WHAT CAN BE CORRECTED IN LLM JUDGE RATINGS? IDENTIFIABILITY, MEASUREMENT, AND TARGETED ADJUSTMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Correcting ratings from LLM judges requires knowing whether associations with features such as verbosity, lexical overlap, or displayed position reflect bias or item quality. Ratings alone can identify differences in these associations across judges, but cannot separate an association shared by the panel from item quality. We formalize this limit with a hierarchical Bayesian measurement model that represents each ordinal rating in terms of latent item quality, judge leniency, and feature associations. We use this model as a diagnostic audit: human labels, controlled counterfactuals, or another external anchor are required to identify a panel-wide association. We therefore vary unidentified panel-wide slopes in sensitivity analyses. Targeted adjustment is a diagnostic control that removes only identifiable judge-specific deviations. On TREC web relevance and SummEval summarization, targeted adjustment has higher Spearman rank-correlation point estimates with human labels than full adjustment, although panel averaging matches or outperforms it. As a practical implementation layer, a 250,000-parameter amortized network matches sampling-based masked-rating reconstruction error (0.302 versus 0.304) while reducing per-dataset inference time from about 240 seconds to 1.3 ms. We recommend selecting a scoring rule through human-label validation and removing a panel-shared association only when an external anchor identifies the corresponding effect.

1 INTRODUCTION

Large language models (LLMs) increasingly grade the outputs of other models: search results, answers, summaries, and pairwise preferences (Zheng et al., 2023; Chiang et al., 2024). Low cost and scalability have made LLM judging common in evaluation pipelines. Yet, LLM judges are not neutral instruments. They differ in leniency, and their ratings can vary systematically with verbosity, position, lexical overlap, and other properties that should not automatically determine quality (Wang et al., 2024; Gu et al., 2025). Two standard responses differ in what they remove: panel averaging retains the feature responses shared across judges, whereas full adjustment removes both shared and judge-specific responses.

The ratings alone may not reveal whether a shared response reflects bias or item quality. Suppose every judge gives higher scores to documents with greater lexical overlap. Every judge may be rewarding overlap as a shortcut. Alternatively, overlap may be associated with higher human-assessed document quality. The rating matrix cannot distinguish these explanations. Retaining or removing the shared response can change which system appears better aligned with human labels (Section 6.4).

Therefore, we treat each LLM judge as a *measurement instrument*, and each ordinal rating as a noisy measurement of latent item quality (Lord, 1980; De Boeck & Wilson, 2004). When multiple judges rate the same items, differences among their slopes show how each judge responds relative to the others: one judge may react more strongly to lexical overlap than another. A panel-wide slope, however, can be absorbed into latent item quality unless human labels, controlled counterfactuals, or another external anchor separates the slope from quality (Proposition 1). Ratings can therefore support comparisons among judges, but cannot by themselves justify removing a panel-wide association.

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

Our experiments separate three distinct goals. For ranking items against human labels, the panel mean does best; for predicting ratings deliberately withheld from judges already in the panel, graded-response item response theory (IRT) does best. Our measurement model addresses a third goal: determining which fitted feature associations can be separated from item quality before scores are adjusted.

We make three contributions. **Diagnostic auditing and identifiability:** we convert the ratings-only limit into an audit that estimates judge-specific deviations, makes explicit that model assumptions determine the panel-wide slope, and reports sensitivity to those assumptions. **Diagnostic control:** targeted adjustment removes only identifiable deviations, illustrating the consequence of retaining versus removing shared associations; it does not outperform panel averaging as a scoring rule. **Practical implementation:** an amortized network closely matches Markov chain Monte Carlo (MCMC) judge-parameter estimates and masked-rating reconstruction while reducing per-dataset inference time from about 240 seconds to 1.3 ms, enabling inexpensive reruns of the same audit rather than changing its scientific conclusion.

2 Related Work

LLM judges and information retrieval. Work on LLM-as-a-judge evaluation has documented systematic effects of position, verbosity, self-preference, and other surface features. These effects are often described as biases and motivate task-specific debiasing procedures and uncertainty estimates (Zheng et al., 2023; Wang et al., 2024; Saad-Falcon et al., 2024; Gu et al., 2025; Ye et al., 2024; Shi et al., 2025; Chen et al., 2024; Wataoka et al., 2024; Dubois et al., 2024). In information retrieval (IR), a related line studies whether LLM-generated relevance judgments agree with human assessments and support reliable system comparisons, including for legal retrieval-augmented generation (RAG) (Voorhees, 2001; Upadhyay et al., 2024; 2025; Arabzadeh & Clarke, 2025; Pradhan et al., 2025; Rahmani et al., 2025). Controlled studies further show that lexical shortcuts, such as inserting query terms, can inflate relevance judgments (Alaofi et al., 2024; Yu et al., 2026). These studies document associations between observable item features and LLM ratings. We ask which components of those associations a crossed rating design can distinguish from latent item quality before any correction is attempted.

Psychometric and annotator models. Models from item response theory (IRT), including graded-response and the Many-Facet Rasch Model (MFRM), represent item difficulty and rater severity (Lord, 1980; Samejima, 1969; Linacre, 1994). Dawid–Skene, the Generative model of Labels, Abilities, and Difficulties (GLAD), Multi-Annotator Competence Estimation (MACE), and Bayesian annotator models infer latent labels or rater reliability from multiple annotations (Dawid & Skene, 1979; Whitehill et al., 2009; Hovy et al., 2013; Carpenter, 2008; Passonneau & Carpenter, 2014). We follow explanatory IRT by including item covariates (De Boeck & Wilson, 2004). Equation 1 is a judge-varying-slope model that can be reparameterized using the random-residual formulation of Cho et al. (2014).

Identification. Differential item functioning examines whether items behave differently across groups at the same latent ability. Without anchors or other identifying assumptions, data identify differences among item-specific group effects but cannot distinguish an effect common to all items from a difference in group ability (Chen et al., 2023; Bechger & Maris, 2015; Doebler, 2019). Results can therefore depend on the anchor choice (Kopf et al., 2015). Likewise, our ratings identify differences among judges' feature slopes, expressed relative to the panel mean by sum-to-zero centering, but cannot separate a panel-shared slope from item quality.

Calibration and efficient inference. Recent evaluator-calibration research includes models of human–LLM disagreement and methods that produce statistically valid aggregate estimates (Polo et al., 2025; Dubois et al., 2025; Gao et al., 2024; Lee et al., 2025; Zhao et al., 2026). Choi et al. (2026) concurrently apply a graded-response model to diagnose LLM judges, but without item covariates or a shared/judge-specific split. Xu et al. (2026) study related non-identifiability for a single pairwise bias coefficient and the role of external anchors. Dorner et al. (2025) prove a complementary limit: an unbiased debiasing estimator can reduce the required ground-truth labels by at most a factor of two. Our contribution is therefore not the identifiability boundary itself, but an audit built around it. We report shared slopes as unidentified from ratings alone, characterize them through explicit breakdown and sensitivity analyses (Imbens & Manski, 2004; Molinari, 2020; Gustafson, 2005;

Cinelli & Hazlett, 2020), adjust only identifiable judge-specific deviations, and amortize the resulting inference procedure (Cranmer et al., 2020; Talts et al., 2018).

Gold-label and aggregation alternatives. A separate line of work uses small gold-labeled samples to anchor aggregate claims. Prediction-powered inference combines gold labels with many auto-rater labels to obtain valid intervals for population functionals (Angelopoulos et al., 2023; 2024). Extensions cover discrete "prompted LLM judge" auto-raters (Hofer et al., 2024) and stratified sampling (Fisch et al., 2024). Oosterhuis et al. (2024) apply this approach to TREC Deep Learning (TREC-DL) relevance using National Institute of Standards and Technology (NIST) labels. These methods are well suited to population means and metric confidence intervals. Our audit instead targets per-item latent quality and per-judge covariate responses, and its deployment setting assumes that gold labels are unavailable (Section 3.3). Panel aggregation provides a label-free alternative (Verga et al., 2024); average pooling of ordinal scores is the panel-mean baseline used throughout our experiments.

3 The Measurement Model

3.1 Covariate-response taxonomy and identifiability

Measurement intuition. An observed rating reflects latent item quality, judge leniency, feature response, and ordinal noise. Consider lexical overlap. If one judge rewards overlap more than another, the gap between their expected ratings changes with overlap and is visible in the rating matrix. If every judge rewards overlap equally, however, the ratings cannot distinguish that shared response from higher item quality. The model formalizes this distinction by writing each feature response as a panel-wide slope plus a judge-specific deviation.

Measurement equation. Let $r_{ij} \in \{0, \dots, K\}$ be judge j's ordinal rating of item i, and let x_i contain five observed item features (covariates): lexical overlap, verbosity, position, formality, and query length. Decompose the covariate slope as

$$\beta_j = \bar{\beta} + \delta_j, \qquad \sum_{j=1}^{J} \delta_j = 0,$$

where $\bar{\beta}$ is the panel-mean slope and δ_j is judge j's centered deviation from it. Let $h_i = \text{position}_i \times \text{verbosity}_i$ denote the position-by-verbosity interaction. The ordered-probit model is

$$\eta_{ij} = \underbrace{\theta_i + \bar{\beta}^\top x_i + \gamma h_i}_{\text{shared across judges}} + \underbrace{\alpha_j + \delta_j^\top x_i}_{\text{judge-specific}}, \tag{1}$$
$$\Pr(r_{ij} \le k \mid \eta_{ij}) = \Phi(\kappa_k - \eta_{ij}), \qquad k = 0, \dots, K-1.$$

Here θ_i is latent item quality, α_j is judge leniency (larger values shift ratings upward), Φ is the standard-normal cumulative distribution function, and $\kappa_0 < \cdots < \kappa_{K-1}$ are the ordinal cutpoints. The coefficient γ is the panel-wide interaction slope. It has no judge index and therefore has the same identification status as the shared feature slopes; we suppress it below.

Definition 1 (Covariate-response classes)**.** For covariate c in the decomposition $\beta_j^c = \bar{\beta}_c + \delta_j^c$, the response is *shared* if $\delta_j^c = 0$ for every judge. It is *partly judge-specific* if $\text{Var}_j(\delta_j^c) > 0$. These labels describe cross-judge heterogeneity, not whether a response is normatively biased.

This is the familiar differential item functioning (DIF) identification problem: without an anchor, common effects are confounded with latent item quality, whereas relative effects can be identified (Bechger & Maris, 2015; Doebler, 2019). We restate the boundary for crossed ordinal LLM-judge panels because it determines what a ratings-only correction may remove.

Proposition 1 (What ratings alone identify)**.** Assume no external anchor is available: no human labels, no counterfactual variation at fixed item quality, and no structural restriction linking item effects to covariates. Assume also the design conditions of Appendix I: fixed global location and unit probit residual variance, a connected judge–item graph, non-degenerate category use, full-rank covariate variation, and at least two judges. Then:

(i) the ratings identify the item-consensus quantity $u_i = \theta_i + \bar{\beta}^\top x_i$, but not its decomposition into θ_i and $\bar{\beta}^\top x_i$;

(ii) the ratings identify the centered judge-specific deviations δ_j.

Proof sketch. For any vector a, the transformation $\bar{\beta}' = \bar{\beta} + a$ and $\theta_i' = \theta_i - a^\top x_i$ leaves every η_{ij}, and hence the entire ratings likelihood, unchanged. Thus the ratings cannot choose among the possible decompositions of u_i into shared covariate response and item quality. By contrast, the judge-varying part is visible through

$$\eta_{ij} - \eta_{ij'} = (\alpha_j - \alpha_{j'}) + (\delta_j - \delta_{j'})^\top x_i.$$

Item quality and the shared slope cancel. Under the stated design conditions, variation in x_i identifies the relative slope contrasts, and centering fixes their panel reference. □

Interpretation. The proposition concerns statistical separability, not normative desirability. Ratings support a panel-relative adjustment of δ_j, but do not establish that the heterogeneous component should be removed or that the shared component is unbiased. Human labels, controlled perturbations, or other counterfactual comparisons can provide the external anchor needed to identify a shared response.

3.2 Generative process and hierarchy

We fit equation 1 as a crossed hierarchical ordered-probit model with item-, judge-, and population-level parameters. Judge intercepts vary around means for predefined capability groups, and item effects use non-centered parameterizations:

$$\alpha_j = \mu_{t(j)} + \sigma_m \epsilon_j, \qquad \theta_i = \sigma_p \tilde{\theta}_i, \qquad \epsilon_j, \tilde{\theta}_i \sim \mathcal{N}(0,1),$$

where $t(j)$ denotes judge j's capability group, σ_m is the within-group scale of judge leniency, and σ_p is the item-effect scale. Each covariate slope is hierarchical as well:

$$\beta_j^c = \mu_c + \sigma_c \tilde{\beta}_j^c, \qquad c \in \{\text{overlap, verbosity, position, formality, query length}\}, \tag{2}$$

with $\tilde{\beta}_j^c \sim \mathcal{N}(0,1)$. Here μ_c is the population mean and σ_c is the between-judge scale. The fit does not impose the zero-sum decomposition in equation 1 directly. For reporting and adjustment, we instead compute the finite-panel mean $\bar{\beta}_c = J^{-1} \sum_j \beta_j^c$ and deviations $\delta_j^c = \beta_j^c - \bar{\beta}_c$, which satisfy $\sum_j \delta_j^c = 0$ by construction. Thus σ_c measures cross-judge heterogeneity, whereas $\bar{\beta}_c$ is the shared component for the observed panel; it need not equal μ_c exactly. Appendix G lists the exact weakly informative priors.

3.3 Audit and deployment settings

The audit and deployment fits use the same rating likelihood. They differ in whether human labels enter the fit:

- **Audit (supervised).** A human-labeled calibration set is available, and the normalized label g_i enters as an additional item-level predictor, $\beta_{gt} g_i$. Including g_i conditions the estimated feature associations on this observed measure of human-assessed quality; we use this fit for diagnosis.
- **Deployment (unsupervised).** The label term is omitted, so θ_i absorbs quality and other residual item variation. This is the label-free fit available at scoring time.

For the audit fit, we report both in-sample and query-level cross-fitted human-label alignment. For the deployment fit, labels are used only after scoring to evaluate alignment. Masked-rating reconstruction is reported separately as a diagnostic of panel behavior.

How the fitted hierarchy selects a shared slope. The conditional ratings likelihood remains flat in the shared slope, yet the implemented hierarchy returns a point estimate because it restricts the distribution of item effects. Suppressing the shared interaction, take the item consensus u_i of Proposition 1. The hierarchy can be written equivalently as

$$\begin{aligned} u_i &= \bar{\beta}^\top x_i + e_i, && \mathbb{E}[e_i \mid x_i] = 0 \quad \text{(deployment)}, \\ u_i &= \bar{\beta}^\top x_i + \beta_{gt} g_i + e_i, && \mathbb{E}[e_i \mid x_i, g_i] = 0 \quad \text{(audit)}, \end{aligned} \qquad e_i \sim \mathcal{N}(0, \sigma_p^2). \tag{3}$$

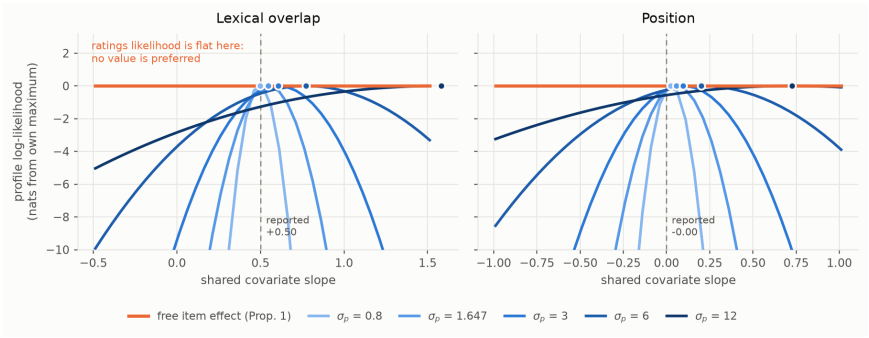


Figure 1: Ratings alone do not select a shared slope. After compensating with item effects, the ratings likelihood is flat (orange). The item-effect restriction creates assumption-dependent peaks (blue), which move as σ_p changes. The dashed line marks the reported fit.

This is the additive multilevel item structure form with random residuals (AMIS-R) (Cho et al., 2014). The identifying assumptions are explicit in its two rows: deployment conditions residual item quality on the observed features x_i, whereas the audit also conditions it on the human label g_i. The ratings identify u_i, while these mean-independence assumptions assign its covariate-predictable part to $\bar{\beta}$ and the remainder to e_i. Thus $\bar{\beta}$ is the L^2 projection of item consensus onto x_i, conditional also on g_i in the audit fit. The ratings inform u_i; the residual restriction, not cross-judge rating variation, selects a point on the likelihood ridge.

The profile likelihood confirms this distinction. As Figure 1 shows, varying $\bar{\beta}$ while compensating with item effects changes the ratings log-likelihood by only 4.5×10^{-13} nats. Relaxing the item-effect restriction moves several shared slopes, whereas the centered judge contrasts remain stable (Kendall's $\tau \geq 0.91$ across the sensitivity fits). This sensitivity matters because NIST relevance is itself associated with the item features ($R^2 = 0.13$; overlap coefficient $+0.077$), so the projection can absorb quality-related variation and a panel-wide feature association may partly reflect item quality. Although the audit includes the observed label g_i, unmeasured dimensions of quality remain unanchored.

Ratings alone therefore provide no finite identified interval for $\bar{\beta}$. For each shared slope, we report how much dependence between the item residual and the covariate would move its point estimate to zero or make its interval include zero. We also examine sensitivity to the item-effect restriction and slope priors (Appendix J). The adjustment below removes only δ_j and retains $\bar{\beta}^\top x_i + \gamma h_i$ in the item score, making the panel-average response the adjustment reference. That choice is a convention, not evidence that the shared response is unbiased.

3.4 Plug-in fusion objectives and targeted adjustment

Given a panel of ratings, we obtain an item score by inverting the fitted ordinal model. Both rules below account for judge leniency. They differ in which feature responses they treat as removable. The *full* rule removes every fitted feature term:

$$\hat{q}_i^{\text{full}} = \arg\max_q \left[\log p(q) + \textstyle\sum_j \log \Pr(r_{ij} \mid q + \hat{\alpha}_j + \hat{\beta}_j^\top x_i + \hat{\gamma} h_i,\ \hat{\kappa}) \right]. \tag{4}$$

Here $p(q) = \mathcal{N}(0, \sigma_p^2)$ is the regularizing latent-score prior used by the fusion routine. The *targeted* rule removes only the identifiable centered deviations:

$$\hat{q}_i^{\text{targ}} = \arg\max_q \left[\log p(q) + \textstyle\sum_j \log \Pr(r_{ij} \mid q + \hat{\alpha}_j + \hat{\delta}_j^\top x_i,\ \hat{\kappa}) \right]. \tag{5}$$

Because the targeted objective omits the shared terms $\hat{\bar{\beta}}^\top x_i + \hat{\gamma} h_i$ from its linear predictor, their contribution remains inside $\hat{q}_i^{\text{targ}}$. We retain the same $\mathcal{N}(0, \sigma_p^2)$ regularizer in both objectives to hold fusion regularization fixed. It is not the transformed generative prior for the targeted score, which would depend on x_i; changing only its variance would not recover that feature-dependent prior. Both objectives plug in posterior point estimates rather than integrating over parameter uncertainty. The

targeted score is therefore an objective-specific plug-in adjustment motivated by Proposition 1, not exact maximum a posteriori (MAP) inference under the original generative model. Propagating posterior uncertainty through the adjustment remains future work.

4 Making the Audit Fast

Because models, prompts, and judge panels change, a practical audit must be inexpensive to rerun. The identifiability result and targeted adjustment do not require neural inference; amortization is an implementation layer, not an additional source of scientific identification. Fitting the hierarchical model with MCMC takes about 240 seconds per dataset, so we amortize inference (Cranmer et al., 2020): a network is trained once on simulated parameter–dataset pairs and then approximates the posterior for a new panel in a single forward pass.

Network and training. A shared multilayer perceptron (MLP) embeds each item's one-hot ratings and covariates, and a masked mean pools the item embeddings into a global code. Three decoders map this code to Gaussian approximate posteriors for population-, judge-, and item-level parameters. The judge decoder also receives a learnable identity embedding; without it, judges in the same capability group collapse to identical estimates. Training minimizes a Gaussian negative-log-likelihood on 50,000 simulations from the prior (about 34 minutes of training). In a learning-curve comparison, posterior-estimation accuracy is near its plateau by $N{=}25{,}000$ simulations. Architecture, loss weighting, and optimizer settings are in Appendix R. We treat amortization as an implementation layer and validate it against MCMC using posterior calibration, agreement of posterior means, masked-rating reconstruction, and robustness to distribution shifts and missing ratings (Section 6.5).

5 Experimental Setup

Datasets and judges. Each panel has a distinct role. The primary audit uses TREC web relevance (485 items, 10 judges), and SummEval summarization (180 items, 5 judges) provides an external-domain check. A separate five-judge TREC web collection on the same items is used for calibration, reconstruction, and amortization. Appendix A gives the complete inventory.

The primary TREC web graph contains 4,846 of 4,850 possible ratings and remains connected. "Position" is a constructed, displayed within-query slot: items were sampled by relevance-label stratum and then assigned slots in that order. It is shown in the prompt but is not an organic retrieval rank. "Formality" is a mean-word-length proxy. The hierarchy uses two capability groups for judge intercepts: frontier/reasoning models and non-reasoning/small models.

Controlled anchors. We run two paired counterfactual experiments, each on 50 fixed TREC web items with four judges. The position experiment changes only the displayed rank (1 versus 5); the overlap experiment appends query terms missing from the passage. Each experiment contains 400 ratings, forming 200 within-item, within-judge pairs. The position manipulation holds content fixed, whereas the overlap manipulation changes the passage wording and length. It therefore does not isolate lexical overlap from accompanying textual changes. Appendices B and C give the complete protocols.

Inference and evaluation. We fit the hierarchical Bayesian measurement (HBM) models in PyMC (Abril-Pla et al., 2023) with the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014). Each full-panel fit uses four chains with 1,500 tuning and 1,500 sampling steps; these fits have $\hat{R} \leq 1.01$ and zero divergences. Across the 50 cross-validation fits, $\hat{R} \leq 1.02$, with three divergences in total (Appendix P).

For human-label alignment on TREC, we use ten repeated five-fold splits by query. Standardization and posterior fitting use only training queries, and held-out items are scored without their labels. We report Spearman correlation with the human labels. Separately, we report masked-rating reconstruction at scoring time: the target rating is omitted from fusion, but the target judge's parameters are estimated once from the complete five-judge panel. That diagnostic is therefore not new-judge generalization (Appendix H).

Table 1: Feature-response taxonomy on 10-judge TREC web. The association μ_c is assumption-dependent; σ_c and $P(\sigma_c{>}0.05)$ quantify identifiable between-judge heterogeneity.

Feature	Association μ_c	σ_c	$P(\sigma_c{>}0.05)$	Class
Lexical overlap	$+0.50$	0.141	1.00	Partly judge-specific
Verbosity	$+0.19$	0.039	0.32	Effectively shared
Formality	$+0.20$	0.041	0.33	Effectively shared
Query length	-0.12	0.063	0.63	Uncertain
Position	-0.00	0.115	0.98	Partly judge-specific

6 RESULTS

6.1 WHICH FEATURE RESPONSES VARY ACROSS JUDGES?

Lexical overlap is partly judge-specific ($\sigma_{ov} = 0.14$ with ten judges and 0.18 with five), and position has the second-largest heterogeneity ($\sigma_{pos} = 0.115$); position responses point in opposing directions across judges even though the fitted population mean is near zero. Verbosity and formality are effectively shared ($\sigma_c \approx 0.04$), while query length remains uncertain. For this classification we use $\varepsilon = 0.05$ latent units as a prespecified threshold for practically meaningful heterogeneity: values of $P(\sigma_c > \varepsilon)$ near one support a partly judge-specific response, low values support an effectively shared response, and intermediate values remain uncertain.

Table 1 reports these results through two quantities with different identification status. Between-judge heterogeneity σ_c is identified by the ratings; the population association μ_c is the assumption-dependent projection described in Section 3.1. Targeted adjustment therefore uses the centered deviations δ_j^c rather than μ_c, although the rating matrix alone does not establish that those deviations should be removed.

The μ_c estimates depend on the item-effect restriction. As that restriction is relaxed, the position mean moves to -0.26 and the formality mean approaches zero; neither is directly measured by the ratings. The positive overlap association is the only one whose direction stays clear across the full sensitivity sweep, with posterior probability of a positive sign at least 0.96 at every point. Appendix J gives the sweep, and Appendix Q examines whether apparent heterogeneity can instead reflect judge-specific response scales.

6.2 WHAT DOES CONTROLLED POSITION VARIATION IDENTIFY?

These experiments are cautionary case studies: they test concrete prompt or text interventions rather than validate observational feature associations as causal biases. The paired position comparison identifies the average effect of changing displayed rank from 1 to 5 for this prompt. Proposition 1 applies when position is observed across different items, where a panel-wide slope can be absorbed into item quality. The controlled panel breaks this ambiguity for one concrete intervention by asking each judge to rate the same item with displayed ranks 1 and 5 (protocol in Appendix B). Because document content is fixed, the within-item, within-judge contrast identifies the shared effect of changing displayed rank; judge-level departures estimate heterogeneous responses on the same scale.

Changing displayed rank from 1 to 5 changes the rating by $+0.035$ points on the 0–3 scale (query-clustered 95% confidence interval (CI) $[-0.015, +0.090]$), or $+0.018$ per standardized position unit (CI $[-0.010, +0.045]$). For this prompt and rank change, the interval is compatible with a small positive, zero, or small negative effect. The common effect is identified experimentally rather than selected by the observational HBM's item-effect restriction. Counterfactual position deviations and observational δ_j^{pos} have positive rank alignment across the four overlapping judges (Spearman 0.63); given the small number of judges, this comparison is descriptive.

The correction diagnostic in Figure 3 shows the operational consequence: subtracting only judge-specific counterfactual deviations leaves the common slope, while subtracting the anchored common-plus-specific effect removes it. Here, unlike in the observational analyses, the shared component is experimentally identifiable. Identification alone does not establish that removing it is desirable. For

Table 2: Spearman rank correlation with human labels under full and targeted adjustment; higher is better. Targeted adjustment has the higher point estimate in all three settings but remains at or below the panel mean. Intervals use setting-specific, cluster-aware resampling procedures (including a query-clustered bootstrap for the deployment row) described in Appendices P and O. The last column is a mean paired difference, not a difference of the columns to its left.

Setting	Panel mean	Full	Targeted	Targeted − full
TREC audit, cross-fitted	0.688	0.661	0.681	$+0.021$ $[-0.002, 0.044]$
TREC deployment, no labels	0.688	0.619	0.688	$+0.069$ [0.040, 0.099]
SummEval audit	0.597	0.553	0.578	$+0.023$ $[-0.037, 0.081]$

observational associations without a corresponding controlled intervention, the targeted rule retains the shared component.

6.3 What does manipulating lexical overlap identify?

Lexical overlap has a larger observational association with ratings than position and accounts for most of the alignment loss under full adjustment. In the five-judge observational fit, its population association is $\hat{\mu}_{ov}$=0.45 and its between-judge scale is σ_{ov}=0.18; Table 1 reports the corresponding 10-judge estimates. We interpret this heterogeneity cautiously: it may partly reflect differences in response scale or discrimination rather than lexical-overlap sensitivity (Appendix Q). We therefore test a paired intervention that increases overlap, although it also changes passage wording and length (protocol in Appendix C; placebo and templated-preamble robustness checks in Appendix D). For 50 items and four judges, the high-overlap arm appends a marked sentence containing the query terms missing from the passage. It raises mean lexical overlap from 0.346 to 0.792 ($\Delta = +0.445$, 95% CI $[+0.378, +0.510]$) and increases mean length from 55.6 to 62.1 words. Unlike the position intervention, it changes the text itself.

Despite the large change in overlap, the mean rating contrast is near zero. The high-overlap minus original contrast is $+0.005$ points on the 0–3 scale (query-clustered 95% CI $[-0.055, +0.060]$), and the two-way fixed-effect slope is $+0.014$ per standardized overlap unit (CI $[-0.015, +0.041]$). For this query-term insertion, the shared-effect point estimate is near zero and its confidence interval includes zero. The experiment does not validate the large observational shared slope: $\hat{\mu}_{ov}$ is an item structure projection and can absorb relevance-related variation associated with overlap. The counterfactual judge deviations also do not reproduce the observational overlap contrasts (Spearman 0.00 across the four judges). This small-sample comparison is descriptive and is consistent with Appendix Q's warning that lexical sensitivity is difficult to separate from response-scale differences. Thus, for this insertion, the interval includes both negative and positive values (Figure 4) and does not justify subtracting the large shared observational overlap association as removable judge behavior. On an additional judge panel, point estimates for displayed-position changes and marked query-term insertion range from -0.032 to $+0.027$ rating points, whereas a templated query-focused preamble raises ratings by $+0.128$ points. An anchor therefore identifies a concrete intervention, not an abstract feature name (Appendix D).

6.4 Does targeted adjustment preserve human-label alignment?

We use targeted adjustment here as a diagnostic control, not as a proposed replacement for panel averaging. It has a higher human-label rank-correlation point estimate than full adjustment in all three settings but does not outperform the panel mean. Full adjustment removes both shared and judge-specific responses; targeted adjustment removes only centered deviations and approximately preserves panel consensus (Table 2).

On TREC, targeted adjustment has a higher out-of-fold Spearman correlation than full adjustment in all ten query-grouped cross-validation repetitions. The mean gain is $+0.021$, but the corrected interval is consistent with no average difference (95% CI $[-0.002, 0.044]$). Targeted adjustment also remains slightly below the panel mean, by -0.004 on average (CI $[-0.010, 0.003]$); that smaller difference is likewise unresolved.

When human labels are excluded from model fitting, correlations are 0.688 under targeted adjustment and 0.619 under full adjustment. Fitting and evaluation use the same items, so this remains an in-sample diagnostic; it shows that the pattern is not created solely by including labels in the audit posterior. The query-clustered bootstrap gives targeted minus full $= +0.069$ (95% CI $[0.040, 0.099]$).

SummEval (Fabbri et al., 2021) provides a public cross-domain check. Because expert scores enter the audit posterior, this is an in-sample mechanism check rather than out-of-sample label generalization. Full adjustment reaches correlation 0.553 with expert scores, compared with 0.597 for the panel mean; targeted adjustment reaches 0.578. Its gain over full adjustment is $+0.023$ (95% CI $[-0.037, 0.081]$), which is compatible with no difference, and targeted adjustment trails the panel mean by 0.020.

The covariate taxonomy itself is domain-dependent. On TREC web relevance, overlap and position show the clearest heterogeneity. On SummEval, displayed position is not a judge-visible feature and is excluded from the fit; heterogeneity is elevated for verbosity and source-article length (Appendix E). The decomposition must therefore be re-estimated rather than transferred across tasks.

Which method best reconstructs withheld ratings? Graded-response IRT best reconstructs masked ratings (mean absolute error (MAE) 0.270), followed by the label-free deployment HBM (0.292) and panel mean (0.414). IRT models judge-specific discrimination but, as implemented here, does not split feature responses into shared and judge-specific components. Because target-judge parameters are fitted to the complete panel, this diagnostic measures within-panel prediction, not human alignment or new-judge generalization. Appendices H and M give the two protocols.

6.5 Does amortization preserve the conclusions?

The amortized network closely reproduces MCMC parameter estimates and masked-rating reconstruction at lower cost, validating a fast implementation of the MCMC audit rather than a separate scoring method. Under the simulator, simulation-based calibration (SBC) ranks are near-uniform (mean Kolmogorov–Smirnov (KS) statistic 0.028), with near-nominal marginal coverage (Figure 5 and Table 18). Fitted judge intercepts and feature slopes correlate > 0.95 with MCMC. Item recovery is lower ($\rho \approx 0.52$–0.54), reflecting the information in five ordinal ratings. Reconstruction MAE is 0.302 versus MCMC's 0.304, while runtime falls from 240 s to 1.3 ms. Appendix S reports the robustness tests.

7 Discussion and Conclusion

The paper's primary contribution is a diagnostic audit of what judge ratings can support. The results distinguish prediction, scoring, and correction: graded-response IRT best predicts withheld judge ratings, the panel mean aligns best with human labels, and the HBM separates identifiable cross-judge contrasts from panel-shared associations confounded with item quality. Masked-rating prediction therefore does not determine human-label alignment or justify removing fitted feature components.

Ratings identify differences among judges' feature responses, but not a panel-wide response separately from item quality. Targeted adjustment is thus a diagnostic control: it removes centered deviations while retaining shared components, making the implications of the two choices explicit. Its human-label rank-correlation point estimate exceeds full adjustment in all three settings but is no higher than the panel mean. On TREC, the corrected targeted-minus-full interval includes zero, and the smaller targeted-minus-mean difference remains unresolved.

The paired displayed-rank and query-term-insertion experiments identify only the tested interventions. The insertion estimate is near zero, whereas the templated query-focused preamble increases ratings by 0.128 points. Neither result identifies a universal effect of an abstract feature. Other limitations include response-scale ambiguity for overlap, plug-in scores that omit parameter uncertainty, and possible omitted features or residual judge dependence. Users should select a scoring rule for the deployment objective and use human-label validation when available. Correction requires substantive justification for judge-specific deviations and an external anchor for a shared response.

Artificial Intelligence Use Statement

Artificial-intelligence-assisted tools supported code and manuscript editing. The authors reviewed and verified all mathematical claims, empirical results, code changes, and final text.

Reproducibility Statement

Appendix G documents model priors, feature computation, baseline implementations, fitting seeds, evaluation protocols, and the settings needed to reproduce the reported analyses. Code and curated artifacts will be released with the camera-ready paper upon acceptance.

References

Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C. Luhmann, Osvaldo A. Martin, Michael Osthege, Ricardo Vieira, Thomas Wiecki, and Robert Zinkov. PyMC: A modern and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. LLMs can be fooled into labelling a document as relevant: Best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 32–41, 2024. doi: 10.1145/3673791.3698431. URL https://doi.org/10.1145/3673791.3698431.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.01453.

Negar Arabzadeh and Charles L. A. Clarke. Benchmarking LLM-based relevance judgment methods. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3194–3204, 2025. doi: 10.1145/3726302.3730305. URL https://doi.org/10.1145/3726302.3730305.

Timo M. Bechger and Gunter Maris. A statistical test for differential item pair functioning. *Psychometrika*, 80(2):317–340, 2015. doi: 10.1007/s11336-014-9408-y.

Bob Carpenter. Multilevel Bayesian models of categorical data annotation. Technical report, Alias-i, 2008.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement biases. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2402.10669.

Wei Chen, Jia Li, Yin Ouyang, and Gongjun Xu. DIF statistical inference without knowing anchoring items. *Psychometrika*, 88(4):1097–1122, 2023. doi: 10.1007/s11336-023-09930-9.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Sun-Joo Cho, Paul De Boeck, Susan Embretson, and Sophia Rabe-Hesketh. Additive multilevel item structure models with random residuals: Item modeling for explanation and item generation. *Psychometrika*, 79(1):84–104, 2014. doi: 10.1007/s11336-013-9360-2.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of LLM-as-a-judge via item response theory. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.00521.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020. doi: 10.1111/rssb.12348.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Alexander P. Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.

Paul De Boeck and Mark Wilson (eds.). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, 2004. doi: 10.1007/978-1-4757-3990-9.

Philipp Doebler. Looking at DIF from a new perspective: A structure-based approach acknowledging inherent indefinability. *Applied Psychological Measurement*, 43(4):303–321, 2019. doi: 10.1177/0146621618795727.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. Oral; arXiv:2410.13341.

Magda Dubois, Harry Coppock, Mario Giulianelli, Timo Flesch, Lennart Luettgau, and Cozmin Ududec. Skewed score: A statistical framework to assess autograders, 2025. URL https://arxiv.org/abs/2507.03772.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.04475.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.04291.

Yicheng Gao, Gonghan Xu, Zhe Wang, and Arman Cohan. Bayesian calibration of win rate estimation with LLM evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4757–4769. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.273. URL https://aclanthology.org/2024.emnlp-main.273/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on LLM-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Paul Gustafson. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, 20(2):111–140, 2005. doi: 10.1214/088342305000000098.

René Hofer, Joshua Maynez, Bhuwan Dhingra, Adam Fisch, Amir Globerson, and William W. Cohen. Bayesian prediction-powered inference. *arXiv preprint arXiv:2405.06034*, 2024.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of NAACL-HLT*, pp. 1120–1130, 2013.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004. doi: 10.1111/j.1468-0262.2004.00555.x.

Julia Kopf, Achim Zeileis, and Carolin Strobl. Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, 75(1):22–56, 2015. doi: 10.1177/0013164414529792.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations, 2025. URL https://arxiv.org/abs/2511.21140.

John Michael Linacre. Many-facet Rasch measurement. *MESA Press*, 1994.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980.

Francesca Molinari. Microeconometrics with partial identification. *Handbook of Econometrics*, 7A: 355–486, 2020. arXiv:2004.11751.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3): 239–281, 2003. doi: 10.1023/A:1024068626366.

Harrie Oosterhuis, Rolf Jagerman, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Reliable confidence intervals for information retrieval evaluation using generative A.I. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024. doi: 10.1145/3637528.3671883.

Rebecca J. Passonneau and Bob Carpenter. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, 2014.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and LLM judgments: Understanding and narrowing the gap. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/160f484617c9e3ce84332d8cbc8ba62b-Abstract-Conference.html.

Anu Pradhan, Alexandra Ortan, Apurv Verma, and Madhavan Seshadri. LLM-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. In *Proceedings of the 2nd Workshop on Evaluating and Applying Recommender Systems with Large Language Models (EARL), co-located with the 19th ACM Conference on Recommender Systems (RecSys)*, 2025. URL https://arxiv.org/abs/2509.12382.

Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Judging the judges: A collection of LLM-generated relevance judgements, 2025. URL https://arxiv.org/abs/2502.13908.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1969.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 2025 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL-IJCNLP)*, 2025. arXiv:2406.07791.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. UMBRELA: UMbrela is the (open-source reproduction of the) bing relevance assessor, 2024. URL https://arxiv.org/abs/2406.06519.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. A large-scale study of relevance assessments with large language models: An initial look, 2025. URL https://arxiv.org/abs/2411.08275.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Ellen M. Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum*, pp. 355–370, 2001.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. In *NeurIPS 2024 Workshop on Safe Generative AI*, 2024. arXiv:2410.21819.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2009.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. When can you debias an LLM judge? identifiability limits, a test, and designs for top-k ranking, 2026. URL https://arxiv.org/abs/2607.02104.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge, 2024. URL https://arxiv.org/abs/2410.02736.

Chuting Yu, Hang Li, Guido Zuccon, Joel Mackenzie, and Teerapong Leelanupab. When LLM judges inflate scores: Exploring overrating in relevance assessment. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026. URL https://arxiv.org/abs/2602.17170.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation, 2026. URL https://arxiv.org/abs/2603.00039.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A Dataset inventory

Table 3 lists every observational rating panel used in the paper. The five-judge calibration panel in the final row is a separate rating collection on the same TREC web items. The controlled position and overlap experiments each use 50 TREC web items and are described with their protocols in Appendices B and C. In the five-judge row, human labels enter only the audit fit.

B Controlled position-counterfactual protocol

The position-anchor experiment uses a fixed sample of 50 TREC web items from the calibration panel. It preserves the query, document identifier, document text, NIST label, and all non-position features. For each item–judge pair, it creates two prompts that differ only in the displayed position string from the original relevance prompt:

```
Document (position {position} in results):
```

Table 3: Datasets, tasks, and rating counts. SummEval carries scores from three independent human experts. MS MARCO denotes Microsoft Machine Reading Comprehension; CNN/DM denotes CNN/Daily Mail.

Dataset	Task	Corpus	Items	Judges	Ratings	Labels	Used for
TREC-DL	Web relevance	MS MARCO	485	10	4,846	NIST 0–3 (human)	Feature taxonomy, audit
SummEval	Summarization	CNN/DM	180	5	900	Human (3 experts)	External check
5-judge calibration panel (TREC web)					2,425	fit only (audit)	Reconstruction, amortized

with `position` $\in \{1, 5\}$. Every item–judge pair receives both prompt variants. We randomize only the order in which the variants are collected, using a fixed seed, and save results incrementally. The reported live run uses four judges: `claude-sonnet-4`, `claude3-haiku`, `gpt-4.1-nano`, and `gpt-5.1`.

The primary estimand is the paired high-minus-low contrast

$$\tau = \mathbb{E}\left[r_{ij}(5) - r_{ij}(1)\right],$$

where the expectation is over the fixed items and judges in the controlled panel. As a scale check, we also regress ratings on standardized displayed position with item and judge fixed effects. Judge-specific effects are $\tau_j - \bar{\tau}$. We evaluate two anchor-informed corrections. The targeted analogue subtracts only the centered judge deviation $\tau_j - \bar{\tau}$. The fully anchored correction subtracts τ_j, which includes the shared effect $\bar{\tau}$. Because the paired intervention identifies $\bar{\tau}$, only the fully anchored correction can remove the shared component; the item-effect restriction cannot.

C Controlled lexical-overlap protocol

The overlap-anchor experiment uses a separate fixed sample of 50 TREC items and four judges. The original arm presents the full passage at its original position. The high-overlap arm appends one marked sentence:

`Additional query terms: {missing query terms}.`

The inserted terms appear in the query but not in the original passage under the paper's lowercased whitespace tokenization. Items with no missing query terms are excluded before sampling. We recompute all document-derived features for each variant, including lexical overlap, verbosity, and formality. Section 6.3 therefore reports the accompanying length change as part of the intervention.

The primary estimand is the paired contrast

$$\tau_{ov} = \mathbb{E}\left[r_{ij}(\text{high overlap}) - r_{ij}(\text{original})\right].$$

The fixed-effect scale check regresses ratings on the realized within-panel overlap_z, with item and judge fixed effects. We evaluate a centered correction that subtracts judge deviations and a fully anchored correction that subtracts the entire counterfactual judge slope. The latter is an experimental diagnostic, not automatically a production-scoring prescription, because the intervention changes the document text and may change perceived relevance.

D Counterfactual robustness checks

Table 4 summarizes live-call robustness checks on a second five-model panel distinct from the primary counterfactual panel: `claude-sonnet-4.5`, `gemini-2.5-pro`, `gpt-4o`, `gpt-5`, and `llama-3.1-70b-chat`. Two caveats apply. First, for two of the five endpoints our parser recovered no rating from a large share of responses under the pairwise and prompt-variant formats; we did not tune those prompts per endpoint, so we attribute this to the prompt–parser combination rather than to the models. Those rows therefore use the three endpoints whose responses parsed under the shared format. Second, the templated-preamble check prepends the following fixed sentence to the original passage:

```
For the query topics {terms}, this passage provides
the following information.
```

It is a less controlled query-focus manipulation, not a rewrite or a pure lexical-overlap intervention.

Table 4: Counterfactual robustness checks. Effects are rating-point changes on the 0–3 scale. The pairwise row instead reports the change in preference for the canonical higher-label document when shown first rather than second; A and B denote the documents in a pair.

Check	Manipulation	Judges/items	Effect (95% CI)
Extra-model position	position 1 vs 5	5 / 50	-0.024 $[-0.053, +0.004]$
Pairwise order	A–B vs. B–A order swap	3 / 50 pairs	-0.060 $[-0.113, -0.020]$
Ignore-position prompt	position 1 vs 5	3 / 50	$+0.000$ $[-0.053, +0.047]$
Search-results prompt	position 1 vs 5	3 / 50	$+0.027$ $[-0.027, +0.080]$
n=100 position	position 1 vs 5	5 / 100	-0.026 $[-0.050, -0.004]$
Marked overlap	query-term append	5 / 50	$+0.020$ $[-0.024, +0.064]$
Placebo append	neutral-term append	5 / 50	$+0.044$ $[+0.012, +0.080]$
Marked vs placebo	query terms vs neutral terms	5 / 50	-0.024 $[-0.060, +0.012]$
Templated preamble	query-focused preamble	5 / 50	$+0.128$ $[+0.088, +0.168]$
n=100 overlap	query-term append	5 / 100	-0.032 $[-0.065, -0.002]$

Across the single-document position checks, point estimates range from -0.026 to $+0.027$ points. The pairwise-order estimate is -0.060 (95% CI $[-0.113, -0.020]$), so the single-document results should not be generalized to pairwise or other multi-document evaluation interfaces. For lexical overlap, the marked-minus-placebo contrast is -0.024 (95% CI $[-0.060, +0.012]$), whereas the templated-preamble estimate is $+0.128$ (95% CI $[+0.088, +0.168]$). These results show that an external anchor identifies the causal effect of a specific intervention; it does not justify treating every shared observational overlap association as removable bias.

E Cross-domain decomposition

Table 5 compares TREC web relevance with SummEval, a public non-IR summarization benchmark with human expert labels. "Overlap" denotes query–document term overlap for TREC and source-article–candidate-summary term overlap for SummEval. The two overlap measures use lowercased whitespace tokens, no stopword removal, and normalization by query or article vocabulary size. "Query/article length" denotes query length for TREC and source-article length for SummEval.

The threshold ε=0.05 is prespecified as practically meaningful heterogeneity. The human-alignment block uses the five-judge TREC calibration panel, rather than the ten-judge panel used above it, and reports in-sample audit values. The cross-fitted TREC result appears in Table 2; its out-of-fold full-adjustment value is 0.661, rather than the in-sample value 0.669 below.

F Per-judge intercepts (10 judges)

The intercept spread is 1.75 latent units. In this panel, the two models labeled "Frontier" have the two most negative fitted intercepts. The hierarchy uses the two capability groups defined in Section 5; the labels in the last column provide a finer descriptive classification.

G Baselines and reproducibility

Baselines. The first Bayesian ordinal baseline is a Many-Facet Rasch Model (MFRM), a Rasch model with a judge-severity facet. The other baselines are graded-response IRT and an ordinal mixed-effects model. MFRM uses $\eta_{ij} = \theta_i - \beta_j^{\text{sev}}$, shared ordered thresholds, and sum-to-zero severities. Graded-response IRT adds per-judge discrimination, $\eta_{ij} = a_j\theta_i - \beta_j^{\text{sev}}$ with $a_j \sim$ LogNormal. Its fits have $\hat{R} \leq 1.02$ and effective sample size (ESS)>470. The ordinal mixed-effects model contains only random judge and item intercepts with an ordered-probit link. We fit these models and the HBM with NUTS using the settings in Section 5.

Table 5: Cross-domain comparison of population associations, between-judge heterogeneity, and in-sample human-label alignment. The shared-versus-judge-specific decomposition is reported rather than population means alone.

	TREC web	SummEval (non-IR)
σ_p/σ_m	3.4×	1.1×
Intercept spread	1.75	2.80
Population association μ_c		
Overlap	$+0.50$	$+0.49$
Verbosity	$+0.19$	$+0.08$
Position	-0.00	n/a
Between-judge heterogeneity σ_c $[P(\sigma_c{>}0.05)]$		
Overlap	0.141 [1.00]	0.108 [0.72]
Verbosity	0.039 [0.32]	0.173 [0.92]
Position	0.115 [0.98]	n/a
Query/article length	0.063 [0.63]	0.190 [0.98]
Human-label alignment, in-sample audit posterior (TREC: 5-judge calibration panel)		
Panel mean	0.688	0.597
Full adjustment	0.669	0.553
Targeted adjustment	0.686	0.578

Table 6: Estimated judge intercepts $\hat{\alpha}_j$. A larger $\hat{\alpha}_j$ shifts ratings upward (greater leniency); the most negative values are the harshest judges.

Judge	Intercept $\hat{\alpha}_j$ (leniency)	Descriptive type
llama-3.2-3b	$+0.40$	Small
claude3-haiku	$+0.04$	Small
gemini-2.5-flash	-0.19	Mid
llama-3.1-405b	-0.31	Large
gpt-4.1-nano	-0.50	Small
o4-mini	-0.74	Reasoning
llama-3.1-70b	-0.79	Mid
claude-sonnet-4	-0.94	Frontier
gpt-4.1	-1.22	Mid
gpt-5.1	-1.34	Frontier

Dawid–Skene is fit by expectation–maximization over per-judge confusion matrices with a class prior. Our inverse-variance baseline, labeled GLAD-lite, weights judges by inverse rating variance; it is not the full generative GLAD model. Z-score uses each judge's marginal mean and standard deviation. All methods use the same five-judge panel and masked-rating reconstruction protocol. The model-based methods predict each masked rating through the same latent-quality fusion step.

Feature computation. Judges were shown the full passage. Verbosity is the full-text word count. Lexical overlap is the fraction of query terms appearing in the document, and formality is mean word length. Both are computed from lowercased whitespace tokens in the full text, without stopword removal. Position is the displayed within-query rank, and query length is the number of query terms. All five features are standardized before fitting.

HBM priors. The audit HBM uses $\mu_t \sim \mathcal{N}([0.2, -0.2], 0.5)$, $\sigma_m \sim \mathrm{HalfNormal}(0.3)$, $\sigma_p \sim \mathrm{HalfNormal}(0.8)$, and $\beta_{gt} \sim \mathcal{N}(1.5, 0.5)$. Feature population slopes use $\mu_{verb} \sim \mathcal{N}(0.2, 0.15)$, $\mu_{pos} \sim \mathcal{N}(-0.3, 0.2)$, $\mu_{ov} \sim \mathcal{N}(0.3, 0.2)$, $\mu_{form} \sim \mathcal{N}(0, 0.2)$, and $\mu_{qlen} \sim \mathcal{N}(0, 0.2)$. The heterogeneity scales $\sigma_{verb}, \sigma_{ov}, \sigma_{form}, \sigma_{qlen}$ follow $\mathrm{HalfNormal}(0.15)$ priors; σ_{pos} follows $\mathrm{HalfNormal}(0.25)$. The position–verbosity interaction has prior $\mathcal{N}(0, 0.1)$, and the ordered Normal cutpoint priors are $\mathcal{N}([-1.5, 0, 1.5], 1.0)$, subject to ordering.

Reproducibility. This manuscript reports the HBM, baselines, simulator, amortized-network settings, fitting seeds (NUTS seed 123), and collection procedures for the 8,171 observational judge ratings used in the fitted-panel analyses. Full code and configurations will be released upon acceptance. The controlled experiments contribute additional ratings documented in Appendices B, C, and D.

The label-free deployment rows in Tables 2 and 13 use the posterior fit without human labels. The controlled position analysis, ablation study, and few-shot study follow the protocols in Appendices B, L, and M, respectively.

Judges were queried through a unified gateway at temperature 0 for non-reasoning models; reasoning models used completion-token limits. We collected all ratings between June and July 2026 from the then-current hosted aliases: `gpt-5.1`, `claude-sonnet-4`, `o4-mini`, `gpt-4.1`, `gpt-4.1-nano`, `gemini-2.5-flash`, `llama-3.1-405b/70b`, `llama-3.2-3b`, and `claude3-haiku`. Hosted endpoints may drift, so the archived ratings used in our analyses fix the empirical results and all sampler and training seeds are set. We recommend recollecting ratings when reproducing on a later application programming interface (API) snapshot.

Upon acceptance, the artifacts will contain ratings, real MS MARCO and SummEval item identifiers, and derived features, but not the underlying query, passage, article, or summary text. This is consistent with the source datasets' terms; the identifiers allow that text to be fetched directly from MS MARCO or SummEval. The judge panels include Meta Llama models; the required attribution is "Built with Meta Llama."

H MASKED-RATING RECONSTRUCTION WITH FITTED JUDGE PARAMETERS

Evaluation setup. For each target rating, the other judges' ratings for the same item are fused to predict the target judge's ordinal score. The target rating is excluded from this fusion calculation, but it contributes to model fitting as described below. The target judge's marginal mean and standard deviation are not used in the primary prediction. The deployment HBM is fitted without human labels; the audit HBM uses a label-informed posterior. All model-based baselines use the same latent-quality fusion step.

The Bayesian models (HBM-audit, HBM-deployment, MFRM, ordinal mixed-effects, and graded-response IRT) are each fit once on the complete five-judge panel; there is no judge-level refitting. The holdout is applied only at scoring time. We infer latent quality from the other four judges' ratings on the item. The target judge's fitted intercept, slopes, severity, or discrimination—depending on the model—then predicts its withheld rating. Because each Bayesian model is fitted to the complete panel, every withheld rating contributes to the fitted target-judge parameters. Each rating is one of 2,425, so its individual influence is small, but the evaluation is not strictly out of sample. Dawid–Skene instead re-estimates its confusion matrices without any ratings from the target judge, and its prediction uses no target-judge parameter.

The label-free deployment HBM reduces MAE from 0.414 for panel-mean aggregation to 0.292, but graded-response IRT performs best at 0.270 (Table 7). The HBM improves on panel averaging but does not achieve the lowest reconstruction error. Its role in the main paper is to expose covariate-specific cross-judge heterogeneity.

The target-independent z-score baseline maps predictions back using only the observed judges' marginal statistics and performs worse than panel-mean aggregation. For comparison, Table 7 also reports an oracle variant that uses the target judge's own mean and standard deviation. It reaches 0.299, but both statistics are computed from the same 485 ratings being predicted. The HBM's fitted target-judge parameters also depend on these ratings, although that dependence is distributed across the joint fit.

Appendix M quantifies the difference. When the target judge is removed from fitting and its intercept is estimated from 50 of its ratings, new-judge MAE is 0.325, compared with 0.304 for the fully in-sample audit posterior. This 0.021 gap reflects the combined advantage of seeing the target judge, its full history, and its feature slopes. It is not a formal estimate of test-rating leakage because this comparison cannot separate those three sources. The adapted new judge retains a 0.088 improvement over panel-mean aggregation.

I IDENTIFIABILITY CONDITIONS (DETAIL)

We make the conditions of Proposition 1 precise. Consider the ordered-probit model $P(r_{ij} \leq k) = \Phi(\kappa_k - \eta_{ij})$, with $\eta_{ij} = \theta_i + \alpha_j + \beta_j^\top x_i$, over J judges and M items.

Table 7: Masked-rating reconstruction on the five-judge TREC panel. Lower MAE and higher accuracy are better. This evaluates reconstruction for known judges, not new-judge generalization; Appendix M reports the new-judge protocol. The deployment HBM uses no human labels, whereas the audit rows use a label-informed posterior.

Method	Reconstruction MAE	Accuracy	Relative MAE vs. panel mean
GLAD-lite	0.448	56.5%	+8.4%
Z-score	0.434	57.6%	+4.9%
Dawid–Skene	0.420	60.5%	+1.5%
Panel mean	0.414	60.0%	n/a
Ordinal mixed-effects	0.319	68.5%	−22.9%
MFRM	0.306	69.7%	−26.0%
HBM-audit (with labels)	0.304	69.9%	−26.6%
HBM-audit (amortized)	0.302	70.0%	−27.0%
HBM-deployment (no labels)	0.292	71.1%	−29.5%
Graded-response IRT	**0.270**	**73.2%**	**−34.8%**
Oracle using the target judge's full-sample statistics:			
Z-score (target statistics)	0.299	70.4%	−27.8%

(C1) Location and scale. Adding a constant c to every η_{ij} and κ_k leaves the likelihood unchanged. In a latent-variable formulation with a free residual standard deviation, multiplying η, the cutpoints, and that standard deviation by $s > 0$ also leaves the probabilities unchanged. Our probit formulation fixes the residual variance to 1 and thereby fixes scale. For location, an abstract identifiability statement can impose sum-to-zero judge effects and slope contrasts. In the PyMC implementation, centered hierarchical priors, the item-quality prior $\theta_i \sim \mathcal{N}(0, \sigma_p^2)$, and cutpoint locations regularize location. We center reported slope contrasts after fitting by subtracting the finite-panel posterior mean. These choices resolve global location and scale for computation and reporting, but not the shared-slope/item-effect invariance in Proposition 1.

(C2) Connected observation graph. Form a bipartite judge–item graph with an edge whenever judge j rates item i. Identifying relative judge intercepts requires this graph to be connected. The TREC web graph is nearly complete: four of 4,850 possible edges are missing across four items, and the graph remains connected. The other panels are complete, so all satisfy this condition.

(C3) Non-degenerate category usage. Each ordinal category must have positive probability, or its cutpoint is not identified. All four categories are populated in every domain.

(C4) Feature variation and contrast identification. The observed features must vary across items and must not be collinear to identify the centered cross-judge slopes δ_j. Full column rank of X prevents confounding among observed features. It does not separate the common item-level effect $\bar{\beta}^\top x_i$ from unrestricted item effects θ_i. At the item level, each reported five-column feature matrix has rank five. The maximum absolute pairwise correlation is 0.30 in the primary TREC panel; pairwise correlations provide a descriptive collinearity check, not a proof of full rank.

(C5) Item and judge variation. The condition $\sigma_p > 0$ ensures item-quality variation. Identifying δ_j also requires $J \geq 2$ judges and feature variation (C4). Once location and scale are fixed (C1), distinct judge intercepts are not required to identify item effects. With identical judges, replicated measurements still identify item effects, but the contrasts δ_j collapse. Here $J = 10$, the fitted item-effect scale is $\sigma_p = 1.65$, and the intercept spread is 1.75.

Under (C1)–(C5), the ratings likelihood identifies relative judge intercepts, cutpoints, item-consensus scores $u_i = \theta_i + \bar{\beta}^\top x_i$, and centered judge-specific slope contrasts. The finer decomposition of u_i into unrestricted item quality θ_i and a shared covariate effect $\bar{\beta}^\top x_i$ remains unidentified without external anchors or additional structural restrictions. The Bayesian prior may regularize this decomposition, but it does not remove the likelihood invariance established in Proposition 1. Clean convergence ($\hat{R} \leq 1.01$, 0 divergences) is evidence about computation, not identification.

J THE SHARED SLOPE IS NOT IDENTIFIED: PROFILE, PROJECTION, SENSITIVITY

This appendix demonstrates Proposition 1 empirically and reports the shared slope as an unidentified parameter. With ratings alone, its identified set is the whole real line. The following results are therefore sensitivity analyses, not bounds on an informative identified set. Each breakdown value states how strong item-residual confounding would have to be to overturn a given interpretation. The profile and breakdown calculations use the fitted 10-judge posterior without refitting; the sensitivity analyses below refit the model as stated.

The likelihood is flat, to machine precision. We fix the judge intercepts, contrasts, cutpoints, and β_{gt} at their posterior means. We then move a shared slope $\bar{\beta}_c + t$ over a grid and profile out the item effect. With θ free, the total log-likelihood varies by only 4.5×10^{-13} nats across $t \in [-1, 1]$ (Figure 1, orange). Relative to a total of $\approx$2540 nats, this is exact flatness up to floating-point precision. Adding the $\mathcal{N}(0, \sigma_p^2)$ penalty turns the same objective into a sharp parabola spanning 306 nats. The item-effect distribution supplies all of this curvature; the ratings supply none.

What the peak represents. Combining the profile curvature with the model's μ_c prior reproduces the sampler's posterior mean for every feature: overlap 0.513 versus 0.502; verbosity 0.191 versus 0.189; formality 0.204 versus 0.202; query length -0.119 versus -0.121; and position 0.009 versus -0.000.

A direct calculation gives the same result. Regressing posterior-mean item consensus $u_i = \theta_i + \bar{\beta}^\top x_i + \beta_{gt} g_i$ jointly on the five features and g_i recovers each μ_c to within 0.015: overlap 0.503 versus 0.502; verbosity 0.182 versus 0.189; position 0.001 versus -0.000; formality 0.217 versus 0.202; and query length -0.134 versus -0.121. Each shared slope is therefore a coefficient in the L^2 projection of item consensus onto the feature vector, conditional on g_i, under $\mathbb{E}[e_i \mid x_i, g_i] = 0$. The mean-independence assumption selects this estimand, and the ratings cannot test that assumption.

Observed quality covariation challenges the assumption. Regressing the normalized NIST label on the five features gives R^2=0.132, with coefficients $+0.077$ (overlap), $+0.097$ (position), $+0.040$ (formality), $+0.025$ (verbosity), and -0.004 (query length). Observed relevance can therefore make the projection absorb quality-related signal. This makes residual mean independence substantively doubtful, even though the restriction itself is not directly testable.

Breakdown analysis. Because the reparameterization is exact, allowing the item residual to have mean $c \cdot x_{ic}$ translates the entire $\bar{\beta}_c$ posterior by $-c$. Two breakdown points follow: the confounding strength that moves the shared slope to zero ($c = \hat{\mu}_c$), and the strength at which its 95% interval first touches zero. We benchmark both against the observed quality channel $\beta_{gt} \times \mathrm{coef}(g \sim x_c)$. The resulting robustness value (RV) expresses the breakdown point in interpretable units (Table 8).

Overlap has the largest breakdown values: $c_{\text{sign}} = 0.502$ and $c_{95} = 0.334$, compared with $c_{95} \leq 0.052$ for every other feature. Its sign is therefore the hardest to overturn. Its $\mathrm{RV} = 1.87$ means that unmeasured quality would need to be $1.87\times$ as strongly associated with overlap as the *entire* NIST channel to flip the sign. Verbosity has a larger RV (2.22) only because its NIST channel is small (0.085), not because its slope is harder to overturn. Position is least robust by both criteria ($\mathrm{RV} = 0$), as expected from $\hat{\mu}_{pos} \approx 0$.

We do not report an Imbens–Manski interval. With ratings alone, the identified set for $\bar{\beta}_c$ is the whole real line, so such an interval is unbounded and uninformative (Imbens & Manski, 2004). The breakdown values, rather than an identified interval, are informative here.

Which conclusions survive the restriction. If the item-effect assumption selects $\bar{\beta}_c$, tightening or relaxing that assumption should move the estimate. We therefore refit the audit model over a two-axis sensitivity grid, holding the MCMC seed fixed to reduce Monte Carlo noise (11 fits).

Axis 1 fixes σ_p over a grid and thereby varies the strength of the restriction. It ranges from the AMIS-R regime, where small σ_p sharply determines $\bar{\beta}$, toward the no-anchor regime of Proposition 1, where $\sigma_p \to \infty$. Axis 2 leaves σ_p free but widens and re-centers the μ_c prior. This control separates sensitivity to the identifying assumption from sensitivity to prior belief.

Let P_{dir} denote the posterior probability that μ_c has the sign of its posterior mean. The axes behave differently (Table 9), and the features separate into three cases:

Table 8: Breakdown analysis for shared feature slopes on 10-judge TREC web. c_{sign} and c_{95} are the item-residual confounding strengths that zero the slope and make its 95% interval touch zero. The NIST channel is $\beta_{gt} \times \text{coef}(g \sim x_c)$; RV expresses c_{sign} in multiples of that channel. Query length's channel is near zero, so its bracketed ratio is not interpretable. Posterior spread is the posterior standard deviation of μ_c.

Covariate	$\hat{\mu}_c$	Posterior spread	c_{sign}	c_{95}	NIST channel	RV
Lexical overlap	$+0.502$	0.086	0.502	0.334	0.269	**1.87**
Verbosity	$+0.189$	0.070	0.189	0.051	0.085	2.22
Formality	$+0.202$	0.076	0.202	0.052	0.138	1.46
Query length	-0.121	0.079	-0.121	0.000	-0.014	[8.56]
Position	-0.000	0.083	-0.000	0.000	0.337	0.00

- **Lexical overlap survives.** It remains positive with $P_{\text{dir}} \geq 0.96$ at every point on both axes.
- **Formality does not.** It falls from $+0.18$ at σ_p=0.8 to $+0.005$ with P_{dir}=0.51 at σ_p=16. The positive association in Table 1 is therefore as much a statement about the item prior as about the judges.
- **Position becomes increasingly negative.** Its population mean moves from -0.00 (P_{dir}=0.50) in the reference fit to -0.135 (P_{dir}=0.78) at σ_p=8 and -0.255 (P_{dir}=0.90) at σ_p=16.

The last position estimate comes from the only poorly sampled row ($\hat{R}$=1.19), but the monotone drift is already clear at σ_p=8. Thus the near-zero population mean in Section 3.1 reflects the restriction strength rather than an identified panel property. Query length decays similarly.

By contrast, widening the μ_c prior 25× on axis 2 moves overlap by only 0.10 and leaves every direction intact. Axis 1 moves overlap by 0.20 and drives position away from zero. The estimates are therefore more sensitive to the item-effect assumption than to their own priors. The judge contrasts remain stable, as Proposition 1 predicts. Across all 11 fits, the largest change in any δ_{cj} is 0.069, against an overlap contrast spread of 0.130. Kendall's τ with the reference run never falls below 0.91 for any feature. The sensitivity sweep therefore moves the shared component but leaves the judge-specific component removed by targeted adjustment stable.

Table 9: Prior sweep on the 10-judge panel. Entries report μ_c with its directional probability in brackets. The top block fixes σ_p; the reference free-σ_p fit has posterior 1.647 [1.53, 1.77]. The bottom block instead scales the re-centered μ_c prior standard deviation. "max $|\Delta\delta|$" is the largest judge-contrast change relative to the reference. All fits have zero divergences and $\hat{R} \leq 1.05$, except the over-dispersed σ_p=16 row ($\hat{R}$=1.19).

σ_p	Overlap	Verbosity	Formality	Query len	Position	max $\lvert\Delta\delta\rvert$	$\hat{R}$
free	$+0.502$ [1.00]	$+0.189$ [1.00]	$+0.202$ [0.99]	-0.121 [0.94]	-0.000 [0.50]	n/a	1.01
0.8	$+0.419$ [1.00]	$+0.147$ [1.00]	$+0.178$ [1.00]	-0.100 [0.99]	-0.002 [0.52]	0.065	1.01
1.65	$+0.504$ [1.00]	$+0.190$ [1.00]	$+0.209$ [1.00]	-0.114 [0.94]	$+0.002$ [0.52]	0.002	1.02
2.5	$+0.540$ [1.00]	$+0.219$ [0.99]	$+0.200$ [0.97]	-0.125 [0.87]	-0.002 [0.51]	0.029	1.01
4	$+0.545$ [1.00]	$+0.226$ [0.97]	$+0.162$ [0.88]	-0.103 [0.76]	-0.024 [0.57]	0.048	1.03
8	$+0.455$ [1.00]	$+0.234$ [0.95]	$+0.091$ [0.68]	-0.076 [0.67]	-0.135 [0.78]	0.062	1.05
16†	$+0.345$ [0.96]	$+0.227$ [0.94]	$+0.005$ [0.51]	-0.021 [0.56]	-0.255 [0.90]	0.069	1.19
σ_p free; μ_c prior standard deviation scaled and re-centered at zero							
×0.25	$+0.092$ [0.96]	$+0.031$ [0.83]	$+0.054$ [0.89]	-0.038 [0.81]	$+0.016$ [0.64]	0.034	1.01
×1	$+0.453$ [1.00]	$+0.154$ [0.98]	$+0.194$ [0.99]	-0.118 [0.94]	$+0.051$ [0.73]	0.002	1.02
×5	$+0.542$ [1.00]	$+0.192$ [0.99]	$+0.241$ [1.00]	-0.131 [0.95]	$+0.057$ [0.74]	0.002	1.01
×25	$+0.555$ [1.00]	$+0.190$ [0.99]	$+0.237$ [1.00]	-0.142 [0.96]	$+0.062$ [0.75]	0.002	1.02

The AMIS-R form, and varying the assumption. Writing the item effect as a feature regression plus a random residual makes the identifying assumption explicit. AMIS-R is a reparameterization, not a rival model. At matched parameter points, the two PyTensor graphs agree on log-density to 1.8×10^{-12}. Refitting AMIS-R reproduces every shared parameter of the reference posterior to within 1.79 pooled Monte Carlo standard errors (MCSEs) across 14 parameters. We therefore do not compare the two forms as models: they assign the same probability to the observed ratings along

the ridge, so any difference would be Monte Carlo noise. Model comparison cannot adjudicate the identifying restriction. The AMIS-R form lets us vary the residual assumption directly; all fits have $\hat{R} \leq 1.03$ with zero divergences. Replacing the normal item residual with Student-$t(4)$ turns the shared slope from an L^2 projection into a robust one. It moves every μ_c by at most 0.037 and leaves judge contrasts unchanged to three decimals, so a few extreme items do not drive the projection.

Dropping the human-label term is more consequential. It changes the restriction from $\mathbb{E}[e_i \mid x_i, g_i] = 0$ to $\mathbb{E}[e_i \mid x_i] = 0$. Lexical overlap rises from $+0.507$ to $+0.711$ (2.4 posterior standard deviations), and position rises from $+0.005$ to $+0.250$ (3.0 standard deviations); query length is unchanged. Conditioning on the observed quality channel therefore removes about 29% of the shared overlap slope and nearly all of the position slope. An observable proxy for quality moves $\bar{\beta}_c$ by two to three posterior standard deviations; ratings cannot condition on the remaining unobserved quality.

What an external anchor selects. We move the shared slope and score the resulting fused quality against NIST labels, using deployed fusion without a β_{gt} term. In this in-sample sensitivity calculation, NIST alignment is maximized when nearly none of the fitted shared overlap association is subtracted: the label-optimal value is $+0.002$, compared with the fitted $+0.502$. Full adjustment subtracts this association, whereas targeted adjustment retains $\bar{\beta}_c x_i$ inside the quality estimate (Table 2).

K ITEM- VERSUS JUDGE-INTERCEPT VARIATION

The estimated item-intercept standard deviation exceeds the judge-intercept standard deviation in every setting. On TREC web with ten judges, σ_p/σ_m is $1.65/0.49$ ($3.4\times$). On SummEval it is approximately $1.1\times$ (Appendix E).

These ratios compare only the two intercept standard deviations. They exclude judge-specific slopes, feature effects, and residual noise. The item term also combines true quality variation, ambiguity, and omitted item characteristics, so the ratio does not directly prescribe budget allocation. It shows only that more modeled cross-rating variation is associated with item intercepts than judge intercepts. Appendix F lists individual judge intercepts. The smaller ratio on summarization indicates substantial task dependence.

L ABLATION DETAIL

Table 10: Ablation ladder under the masked-rating reconstruction protocol of Table 7. "Relative MAE increase" is the increase relative to the full audit model, expressed as a percentage of the panel-mean MAE (0.414).

Remove	Reconstruction MAE	Relative MAE increase
Nothing (full audit)	0.304	n/a
Judge intercepts (α_j=0)	0.395	$+22.0\%$
All covariate slopes	0.327	$+5.6\%$
Overlap/formality slopes	0.319	$+3.7\%$
Verbosity/position slopes	0.309	$+1.4\%$

M FEW-SHOT ADAPTATION TO A NEW JUDGE

Unlike Table 7, this is a held-out new-judge test. For each held-out judge h, we refit the audit posterior after removing *all* of h's ratings (five fits, $\hat{R} \leq 1.01$, zero divergences). The resulting item qualities, other-judge parameters, cutpoints, and panel-average slopes are estimated without data from h.

We then estimate h's intercept from k calibration ratings. We sample k of its ratings, infer those items' latent quality using only the other judges, and choose $\hat{\alpha}_h$ to maximize the ordered-probit likelihood of the k observed ratings. Feature slopes remain fixed at the average over known judges because k ratings do not support new per-judge slopes. We then predict and score the judge's remaining items.

"Percentage of in-sample gain" is $100(0.414 - \mathrm{MAE}_k)/(0.414 - 0.304)$. It compares the panel mean (0.414, no judge knowledge, 0%) with the full audit model using h's own fitted parameters (0.304,

100%); the latter reference is not itself out of sample. Results are averaged over five held-out judges and 20 resamples for each value of k. Finite-resample variation means that the averages need not be monotone in k.

Table 11: Few-shot reconstruction for a held-out new judge. Lower MAE is better.

Ratings from the new judge	5	10	25	50	100
New-judge MAE	0.370	0.355	0.338	0.325	0.329
Percentage of in-sample gain	39.7	53.7	69.7	81.1	77.6

Query-split generalization. A complementary check asks whether responses generalize across queries rather than judges. We split the 97 queries into two halves, the $K=2$ case of the query-grouped split in Appendix P. We fit the audit model on each half and reconstruct ratings on the other half using response parameters that never saw those queries. Both fits converge cleanly ($\hat{R} \leq 1.01$, zero divergences).

Table 12: Masked-rating reconstruction across held-out query halves. Lower MAE is better.

	Panel mean	Cross-half (held-out queries)	Full-panel audit
Reconstruction MAE	0.414	0.318	0.304

The cross-half result retains 87.3% of the gap between the panel mean and the full-panel audit fit, $(0.414-0.318)/(0.414-0.304)$. This result suggests that the estimated judge responses are not specific to the queries used to fit them.

N IN-SAMPLE ADJUSTMENT CONTROLS AND SCORE MAPPINGS

The main text gives the cross-fitted TREC comparison, the label-free deployment comparison, and the SummEval result. This section reports the complete in-sample TREC diagnostic. The audit posterior is fitted using NIST labels from these same items, so these results are mechanism checks, not out-of-sample estimates.

Table 13: Complete in-sample TREC adjustment comparison. "Native MAE" uses each method's original score-to-label mapping. "Matched MAE" maps all methods to the NIST label marginal using the same monotone transform. The audit-posterior rows share one fitted posterior and differ only in which covariate components are removed.

Method / fusion mode	Spearman	Native MAE	Matched MAE
Panel mean	**0.688**	0.623	**0.569**
Z-score	0.686	0.625	0.581
Rasch / MFRM	0.683	0.633	n/a
Ordinal mixed-effects	0.683	0.631	n/a
Full adjustment	0.669	0.773	0.623
Shared-only removal	0.669	0.775	0.627
Targeted, overlap only	0.687	0.722	0.590
Targeted, all judge-specific deviations	0.686	0.722	0.577
Deployment full (no labels)	0.619	0.746	n/a
Deployment targeted (no labels)	0.688	0.664	n/a

Within the audit posterior, the Spearman point estimate rises from 0.669 under full adjustment to 0.686 under targeted adjustment. The paired item-bootstrap difference is $+0.017$ with 95% CI $[-0.004, 0.040]$; 94% of bootstrap replicates are positive. The targeted-minus-panel-mean difference is -0.002 with CI $[-0.008, 0.004]$. Removing only the shared component yields 0.669, whereas adjusting only the overlap deviations yields 0.687. These controls indicate that retaining the shared overlap component accounts for most of the in-sample full-versus-targeted difference.

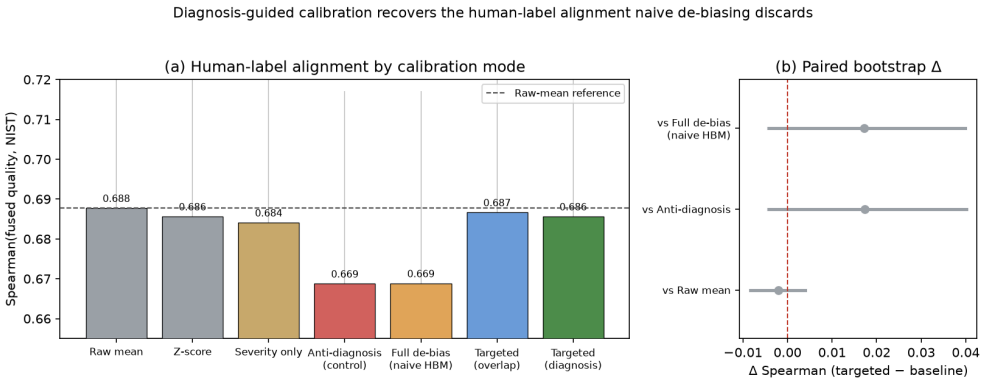


Figure 2: In-sample human-label alignment under alternative adjustment rules. **(a)** Rank correlation with NIST labels by fusion mode. **(b)** Paired-bootstrap differences between modes. Numerical details appear in Appendix N; Table 2 reports the cross-fitted comparison.

Targeted adjustment also has the higher correlation when human labels are excluded from model fitting: the point estimate rises from 0.619 to 0.688, while the corresponding masked-rating reconstruction MAE changes from 0.292 to 0.295. This comparison remains in-sample with respect to the NIST evaluation labels, but it shows that the full-versus-targeted ordering is not created solely by including NIST labels in the audit posterior.

Under the masked-rating reconstruction metric of Appendix H, full audit adjustment has MAE 0.304 and accuracy 69.9%, while targeted adjustment has MAE 0.306 and accuracy 69.6%. The overlap-only and shared-only variants reach $0.313/68.9\%$ and $0.318/68.6\%$, respectively. Deployment targeted reaches $0.295/70.7\%$ versus $0.292/71.1\%$ for deployment full. Thus retaining the shared component changes reconstruction performance little while changing human-label rank alignment. The reconstruction difference is small because moving $\bar{\beta}^\top x_i$ between the fused quality $\hat{q}_i$ and the feature term shifts both by the same item-level amount. The reconstructed predictor $\hat{q}_i + \hat{\alpha}_j + \hat{\delta}_j^\top x_i$ therefore changes little. The quality estimate $\hat{q}_i$ does change, and that estimate is ranked against human labels.

The z-score results further show that human-label alignment and masked-rating reconstruction measure different objectives. Z-score correlation with human labels is 0.686, close to the panel mean's 0.688. When prediction excludes the target judge's own marginal statistics, however, z-score reconstruction MAE is 0.434, compared with 0.414 for the panel mean. We therefore report the two metrics separately.

Absolute-error mappings. Native MAE is not directly comparable across all methods because the panel mean already lies on the 0–3 label scale, whereas HBM scores are latent quantities mapped through fitted cutpoints. Under the native mappings, targeted adjustment has MAE 0.722 versus 0.623 for the panel mean. Mapping every method to the NIST marginal using the same monotone transformation reduces this difference to 0.008 (0.577 versus 0.569). A separately fitted best affine map gives a difference of -0.006. These are in-sample score-mapping diagnostics; the main-text conclusions therefore rely on the scale-free Spearman comparisons. Figure 2 plots the alignment-by-fusion-mode and paired-bootstrap views of the same in-sample comparison.

O SummEval adjustment detail

The SummEval result in Table 2 uses the five-judge SummEval audit posterior and evaluates the fused scores against the three-expert reference scores. This is a cross-domain replication but, because the expert scores enter the audit posterior, it remains an in-sample diagnostic with respect to those labels.

Displayed position is not a judge-visible feature in this panel, so the SummEval fit excludes it. Panel-mean correlation is 0.597. Full adjustment reaches 0.553, whereas targeted adjustment reaches 0.578. The article-clustered paired bootstrap gives a targeted-minus-full difference of $+0.023$ with 95% CI

$[-0.037, 0.081]$. Targeted adjustment remains below the panel mean by 0.020 (CI $[-0.042, -0.001]$). Thus this in-sample check is consistent with retaining more label-relevant information than full adjustment, but does not resolve their difference.

P CROSS-FITTED EVALUATION DETAIL

This section supplies the complete procedure and uncertainty analysis for the cross-fitted TREC row in Table 2. We partition the 97 queries into five folds, compute covariate standardization and fit the audit posterior using training queries only, and score each held-out item without its human label. We repeat the five-fold partition ten times, yielding 50 fitted fold models. All satisfy $\hat{R} \leq 1.02$, with three divergent transitions in total.

The implementation separates query-grouped splitting and train-only standardization from posterior fusion and uncertainty aggregation. Each fold uses 1,500 tuning and 1,500 retained sampling draws. We test the fusion calculation against the in-sample calculation rather than assuming equivalence. Disabling the query split reproduces each in-sample Spearman value to within 10^{-6}. Changes in the cross-fitted results therefore come from the split, not a different fusion calculation.

Every fusion mode adjusts for the judge intercept $\hat{\alpha}_j$. "Intercept" adjusts for nothing else. "Full" also subtracts the shared $\bar{\beta}_c x_i$, the judge-specific $\delta_{cj} x_i$, and the shared position–verbosity interaction. "Shared-only" removes the shared terms but retains $\delta_{cj} x_i$; it is the reverse of targeted adjustment and corresponds to "Shared-only removal" in Table 13. "Targ. (ov.)" adjusts for only the judge intercept and overlap deviation. "Targeted" adjusts for the judge intercept and every $\delta_{cj} x_i$, while retaining all shared terms.

Table 14: Out-of-fold Spearman correlation with NIST labels across all ten repeats; higher is better. The panel mean has no fitted parameters, so its out-of-fold and in-sample values agree to four decimals in every repeat. The final row reports the standard deviation across repeats.

	Panel mean	Z-score	Intercept	Shared-only	Full	Targ. (ov.)	Targeted
In-sample	0.6877	0.6856	0.6841	0.6687	0.6688	0.6866	0.6856
Repeat 0	0.6877	0.6816	0.6816	0.6597	0.6601	0.6829	0.6826
Repeat 1	0.6877	0.6800	0.6792	0.6619	0.6620	0.6797	0.6797
Repeat 2	0.6877	0.6797	0.6805	0.6616	0.6612	0.6808	0.6805
Repeat 3	0.6877	0.6774	0.6789	0.6546	0.6548	0.6785	0.6789
Repeat 4	0.6877	0.6814	0.6814	0.6638	0.6643	0.6815	0.6820
Repeat 5	0.6877	0.6832	0.6842	0.6632	0.6634	0.6839	0.6834
Repeat 6	0.6877	0.6780	0.6763	0.6603	0.6599	0.6766	0.6765
Repeat 7	0.6877	0.6819	0.6812	0.6584	0.6589	0.6821	0.6819
Repeat 8	0.6877	0.6817	0.6803	0.6605	0.6608	0.6812	0.6810
Repeat 9	0.6877	0.6804	0.6811	0.6641	0.6644	0.6823	0.6827
Mean	0.6877	0.6805	0.6805	0.6608	0.6610	0.6809	0.6809
Spread	0.0000	0.0018	0.0021	0.0028	0.0029	0.0022	0.0021

Table 15: Corrected resampled t-test over all 50 fold-level differences (Nadeau & Bengio, 2003). The naive interval ignores training-set overlap; the correction inflates the standard error by $3.67\times$. Mean Δ averages fold-level differences and need not equal a difference between pooled correlations. The last column counts per-repeat query-clustered bootstraps whose intervals exclude zero (2,000 resamples of 97 queries).

Comparison	Mean Δ	Naive 95% CI	Corrected 95% CI	p	Boot. excl. 0
Targeted − Full	$+0.0210$	$[+0.0146, +0.0274]$	$[-0.0024, +0.0444]$	0.078	4/10
Targeted − Shared-only	$+0.0210$	$[+0.0147, +0.0273]$	$[-0.0022, +0.0442]$	0.075	4/10
Targeted − Panel mean	-0.0037	$[-0.0055, -0.0019]$	$[-0.0103, +0.0029]$	0.265	6/10
Targeted − Z-score	-0.0000	$[-0.0020, +0.0020]$	$[-0.0073, +0.0073]$	0.997	0/10

Table 14 lists all ten repeats, and Table 15 reports the corrected test. The corrected interval is wider because folds share training data. Under this correction, targeted-minus-full is $+0.0210$

with 95% CI $[-0.0024, 0.0444]$ ($p = 0.078$), while targeted-minus-panel-mean is -0.0037 with CI $[-0.0103, 0.0029]$. Targeted adjustment has the higher point estimate than full adjustment in every repetition, but the available TREC data do not resolve either comparison at the 5% level.

Why additional repeats do not resolve the comparison. Repeated partitions reuse the same 97 queries, so increasing R stabilizes the variance estimate but does not add independent observations. At R=2, the variance is estimated from ten folds and the corrected test returns p=0.026; with 50 folds, it returns p=0.078. We report the ten-repeat result and do not interpret the R=2 value as independent evidence.

What the design can resolve. The targeted-versus-mean gap of -0.0037 is below the design's resolution. With 97 query clusters, the query-clustered bootstrap's 95% half-width on a paired difference is 0.0064. At the observed uncertainty, this analysis cannot distinguish a difference of 0.0037 from zero. The in-sample query-clustered bootstrap has an NIST-label intraclass correlation of 0.049, reflecting the stratified-by-relevance sampling design of Voorhees (2001).

Q Is the slope heterogeneity a response-scale artifact?

Analytic pre-check. The main model fixes a shared probit scale and shared cutpoints. Suppose instead that judges differ only in *discrimination*, reading the latent axis more or less steeply. Under this pure-scale null, $\beta_{cj} = a_j \mu_c$, $\bar{\beta}_c = \bar{a}\mu_c$, and $\delta_{cj} = (a_j - \bar{a})\mu_c$. Consequently, $\mathrm{sd}_j(\delta_c)/|\bar{\beta}_c| = \mathrm{sd}_j(a)/\bar{a}$ should be constant across features. Here, sd_j denotes the standard deviation across judges. Two patterns in the 10-judge posterior depart from this null:

(i) The implied relative discrimination spread is 0.17 for verbosity, 0.17 for formality, 0.25 for overlap, 0.45 for query length, and 1.91 for position—an 11× range.

(ii) Let $\Delta \in \mathbb{R}^{C \times J}$ be the feature-by-judge matrix whose cth row contains the judge deviations for feature c. Under the null, Δ should be rank one and its leading left singular vector u_1 should be parallel to μ. The leading component carries 63.3% of the variance, and $|\cos(u_1, \mu)| = 0.84$ rather than ≈1.

This is a motivating diagnostic, not a definitive test. Each feature has its own σ_c hyperprior, so feature-specific shrinkage contributes to the non-constant ratio in (i).

Discrimination-augmented model. We fit a robustness model on the 10-judge panel that nests the pure-scale null:

$$z_i = \theta_i + \beta_{gt}\, g_i + \textstyle\sum_c \mu_c x_{ic} + \beta_{pv}(\mathrm{pos{\cdot}verb})_i, \qquad \eta_{ij} = \alpha_j + a_j z_i + \textstyle\sum_c d_{cj} x_{ic}. \tag{6}$$

We constrain $a_j > 0$ and $\sum_j \log a_j$=0, closing an otherwise exact likelihood ridge with σ_p. Writing $\ell_j = \log a_j$, the model uses a zero-sum Normal prior with scale τ_a, where $\tau_a \sim \mathrm{HalfNormal}(0.5)$. The residual feature responses d_{cj} are zero-sum and use the same σ_{d_c} hyperpriors as the main model's σ_c. Centering θ_i, g_i, and the interaction keeps α_j identified as judge j's intercept at the average item.

Setting $a_j \equiv 1$ recovers the main model (M0). Setting $d_{cj} \equiv 0$ gives the pure-scale null (M1). Letting both vary (M2) decomposes the two sources. All three models converge cleanly ($\hat{R} \leq 1.03$, zero divergences).

Under M2, the total feature slope and centered deviation are $\beta_{cj} = a_j \mu_c + d_{cj}$ and $\delta_{cj} = (a_j - \bar{a})\mu_c + d_{cj}$. Because the discrimination and residual terms can trade off, we report the per-draw identity

$$\begin{aligned}\mathrm{Var}_j(\delta_{cj}) = \mathrm{Var}_j[(a_j - \bar{a})\mu_c] + \mathrm{Var}_j(d_{cj}) \\ + 2\,\mathrm{Cov}_j((a_j - \bar{a})\mu_c, d_{cj})\end{aligned}$$

as three variance shares, together with the underlying correlation. A large negative correlation indicates that the two terms are not separately identified for that feature.

In Table 16, "Sep." is a heuristic separability indicator: it is "no" when the scale/residual correlation exceeds 0.3 in magnitude. The three variance shares sum to 100% up to rounding; the covariance share is the explicit cross term above divided by $\mathrm{Var}_j(\delta_c)$. The covariance share and correlation are averaged under different per-draw normalizations and can therefore differ in sign. The reported $\mathrm{sd}_j(\delta)$ is the finite-panel posterior spread, not the hyperparameter σ_c in Table 1. Finally, τ is Kendall's rank correlation between the M0 and M2 judge deviations.

Table 16: Response-scale decomposition under M2 on the 10-judge panel. The three variance shares distinguish scale, residual feature response, and their covariance. "Sep." reports the heuristic separability indicator defined in the text.

Covariate	$\mathrm{sd}_j(\delta)$ M0	$\mathrm{sd}_j(d)$ M2	%scale	%resid.	%cov	corr	Sep.	τ
Position	0.100	0.070	17%	**85%**	-2%	$+0.03$	yes	0.64
Query length	0.057	0.080	27%	**71%**	$+2\%$	$+0.07$	yes	0.87
Verbosity	0.034	0.039	78%	35%	-13%	-0.07	yes	0.64
Formality	0.035	0.030	88%	25%	-14%	-0.08	yes	0.38
Overlap	0.131	0.115	155%	73%	$\mathbf{-128\%}$	$\mathbf{-0.58}$	**no**	0.16

Position and query length retain most of their heterogeneity as differential response after accounting for discrimination. The posterior-mean M0 deviation and M2 residual have the same sign for seven of ten judges on position and all ten judges on query length. Their rank agreement is moderate to high (τ=0.64–0.87). Verbosity and formality are effectively shared in the main taxonomy, so their scale-dominated decomposition is not consequential for targeted adjustment.

Lexical overlap remains unresolved. Its scale and residual components have posterior correlation -0.58, so M2 cannot separate them. Its judge-ranking agreement with M0 is also the weakest of the five (τ=0.16). Because overlap accounts for most of the full-versus-targeted difference (Section 6.4), we treat this as an open question. M0's between-judge spread for overlap remains sizeable under every specification tested, but some of it may reflect response scale rather than differential lexical sensitivity.

Discrimination-augmented targeted adjustment. We also run targeted fusion with the M2 posterior. This replaces the main model's centered slopes with M2's residual d_{cj} terms while keeping shared responses inside the quality score. Against NIST labels on the 10-judge panel, targeted adjustment has a higher rank-correlation point estimate than full adjustment (ρ=0.690 versus 0.671) but does not outperform the panel mean (ρ=0.691).

The query-clustered paired bootstrap gives targeted $-$ full $= +0.019$ with 95% CI $[-0.001, +0.039]$, and targeted $-$ panel mean $= -0.001$ with CI $[-0.005, +0.003]$. The point estimate favors retaining shared components, but the targeted-minus-full interval includes zero. Allowing judge discrimination does not make targeted adjustment a better scorer than the panel mean.

Sensitivity and convergent-model check. The posterior mean of τ_a under M2 is 0.347. We refit M2 with the more permissive fixed value $\tau_a = 1.0$. This changes each posterior-mean a_j by at most 0.0093 and any variance share by at most 4.32 percentage points; all five heuristic separability classifications remain unchanged.

Independently, a feature-free graded-response model estimates per-judge discrimination on the five-judge calibration panel (Appendix G). For the five overlapping judges, these discrimination estimates correlate with M2's discrimination estimates (ρ=0.70, Kendall τ=0.60). This agreement indicates that the estimated discrimination ordering is not unique to the feature-augmented specification. The analysis is standalone; it does not change the main pipeline, targeted rule, or amortized network.

R AMORTIZED NETWORK: ARCHITECTURE AND TRAINING DETAIL

Architecture (250,000 parameters). A shared item-encoder MLP embeds each item's one-hot ratings ($J\times(K+1)$) and six features: the five item features plus audit label g_i, which is set to zero at deployment. A masked mean pools the item embeddings into one global code $\mathbf{h}$.

Three decoders map $\mathbf{h}$ to Gaussian approximate posteriors. A global decoder predicts population parameters, a judge decoder predicts $\{\alpha_j, \beta_j\}$, and an item decoder predicts item qualities $\{\theta_i\}$. The judge decoder also receives a learnable identity embedding. Without it, judges in the same capability group collapse to identical estimates. Per-judge encoder branches instead overfit the simulator and hurt transfer, so the final network uses one shared encoder.

Training. We simulate N datasets from the prior and minimize a Gaussian negative-log-likelihood. The global, judge, and item losses receive weights $10:5:1$, preventing the many item terms from

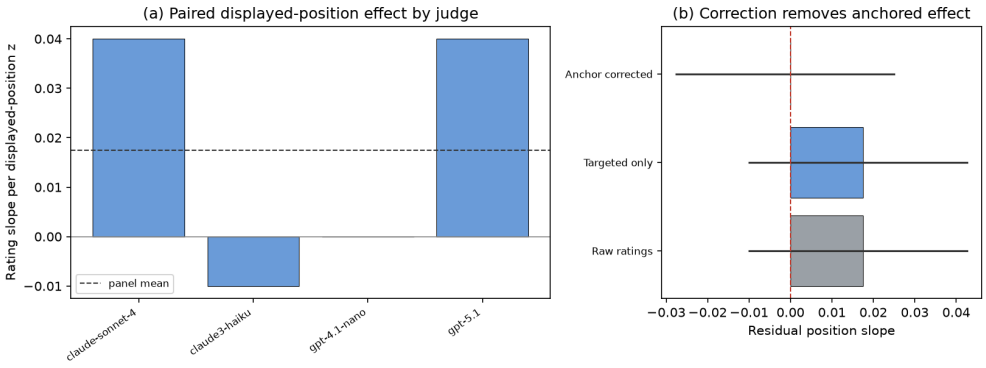


Figure 3: Paired position variation at fixed content; only collection order is randomized. **(a)** Judge-level slopes in the four-judge controlled panel. **(b)** Residual slopes after no correction, judge-specific correction, and anchored common-plus-specific correction (Section 6.2).

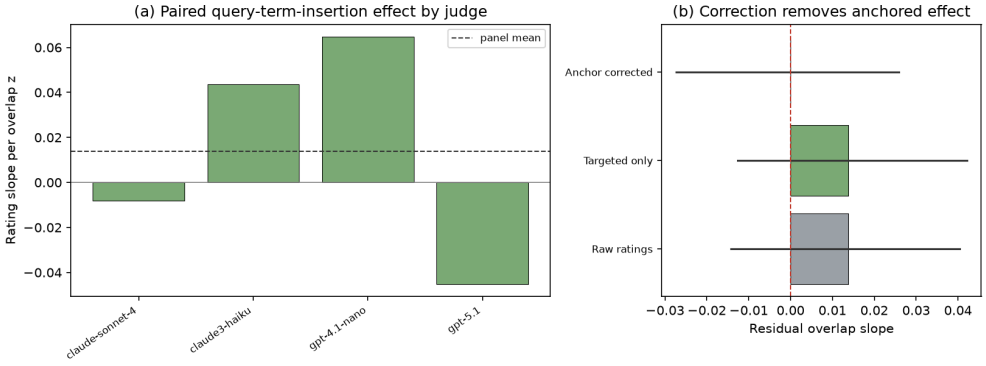


Figure 4: Marked query-term insertion at fixed item and judge. **(a)** Judge-level paired slopes after insertion. **(b)** Residual slopes after no correction, judge-specific correction, and anchored common-plus-specific correction (Section 6.3).

dominating. Optimization uses AdamW with a cosine learning-rate schedule on an Apple Metal Performance Shaders (MPS) device. The primary network uses 50,000 simulations and trains in about 34 minutes; a learning-curve comparison shows that accuracy is near its plateau by $N{=}25{,}000$. We evaluate posterior calibration, agreement with MCMC posterior means, masked-rating reconstruction, and robustness rather than reducing performance to one number (Section 6.5).

S SUPPLEMENTARY FIGURES AND TABLES

Amortized robustness detail. The $3\times$ spread intervention widens observed judge intercepts after simulation rather than resampling σ_m. It therefore probes extreme panels, not full prior misspecification. Under missing-at-random ratings, a missing rating is represented by an all-zero vector in place of its one-hot rating encoding; the other item inputs remain unchanged.

The primary network is trained only on complete inputs. Across 300 paired fresh simulations, its recovery is $\rho{=}0.54$ with complete data, 0.48 with 20% dropped, and 0.32 with 60% dropped. A second network is trained with a per-dataset drop fraction drawn from $\mathcal{U}(0, 0.6)$ and uses the same test-time encoding. It retains complete-data accuracy ($\Delta\rho{=}-0.004$) and calibration (mean SBC KS 0.038; 90% coverage 91.0%). Recovery rises to $\rho{=}0.50$ at 20% dropped ($\Delta{=}+0.023$, 95% CI $[+0.020, +0.026]$) and $\rho{=}0.41$ at 60% dropped ($\Delta{=}+0.087$, CI $[+0.081, +0.093]$).

Table 17: Method comparison as implemented in this work. Masked-rating mean absolute error (MAE) is computed on the same five-judge calibration panel. The audit HBM values are from the MCMC and amortized fits, respectively; the deployment HBM is fit without human labels. Other methods may admit extensions beyond the implementations compared here.

Method	Masked MAE ↓	Fitted structure	Shared/specific split shown	Few-shot evaluated	Amortized here
Panel mean	0.414	aggregation only	no	no	no
Z-score	0.434	judge location/scale	no	no	no
Dawid–Skene	0.420	judge confusion matrices	no	no	no
MFRM	0.306	item quality + judge severity	no	no	no
Graded-response IRT	0.270	item quality + severity/discrimination	no	no	no
Ordinal mixed-effects	0.319	item + judge random intercepts	no	no	no
HBM audit (ours)	0.304 / 0.302	item quality + judge-varying slopes	yes	yes	yes (1.3 ms)
HBM deployment (no labels)	0.292	item quality + judge-varying slopes	yes	no	no

Table 18: Central-interval coverage of the amortized posterior (2,000 simulated datasets).

Nominal	50%	60%	70%	80%	90%	95%
Empirical	52.1%	62.2%	72.0%	81.5%	90.9%	95.3%

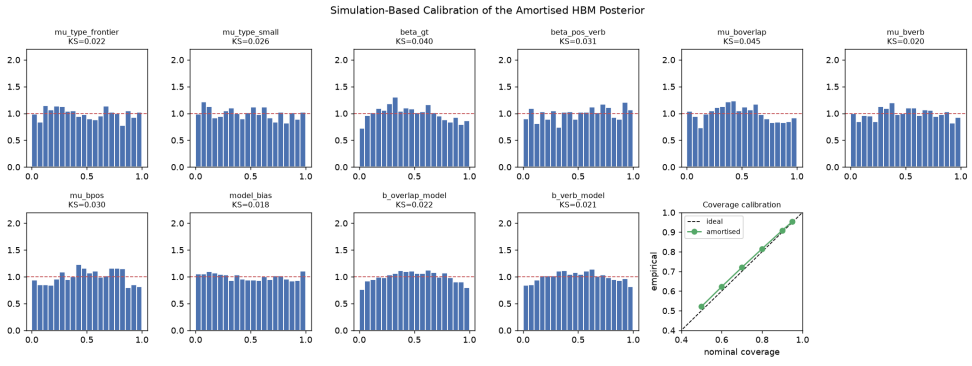


Figure 5: Simulation-based calibration under the training simulator. The reported marginal rank histograms are near-uniform (dashed red = ideal), and marginal interval coverage is near nominal.