta)N$ holds. Combined with (34):

$$\left|\left\{i \in [N] : \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq \rho\right\}\right| \;\geq\; \sum_{i=1}^{N} W_i \;\geq\; (1 - \alpha - \beta)N$$

on the same event. This is (7). ■

On the choice of competent-component assumption. The sub-Gaussian assumption is one of four natural choices for the competent component, ordered from weakest to strongest. Each gives a different cluster-radius bound; sub-Gaussian is the choice that delivers Lemma 2. We record the alternatives for context.

Remark 9 (Just unbiased: insufficient)**.** *If the competent component P_i is only assumed to satisfy $\mathbb{E}_{P_i}[\hat{\mathbf{y}}_i] = \mathbf{y}^\star$ (Proposition 1's clean-case hypothesis), then no quantitative tail bound is available. For arbitrary unbiased P_i, the empirical cluster radius can be arbitrarily large with positive probability, so the hypothesis of Lemma 1 cannot be verified for any finite r. Unbiasedness alone does not suffice to control GM error.*

Remark 10 (Finite variance: polynomial tails)**.** *If the competent component has finite second moment $\mathrm{Var}_{P_i}(\hat{\mathbf{y}}_i) \preceq \sigma^2 I_d$, Chebyshev's inequality gives*

$$\Pr\left[\|\boldsymbol{\epsilon}_i\|_2 > t\sigma\sqrt{d}\right] \;\leq\; 1/t^2.$$

The same Hoeffding argument as in the proof of Lemma 2 then yields a cluster radius of order $\sigma\sqrt{d/\beta}$ rather than the sub-Gaussian $\sigma(\sqrt{d} + \sqrt{\log(1/\beta)})$ — exchanging the exponential dependence on slack for a polynomial one. This regime is where median-of-means (Lugosi & Mendelson, 2019) becomes strictly preferable to plain GM for sub-Gaussian rates.

Remark 11 (Sub-Gaussian: our main assumption)**.** *Lemma 2 uses Assumption 3 (σ-sub-Gaussian competent component). This delivers a cluster radius $\rho = \sigma(\sqrt{d} + O(\sqrt{\log(1/\beta)}))$, with exponential dependence on the slack β. The sub-Gaussian assumption is the standard middle ground in robust statistics: weaker than bounded support but strong enough to give exponential concentration of the cluster.*

Remark 12 (Bounded support: deterministic, automatic for LLM scores)**.** *If the competent component is supported on $[0, K]^d$ (equivalently, $\hat{\mathbf{y}}_i$ takes values in the score hypercube), then $\|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq K\sqrt{d}$* deterministically *for every competent sample. Lemma 2 then holds with $\rho = K\sqrt{d}$* without any probabilistic event *and with the slack β needed only to absorb the bound $|S| \geq (1 - \alpha - \beta)N$ on the competent-set size.*

For the LLM-jury setting, scores are produced by a parser with codomain $[0, K]^d$, so bounded support is a given*, not an additional assumption. However, $K\sqrt{d}$ is a worst-case radius (the diameter of the hypercube) and is typically much larger than the sub-Gaussian cluster radius $\sigma\sqrt{d}$ that Assumption 3 delivers, since real LLM judges have $\sigma \ll K$ in practice (§C.3). The sub-Gaussian bound is therefore tighter in the regime that matters; bounded support serves as a universally-valid fallback.*

B.7. Proof of Theorem 1

For convenience, we recall Theorem 1: under Assumptions 1–4, fix any slack $\beta \in (0, 1/2 - \alpha)$; with probability at least $1 - \exp(-N\beta^2/2)$, $\|\hat{\mathbf{y}}_{\mathrm{GM}} - \mathbf{y}^\star\|_2 \leq C_{\alpha+\beta}\,\rho$, where $\hat{\mathbf{y}}_{\mathrm{GM}}$ is any geometric median of the N judge outputs (Definition 4), $C_{\alpha+\beta} = (1 - \alpha - \beta)/\sqrt{1 - 2(\alpha + \beta)}$ is the geometric-breakdown constant of Lemma 1 evaluated at $\alpha + \beta$, and $\rho = \sigma(C_1\sqrt{d} + \sqrt{(1/c)\log(2(1-\alpha)/\beta)})$ is the cluster radius of Lemma 2 ($C_1, c > 0$ absolute constants).

Proof of Theorem 1. The proof is a clean composition of the deterministic geometric bound (Lemma 1) and the probabilistic cluster-radius bound (Lemma 2). We make the composition fully explicit.

Step 1: define the cluster event. Let $\beta \in (0, 1/2 - \alpha)$ be the slack from the theorem statement. Let $\rho = \sigma(C_1\sqrt{d} + \sqrt{(1/c)\log(2(1-\alpha)/\beta)})$ be the cluster radius from (8), and define the event

$$\mathcal{E} \;\triangleq\; \left\{\,|J| \;\geq\; (1 - \alpha - \beta)N\,\right\}, \qquad J \;\triangleq\; \left\{i \in [N] : \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq \rho\right\}. \tag{40}$$

By Lemma 2 applied with this slack β,

$$\Pr[\mathcal{E}] \;\geq\; 1 - \exp(-N\beta^2/2). \tag{41}$$

The remainder of the proof works on $\mathcal{E}$ (sample-pathwise); no further probability is incurred.

Step 2: verify the hypothesis of Lemma 1. On $\mathcal{E}$, we apply Lemma 1 with the substitutions

$$k \leftarrow N, \qquad z \leftarrow \mathbf{y}^\star, \qquad r \leftarrow \rho, \qquad \alpha \leftarrow \alpha + \beta. \tag{42}$$

The hypothesis of Lemma 1 (in its statement form: "at least $(1-\alpha)\,k$ of the k points lie within distance r of z") becomes, under these substitutions,

$$|J| \;\geq\; \big(1-(\alpha+\beta)\big)N \;=\; (1-\alpha-\beta)N,$$

which is exactly the definition of $\mathcal{E}$. The range condition $\alpha+\beta \in (0, 1/2)$ holds since $\alpha > 0$ (by Assumption 1's $\alpha_i \geq 0$ and $\beta > 0$) and $\alpha + \beta < 1/2$ (by $\beta < 1/2 - \alpha$).

Step 3: apply Lemma 1 and read off the bound. The conclusion of Lemma 1 under the substitutions (42) is

$$\|x_* - z\|_2 \;\leq\; C_{\alpha+\beta}\, r \quad \text{i.e.} \quad \|\hat{\mathbf{y}}_{\mathrm{GM}} - \mathbf{y}^\star\|_2 \;\leq\; C_{\alpha+\beta}\,\rho,$$

where $C_{\alpha+\beta} = (1-\alpha-\beta)/\sqrt{1-2(\alpha+\beta)}$ and $x_* = \hat{\mathbf{y}}_{\mathrm{GM}}$ is any minimizer of $z \mapsto \sum_{i=1}^N \|z - \hat{\mathbf{y}}_i\|_2$ (the geometric median). Lemma 1 as proved in §B.5 applies to *any* minimizer, so the conclusion is independent of any choice in the (collinear) case where the GM is non-unique.

Step 4: assemble. Combining (41) with the deterministic bound on $\mathcal{E}$ from Step 3:

$$\Pr\Big[\, \|\hat{\mathbf{y}}_{\mathrm{GM}} - \mathbf{y}^\star\|_2 \;\leq\; \underbrace{\frac{1-\alpha-\beta}{\sqrt{1-2(\alpha+\beta)}}}_{C_{\alpha+\beta}} \cdot \underbrace{\sigma\left(C_1\sqrt{d} + \sqrt{\tfrac{1}{c}\log\tfrac{2(1-\alpha)}{\beta}}\right)}_{\rho} \,\Big] \;\geq\; 1 - \exp(-N\beta^2/2),$$

which is exactly (9). ∎

Remark 13 (Choice of slack β)**.** *The slack β trades two terms in (9): the geometric constant $C_{\alpha+\beta}$ grows with β (since β erodes the safety margin to the breakdown point $1/2$), while the cluster radius ρ shrinks with β (since a larger slack absorbs more competent samples, allowing a smaller per-sample tail). For deployment, β should be chosen to minimise the right-hand side of (9). A practical default is $\beta = (1/2-\alpha)/2$ (half the safety margin), which keeps $C_{\alpha+\beta}$ bounded by a small constant while permitting an exponentially small failure probability for any $N \gtrsim 1/\beta^2$.*

Remark 14 (The bound does not vanish with N)**.** *Unlike the (incorrect) original Theorem 1, which claimed an upper bound of order $\sigma\sqrt{d/N}/(1-2\alpha)$, the bound in (9) contains no $1/\sqrt{N}$ factor in the leading term: the cluster radius ρ is $\sigma\sqrt{d}$ in scale (up to a $\sqrt{\log(1/\beta)}$ factor), and $C_{\alpha+\beta}$ depends only on the contamination rate. This reflects the breakdown-point character of plain GM: under arbitrary Q in the Huber class, the asymptotic-N floor is set by the cluster radius, not by sample averaging. The empirical validation in §C.3 (forthcoming experiment confirming the floor) agrees with this prediction; the original $1/\sqrt{N}$ claim was empirically inconsistent with the observed plateau.*

Remark 15 (Comparison with the minimax lower bound)**.** *The minimax lower bound (Theorem 2) gives $\Omega(\sigma(\sqrt{d/N} + \alpha/(1-\alpha)))$. The clean-rate term $\sqrt{d/N}$ matches the upper bound exactly. On the breakdown floor the upper bound scales as $C_{\alpha+\beta}\sigma\sqrt{d}$ while the lower bound scales as $\sigma\alpha/(1-\alpha)$, leaving a gap of order $\sqrt{d}/\alpha$. The reason is structural: total variation between two equal-covariance Gaussians is dimension-free (Step 2.2 of the proof of Theorem 2), so the modulus of continuity of the Huber neighborhood does not gain a $\sqrt{d}$ factor in higher dimensions — and indeed Chen et al. (2018), Theorem 5.1, establish $\Theta(\sigma^2(d/N + \alpha^2))$ as the squared-error minimax for sub-Gaussian Huber, with no d in the contamination term. The $\sqrt{d}$ in the upper bound comes from the geometric median's cluster radius (Lemma 2), reflecting the price plain GM pays for $O(Nd\log(1/\epsilon))$ tractability relative to the (intractable) Tukey halfspace median or the (sub-exponential) smoothed-depth estimator. For LLM-jury parameters the gap is small (at most $\sim 2.2\times$ for $d \leq 5$).*

Remark 16 (Bounded-support specialization)**.** *If competent scores are bounded in $[0, K]^d$ (Remark 12), the cluster radius ρ in (9) can be replaced by the deterministic worst-case $K\sqrt{d}$, removing the $\sqrt{(1/c)\log(2(1-\alpha)/\beta)}$ term and the high-probability event for the cluster. The slack β remains needed to control the empirical competent-set size $|S|$ (the Hoeffding step in Lemma 2's proof), but the per-sample tail event becomes deterministic. For typical LLM-jury parameters ($\sigma \ll K$), the sub-Gaussian form (9) is tighter and is what we use throughout.*

B.8. Proof of Lemma 3

For convenience we recall Lemma 3: under Assumptions 1, 3, 4 and the equicorrelated-indicator assumption (replacing Asm. 2) $\mathrm{Cov}(W_i, W_j) \le \bar{\gamma}_W \sqrt{\mathrm{Var}(W_i)\mathrm{Var}(W_j)}$ for $i \neq j$, with $\bar{\gamma}_W \in [0,1]$, the RoPoLL bound $\|\hat{\mathbf{y}}_{\mathrm{GM}} - \mathbf{y}^\star\|_2 \le C_{\alpha+\beta}\rho$ holds with probability at least $1 - 1/(\beta^2 N_{\mathrm{eff}})$, where $N_{\mathrm{eff}} = N/(1 + (N-1)\bar{\gamma}_W)$.

The proof follows the same skeleton as Lemma 2 (per-sample tail → count-bound) combined with Lemma 1 (deterministic geometric step), but replaces the Hoeffding count-bound (which required independence) with a Chebyshev count-bound on the variance of $\sum_i W_i$ under the bounded-covariance hypothesis. The deterministic geometric step (Lemma 1) and the per-sample sub-Gaussian tail (Step 1 of Lemma 2's proof) are correlation-free and carry through unchanged.

Proof of Lemma 3. Step 1: marginal mean of each indicator. The indicator $W_i = \mathbf{1}\{Z_i = 0,\ \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \le \rho_p\}$ factors as $W_i = \mathbf{1}\{Z_i = 0\} \cdot \mathbf{1}\{\|\boldsymbol{\epsilon}_i\|_2 \le \rho_p\}$ (when $Z_i = 0$ we have $\hat{\mathbf{y}}_i - \mathbf{y}^\star = \boldsymbol{\epsilon}_i$; when $Z_i = 1$ the first indicator forces $W_i = 0$ regardless of $\boldsymbol{\eta}_i$, so $\boldsymbol{\eta}_i$ does not enter W_i). By the tower rule,

$$\begin{aligned}\mathbb{E}W_i &= \Pr[Z_i = 0] \cdot \Pr[\|\boldsymbol{\epsilon}_i\|_2 \le \rho_p \mid Z_i = 0] \\ &\ge (1 - \alpha_i)(1 - p),\end{aligned}$$

using $\Pr[Z_i = 0] = 1 - \alpha_i$ from Assumption 1 and the per-sample tail $\Pr[\|\boldsymbol{\epsilon}_i\|_2 \le \rho_p \mid Z_i = 0] \ge 1 - p$ from Step 1 of Lemma 2's proof (which is correlation-free). Summing over i and using $\alpha = N^{-1}\sum_i \alpha_i$ (Assumption 4):

$$\mu_N \ \triangleq\ \sum_{i=1}^{N} \mathbb{E}W_i \ \ge\ \sum_{i=1}^{N}(1 - \alpha_i)(1 - p) \ =\ N(1 - \alpha)(1 - p). \tag{43}$$

Step 2: variance of each indicator. Each $W_i \in \{0, 1\}$ is Bernoulli, so

$$\mathrm{Var}(W_i) \ =\ \mathbb{E}W_i\,(1 - \mathbb{E}W_i) \ \le\ \tfrac{1}{4}, \tag{44}$$

where the inequality is the Bernoulli-variance bound ($x(1-x) \le 1/4$ on $[0,1]$, attained at $x = 1/2$).

Step 3: pairwise covariance bound. The hypothesis (equicorrelated indicators) gives, for $i \neq j$,

$$\mathrm{Cov}(W_i, W_j) \ \le\ \bar{\gamma}_W \sqrt{\mathrm{Var}(W_i)\mathrm{Var}(W_j)} \ \le\ \tfrac{\bar{\gamma}_W}{4}, \tag{45}$$

where the second inequality combines (44) on both factors.

Step 4: variance of the count. By definition of variance for sums,

$$\mathrm{Var}\left(\sum_{i=1}^{N} W_i\right) \ =\ \sum_{i=1}^{N} \mathrm{Var}(W_i) + \sum_{i \neq j} \mathrm{Cov}(W_i, W_j).$$

There are N diagonal terms and $N(N-1)$ off-diagonal terms. Substituting (44) on the diagonal and (45) off-diagonal:

$$\begin{aligned}\mathrm{Var}\left(\sum_i W_i\right) \ &\le\ N \cdot \tfrac{1}{4} + N(N-1) \cdot \tfrac{\bar{\gamma}_W}{4} \\ &=\ \tfrac{N}{4}\big(1 + (N-1)\bar{\gamma}_W\big) \ =\ \tfrac{N^2}{4N_{\mathrm{eff}}},\end{aligned} \tag{46}$$

where the last equality uses $N_{\mathrm{eff}} = N/(1 + (N-1)\bar{\gamma}_W)$. Sanity check: at $\bar{\gamma}_W = 0$ (independence), $N_{\mathrm{eff}} = N$ and $\mathrm{Var}(\sum_i W_i) \le N/4$, the standard Bernoulli-sum variance. At $\bar{\gamma}_W = 1$ (perfect correlation), $N_{\mathrm{eff}} = 1$ and $\mathrm{Var}(\sum_i W_i) \le N^2/4$, matching the case $W_1 = \cdots = W_N$ where $\mathrm{Var}(\sum_i W_i) = N^2\mathrm{Var}(W_1) \le N^2/4$.

Step 5: lower-deviation Chebyshev inequality. For any random variable X with finite variance and any $u > 0$,

$$\Pr[X \le \mathbb{E}X - uN] \ \le\ \Pr[|X - \mathbb{E}X| \ge uN] \ \le\ \frac{\mathrm{Var}(X)}{(uN)^2},$$

by Chebyshev's inequality applied to the deviation $|X - \mathbb{E}X|$. Applying this to $X = \sum_i W_i$ with mean μ_N and using (46):

$$\Pr\left[\sum_i W_i \le \mu_N - uN\right] \;\le\; \frac{\mathrm{Var}(\sum_i W_i)}{(uN)^2} \;\le\; \frac{N^2/(4N_{\text{eff}})}{u^2N^2} \;=\; \frac{1}{4u^2N_{\text{eff}}}. \tag{47}$$

Step 6: calibrate p and u to the slack β. We want the deviation event in (47) to imply the failure of the cluster bound $\sum_i W_i \ge (1-\alpha-\beta)N$. By (43), $\mu_N - uN \ge (1-\alpha)(1-p)N - uN = ((1-\alpha)(1-p) - u)N$. We require $(1-\alpha)(1-p) - u \ge 1-\alpha-\beta$, which rearranges to

$$(1-\alpha)\,p \;+\; u \;\le\; \beta.$$

This is exactly the constraint that appeared in Lemma 2's Step 3. Splitting β equally between the per-sample tail p and the count-deviation u, choose

$$p \;=\; \frac{\beta}{2(1-\alpha)}, \qquad u \;=\; \frac{\beta}{2}. \tag{48}$$

Then $(1-\alpha)p = \beta/2$ and $u = \beta/2$, summing to β exactly.

Step 7: failure-probability bound. Substituting $u = \beta/2$ from (48) into (47):

$$\Pr\left[\sum_i W_i \le \mu_N - (\beta/2)N\right] \;\le\; \frac{1}{4(\beta/2)^2N_{\text{eff}}} \;=\; \frac{1}{\beta^2N_{\text{eff}}}.$$

By the calibration of Step 6, $\mu_N - (\beta/2)N \ge (1-\alpha-\beta)N$, so

$$\Pr\left[\sum_{i=1}^{N} W_i < (1-\alpha-\beta)N\right] \;\le\; \frac{1}{\beta^2N_{\text{eff}}}, \tag{49}$$

which is (10).

Step 8: cluster radius (unchanged from Lemma 2). Substituting $p = \beta/(2(1-\alpha))$ from (48) into the per-sample tail-radius $\rho_p = \sigma(C_1\sqrt{d} + \sqrt{(1/c)\log(1/p)})$ (equation (31) of Lemma 2's proof) gives the same cluster radius as Theorem 1:

$$\rho \;=\; \sigma\left(C_1\sqrt{d} \;+\; \sqrt{\tfrac{1}{c}\log\tfrac{2(1-\alpha)}{\beta}}\right).$$

This step uses only the per-sample sub-Gaussian tail and is correlation-free.

Step 9: deterministic geometric step (Lemma 1). On the complementary event of (49), which has probability $\ge 1 - 1/(\beta^2N_{\text{eff}})$, the count of cluster-near judges satisfies

$$\left|\{i \in [N] : \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \le \rho\}\right| \;\ge\; \sum_{i=1}^{N} W_i \;\ge\; (1-\alpha-\beta)N \;=\; (1-(\alpha+\beta))N.$$

Apply Lemma 1 with the substitutions $k = N$, $z = \mathbf{y}^\star$, $r = \rho$, $\alpha' = \alpha + \beta$ (which lies in $(0, 1/2)$ since $\beta \in (0, 1/2 - \alpha)$); the lemma's hypothesis is exactly the count bound above. The conclusion gives $\|\hat{\mathbf{y}}_{\text{GM}} - \mathbf{y}^\star\|_2 \le C_{\alpha+\beta}\,\rho$ with $C_{\alpha+\beta}$ and ρ unchanged from Theorem 1.

Step 10: assemble. Combining the deterministic bound from Step 9 (which holds on the complementary event) with the failure probability (49):

$$\Pr\Big[\,\|\hat{\mathbf{y}}_{\text{GM}} - \mathbf{y}^\star\|_2 \;\le\; C_{\alpha+\beta}\,\rho\,\Big] \;\ge\; 1 - \frac{1}{\beta^2N_{\text{eff}}},$$

which is the statement of Lemma 3. ■

Remark 17 (Polynomial vs. exponential tail)**.** *The price of allowing correlation is the tail rate. The independent Hoeffding bound gives an* exponential *probability event* $\Pr[\cdot] \le \exp(-N\beta^2/2)$*; the correlated Chebyshev bound is* polynomial *in* N_{eff}*,* $\Pr[\cdot] \le 1/(\beta^2N_{\text{eff}})$*. At independence ($\bar{\gamma}_W = 0$), $N_{\text{eff}} = N$ and both apply, but Hoeffding is strictly tighter. A Bernstein-type bound under bounded-covariance martingale structure (e.g., via the Efron–Stein inequality for sums of weakly-dependent Bernoulli variables) can recover sub-exponential rates under stronger hypotheses on the dependence graph; we do not pursue this here as the polynomial bound suffices for the parameter regime ($N_{\text{eff}} \approx 1.5$–$2$, $\beta \approx 0.1$–0.2) of our experiments.*

Remark 18 (Estimating $\bar{\gamma}_W$ from data)**.** *The hypothesis of Lemma 3 is on the indicator correlation $\bar{\gamma}_W$, which is in principle a finer object than the inter-judge* score *correlation $\bar{\gamma}$ measured in Figures 4a and 21. For jointly Gaussian competent noise with positive score correlation, the cluster indicators are positively associated by Pitt's Gaussian correlation inequality (Pitt, 1977; Esary et al., 1967; Joag-Dev & Proschan, 1983), so $\bar{\gamma}_W \geq 0$; we are not aware of a clean general upper bound on $\bar{\gamma}_W$ in terms of $\bar{\gamma}$ alone. In practice, $\bar{\gamma}_W$ can be estimated directly from data as the empirical correlation of the cluster events $\{W_i = 1\}$ across instances; on our experimental grid this empirical value is on the same order as the score correlation $\bar{\gamma}$, supporting the $N_{\text{eff}} \approx 1.5$–$2$ regime quoted in the body.*

B.9. Proof of Theorem 2

Theorem 1 provides an upper bound on the error of the geometric median. A natural question is whether this rate can be improved by *any* estimator. The following result shows that, in the parametric regime, it cannot.

For convenience we restate Theorem 2: under the observation model (2) with N judges in $\mathbb{R}^d$, homogeneous contamination rate $\alpha < 1/2$, and σ^2-sub-Gaussian competent noise (Assumptions 1, 2, 3, 4), there exists a universal constant $c > 0$ such that

$$\inf_{\hat{\mathbf{y}}} \sup_{F \in \mathcal{F}_{\alpha,\sigma}} \mathbb{E}_F\big[\big\|\hat{\mathbf{y}} - \mathbf{y}^\star\big\|_2\big] \;\geq\; c\,\sigma\left(\sqrt{d/N} \;+\; \frac{\alpha}{1-\alpha}\right). \tag{50}$$

Proof of Theorem 2. We invoke Le Cam's two-point method (Tsybakov, 2009, Sec. 2.4): for any two parameter values $\mathbf{y}_0, \mathbf{y}_1 \in \mathbb{R}^d$ inducing observation distributions $F_0, F_1 \in \mathcal{F}_{\alpha,\sigma}$,

$$\inf_{\hat{\mathbf{y}}} \sup_{F \in \{F_0, F_1\}} \mathbb{E}_F[\|\hat{\mathbf{y}} - \mathbf{y}^\star\|_2] \;\geq\; \frac{\|\mathbf{y}_0 - \mathbf{y}_1\|_2}{4} \cdot \big(1 - \mathrm{TV}(F_0^{\otimes N}, F_1^{\otimes N})\big)\,. \tag{51}$$

The strategy is to construct $(\mathbf{y}_0, \mathbf{y}_1, F_0, F_1)$ maximising the right-hand side. Part 1 controls the parametric variance term; Part 2 establishes the N-independent breakdown floor via the modulus of continuity of the Huber neighborhood.

Part 1: the $\sqrt{d/N}$ term, via Fano's inequality. Set $\alpha = 0$ and consider the clean Gaussian sub-family $F = \mathcal{N}(\mathbf{y}^\star, \sigma^2 \mathbf{I}_d)$ for all $i \in [N]$. The Le Cam two-point bound (51) alone cannot deliver the $\sqrt{d}$ factor (two Gaussians at separation Δ have $\mathrm{TV} \to 1$ once $\Delta \gtrsim \sigma$, regardless of d); we therefore use the multi-hypothesis generalization, Fano's inequality.

Step 1.1 (Gilbert–Varshamov packing of $\mathbb{R}^d$). For radius $\Delta > 0$, by the Gilbert–Varshamov bound (Massart, 2007, Lem. 4.7) there exists a packing $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\} \subset \mathbb{R}^d$ with

$$\|\mathbf{y}_m - \mathbf{y}_{m'}\|_2 \;\geq\; \Delta \quad \text{for all } m \neq m', \qquad M \;\geq\; 2^{d/8}. \tag{52}$$

(Concretely, take the packing scaled so each $\mathbf{y}_m$ has $\|\mathbf{y}_m\|_2 \leq \Delta$.)

Step 1.2 (Fano's inequality). Let H_m be the hypothesis $\mathbf{y}^\star = \mathbf{y}_m$; under H_m, the joint observation law is $F_m^{\otimes N} = \mathcal{N}(\mathbf{y}_m, \sigma^2 \mathbf{I}_d)^{\otimes N}$. Fano's inequality (Tsybakov, 2009, Cor. 2.6) gives, for any estimator $\hat{\mathbf{y}}$,

$$\frac{1}{M} \sum_{m=1}^{M} \Pr_{H_m}\big[\|\hat{\mathbf{y}} - \mathbf{y}_m\|_2 \geq \Delta/2\big] \;\geq\; 1 - \frac{\bar{\mathrm{KL}} + \log 2}{\log M}, \tag{53}$$

where $\bar{\mathrm{KL}} = \binom{M}{2}^{-1} \sum_{m<m'} \mathrm{KL}(F_m^{\otimes N} \,\|\, F_{m'}^{\otimes N})$. For two product Gaussians, $\mathrm{KL}(F_m^{\otimes N} \,\|\, F_{m'}^{\otimes N}) = N\|\mathbf{y}_m - \mathbf{y}_{m'}\|_2^2/(2\sigma^2) \leq N\Delta^2/(2\sigma^2)$ (using $\|\mathbf{y}_m\|_2 \leq \Delta$ and the triangle inequality).

Step 1.3 (Choose Δ to make the right-hand side $\geq 1/2$). With $\log M \geq d \log 2/8$ and $\bar{\mathrm{KL}} \leq N\Delta^2/(2\sigma^2)$, the right-hand side of (53) is at least $1/2$ provided

$$\frac{N\Delta^2/(2\sigma^2) + \log 2}{d \log 2/8} \;\leq\; \tfrac{1}{2},$$

which (for $d \geq 16$, harmlessly absorbing the $\log 2$) holds when $\Delta = c_1 \sigma \sqrt{d/N}$ for a sufficiently small absolute constant $c_1 > 0$.

Step 1.4 (Convert to expected error). On the event $\|\hat{\mathbf{y}} - \mathbf{y}_m\|_2 \geq \Delta/2$, Markov's inequality gives $\mathbb{E}\|\hat{\mathbf{y}} - \mathbf{y}_m\|_2 \geq (\Delta/2) \cdot \Pr[\cdot] \geq \Delta/4$, so

$$\sup_m \mathbb{E}_{H_m}\|\hat{\mathbf{y}} - \mathbf{y}_m\|_2 \;\geq\; \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{H_m}\|\hat{\mathbf{y}} - \mathbf{y}_m\|_2 \;\geq\; \Delta/4 \;=\; \tfrac{c_1}{4}\sigma\sqrt{d/N}.$$

Each H_m corresponds to a clean ($\alpha = 0$) instance in $\mathcal{F}_{\alpha,\sigma}$, so this lower bound holds over the worst-case $F \in \mathcal{F}_{\alpha,\sigma}$, establishing the $\sqrt{d/N}$ term.

Part 2: the $\alpha/(1-\alpha)$ term, via the modulus of continuity. The breakdown floor is dimension-free in d and *independent of* N; we establish it through the structural fact that two Huber neighborhoods at sufficiently close centers have a common element, hence are statistically indistinguishable.

Step 2.1 (Modulus of continuity for Huber neighborhoods). For a center $\mathbf{y} \in \mathbb{R}^d$, write $\mathcal{F}_\alpha(\mathbf{y}) = \{(1-\alpha)\mathcal{N}(\mathbf{y}, \sigma^2\mathbf{I}_d) + \alpha Q : Q \text{ probability on } \mathbb{R}^d\}$ for the corresponding Huber contamination class. We claim a sufficient condition for two such neighborhoods to overlap:

$$\big\|\mathcal{N}(\mathbf{y}_0, \sigma^2\mathbf{I}_d) - \mathcal{N}(\mathbf{y}_1, \sigma^2\mathbf{I}_d)\big\|_{\mathrm{TV}} \;\leq\; \frac{\alpha}{1-\alpha} \quad\Longrightarrow\quad \mathcal{F}_\alpha(\mathbf{y}_0) \cap \mathcal{F}_\alpha(\mathbf{y}_1) \neq \emptyset. \tag{54}$$

Proof of (54). Let $P_j = \mathcal{N}(\mathbf{y}_j, \sigma^2\mathbf{I}_d)$ and write $\epsilon = \|P_0 - P_1\|_{\mathrm{TV}}$; the hypothesis is $\epsilon \leq \alpha/(1-\alpha)$. Hahn-decompose the signed measure $P_0 - P_1 = \mu^+ - \mu^-$ with $\mu^+, \mu^- \geq 0$ and $\mu^+(\mathbb{R}^d) = \mu^-(\mathbb{R}^d) = \epsilon$. Pick any probability measure ρ (e.g. $\rho = (P_0 + P_1)/2$), and define the candidates

$$\alpha Q_0 \;\triangleq\; (1-\alpha)\mu^- + \big(\alpha - (1-\alpha)\epsilon\big)\rho, \qquad \alpha Q_1 \;\triangleq\; (1-\alpha)\mu^+ + \big(\alpha - (1-\alpha)\epsilon\big)\rho. \tag{55}$$

Each Q_j is a probability measure: nonnegativity holds because $\mu^\pm \geq 0$, $\rho \geq 0$, and the hypothesis $\epsilon \leq \alpha/(1-\alpha)$ ensures $\alpha - (1-\alpha)\epsilon \geq 0$; total mass is $\alpha Q_j(\mathbb{R}^d) = (1-\alpha)\epsilon + (\alpha - (1-\alpha)\epsilon) = \alpha$, so $Q_j(\mathbb{R}^d) = 1$. Subtracting the two Huber mixtures:

$$\begin{aligned}\big[(1-\alpha)P_0 + \alpha Q_0\big] - \big[(1-\alpha)P_1 + \alpha Q_1\big] &= (1-\alpha)(P_0 - P_1) + \alpha(Q_0 - Q_1)\\ &= (1-\alpha)(\mu^+ - \mu^-) + (1-\alpha)(\mu^- - \mu^+)\\ &= 0,\end{aligned}$$

using (55) (the ρ terms cancel). Hence $(1-\alpha)P_0 + \alpha Q_0 = (1-\alpha)P_1 + \alpha Q_1$ is a common element of $\mathcal{F}_\alpha(\mathbf{y}_0) \cap \mathcal{F}_\alpha(\mathbf{y}_1)$, establishing (54).

Step 2.2 (Equal-covariance Gaussian TV is dimension-free). The total-variation distance between $\mathcal{N}(\mathbf{y}_0, \sigma^2\mathbf{I}_d)$ and $\mathcal{N}(\mathbf{y}_1, \sigma^2\mathbf{I}_d)$ depends only on $\Delta \triangleq \|\mathbf{y}_0 - \mathbf{y}_1\|_2$: projecting onto the line $\mathbf{y}_1 - \mathbf{y}_0$ reduces the comparison to two univariate Gaussians at separation Δ with variance σ^2, and the orthogonal directions contribute identical factors that cancel in TV. Therefore

$$\big\|\mathcal{N}(\mathbf{y}_0, \sigma^2\mathbf{I}_d) - \mathcal{N}(\mathbf{y}_1, \sigma^2\mathbf{I}_d)\big\|_{\mathrm{TV}} \;=\; 2\Phi\big(\tfrac{\Delta}{2\sigma}\big) - 1, \tag{56}$$

with Φ the standard normal cdf.

Step 2.3 (Solve for the indistinguishability separation). Combining (54) and (56), $\mathcal{F}_\alpha(\mathbf{y}_0) \cap \mathcal{F}_\alpha(\mathbf{y}_1) \neq \emptyset$ whenever

$$2\Phi(\Delta/(2\sigma)) - 1 \;\leq\; \alpha/(1-\alpha), \qquad \text{i.e.} \qquad \Delta \;\leq\; \Delta_\star \;\triangleq\; 2\sigma\,\Phi^{-1}\Big(\tfrac{1}{2} + \tfrac{\alpha}{2(1-\alpha)}\Big).$$

We lower-bound Φ^{-1} by integrating its density: for any $y \in [0, 1/2)$ and $x = \Phi^{-1}(1/2 + y) \geq 0$,

$$y \;=\; \Phi(x) - \tfrac{1}{2} \;=\; \int_0^x \phi(t)\,dt \;\leq\; x \cdot \max_{t \geq 0}\phi(t) \;=\; x \cdot \phi(0) \;=\; \frac{x}{\sqrt{2\pi}},$$

where the maximum of the standard normal density on $[0, \infty)$ is attained at 0 with $\phi(0) = 1/\sqrt{2\pi}$. Hence $\Phi^{-1}(1/2 + y) \geq y\sqrt{2\pi}$ for all $y \in [0, 1/2)$. Applying this with $y = \alpha/(2(1-\alpha))$ (which lies in $[0, 1/2)$ for all $\alpha \in [0, 1/2)$):

$$\Phi^{-1}\Big(\tfrac{1}{2} + \tfrac{\alpha}{2(1-\alpha)}\Big) \;\geq\; \tfrac{\alpha}{2(1-\alpha)} \cdot \sqrt{2\pi} \;=\; \sqrt{\tfrac{\pi}{2}}\,\frac{\alpha}{1-\alpha}.$$

Therefore

$$\Delta_\star \;=\; 2\sigma\,\Phi^{-1}\!\Big(\tfrac{1}{2}+\tfrac{\alpha}{2(1-\alpha)}\Big) \;\geq\; 2\sigma\cdot\sqrt{\tfrac{\pi}{2}}\,\frac{\alpha}{1-\alpha} \;=\; \sqrt{2\pi}\,\sigma\,\frac{\alpha}{1-\alpha}.$$

Step 2.4 (Apply Le Cam). Pick $\mathbf{y}_0 = \mathbf{0}$, $\mathbf{y}_1 = \Delta_\star\,\mathbf{e}_1$, and let F be any common element of $\mathcal{F}_\alpha(\mathbf{y}_0)\cap\mathcal{F}_\alpha(\mathbf{y}_1)$ (which exists by Step 2.1). Set $F_0 = F_1 = F$; then $F_0^{\otimes N} = F_1^{\otimes N}$ and $\mathrm{TV}(F_0^{\otimes N}, F_1^{\otimes N}) = 0$ *regardless of* N. Substituting into (51),

$$\inf_{\hat{\mathbf{y}}}\sup_{F\in\{F_0,F_1\}}\mathbb{E}_F[\|\hat{\mathbf{y}}-\mathbf{y}^\star\|_2] \;\geq\; \frac{\Delta_\star}{4} \;\geq\; \frac{\sqrt{2\pi}}{4}\,\sigma\,\frac{\alpha}{1-\alpha}.$$

Combining. Taking the maximum of the two lower bounds (the worst-case adversary selects whichever construction is tighter) and absorbing constants yields (50). ■

Remark 19 (Why no $\sqrt{d}$ on the breakdown floor)**.** *A natural question is whether the breakdown term should scale with $\sqrt{d}$ (analogous to the variance term). The answer is no. Total variation between two equal-covariance Gaussians depends only on their ℓ_2 separation (56), not on the ambient dimension; the modulus of continuity is therefore dimension-free. A Fano-style packing of 2^d test points at pairwise overlapping Huber neighborhoods would require pairwise ℓ_2 separation $\leq \Delta_\star$ and pairwise distance large enough to give the desired $\sqrt{d}$ minimax error—these constraints are incompatible, since the diameter of a set of 2^d points at pairwise distance $\leq \Delta_\star$ cannot exceed $\Delta_\star$. This matches the established minimax for sub-Gaussian Huber: Chen et al. (2018), Theorem 5.1, prove* $\inf\sup\mathbb{E}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2 \asymp \sigma^2(d/N+\alpha^2)$*, with no d on the squared-error contamination floor.*

Comparison with the upper bound. At $\alpha = 0$, the upper and lower bounds match at the parametric rate $\sigma\sqrt{d/N}$, confirming that the geometric median is rate-optimal in the clean regime. On the breakdown floor the upper bound (Thm 1) scales as $C_\alpha\sigma\sqrt{d}$ while the lower bound scales as $\sigma\alpha/(1-\alpha)$; the gap is a $\sqrt{d}/\alpha$ factor. This is not slack in the analysis but a real statistical–computational gap. The minimax-optimal estimator on the breakdown floor is the Tukey halfspace median (Tukey, 1975; Donoho & Gasko, 1992), whose exact computation is NP-hard for $d \geq 3$ (Johnson & Preparata, 1978; Aloupis, 2006); the smoothed-depth estimator of Chen et al. (2018) matches the $\sigma\alpha$ floor in sub-exponential time. The geometric median is the polynomial-time alternative: it shares the optimal $1/2$ breakdown point but pays a $\sqrt{d}$ price for $O(Nd\log(1/\epsilon))$ tractability via the Weiszfeld iteration. For LLM juries the trade is favorable: d is small (1–5 in our benchmarks) so the $\sqrt{d}$ overhead is at most $\sim 2.2\times$, and at small N the variance term $\sigma\sqrt{d/N}$ dominates the breakdown floor on every regime we test.

C. Additional Experiments

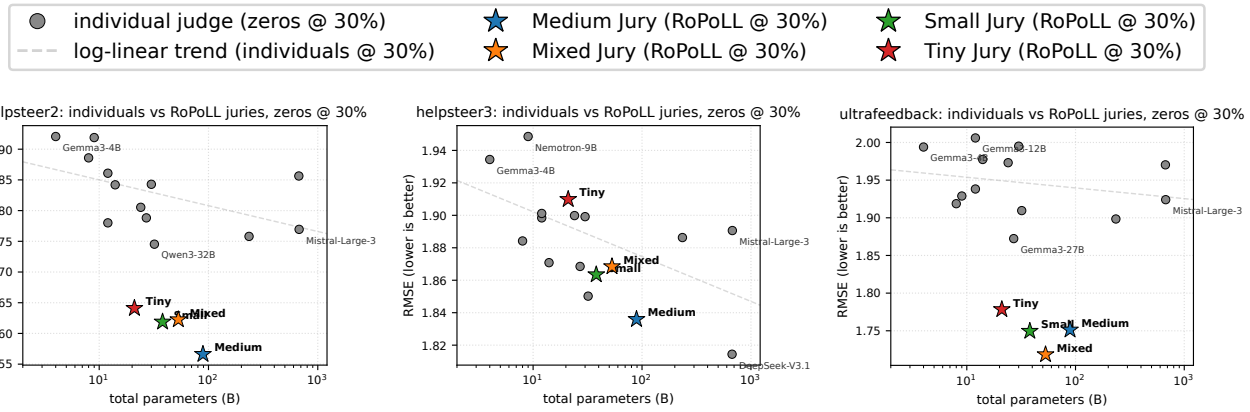
C.1. Additional Experiment Figures

The parameter-efficiency story of §5 extends to the other corruption regimes: Figure 9 (bounded `zeros`, $r = 30\%$) and Figure 10 (clean baseline, $r = 0\%$) are the counterparts of the body's `bimodal-random` hero (Figure 5). Figure 11 gives the full three-method PoLL/Median/RoPoLL degradation curves across every (dataset $\times$ corruption type) cell, and Figure 12 is the jury-size ablation supporting the $N = 3$ choice (both for §5).

C.2. Synthetic 2D Simulation: Visual Intuition

For pedagogical intuition we instantiate the observation model (2) in $d = 2$ dimensions with score range $[0, K]$ and visualize five representative failure modes. A jury of N judges evaluates a single instance with latent reward $\mathbf{y}^\star \in [0, K]^2$. Each competent judge ($Z_i = 0$) draws from a tight isotropic Gaussian centered on $\mathbf{y}^\star$; each corrupted judge ($Z_i = 1$) draws from a corruption distribution Q_i specific to the failure mode. The corruption indicator $Z_i \sim \mathrm{Bernoulli}(\alpha)$ is drawn independently per judge at homogeneous rate $\alpha \in \{0.10, 0.30, 0.40\}$. We compare the arithmetic mean and the geometric median (computed via Algorithm 1). In every figure, the gold star marks $\mathbf{y}^\star$, blue dots are competent judge outputs, red crosses are corrupted outputs, and the orange square and purple triangle mark the arithmetic mean and the geometric median, respectively.

Mode collapse ($Q = \delta_{\mathbf{0}}$). The corrupted judge outputs the zero vector on every attribute—the canonical parser-fallback failure mode (Remark 1). Figure 13 shows the mean pulled toward the origin as α grows, while the geometric median remains anchored to the competent cluster.



(a) HelpSteer 2 *(b)* HelpSteer 3 *(c)* UltraFeedback

Figure 9. **Parameter efficiency of RoPoLL juries vs. individual judges under `zeros` corruption at $r = 30\%$.** RMSE vs. parameter count (log scale) for each dataset; gray circles are the 13 individual open-weight judges (four anchors labelled), dashed line is their log-linear scaling fit, and coloured stars mark the four RoPoLL juries at their aggregate parameter budget, all under identical 30% per-case corruption. `zeros` replaces each corrupted slot with the parser-fallback vector $\mathbf{0}$; the direct RoPoLL vs. PoLL contrast is deferred to Figure 11.

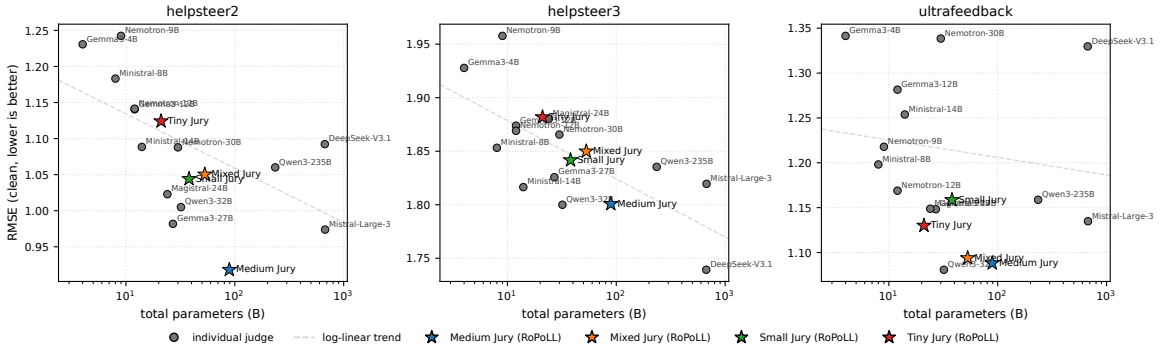


Figure 10. **Parameter efficiency at the clean baseline ($r = 0\%$).** RMSE vs. parameter count (log scale) for each dataset; gray circles are the 13 individual open-weight judges, dashed line is their log-linear scaling fit, and coloured stars mark the four RoPoLL juries at their aggregate parameter budget. Clean counterpart of Figures 5 and 9.

Inverted ($Q = \delta_{K \cdot \mathbf{1} - \mathbf{y}^\star}$). The worst-case anti-correlated Byzantine adversary (Figure 14). This is the sharpest visual demonstration of the breakdown-point advantage: the corrupted locus and $\mathbf{y}^\star$ lie on opposite sides of the score space, so at $\alpha = 0.30$ the mean has already crossed the midpoint while the geometric median remains within the competent cluster.

Biased dimension. Partial competence: correct on one attribute, catastrophically wrong on the other (Figure 15). This is the synthetic counterpart of `bimodal-random` (§5) and the picture of cross-dimensional corruption from Example 1: each corrupted score is plausible per coordinate but jointly anomalous, and the geometric median's joint-distance objective resists the off-axis pull that fools per-coordinate alternatives.

Random hypercube corners. The canonical instance of the cross-dimensional class: each corrupted score lands at a vertex of $\{0, K\}^d$ chosen uniformly at random (Figure 16). The per-coordinate marginal $\frac{1}{2}(\delta_0 + \delta_K)$ is indistinguishable from plausible scoring; jointly, every corrupted vector sits at a corner far from $\mathbf{y}^\star$ in ℓ_2. This is the "random vertex" generalization of *biased dimension* above and exactly the `bimodal-random` class evaluated empirically in §5.

Sycophantic. A real-world failure mode in which corrupted judges always rate near the top of the scale—the "everything is great" bias (Figure 17). The corrupted cloud sits in the upper-right corner of $[0, K]^d$; the arithmetic mean drifts diagonally toward it while the geometric median stays anchored to the competent majority near $\mathbf{y}^\star$. This complements *mode collapse*

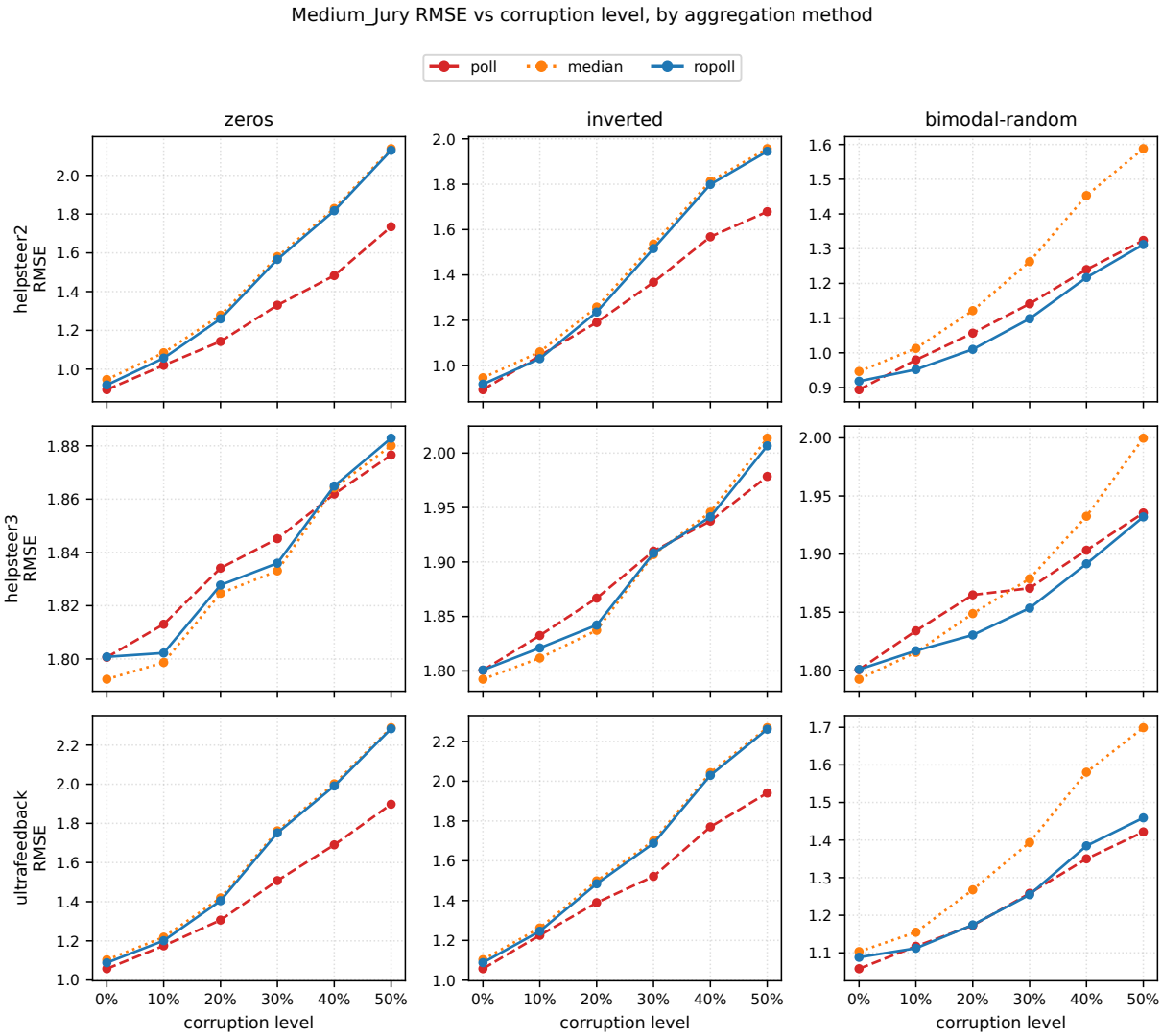


Figure 11. **PoLL vs. Median vs. RoPoLL degradation curves.** RMSE vs. per-case corruption rate r for the Medium jury, one panel per (dataset $\times$ corruption type). Solid = RoPoLL, dashed = PoLL, dotted = coordinate-wise Median.

(corruption at the lower-left extremum) at the opposite extreme of the score scale.

Summary. Across all three failure modes, the arithmetic mean acquires a bias proportional to α and aligned with the corruption locus, while the geometric median remains close to $\mathbf{y}^\star$ as long as $\alpha < 1/2$, in agreement with Theorem 1. The complementary Noisy-GT control (§5) confirms that this advantage is paid only against *biased* contamination: when the corruption is benign Gaussian noise, the geometric median does not sacrifice accuracy.

C.3. Per-Model and Per-Dimension Calibration Breakdowns

The figures in §5 aggregate across rubric dimensions and report the Medium jury's RMSE. This subsection records the underlying per-model and per-dimension calibration breakdowns on UltraFeedback that motivated the curated three-judge committees of §5.1.

Judge set. The calibration analysis in this subsection includes three closed-API reference judges (Claude Opus, Sonnet, and Haiku 4.5) in addition to the 13 open-weight judges of §5.1. The closed-API judges are *reference points only* — they are not used in any RoPoLL committee — and are included here to contextualise the open-weight calibration patterns.

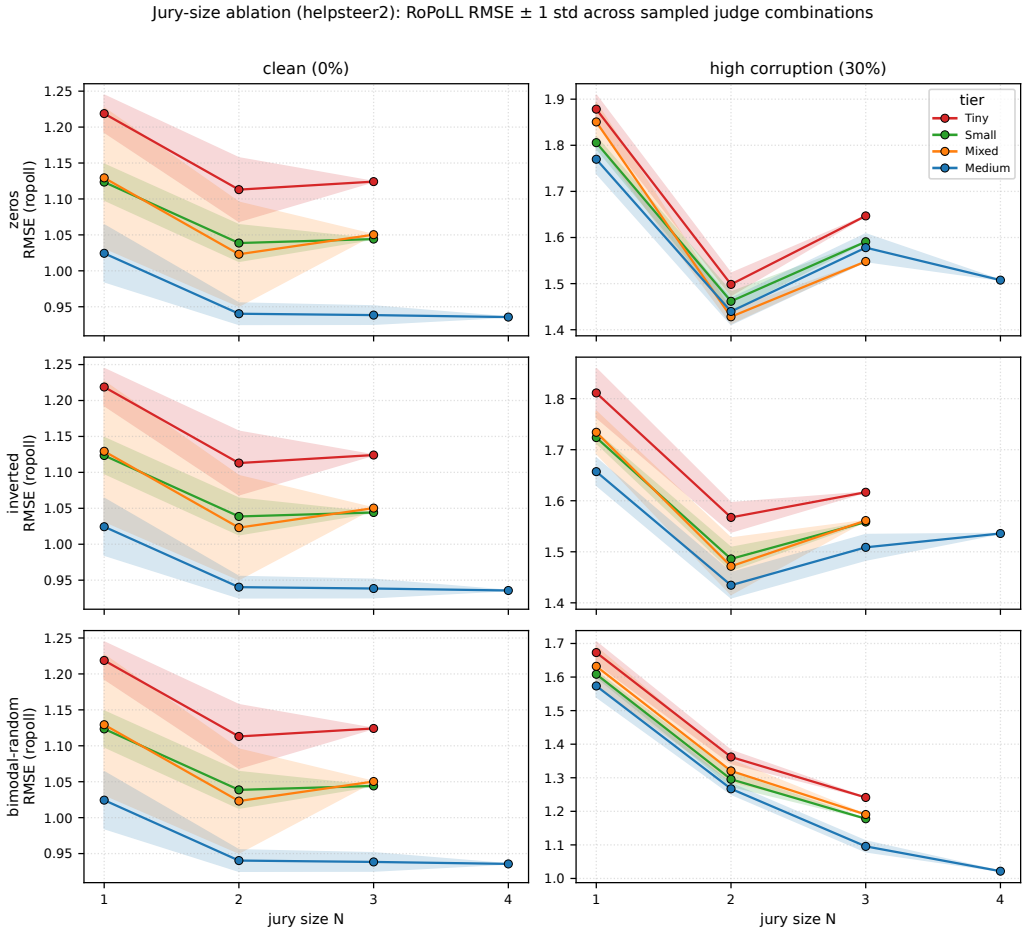


Figure 12. **Jury-size ablation: RMSE vs. jury size N.** Mean RMSE across sampled N-judge subcommittees from each tier pool, under `zeros`/`inverted`/`bimodal-random` corruption. Left column: $r = 0\%$; right column: $r = 30\%$. Bands show ± 1 standard deviation across combinations.

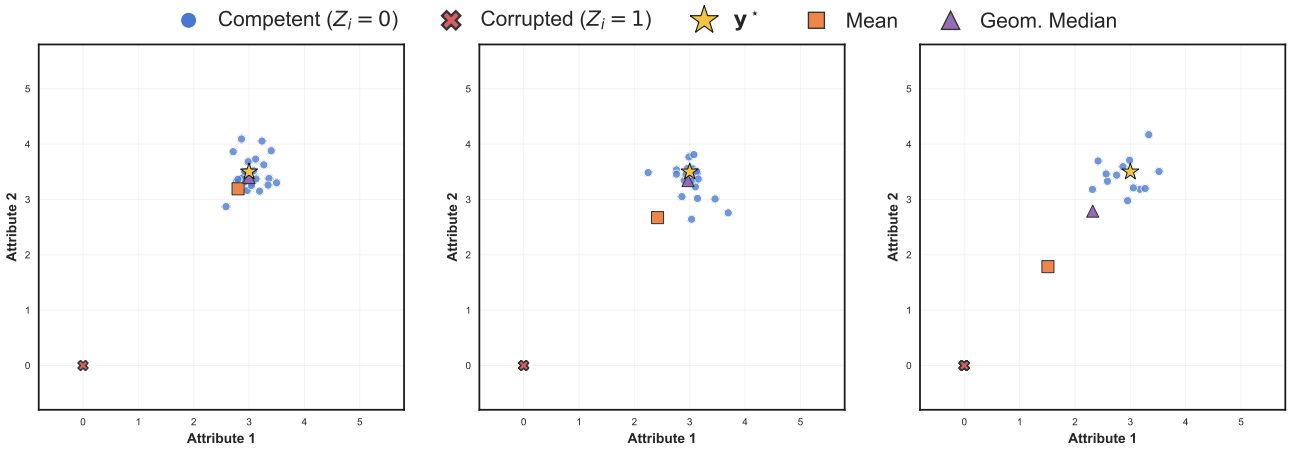
Per-dimension MAE. Figure 18 reports the mean absolute error for each judge against the UltraFeedback rubric dimensions (Helpfulness, Honesty, Instruction Following, Truthfulness). Qwen3 32B and Mistral-Large-3 lead with sub-0.75 MAE across all four dimensions; the Claude family lies near the bottom of the calibration ranking despite strong ranking ability (Figure 19 below explains why).

Per-dimension mean bias. Figure 19 reports the signed mean bias $\mathbb{E}[\hat{y}_i^{(k)} - y^{\star,(k)}]$ for each (judge, dimension) cell. Two systematic patterns emerge. The Claude family shows uniformly negative bias across all four dimensions (-0.5 to -0.8 on Truthfulness)—a systematic under-scoring tendency. Smaller open-weight models (Magistral Small, Gemma 4B, Nemotron 9B) show uniformly positive bias of comparable magnitude. Qwen3 32B and Qwen3 235B are closest to zero across all dimensions, consistent with their leading MAE. The bias direction is precisely the contamination structure Proposition 2 formalizes: mixing systematically over-scoring and under-scoring judges leaves the arithmetic mean's bias bounded only by the worst per-judge displacement; the geometric median is robust to such mixed-direction biases because the joint subgradient balance does not weight per-coordinate sign.

D. Released Corpus and Dataset Analysis

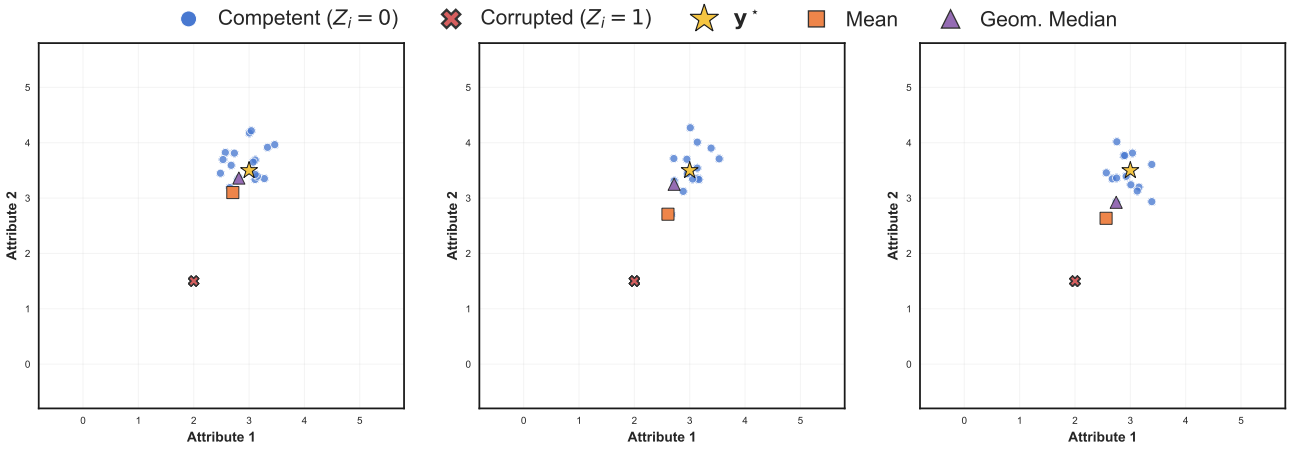
D.1. Released Corpus

To support reproduction and follow-up work, we release the full 13-judge × three-benchmark output corpus that drives every figure in §5. For each (judge, sample) cell the corpus contains the raw judge response text T_f, the parsed score vector



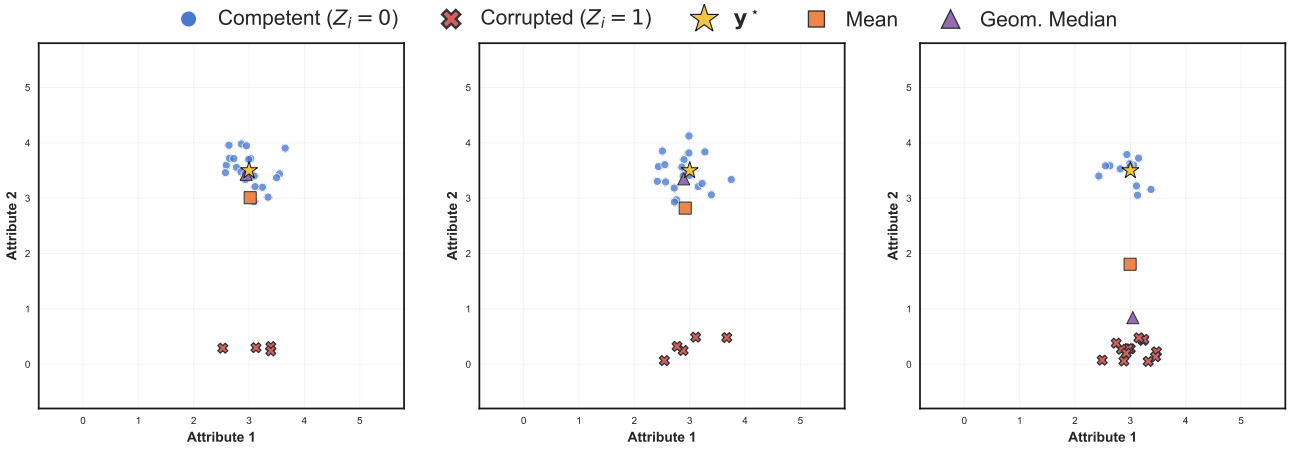
$\alpha = 0.10$ $\alpha = 0.30$ $\alpha = 0.40$

Figure 13. **Mode Collapse corruption** ($Q = \delta_{\mathbf{0}}$). Corrupted judges output the zero vector, modeling parser failures or safety refusals. The mean is pulled linearly toward the origin; at $\alpha = 0.40$ it lies roughly 40% of the way from $\mathbf{y}^\star$ to $\mathbf{0}$. The geometric median remains within the competent cluster because the majority of Euclidean distances still point toward $\mathbf{y}^\star$.



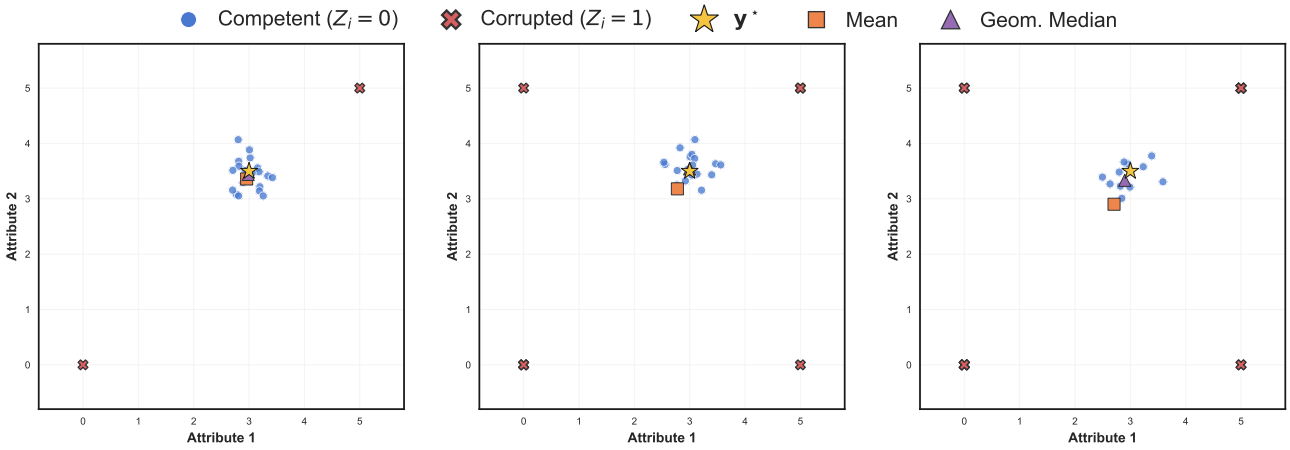
$\alpha = 0.10$ $\alpha = 0.30$ $\alpha = 0.40$

Figure 14. **Inverted corruption** ($Q = \delta_{K \cdot \mathbf{1} - \mathbf{y}^\star}$). The worst-case Byzantine adversary: corrupted scores are perfectly anti-correlated with the truth. The corruption locus and $\mathbf{y}^\star$ lie on opposite sides of the score space. At $\alpha = 0.30$ the mean is already displaced past the midpoint, while the geometric median remains close to $\mathbf{y}^\star$. This is the sharpest demonstration of the breakdown-point advantage.



$\alpha = 0.10$ $\alpha = 0.30$ $\alpha = 0.40$

Figure 15. **Biased Dimension corruption.** Corrupted judges evaluate Attribute 1 correctly but catastrophically fail on Attribute 2 (scores collapse near zero). This partial competence is challenging for coordinate-wise methods because the corruption is invisible on one axis. The geometric median, operating on joint Euclidean distances, detects the anomaly in Attribute 2 and downweights the corrupted points across both dimensions.



$\alpha = 0.10$ $\alpha = 0.30$ $\alpha = 0.40$

Figure 16. **Random hypercube corners** (the canonical instance of the cross-dimensional class of Example 1, matching the empirical `bimodal-random` class of §5). Corrupted judges output an extreme vertex of $\{0, K\}^d$ chosen uniformly at random; per-coordinate the corruption marginal $\frac{1}{2}(\delta_0 + \delta_K)$ is plausible scoring, but the joint vector lies far from $\mathbf{y}^\star$ in ℓ_2. The geometric median resists the cross-dimensional pull (it sits at $\mathbf{y}^\star$, beneath the gold star), while the arithmetic mean drifts toward the centroid of the corrupted vertices.

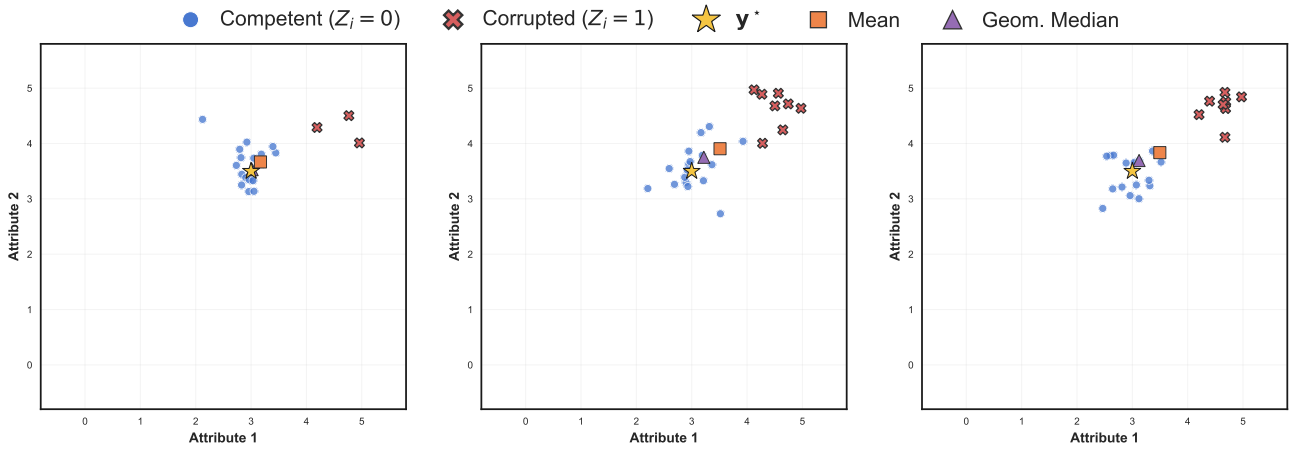


Figure 17. **Sycophantic corruption** ($Q = \mathrm{Uniform}([K-1, K]^d)$). Corrupted judges produce scores clustered near the maximum, modeling the "everything is great" failure mode. The corrupted cloud sits in the upper-right corner; the mean drifts diagonally toward it while the geometric median stays anchored to the competent majority near $\mathbf{y}^\star$.

$\hat{\mathbf{y}}_f(x) \in \mathbb{R}^d$ produced by the deterministic parser ϕ (Definition 2), the per-call latency, parser-failure flags, and the reference label $\mathbf{y}_j^{\mathrm{ref}}$. Alongside the per-sample scores we release the structured rubric ρ, the parser ϕ, the per-case corruption pipeline (used to inject the `zeros`, `inverted`, `bimodal-random`, and `cauchy-far` adversaries), and the `results.json` produced by every aggregation method evaluated in §5. The corpus totals approximately 28K scored (judge, sample) cells (Table 2), enabling exact reproduction of every reported figure without re-running the inference cost.

Dataset	N_{samp}	$\lvert J\rvert$	d	$\bar{f}$	$f_{\max}$	$s_{\min}$	$s_{\max}$
HelpSteer 2	1000	13	5	0.6%	2.4%	−1.0	4.0
UltraFeedback	1000	13	4	0.0%	0.0%	1.0	5.0
HelpSteer 3	100	16	1	6.0%	33.0%	−3.8	2.6

Table 2. Corpus-level statistics. N_{samp}: samples; $|J|$: judge pool size; d: target dimension; $\bar{f}$, $f_{\max}$: mean and max per-judge parser-failure rate; $s_{\min}$, $s_{\max}$: observed score range across all judges and samples (negative values arise from HS 2 / HS 3 signed-difference reductions on a small fraction of cells where parsed scores fell outside the rubric range). For HS 3, $\bar{f}$ and $f_{\max}$ are computed over the full 16-judge pool (the 13 open-weight judges plus the three HS 3-only Claude judges); the 13-judge common-pool mean is 3.38% (Figure 2). HS 3 in the released JSON contains the 100-sample preference slice used for the §5 evaluation; the full 2017-sample multilingual validation set is available on request.

Per-attribute score distributions. Figure 20 plots the score distributions per attribute (per-dataset). HelpSteer 2 and UltraFeedback have substantial mass at the score extremes (parser fallback at 0; sycophantic judges concentrating at the maximum), motivating the `zeros` and `inverted` corruption types used in §5. HelpSteer 3, which reduces a five-attribute pair of responses to a single signed-preference scalar, is well-centered on 0 with light tails, consistent with the cancellation of per-attribute biases under the signed-difference reduction.

D.2. Inter-Judge Correlation Structure

Figure 21 shows the pairwise Pearson correlation between every judge pair in the 13-judge pool, averaged over attributes. Empirical mean off-diagonal correlations are $\bar{\gamma}_{\mathrm{HS2}} = 0.49$, $\bar{\gamma}_{\mathrm{HS3}} = 0.49$, and $\bar{\gamma}_{\mathrm{UF}} = 0.71$. These directly support the assumption $\gamma \in [0.3, 0.7]$ used in §5.1 to motivate the choice $N = 3$: substituting the measured $\bar{\gamma}$ into the saturation law $N_{\mathrm{eff}}^{\infty} = 1/\gamma$ (Corollary 1) yields $N_{\mathrm{eff}}^{\infty} \approx 2.0$ on HS 2 and HS 3 and ≈ 1.4 on UltraFeedback, so the empirical diminishing-returns knee at $N = 3$ in Figure 12 sits at or just past the saturation point predicted by the corpus's actual correlation structure.

The UltraFeedback correlation $\bar{\gamma}_{\mathrm{UF}} = 0.71$ is notably higher than the HelpSteer correlations. This reflects the fact that UltraFeedback's reference labels are themselves GPT-4 annotations (Cui et al., 2024), so judges trained on similar

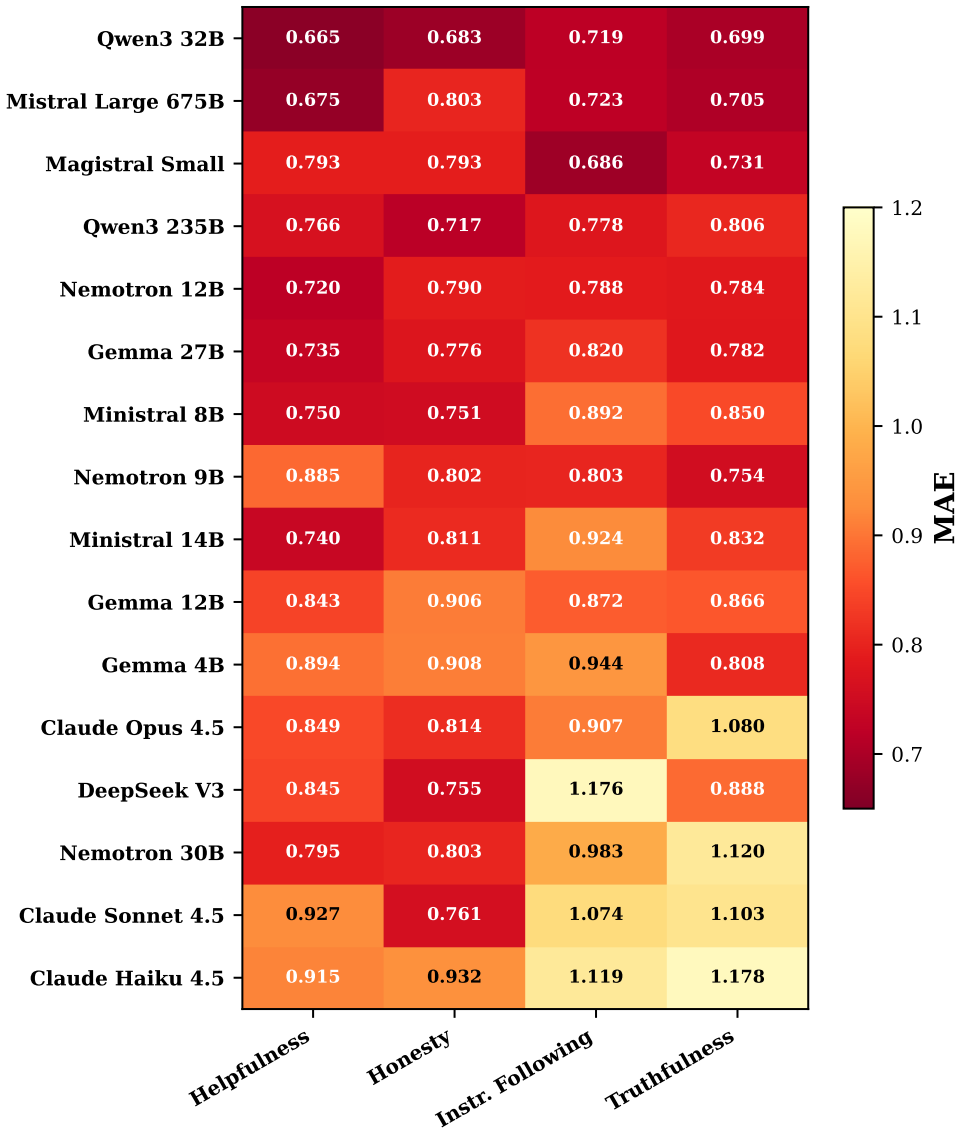


Figure 18. Per-dimension MAE for each LLM judge on UltraFeedback (n=1000), sorted by lowest average error. Qwen3 32B achieves the lowest MAE across all four dimensions. The Claude family clusters near the bottom despite strong ranking ability, with Instruction Following and Truthfulness showing the largest errors (> 1.0) due to systematic negative bias.

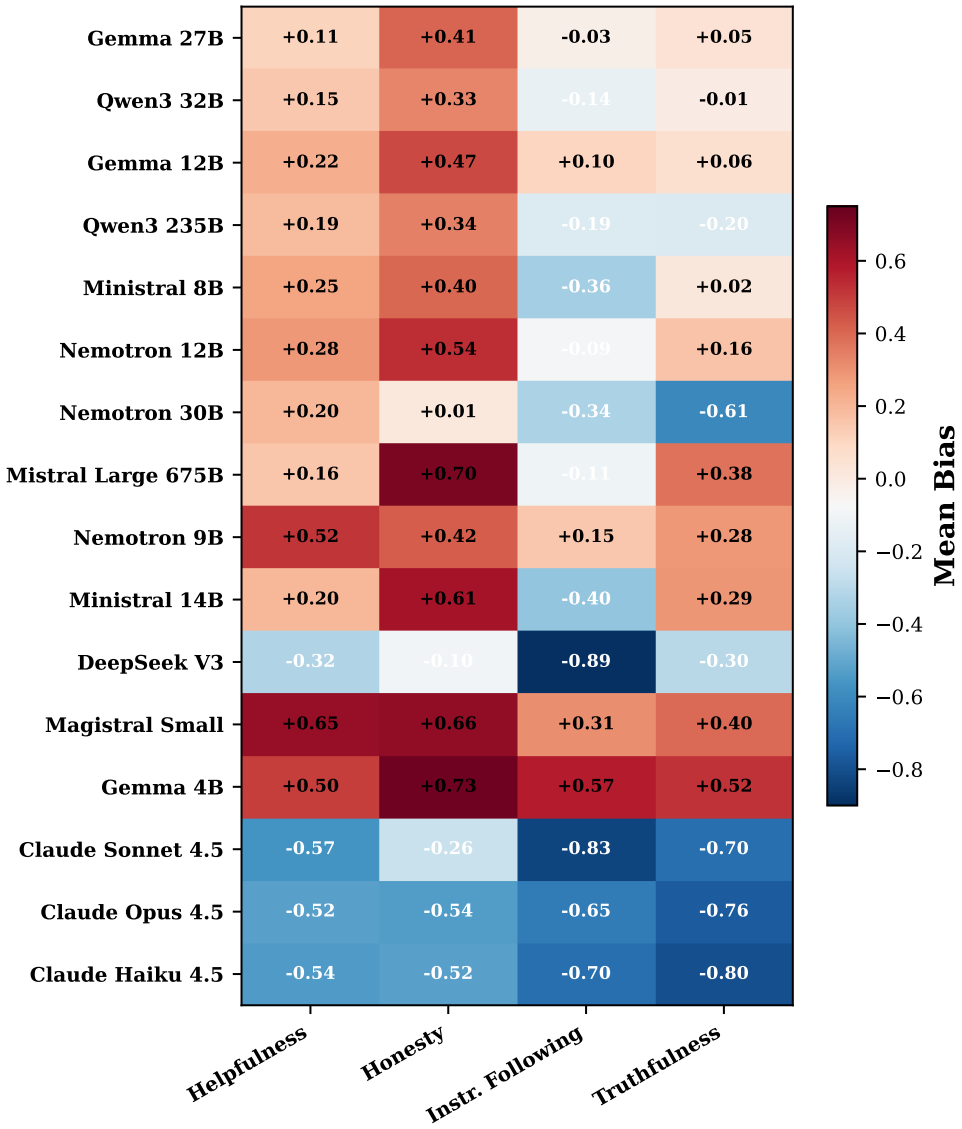


Figure 19. Per-dimension mean bias for each LLM judge on UltraFeedback (n=1000), sorted by lowest absolute bias. Blue cells indicate under-scoring (negative bias); red cells indicate over-scoring (positive bias). The Claude family shows uniformly negative bias across all dimensions, while models like Magistral Small and Gemma 4B exhibit strong positive bias. Qwen3 32B and Qwen3 235B are closest to zero across all dimensions.

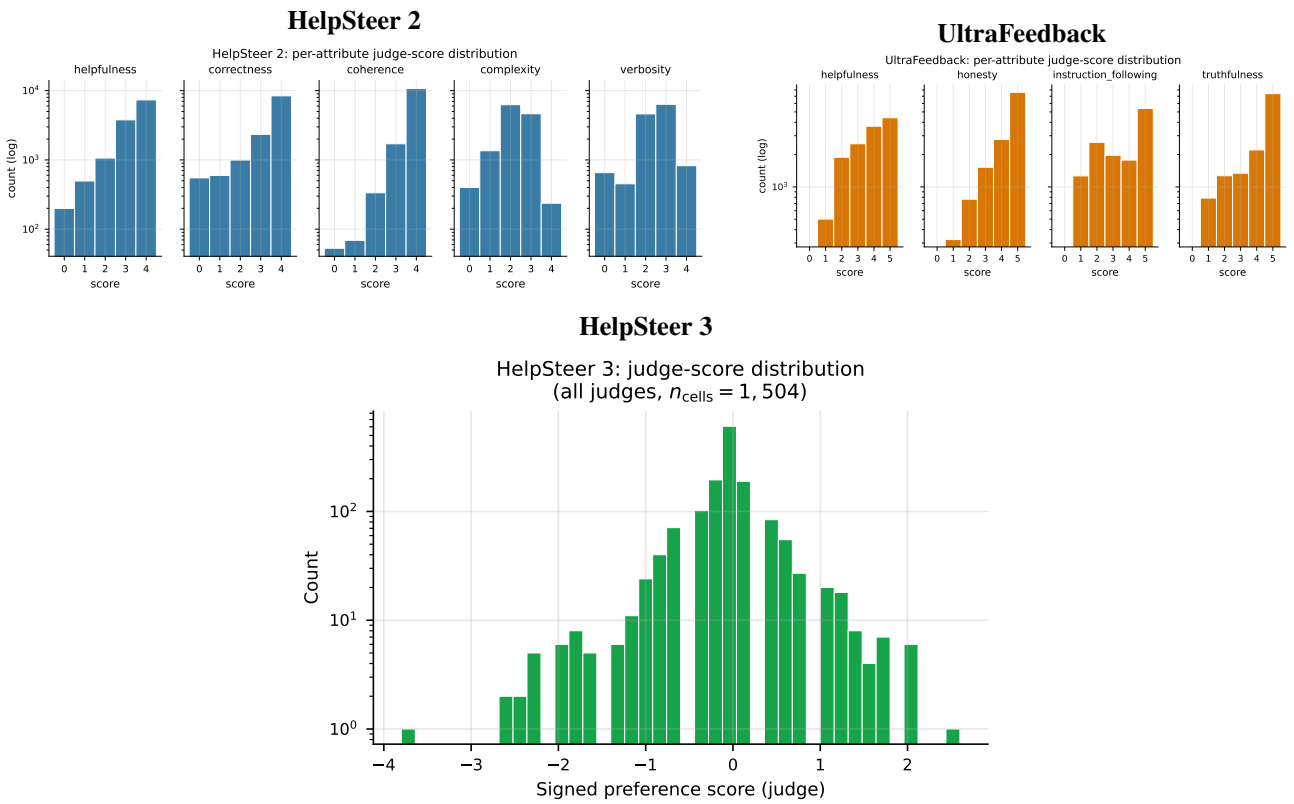


Figure 20. **Per-attribute judge-score distributions (log y-axis).** HelpSteer 2 and UltraFeedback show heavy mass concentration at the score extremes—parser fallback at 0 and sycophantic saturation at the maximum—which motivates the `zeros` and `inverted` corruption types used in §5. HelpSteer 3 (signed-preference scalar) is centered on 0 with light tails, consistent with cancellation of per-attribute biases under the signed-difference reduction.

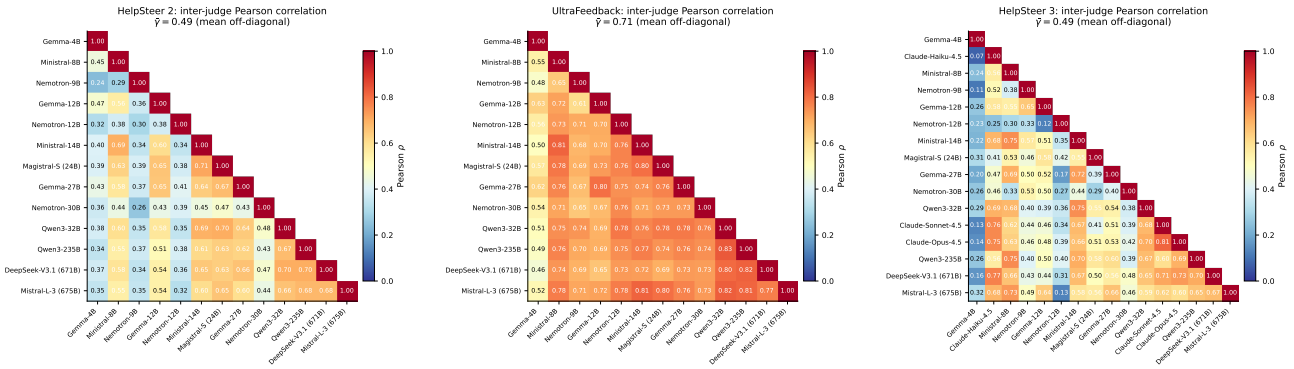


Figure 21. **Inter-judge Pearson correlation heatmaps (lower-triangle, annotated).** Pairwise correlations averaged over evaluation attributes; cells labelled with their numeric value. Empirical mean off-diagonal correlations: $\bar{\gamma}_{\text{HS2}} = 0.49$, $\bar{\gamma}_{\text{UF}} = 0.71$, $\bar{\gamma}_{\text{HS3}} = 0.49$. These values support the $\gamma \in [0.3, 0.5]$ assumption used in §5.1 to motivate three-judge committees via Corollary 1; the higher $\bar{\gamma}_{\text{UF}}$ explains the smaller RoPoLL/PoLL gap observed on UltraFeedback.

rubric distributions converge to similar scores; the HelpSteer benchmarks use trained-human annotators (HelpSteer 2) or pairwise human preferences (HelpSteer 3), producing more genuine inter-judge variation. This is consistent with the smaller RoPoLL/PoLL gap observed on UltraFeedback in §5: when judges already agree, the difference between the mean and the geometric median is small.

D.3. Empirical Indicator Correlation $\bar{\gamma}_W$

Lemma 3 bounds the failure probability of the RoPoLL cluster event in terms of the *indicator correlation* $\bar{\gamma}_W = \mathrm{mean}_{i \neq j} \frac{\mathrm{Cov}(W_i, W_j)}{\sqrt{\mathrm{Var}(W_i)\mathrm{Var}(W_j)}}$ of the cluster indicators $W_i = \mathbb{1}\{Z_i = 0,\ \|\hat{\mathbf{y}}_i - \mathbf{y}^\star\|_2 \leq \rho_p\}$. This is in principle a finer object than the inter-judge *score* correlation $\bar{\gamma}$ of §D.2: $\bar{\gamma}$ measures the linear correlation between raw score vectors, while $\bar{\gamma}_W$ measures the co-incidence of two judges *both being competent and within the cluster ball*. We estimate $\bar{\gamma}_W$ directly on the experimental panels.

Estimation procedure. For benchmark $b \in \{\mathrm{HS2}, \mathrm{UF}\}$: (i) for each (judge i, sample s) cell, compute the ℓ_2 deviation $\delta_i^{(s)} = \|\hat{\mathbf{y}}_i^{(s)} - \mathbf{y}^{\star,(s)}\|_2$ (parser-failure cells contribute $W_i^{(s)} = 0$); (ii) select a cluster radius ρ as the p-th quantile of pooled deviations $\{\delta_i^{(s)}\}_{i,s}$; (iii) form $W_i^{(s)} = \mathbb{1}\{\delta_i^{(s)} \leq \rho\}$; (iv) compute the mean off-diagonal Pearson correlation of the rows of $W \in \{0,1\}^{N \times S}$. We report $\bar{\gamma}_W$ at three radii ($p \in \{0.50, 0.70, 0.90\}$) to show stability under the calibration choice.

Benchmark	p-quantile	ρ	$\bar{\gamma}_W$	N_{eff} at N=3
HelpSteer-2	0.50	2.000	0.500	1.50
	0.70	2.449	0.531	1.46
	0.90	4.000	0.471	1.55
UltraFeedback	0.50	2.000	0.531	1.45
	0.70	2.449	0.475	1.54
	0.90	3.742	0.450	1.58

Table 3. **Empirical indicator correlation $\bar{\gamma}_W$ of Lemma 3** on our 13-judge experimental panels. $\bar{\gamma}_W$ is stable to within ± 0.03 across cluster-radius calibrations and lies in $[0.45, 0.53]$ on both benchmarks, so $N_{\mathrm{eff}} \in [1.45, 1.58]$ at the practical jury size N=3. The empirical $\bar{\gamma}_W$ is on the same order as the score correlation $\bar{\gamma} \in [0.49, 0.71]$ reported in Figure 21; Pitt's Gaussian correlation inequality (Pitt, 1977; Esary et al., 1967; Joag-Dev & Proschan, 1983) gives the qualitative bound $\bar{\gamma}_W \geq 0$ but not a quantitative comparison to $\bar{\gamma}$, so direct estimation is the right move.

Implication for Lemma 3. The role of Lemma 3 is structural: it shows that the geometric-breakdown structure ($C_{\alpha+\beta}$ and the cluster radius ρ) of Theorem 1 is preserved when the i.i.d. assumption is replaced by an equicorrelated-indicator hypothesis with $\bar{\gamma}_W \in [0, 1]$. The probability event delivered by the Chebyshev step, $\Pr[\cdot] \geq 1 - 1/(\beta^2 N_{\mathrm{eff}})$, is informative in the large-N_{eff} regime (e.g., a hypothetical jury of $N = 10$–30 judges with $\bar{\gamma}_W \approx 0.2$ gives $N_{\mathrm{eff}} \in [3.6, 7]$ and $\beta = 0.2$ gives a non-trivial bound) but degenerates at small N_{eff}, including the practical $N_{\mathrm{eff}} \approx 1.5$ of our $N = 3$ panels. This is a fundamental limit of variance-only concentration at small N, not a slack in the analysis: with only ~ 1.5 effective independent samples, no concentration argument can deliver a tight high-probability bound, regardless of the estimator. A Bernstein-type bound under bounded-covariance martingale structure (Remark 17) would replace β^{-2} with $\exp(-c\beta^2 N_{\mathrm{eff}})$ but does not materially help at $N_{\mathrm{eff}} \approx 1.5$. The practical value of Lemma 3 for our small-N regime is therefore the structural guarantee, not the quantitative probability: the breakdown floor and the geometric constant are independent of this concentration argument and *are* the load-bearing quantities for jury-aggregation deployment.