IDENTIFIABILITY-AWARE UNCERTAINTY AUDITS FOR SPARSE LEADERBOARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern evaluation leaderboards often combine noisy results from partially overlapping views, yet uncertainty is usually assessed only after the evaluation table is fixed. We show that the cross-view covariances may not determine a unique shared scale for confidence intervals and ranking probabilities. We represent view overlap as a graph and prove that each bipartite connected component leaves one scale ambiguity, while a component with an odd cycle does not. We also derive an information bound under a Gaussian model for the estimated log cross-view moments, showing how weak graph conditioning limits the precision with which the scale can be recovered. The same graph structure also provides model-consistency checks and guides which new overlaps would be most informative to add. Building on our analysis, we propose anchored cross-view empirical Bayes (ACEB), a target-label-free uncertainty audit that combines graph diagnostics, fixed-anchor latent-score estimation, leave-view residual-noise estimation, and selective decisions that abstain when confidence is low. In sparse heterogeneous simulations, ACEB improves uncertainty diagnostics over several target-free baselines. After same-coverage recalibration, ACEB variants achieve the lowest pairwise Brier and interval scores among the six compared methods. On 13,077 MedHELM pairwise-gap records, coverage recalibration gives ACEB 22% fewer false signed declarations per record than random effects (0.147 *v.s.* 0.189). ACEB makes signed decisions on 78.6% rather than 98.3% of records, indicating that most of reduction comes from greater abstention[1]. Together, these results show that accounting for overlap structure can inform both leaderboard uncertainty quantification and selective decision making.

1 INTRODUCTION

Modern evaluation leaderboards are used not only to rank systems, but also to judge whether one system is reliably better than another. Such judgments can change with data sampling, random initialization, or hyperparameter choices (Reimers & Gurevych, 2017; Bouthillier et al., 2021). Comparisons are further complicated by incomplete evaluation coverage: the original HELM study noted that some prominent models shared no evaluation scenario before its standardized assessment (Liang et al., 2023). We use a *view* to mean one source or slice of evaluation, such as a prompt block, judge, scenario, seed, or metric. A system may be evaluated on some views but not others, and the views can differ in sensitivity and noise.

Existing work improves reliability through score distributions, resampling-based intervals, and hierarchical benchmark aggregation (Reimers & Gurevych, 2017; Agarwal et al., 2021; Longjohn et al., 2025). Empirical Bayes stabilizes noisy estimates (Stephens, 2017), while post-hoc calibration adjusts predictive uncertainty using observed outcomes (Kuleshov et al., 2018). However, uncertainty estimation and calibration do not by themselves establish what the overlap design identifies. With unknown view sensitivities, the same cross-view moments can imply different latent scales and hence different confidence in system comparisons. This ambiguity remains even when the moments are known exactly, so more precise estimates alone cannot resolve it. We therefore connect uncertainty reporting to measurement design: which overlaps identify the scale, how stably it can be recovered, and which additional evaluations provide the missing information.

[1]Code is available at: https://anonymous.4open.science/r/aceb-0DEA/

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.

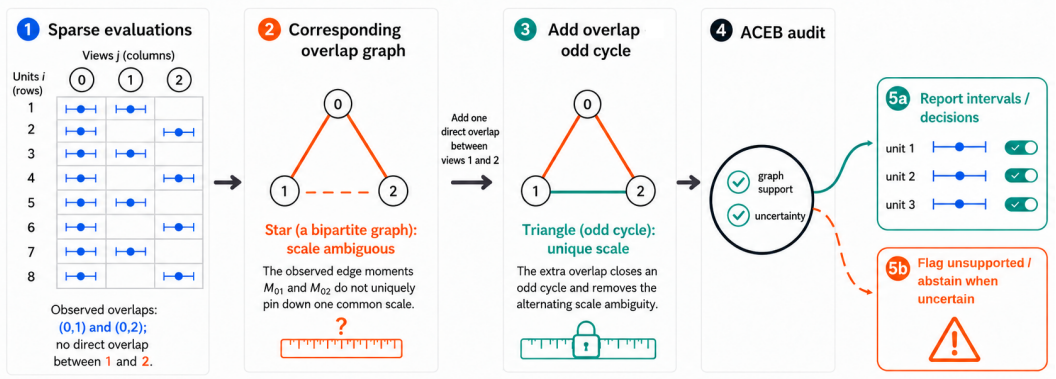


Figure 1: From sparse evaluations to an identified scale. (1) Rows are units i and columns are views j (e.g., prompt blocks, judges, scenarios, seeds, or metrics). (2) Overlaps $(0,1)$ and $(0,2)$ form a bipartite star with an ambiguous scale. (3) Adding overlap $(1,2)$ closes a triangle and removes this ambiguity under the working model. (4) ACEB audits support and uncertainty, with (5a) fixed-anchor intervals and decisions and (5b) unsupported cases or abstentions when uncertain.

A three-view example in Figure 1 illustrates the issue. For this illustration, we use y_{ij} to denote the observed summary for unit i under view j after subtracting that view's mean across units. Suppose $y_{ij} = a_j\theta_i + \varepsilon_{ij}$, where θ_i is a scalar latent score, $a_j > 0$ is the sensitivity of view j, and ε_{ij} is residual noise. We assume the residuals are uncorrelated across views and with θ_i. Let $v = \mathrm{Var}(\theta_i) > 0$. Then, for $j \neq k$, the cross-view covariance is $M_{jk} := \mathrm{Cov}(y_{ij}, y_{ik}) = v a_j a_k$. Write $u_j = \sqrt{v}\, a_j$. Suppose view 0 overlaps with views 1 and 2, but views 1 and 2 never overlap. The observed moments reveal only u_0u_1 and u_0u_2, which are unchanged under

$$(u_0, u_1, u_2) \mapsto (cu_0, c^{-1}u_1, c^{-1}u_2), \qquad c > 0.$$

Thus the cross-view moments alone do not identify the scale decomposition. Adding overlap $(1,2)$ closes a triangle and removes this ambiguity, whereas more units on the original overlaps do not.

This example is naturally a graph problem: views are vertices, and an edge connects two views that share evaluations. A component is *bipartite* when its vertices split into two groups with every edge crossing between them. Stars, chains, trees, and even cycles are examples. In such a component, multiplying node scales in one group by c and those in the other by c^{-1} leaves all edge products unchanged, so the scale remains ambiguous. An odd cycle breaks this rescaling and removes the ambiguity. Taking logs turns these edge products into a signless-incidence system, whose nullspace formalizes the same ambiguity.

This graph perspective is closely related to sparse rank-one completion. Ma et al. (2020) connect worker-skill identification to non-bipartite overlap and derive estimation bounds involving the signless Laplacian. We build on this connection for continuous leaderboard summaries with reported standard errors and additional view noise. The same structure also lets us assess conditioning, check moment compatibility, and guide overlap repair. These diagnostics complement a partial order, which records asserted or withheld comparisons but does not itself establish measurement support. Appendix A discusses additional related work.

In this paper, we propose *anchored cross-view empirical Bayes* (ACEB), a target-label-free uncertainty audit. ACEB reports graph diagnostics, fits a fixed-anchor latent-score model, estimates residual view noise from leave-view predictions, and abstains when posterior confidence is low. Structural warnings are reported alongside posterior summaries and do not gate the decisions. Figure 1 summarizes the procedure.

Contributions.

1. **Measurement-design audit.** Building on the non-bipartite identification condition for sparse rank-one recovery (Ma et al., 2020), we study the corresponding structure for continuous cross-view moments and characterize its componentwise identified set. We further connect this structure to conditioning, model-consistency checks, and overlap repair.

2. **Operational uncertainty audit.** ACEB combines graph diagnostics with fixed-anchor empirical-Bayes estimation, leave-view residual-noise estimation, and selective decisions. Its base fitting procedure does not require target labels.
3. **Mechanism-focused evaluation.** Synthetic graph interventions, component ablations, and negative controls examine the roles of overlap structure and residual-noise estimation. MedHELM and SummEval show the tradeoffs among uncertainty calibration, interval score, and abstention.

Scope. Leaderboards are one setting for estimating latent quantities from noisy, partially overlapping sources. Our graph results concern off-diagonal moments under a scalar latent score, positive loadings shared across units, uncorrelated cross-view residuals, and common latent-score variance across overlap subsets. ACEB adds plug-in posterior inference under the working model using reported standard errors. Section 3 and Appendix C.1 distinguish moment identification from information in the full observation model.

2 Model and ACEB audit

We consider n evaluation units and J views. Let $\mathcal{I} \subseteq \{1, \dots, n\} \times \{0, \dots, J-1\}$ be the observed unit-view pairs, and let $\mathcal{J}_i := \{j : (i,j) \in \mathcal{I}\}$ be the views observed for unit i. For each $(i,j) \in \mathcal{I}$, the benchmark reports a summary y_{ij} and standard error $s_{ij} > 0$. We assume each unit i has an unobserved scalar score $\theta_i \in \mathbb{R}$, oriented so that larger values indicate better performance. Our aim is to estimate these latent scores and quantify uncertainty in the resulting leaderboard comparisons.

Probability model. We model each observed summary y_{ij} as

$$y_{ij} \mid \theta_i, a_j, b_j, \tau_j \sim \mathcal{N}\big(b_j + a_j\theta_i,\, s_{ij}^2 + \tau_j^2\big), \qquad (i,j) \in \mathcal{I}, \tag{1}$$

where $a_j > 0$ is the sensitivity of view j and b_j is its additive bias. Here, s_{ij} captures sampling uncertainty in y_{ij}, and τ_j denotes additional view noise. We write $\sigma_{ij}^2 := s_{ij}^2 + \tau_j^2$ as the total conditional observation variance. The observed summaries y_{ij} are conditionally independent given the latent scores θ_i and view parameters a_j, b_j, τ_j. Sharing a_j across units allows overlaps to inform the relative view scales. We refer to Section 3 for further analysis on the cross-view moments.

Prior. Let $\mathbf{x}_i$ contain observed covariates for unit i, with a constant 1 included for the intercept. For each latent score θ_i, we use an empirical-Bayes prior (Efron, 2010):

$$\theta_i \mid \mathbf{x}_i \sim \mathcal{N}(\mu_i, v_i), \qquad \mu_i = f_\mu(\mathbf{x}_i), \quad v_i = f_v(\mathbf{x}_i) > 0. \tag{2}$$

We model the latent scores as independent across units, with prior mean μ_i and variance v_i for unit i. Without additional covariates, $\mathbf{x}_i = 1$, and ACEB estimates a common prior mean and variance across units. Equation (1) allows view-specific biases for generality.

All experiments fix $b_j = 0$, so $\theta_i = 0$ implies zero expected score in every view. A pre-specified reference view j_0 fixes the measurement unit through $a_{j_0} = 1$. This anchoring convention alone does not remove a scale ambiguity left by the observed overlap graph.

Posterior. ACEB fits the prior parameters and non-reference loadings by staged penalized marginal-likelihood optimization, integrating out the latent scores. It then estimates residual variances using the leave-view moments below. We refer to Appendix B for fitting objectives and numerical settings.

For fixed parameter values, the posterior is Gaussian with variance V_i and mean m_i:

$$V_i = \left(v_i^{-1} + \sum_{j \in \mathcal{J}_i} \frac{a_j^2}{\sigma_{ij}^2}\right)^{-1}, \qquad m_i = V_i \left(\frac{\mu_i}{v_i} + \sum_{j \in \mathcal{J}_i} \frac{a_j(y_{ij} - b_j)}{\sigma_{ij}^2}\right). \tag{3}$$

Here, the posterior mean m_i averages the prior mean μ_i and adjusted summaries $(y_{ij} - b_j)/a_j$ with inverse-variance weights. Let $\mathcal{D}$ denote the observed summaries $\{y_{ij}\}$, reported standard errors $\{s_{ij}\}$, and covariates $\{\mathbf{x}_i\}$.

At fitted parameter values, ACEB reports the plug-in approximation $\theta_i \mid \mathcal{D} \approx \mathcal{N}(m_i, V_i)$. Under this plug-in model, independence across units gives

$$p_{i,i'} := \mathbb{P}(\theta_i > \theta_{i'} \mid \mathcal{D}) \approx \Phi\left(\frac{m_i - m_{i'}}{\sqrt{V_i + V_{i'}}}\right), \quad i \neq i' \tag{4}$$

where Φ is the standard normal cumulative distribution function. When each unit is an individual system, $p_{i,i'}$ is the posterior probability that system i outperforms i'. Alternatively, when a unit represents a model-A-minus-model-B score gap, the probability that A outperforms B is $\mathbb{P}(\theta_i > 0 \mid \mathcal{D}) \approx \Phi(m_i/\sqrt{V_i})$.

Leave-view residual-noise estimation. A residual based on m_i reuses y_{ij} in the prediction, which can make the residual artificially small. We use $\mathcal{D}_{i,-j}$ to denote the observations for unit i excluding view j, and $(m_{i,-j}, V_{i,-j})$ represent the corresponding posterior moments. Conditional on fixed parameters,

$$y_{ij} \mid \mathcal{D}_{i,-j} \sim \mathcal{N}\big(b_j + a_j m_{i,-j},\, a_j^2 V_{i,-j} + s_{ij}^2 + \tau_j^2\big)\,. \tag{5}$$

Let $\mathcal{I}_j := \{i : (i,j) \in \mathcal{I}\}, n_j := |\mathcal{I}_j|$. ACEB estimates the additional residual variance by

$$\widehat{\tau}_j^2 = \left[\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\Big\{(y_{ij} - \widehat{b}_j - \widehat{a}_j\widehat{m}_{i,-j})^2 - \widehat{a}_j^2\widehat{V}_{i,-j} - s_{ij}^2\Big\}\right]_+ . \tag{6}$$

In the reported experiments, $\widehat{b}_j = 0$. The leave-view construction avoids direct unit-level reuse of y_{ij} in its predictive residual. This is a moment estimator, not a finite-sample coverage guarantee. The global loadings and prior parameters are treated as plug-in estimates, so their uncertainty is not propagated. Appendix B gives the full implementation and a consistency result under explicit convergence conditions.

Audit and selective decisions. ACEB reports graph rank, conditioning, and moment-positivity diagnostics alongside fixed-anchor intervals and directional probabilities. These diagnostics do not gate the fitted decisions. In Section 3, we provide additional cycle-consistency checks. The reported experiments use one pre-specified reference view. If a component is structurally deficient, ACEB instead reports a warning for scale-sensitive interpretation while still computing fixed-anchor summaries. For an oriented pairwise gap, let $\pi_i = \Phi(m_i/\sqrt{V_i})$. At confidence threshold κ, ACEB declares a positive sign when $\pi_i \geq \kappa$, a negative sign when $\pi_i \leq 1-\kappa$, and otherwise abstains. Anchor sweeps are separate sensitivity analyses and do not alter the reported decisions. We provide the complete staged procedure in Appendix B.3.

3 Measurement-design identifiability

3.1 Identified scale under the overlap model

We first ask whether the observed view overlaps identify a unique common measurement scale. Since additive view biases b_j do not affect the covariance between summaries for the same unit across two views, we subtract them and rewrite the observed summary of unit i under view j as

$$y_{ij} = a_j\theta_i + \varepsilon_{ij}, \qquad a_j > 0. \tag{7}$$

For each pair of views (j,k) that share observed units, let $\mathcal{I}_{jk} = \{i : (i,j),(i,k) \in \mathcal{I}\}$. We assume

$$\begin{aligned} &\mathbb{E}(\theta_i \mid i \in \mathcal{I}_{jk}) = 0, \qquad \mathrm{Var}(\theta_i \mid i \in \mathcal{I}_{jk}) = v > 0, \\ &\mathbb{E}(\varepsilon_{iq} \mid \theta_i, i \in \mathcal{I}_{jk}) = 0\ (q \in \{j,k\}), \qquad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik} \mid i \in \mathcal{I}_{jk}) = 0. \end{aligned} \tag{8}$$

Here, the same latent-score variance v applies to every shared-unit set $\mathcal{I}_{jk}$. This holds, for example, if membership in $\mathcal{I}_{jk}$ is independent of θ_i, analogous to ignorable missingness (Rubin, 1976). Under the residual assumptions in the second line of (8), we have

$$M_{jk} := \mathrm{Cov}(y_{ij}, y_{ik} \mid i \in \mathcal{I}_{jk}) = v a_j a_k. \tag{9}$$

If the latent-score variance differs across the shared-unit sets $\mathcal{I}_{jk}$, this factorization can fail, and the overlap structure alone cannot identify one common scale.

From overlap to a graph. We define the co-observation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, we consider views as vertices and connect vertices j and k with an edge when the two views share enough units to estimate M_{jk}. We define the node scale $u_j := \sqrt{v}\, a_j$ which combines the common latent-score variance v and the sensitivity a_j of view j. (9) then becomes $M_{jk} = u_j u_k$ on every edge.

A connected component $\mathcal{C} = (\mathcal{V}_C, \mathcal{E}_C)$ is *bipartite* if its vertices can be partitioned into two sets such that every edge connects the two sets, with no edge inside each vertex set. Equivalently, a component

is bipartite if and only if it contains no odd cycle. Thus stars, chains, trees, and even cycles are bipartite. For a bipartite component, multiplying node scale u_j by c on one side and by c^{-1} on the other leaves every edge product $u_j u_k$ unchanged, so the scale is not uniquely identified. An odd cycle rules out this nontrivial rescaling.

To express this algebraically, let $\gamma_j = \log u_j$ and $\ell_{jk} = \log M_{jk}$. Let $\boldsymbol{\gamma}$ collect the γ_j and $\boldsymbol{\ell}$ the ℓ_{jk} over all edges. We define the signless incidence matrix $A_{\mathcal{G}} \in \{0,1\}^{|\mathcal{E}|\times J}$, whose row for edge (j,k) is $\mathbf{e}_j^\top + \mathbf{e}_k^\top$, where $\mathbf{e}_j$ is the jth standard basis vector in $\mathbb{R}^J$. Then, it follows that

$$\ell_{jk} = \gamma_j + \gamma_k, \qquad \boldsymbol{\ell} = A_{\mathcal{G}}\boldsymbol{\gamma}. \tag{10}$$

Let $b(\mathcal{G})$ be the number of bipartite connected components, including isolated vertices. For $r = 1, \ldots, b(\mathcal{G})$, write the rth such component as $\mathcal{C}_r = (\mathcal{V}_r, \mathcal{E}_r)$ and choose a bipartition $\mathcal{V}_r = \mathcal{A}_r \dot\cup \mathcal{B}_r$. We define the alternating vector $\mathbf{s}_r \in \{0, \pm 1\}^J$ to be $+1$ on $\mathcal{A}_r$, -1 on $\mathcal{B}_r$, and zero outside $\mathcal{V}_r$.

The same non-bipartite graph condition appears in sparse rank-one worker-skill recovery (Ma et al., 2020). Here, we study the positive cross-view moment model $M_{jk} = u_j u_k$ under the overlap assumptions in (8), and make its componentwise nullspace and identified set explicit.

Theorem 3.1 (Graph identifiability under the one-factor overlap model)**.** *Suppose $M_{jk} = u_j u_k > 0$ for every edge $(j,k) \in \mathcal{E}$. Then*

$$\operatorname{rank}(A_{\mathcal{G}}) = J - b(\mathcal{G}), \qquad \operatorname{Null}(A_{\mathcal{G}}) = \operatorname{span}\{\mathbf{s}_1, \ldots, \mathbf{s}_{b(\mathcal{G})}\}, \tag{11}$$

where the span is $\{\mathbf{0}\}$ when $b(\mathcal{G}) = 0$. If $\boldsymbol{\gamma}^\star$ solves (10)*, then*

$$\mathrm{ID}_\gamma(\mathcal{G}, \boldsymbol{\ell}) = \{\boldsymbol{\gamma}^\star + \mathbf{h} : \mathbf{h} \in \operatorname{Null}(A_{\mathcal{G}})\}. \tag{12}$$

Thus each bipartite component contributes one alternating scale direction, whereas a component containing an odd cycle contributes none.

If the anchor view j_0 lies in a non-bipartite component, then u_{j_0} is identified. With $a_{j_0} = 1$, this gives $v = u_{j_0}^2$ and $a_j = u_j / u_{j_0}$ for every identified u_j. Bipartite components retain their alternating ambiguity no matter how precisely the existing edge moments are estimated. Information beyond these off-diagonal moments may further restrict latent-score variance v and the view sensitivities a_j.

3.2 Conditioning, model checks, and repair

We have established when the scale is identified. We now study how well it is determined, whether the observed moments fit the model, and which additional overlaps would improve the measurement design. Each overlap gives one equation relating the latent-score variance v and the relevant view sensitivities a_j. We collect these equations in a matrix to study rank and conditioning.

Fix the anchor view j_0. We define $\beta = \log v$, $\alpha_j = \log a_j$, and $\alpha_{j_0} = 0$, and collect the free parameters as $\boldsymbol{\eta} = \big(\beta, (\alpha_q)\big)^\top \in \mathbb{R}^J, \quad q \in \mathcal{V} \setminus \{j_0\}$. For each edge $e = (j,k)$, we define its row of the *anchored edge matrix* $B_{\mathcal{G},j_0}$ by $(B_{\mathcal{G},j_0})_{e,\beta} = 1$, and $(B_{\mathcal{G},j_0})_{e,\alpha_q} = \mathbf{1}\{q \in \{j,k\}\}$. Thus, an edge (j_0,k) represents $\ell_{j_0 k} = \beta + \alpha_k$, while an edge with neither endpoint equal to j_0 represents $\ell_{jk} = \beta + \alpha_j + \alpha_k$. Stacking these rows gives $\boldsymbol{\ell} = B_{\mathcal{G},j_0}\boldsymbol{\eta}$.

The invertible change of variables $\gamma_{j_0} = \beta/2$ and $\gamma_j = \beta/2 + \alpha_j$ shows that $\operatorname{rank}(B_{\mathcal{G},j_0}) = \operatorname{rank}(A_{\mathcal{G}})$. Therefore, $\delta(\mathcal{G}, j_0) := J - \operatorname{rank}(B_{\mathcal{G},j_0}) = b(\mathcal{G})$. When $\delta > 0$, the anchored parameters are not uniquely determined by the edge moments alone. When $\delta = 0$, they are identified from those moments under the working factorization, but their estimation can still be unstable. For $\widehat{\boldsymbol{\ell}} = \boldsymbol{\ell} + \boldsymbol{\xi}$, Lemma C.3 gives

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \le \frac{\|\boldsymbol{\xi}\|_2}{\sigma_{\min}(B_{\mathcal{G},j_0})}, \tag{13}$$

so a small singular value amplifies edge-moment error in the recovered scale parameters.

To quantify precision in the retained moments, we consider a Gaussian model for the estimated log edge moments, with known, parameter-independent covariance factor $\boldsymbol{\Omega} \succ 0$. Write $F_{\mathcal{G},j_0} := B_{\mathcal{G},j_0}^\top \boldsymbol{\Omega}^{-1} B_{\mathcal{G},j_0}$.

Theorem 3.2 (Information bound in the Gaussian edge-moment model)**.** *Suppose*

$$\widehat{\boldsymbol{\ell}} \mid \boldsymbol{\eta} \sim \mathcal{N}\left(B_{\mathcal{G},j_0}\boldsymbol{\eta}, \frac{\boldsymbol{\Omega}}{N}\right), \qquad \boldsymbol{\eta} \in \mathbb{R}^J,$$

where $N > 0$ and $\mathbf{\Omega}$ is known and does not depend on $\boldsymbol{\eta}$. If $B_{\mathcal{G},j_0}$ has full column rank, every unbiased estimator $\widetilde{\boldsymbol{\eta}}$ based on $\widehat{\boldsymbol{\ell}}$, with finite covariance and valid differentiation under expectation, satisfies

$$\mathrm{Var}_{\boldsymbol{\eta}}\big(\mathbf{d}^\top \widetilde{\boldsymbol{\eta}}\big) \geq \frac{1}{N}\mathbf{d}^\top F_{\mathcal{G},j_0}^{-1}\mathbf{d}$$

for every unit vector $\mathbf{d} \in \mathbb{R}^J$. The generalized least-squares estimator attains this bound. In the least-informed unit direction, the bound is $\{N\lambda_{\min}(F_{\mathcal{G},j_0})\}^{-1}$. If $B_{\mathcal{G},j_0}$ is rank deficient, its nonzero null directions are unidentifiable in this experiment.

The bound describes precision in this Gaussian edge-moment model. It is not a universal lower bound for estimators using the full unit-view table, and it does not provide a coverage guarantee for ACEB's plug-in posterior intervals.

Model checks. Identification and model checking are distinct. The star and triangle in Figure 1 illustrate identification: the star leaves the node factors ambiguous, whereas the triangle identifies them under the working model. Neither graph provides an additional relation that can be used to check the model itself. A four-cycle provides the simplest such check (Figure 5). Under the one-factor model, $M_{12}M_{34} = M_{23}M_{41}$, because both sides equal $u_1u_2u_3u_4$. A discrepancy beyond sampling error therefore indicates that no single positive node-factor model can explain all four edge moments. Proposition C.5 gives the general left-nullspace condition.

Repair. For a candidate edge $e = (j, k)$, let $\mathbf{z}_e^\top$ denote the row that this edge would add to $B_{\mathcal{G},j_0}$: it has one in the β column and in each non-anchor endpoint column, zero elsewhere. For a positive diagonal edge-precision matrix W, define $F_{\mathcal{G},j_0}(W) = B_{\mathcal{G},j_0}^\top W B_{\mathcal{G},j_0}$. With anticipated precision $w_e > 0$, adding edge e updates this matrix by $w_e\mathbf{z}_e\mathbf{z}_e^\top$. While $\delta > 0$, prioritize edges that increase rank. Within a connected bipartite component, such an edge joins two vertices in the same bipartition class. This closes an odd cycle and removes the corresponding alternating null direction. Once full rank is reached, let $\mathbf{d}_{\min}$ be a unit eigenvector associated with the smallest eigenvalue of $F_{\mathcal{G},j_0}(W)$. It is the least precisely measured linear combination of β and the non-anchor log loadings. If that eigenvalue has multiplicity one, the first-order gain from edge e is $g_e = w_e(\mathbf{d}_{\min}^\top \mathbf{z}_e)^2$. We therefore prioritize the feasible edge with the largest g_e. If the smallest eigenvalue is repeated, we instead compare the exact post-addition smallest eigenvalues. We refer to Proposition C.6 and Appendix C.3.8 for the formal statement and proof.

4 Experiments

We show in Section 3 which overlap patterns identify a common scale, when that scale may be weakly determined, and which new overlaps can repair deficient designs. Guided by these results, we test whether the predicted graph repairs improve finite-sample recovery, whether ACEB improves uncertainty estimates, and how these gains translate to selective decisions on real data.

4.1 Graph repair improves structural support

We find that repairing deficient overlap graphs consistently improves structural support. When the repair adds new observations, it also improves scale recovery on average.

We simulate observations from (1) with known latent scores θ_i, varying view sensitivities a_j, residual noise τ_j, and the overlap graph. The main benchmark contains 840 parameter-graph-seed combinations, with $n \in \{1000, 5000\}$, $J \in \{5, 10\}$, and five seeds. It includes bipartite and odd-cycle designs. We refer to Appendix D.1 for the full simulation grid.

In Figure 6, we confirm the structural prediction: all constructed deficient designs have positive rank deficiency, whereas the identified designs do not. To test the effect of repair, we compare paired designs with the same θ_i, a_j, τ_j, and base observations. Every repair reduces rank deficiency $\delta(\mathcal{G}, j_0)$ and improves conditioning. Adding observations improves ACEB leave-view's recovery of a_j, estimation of θ_i, and pairwise probabilities $p_{i,i'}$ on average. We find more variable results when the same repair is made by reallocating a fixed observation budget (Appendix D.10). These repairs improve structural support, but finite-sample gains also depend on the measurement budget.

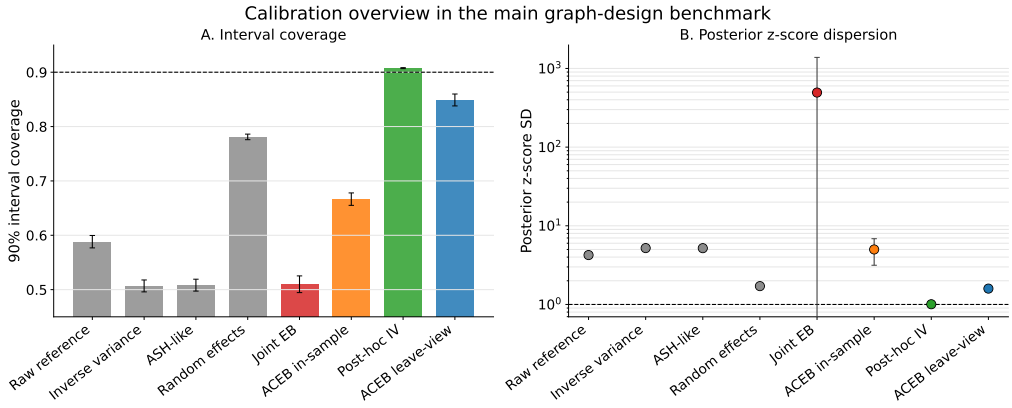


Figure 2: Synthetic uncertainty diagnostics. (A) 90% interval coverage; (B) Posterior z-score dispersion on a log scale. Bars in A and points in B show cell means with ± 2 SE. Nonpositive lower endpoints in B cannot be shown on the log scale. Post-hoc IV uses target labels for calibration.

4.2 Leave-view ACEB reduces overconfidence

We find that ACEB leave-view substantially reduces overconfidence in sparse heterogeneous settings, with leave-view residual-noise estimation explaining an important part of the improvement.

On the same synthetic benchmark, we compare raw reference scoring, inverse variance (IV), an adaptive-shrinkage-like baseline (ASH-like) based on a zero-centered normal-mixture prior (Stephens, 2017), random effects, joint empirical Bayes (EB), post-hoc IV, and ACEB with in-sample or leave-view residuals. Post-hoc IV uses targets for calibration; the other methods do not. We assess 90% interval coverage of θ_i and the standard deviation of $(m_i - \theta_i)/\sqrt{V_i}$, whose nominal value is one.

We summarize the main uncertainty results in Figure 2 and report the full metrics in Appendix D. Among target-free methods, ACEB leave-view is closest to nominal on both metrics, with coverage 0.849 and z-score dispersion 1.590 (Figure 2). Post-hoc IV reaches 0.908 coverage using target labels. ACEB leave-view also has fewer confident declarations of the wrong latent winner than post-hoc IV (false-best rate 0.006 versus 0.030). Joint EB achieves lower RMSE for θ_i (0.347 versus 0.601), but only 0.510 coverage and a false-best rate of 0.092 (Table 3). The source of its extreme dispersion remains unresolved. We discuss these optimization limitations in Appendix F.

Our ablations show that residual-noise estimation explains part of this improvement. Setting $\tau_j = 0$ or using in-sample residuals worsens calibration. Fixing $a_j = 1$ can still give good coverage despite poor sensitivity recovery, while using the true a_j and τ_j approaches nominal calibration (Appendix D.4).

We also test whether ACEB's results persist after aligning marginal coverage across methods. In this separate experiment, we fit a scalar variance multiplier on calibration units and evaluate it on held-out units. Among six methods, ACEB in-sample has the lowest mean pairwise Brier score, while ACEB leave-view has the lowest mean interval score (Figure 3); joint EB is not included. The pairwise Brier score compares $p_{i,i'}$ with whether $\theta_i > \theta_{i'}$, while the interval score balances interval width against failure to cover the true θ_i. Lower values are better for both metrics. In the dense negative control with $a_j = 1$ and $\tau_j = 0$, we find that ACEB performs similarly to simpler methods (Appendix D.6).

4.3 ACEB reduces false declarations through abstention on MedHELM

Using the public MedHELM instance-level outputs (Bedi et al., 2026), we construct 13,077 pairwise comparison records. Each record is the score difference between model A and model B within the same scenario and metric. For each record i, we split the available instances into observed blocks and a disjoint held-out set. We use about five observed blocks to form the summaries y_{ij} and a disjoint held-out set to form the target T_i with standard error q_i. We fit ACEB without these targets. We compare the held-out target T_i with the posterior mean m_i. For coverage, z-score dispersion, and interval score, we account for uncertainty in both quantities using $V_i + q_i^2$. For posterior interval width and sign decisions, we use the posterior variance V_i (Appendix E.1.1).

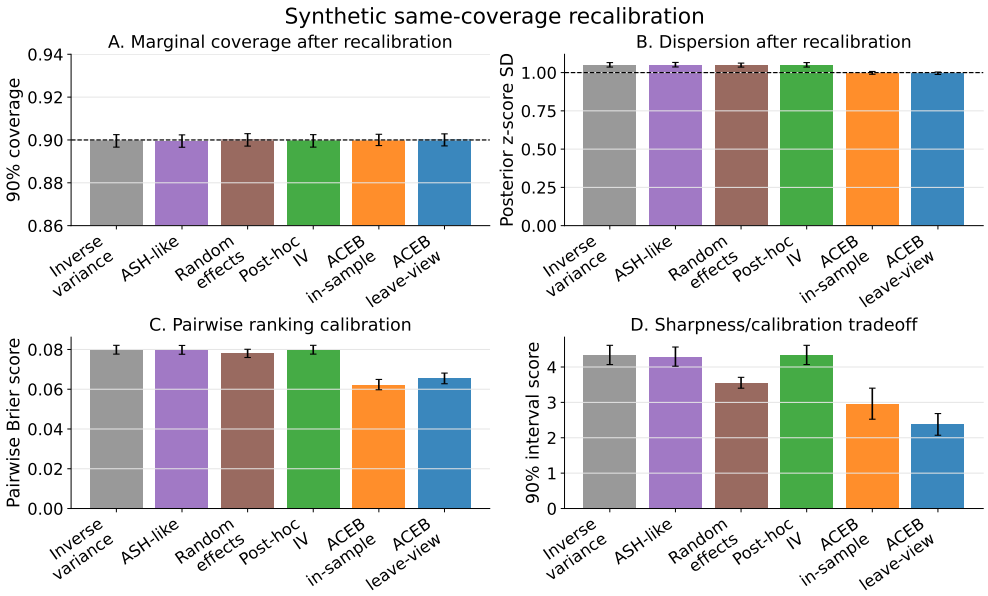


Figure 3: Synthetic same-coverage recalibration. Each method receives a scalar variance multiplier fitted on calibration items and evaluated on held-out items. (A) Marginal coverage. (B) Posterior z-score dispersion. (C) Pairwise Brier score. (D) Interval score; lower is better in C-D. Bars show means across cells; error bars show ± 2 SE.

Table 1: All-view MedHELM results. (A) Unsupervised audit without target calibration. Coverage and interval score (Int. score) use $V_i + q_i^2$, while posterior-only 90% interval width (Post. width) is based on V_i. (B) Scenario-held-out coverage-aligned decision audit.

A. Unsupervised, without target calibration

Method	Coverage	Int. score	Post. width	Decl. rate	Sel. acc.
Inverse variance	0.953	0.803	0.266	0.560	0.940
ASH-like	0.933	0.786	0.194	0.310	0.993
Random effects	0.962	0.774	0.297	0.559	0.941
ACEB leave-view	0.993	1.459	1.050	0.155	0.986

B. Scenario-held-out coverage-aligned decision audit

Method	Coverage	Int. score	False-sign	Decl. rate	False signed/record
Inverse variance	0.938	0.826	0.195	0.983	0.192
ASH-like	0.915	0.816	0.195	0.982	0.191
Random effects	0.943	0.792	0.192	0.983	0.189
ACEB leave-view	0.899	0.937	0.187	0.786	0.147

We summarize the main MedHELM results in Table 1 and Figure 4, with additional analyses in Appendix E.1.1. Without target calibration, ACEB produces conservative intervals. Leave-view coverage is 0.993, but its interval score is worse than random effects (1.459 versus 0.774; Table 1A). We observe the same conservative pattern across view subsets (Figure 4A-B).

After coverage alignment, ACEB leave-view makes 22% fewer false signed declarations per record than random effects (0.147 versus 0.189; Table 1B). We calibrate each method's posterior variance V_i using separate scenario-metric groups, then apply the learned variance scaling to held-out groups. We make sign decisions using the recalibrated variance at confidence threshold $\kappa = 0.95$. A false sign means disagreement with the sign of T_i, not with a known latent truth. ACEB declares on 78.6% rather than 98.3% of records, while false-sign rates among declarations are similar (0.187 versus 0.192). Thus, most of the reduction comes from abstention. Under this coverage-aligned comparison at $\kappa = 0.95$, ACEB has higher utility than random effects when an incorrect declaration costs more than about 3.7 times the value of a correct one (Appendix E.1.2). We also show how the error-abstention tradeoff changes with the confidence threshold in Figure 4C.

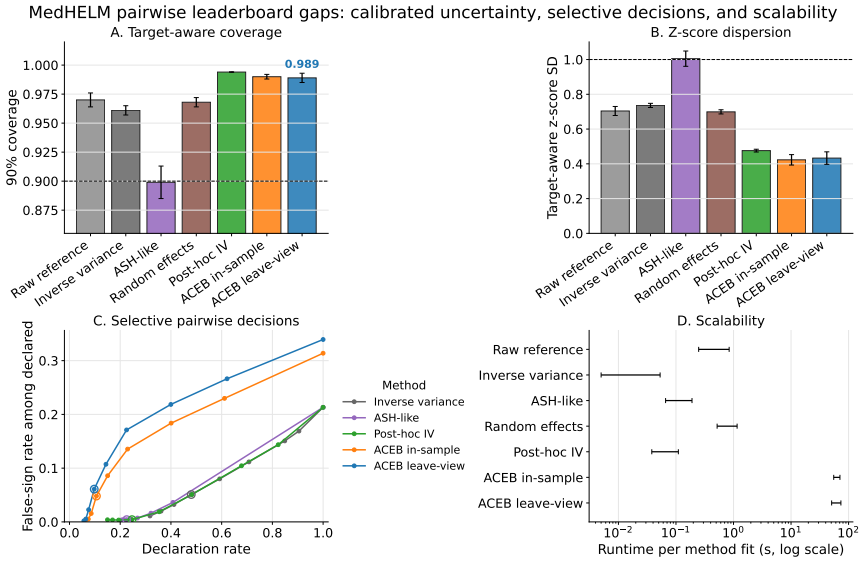


Figure 4: MedHELM uncertainty, selective decisions, and runtime. (A-B) Mean target-aware coverage and z-score dispersion across view subsets. (C) Raw-audit false-sign rate versus declaration rate across confidence thresholds; open circles mark $\kappa = 0.95$. (D) Per-fit runtime on a log scale.

We run two additional checks to see whether the findings extend beyond the MedHELM analysis. (1) Using different MedHELM view subsets, we find that star and chain designs are rank deficient, while odd-cycle, star-plus-edge, all-view, and random-sparse designs are full rank. ACEB also gives wider intervals in the deficient designs, consistent with weaker measurement support (Appendix E.1.3). (2) We repeat the uncertainty evaluation on SummEval (Fabbri et al., 2021), using crowd ratings as views and expert averages as held-out targets. ACEB improves raw coverage over IV and ASH-like, while the simpler methods achieve lower interval scores after same-coverage recalibration (Appendix E.2).

5 Discussion

Evaluation on leaderboards is challenging because of missing data and sparse measurement views. In this study, we show that the overlap pattern determines whether latent scores θ_i can be compared on one shared scale and how reliably that scale can be estimated. We model such overlapping patters as graphs, allowing us to distinguish structural scale ambiguity from finite-sample estimation uncertainty, checks model consistency, and guides which new overlaps to add. Building on these analysis, we propose ACEB, which combines graph-based diagnostics with empirical Bayes estimation, leave-view residual-noise estimation, and selective decisions. We validate in sparse heterogeneous simulations that ACEB reduces overconfidence relative to target-free baselines, and on MedHELM we observe a clear error-abstention tradeoff. These results motivate reporting uncertainty alongside what the observed cross-view moments identify.

This measurement-design perspective also matters for benchmark construction. Under the proposed moment model, strategically adding overlaps can resolve scale ambiguities that cannot be eliminated simply by collecting more observations on existing overlaps. This perspective also extends beyond leaderboards to related measurement problems in genomic effect estimation across conditions (Urbut et al., 2019) and sensor calibration without known ground truth (Stoffelen, 1998), where views may correspond to studies, tissues, or instruments. In these settings, the audit can help identify which shared measurements would be most informative to collect.

Our work has limitations. Our proposed model can fail with multidimensional scores, nonpositive or unit-specific sensitivities, correlated residuals, or performance-dependent observation. Cycle checks detect only some violations. In ACEB, we use plug-in parameters without full sample splitting, and the method can be conservative or sensitive to the reference view (Appendix D.7). The MedHELM records are dependent, and their target standard errors ignore cross-model covariance.

ETHICS STATEMENT

This study analyzes public benchmark outputs and previously collected human ratings; it collects no new human-subject or patient data. MedHELM is used only as an evaluation benchmark, and our results should not be interpreted as clinical-safety or deployment validation. The proposed audit is intended to reduce unsupported comparative claims, but its modeling assumptions and stated limitations remain material in any high-stakes use.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs are in Appendix C. The exact ACEB objectives, thresholds, numerical safeguards, and optimizer settings are recorded in Appendices B and F. The synthetic data-generating process and full experiment grids appear in Appendix D. Appendix E documents the MedHELM and SummEval construction and target-aware evaluation formulas.

AI USE STATEMENT

We used ChatGPT to assist with editorial refinement of the manuscript, including proofreading, grammar correction, and polishing writing. We used ChatGPT and Gemini to help generate the schematic illustrations in Figure 1 and Figure 5. All AI-assisted text and figures were reviewed, edited where necessary, and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32(3):943–951, 2026. doi: 10.1038/s41591-025-04151-2.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c7e1249ffc03eb9ded908c236bd1996d-Abstract-round2.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Dragoš Cvetković, Peter Rowlinson, and Slobodan K. Simić. Signless Laplacians of finite graphs. *Linear Algebra and its Applications*, 423(1):155–171, 2007. doi: 10.1016/j.laa.2007.01.009.

A. Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77 (379):605–610, 1982. doi: 10.1080/01621459.1982.10477856.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28, 1979. doi: 10.2307/2346806.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986. doi: 10.1016/0197-2456(86)90046-2.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance—how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1266. URL https://aclanthology.org/P19-1266/.

Bradley Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 2010.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4846–4853. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393/.

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. doi: 10.1162/tacl_a_00373.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Arpita Ghosh and Stephen Boyd. Growing well-connected graphs. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 6605–6611, 2006. doi: 10.1109/CDC.2006.377282.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379, 1961.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016. doi: 10.1145/2897350.2897352.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

Rachel Longjohn, Giri Gopalan, and Emily Casleton. Statistical uncertainty quantification for aggregate performance metrics in machine learning benchmarks. *arXiv preprint arXiv:2501.04234*, 2025. URL https://arxiv.org/abs/2501.04234.

Yao Ma, Alex Olshevsky, Venkatesh Saligrama, and Csaba Szepesvari. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. *Journal of Machine Learning Research*, 21(133):1–36, 2020. URL https://jmlr.org/papers/v21/19-359.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43): 1297–1322, 2010. URL https://jmlr.org/papers/v11/raykar10a.html.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1035. URL https://aclanthology.org/D17-1035/.

Herbert Robbins. An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4486–4503. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.346. URL https://aclanthology.org/2021.acl-long.346/.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Vicente Ivan Sanchez Carmona, Shanshan Jiang, and Bin Dong. Towards robust comparisons of NLP models: A case study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4973–4979. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.332/.

Matthew Stephens. False discovery rates: A new deal. *Biostatistics*, 18(2):275–294, 2017. doi: 10.1093/biostatistics/kxw041.

Ad Stoffelen. Toward the true near-surface wind speed: Error modeling and calibration using triple collocation. *Journal of Geophysical Research: Oceans*, 103(C4):7755–7766, 1998. doi: 10.1029/97JC03180.

Sarah M. Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, 51: 187–195, 2019. doi: 10.1038/s41588-018-0268-8.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

This supplement begins with a notation guide and is then organized as follows. Appendix A expands the related-work positioning; Appendix B gives additional ACEB details; Appendix C states auxiliary theoretical results and provides proofs; Appendix D reports the full synthetic design and additional results; Appendix E gives the MedHELM and SummEval details; and Appendix F records reproducibility checks.

NOTATION GUIDE

Scalars are italic, vectors are bold lowercase, matrices are uppercase, and sets are calligraphic. A hat denotes an estimate, a tilde a competing estimator, and $\dot{\sim}$ a working approximation. The table gives formal definitions for the global symbols used repeatedly in the paper; symbols local to one proof or experiment are defined at first use. We use $\mathbb{E}$, Var, Cov, rank, Null, and span for expectation, variance, covariance, rank, nullspace, and linear span.

Symbol	Definition or role
Data and measurement model	
$i, i';\ j, k;\ n, J$	Unit indices, view indices, and their counts; i' denotes the second unit in a pairwise comparison.
$\mathcal{V}, \mathcal{I}, \mathcal{J}_i$	$\mathcal{V} = \{0, \dots, J-1\}$; $\mathcal{I} = \{(i,j) : y_{ij} \text{ is observed}\}$; $\mathcal{J}_i = \{j : (i,j) \in \mathcal{I}\}$.
$\mathcal{I}_j, n_j;\ \mathcal{I}_{jk}, n_{jk}$	$\mathcal{I}_j = \{i : (i,j) \in \mathcal{I}\}$, $n_j = \lvert\mathcal{I}_j\rvert$; $\mathcal{I}_{jk} = \{i : (i,j), (i,k) \in \mathcal{I}\}$, $n_{jk} = \lvert\mathcal{I}_{jk}\rvert$.
$y_{ij}, s_{ij}, \mathbf{x}_i,$ $\mathcal{D}, \mathcal{D}_{i,-j}$	Reported summary, reported standard error, covariates, full data, and unit-i data with view j omitted.
$\theta_i, a_j, b_j, \tau_j, \varepsilon_{ij}$	Latent score, positive loading, additive bias, extra residual standard deviation, and centered residual in $y_{ij} = b_j + a_j\theta_i + \varepsilon_{ij}$. Reported fits set $b_j = 0$.
$\sigma_{ij}^2, \mu_i, v_i, v$	$\sigma_{ij}^2 = s_{ij}^2 + \tau_j^2$; (μ_i, v_i) are the prior moments of θ_i; $v = \mathrm{Var}(\theta_i \mid i \in \mathcal{I}_{jk})$ is the common edge-subpopulation variance.
$f_\mu, f_v, \mathbf{w}_\mu, \mathbf{w}_v, \zeta_j$	Prior functions and implemented coefficients; $\mu_i = \mathbf{x}_i^\top \mathbf{w}_\mu$, $v_i = \mathrm{softplus}(\mathbf{x}_i^\top \mathbf{w}_v) + 10^{-8}$, and $a_j = e^{\zeta_j}$ for $j \neq j_0$.
$m_i, V_i;$ $m_{i,-j}, V_{i,-j}$	Plug-in posterior mean and variance using all observed views, and the corresponding leave-view moments.
$p_{i,i'}, \pi_i, \Phi, \kappa$	Pairwise win probability, sign probability $\pi_i = \Phi(m_i/\sqrt{V_i})$ for an oriented gap, standard normal cdf, and decision threshold.
$[z]_+, \rho_{ij}$	$[z]_+ = \max\{z, 0\}$; $\rho_{ij} = a_j^2 v_i / (a_j^2 v_i + s_{ij}^2 + \tau_j^2)$.
Overlap graph and structural quantities	
$\mathcal{G} = (\mathcal{V}, \mathcal{E}), n_{\min}$	View-overlap graph; $(j,k) \in \mathcal{E}$ when $n_{jk} \geq n_{\min}$ in the diagnostic rule.
$b(\mathcal{G}), \mathcal{C}_r, \mathcal{A}_r, \mathcal{B}_r, \mathbf{s}_r$	$b(\mathcal{G})$ counts bipartite connected components, including isolated vertices; $\mathcal{C}_r = (\mathcal{V}_r, \mathcal{E}_r)$ is the rth such component, $\mathcal{V}_r = \mathcal{A}_r \dot\cup \mathcal{B}_r$, and $\mathbf{s}_r$ is $+1/-1/0$ on $\mathcal{A}_r/\mathcal{B}_r/\mathcal{V}_r^c$.
$M_{jk}, u_j, \gamma_j, \ell_{jk}$	$M_{jk} = \mathrm{Cov}(y_{ij}, y_{ik} \mid i \in \mathcal{I}_{jk}) = v a_j a_k$; $u_j = \sqrt{v}\, a_j$; $\gamma_j = \log u_j$; $\ell_{jk} = \log M_{jk}$.
$\boldsymbol{\gamma}, \boldsymbol{\ell}, A_{\mathcal{G}}$	Vectors collecting γ_j and ordered ℓ_{jk}; $A_{\mathcal{G}}$ is the signless incidence matrix, so $\boldsymbol{\ell} = A_{\mathcal{G}}\boldsymbol{\gamma}$.
$\mathsf{ID}_\gamma, \mathbf{M}_{\mathcal{E}}, \mathsf{ID}_{j_0}, \Psi$	Identified sets in the log-node and anchored variance-loading parameterizations, ordered edge moments, and a generic scale-sensitive functional.
$j_0, \beta, \alpha_j, \boldsymbol{\eta}$	Fixed reference view; $\beta = \log v$, $\alpha_j = \log a_j$, $\alpha_{j_0} = 0$, and $\boldsymbol{\eta} = (\beta, (\alpha_j)_{j \neq j_0})^\top$.
$\mathbf{z}_e, B_{\mathcal{G},j_0}$	For edge $e = (j,k)$, $\mathbf{z}_e^\top$ has one in the β column and in each non-anchor endpoint column, and zero elsewhere; $B_{\mathcal{G},j_0}$ stacks these rows, so $\boldsymbol{\ell} = B_{\mathcal{G},j_0}\boldsymbol{\eta}$.
$\delta(\mathcal{G}, j_0), \boldsymbol{\xi}, \sigma_{\min}$	$\delta = J - \mathrm{rank}(B_{\mathcal{G},j_0}) = b(\mathcal{G})$; $\widehat{\boldsymbol{\ell}} = \boldsymbol{\ell} + \boldsymbol{\xi}$; $\sigma_{\min}$ is the smallest full-column singular value.
$N, \boldsymbol{\Omega}, F_{\mathcal{G},j_0}$	Effective information scale, covariance factor of the Gaussian edge-moment experiment, and $F_{\mathcal{G},j_0} = B_{\mathcal{G},j_0}^\top \boldsymbol{\Omega}^{-1} B_{\mathcal{G},j_0}$.
$\mathbf{d}, \mathbf{d}_{\min}, \lambda_{\min}$	Unit parameter direction, least-informed unit direction, and smallest eigenvalue of the information matrix.

Symbol	Definition or role
$d_{\mathcal{G}}, L_{\mathcal{G}}$	$d_{\mathcal{G}} = \lvert\mathcal{E}\rvert - \operatorname{rank}(A_{\mathcal{G}})$; columns of $L_{\mathcal{G}}$ span $\operatorname{Null}(A_{\mathcal{G}}^{\top})$.
W, w_e, g_e	Diagonal log-moment precision matrix, anticipated precision of candidate edge e, and $g_e = w_e(\mathbf{d}_{\min}^{\top}\mathbf{z}_e)^2$.
Recalibration and real-data evaluation	
$\widehat{\theta}_{i,h}, \widehat{\sigma}_{i,h}, c, \widehat{c}_h$	Method-h posterior mean and standard deviation, a candidate scalar inflation, and the learned inflation.
$T_i, q_i, \sigma^2_{i,\mathrm{TA}}$	Aggregate held-out target, its standard error, and $\sigma^2_{i,\mathrm{TA}} = \max\{V_i, 10^{-18}\} + q_i^2$.
T_M, q_M; T_{AB}, q_{AB}	Model-level held-out target and standard error; oriented model-pair target gap and quadrature standard error.
$z_{.95}, L_i, U_i, \mathrm{IS}_i$	$z_{.95} = \Phi^{-1}(0.95)$; target-aware interval endpoints and 90% interval score.
$d_h, f_h, \lambda, U_\lambda(h)$	Declaration rate, false-sign rate among declarations, relative false-claim cost, and $U_\lambda(h) = d_h(1 - f_h) - \lambda d_h f_h$.
$O_p(1), o_p(1), \xrightarrow{p}$, $\Rightarrow$	Standard boundedness and convergence notation used in the consistency and information results.

A Extended related work

Benchmark variance, comparison, and leaderboards. Repeated runs, examples, tasks, and hyperparameter choices can materially alter reported performance, motivating score distributions, resampling, hierarchical comparisons, and uncertainty-aware leaderboards (Reimers & Gurevych, 2017; Bouthillier et al., 2021; Agarwal et al., 2021; Dror et al., 2019; Rodriguez et al., 2021; Sanchez Carmona et al., 2025; Longjohn et al., 2025). Leaderboards can also optimize a proxy that differs from user utility (Ethayarajh & Jurafsky, 2020). This literature asks how to compare systems once an evaluation table is specified. We ask the logically prior question of whether the table's overlap pattern identifies the latent scale on which interval and ranking claims are expressed.

Calibration, uncertainty, and selective prediction. Classical and modern work studies probabilistic calibration, calibrated regression, proper scoring rules, and uncertainty under shift (Brier, 1950; Dawid, 1982; Guo et al., 2017; Kuleshov et al., 2018; Gneiting & Raftery, 2007; Ovadia et al., 2019). Selective classification explicitly trades coverage against error by allowing abstention (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). ACEB adopts that reporting principle, but its graph layer addresses a different failure: before assessing calibration or a risk-coverage curve, it asks whether the observed overlaps support a unique scale-sensitive quantity under the working model.

Empirical Bayes and heterogeneous aggregation. Empirical-Bayes and shrinkage estimators stabilize noisy effects by borrowing strength across units (Robbins, 1956; James & Stein, 1961; Efron, 2010; Stephens, 2017); multivariate shrinkage extends this idea across conditions (Urbut et al., 2019), and random-effects meta-analysis models excess heterogeneity (DerSimonian & Laird, 1986). ACEB uses empirical Bayes as a computational layer rather than claiming a new shrinkage prior. Its distinct role is to couple view-specific loadings and leave-view residual-noise estimation to an explicit audit of measurement-design support.

Latent truth, annotator reliability, and identifiability. Dawid-Skene-style models, learning-from-crowds methods, and truth discovery jointly infer latent quantities and source reliability (Dawid & Skene, 1979; Raykar et al., 2010; Li et al., 2016). General latent-structure identifiability is a well-developed subject (Allman et al., 2009). Our theorem is narrower and more operational: for continuous positive cross-view moments with a shared one-factor loading, it gives the complete identified set directly from the view-overlap graph. The common-overlap condition is also related to classical ignorability questions for missing data (Rubin, 1976); outcome-dependent coverage can invalidate the shared node factorization even when the topology is unchanged.

Evaluation infrastructure and human judgments. OpenML and its benchmark suites support reproducible task collections (Vanschoren et al., 2014; Bischl et al., 2021); HELM and MedHELM

organize broad, heterogeneous language-model evaluations (Liang et al., 2023; Bedi et al., 2026); and SummEval provides multi-annotator human judgments for summarization (Fabbri et al., 2021). These resources motivate our multi-view setting. ACEB complements them by auditing whether co-observation structure supports the uncertainty claims made from their scores, rather than proposing a replacement benchmark.

Graph structure, model checks, and design. The edge equations contain sums of endpoint log factors, giving a signless-incidence representation. The identification argument is closely related to sparse rank-one completion (Ma et al., 2020), and the relevant graph spectra are classical (Cvetković et al., 2007). Our graph analysis specializes these ideas to the stated continuous measurement model and makes the componentwise identified set explicit. Left-nullspace restrictions describe compatibility of the edge moments, while spectral criteria guide overlap repair (Ghosh & Boyd, 2006). These results form the diagnostic layer of ACEB; the fitting and reporting layer additionally uses reported standard errors, fixed-anchor empirical Bayes, and leave-view residual-noise estimation. Triple collocation similarly estimates calibration coefficients and error variances from multiple noisy measurement systems without an error-free reference (Stoffelen, 1998). Its linear three-view setting has the same pairwise covariance factorization; our analysis considers sparse overlap graphs, their componentwise ambiguities, model checks, and the design of additional overlaps.

Partial orders and selective reporting. A partial order is a complementary reporting object: it records which comparisons are asserted or withheld. It does not determine whether noisy heterogeneous views identify a common probabilistic scale, quantify weak support, or select an informative new overlap. ACEB's abstentions can be read as a data-dependent partial report, while the graph audit explains why a claim is or is not supported in the first place.

B ACEB IMPLEMENTATION AND METHOD DETAILS

B.1 LOCATION AND SCALE CONVENTIONS, PRIOR SPECIFICATION, AND RELIABILITY

With freely estimated view biases, the general likelihood in (1) is invariant under

$$\theta_i' = \theta_i + c, \qquad b_j' = b_j - a_j c.$$

The reported implementation does not estimate this location transformation: it fixes $b_j = 0$, includes an intercept in $\mathbf{x}_i$, and learns the corresponding prior-mean intercept. Thus the oriented benchmark scores supply the location convention. Exactly one reference view fixes the measurement unit through $a_{j_0} = 1$. This scale convention is distinct from graph identification: in a deficient overlap graph, multiple latent variances and non-reference loadings can still satisfy the same edge moments after $a_{j_0} = 1$ is imposed.

A common positive change of units is not itself an identification problem: if $m_i' = cm_i$ and $V_i' = c^2 V_i$ for $c > 0$, then $m_i'/\sqrt{V_i'} = m_i/\sqrt{V_i}$, and pairwise z-scores are likewise unchanged. The relevant ambiguity is variation in a scale-sensitive functional across admissible variance–loading decompositions after anchoring.

Without non-intercept covariates, $\mathbf{x}_i = 1$, so the reported score-only fit learns a common prior mean and a common prior variance. With covariates, an intercept is added and the non-intercept features are standardized before fitting. The unit-view signal-to-total-variance ratio

$$\rho_{ij} := \frac{a_j^2 v_i}{a_j^2 v_i + s_{ij}^2 + \tau_j^2} \in (0, 1) \tag{14}$$

is a derived reliability diagnostic. It is not an additional model parameter and does not enter the graph-identifiability theorem.

The item-specific prior variance v_i in (2) should not be confused with the marginal variance $v = \mathrm{Var}(\theta_i)$ used in the population moment analysis of Section 3. The former controls unit-level posterior shrinkage; the latter defines the cross-view second-moment scale.

B.2 Posterior summaries and decision probabilities

Let

$$\mathcal{D}_i = \{(y_{ij}, s_{ij}) : j \in \mathcal{J}_i\}.$$

Conditional on the fitted hyperparameters, the posterior kernel is

$$p(\theta_i \mid \mathcal{D}_i) \propto \exp\left[-\frac{(\theta_i - \mu_i)^2}{2v_i} - \frac{1}{2}\sum_{j\in\mathcal{J}_i}\frac{(y_{ij} - b_j - a_j\theta_i)^2}{s_{ij}^2 + \tau_j^2}\right].$$

Completing the square gives the moments in (3).

More generally, let

$$C_{i,i'} = \mathrm{Cov}(\theta_i, \theta_{i'} \mid \mathcal{D})$$

denote posterior covariance between two units. Their Gaussian pairwise probability is

$$\mathbb{P}(\theta_i > \theta_{i'} \mid \mathcal{D}) \approx \Phi\left(\frac{m_i - m_{i'}}{\sqrt{V_i + V_{i'} - 2C_{i,i'}}}\right). \tag{15}$$

The plug-in ACEB model factorizes over units conditional on fitted hyperparameters, giving $C_{i,i'} = 0$ and hence (4). This plug-in approximation does not propagate hyperparameter uncertainty; that limitation is discussed in the main paper.

B.3 Implementation contract for the reported experiments

This subsection records the staged fixed-anchor implementation used to generate the reported results. Algorithm 1 summarizes the procedure, and Table 2 records the exact implementation settings.

For two views j and k, let $\mathcal{I}_{jk} := \{i : (i,j), (i,k) \in \mathcal{I}\}$ and $n_{jk} := |\mathcal{I}_{jk}|$.

Algorithm 1 ACEB: identifiability-aware uncertainty audit

1: **Input:** $\{(y_{ij}, s_{ij})\}_{(i,j)\in\mathcal{I}}$, optional $\{\mathbf{x}_i\}$, reference view j_0, overlap threshold $n_{\min}$, fixed penalties, and decision threshold κ.
2: Build the view-overlap graph from pairs with $n_{jk} \geq n_{\min}$; record centered edge moments, positivity, rank deficiency, and conditioning. These diagnostics do not gate decisions.
3: Fit the prior and positive non-reference loadings with $a_{j_0} = 1$.
4: Estimate τ_j^2 from the leave-view moments in (6); recompute m_i, V_i, and directional probabilities.
5: Apply κ to obtain positive, negative, or abstain decisions. Return the posterior summaries together with the graph diagnostics.

Staged objectives. Let $\mathbf{x}_i$ include an intercept. The prior mean and variance are

$$\mu_i = \mathbf{x}_i^\top \mathbf{w}_\mu, \qquad v_i = \mathrm{softplus}(\mathbf{x}_i^\top \mathbf{w}_v) + 10^{-8}.$$

The first stage minimizes the negative Gaussian marginal log likelihood plus

$$10^{-4}\big(\|\mathbf{w}_\mu\|_2^2 + \|\mathbf{w}_v\|_2^2\big). \tag{16}$$

The reference loading is fixed at $a_{j_0} = 1$. For $j \neq j_0$, the implementation sets $a_j = \exp(\zeta_j)$ and minimizes the negative Gaussian marginal log likelihood plus

$$10^{-4}\sum_{j\neq j_0}(a_j - 1)^2. \tag{17}$$

View biases are fixed at zero in all reported experiments. Per-view residual noise is then estimated using the leave-view moment in (6), positive-part truncated, clipped to $\tau_j \in [0, 10]$, and followed by a final posterior recomputation. Table 2 summarizes the exact implementation settings used in the reported experiments.

Table 2: Implementation contract used to generate the reported ACEB results. Experiment-specific optimizer budgets are listed in Appendix F.

Component	Contract used in the reported experiments
Diagnostic graph	Edge (j, k) is included when $n_{jk} \geq n_{\min}$. Cross-view moments are centered within view. Positivity above 10^{-12} is recorded as a compatibility flag but does not gate topological edge inclusion; no edge standard error or significance test is used. Main synthetic diagnostics use $n_{\min} = 1$, and the reported MedHELM run-level graph audit uses $n_{\min} = 10$.
Graph numerics	Rank uses tolerance $\max\{10^{-7},\ 10^{-7}\sigma_{\max}\}$. The reported conditioning statistic is the smallest full-column singular value divided by $\sqrt{\max\{\lvert\mathcal{E}\rvert, 1\}}$.
Reference view	Exactly one reference view is supplied. Synthetic experiments use view 0; the MedHELM adapter marks the first sorted view ID as reference. The reported SummEval run also supplies exactly one fixed reference view. There is no degree-, overlap-, rank-, or conditioning-based anchor admissibility rule in the reported fit.
Prior and loading fit	The penalties are given in (16)-(17). Non-reference loadings are optimized on the log scale; view biases are fixed at zero.
Initialization	Reference loading 1, non-reference loadings 0.5, residual noise 10^{-4}, and $\mathbf{w}_\mu = 0$. The intercept of $\mathbf{w}_v$ is initialized by the inverse-softplus of the centered input variance, with remaining coefficients zero.
Optimization	SciPy L-BFGS-B with finite-difference gradients, no restarts, and only the experiment-specific maximum iteration count set explicitly. Other stopping tolerances use SciPy defaults.
Numerical safeguards	Input standard errors are floored at 10^{-8} before dataset-specific floors; prior variances include a 10^{-8} floor, leave-view precision uses a 10^{-12} variance floor, and reported posterior variances are floored at 10^{-18}. Pairwise probabilities are clipped to $[10^{-8}, 1 - 10^{-8}]$ only for ECE computation.
Dataset-specific SEs	MedHELM observation standard errors are floored at 0.02. Its model-level target standard errors use a 0.01 floor and pairwise targets a 0.02 floor. SummEval uses a fixed crowd-view standard error of 0.75 and a target-SE floor of 0.10.
Decisions	The principal leaderboard threshold is $\kappa = 0.95$. Selective curves additionally use $\kappa \in \{0.50, 0.55, 0.60, 0.70, 0.80, 0.90, 0.95, 0.975, 0.99\}$.

B.4 Anchor-sensitivity diagnostic

The reported main results use one fixed reference view and do not intersect decisions across multiple anchors. The separate anchor-sensitivity experiment refits the model under alternative reference views solely to quantify finite-sample dependence on that convention.

To compare score summaries across fits, let j_0 be the primary reference and let $a_{j_0}^{(r)}$ be the estimated loading of view j_0 in a fit anchored at view r. We map that fit to the primary-reference scale by

$$m_i^{(r\to j_0)} = a_{j_0}^{(r)} m_i^{(r)}, \qquad V_i^{(r\to j_0)} = \big(a_{j_0}^{(r)}\big)^2 V_i^{(r)}. \tag{18}$$

Variation after this mapping is reported as estimator sensitivity. The anchor sweep in Appendix D.7 is diagnostic only and does not change the anchor or the main decisions.

B.5 Leave-view moment consistency

Fix a view j. For $i \in \mathcal{I}_j$, define the oracle predictive residual

$$r_{ij} := y_{ij} - b_j - a_j m_{i,-j}$$

and the oracle residual-moment contribution

$$R_{ij} := r_{ij}^2 - a_j^2 V_{i,-j} - s_{ij}^2. \tag{19}$$

Under the working model,

$$\mathbb{E}(R_{ij} \mid \mathcal{D}_{i,-j}) = \tau_j^2.$$

Proposition B.1 (Consistency of the leave-view residual moment)**.** *For a fixed view j, suppose that as $n_j \to \infty$:*

(i)

$$\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}(R_{ij}-\tau_j^2)\xrightarrow{p}0;$$

(ii) $\widehat{a}_j\xrightarrow{p}a_j$, $\widehat{b}_j\xrightarrow{p}b_j$,

$$\max_{i\in\mathcal{I}_j}|\widehat{m}_{i,-j}-m_{i,-j}|\xrightarrow{p}0,\qquad \max_{i\in\mathcal{I}_j}|\widehat{V}_{i,-j}-V_{i,-j}|\xrightarrow{p}0;$$

(iii)

$$\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\left(r_{ij}^2+m_{i,-j}^2+V_{i,-j}\right)=O_p(1).$$

Then the estimator in (6) *satisfies*

$$\widehat{\tau}_j^2\xrightarrow{p}\tau_j^2$$

for every $\tau_j^2\geq 0$.

Proof. Define

$$\Delta_{ij}:=(\widehat{b}_j-b_j)+(\widehat{a}_j-a_j)m_{i,-j}+\widehat{a}_j(\widehat{m}_{i,-j}-m_{i,-j}).$$

The fitted leave-view residual can then be written as

$$y_{ij}-\widehat{b}_j-\widehat{a}_j\widehat{m}_{i,-j}=r_{ij}-\Delta_{ij}.$$

Assumptions (ii)-(iii), together with $\widehat{a}_j=O_p(1)$, imply

$$\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\Delta_{ij}^2=o_p(1).$$

By the Cauchy-Schwarz inequality,

$$\left|\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\left\{(r_{ij}-\Delta_{ij})^2-r_{ij}^2\right\}\right|$$
$$\leq 2\left(\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}r_{ij}^2\right)^{1/2}\left(\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\Delta_{ij}^2\right)^{1/2}+\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\Delta_{ij}^2=o_p(1).$$

Moreover,

$$\left|\widehat{a}_j^2\widehat{V}_{i,-j}-a_j^2V_{i,-j}\right|\leq|\widehat{a}_j^2-a_j^2|V_{i,-j}+\widehat{a}_j^2|\widehat{V}_{i,-j}-V_{i,-j}|,$$

so assumptions (ii)-(iii) imply

$$\frac{1}{n_j}\sum_{i\in\mathcal{I}_j}\left|\widehat{a}_j^2\widehat{V}_{i,-j}-a_j^2V_{i,-j}\right|=o_p(1).$$

Therefore, the untruncated average in (6) equals $n_j^{-1}\sum_{i\in\mathcal{I}_j}R_{ij}+o_p(1)$, which converges in probability to τ_j^2 by assumption (i). Because $z\mapsto[z]_+$ is continuous on $\mathbb{R}$, the continuous mapping theorem gives $\widehat{\tau}_j^2\xrightarrow{p}\tau_j^2$. □

A sufficient condition for assumption (i) is an appropriate weak law of large numbers across units, for example conditional independence together with uniformly bounded $2+\epsilon$ moments for some $\epsilon>0$.

C Additional structural theory and proofs

C.1 Scope of the view-loading moment theorem

The structural results in Section 3 apply to the centered view-loading model in (7). This model is the second-moment layer of the ACEB likelihood in (1): additive biases are removed or estimated, reported standard errors and residual view variances affect diagonal uncertainty and posterior weighting, and the off-diagonal cross-view moments factor through the shared loading vector $\mathbf{a}$ when residuals are mean-independent of the latent score and uncorrelated across the two views.

The theorem does not require the empirical-Bayes prior variance to be constant across units. The quantities v_i in (2) govern unit-level posterior shrinkage, whereas v in (8) is the common marginal second moment of the centered latent scores on every edge-specific co-observed subset.

This common-overlap condition is essential. Without it, an edge can satisfy

$$M_{jk} = v_{jk} a_j a_k, \qquad v_{jk} = \mathrm{Var}(\theta_i \mid i \in \mathcal{I}_{jk}),$$

with a different v_{jk} for each overlap subset. Such edge-specific factors cannot generally be absorbed into node parameters, so the signless-incidence system no longer follows. Missingness independent of the latent score is a simple sufficient condition for common v; outcome-dependent benchmark coverage is a concrete failure mode. Cycle restrictions can reveal some such incompatibility, but satisfying them does not prove ignorability.

The view-loading restriction is essential. If loadings vary freely across both units and views,

$$y_{ij} = a_{ij}\theta_i + \varepsilon_{ij},$$

then, in general,

$$\mathrm{Cov}(y_{\cdot j}, y_{\cdot k}) = \mathbb{E}[a_{ij}a_{ik}\theta_i^2] - \mathbb{E}[a_{ij}\theta_i]\mathbb{E}[a_{ik}\theta_i], \tag{20}$$

which need not factor into one node-specific quantity for view j and one for view k. The overlap topology alone therefore cannot imply (10). Stratified loadings, low-rank loading models, and multidimensional latent scores are important extensions, but they are outside the formal claim considered here.

Additional information from diagonal variances. The edge-moment identified set need not equal the identified set from the full observation model. For example, consider a three-view star with edges $(0,1)$ and $(0,2)$, $a_0 = 1$, $s_{ij}^2 = 1$, $\mathrm{Var}(y_{ij}) = 2$, and $M_{01} = M_{02} = 1$. The edge equations give $a_1 = a_2 = 1/v$, while the diagonal variances imply

$$\tau_0^2 = 1 - v, \qquad \tau_1^2 = \tau_2^2 = 1 - \frac{1}{v}.$$

Since $\tau_j^2 \geq 0$, we must have $v = 1$. Thus the full variance structure can identify parameters even when the star's off-diagonal moment system is rank deficient. The graph theorem therefore concerns identification from the specified cross-view moments, not universal full-data non-identifiability.

C.2 Additional identified-set consequences

Corollary C.1 (Odd cycles, bipartite ridges, and anchored scale)**.** *Let $\mathcal{C} = (\mathcal{V}_C, \mathcal{E}_C)$ be a connected component of $\mathcal{G}$.*

(a) If $\mathcal{C}$ contains an odd cycle, then $(u_j)_{j\in\mathcal{V}_C}$ is uniquely identified from its edge moments.

(b) If $\mathcal{C}$ is bipartite with $\mathcal{V}_C = \mathcal{A}_C \,\dot\cup\, \mathcal{B}_C$, then for every $c > 0$,

$$u'_j = \begin{cases} cu_j, & j \in \mathcal{A}_C, \\ c^{-1}u_j, & j \in \mathcal{B}_C, \end{cases}$$

preserves every edge moment in $\mathcal{C}$.

(c) Suppose anchor j_0 lies in a component for which u_{j_0} is identified and ACEB imposes $a_{j_0} = 1$. Then

$$v = u_{j_0}^2, \qquad a_j = \frac{u_j}{u_{j_0}}$$

for every identified node factor u_j. Bipartite or isolated components retain their alternating scale directions unless additional scale information is supplied.

Write

$$\mathbf{M}_{\mathcal{E}} := (M_{jk})_{(j,k)\in\mathcal{E}}$$

for the ordered population edge moments. Fix j_0 and impose $a_{j_0} = 1$. The identified set in the variance-loading parameterization is

$$\mathsf{ID}_{j_0}(\mathcal{G}, \mathbf{M}_{\mathcal{E}}) := \left\{(v, \mathbf{a}) \in \mathbb{R}_{>0} \times \mathbb{R}^J_{>0} : a_{j_0} = 1,\ M_{jk} = v a_j a_k \text{ for every } (j,k) \in \mathcal{E}\right\}. \tag{21}$$

Corollary C.2 (No unique scale-sensitive inference along a ridge)**.** *Suppose $\mathsf{ID}_{j_0}(\mathcal{G}, \mathbf{M}_{\mathcal{E}})$ contains two parameter values $(v, \mathbf{a})$ and $(v', \mathbf{a}')$ that induce different latent-score scales or decision probabilities. Then those quantities are not identified by the observed edge moments alone. Identifying such a quantity requires information or restrictions that distinguish the relevant elements of the moment identified set. A prior or penalty can select a numerical representative without making that representative identified by the edge moments.*

Lemma C.3 (Finite-sample amplification by graph conditioning)**.** *Suppose $B_{\mathcal{G},j_0}$ has full column rank and $\widehat{\boldsymbol{\ell}} = \boldsymbol{\ell} + \boldsymbol{\xi}$. Define*

$$\widehat{\boldsymbol{\eta}} = (B_{\mathcal{G},j_0}^\top B_{\mathcal{G},j_0})^{-1} B_{\mathcal{G},j_0}^\top \widehat{\boldsymbol{\ell}}. \tag{22}$$

Then

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \le \frac{\|\boldsymbol{\xi}\|_2}{\sigma_{\min}(B_{\mathcal{G},j_0})}. \tag{23}$$

If $\widehat{\boldsymbol{\ell}}$ is formed by log-transforming noisy positive moment estimates, the same inequality applies to the first-order term, with the Taylor remainder added to the right-hand side.

Corollary C.4 (No-free leaderboard decisions)**.** *Let Ψ be a scale-sensitive functional of the anchored parameters, such as an interval width, posterior z-scale, pairwise win probability, or sign-decision set. If $B_{\mathcal{G},j_0}\boldsymbol{\eta}_1 = B_{\mathcal{G},j_0}\boldsymbol{\eta}_2$ but $\Psi(\boldsymbol{\eta}_1) \neq \Psi(\boldsymbol{\eta}_2)$, then Ψ is not identified by the observed edge moments. No target-label-free procedure whose scale information is restricted to $\boldsymbol{\ell}$ can certify a unique value of Ψ on that graph.*

Four-cycle example. The four-cycle makes the distinction between identification and model checking explicit. Because it is an even cycle, it is bipartite and still retains the alternating scale ambiguity. Nevertheless, under $M_{jk} = u_j u_k$, multiplying opposite edges gives

$$M_{12}M_{34} = (u_1u_2)(u_3u_4) = u_1u_2u_3u_4 = (u_2u_3)(u_4u_1) = M_{23}M_{41}.$$

Thus disagreement between the two products beyond sampling variability rejects the one-factor moment representation on these four views, although agreement does not validate the full model.

Proposition C.5 (Cycle-consistency restrictions)**.** *Let $L_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{E}|\times d_{\mathcal{G}}}$ have columns forming a basis of $\mathrm{Null}(A_{\mathcal{G}}^\top)$, where $d_{\mathcal{G}} := |\mathcal{E}| - \mathrm{rank}(A_{\mathcal{G}})$. Under* (9)*,*

$$L_{\mathcal{G}}^\top \boldsymbol{\ell} = 0. \tag{24}$$

For the ordered four-cycle $(1,2),(2,3),(3,4),(4,1)$, one restriction is

$$\log M_{12} + \log M_{34} - \log M_{23} - \log M_{41} = 0. \tag{25}$$

Systematic violations diagnose departures such as multidimensional abilities, correlated view errors, nonlinear metrics, nonconstant loadings, or edge-dependent overlap selection.

Proposition C.6 (One-edge graph repair and E-optimality)**.** *Let $W \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{E}|}$ be diagonal with strictly positive entries and define*

$$F_{\mathcal{G},j_0}(W) := B_{\mathcal{G},j_0}^\top W B_{\mathcal{G},j_0}. \tag{26}$$

For a feasible missing edge $e = (j,k)$, let $\mathbf{z}_e^\top$ have one in the β column and in each non-anchor endpoint column, and zero elsewhere; let $w_e > 0$ be its anticipated precision. Adding the edge gives

$$F_{\mathcal{G}+e,j_0} = F_{\mathcal{G},j_0}(W) + w_e \mathbf{z}_e \mathbf{z}_e^\top. \tag{27}$$

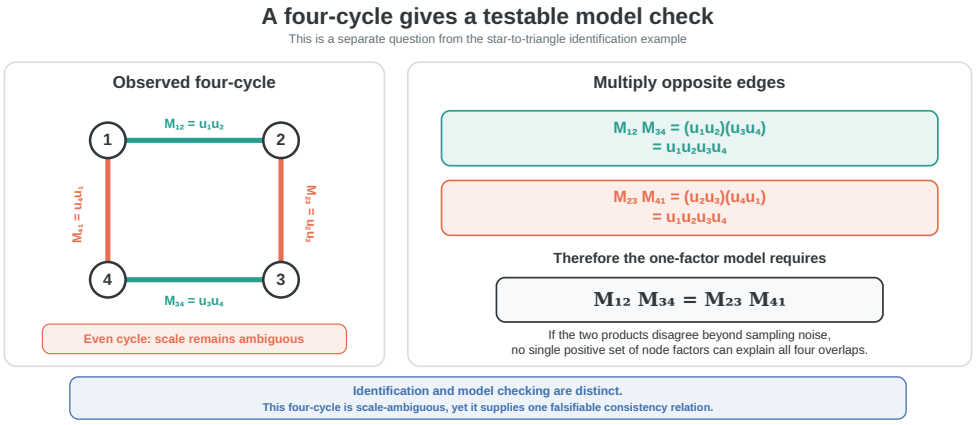


Figure 5: Four-cycle model check. The four-cycle is bipartite, so it still leaves an alternating scale ambiguity and is not an identification example. However, under $M_{jk} = u_j u_k$, both products of opposite edge moments equal $u_1u_2u_3u_4$, forcing $M_{12}M_{34} = M_{23}M_{41}$. A discrepancy beyond sampling variability indicates that no single positive node-factor model fits all four edge moments. This is the four-cycle's single independent left-nullspace restriction.

The rank of $B_{\mathcal{G},j_0}$ increases if and only if $\mathbf{z}_e^\top \notin \mathrm{Row}(B_{\mathcal{G},j_0})$. Within a connected bipartite component, this occurs exactly when e joins two vertices on the same side of the bipartition, equivalently when it closes an odd cycle.

If $F_{\mathcal{G},j_0}(W)$ is positive definite and its smallest eigenvalue is simple, let $\mathbf{d}_{\min}$ be the associated unit eigenvector. Then

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\lambda_{\min}\left(F_{\mathcal{G},j_0}(W) + \epsilon w_e \mathbf{z}_e \mathbf{z}_e^\top\right)\Big|_{\epsilon=0} = w_e(\mathbf{d}_{\min}^\top \mathbf{z}_e)^2. \tag{28}$$

Thus edge selection is two-stage: first choose an edge that reduces rank deficiency; after full rank, choose the feasible edge with the largest first-order increase in the smallest information eigenvalue.

C.3 Proofs

C.3.1 Proof of Theorem 3.1

After permuting rows and columns by connected component, $A_{\mathcal{G}}$ is block diagonal, so it suffices to characterize the nullspace on one connected component $\mathcal{C} = (\mathcal{V}_C, \mathcal{E}_C)$. Let A_C be its signless incidence matrix. A vector $\mathbf{h}_C \in \mathbb{R}^{|\mathcal{V}_C|}$ lies in $\mathrm{Null}(A_C)$ if and only if

$$h_j + h_k = 0 \qquad \text{for every } (j,k) \in \mathcal{E}_C. \tag{29}$$

Suppose first that $\mathcal{C}$ is bipartite with $\mathcal{V}_C = \mathcal{A}_C \dot\cup \mathcal{B}_C$. Define $\mathbf{s}_C$ to be $+1$ on $\mathcal{A}_C$ and -1 on $\mathcal{B}_C$. Every edge crosses the two classes, so $\mathbf{s}_C \in \mathrm{Null}(A_C)$. Conversely, fix any vertex $j_\star \in \mathcal{V}_C$. Equation (29) forces the entries of every $\mathbf{h}_C \in \mathrm{Null}(A_C)$ to alternate in sign along paths from $j_\star$. Connectivity therefore gives $\mathbf{h}_C = (h_{j_\star}/s_{j_\star})\mathbf{s}_C$, and hence $\mathrm{Null}(A_C) = \mathrm{span}\{\mathbf{s}_C\}$. The same conclusion holds for an isolated vertex.

Now suppose that $\mathcal{C}$ is non-bipartite. It then contains an odd cycle $q_0, q_1, \ldots, q_{2m}, q_0$. Repeated use of (29) gives

$$h_{q_0} = -h_{q_1} = h_{q_2} = \cdots = h_{q_{2m}} = -h_{q_0},$$

so $h_{q_0} = 0$. Connectivity then forces every entry of $\mathbf{h}_C$ to be zero. Thus a connected non-bipartite component has trivial nullspace.

By construction, $\mathcal{C}_1, \ldots, \mathcal{C}_{b(\mathcal{G})}$ are exactly the bipartite connected components. Their alternating vectors have disjoint supports, so they are linearly independent and

$$\mathrm{Null}(A_{\mathcal{G}}) = \mathrm{span}\{\mathbf{s}_1, \ldots, \mathbf{s}_{b(\mathcal{G})}\}, \qquad \dim \mathrm{Null}(A_{\mathcal{G}}) = b(\mathcal{G}).$$

Rank-nullity gives $\mathrm{rank}(A_{\mathcal{G}}) = J - b(\mathcal{G})$. Finally, if $\boldsymbol{\gamma}^\star$ is one solution of $A_{\mathcal{G}}\boldsymbol{\gamma} = \boldsymbol{\ell}$, the complete solution set is $\boldsymbol{\gamma}^\star + \mathrm{Null}(A_{\mathcal{G}})$, proving (12).

C.3.2 Proof of Corollary C.1

If $\mathcal{C}$ contains an odd cycle, it is non-bipartite, so Theorem 3.1 gives $\mathrm{Null}(A_C) = \{\mathbf{0}\}$. Hence $(\gamma_j)_{j\in\mathcal{V}_C}$ and the positive factors $(u_j)_{j\in\mathcal{V}_C}$ are unique.

If $\mathcal{C}$ is bipartite with $\mathcal{V}_C = \mathcal{A}_C \,\dot\cup\, \mathcal{B}_C$, all log-factor solutions have the form

$$\gamma'_j = \begin{cases} \gamma_j + t, & j \in \mathcal{A}_C, \\ \gamma_j - t, & j \in \mathcal{B}_C, \end{cases} \qquad t \in \mathbb{R}.$$

Exponentiating and setting $c = e^t$ gives the stated rescaling. Every edge crosses the two classes, so all edge products are unchanged.

For the anchor statement, $a_{j_0} = 1$ and identified u_{j_0} imply $u_{j_0} = \sqrt{v}$, hence $v = u_{j_0}^2$ and $a_j = u_j/u_{j_0}$ for every identified u_j.

C.3.3 Proof of Corollary C.2

Let $(v, \mathbf{a})$ and $(v', \mathbf{a}')$ be two elements of $\mathsf{ID}_{j_0}(\mathcal{G}, \mathbf{M}_\mathcal{E})$ that induce different latent scales or decision probabilities. By definition of the identified set, the two parameter values produce the same edge moments M_{jk} for every $(j,k) \in \mathcal{E}$. A quantity that differs between these two parameter values therefore cannot be a function of the observed edge moments alone. Any procedure returning a unique value must use information not contained in those moments.

C.3.4 Proof of Lemma C.3

For brevity, write

$$B = B_{\mathcal{G},j_0}.$$

Because B has full column rank,

$$\boldsymbol{\ell} = B\boldsymbol{\eta}$$

and

$$\begin{aligned} \widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} &= (B^\top B)^{-1} B^\top \left(\widehat{\boldsymbol{\ell}} - \boldsymbol{\ell}\right) \\ &= B^+ \boldsymbol{\xi}, \end{aligned} \tag{30}$$

where

$$B^+ := (B^\top B)^{-1} B^\top$$

is the Moore-Penrose left inverse. Its Euclidean operator norm is

$$\|B^+\|_{\mathrm{op}} = \frac{1}{\sigma_{\min}(B)}.$$

Therefore

$$\|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \le \|B^+\|_{\mathrm{op}} \|\boldsymbol{\xi}\|_2 = \frac{\|\boldsymbol{\xi}\|_2}{\sigma_{\min}(B)}.$$

Suppose instead that $\widehat{\ell}_{jk} = \log \widehat{M}_{jk}$ for positive moment estimates $\widehat{M}_{jk}$. A Taylor expansion around $M_{jk} > 0$ gives

$$\log \widehat{M}_{jk} - \log M_{jk} = \frac{\widehat{M}_{jk} - M_{jk}}{M_{jk}} + R_{jk},$$

where R_{jk} is second order in the relative moment error. Applying the preceding argument to the linear term yields the stated first-order bound, with the transformed remainder propagated through B^+.

C.3.5 PROOF OF THEOREM 3.2

Write $B = B_{\mathcal{G},j_0}$ and $F = B^\top \mathbf{\Omega}^{-1} B$. The observation model assumed for this result is

$$\widehat{\boldsymbol{\ell}} \mid \boldsymbol{\eta} \sim \mathcal{N}\left(B\boldsymbol{\eta}, \frac{\mathbf{\Omega}}{N}\right), \qquad \mathbf{\Omega} \succ 0, \tag{31}$$

where $\mathbf{\Omega}$ is known and independent of $\boldsymbol{\eta}$. This experiment is assumed explicitly, rather than inferred solely from a central limit theorem for an empirical moment estimator.

Up to a constant independent of $\boldsymbol{\eta}$, the log likelihood is

$$\mathcal{L}_N(\boldsymbol{\eta}) = -\frac{N}{2}(\widehat{\boldsymbol{\ell}} - B\boldsymbol{\eta})^\top \mathbf{\Omega}^{-1}(\widehat{\boldsymbol{\ell}} - B\boldsymbol{\eta}).$$

Its score is

$$\mathbf{S}_N = NB^\top \mathbf{\Omega}^{-1}(\widehat{\boldsymbol{\ell}} - B\boldsymbol{\eta}).$$

Consequently,

$$\mathbb{E}_{\boldsymbol{\eta}}[\mathbf{S}_N] = \mathbf{0}, \qquad \mathrm{Cov}_{\boldsymbol{\eta}}(\mathbf{S}_N) = NF.$$

Suppose first that B has full column rank. For an unbiased estimator $\widetilde{\boldsymbol{\eta}}$ satisfying the stated regularity conditions, differentiating $\mathbb{E}_{\boldsymbol{\eta}}[\widetilde{\boldsymbol{\eta}}] = \boldsymbol{\eta}$ gives

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\widetilde{\boldsymbol{\eta}}, \mathbf{S}_N) = I_J.$$

It follows that

$$\begin{aligned} 0 &\preceq \mathrm{Cov}_{\boldsymbol{\eta}}\big(\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta} - (NF)^{-1}\mathbf{S}_N\big) \\ &= \mathrm{Cov}_{\boldsymbol{\eta}}(\widetilde{\boldsymbol{\eta}}) - (NF)^{-1}. \end{aligned}$$

Taking the quadratic form in a unit vector $\mathbf{d}$ gives

$$\mathrm{Var}_{\boldsymbol{\eta}}(\mathbf{d}^\top \widetilde{\boldsymbol{\eta}}) \geq \frac{1}{N}\mathbf{d}^\top F^{-1}\mathbf{d}.$$

The generalized least-squares estimator is

$$\widehat{\boldsymbol{\eta}}_{\mathrm{GLS}} = F^{-1}B^\top \mathbf{\Omega}^{-1}\widehat{\boldsymbol{\ell}}.$$

It is unbiased and has covariance

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\widehat{\boldsymbol{\eta}}_{\mathrm{GLS}}) = \frac{1}{N}F^{-1},$$

so it attains the bound. Moreover,

$$\sup_{\|\mathbf{d}\|_2=1} \frac{1}{N}\mathbf{d}^\top F^{-1}\mathbf{d} = \frac{1}{N\lambda_{\min}(F)}.$$

If B is rank deficient, choose $\mathbf{0} \neq \mathbf{h} \in \mathrm{Null}(B)$. For every $t \in \mathbb{R}$,

$$B(\boldsymbol{\eta} + t\mathbf{h}) = B\boldsymbol{\eta}.$$

The mean and covariance in (31) are therefore identical at $\boldsymbol{\eta}$ and $\boldsymbol{\eta} + t\mathbf{h}$. Hence these null directions are unidentifiable in this experiment.

Relation to an empirical moment central limit theorem. If the empirical log moments instead satisfy

$$\sqrt{N}(\widehat{\boldsymbol{\ell}} - B\boldsymbol{\eta}) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{\Omega}),$$

then the fixed-weight generalized least-squares estimator above satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\eta}}_{\mathrm{GLS}} - \boldsymbol{\eta}) \Rightarrow \mathcal{N}(\mathbf{0}, F^{-1}).$$

This follows by applying the linear map $F^{-1}B^\top \mathbf{\Omega}^{-1}$ to the moment limit. It establishes the limiting distribution of this estimator, not an information lower bound for every estimator based on the original observations.

C.3.6 Proof of Corollary C.4

Because

$$B_{\mathcal{G},j_0}\boldsymbol{\eta}_1 = B_{\mathcal{G},j_0}\boldsymbol{\eta}_2,$$

the two parameter values induce the same log-edge moment vector $\boldsymbol{\ell}$. Any target-label-free procedure whose scale information is restricted to $\boldsymbol{\ell}$ must therefore have the same population input at $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. Since $\Psi(\boldsymbol{\eta}_1) \neq \Psi(\boldsymbol{\eta}_2)$, no single output can equal the true functional value at both parameter points. Thus Ψ is not identified by the observed edge moments.

C.3.7 Proof of Proposition C.5

Under the rank-one view-loading moment model,

$$\boldsymbol{\ell} = A_{\mathcal{G}}\boldsymbol{\gamma}.$$

Because the columns of $L_{\mathcal{G}}$ lie in $\mathrm{Null}(A_{\mathcal{G}}^{\top})$,

$$L_{\mathcal{G}}^{\top}A_{\mathcal{G}} = 0.$$

It follows immediately that

$$L_{\mathcal{G}}^{\top}\boldsymbol{\ell} = L_{\mathcal{G}}^{\top}A_{\mathcal{G}}\boldsymbol{\gamma} = 0.$$

For the four-cycle ordered as $(1,2),(2,3),(3,4),(4,1)$, the vector

$$(1,-1,1,-1)^{\top}$$

lies in the left nullspace of $A_{\mathcal{G}}$ because each vertex appears once with coefficient $+1$ and once with coefficient -1. Multiplication by the ordered log-edge vector gives

$$\log M_{12} - \log M_{23} + \log M_{34} - \log M_{41} = 0,$$

which is equivalent to (25).

C.3.8 Proof of Proposition C.6

Write $B = B_{\mathcal{G},j_0}$. Appending candidate edge e adds the row $\mathbf{z}_e^{\top}$, so

$$B_{\mathrm{new}} = \begin{pmatrix} B \\ \mathbf{z}_e^{\top} \end{pmatrix}, \qquad W_{\mathrm{new}} = \begin{pmatrix} W & 0 \\ 0 & w_e \end{pmatrix}.$$

Therefore

$$B_{\mathrm{new}}^{\top}W_{\mathrm{new}}B_{\mathrm{new}} = B^{\top}WB + w_e\mathbf{z}_e\mathbf{z}_e^{\top},$$

which proves (27). Appending one row increases rank exactly when that row is outside the existing row space.

For the odd-cycle claim, let $\mathcal{C}$ be a connected bipartite component with $\mathcal{V}_C = \mathcal{A}_C \,\dot\cup\, \mathcal{B}_C$, and let $\mathbf{s}_C$ be its alternating null vector. In the signless-incidence parameterization, the candidate edge $e = (j,k)$ adds the row $(\mathbf{e}_j + \mathbf{e}_k)^{\top}$. It removes the component's null direction exactly when

$$(\mathbf{e}_j + \mathbf{e}_k)^{\top}\mathbf{s}_C \neq 0.$$

This inner product is zero for opposite-side endpoints and equals 2 or -2 for same-side endpoints. Thus the edge increases rank exactly in the latter case. Every path between same-side vertices in a connected bipartite graph has even length, so adding the edge creates an odd cycle. Because the map between $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ is invertible, the same rank conclusion holds for the anchored edge matrix $B_{\mathcal{G},j_0}$.

Finally, let $\mathbf{d}_{\min}$ be the unit eigenvector of the simple smallest eigenvalue of $F_{\mathcal{G},j_0}(W)$. Standard first-order perturbation gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\epsilon}\lambda_{\min}\big(F_{\mathcal{G},j_0}(W) + \epsilon w_e\mathbf{z}_e\mathbf{z}_e^{\top}\big)\Big|_{\epsilon=0} &= \mathbf{d}_{\min}^{\top}(w_e\mathbf{z}_e\mathbf{z}_e^{\top})\mathbf{d}_{\min} \\ &= w_e(\mathbf{d}_{\min}^{\top}\mathbf{z}_e)^2.\end{aligned}$$

This proves the proposition.

D SYNTHETIC EXPERIMENTS: DESIGN AND ADDITIONAL RESULTS

D.1 SYNTHETIC BENCHMARK DESIGN

The synthetic experiments follow the failure modes isolated by the theory. The graph benchmark tests whether rank deficiency and weak conditioning predict scale-recovery and uncertainty failures. Component and oracle ablations test whether ACEB's behavior comes from graph-scale estimation and leave-view residual-noise estimation rather than generic interval inflation. Same-coverage recalibration gives the included methods the same scalar variance-rescaling opportunity. Paired graph interventions change only the view-overlap design to separate structural identifiability from random simulation variation. Dense homogeneous negative controls, anchor diagnostics, misspecification stress tests, and structural edge-selection experiments delimit the intended operating regime.

We generate observations from the multi-view measurement model with known latent scores, view loadings, and residual noises. The co-observation graph is controlled directly. Deficient designs include a star, chain, even cycle, and disconnected graph. Identified designs include a star plus one leaf edge, a triangle plus tails, an odd cycle, and a complete graph. Invalid cycle designs, such as an even cycle with fewer than four views, are skipped rather than silently replaced by a different topology.

The main graph benchmark uses

$$n \in \{1000, 5000\}, \qquad J \in \{5, 10\}, \qquad \tau/s \in \{0, 1, 2\},$$

loading spread in $\{0.3, 1.0\}$, and random seeds 0-4. This produces 840 valid benchmark cells and no failed cells. Results are summarized by the mean and standard error across valid cells or seeds.

The comparison methods are:

- raw reference-view scoring;
- inverse-variance averaging;
- an adaptive-shrinkage-like baseline;
- random-effects meta-analysis;
- joint empirical Bayes;
- post-hoc inverse-variance calibration;
- ACEB with in-sample residual moments; and
- ACEB with leave-view residual moments.

Baseline implementations. Inverse variance uses per-unit reciprocal-variance weighting. Random effects uses a per-unit DerSimonian-Laird-style heterogeneity estimate. The ASH-like baseline first collapses the views by inverse variance and then fits a zero-centered finite normal mixture by EM. Joint EB jointly optimizes the prior, positive non-reference loadings, and residual noises with the reference loading fixed to one. Post-hoc inverse variance multiplies the inverse-variance posterior variance by one learned scalar; it uses evaluation targets in the main synthetic and MedHELM runs, and is therefore a target-calibrated reference there. It can restore marginal coverage through target-based scale inflation, but does not estimate the view-loading structure or diagnose graph identifiability.

D.2 STRUCTURAL VERIFICATION AND OVERALL PERFORMANCE

D.2.1 MAIN GRAPH BENCHMARK TABLE

Figure 7 and Figure 8 illustrate why the comparison is not organized by RMSE alone. Joint EB can estimate latent means accurately while badly underestimating uncertainty. Numerically, ACEB leave-view reaches 0.849 coverage and a false-best rate of 0.006; joint EB attains lower RMSE, 0.347, but only 0.510 coverage and a false-best rate of 0.092. Post-hoc inverse variance reaches 0.908 coverage and a false-best rate of 0.030, but does not estimate loadings or diagnose the graph. ACEB is not best on every scalar metric; its purpose is graph-aware uncertainty and false-best control under the multi-view graph model.

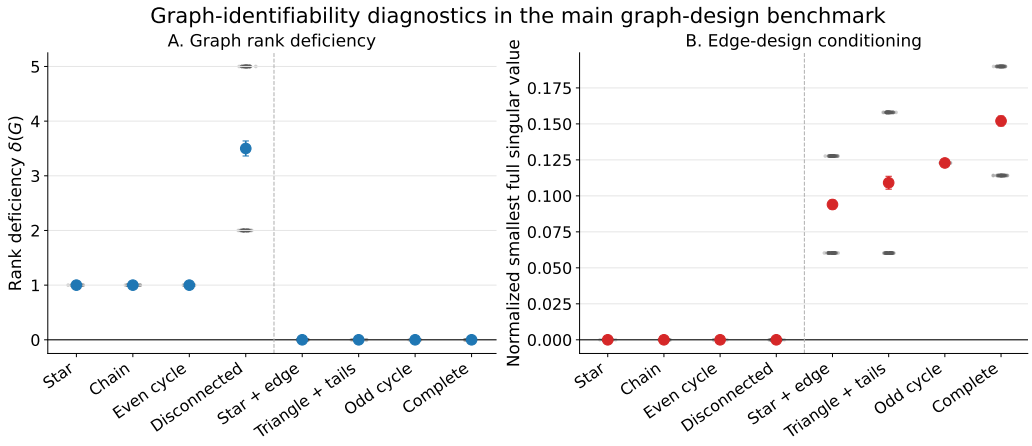


Figure 6: Graph identifiability diagnostics for the synthetic benchmark. Panel A reports the rank deficiency $\delta(G)$ of the edge-design system; Panel B reports its normalized smallest full-column singular value, for which larger positive values indicate stronger conditioning. Star, chain, even-cycle, and disconnected designs retain unidentified loading directions and therefore have $\delta(G) > 0$ (A) and singular value zero (B). Star+edge and triangle+tails designs remove this deficiency by closing an odd cycle; odd-cycle and complete designs are likewise full rank. Colored circles summarize the configured view counts, and gray marks show the constituent configurations. The dashed line separates deficient from identified graph designs.

Table 3: Overall performance on the main graph-design benchmark (mean $\pm$ standard error). Panel A reports estimation and interval metrics; Panel B reports decision calibration and runtime. No method dominates every metric.

A. Estimation and interval metrics

Method	Coverage 90%	z-score SD	RMSE	Interval score
Raw reference	0.588 ± 0.006	4.244 ± 0.179	0.668 ± 0.025	5.251 ± 0.254
Inverse variance	0.507 ± 0.005	5.206 ± 0.200	0.741 ± 0.028	6.402 ± 0.288
ASH-like	0.508 ± 0.005	5.194 ± 0.200	0.735 ± 0.028	6.343 ± 0.287
Random effects	0.781 ± 0.003	1.710 ± 0.024	0.706 ± 0.025	2.844 ± 0.081
Joint EB	0.510 ± 0.008	493.951 ± 444.185	0.347 ± 0.005	2.988 ± 0.067
ACEB in-sample	0.666 ± 0.006	5.001 ± 0.925	0.636 ± 0.020	4.743 ± 0.231
Post-hoc IV	0.908 ± 0.000	1.006 ± 0.002	0.741 ± 0.028	3.777 ± 0.159
ACEB leave-view	0.849 ± 0.005	1.590 ± 0.077	0.601 ± 0.019	3.574 ± 0.168

B. Decision calibration and computation

Method	False-best	Close-pair ECE	Pairwise Brier	Runtime (s)
Raw reference	0.235 ± 0.015	0.312 ± 0.002	0.086 ± 0.001	0.077 ± 0.002
Inverse variance	0.318 ± 0.016	0.291 ± 0.002	0.080 ± 0.001	0.007 ± 0.000
ASH-like	0.319 ± 0.016	0.292 ± 0.002	0.080 ± 0.001	0.050 ± 0.001
Random effects	0.000 ± 0.000	0.325 ± 0.002	0.073 ± 0.001	0.167 ± 0.005
Joint EB	0.092 ± 0.010	0.265 ± 0.002	0.064 ± 0.001	53.71 ± 1.33
ACEB in-sample	0.042 ± 0.007	0.279 ± 0.002	0.062 ± 0.001	26.19 ± 0.667
Post-hoc IV	0.030 ± 0.006	0.365 ± 0.002	0.080 ± 0.001	0.024 ± 0.001
ACEB leave-view	0.006 ± 0.003	0.321 ± 0.002	0.065 ± 0.001	26.05 ± 0.657

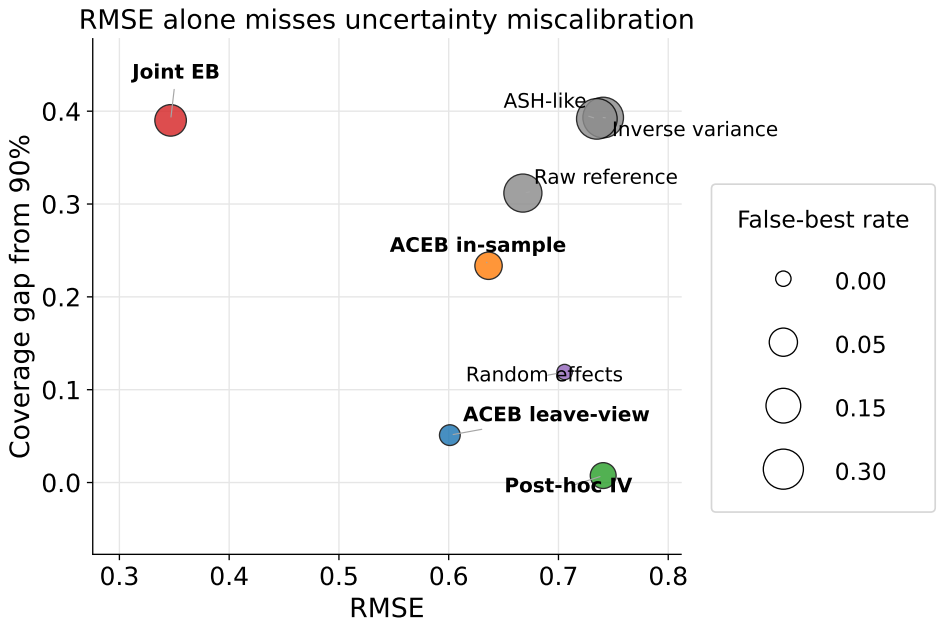


Figure 7: Accuracy-calibration tradeoff in the main synthetic graph-design benchmark. RMSE (x-axis) is plotted against the absolute deviation of 90% coverage from its nominal value (y-axis); lower is better on both axes. Circle size represents the false-best rate. Joint EB attains the lowest RMSE but is severely undercovered. Among target-free methods, ACEB leave-view has the smallest coverage gap and a low false-best rate; post-hoc IV is closest to nominal coverage but uses target labels for calibration.

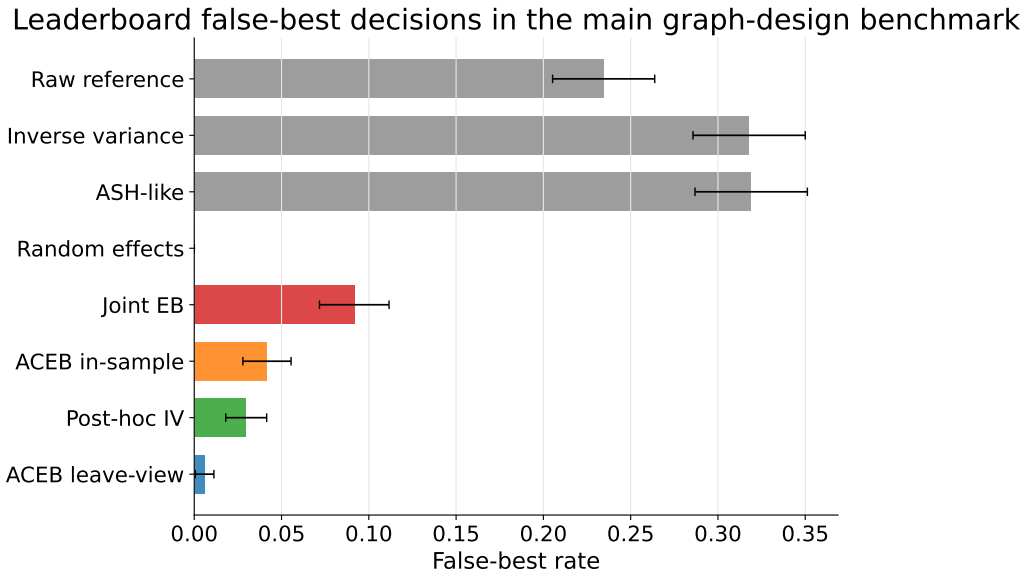


Figure 8: False-best decisions in the main synthetic graph-design benchmark. Lower is better when methods name a winner at comparable rates; bars show means across benchmark cells and whiskers show ± 2 standard errors. ACEB leave-view has the lowest nonzero false-best rate. Random effects is zero because it never assigns any model more than 95% probability of being best, not because it always selects the correct model. Because winner-selection rates differ across methods, this is a descriptive diagnostic rather than an equal-decision comparison. Post-hoc IV uses target labels for calibration.

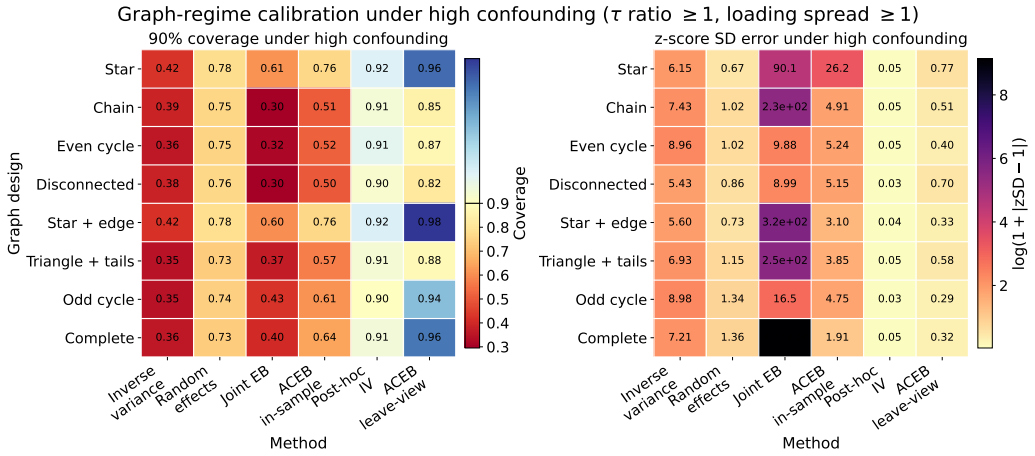


Figure 9: High-confounding stress test (τ ratio ≥ 1 and loading spread ≥ 1). Cells report mean 90% coverage (left; closer to 0.90 is better) and absolute z-score SD error (right; lower is better); the right-hand color scale uses $\log(1 + x)$, while annotations show the untransformed values. Across graph designs, ACEB leave-view is substantially better calibrated than joint EB and in-sample ACEB, although its coverage remains graph-dependent and is not uniformly nominal. Post-hoc IV is closest to nominal but uses target labels for calibration and does not estimate the graph-loading structure.

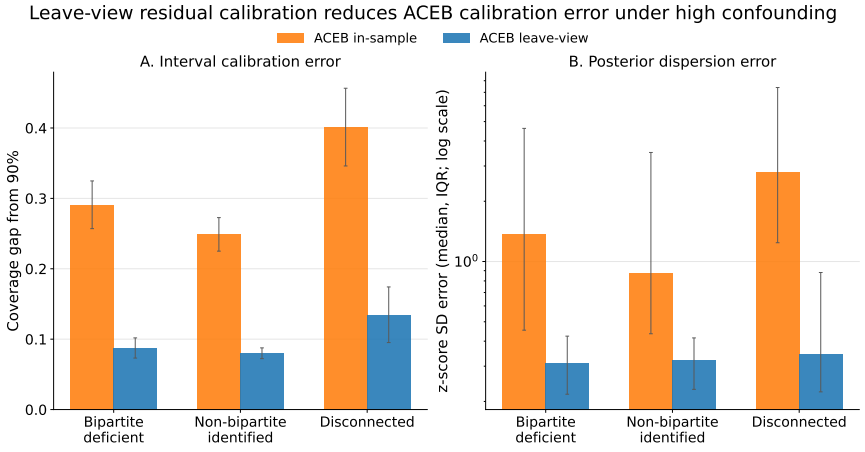


Figure 10: Leave-view residual calibration under high confounding. Panel A shows the mean absolute gap from 90% coverage with ± 2 SEs; Panel B shows the median absolute deviation of the z-score SD from one, with interquartile ranges on a log scale. Lower is better in both panels. Leave-view ACEB reduces both errors across all graph classes, with the largest gain in disconnected designs, indicating that reuse of in-sample residuals produces overly optimistic uncertainty.

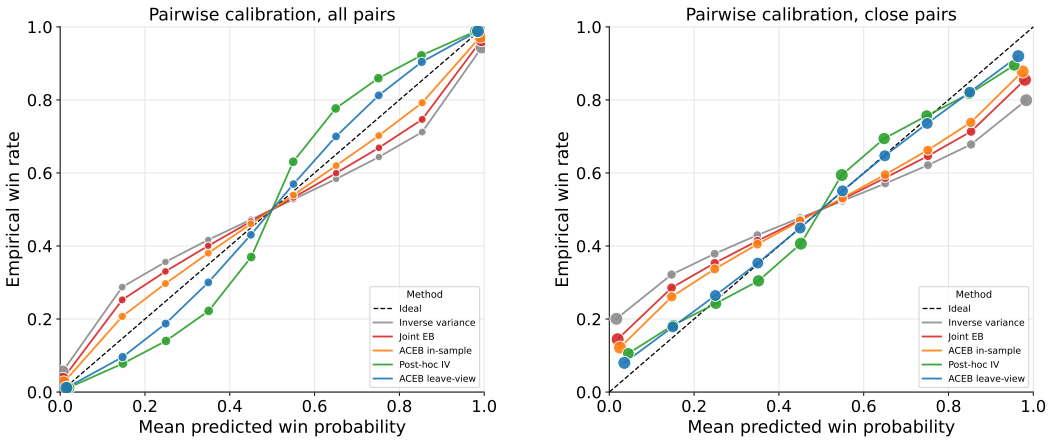


Figure 11: Pairwise calibration curves for all pairs (left) and close pairs (right). Each curve compares binned posterior win probabilities with empirical win frequencies; closer agreement with the dashed diagonal indicates better calibration. All-pair calibration can be dominated by easy comparisons, whereas close-pair calibration is noisier but more relevant to leaderboard ties. ACEB leave-view tracks the ideal line closely, particularly for close pairs, while the close-pair curves are necessarily noisier because fewer comparisons are available. Post-hoc IV uses target labels for calibration.

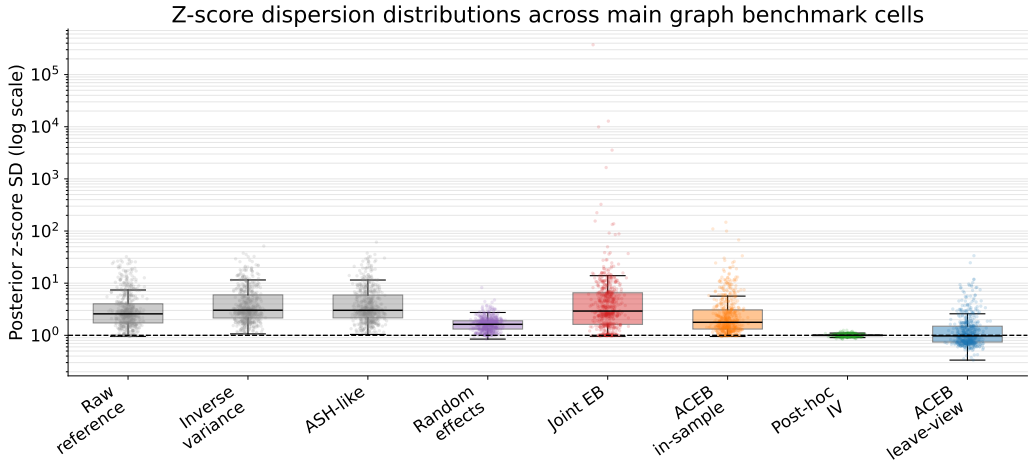


Figure 12: Distribution of posterior z-score SD across synthetic graph-benchmark cells. Ideal value of 1 is marked by the dashed line (values above indicate underestimated posterior uncertainty; values below indicate conservative uncertainty). Boxes show the interquartile range and median, and jittered points show individual benchmark cells. No values are clipped. Joint EB can suffer extreme z-score dispersion even when its RMSE is small. ACEB leave-view is concentrated near the ideal value among the target-free methods; post-hoc IV uses target labels for calibration.

D.3 Parameter recovery diagnostics

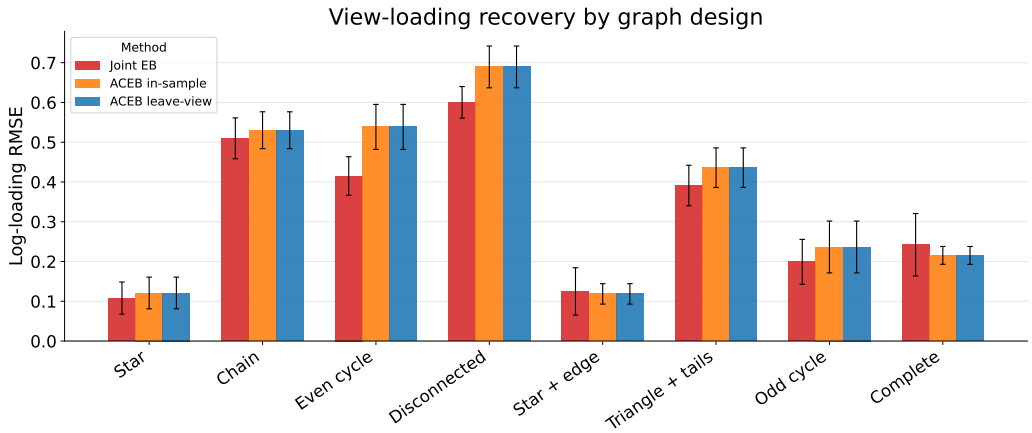


Figure 13: View-loading recovery by graph design. Bars show mean RMSE between the estimated and true anchor-normalized log loadings with ± 2 SEs; lower is better. Log-loading RMSE measures multiplicative recovery without allowing raw-scale explosions under confounding to dominate the comparison. Recovery remains topology-dependent: disconnected designs are the most difficult, while star and star+edge designs have the lowest errors. The two ACEB variants are nearly identical because leave-view residual calibration changes residual-noise and posterior uncertainty estimates, not the fitted loadings. Full-rank identifiability therefore does not imply uniform finite-sample loading precision.

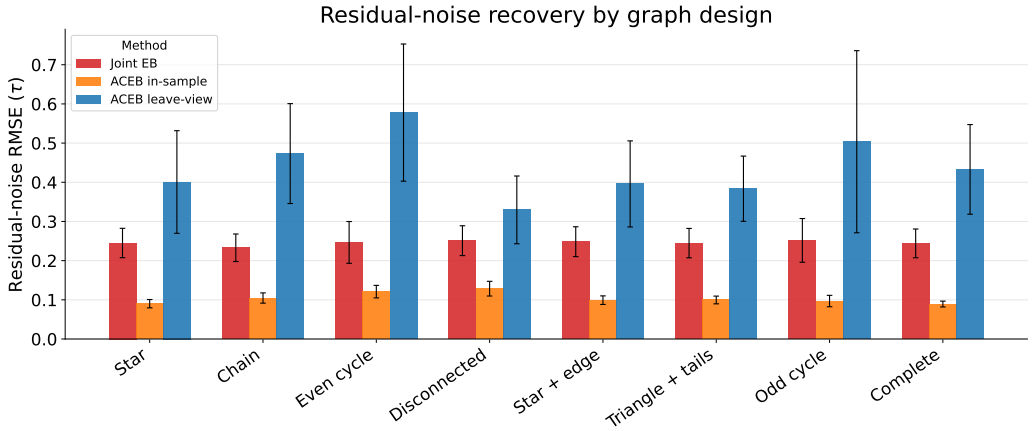


Figure 14: Residual-noise recovery by graph design. Bars show mean RMSE between estimated and true per-view residual noise τ_j with ± 2 SEs; lower is better. In-sample ACEB recovers the generative τ_j most closely, whereas leave-view ACEB estimates a larger residual variance and has greater τ-RMSE. Despite this, leave-view ACEB gives better interval calibration in the synthetic benchmark. Thus, exact recovery of the generative noise parameter and calibrated plug-in posterior uncertainty are distinct empirical objectives.

D.4 Component and oracle ablations

The main benchmark establishes the calibration behavior of ACEB, but does not by itself determine whether that behavior arises from the proposed components or from generic interval widening. We therefore run a compact component and oracle ablation on high-confounding graph cells. Variants fix all view loadings to one, force a single global residual noise, set residual noise to zero, replace leave-view residual moments with in-sample moments, or use oracle loadings and/or oracle residual

noises from the data-generating process. The ablation uses $n = 1000$, $J \in \{5, 10\}$, six graph designs, $\tau/s \in \{1, 2\}$, and seeds 0-4.

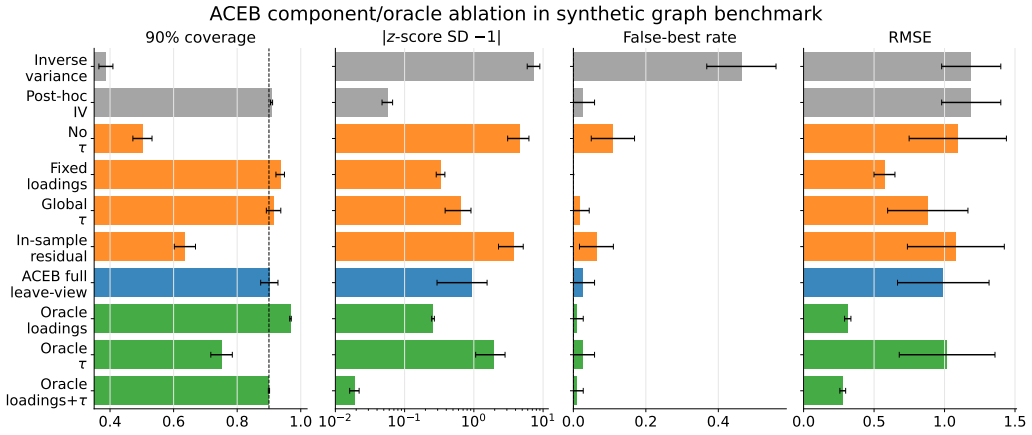


Figure 15: ACEB component and oracle ablation. From left to right, the panels report 90% coverage, absolute z-score SD error on a log scale, false-best rate, and RMSE. Lower is better for the latter three metrics, with false-best comparisons requiring similar declaration rates. Forcing residual noise to zero or using in-sample residuals worsens calibration in these settings. Oracle loadings primarily improve RMSE, while oracle loadings together with oracle noise approach nominal calibration. Bars show means across ablation cells; error bars show ± 2 SE.

The ablation supports three conclusions in these settings. First, leave-view residual calibration reduces overconfidence relative to the in-sample variant. Second, view-loading estimation matters: fixed-loading variants can restore coverage in some cells but do not recover the graph-scale mechanism. Third, oracle residual noise alone does not close the gap when loadings are poorly estimated, whereas the oracle-loading-plus-oracle-noise variant behaves as expected. The synthetic result therefore reflects an interaction between graph-aware loading estimation and predictive residual-noise estimation.

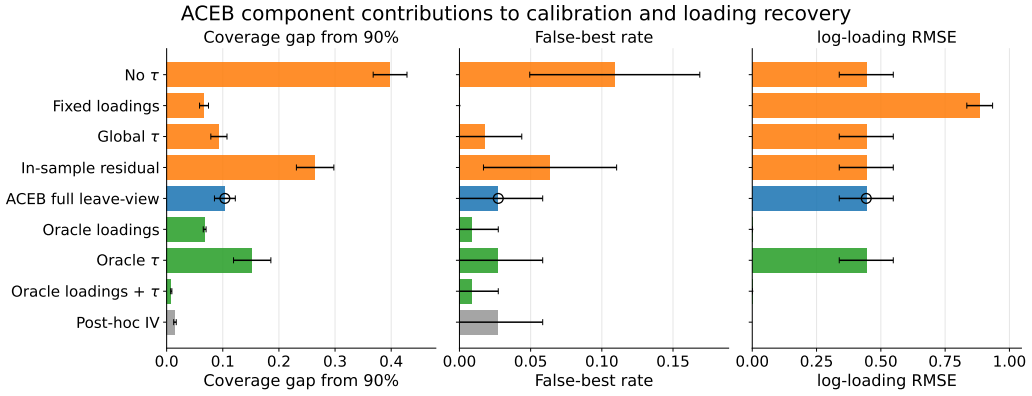


Figure 16: ACEB component contributions to calibration and loading recovery. The panels report absolute coverage gap from 90%, false-best rate, and log-loading RMSE (lower is better throughout). The full leave-view protocol reduces calibration gap and false-best rate relative to no-τ and in-sample variants. Combining oracle loadings with oracle residual noise nearly closes both gaps, showing that both components matter. Fixed loadings can appear marginally calibrated despite poor loading recovery. Oracle-loading variants have zero loading error by construction, while target-calibrated post-hoc IV does not estimate loadings. Bars show means across ablation cells; error bars show ± 2 SE.

D.5 SAME-COVERAGE RECALIBRATION

For method h, let $\widehat{\theta}_{i,h}$ and $\widehat{\sigma}_{i,h}$ denote its posterior mean and standard deviation for unit i. For a scalar inflation factor $c > 0$, define

$$C_{i,h}(c) = \left[\widehat{\theta}_{i,h} - z_{0.95}c\widehat{\sigma}_{i,h},\ \widehat{\theta}_{i,h} + z_{0.95}c\widehat{\sigma}_{i,h}\right]. \tag{32}$$

On calibration units, we choose the smallest $\widehat{c}_h$ that attains 90% marginal coverage and then evaluate the inflated intervals on held-out units. This equalizes the simplest form of conservatism: if a method's only advantage is wider intervals, that advantage should disappear after recalibration.

The comparison includes inverse variance, adaptive shrinkage, random effects, post-hoc IV, and both ACEB variants. After coverage alignment, ACEB retains favorable scoring metrics among these methods (Figure 3); recalibrated joint EB is not evaluated. ACEB uses a 100-iteration cap here, versus 8 in the main benchmark, so this is a separate fitted comparison rather than recalibration of the main benchmark's outputs.

D.6 NEGATIVE CONTROL: DENSE HOMOGENEOUS VIEWS

A useful falsification check is to evaluate ACEB in a regime where its graph-aware and residual-calibration machinery should not provide a special advantage. We construct a dense homogeneous benchmark in which every unit is observed under every view, all view loadings are equal, there is no unreported residual view noise, and reported standard errors are correctly specified:

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \mathcal{N}(0, s_{ij}^2), \qquad a_j = 1, \qquad \tau_j^2 = 0. \tag{33}$$

The co-observation graph is complete, so the measurement scale is structurally identified and there is no hidden view-specific residual variance to estimate. This setting is intentionally favorable to simple inverse-variance and shrinkage estimators. Its purpose is not to show ACEB dominance, but to test whether ACEB spuriously widens intervals or creates an artificial calibration advantage when the simpler model is correct.

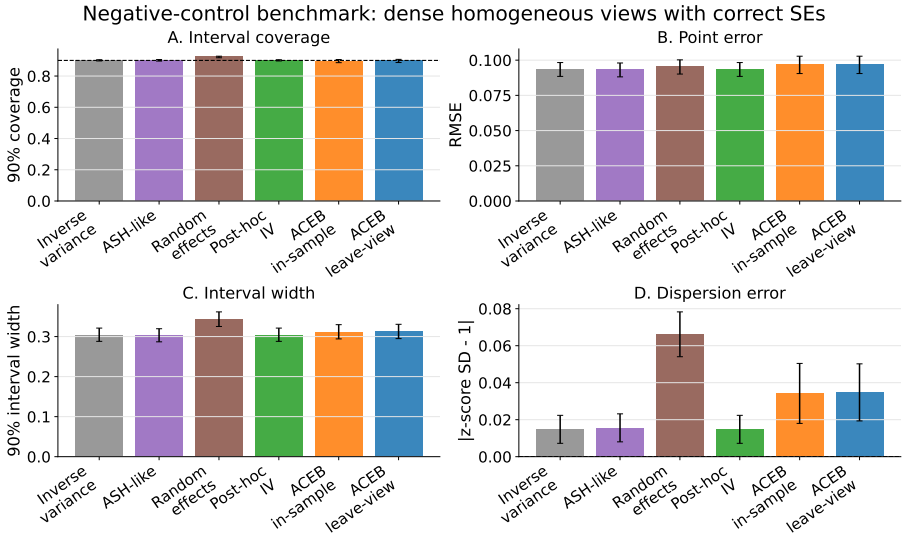


Figure 17: Negative-control benchmark with dense homogeneous views. All views have unit loadings, zero residual view noise and correctly specified SEs. Panel A reports 90% coverage (ideal 0.90). Panels B-D report RMSE, interval width, and absolute z-score SD error; lower is better, with width interpreted at comparable coverage. Bars show means with ± 2 SEs. ACEB behaves similarly to inverse variance and the shrinkage baselines, showing that it does not create a large artificial advantage or inflate uncertainty when the simple homogeneous model is correct. Post-hoc IV is target-calibrated.

This control clarifies the intended operating regime. When the benchmark is dense, homogeneous, and correctly reported, the graph and residual-noise diagnostics are less consequential and simpler estimators may be preferable on computational grounds.

D.7 ANCHOR SENSITIVITY

We run a separate anchor-sensitivity ablation by refitting under alternative reference views and by constructing weak and noisy defaults. These refits do not alter the anchor or decisions in the reported main experiments. Weak anchors substantially damage calibration, whereas normal and noisy anchors are more stable. Odd-cycle structure removes the population scale ridge but does not eliminate finite-sample dependence on the reference convention; we report that variation as estimator sensitivity rather than as a second decision rule.

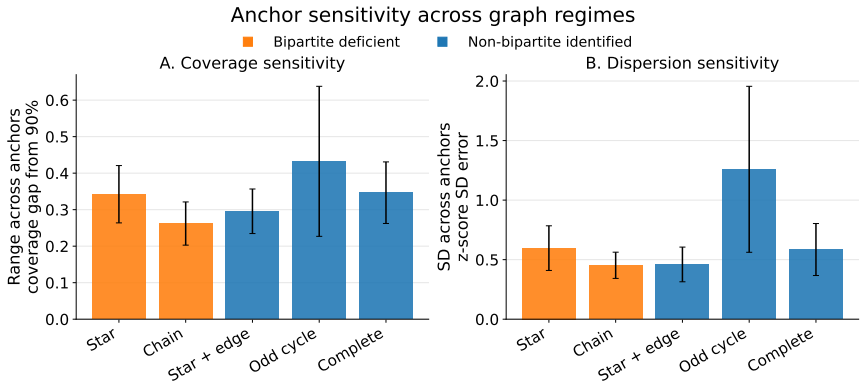


Figure 18: Anchor sensitivity across graph regimes. Each synthetic dataset is refitted using every view in turn as the anchor. Panel A reports the range across anchors of the absolute 90% coverage gap. Panel B reports the across-anchor SD of the z-score dispersion error. Lower values indicate greater anchor stability. Sensitivity persists in both deficient and identified graphs and is especially variable for odd-cycle designs. Thus, population identifiability does not eliminate finite-sample dependence on the chosen reference view, motivating explicit anchor-sensitivity checks. Bars show mean anchor-sensitivity statistics across benchmark cells; error bars show ± 2 SE.

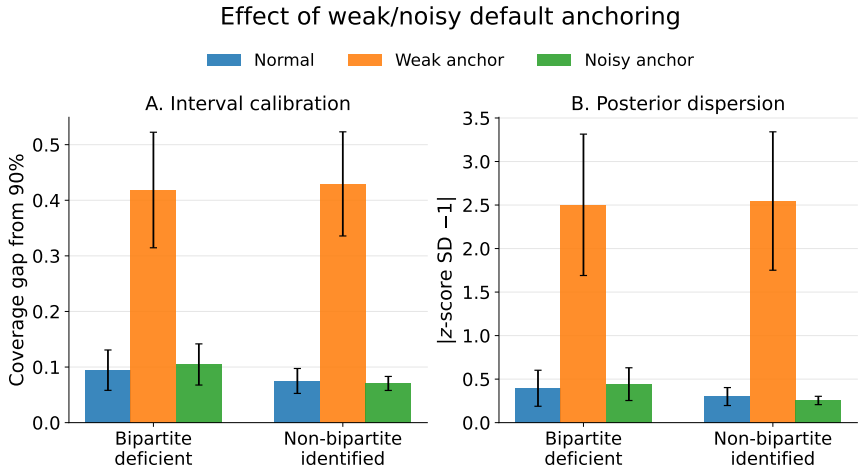


Figure 19: Effect of default-anchor quality. A weak anchor has a low true loading, whereas a noisy anchor has elevated residual noise. Panel A reports the absolute gap from 90% coverage. Panel B reports the absolute deviation of the z-score SD from one; lower is better in both panels. Bars show means with ± 2 SEs. Weak anchors substantially degrade calibration in both deficient and identified graphs, while noisy anchors remain close to normal anchoring in this simulation. Thus, graph identification alone does not protect against a poorly scaled reference view.

D.8 COMPACT MISSPECIFICATION ROBUSTNESS

We run a compact robustness grid with heavy-tailed residuals, underreported standard errors, correlated views, and bounded or binomial observations analyzed on the logit scale. The purpose is not to establish universal robustness, but to check that ACEB's behavior is not solely an artifact of the exact Gaussian data-generating model. ACEB leave-view remains competitive on calibration and false-best behavior under heavy-tailed and underreported-standard-error stress, while remaining conservative and sometimes less sharp than post-hoc inverse variance or random effects.

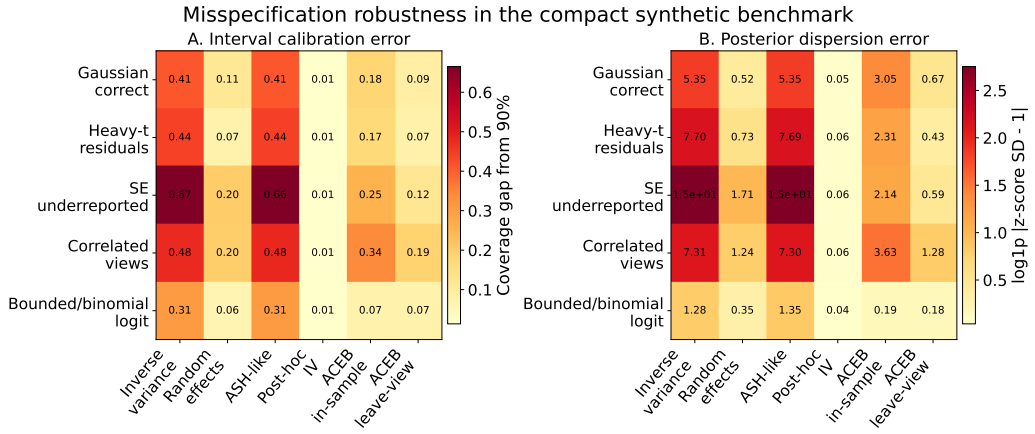


Figure 20: Robustness to model misspecification in the compact synthetic benchmark. Rows denote the data-generating stressor and columns the estimation method. Panel A shows the absolute gap from 90% coverage. Panel B shows the absolute deviation of the z-score SD from one; lower is better in both panels. Panel B uses a $\log(1 + x)$ color scale, while cell annotations report untransformed errors. Leave-view ACEB consistently improves on its in-sample ablation and substantially reduces the failures of inverse variance and ASH-like methods, although correlated views remain challenging. Post-hoc IV uses target labels and serves as a target-calibrated reference.

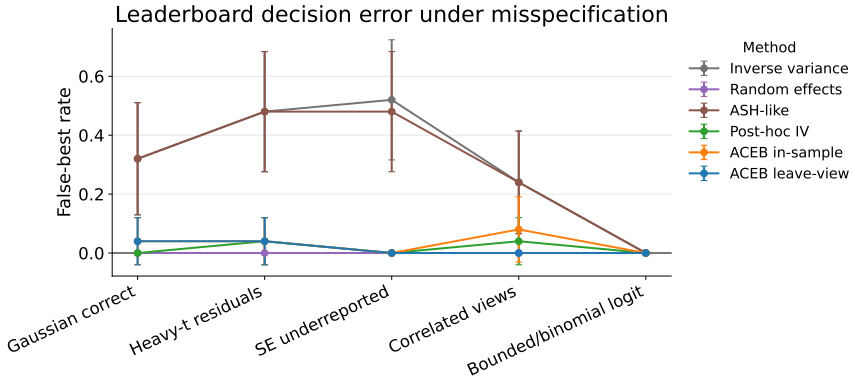


Figure 21: False-best diagnostics across misspecification modes. Points show mean rates across graph and seed cells with ± 2 SEs. Inverse variance and ASH-like are highly sensitive to heavy-tailed residuals and underreported SEs, whereas leave-view ACEB remains near zero across stressors. Lower is better only when methods select winners at comparable rates: no-declaration cases are encoded as zero, so this is not a complete selective-risk comparison. Post-hoc IV is target-calibrated.

D.9 STRUCTURAL EDGE-SELECTION ABLATION

The graph theorem can also be read as a design rule: select new co-observation edges that reduce rank deficiency or improve the full-column singular value of the anchored edge matrix. We enumerate

candidate missing edges in star, chain, and disconnected base graphs, classify candidates as best, median, or worst by predicted structural gain, and add the selected edge in paired simulations. Structural gain is a useful design heuristic, especially in chains, but does not deterministically imply improvement in every downstream metric.

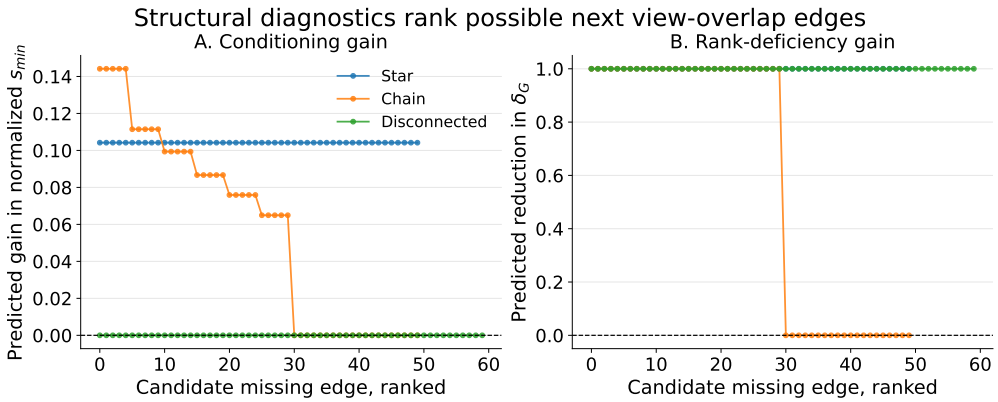


Figure 22: Structural ranking of candidate view-overlap edges. Each point represents one missing edge, ordered from largest to smallest predicted gain in the normalized smallest full-column singular value. Panel A reports conditioning gain. Panel B reports the reduction in rank deficiency δ_G; higher is better in both panels. All star repairs are equivalent as each closes a triangle, whereas chain repairs form a clear hierarchy. A single disconnected-graph repair reduces δ_G by one but leaves another unidentified direction, so the smallest full-column singular value remains zero.

D.10 Paired graph interventions

To separate structural identification from random simulation variation, each intervention pair shares the same latent scores, view loadings, residual noises, and base observations. The repaired design either adds the observations needed for the structural repair without removing existing observations, referred to as an additive repair, or reallocates a fixed observation budget to realize the same repair, referred to as a budget-matched repair. We study star-to-triangle, chain-to-triangle-plus-tail, and disconnected-to-connected-with-odd-cycle repairs over seeds 0-9.

The structural result is unambiguous: every repair improves rank deficiency and the full-column singular value. Downstream results are more nuanced, as expected. Additive repair improves log-loading recovery, RMSE, close-pair expected calibration error, and pairwise Brier score for ACEB leave-view on average. Budget-matched repair improves several recovery metrics but can worsen coverage-gap or dispersion metrics because the repair competes with the fixed measurement budget. The theorem determines whether scale is identifiable; it does not imply that every finite-sample downstream metric improves uniformly.

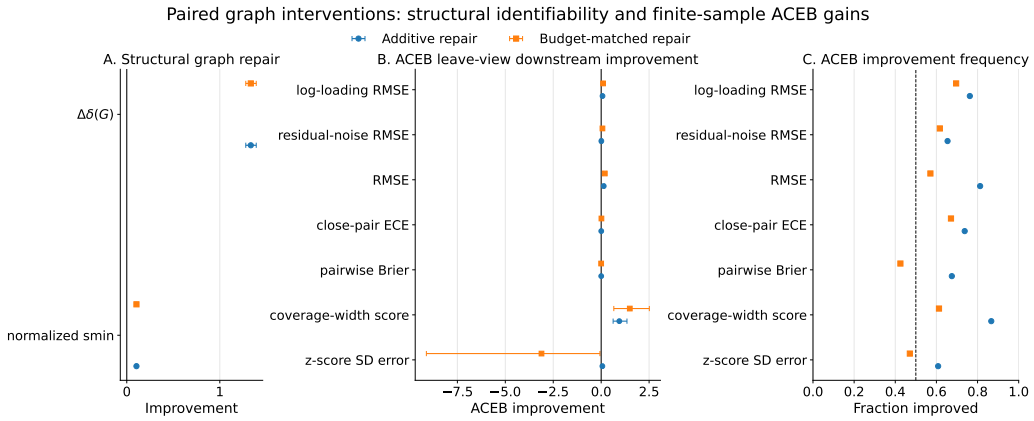


Figure 23: Paired graph-repair interventions with the latent data-generating quantities held fixed. Panel A reports the reduction in rank deficiency $\Delta\delta(G)$ and the gain in normalized smallest full-column singular value (positive values indicate improved identifiability and conditioning). Panel B shows mean paired changes in ACEB leave-view metrics, oriented so that positive values indicate improvement. Points in A-B show mean paired changes across benchmark cells with ± 2 SE. Panel C reports the fraction of paired cells improved. Additive repair gives the cleanest gains. Budget-matched repair is more variable as topology changes without an increase in the total measurement budget.

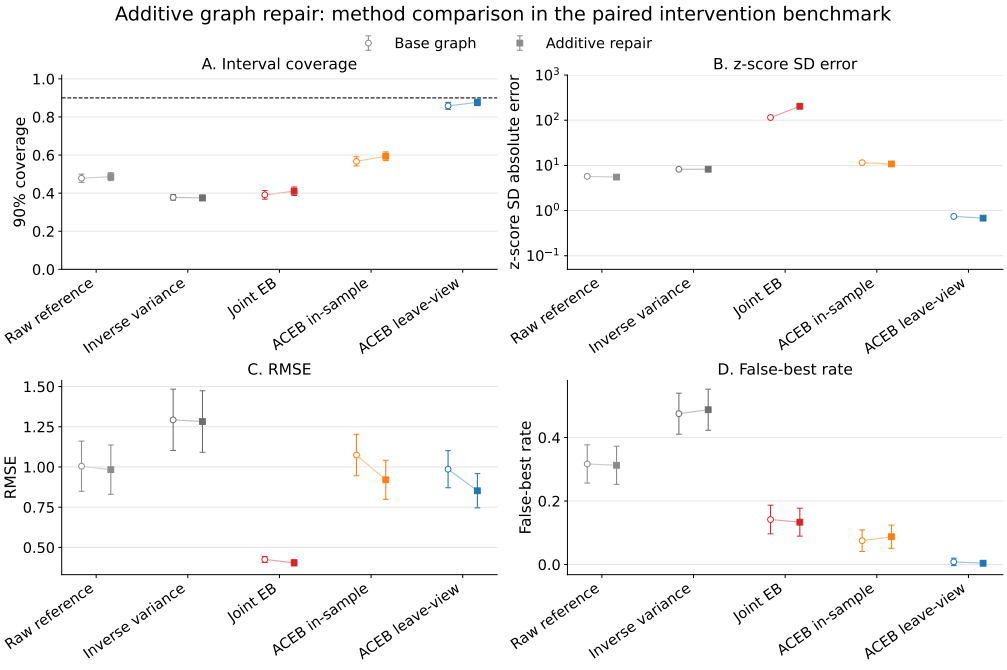


Figure 24: Method behavior before and after additive graph repair. Open circles denote the base graph, and filled squares the paired repaired graph. Panels A-D report 90% coverage (ideal 0.90), absolute z-score SD error (lower is better; y-axis on log scale), RMSE (lower is better), and false-best rate (lower is better at comparable declaration rates). Repair improves structural identifiability but does not eliminate method-specific behavior. ACEB leave-view remains best calibrated and has the lowest false-best rate, whereas joint EB retains low RMSE but remains overconfident. Points show means across paired benchmark cells; error bars show ± 2 SE.

E REAL-DATA DETAILS AND ADDITIONAL RESULTS

E.1 MEDHELM AUDIT

E.1.1 DATA CONSTRUCTION AND RAW TARGET-AWARE AUDIT

We evaluate the MedHELM/HELM-style pairwise-gap benchmark as a deployment audit rather than as a winner table. For each model, scenario, and metric, instance scores are deterministically ordered by a SHA-256 hash of the instance identifier and seed 0. Approximately 20% of the instances, with at least five instances, form a held-out aggregate target; the remaining instances are split into five observed view blocks. All metrics are oriented so that larger is better before model-pair differences are formed.

For model M, let $\mathcal{T}$ denote its held-out instance set. Let x_{Mr} denote the oriented score of model M on held-out instance r, $r \in \mathcal{T}$. The stored target artifact uses

$$T_M = \frac{1}{|\mathcal{T}|}\sum_{r\in\mathcal{T}} x_{Mr}, \qquad q_M = \max\left\{\frac{\mathrm{sd}(x_{Mr} : r \in \mathcal{T})}{\sqrt{|\mathcal{T}|}}, 0.01\right\}. \tag{34}$$

For an oriented model pair (A, B), the implementation then uses

$$T_{AB} = T_A - T_B, \qquad q_{AB} = \max\left\{\sqrt{q_A^2 + q_B^2},\ 0.02\right\}. \tag{35}$$

The quadrature expression does not estimate $\mathrm{Cov}(T_A, T_B)$ even when the two models are evaluated on the same held-out instances. Consequently, the target-aware coverage and interval scores inherit this working approximation. A positive paired covariance would make the quadrature standard error conservative, but its sign and magnitude cannot be recovered from the retained aggregate tables.

Let m_i and V_i be the posterior mean and variance produced by a method for pairwise record i, and let

$$\sigma^2_{i,\mathrm{TA}} = \max\{V_i, 10^{-18}\} + q_i^2. \tag{36}$$

The benchmark evaluates

$$T_i \mid \mathcal{D}_{\mathrm{obs}} \dot{\sim} \mathcal{N}(m_i, \sigma^2_{i,\mathrm{TA}}). \tag{37}$$

With $z_{.95} = \Phi^{-1}(0.95)$, target-aware coverage uses $m_i \pm z_{.95}\sigma_{i,\mathrm{TA}}$, standardized residuals use $(m_i - T_i)/\sigma_{i,\mathrm{TA}}$, and the reported z-score dispersion is their population standard deviation. RMSE uses $m_i - T_i$ and is independent of the variance; the 90% interval score uses the same endpoints:

$$\mathrm{IS}_i = (U_i - L_i) + 20(L_i - T_i)_+ + 20(T_i - U_i)_+. \tag{38}$$

No additional target-side residual term τ_t^2 is added. Fitted view residual noise already contributes to ACEB's posterior variance V_i, while q_i^2 represents finite-sample uncertainty in the aggregate target.

Scores and targets are sign-oriented before pairwise differencing and are compared directly on the fixed-reference scale; no fitted target-view loading or bias is applied. This assumes that the aggregate target and the reference block share the same oriented metric scale, as expected for disjoint random blocks of the same metric.

The column labeled *posterior width* is $2z_{.95}\sqrt{V_i}$ and is therefore distinct from the target-aware width $2z_{.95}\sqrt{V_i + q_i^2}$ implicit in coverage and interval score. Recomputation from the stored item-level predictions reproduced coverage, target-aware dispersion, interval score, RMSE, and posterior width to machine precision without refitting any model.

The full run contains 13,077 dependent pairwise-gap records with approximately five observed views per record. Joint EB is computationally infeasible on the full pairwise table, so its analysis remains restricted to the capped feasibility subset. A separate held-out-view diagnostic is distinct: it projects ACEB to a view-specific mean $b_t + a_t m_i$ and variance $a_t^2 V_i + s_{it}^2 + \tau_t^2$.

Figure 4 in the main paper summarizes the raw aggregate-target uncertainty audit, selective-decision curves, and runtime. The table below gives the corresponding numerical audit and distinguishes unsupervised from target-calibrated comparisons.

Table 4: MedHELM post-hoc fairness audit. Target-calibrated post-hoc IV uses evaluation targets during calibration and should not be described as deployable in the same sense as ACEB. Scenario-held-out post-hoc IV calibrates on distinct scenario-metric folds and reduces to inverse variance in this audit. Coverage and interval score are target-aware; width is posterior-only.

A. Uncertainty and estimation

Method	Calibration type	Coverage	z-score SD	RMSE	Interval score
Inverse variance	unsupervised	0.953	0.756	0.540	0.803
ASH-like	unsupervised	0.933	0.896	0.532	0.786
Random effects	unsupervised	0.962	0.722	0.522	0.774
Post-hoc IV	target-calibrated	0.994	0.462	0.540	1.108
Post-hoc IV	scenario-held-out	0.953	0.756	0.540	0.803
ACEB in-sample	unsupervised	0.992	0.399	0.594	1.368
ACEB leave-view	unsupervised	0.993	0.388	0.594	1.459

B. Posterior width and selective decisions

Method	Calibration type	Posterior width	Decl. rate	Selective acc.
Inverse variance	unsupervised	0.266	0.560	0.940
ASH-like	unsupervised	0.194	0.310	0.993
Random effects	unsupervised	0.297	0.559	0.941
Post-hoc IV	target-calibrated	0.900	0.257	0.994
Post-hoc IV	scenario-held-out	0.266	0.560	0.940
ACEB in-sample	unsupervised	0.947	0.164	0.986
ACEB leave-view	unsupervised	1.050	0.155	0.986

Unsupervised methods include ACEB leave-view, ACEB in-sample, adaptive shrinkage, inverse variance, and random effects; none uses held-out target labels during fitting or calibration. Target-calibrated post-hoc inverse variance uses evaluation targets and is therefore transductive. Scenario-held-out post-hoc inverse variance calibrates on separate scenario-metric groups, but learns no useful additional inflation in this audit and reduces to inverse variance.

Fixed-declaration-rate diagnostics and aggregate-target evaluation emphasize that ACEB operates at a different point from sharp ranking baselines. It is substantially more conservative and wider than adaptive shrinkage or inverse variance. The scenario-held-out audit below shows that this conservatism can reduce false signed declarations when abstention is allowed. We use MedHELM to evaluate target-label-free base fitting and target-calibrated selective decisions. The synthetic experiments isolate the graph-identification and residual-noise mechanisms.

E.1.2 Scenario-held-out coverage-aligned audit

Coverage-aligned recalibration. For each method h and fold f, a scalar standard-deviation multiplier $\widehat{c}_{hf}$ is estimated on calibration scenario–metric groups and applied to held-out groups. Posterior means and target standard errors remain fixed, while

$$V_{ih}^{\mathrm{cal}} = \widehat{c}_{hf}^{2} V_{ih}.$$

Coverage and interval score use $V_{ih}^{\mathrm{cal}} + q_i^2$, whereas posterior width and signed decisions use V_{ih}^{cal}, with sign probability $\Phi(m_{ih}/\sqrt{V_{ih}^{\mathrm{cal}}})$. Multipliers are fit separately by method and fold.

Held-out coverages range from 0.899 to 0.943, so the comparison is coverage-aligned rather than exactly matched. ACEB leave-view attains coverage 0.899, the lowest false-sign rate among declarations (0.187), and the lowest rate of false signed declarations per record (0.147). The tradeoff is a lower declaration rate and a worse interval score than the sharper baselines.

To express the abstention tradeoff explicitly, let d_h denote the declaration rate of method h, let f_h denote its false-sign rate among declared pairs, and let $\lambda \geq 0$ be the cost of a false signed claim relative to a correct signed declaration. Define

$$U_\lambda(h) = d_h(1 - f_h) - \lambda d_h f_h. \tag{39}$$

Table 5: Scenario-held-out coverage-aligned recalibration on MedHELM pairwise gaps. A method- and fold-specific scalar rescales posterior variance while posterior means and target standard errors remain fixed. Panel A reports target-aware uncertainty metrics; Post. width uses the recalibrated posterior variance only. Panel B reports selective-decision metrics. Held-out coverages are close but not identical.

A. Uncertainty and estimation

Method	Coverage	z-score SD	RMSE	Int. score	Post. width
Inverse variance	0.938	1.501	0.540	0.826	0.571
ASH-like	0.915	1.311	0.532	0.816	0.571
Random effects	0.943	1.489	0.522	0.792	0.571
Post-hoc IV	0.938	1.501	0.540	0.826	0.571
ACEB in-sample	0.901	1.495	0.594	0.915	0.573
ACEB leave-view	0.899	1.421	0.594	0.937	0.579

B. Selective decisions

Method	False-sign	Decl. rate	Sel. acc.	False signed per item
Inverse variance	0.195	0.983	0.805	0.192
ASH-like	0.195	0.982	0.805	0.191
Random effects	0.192	0.983	0.808	0.189
Post-hoc IV	0.195	0.983	0.805	0.192
ACEB in-sample	0.212	0.936	0.788	0.198
ACEB leave-view	0.187	0.786	0.813	0.147

At the reported operating points, Table 5 gives $U_\lambda(\mathrm{ACEB}) > U_\lambda(\mathrm{RE})$ for $\lambda \gtrsim 3.7$. This utility is evaluated against the noisy held-out target sign. It does not establish superiority after method-specific threshold tuning or at matched declaration rates.

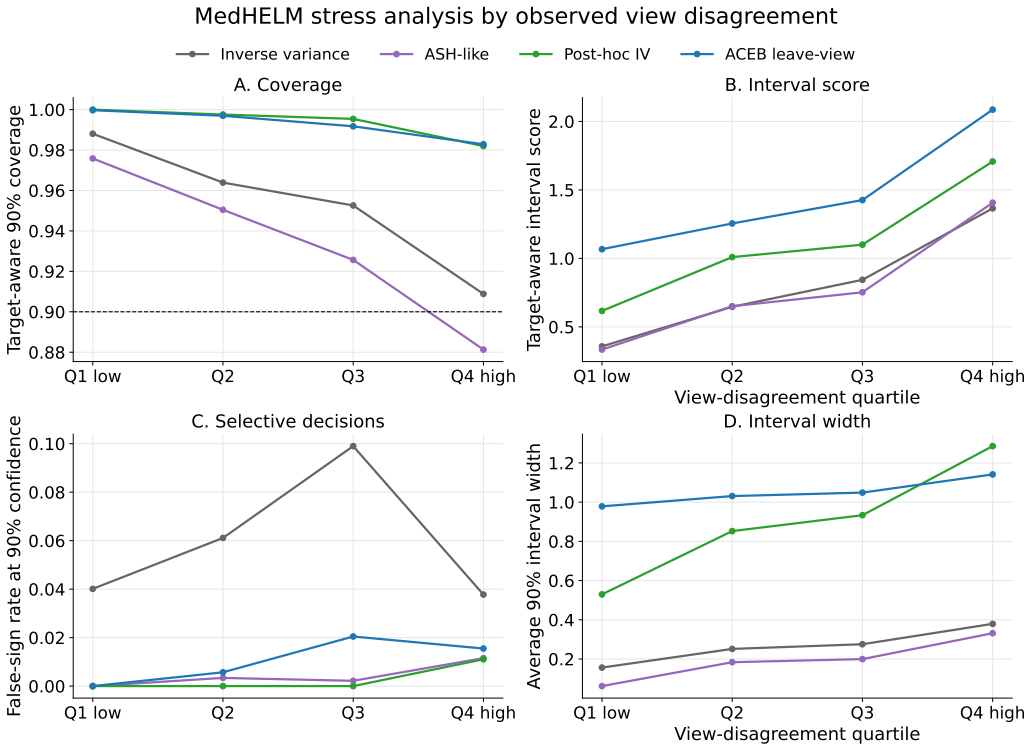


Figure 25: MedHELM stress analysis by observed-view disagreement. Quartiles are defined from observed-view disagreement, independently of method outcomes. Panel A reports target-aware 90% coverage (ideal 0.90). Panel B reports interval score (lower is better). Panel C reports false-sign rate among pairs declared at 90% confidence (lower is better only at comparable declaration rates). Panel D reports interval width. ACEB remains conservative in high-disagreement quartiles. ASH-like and inverse variance (IV) remain sharper; coverage declines with disagreement, with ASH-like becoming undercovered in the highest quartile and IV approaching nominal coverage from above. Post-hoc IV is target-calibrated.

E.1.3 MEDHELM GRAPH-REGIME AUDIT

The main MedHELM run also permits an audit of graph diagnostics on real view-subset designs. Each regime is fitted separately using only its retained observations. The present analysis reuses the stored regime-specific predictions without refitting them. Star and chain are rank deficient, whereas all-view, odd-cycle, star-plus-edge, and random-sparse subsets are topologically full rank under the anchored diagnostic. The normalized smallest singular value varies among those full-rank subsets, showing that structurally supported designs can still differ in conditioning.

ACEB is conservative across regimes. Its mean target-aware coverage is 0.992 in deficient designs and 0.988 in full-rank designs. Mean posterior width is larger in deficient designs, 3.939 rather than 2.654, consistent with cautious uncertainty under weaker measurement structure. Downstream results remain mixed: ACEB is wider than adaptive shrinkage and has a worse interval score, while target-calibrated post-hoc inverse variance remains conservative when evaluation targets are used. This is therefore measurement-design evidence rather than a dominance result.

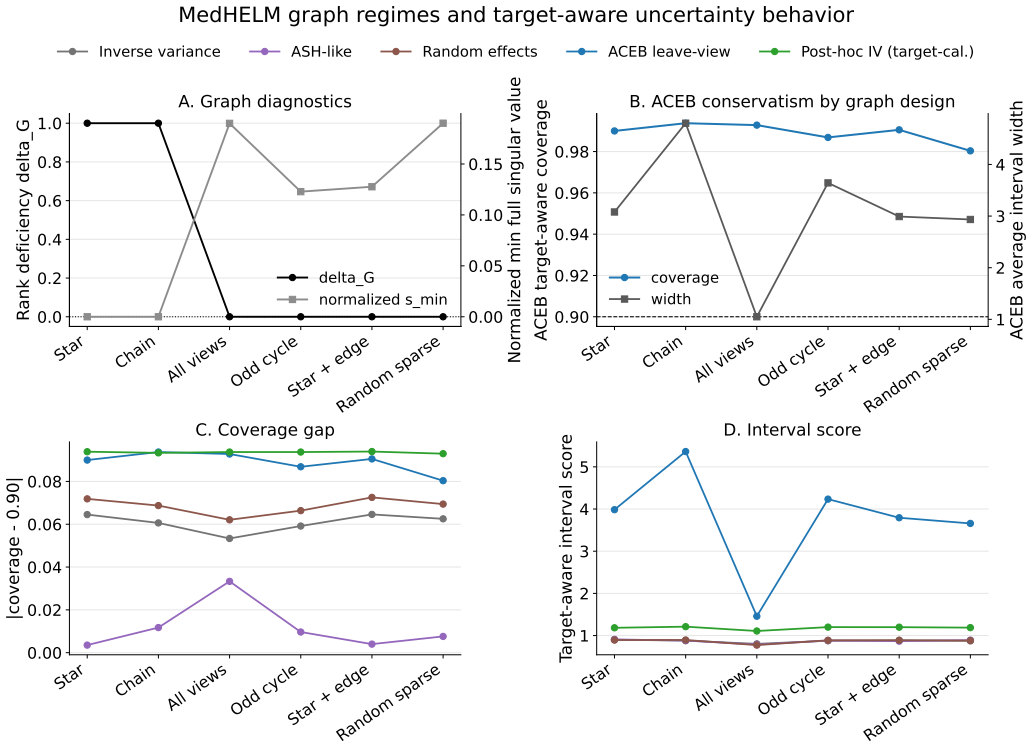


Figure 26: MedHELM graph regimes and target-aware uncertainty. Panel A shows that star and chain view subsets are rank deficient, whereas all-view, odd-cycle, star+edge, and random-sparse subsets have positive smallest full-column singular values and are full rank under the working model. Panel B reports ACEB coverage (ideal 0.90) and posterior-only interval width across designs. Panels C and D compare methods using absolute coverage gap and target-aware interval score (lower is better in both). Graph diagnostics are method-independent, but downstream performance is mixed: ACEB produces more conservative intervals at a substantial sharpness cost. Post-hoc IV is target-calibrated and transductive.

E.2 SUMMEVAL HUMAN-EVALUATION PILOT

We include a second public human-evaluation pilot on SummEval (Fabbri et al., 2021). SummEval contains system summaries scored by crowd and expert annotators along multiple quality dimensions. We construct units as summary-system-dimension triples, use the five anonymous crowd annotation slots as observed views, and use the mean of three expert annotations as the held-out target. The implementation uses a fixed crowd-view standard error of 0.75 and floors the expert-target standard

error at 0.10. Expert targets from evaluation units are not used for model fitting or same-coverage recalibration.

For SummEval, base post-hoc IV is calibrated from held-out views rather than expert targets; the later same-coverage recalibration for all methods uses expert targets from separate calibration articles.

This construction is narrower than MedHELM because the crowd slots are anonymous and exchangeable rather than persistent named annotators. We therefore treat SummEval as supplementary validation rather than as a second main benchmark. Before same-coverage recalibration, all methods are undercovered, but ACEB improves raw target-aware coverage and z-score dispersion relative to inverse variance and adaptive shrinkage. The same pattern occurs among high-disagreement units. After scalar same-coverage recalibration, the simpler baselines achieve lower interval scores than ACEB (Table 6). This supports ACEB as a target-label-free uncertainty protocol rather than a uniformly sharp estimator.

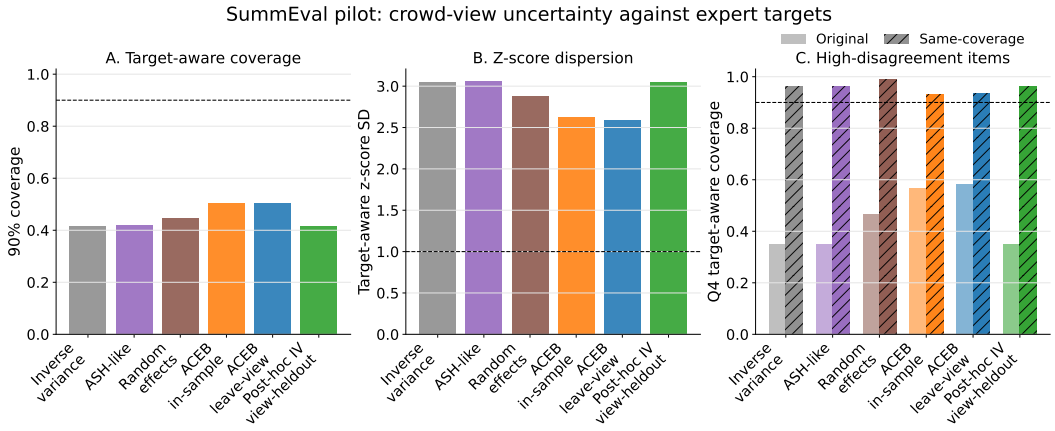


Figure 27: SummEval pilot. Units are summary-system-dimension triples, the five crowd annotation slots are treated as noisy views, and three-expert consensus is used as the target. Panels A-B report raw target-aware 90% coverage and z-score dispersion. Panel C restricts evaluations to the highest crowd-disagreement quartile and compares target-aware coverage before and after scalar same-coverage recalibration learned on separate calibration articles. Table 6 reports the corresponding same-coverage interval scores.

Table 6: SummEval pilot. Before same-coverage recalibration, ACEB improves raw coverage and dispersion relative to inverse variance and adaptive shrinkage. After recalibration, ACEB reaches nominal coverage but has a larger interval score.

Method	Raw coverage	Raw z-SD	RMSE	Raw interval score	Same-cov. coverage	Same-cov. interval score
Inverse variance	0.417	3.044	1.294	11.298	0.927	5.148
ASH-like	0.418	3.055	1.299	11.446	0.928	5.139
Random effects	0.446	2.876	1.294	10.608	0.936	5.581
ACEB in-sample	0.503	2.622	1.373	10.430	0.901	6.004
ACEB leave-view	0.503	2.589	1.385	10.499	0.900	6.060
Post-hoc IV view-held-out	0.417	3.044	1.294	11.298	0.927	5.148

F Reproducibility

All synthetic runs were executed as scenario cells with per-cell status files. The main graph benchmark contained 840 valid cells and 120 intentionally skipped invalid graph-design cells, with zero failed cells. Observed graphs were checked against intended graphs, and full-column singular values were computed from $B_{\mathcal{G}}^{\top} B_{\mathcal{G}}$ so that implicit zero singular values were retained.

Table 7: Maximum L-BFGS-B iteration counts used in the reported staged ACEB fits. No restart or optimizer-success-based exclusion rule was used; all other stopping tolerances use SciPy defaults.

Experiment family	ACEB `max_iter`
Main synthetic graph benchmark	8
Paired synthetic graph interventions	8
Component/oracle ablations	40
Anchor-sensitivity analysis	40
Misspecification analysis	40
Structural edge-selection analysis	35
Synthetic same-coverage analysis	100
Dense negative control	100
MedHELM full and joint-subset runs	3
SummEval pilot	80

Interpretation of optimization budgets. The iteration counts in Table 7 are computational caps, not convergence guarantees. Successful completion of a benchmark cell means that the required outputs were produced; it does not establish that every optimization stage reached a stationary point. The reported comparisons therefore concern the specified numerical implementations. Separating statistical behavior from optimization sensitivity requires within-method comparisons of fitting objectives, termination status, iteration-cap frequency, and stability under larger budgets and alternative initializations.

The paired intervention benchmark used shared latent variables, loadings, residual noises, and base observations across paired arms. The component/oracle, same-coverage, negative-control, anchor, misspecification, edge-selection, and MedHELM graph-regime analyses were run on compact grids or existing outputs selected to answer mechanism questions rather than to maximize benchmark size.

The MedHELM target-aware metrics were independently recomputed from stored item-level predictions, with maximum absolute discrepancies below 7×10^{-16} and no model refitting.