Can Ideologues Agree on Quality? From Non-identifiable Latent Factors to Collective Outcomes

David Gamba Seura Ha Daniel M. Romero Grant Schoenebeck
University of Michigan
gamba@umich.edu, seura@umich.edu, drom@umich.edu, schoeneb@umich.edu

Abstract

To surface "objective" fact-check notes, X's Community Notes uses Biased Matrix Factorization (BMF) on crowd-sourced ratings to separate note quality from reviewer ideology. We show that for any correctly specified model that decomposes ratings into quality and ideological alignment, this separation requires fixing an ideological anchor: a reference point for what counts as neutral. We give a characterization for the standard implementation: with L_2 regularization, BMF anchors quality at a point proportional to the mean reviewer ideology, recovering objective quality if and only if the mean reviewer is neutral, assuming note quality is uncorrelated with note leaning. This parallels the Condorcet Jury Theorem: collective judgment recovers truth when the pool is unbiased on average. When that assumption fails, BMF's recovered ranking shifts toward maximizing aggregate reviewer satisfaction, which coincides with social welfare only when the pool is representative. This is not specific to BMF; the rating matrix admits a continuous family of gauge-equivalent decompositions, each implying a different quality ranking, so no mechanism based on ratings alone can recover quality without implicitly choosing such an anchor. What appears to be routine hyperparameter tuning is in fact a normative choice about whose preferences define quality.

1 Introduction

When a post goes viral on X (formerly Twitter), a crowdsourced fact-checking note may appear beneath it. Whether that note is shown to millions of users is determined by an algorithm that at its core uses Biased Matrix Factorization (BMF), which aggregates ratings from volunteer reviewers to estimate the "quality" of each proposed note. The system, called Community Notes (formerly Birdwatch), has become one of the most prominent examples of algorithmic governance at scale [Wojcik et al., 2022, Allen et al., 2022]. Similar systems are being adopted elsewhere: Meta has announced plans to replace professional fact-checkers with community-based evaluation [Vallance, 2025, Meta, 2025]. As social media platforms have become major political arenas where algorithmic curation mediates discourse and often intensifies polarization [Pathak and Spezzano, 2024, Corsi, 2023], understanding these mechanisms becomes a pressing concern.

BMF's design aims to extract "objective quality"[1] from ideologically biased ratings [Wojcik et al., 2022]. The intuition is that if a note receives high ratings from both left-leaning and right-leaning reviewers, it is probably genuinely good, whereas a note praised only by one side may simply be ideologically aligned. BMF formalizes this by learning latent ideology vectors for both reviewers and notes, then defining quality as the residual, what persists after accounting for ideological alignment. **Whether this choice recovers objective quality is the central question of this paper.**

[1] Community Notes documentation refers to "idiosyncratic helpfulness." We interpret this as a quality measure used to decide whether to display a note.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

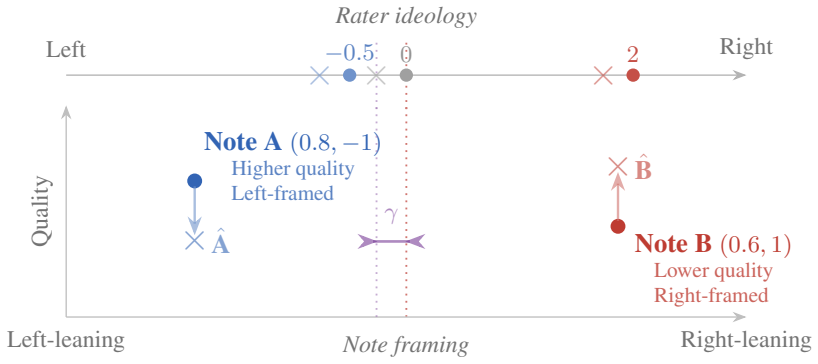


Figure 1: Running example. Three reviewers ($u \in \{-0.5, 0, 2\}$, dots) evaluate Note A (left-leaning, higher-quality) and Note B (right-leaning, lower-quality). Crosses show the gauge-shifted representation; reviewer ideologies shift left by γ while note qualities adjust proportionally, leaving all ratings unchanged but reversing the ranking ($\hat{q}_B > \hat{q}_A$). BMF implicitly selects this shifted frame.

To answer this question, we assume ratings are generated by a ground truth model in which BMF is correctly specified. That is, each reviewer i has a latent ideology $u_i \in \mathbb{R}$ and each note j has a true quality $q_j \in \mathbb{R}$ and an ideological leaning $v_j \in \mathbb{R}$, with ratings given by:

$$r_{ij} = q_j + u_i\, v_j + \varepsilon_{ij}, \tag{1}$$

where ε_{ij} are i.i.d. noise terms. This matches the functional form that BMF uses to model rating generation (Equation (2)), with 'true' parameters in place of estimated ones. For the algorithm analysis, we take the noiseless limit: if BMF fails to recover quality here, it cannot do so in more general conditions.

Our analysis characterizes precisely what BMF recovers, starting from a simple observation: the decomposition of a rating into "quality" and "ideological alignment" is not unique. One can shift all reviewer ideologies by a constant and compensate by adjusting quality scores, without changing any observed or predicted rating (as illustrated in Figure 1). This *gauge symmetry*, or *non-identifiability*, means the data alone cannot determine which reviewers are neutral and which are biased; different anchor choices lead to different quality rankings. We show that for extracting quality from ratings, BMF implicitly designates some ideology as the reference point—an *anchor* defining which perspective is "neutral". More specifically, we find that regularization plays a major role in this:

1. L_2 **regularization anchors at the mean ideology.** When BMF resolves the gauge symmetry ambiguity via L_2 regularization, the recovered qualities satisfy $\hat{q}_j = q_j + \gamma^* v_j$ up to a constant, where $\gamma^* = s\,\mu_u$, μ_u is the mean reviewer ideology, and $s \in (0, 1)$ (Theorem 10). Quality is recovered when $\mu_u = 0$; otherwise rankings are shifted toward the average reviewer's perspective (Corollary 17).
2. **No mechanism on ratings alone can do better.** For any two notes with different ideological leanings, there exists a gauge-equivalent parameterization that preserves all ratings but reverses their quality ordering (Theorem 15). This is a property of the rating data itself: *no* mechanism operating on ratings alone can recover quality (Corollary 16).

Just as the Condorcet Jury Theorem guarantees majority vote recovers truth when the voter pool is unbiased on average [de Condorcet, 1785], BMF recovers quality exactly when the reviewer pool is ideologically balanced ($\mu_u = 0$), assuming note quality is uncorrelated with note leaning. This limitation is not an artifact of BMF's particular algorithm but a property of the rating data itself, so no mechanism can escape it. What appears to be a routine choice of regularization weights is in fact a normative decision about whose ideological perspective defines quality; we discuss implications for platform design in Section 6.

Related work. BMF is a standard recommender-systems technique [Koren et al., 2009, Bell and Koren, 2007, He et al., 2017] adapted for content evaluation in Community Notes [Wojcik et al., 2022]. Our model (1) is closely related to ideal-point models in political science [Poole and Rosenthal,

1985, Clinton et al., 2004] and item-response theory in psychometrics, both of which have long required explicit anchoring for interpretation. Non-identifiability in factor models has been studied since Anderson and Rubin [1956], Bai and Li [2012]; conditions for identifiable representation learning remain an active area [Ahuja et al., 2021, Locatello et al., 2019, Park et al., 2024, Zou et al., 2025]. Our result parallels Locatello et al. [2019] and D'Amour [2019]: just as disentanglement is impossible without inductive biases and latent confounders are not identified by treatment structure alone, quality recovery is impossible without an ideological anchor.

Content moderation can also be viewed as a social choice problem [Roberts, 2019, Wang and Lucas, 2024], connecting to classical and computational social choice [Arrow, 2012, Brandt et al., 2016, Boutilier et al., 2015] and to recent work framing ML systems as preference aggregators [Qiu, 2025, Siththaranjan et al., 2024, Fish et al., 2024, Christiano et al., 2023], which face well-known strategic vulnerabilities [Gibbard, 1973]. The tension between aggregation and fairness in recommender systems [Ekstrand et al., 2018, Ge et al., 2022, Dwork et al., 2012], together with calls for algorithmic accountability [Kroll et al., 2017, Mitchell et al., 2019], motivates making such implicit normative choices explicit.

2 Biased Matrix Factorization and the Correctly Specified Setting

2.1 The BMF Mechanism

We consider n reviewers indexed by $i \in [n]$ and m notes indexed by $j \in [m]$. The rating of reviewer i on note j is $r_{ij} \in \mathbb{R}$ (with $r_{ij} = \perp$ if unobserved).

Definition 1 (Biased Matrix Factorization)**.** BMF models the rating of reviewer i on note j as:

$$\hat{r}_{ij} = \hat{c} + \hat{b}_i + \hat{q}_j + \hat{u}_i \cdot \hat{v}_j, \tag{2}$$

where $\hat{c} \in \mathbb{R}$ is a global intercept, $\hat{b}_i \in \mathbb{R}$ is a reviewer-specific intercept, $\hat{q}_j \in \mathbb{R}$ is interpreted as the quality of note j, and $\hat{u}_i, \hat{v}_j \in \mathbb{R}^d$ are latent ideology factors. Parameters are estimated by minimizing:

$$\mathcal{L}(\hat{\theta}) = R(\hat{\theta}) + \Lambda(\hat{\theta}) \; = \sum_{i,j:\, r_{ij} \neq \perp} (r_{ij} - \hat{r}_{ij})^2 + \Lambda(\hat{\theta}), \tag{3}$$

where $\hat{\theta} = \{\hat{c}, \hat{b}_i, \hat{q}_j, \hat{u}_i, \hat{v}_j\}$, $R(\hat{\theta})$ is the empirical risk and $\Lambda(\hat{\theta})$ the regularization term.

We focus on the model with scalar ideology factors ($d = 1$), matching Community Notes' implementation. The $d > 1$ case introduces one latent dimension per ideological axis; the non-identifiability results of Sections 2.3 and 4 extend, since each axis admits a shift ambiguity of the kind formalized in Section 2.3, while the anchor characterization of Section 3 is specific to $d = 1$ (Section 6). The key interpretive claim is that $\hat{q}_j$ captures intrinsic quality independent of ideology; notes are selected for display when $\hat{q}_j$ exceeds a threshold.

We note that Community Notes uses categorical ratings (helpful/not helpful/somewhat helpful) rather than continuous scores, and employs logistic loss rather than squared error. Our analysis works in the underlying *score space* (given by continuous $\hat{r}_{i,j}$), where BMF fits parameters in both cases; the logistic link $\sigma(\cdot)$ enters only when scores are mapped to rating probabilities for the categorical loss, leaving the structure we study unchanged. Community Notes also includes additional filtering steps (minimum rating counts, tag-based weighting) that we do not model. Our analysis focuses on the core factorization mechanism, and our empirical results (Section 5) verify that the main findings are robust when extended to a binary setting (helpful/not helpful).

2.2 The Correctly Specified Setting

To evaluate whether BMF recovers objective quality, we need a setting with well-defined ground truth. We assume ratings are generated by a model that matches BMF's functional form (Equation (2)), with the same global intercept c, per-reviewer intercepts b_i, quality terms q_j which represent 'objective' quality, and ideological alignment $u_i v_j$. If reviewer and note share the same ideological sign, the reviewer rates it above quality alone.

Definition 2 (Ground Truth Model)**.** Each reviewer i has a true ideology $u_i \in \mathbb{R}$ and intercept $b_i \in \mathbb{R}$. Each note j has a true quality $q_j \in \mathbb{R}$ and ideological leaning $v_j \in \mathbb{R}$. With a global intercept $c \in \mathbb{R}$,

ratings are generated as:

$$r_{ij} = c + b_i + q_j + u_i\, v_j + \varepsilon_{ij}, \tag{4}$$

where ε_{ij} are i.i.d. noise terms with mean zero and variance Δ.

For the analysis of what BMF recovers in Section 3 we additionally set $c = b_i = 0$, which we call the *simplified ground truth model*, isolating the two elements central to our question: note quality q_j and ideological alignment $u_i v_j$. This simplifies the proof for Theorem 10 while keeping the interaction between quality and ideological alignment; the result extends under weaker conditions on c and b_i (Section B.2). The gauge and tightness results in Sections 2.3 and 4 hold for the full ground truth model without this simplification.

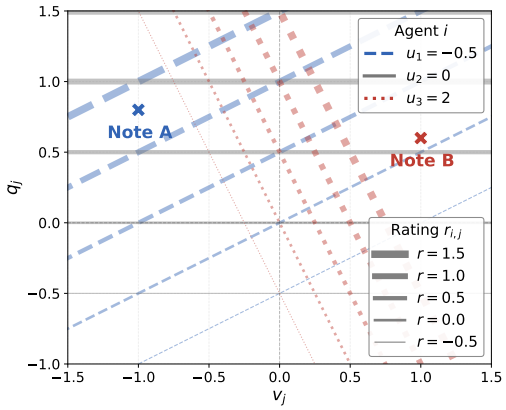


Figure 2: Iso-rating contours for the running example (Figure 1). Contour lines show constant rating for reviewers. The neutral reviewer's contours are horizontal (rating equals quality); ideological reviewers have tilted contours, systematically over- or under-rating ideologically aligned notes.

We analyze the noiseless setting ($\Delta = 0$) under correct specification, the most favorable setting for the algorithm: if BMF fails to recover quality on data its own model generated, noise does not help it (Corollary 12), and misspecification did not help it in the three departures we tested (Section D.8). On real ratings the true ideologies and qualities are unobserved, so there is no standard against which to check the anchor; the synthetic setting supplies one.

A reviewer with $u_i = 0$ is *ideologically neutral*; their ratings rank notes by quality. This leads to a key observation:

Proposition 3 (Neutral-reviewer ratings rank by quality)**.** *In the ground truth model, for any reviewer i with $u_i = 0$ and any two notes j_1, j_2: $q_{j_1} > q_{j_2}$ if and only if $\mathbb{E}[r_{ij_1}] > \mathbb{E}[r_{ij_2}]$.*

Proof. $\mathbb{E}[r_{ij} \mid u_i = 0] = c + b_i + q_j$. Since $c + b_i$ is constant across notes, $\mathbb{E}[r_{ij_1}] - \mathbb{E}[r_{ij_2}] = q_{j_1} - q_{j_2}$. □

Thus, recovering the ranking of items according to objective quality is equivalent to recovering the perspective of the neutral reviewer. The question is whether BMF can identify which perspective is neutral from ratings alone.

Running example (Figure 1). Throughout, the running example uses the simplified model setting ($c = b_i = 0$). Two candidate fact-checking notes have been proposed: Note A is higher quality with left-leaning framing ($q_A = 0.8$, $v_A = -1$); Note B is of lower quality but right-leaning ($q_B = 0.6$, $v_B = 1$). Three reviewers span the ideological spectrum: $u_1 = -0.5$ (left), $u_2 = 0$ (neutral), $u_3 = 2$ (far-right). The pool skews right (average ideology 0.5). Ratings are tabulated in section C.

2.3 Gauge Symmetries of BMF

The running example illustrates a key property of the rating model: shifting all reviewer ideologies by a constant while adjusting qualities to compensate leaves every rating unchanged. We characterize the full family of such transformations and identify which one affects quality rankings.

Lemma 4 (Gauge Invariance of Ratings)**.** *In the ground truth model, for any $\gamma \in \mathbb{R}$, define $u'_i = u_i - \gamma$, $q'_j = q_j + \gamma v_j$, $v'_j = v_j$. Then for all i, j: $r'_{ij} = c + b_i + q'_j + u'_i v'_j = r_{ij}$.*

Proof. $c + b_i + q'_j + u'_i v'_j = c + b_i + (q_j + \gamma v_j) + (u_i - \gamma) v_j = c + b_i + q_j + u_i v_j$. □

The two parameter sets generate identical ratings for all reviewer-note pairs. The term "gauge invariance" is borrowed from physics, where it describes symmetries under which observable quantities are unchanged despite changes in representation. Here, ratings are observable; latent factors are not.

This symmetry extends to the full BMF model (2), which admits a richer family of transformations. We characterize all of them, then identify which one matters for quality recovery.

Theorem 5 (Characterization of BMF's Gauge Symmetries)**.** *Every gauge transformation of BMF (i.e., every parameter mapping that preserves all predicted ratings $\hat{r}_{ij}$) can be parameterized by $(\alpha, \beta, \gamma, \eta, \tau)$ with $\alpha, \beta, \gamma, \eta \in \mathbb{R}$ and $\tau \in \mathbb{R} \setminus \{0\}$:*

$$\begin{aligned} \hat{u}'_i &= \tau(\hat{u}_i - \gamma), \qquad \hat{q}'_j = \hat{q}_j + \gamma\hat{v}_j + \beta, \\ \hat{v}'_j &= \hat{v}_j/\tau - \eta/\tau, \qquad \hat{b}'_i = \hat{b}_i + \hat{u}_i\eta + \alpha, \qquad \hat{c}' = \hat{c} - \alpha - \beta - \gamma\eta. \end{aligned} \tag{5}$$

These five parameters exhaust all degrees of non-identifiability.

Proof sketch. A derivative argument on the bilinear constraint $\hat{r}'_{ij} = \hat{r}_{ij}$ shows that rating-preserving maps are affine in each factor. Requiring the same form to hold for every reviewer-note pair forces coefficients to be index-independent, leaving five universal scalars. □

Each of the five parameters in (5) corresponds to a distinct rating-preserving symmetry. γ is the agent-centric shift of Lemma 4, moving the ideological zero among reviewers; η is its item-side analogue, a shift in note ideology absorbed into the per-reviewer intercept; τ rescales reviewer and note ideologies reciprocally, leaving the bilinear product $\hat{u}_i\hat{v}_j$ unchanged; and α, β are additive offsets that redistribute constants among $\hat{c}, \hat{b}_i, \hat{q}_j$. Among these five symmetries, only one affects the quality ranking.

Proposition 6 (Only γ Affects Quality Ordering)**.** *The relative ordering of $\hat{q}_j$ is invariant under the item-centric shift (η), scaling (τ), and global offsets (α, β). Only the agent-centric shift γ can change the ordering.*

Proof sketch. From (5), $\hat{q}'_j = \hat{q}_j + \gamma\hat{v}_j + \beta$. For two notes: $\hat{q}'_{j_1} - \hat{q}'_{j_2} = (\hat{q}_{j_1} - \hat{q}_{j_2}) + \gamma(\hat{v}_{j_1} - \hat{v}_{j_2})$. Only γ appears in this difference; β cancels. The item transformation changes $\hat{v}_j$ but not $\hat{q}_j$; scaling by τ preserves the inner product $\hat{u}_i\hat{v}_j$ and does not enter $\hat{q}_j$ directly. Full proof in Section B.1. □

In the running example, choosing $\gamma = 1$ shifts the neutral reference from $u = 0$ to $u = 1$: the quality estimates become $\hat{q}_A = -0.20$ and $\hat{q}_B = 1.60$, re-ranking Note B above Note A despite its lower true quality, while leaving every entry of the rating matrix unchanged (Table 2 in Section C).

This result simplifies the analysis considerably: for quality rankings, the only consequential degree of freedom is γ, i.e., which ideology is treated as neutral. By Proposition 3, this is the ideology whose perspective ranks notes by quality. The remaining gauge parameters either cancel in pairwise quality differences or are absorbed by the constant offset β. The question then becomes: how does BMF, in practice, choose γ?

3 How L_2 Regularization Resolves the Ambiguity

We have established that the risk component R of BMF's loss is gauge-invariant, so its minimizers form a manifold of equivalent solutions. In practice, however, BMF returns a single answer because the regularizer resolves this ambiguity. Proposition 6 reduces BMF's ranking ambiguity to a single scalar γ. We now derive where L_2 regularization places it.

The intuition is as follows. L_2 regularization balances two forces: penalizing $\sum_i \hat{u}_i^2$ favors centering reviewer ideologies at zero, which pulls the anchor γ toward the population mean μ_u; penalizing $\sum_j \hat{q}_j^2$ favors small quality estimates, which pulls γ toward zero. The ratio $\kappa = \lambda_q/\lambda_u$ controls this trade-off. In the remainder of this section, we formalize this intuition through a set of balance equations and solve for the resulting anchor.

We will show (Theorem 10) that the selected anchor is $\gamma^* = s\,\mu_u$ for some $s \in (0, 1)$ depending on regularization weights. This implies our main result: BMF recovers quality only if $\mu_u = 0$.

Notation. We write $\mu_u = \frac{1}{n}\sum_i u_i$ and $\mu_v = \frac{1}{m}\sum_j v_j$ for the empirical means of reviewer and note ideologies, $\sigma_u^2 = \frac{1}{n}\sum_i (u_i - \mu_u)^2$ and $\sigma_v^2 = \frac{1}{m}\sum_j v_j^2$ for the corresponding empirical variances (σ_v^2 is uncentered; it is the variance when $\mu_v = 0$, as Theorem 10 assumes), and $\rho = m/n$ for the note-to-reviewer ratio. Analogous shorthands μ_q, μ_b are used as needed.

3.1 The Min-Min Solution

We begin by formalizing how regularization interacts with the gauge orbit. Recall that Δ denotes the noise variance from Definition 2. In the noiseless case ($\Delta = 0$), the true parameters achieve zero reconstruction loss. We write $\mathcal{G}(\theta_{\text{true}})$ for the *gauge orbit* of the true parameters — the set of all parameter vectors producing the same predictions as θ_{true}:

$$\mathcal{G}(\theta_{\text{true}}) = \{T(\theta_{\text{true}}) : T \text{ is a gauge transformation (5)}\}.$$

This is the set of all zero-risk parameterizations. Regularization selects from this set by preferring "simpler" solutions.

Definition 7 (Min-Min Solution)**.** The *min-min solution* θ^*_Λ is the element of $\mathcal{G}(\theta_{\text{true}})$ minimizing $\Lambda(\theta)$ (unique under L_2, Lemma 18). The reflection $(\hat{u}, \hat{v}) \mapsto (-\hat{u}, -\hat{v})$, the gauge element $\tau = -1$, leaves R, Λ_{L_2}, and $\hat{q}$ unchanged; we identify parameter vectors that differ by it, which amounts to taking $\tau > 0$ on the orbit. A starred quantity for a parameter $(\gamma^*, \tau^*, \alpha^*, \dots)$ denotes its value at θ^*_Λ.

Proposition 8 (Regularized BMF converges to the min-min solution)**.** *In the noiseless setting, write the regularizer as $\lambda\Lambda(\theta)$, with $\lambda > 0$ a common factor and the ratios of $\lambda_u, \lambda_q, \lambda_v, \lambda_b, \lambda_c$ fixed. As $\lambda \to 0$, the minimizer θ_λ of $R(\theta) + \lambda\Lambda(\theta)$ converges to θ^*_Λ.*

Proof. By optimality of θ_λ against θ^*_Λ (which has zero risk): $R(\theta_\lambda) + \lambda\Lambda(\theta_\lambda) \le \lambda\Lambda(\theta^*_\Lambda)$. Hence $R(\theta_\lambda) \to 0$ as $\lambda \to 0$, forcing θ_λ onto the zero-risk manifold $\mathcal{G}(\theta_{\text{true}})$. Among zero-risk solutions, θ^*_Λ uniquely minimizes Λ (Lemma 18). By standard results in perturbation theory, the minimizer of a penalized objective converges to the minimizer of the penalty on the limiting constraint set as the penalty weight goes to zero. Hence $\theta_\lambda \to \theta^*_\Lambda$. □

3.2 Balance Equations

The min-min solution must satisfy first-order optimality conditions along each gauge direction. We call these *balance equations*. Since the risk is constant on the gauge orbit, the gradient of the regularizer must vanish along each symmetry direction. All five balance equations hold; the intercept and item-centric ones can be shown to force $\hat{b}_i, \hat{c} \to 0$ under the simplified ground truth model ($c = b_i = 0$). The two equations that determine γ^* and hence quality rankings are the agent-centric balance (by Proposition 6) and the scale balance, which couples with it.

We consider L_2 regularization of the form

$$\Lambda_{L_2}(\theta) = \lambda_u \sum_i \hat{u}_i^2 + \lambda_q \sum_j \hat{q}_j^2 + \lambda_v \sum_j \hat{v}_j^2 + \lambda_b \sum_i \hat{b}_i^2 + \lambda_c \hat{c}^2. \tag{6}$$

Lemma 9 (Balance Equations for L_2)**.** *Any minimizer of the L_2-regularized BMF objective satisfies:*

$$\begin{aligned} \lambda_u \textstyle\sum_i \hat{u}_i = \lambda_q \sum_j \hat{q}_j \hat{v}_j \quad &(\textit{Agent-centric}) & \qquad \lambda_u \textstyle\sum_i \hat{u}_i^2 = \lambda_v \sum_j \hat{v}_j^2 \quad &(\textit{Scale}) \\ \lambda_b \textstyle\sum_i \hat{b}_i = \lambda_q \sum_j \hat{q}_j = \lambda_c \hat{c} \quad &(\textit{Intercept}) & \qquad \lambda_b \textstyle\sum_i \hat{b}_i \hat{u}_i = \lambda_v \sum_j \hat{v}_j \quad &(\textit{Item-centric}) \end{aligned} \tag{7}$$

Proof sketch. For each gauge symmetry, the risk is invariant along the corresponding orbit, so the regularization term Λ must be stationary at the minimum in each gauge direction. Each equation follows by differentiating Λ with respect to the corresponding gauge parameter and setting the derivative to zero. Full proof in Section B.2. □

The agent-centric balance has a natural interpretation: if $\sum_j \hat{q}_j \hat{v}_j = 0$ (quality and ideology uncorrelated), it requires $\sum_i \hat{u}_i = 0$, centering reviewer ideologies at zero. The right-hand side pulls the anchor away from this centering by an amount proportional to the quality-ideology covariance. The scale equation balances the norms of reviewer and note ideology vectors; together with the agent-centric it forms a coupled system whose unique solution is (γ^*, τ^*).

3.3 Main Result: BMF Anchors Proportional to the Mean

We now characterize the anchor selected by L_2 regularization. The main result is that the anchor lands at a fraction of the mean reviewer ideology: $\gamma^* = s\mu_u$ for some $s \in (0, 1)$.

Theorem 10 (BMF Anchors proportional to μ_u)**.** *Consider ratings generated by the simplified ground truth model ($c = b_i = 0$[2]*, *Definition 2) with mean reviewer ideology μ_u and variance $\sigma_u^2 > 0$, zero-mean notes ($\mu_q = \mu_v = 0$), and $\mathrm{Cov}(q_j, v_j) = 0$. Under L_2 regularization, as $\lambda \to 0$, BMF recovers $\hat{q}_j = q_j + \gamma^* v_j + \beta$ where β is a ranking-irrelevant constant and*

$$\gamma^* = \frac{\mu_u}{1 + \kappa\rho\sigma_v^2/\tau^{*2}}, \qquad \tau^{*4} = \frac{\lambda_v m \sigma_v^2}{\lambda_u n[\sigma_u^2 + (\mu_u - \gamma^*)^2]}, \tag{8}$$

with $\kappa = \lambda_q/\lambda_u$ and $\rho = m/n$. In particular, $\gamma^* = s\mu_u$ for some $s \in (0, 1)$.

Proof sketch. The proof has three steps:

Step 1: Reduce to two unknowns. By Proposition 8, the solution converges to the min-min solution, which minimizes Λ_{L_2} over the gauge orbit $\mathcal{G}(\theta_{\mathrm{true}})$. On this orbit, the agent-centric shift transforms parameters as $\hat{u}_i = \tau(u_i - \gamma)$ and $\hat{q}_j = q_j + \gamma v_j$. The item-centric shift η and global offsets α, β can be shown to vanish at the minimum: for fixed (γ, τ), the regularizer is strictly convex and quadratic in (η, α, β), and the moment conditions $\mu_q = \mu_v = 0$ eliminate the linear terms, forcing $\eta^* = \alpha^* = \beta^* = 0$ (Section B.2, Step 2).

This reduces the problem to two equations in (γ, τ).

Step 2: Apply the balance equations. The two remaining free parameters (γ, τ) must satisfy the agent-centric and scale balance equations (Lemma 9). Substituting $\hat{u}_i = \tau(u_i - \gamma)$, $\hat{q}_j = q_j + \gamma v_j$ into the agent-centric balance, using $\mathrm{Cov}(q, v) = 0$, and solving for γ yields γ^* in (8). Similarly, the scale balance yields an expression for τ^*.

Step 3: $s \in (0, 1)$. The coupled system reduces to a single monotone equation in τ^2 with exactly one root, so the fixed point (γ^*, τ^*) is unique (Lemma 18). Since $\kappa, \rho, \sigma_v^2, \tau^{*2} > 0$ ensure the denominator in (8) exceeds 1, $\gamma^* = s\mu_u$ with $s \in (0, 1)$. By Proposition 6, only γ affects quality rankings. Full proof in Section B.2. □

The theorem has direct implications for when quality can be recovered. When the reviewer pool is ideologically neutral, the anchor vanishes, and BMF recovers quality:

Corollary 11 (Quality Recovery under Neutrality)**.** *Under the conditions of Theorem 10, if $\mu_u = 0$ then BMF recovers the correct quality ranking for every configuration (q, v). Proof in Section B.2.*

The converse, whether $\mu_u \neq 0$ always breaks recovery, requires additional arguments and is established in Section 4. Missing the 'objective' ranking is also not a noiseless artifact: it persists as ratings become precise.

Corollary 12 (Recovery Under Noise)**.** *For any reviewer pool with $\mu_u \neq 0$, the recovery failure is structural: as $\Delta \to 0$, the L_2 anchor satisfies $\gamma^*(\Delta) \to s\mu_u$ with $s \in (0, 1)$, so for all sufficiently small $\Delta > 0$ BMF still misorders, with probability tending to one as $\Delta \to 0$, the configurations (q, v) that it misorders at $\Delta = 0$. Proof in Section B.2.*

When $\mu_u \neq 0$, it is natural to ask what BMF does recover. The answer is the average rating:

Remark 13 (Average rating and aggregate welfare). The average expected rating of note j across all reviewers is $q_j + \mu_u v_j$, which equals the rating of a hypothetical reviewer with ideology μ_u. Recovering objective quality (anchoring at $\gamma = 0$) and maximizing average reviewer satisfaction (anchoring at $\gamma = \mu_u$) therefore coincide only when $\mu_u = 0$.

When $\mu_u \neq 0$, BMF's recovered ranking shifts away from objective quality and toward the ranking that maximizes average reviewer satisfaction, exactly to the latter in the $\kappa \to 0$ limit (Corollary 14). Whether this is desirable depends on whether the average reviewer is representative of the population whose welfare we care about, a normative question. The regularization weights determine where on this spectrum BMF lands:

Corollary 14 (Regularization Encodes Design Choices)**.** *As $\kappa \to \infty$ (strong quality regularization relative to user regularization), $\gamma^* \to 0$: the* ordering *of estimated qualities approaches the true quality ordering, though the absolute quality estimates are shrunk toward zero and may be uninformative. As $\kappa \to 0$ (strong user regularization), $\gamma^* \to \mu_u$: by Remark 13, the ordering matches the perspective of the average reviewer, maximizing aggregate reviewer satisfaction. Proof in Section B.2.*

[2]The hypothesis $c = b_i = 0$ can be relaxed to $c + \mu_b = 0$ and $\mathrm{Cov}(b_i, u_i) = 0$ (Section B.2); we keep the simplified form in the main statement to streamline Step 1 of the proof sketch.

The two poles of Corollary 14 are not equally robust to relaxing Theorem 10's moment conditions: the welfare pole at μ_u survives nonzero $\mathrm{Cov}(q, v)$, while the recovery pole at 0 shifts: with $\mathrm{Cov}(q, v) \neq 0$ the agent-centric balance gives $\gamma^* = \big(t^*\mu_u - A\,\mathrm{Cov}(q, v)/\sigma_v^2\big)/(t^* + A)$ in the notation of Lemma 18, whose limit is μ_u as $\kappa \to 0$ and $-\mathrm{Cov}(q, v)/\sigma_v^2$ as $\kappa \to \infty$. Figure 8 confirms the $\gamma^* \propto \mu_u$ relationship is robust across regularization strengths. Either way, the choice of κ via regularization is a design decision about whose preferences define quality.

4 Non-Identifiability: No Mechanism on Ratings Alone

The recovery condition established in Section 3 might appear specific to L_2 regularization. We now show that the gauge symmetry of Section 2.3 lives in the rating data itself, so any mechanism on ratings alone faces the same fundamental ambiguity.

Theorem 15 (Non-Identifiability of Quality)**.** *In the ground truth model, for any two notes j_1, j_2 with $q_{j_1} > q_{j_2}$ and $v_{j_1} \neq v_{j_2}$, there exists a gauge-equivalent parameterization in which $q'_{j_1} < q'_{j_2}$. The quality ordering is reversed while all ratings remain unchanged.*

Proof. Let $\delta = v_{j_2} - v_{j_1} \neq 0$. Choose γ with $\gamma\delta > q_{j_1} - q_{j_2}$. By Lemma 4: $q'_{j_2} - q'_{j_1} = (q_{j_2} - q_{j_1}) + \gamma\delta > 0$. □

Corollary 16 (No mechanism can recover quality from ratings alone)**.** *Let $M : \mathbb{R}^{n\times m} \to \mathbb{R}^m$ be any mechanism that maps a rating matrix to quality estimates. There exist ground-truth parameterizations (u, q, v) and (u', q', v') that produce identical rating matrices but with $q_{j_1} > q_{j_2}$ in one parameterization and $q'_{j_1} < q'_{j_2}$ in the other. Since M must produce the same output for both, it cannot recover the correct quality ordering in both cases. Proof in Section B.3.*

The ambiguity is not an artifact of BMF's particular algorithm but an inherent property of the rating data under the ground truth model. Without additional information about which reviewer ideology is truly neutral, the quality ordering is inherently ambiguous. Together with Corollary 11, this gives the full characterization:

Corollary 17 (Universal Quality Recovery)**.** *Under the conditions of Theorem 10, BMF recovers the correct quality ranking for every configuration (q, v) if and only if $\mu_u = 0$.*

Proof. ($\Leftarrow$) Corollary 11. ($\Rightarrow$) If $\mu_u \neq 0$ then $\gamma^* \neq 0$. Fix v with $\mu_v = 0$; since γ^* depends on the configuration only through σ_v^2, this fixes $\gamma^* \neq 0$. Take two notes with $v_{j_1} \neq v_{j_2}$, labelled so that $\gamma^*(v_{j_2} - v_{j_1}) > 0$. Any q with $\mu_q = 0$, $\mathrm{Cov}(q, v) = 0$, and $0 < q_{j_1} - q_{j_2} < \gamma^*(v_{j_2} - v_{j_1})$ then gives $\hat{q}_{j_2} > \hat{q}_{j_1}$, reversing the true order. □

5 Empirical Validation

We validate the theoretical predictions using simulations of the ground truth model with $n = 101$, $m = 100$, $d = 1$. Reviewer ideologies are drawn $u_i \sim \mathcal{N}(\mu_u, 1)$, notes from $q_j, v_j \sim \mathcal{N}(0, 1)$ independently, and ratings as $r_{ij} = q_j + u_i v_j + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \Delta)$, $\Delta = 0.25$ ($\sigma = 0.5$). BMF is fitted with L_2 regularization using the Adam optimizer [Kingma and Ba, 2015]. The anchor γ is estimated by two complementary regressions: item side ($\hat{q}_j$ on v_j) and user side (u_i on $\hat{u}_i$), which agree under correct specification. Full experimental details are in Section D.4.

γ tracks μ_u. We vary $\mu_u \in [-2, 2]$ with $\lambda_u = \lambda_q = 0.01$ ($\kappa = 1$). In Figure 3 both estimators track γ^* of Theorem 10, the root of (10), which follows $\mu_u/2$ near zero and bends below it at the extremes; at $\mu_u = 2$ the prediction is 0.80 against 0.72 (item-based) and 0.78 (user-based). Individual estimates scatter around the line at these pool sizes, a finite-sample effect quantified in Section D.6. The relationship is robust across regularization strengths, ratio κ, and observation noise levels (Section D.6).

Quality recovery degrades with $|\mu_u|$. Figure 4 shows that when $\mu_u = 0$, BMF achieves near-optimal quality recovery. As $|\mu_u|$ increases, quality is increasingly influenced by ideological alignment: notes leaning in the direction of the average reviewer are overrated.

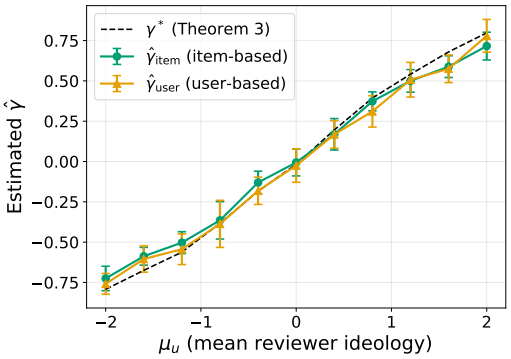


Figure 3: Estimated anchor γ vs. mean reviewer ideology μ_u ($\kappa = 1$). Item-based ($\hat{\gamma}_{\text{item}}$, circles) and user-based ($\hat{\gamma}_{\text{user}}$, triangles) estimators (Section D) both track γ^* of Theorem 10 (dashed; the root of (10) at each replication's empirical moments, averaged within each μ_u), which follows $\mu_u/2$ near zero and bends below it at the extremes. Error bars: ± 1 SD over 10 replications.

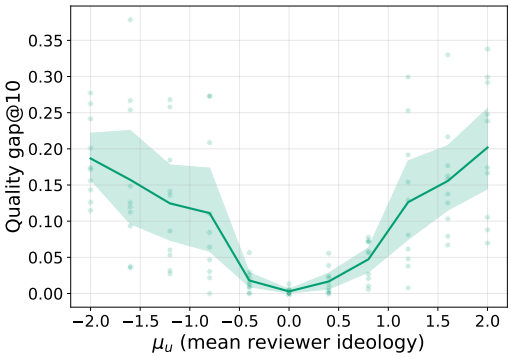


Figure 4: Quality gap@10 (top-10 quality vs. oracle, Section D) versus μ_u. Near-zero at $\mu_u = 0$; grows monotonically with $|\mu_u|$. 10 replications (± 1 SE shaded).

Misspecified utilities. We also generate ratings from three departures from the bilinear form: a Euclidean utility $-(u_i - v_j)^2$, a higher-order interaction $u_i v_j + \varphi\, u_i^2 v_j^2$, and an unmodeled confounder $u_i v_j + w_i z_j$. The anchor tracks μ_u under all three. Quality recovery at $\mu_u = 0$ survives the confounder and fails under the Euclidean utility and the higher-order interaction, which the theory allows since Corollary 11 is proved for the bilinear model. These are preliminary checks on three specific departures. Details in Section D.8.

Binary ratings. Community Notes uses categorical (helpful/not helpful) ratings. We test robustness by passing scores through a logistic link and sampling Bernoulli ratings. The user-based estimate tracks the curve of Theorem 10 (0.77 against 0.79 at $\mu_u = 2$), while the item-based estimate compresses at large $|\mu_u|$ (0.57), likely due to saturation of the logistic link. This indicates that the anchoring to the population mean persists in the categorical setting used by Community Notes. Additional details in Section D.7.

6 Discussion and Conclusion

Under BMF, the decomposition of ratings into "quality" and "ideology" is not unique: shifting all reviewer ideologies by a constant while adjusting quality scores leaves every rating unchanged. BMF's regularization resolves this ambiguity by anchoring at a constant times the mean reviewer ideology (Theorem 10), recovering true quality if and only if that mean is zero. Like the Condorcet Jury Theorem, our result identifies conditions for when a collective judgment procedure recovers truth. This limitation is not an L_2 artifact: no mechanism based on ratings alone can do better (Corollary 16). The rest of this section draws out what these results imply for platform design and the broader question of what Community Notes is actually optimizing.

Indeterminacy and subjective vs. objective interpretations. The gauge symmetries we identify reveal a core indeterminacy: ratings are consistent with the platform recovering an objective ground truth *or* aggregating subjective preferences, depending solely on which ideology is declared neutral. Both interpretations map onto the same mathematical structure, differing only in the anchor. Our model assumes a single ground truth and asks whether BMF recovers it, but the indeterminacy suggests a more general framing in which there is no uniquely privileged vantage point. One could instead model quality as a family of equally valid standards, each anchored to a different perspective, with the platform's problem recast as choosing among them according to an explicit social welfare criterion [Arrow, 2012, Boutilier et al., 2015]. Nonetheless, our analysis reveals an asymmetry between these readings: the welfare anchor at μ_u is robust to moment conditions on (q, v), while the truth-recovery

anchor at 0 requires them. In this sense BMF natively performs preference aggregation, with truth recovery available as a special case.

Implications for Community Notes. A skewed reviewer pool (likely, given self-selection [Allen et al., 2022]) means quality estimates inherit that skew: notes aligned with the majority ideology will rank higher not because they are better, but because BMF's anchor drifts toward the mean reviewer. Empirical assessments of "helpfulness" [Wojcik et al., 2022] carry the same confound, which follows from using ratings alone (Corollary 16). The platform therefore faces a normative choice: whose perspective should define quality? External information could help where ratings alone cannot: expert-labeled ground truth examples fix the anchor to the expert's perspective [Zhao and Naaman, 2023]; requiring high approval from both sides of the spectrum [Ovadya and Thorburn, 2023] can be read in the model as requiring $r_{ij} = q_j + u_i v_j$ to be high for reviewers with $u_i < 0$ and $u_i > 0$ alike, which holds only for notes with high q_j and small $|v_j|$, whose estimate $q_j + \gamma^* v_j$ then barely depends on the anchor; and note-level markers of quality, such as whether a note cites sources, pin q_j directly, which no gauge shift can absorb.

Limitations and future work. We assume $d = 1$, matching Community Notes but potentially missing higher-dimensional disagreement. The non-identifiability extends to any d (Theorem 15), while characterizing the anchor there needs a separate analysis, since the balance equations become matrix equations. We focus on the correctly specified setting; a general treatment of misspecification remains open. We also analyze the core factorization on its own; a deployment adds missing ratings, weighting, and display thresholds on top of it, and filtering which reviewers count moves μ_u and with it the anchor. We analyze a static snapshot; dynamic effects including reviewer entry and exit and feedback loops [Akpinar et al., 2022] are left to future work. Developing adversarially robust BMF variants [Schoenebeck et al., 2021], using richer data to better disentangle quality from ideology, validating on real Community Notes data, and fully characterizing strategic incentives under BMF are important next steps. More broadly, our framework applies to any system learning latent representations to aggregate opinions: making normative and design choices explicit is a prerequisite for accountable algorithmic governance.

Acknowledgments and Disclosure of Funding

We thank Maggie Makar for helpful suggestions on the theoretical approach. This work was supported by the National Science Foundation under awards 2313137 and 2208662. The authors declare no competing interests.

References

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Properties from Mechanisms: An Equivariance Perspective on Identifiable Representation Learning, October 2021.

Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. Long-term Dynamics of Fairness Intervention in Connection Recommender Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 22–35, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534173.

Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–19, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502040.

T. W. Anderson and Herman Rubin. Statistical Inference in Factor Analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, volume 3.5, pages 111–151. University of California Press, January 1956.

Kenneth J. Arrow. *Social Choice and Individual Values: Third Edition*. Yale University Press, June 2012. ISBN 978-0-300-18698-7.

Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465, February 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/11-AOS966.

Robert M. Bell and Yehuda Koren. Lessons from the Netflix prize challenge. *SIGKDD Explor. Newsl.*, 9(2):75–79, December 2007. ISSN 1931-0145. doi: 10.1145/1345448.1345465.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. *Artificial Intelligence*, 227:190–213, October 2015. ISSN 0004-3702. doi: 10.1016/j.artint.2015.06.003.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. Introduction to Computational Social Choice. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 1–29. Cambridge University Press, 2016. doi: 10.1017/CBO9781107446984.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, February 2023.

Joshua Clinton, Simon Jackman, and Douglas Rivers. The Statistical Analysis of Roll Call Data. *American Political Science Review*, 98(2):355–370, May 2004. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055404001194.

Giulio Corsi. Evaluating Twitter's Algorithmic Amplification of Low-Credibility Content: An Observational Study, September 2023.

Alexander D'Amour. On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives, March 2019.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, Paris, 1785.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, J. Ekstrand, Oghenemaro Anuyah, David McNeill, and M. S. Pera. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *FAT*, January 2018.

Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24)*, page 985. Association for Computing Machinery, 2024. doi: 10.1145/3670865.3673547.

Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. Explainable Fairness in Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 681–691, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-8732-3. doi: 10.1145/3477495.3531973.

Allan Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–601, 1973. ISSN 0012-9682. doi: 10.2307/1914083.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 173–182, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052569.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. ISSN 1558-0814. doi: 10.1109/MC.2009.263.

Joshua Kroll, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu. Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3):633, January 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, May 2019.

Meta. Testing begins for community notes on Facebook, Instagram and Threads. Meta Newsroom, `https://about.fb.com/news/2025/03/testing-begins-community-notes-facebook-instagram-threads/`, March 2025.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287596.

Aviv Ovadya and Luke Thorburn. Bridging Systems: Open Problems for Countering Destructive Divisiveness across Ranking, Recommenders, and Governance. arXiv:2301.09976, July 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 39643–39666, Vienna, Austria, July 2024. JMLR.org.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 8026–8037, Red Hook, NY, USA, December 2019. Curran Associates Inc.

Royal Pathak and Francesca Spezzano. Analyzing the Interplay between Diversity of News Recommendations and Misinformation Spread in Social Media. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '24, pages 80–85, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0466-6. doi: 10.1145/3631700.3664870.

Keith T. Poole and Howard Rosenthal. A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384, 1985. ISSN 0092-5853. doi: 10.2307/2111172.

Tianyi Qiu. Representative Social Choice: From Learning Theory to AI Alignment. arXiv:2410.23953, October 2025.

Sarah T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. ISBN 978-0-300-23588-3. doi: 10.2307/j.ctvhrcz0v.

Grant Schoenebeck, Fang-Yi Yu, and Yichi Zhang. Information Elicitation from Rowdy Crowds. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3974–3986, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3449840.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *International Conference on Learning Representations (ICLR)*, 2024.

Chris Vallance. Meta wants X-style community notes to replace fact checkers - can it work? *BBC*, January 2025.

Chenlong Wang and Pablo Lucas. Efficiency of Community-Based Content Moderation Mechanisms: A Discussion Focused on Birdwatch. *Group Decision and Negotiation*, 33(3):673–709, June 2024. ISSN 1572-9907. doi: 10.1007/s10726-024-09881-1.

Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, M. B. Fallin Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. *arXiv preprint arXiv:2210.15723*, 2022. doi: 10.48550/ARXIV.2210.15723.

Andy Zhao and Mor Naaman. Insights from a Comparative Study on the Variety, Velocity, Veracity, and Viability of Crowdsourced and Professional Fact-Checking Services. *Journal of Online Trust and Safety*, 2(1), September 2023. ISSN 2770-3142. doi: 10.54501/jots.v2i1.118.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency, March 2025.

A Notation Table

For convenience, we collect the paper-specific shorthand used across sections. Symbols defined at their only meaningful use site (e.g., $\hat{r}_{ij}, \varepsilon_{ij}, \perp$), and standard mathematical notation, are not listed here. The full BMF and ground-truth parameters appear in Definitions 1 and 2; the gauge family in Theorem 5.

Symbol	Meaning
Empirical statistics of the reviewer/note pool	
μ_u, μ_v	Mean reviewer ideology and mean note ideology, $\mu_u = \frac{1}{n}\sum_i u_i$, etc.
σ_u^2, σ_v^2	Empirical variances of u_i and v_j.
$\rho = m/n$	Note-to-reviewer ratio.
Loss decomposition (Definition 1)	
$R(\hat{\theta})$	Squared-error risk; the total loss is $\mathcal{L} = R + \Lambda$.
$\Lambda_{L_2}(\hat{\theta})$	L_2 regularizer with per-block weights $\lambda_u, \lambda_q, \lambda_v, \lambda_b, \lambda_c$ (Section 3).
$\lambda > 0$	Overall regularization strength (common factor on all five weights, Proposition 8).
$\kappa = \lambda_q/\lambda_u$	Quality-to-user weight ratio.
Δ	Noise variance ($\mathbb{E}[\varepsilon_{ij}^2] = \Delta$); noiseless setting is $\Delta = 0$.
Gauge parameters (Theorem 5, Equation (5)*)*	
γ	Agent-centric shift; the only ranking-relevant gauge parameter (Proposition 6).
η	Item-centric shift.
τ	Rank-1 rescaling factor; $\tau > 0$ by the convention of Definition 7.
α, β	Additive offsets absorbed into intercepts and quality.
$\mathcal{G}(\theta_{\text{true}})$	Gauge orbit: the parameterizations predicting exactly as the true model, which minimize the risk for every Δ and achieve zero risk when $\Delta = 0$.
Solution-related notation	
θ^*_Λ	Min-min solution: the element of the gauge orbit minimizing Λ (Definition 7).
x^*	For any quantity x derived from θ, x^* denotes its value at θ^*_Λ.
$\gamma^* = s\,\mu_u$, $s \in (0,1)$	Anchor selected by L_2 regularization (Theorem 10).
Reduced system (Lemmas 18, 19 and 23)	
$t = \tau^2$	Squared scale parameter, the variable of the reduced scalar equation.
$A = \kappa\rho\sigma_v^2$	Damping constant in $\gamma(t) = \mu_u t/(t + A)$.
$K = (\lambda_v/\lambda_u)\rho\sigma_v^2$	Right side of the reduced scalar equation (10).
ψ	Left side of (10); strictly increasing from 0 to ∞.
w	Moments of the base point, $(\mu_u, \sigma_u^2, \sigma_v^2, \mathrm{Cov}(q, v))$.
ζ	Deviation of the fitted moments from the true ones.

Note on overloaded σ. In the experimental appendices, σ denotes the observation-noise standard deviation (so $\Delta = \sigma^2$), while $\sigma(\,\cdot\,)$ denotes the logistic link in the binary-rating extension. The two uses are distinguished by the presence of an argument.

B Additional Proofs

B.1 Gauge Symmetries (Section 2.3)

Theorem 5 (Characterization of BMF's Gauge Symmetries)**.** *Every gauge transformation of BMF (i.e., every parameter mapping that preserves all predicted ratings $\hat{r}_{ij}$) can be parameterized by $(\alpha, \beta, \gamma, \eta, \tau)$ with $\alpha, \beta, \gamma, \eta \in \mathbb{R}$ and $\tau \in \mathbb{R} \setminus \{0\}$:*

$$\begin{aligned} \hat{u}'_i &= \tau(\hat{u}_i - \gamma), \qquad \hat{q}'_j = \hat{q}_j + \gamma\hat{v}_j + \beta, \\ \hat{v}'_j &= \hat{v}_j/\tau - \eta/\tau, \qquad \hat{b}'_i = \hat{b}_i + \hat{u}_i\eta + \alpha, \qquad \hat{c}' = \hat{c} - \alpha - \beta - \gamma\eta. \end{aligned} \tag{5}$$

These five parameters exhaust all degrees of non-identifiability.

Proof. We prove the result for $d = 1$; the general case follows componentwise.

Consider any gauge transformation, assumed defined on all of parameter space and twice differentiable, satisfying

$$\hat{c}' + \hat{b}'_i + \hat{q}'_j + \hat{u}'_i\hat{v}'_j = \hat{c} + \hat{b}_i + \hat{q}_j + \hat{u}_i\hat{v}_j \tag{9}$$

for all $i \in [n]$ and $j \in [m]$.

Step 1: Affine structure. Since the right-hand side is linear in $\hat{v}_j$, taking $\partial^2/\partial\hat{v}_j^2$ of the left-hand side yields $\hat{u}'_i \frac{\partial^2 \hat{v}'_j}{\partial \hat{v}_j^2} + 2\frac{\partial \hat{u}'_i}{\partial \hat{v}_j}\frac{\partial \hat{v}'_j}{\partial \hat{v}_j} = 0$. For this to hold for all possible values of $\hat{u}'_i$, we need $\frac{\partial^2 \hat{v}'_j}{\partial \hat{v}_j^2} = 0$, so $\hat{v}'_j = A_j\hat{v}_j + B_j$. By symmetry, $\hat{u}'_i = C_i\hat{u}_i + D_i$.

Step 2: Universal coefficients. Substituting the affine forms into (9) and expanding:

$$\hat{c}' + \hat{b}'_i + \hat{q}'_j + C_iA_j\hat{u}_i\hat{v}_j + C_iB_j\hat{u}_i + D_iA_j\hat{v}_j + D_iB_j = \hat{c} + \hat{b}_i + \hat{q}_j + \hat{u}_i\hat{v}_j.$$

Taking $\frac{\partial}{\partial \hat{u}_i}$: $\frac{\partial \hat{b}'_i}{\partial \hat{u}_i} + C_iA_j\hat{v}_j + C_iB_j = \hat{v}_j$. For this to hold for all $\hat{v}_j$, we need $C_iA_j = 1$ for all i, j. Considering two items: $C_iA_{j_1} = 1 = C_iA_{j_2}$ implies $A_{j_1} = A_{j_2}$. Thus $A_j = A$ and $C_i = 1/A := C$ are universal. By the symmetric argument (differentiating with respect to $\hat{v}_j$), $B_j = B$ and $D_i = D$ are also universal.

Step 3: Parametrization. With $\hat{u}'_i = C\hat{u}_i + D$ and $\hat{v}'_j = \frac{1}{C}\hat{v}_j + B$, set $\tau = C$, $\gamma = -D/C$, $\eta = -\tau B$. The constant $DB = \gamma\eta$ from the product $\hat{u}'_i\hat{v}'_j$ goes into $\hat{c}'$, and the remaining terms into α, β, which yields the five-parameter family in (5). The total degrees of freedom are $3 + 2d = 5$ for $d = 1$. □

Proposition 6 (Only γ Affects Quality Ordering)**.** *The relative ordering of $\hat{q}_j$ is invariant under the item-centric shift (η), scaling (τ), and global offsets (α, β). Only the agent-centric shift γ can change the ordering.*

Proof. By the gauge invariance property, any gauge transformation satisfies $\hat{c} + \hat{b}_i + \hat{q}_j + \hat{u}_i\hat{v}_j = \hat{c}' + \hat{b}'_i + \hat{q}'_j + \hat{u}'_i\hat{v}'_j$ for all i, j.

Since $\hat{q}'_j$ depends only on j (not on i), rearranging shows $\hat{q}'_j$ must have the form $\hat{q}'_j = \hat{q}_j + \langle\omega, \hat{v}_j\rangle + \beta$ for some $\omega \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$ (by bilinearity of the interaction term and independence from agent index).

Now examine each basic gauge transformation:

Agent-centric transformation: $\hat{q}_j \mapsto \hat{q}_j + \gamma\hat{v}_j$. This gives $\omega = \gamma$, which is nonzero for $\gamma \neq 0$.

Item-centric transformation: $\hat{v}_j \mapsto \hat{v}_j - \eta$, $\hat{b}_i \mapsto \hat{b}_i + \hat{u}_i\eta$, $\hat{q}_j \mapsto \hat{q}_j$. Quality is unchanged, so $\omega = 0$.

Scaling: $\hat{u}_i \mapsto \tau\hat{u}_i$, $\hat{v}_j \mapsto \hat{v}_j/\tau$. This preserves $\hat{u}_i\hat{v}_j$ and does not alter $\hat{q}_j$, so $\omega = 0$.

Global offsets: $\hat{q}_j \mapsto \hat{q}_j + \beta$. This shifts all qualities by a constant, so $\omega = 0$.

Therefore, among the gauge transformations, only the agent-centric shift with $\gamma \neq 0$ produces $\omega \neq 0$ and can affect the relative ordering of qualities. □

B.2 Quality Recovery (Section 3)

Lemma 9 (Balance Equations for L_2)**.** *Any minimizer of the L_2-regularized BMF objective satisfies:*

$$\begin{aligned} \lambda_u \textstyle\sum_i \hat{u}_i = \lambda_q \sum_j \hat{q}_j \hat{v}_j \quad &(\textit{Agent-centric}) & \lambda_u \textstyle\sum_i \hat{u}_i^2 = \lambda_v \sum_j \hat{v}_j^2 \quad &(\textit{Scale}) \\ \lambda_b \textstyle\sum_i \hat{b}_i = \lambda_q \sum_j \hat{q}_j = \lambda_c \hat{c} \quad &(\textit{Intercept}) & \lambda_b \textstyle\sum_i \hat{b}_i \hat{u}_i = \lambda_v \sum_j \hat{v}_j \quad &(\textit{Item-centric}) \end{aligned} \tag{7}$$

Proof. For each gauge symmetry, the risk is invariant along the corresponding one-parameter family. Since the total objective $R + \lambda\Lambda$ must be stationary at the minimum, and the risk contribution is flat along the orbit, the derivative of Λ alone must vanish in each gauge direction.

For the agent-centric shift $\hat{u}_i \mapsto \hat{u}_i - \gamma$, $\hat{q}_j \mapsto \hat{q}_j + \gamma\hat{v}_j$: the γ-dependent regularization is $\Lambda(\gamma) = \lambda_u \sum_i (\hat{u}_i - \gamma)^2 + \lambda_q \sum_j (\hat{q}_j + \gamma\hat{v}_j)^2 + \text{const}$. Setting $\frac{d\Lambda}{d\gamma}\big|_{\gamma=0} = 0$ yields (Agent-centric).

For the scaling $\hat{u}_i \mapsto \tau\hat{u}_i$, $\hat{v}_j \mapsto \hat{v}_j/\tau$: the τ-dependent part is $\Lambda(\tau) = \lambda_u \sum_i (\tau\hat{u}_i)^2 + \lambda_v \sum_j (\hat{v}_j/\tau)^2$. Setting $\frac{d\Lambda}{d\tau}\big|_{\tau=1} = 0$ yields (Scale).

For the intercept shift $\hat{b}_i \mapsto \hat{b}_i + \alpha$, $\hat{q}_j \mapsto \hat{q}_j + \beta$, $\hat{c} \mapsto \hat{c} - \alpha - \beta$: differentiating with respect to α and β at zero yields (Intercept).

For the item-centric shift $\hat{b}_i \mapsto \hat{b}_i + \hat{u}_i\eta$, $\hat{v}_j \mapsto \hat{v}_j - \eta$: differentiating at $\eta = 0$ yields (Item-centric). □

Theorem 10 (BMF Anchors proportional to μ_u)**.** *Consider ratings generated by the simplified ground truth model ($c = b_i = 0$[3], Definition 2) with mean reviewer ideology μ_u and variance $\sigma_u^2 > 0$, zero-mean notes ($\mu_q = \mu_v = 0$), and $\mathrm{Cov}(q_j, v_j) = 0$. Under L_2 regularization, as $\lambda \to 0$, BMF recovers $\hat{q}_j = q_j + \gamma^* v_j + \beta$ where β is a ranking-irrelevant constant and*

$$\gamma^* = \frac{\mu_u}{1 + \kappa\rho\sigma_v^2/\tau^{*2}}, \qquad \tau^{*4} = \frac{\lambda_v m \sigma_v^2}{\lambda_u n[\sigma_u^2 + (\mu_u - \gamma^*)^2]}, \tag{8}$$

with $\kappa = \lambda_q/\lambda_u$ and $\rho = m/n$. In particular, $\gamma^* = s\mu_u$ for some $s \in (0, 1)$.

Proof. We characterize the min-min solution under the ground truth model with L_2 regularization. The proof proceeds in three steps.

Step 1: Min-min characterization under the ground truth model

Under the simplified ground truth model with $\mu_q = \mu_v = 0$ and $\mathrm{Cov}(q, v) = 0$, the min-min solution under L_2 regularization satisfies:

$$\lambda_b\big[\tau n(\mu_u - \gamma)\eta + n\alpha\big] = \lambda_q m\beta = -\lambda_c(\alpha + \beta) \tag{S1}$$

$$\lambda_u\tau^2 n(\mu_u - \gamma) = \lambda_q m(\gamma\sigma_v^2 - \beta\eta) \tag{S2}$$

$$\lambda_u\tau^2 n[\sigma_u^2 + (\mu_u - \gamma)^2] = \frac{\lambda_v m}{\tau^2}[\sigma_v^2 + \eta^2] \tag{S3}$$

These follow from substituting the gauge transformation $\hat{u}_i = \tau(u_i - \gamma)$, $\hat{q}_j = q_j + \gamma v_j + \beta$, $\hat{v}_j = v_j/\tau - \eta/\tau$, $\hat{b}_i = u_i\eta + \alpha$, $\hat{c} = -\alpha - \beta - \gamma\eta$ into the balance equations of Lemma 9, then applying the moment assumptions.

Step 2: $\eta = \alpha = \beta = 0$ at the minimum

By the law of iterated minimization: $\min_{\gamma,\eta,\tau} \Lambda(\gamma, \eta, \tau) = \min_\tau[\min_{\gamma,\eta} \Lambda(\gamma, \eta, \tau)]$.

For fixed τ, the gauge orbit is affine in (γ, η) and Λ is strictly convex, giving a unique inner minimizer. From (S1), α and β are linear in η: $\alpha(\eta) = k_1\eta$ and $\beta(\eta) = k_2\eta$ for constants k_1, k_2. Substituting, the η-dependent part of Λ is quadratic:

$$\Lambda(\eta) = \Lambda(0) + \frac{2\lambda_v m\mu_v}{\tau^2}\eta + C\eta^2$$

[3]The hypothesis $c = b_i = 0$ can be relaxed to $c + \mu_b = 0$ and $\mathrm{Cov}(b_i, u_i) = 0$ (Section B.2); we keep the simplified form in the main statement to streamline Step 1 of the proof sketch.

where $C > 0$. Since $\mu_v = 0$ under the ground truth model, the linear term vanishes and $\eta^* = 0$ is the unique minimizer for every fixed τ. Therefore $\eta = 0$ at the joint minimum. Since $b_i = c = 0$ in the ground truth, equation (S1) then gives $\alpha = \beta = 0$.

Remark (weaker conditions). The conclusion above extends to ground-truth models with $c + \mu_b = 0$ and $\mathrm{Cov}(b_i, u_i) = 0$, without requiring $c = b_i = 0$ pointwise. Under these conditions the inner minimizer becomes $(\alpha^*, \beta^*, \eta^*) = (-\mu_b, 0, 0)$ rather than $(0, 0, 0)$: the α-shift absorbs the mean intercept into $\hat{c}$ (a residual gauge freedom under $c + \mu_b = 0$), while $\mathrm{Cov}(b_i, u_i) = 0$ removes the cross-term in the η balance that arbitrary b_i would otherwise contribute. The reduced system (S2')–(S3') for (γ, τ) contains no b- or c-dependent terms, so γ remains a fraction of μ_u.

Step 3: Solving the reduced system

With $\eta = \alpha = \beta = 0$, the system reduces to:

$$\lambda_u \tau^2 n(\mu_u - \gamma) = \lambda_q m \gamma \sigma_v^2 \tag{S2'}$$

$$\lambda_u \tau^2 n[\sigma_u^2 + (\mu_u - \gamma)^2] = \frac{\lambda_v m \sigma_v^2}{\tau^2} \tag{S3'}$$

Solving for γ: Rearranging (S2'): $\lambda_u \tau^2 n \mu_u = \gamma(\lambda_u \tau^2 n + \lambda_q m \sigma_v^2)$, giving $\gamma = \mu_u / (1 + \kappa\rho\sigma_v^2/\tau^2)$.

Solving for τ: Rearranging (S3'): $\tau^4 = \lambda_v m \sigma_v^2 / (\lambda_u n[\sigma_u^2 + (\mu_u - \gamma)^2])$.

Existence, uniqueness, and $s \in (0, 1)$: Substituting $\gamma(\tau^2)$ into (S3') leaves a single equation in τ^2 whose left side is continuous and strictly increasing from 0 to ∞. It has exactly one root, which gives the unique fixed point (γ^*, τ^*) (Lemma 18). Writing $\gamma^* = s\mu_u$: $s = 1/(1 + \kappa\rho\sigma_v^2/\tau^{*2})$. Since $\kappa, \rho, \sigma_v^2, \tau^{*2} > 0$, the denominator exceeds 1, so $s \in (0, 1)$.

Quality rankings: By Proposition 6, only γ affects the ordering of estimated qualities $\hat{q}_j = q_j + \gamma v_j$. The scale τ does not enter $\hat{q}_j$. □

Solving the reduced system

Throughout, $t := \tau^2$ and

$$A := \kappa\rho\sigma_v^2 = \frac{\lambda_q}{\lambda_u}\frac{m}{n}\sigma_v^2 > 0, \qquad K := \frac{\lambda_v}{\lambda_u}\frac{m}{n}\sigma_v^2 > 0.$$

Lemma 18 (Uniqueness of the reduced system)**.** *Assume $\sigma_u^2 > 0$ and $\mu_u > 0$. The system* (S2')–(S3') *has exactly one solution $(\gamma^*, \tau^*) \in (0, \mu_u) \times (0, \infty)$. It is given by $\gamma^* = \mu_u t^*/(t^* + A)$, where t^* is the unique root of*

$$\psi(t) := t^2\sigma_u^2 + \mu_u^2 A^2 \Big(\frac{t}{t + A}\Big)^2 = K, \tag{10}$$

and $\psi'(t) > 0$ for every $t > 0$. For $\mu_u < 0$ the solution is the sign reflection of the one for $|\mu_u|$; for $\mu_u = 0$ it is $\gamma^* = 0$ with $\tau^{*4} = K/\sigma_u^2$.

Proof. Equation (S2') is linear in γ. Dividing by $\lambda_u n$ and using $\lambda_q m \sigma_v^2 = \lambda_u n A$ gives $t(\mu_u - \gamma) = A\gamma$, so

$$\gamma(t) = \frac{\mu_u t}{t + A}, \qquad \mu_u - \gamma(t) = \frac{\mu_u A}{t + A}, \tag{11}$$

and γ increases strictly from $\gamma(0) = 0$ to μ_u. Equation (S3') states $t^2[\sigma_u^2 + (\mu_u - \gamma)^2] = K$; substituting (11) reduces the system to the scalar equation (10).

Both terms of ψ increase strictly on $(0, \infty)$, because $t \mapsto t^2$ and $t \mapsto t/(t + A)$ do, and

$$\psi'(t) = 2t\sigma_u^2 + \frac{2\mu_u^2 A^3 t}{(t + A)^3} > 0.$$

Also $\psi(0) = 0$, and $\psi(t) \to \infty$ through its first term, which uses $\sigma_u^2 > 0$. A continuous strictly increasing function from 0 to ∞ takes the value $K > 0$ exactly once, so (10) has a unique root $t^* \in (0, \infty)$ and $\gamma^* = \gamma(t^*) \in (0, \mu_u)$ is the unique solution of the system. Finally $s = \gamma^*/\mu_u = t^*/(t^* + A) \in (0, 1)$, which is the form appearing in (8).

For $\mu_u < 0$, ψ depends on μ_u only through μ_u^2, so t^* is unchanged while $\gamma(t)$ changes sign with μ_u. For $\mu_u = 0$, (11) gives $\gamma \equiv 0$ and (10) becomes $t^2\sigma_u^2 = K$. □

Lemma 19 (Nonsingularity of the Jacobian)**.** *Write the reduced system as $F(\gamma, t) = 0$ with*

$$F_1(\gamma,t) = \lambda_u nt(\mu_u - \gamma) - \lambda_q m\gamma\sigma_v^2, \qquad F_2(\gamma,t) = \lambda_u nt\big[\sigma_u^2 + (\mu_u-\gamma)^2\big] - \frac{\lambda_v m\sigma_v^2}{t}.$$

At the solution of Lemma 18,

$$\det J(\gamma^*, t^*) = -\frac{(\lambda_u n)^2(t^*+A)}{t^*}\,\psi'(t^*) < 0.$$

The same conclusion holds in the coordinates (γ, τ), since $\tau \mapsto \tau^2$ is a diffeomorphism of $(0,\infty)$.

Proof. Using $\lambda_q m\sigma_v^2 = \lambda_u nA$, the partial derivatives are

$$\frac{\partial F_1}{\partial \gamma} = -\lambda_u n(t+A), \qquad \frac{\partial F_1}{\partial t} = \lambda_u n(\mu_u - \gamma),$$
$$\frac{\partial F_2}{\partial \gamma} = -2\lambda_u nt(\mu_u-\gamma), \qquad \frac{\partial F_2}{\partial t} = \lambda_u n\big[\sigma_u^2 + (\mu_u-\gamma)^2\big] + \frac{\lambda_v m\sigma_v^2}{t^2}.$$

Let $\gamma_1(t) = \mu_u t/(t+A)$ be the curve $F_1 = 0$ from (11), and set $g(t) := F_2(\gamma_1(t), t)$. Factoring out $\lambda_u n/t$,

$$g(t) = \frac{\lambda_u n}{t}\Big(t^2\big[\sigma_u^2 + (\mu_u - \gamma_1(t))^2\big] - K\Big) = \frac{\lambda_u n}{t}\big(\psi(t) - K\big).$$

Since $F_1(\gamma_1(t), t) \equiv 0$, implicit differentiation gives $\gamma_1'(t) = -(\partial F_1/\partial t)/(\partial F_1/\partial\gamma)$, and the chain rule gives $g'(t) = (\partial F_2/\partial\gamma)\,\gamma_1'(t) + \partial F_2/\partial t$. Substituting γ_1' and collecting over $\partial F_1/\partial\gamma$,

$$\det J = \frac{\partial F_1}{\partial\gamma}\frac{\partial F_2}{\partial t} - \frac{\partial F_1}{\partial t}\frac{\partial F_2}{\partial\gamma} = \frac{\partial F_1}{\partial\gamma}\,g'(t).$$

At $t = t^*$ the identity $\psi(t^*) = K$ removes the derivative of the prefactor, leaving $g'(t^*) = \lambda_u n\,\psi'(t^*)/t^*$. With $\partial F_1/\partial\gamma = -\lambda_u n(t^*+A)$ and $\psi'(t^*) > 0$ from Lemma 18, the determinant takes the stated form. □

Remark 20. The negative sign states that the two curves $F_1 = 0$ and $F_2 = 0$ cross transversally at (γ^*, t^*): the rows of J are their normals, and parallel normals would force $\psi'(t^*) = 0$. Strict monotonicity of ψ is the only input, and Lemma 18 establishes it before it establishes uniqueness.

Lemma 21 (Noise leaves the population anchor fixed)**.** *Let $p_{ij}(\theta)$ denote the prediction of BMF at parameters θ, and let $\mathcal{L}_0(\theta) = \sum_{i,j}\big(p_{ij}(\theta_{\mathrm{true}}) - p_{ij}(\theta)\big)^2$. For ratings $r_{ij} = p_{ij}(\theta_{\mathrm{true}}) + \varepsilon_{ij}$ with ε_{ij} i.i.d. of mean zero and variance Δ, drawn independently of θ,*

$$\mathbb{E}\Big[\sum_{i,j}\big(r_{ij} - p_{ij}(\theta)\big)^2\Big] = \mathcal{L}_0(\theta) + nm\Delta \qquad \textit{for every } \theta.$$

Hence for every $\Delta \geq 0$ the minimizers of the expected risk are exactly $\mathcal{G}(\theta_{\mathrm{true}})$, the min-min problem is the one solved in Theorem 10, and the anchor is $\gamma^* = s\mu_u$.

Proof. Expanding the square,

$$\sum_{i,j}\big(r_{ij} - p_{ij}(\theta)\big)^2 = \mathcal{L}_0(\theta) + 2\sum_{i,j}\varepsilon_{ij}\big(p_{ij}(\theta_{\mathrm{true}}) - p_{ij}(\theta)\big) + \sum_{i,j}\varepsilon_{ij}^2.$$

The middle term has mean zero, and the last has mean $nm\Delta$ and does not depend on θ, so the expected loss differs from $\mathcal{L}_0$ by a constant. The risk is therefore minimized exactly when $p(\theta) = p(\theta_{\mathrm{true}})$, which is the definition of $\mathcal{G}(\theta_{\mathrm{true}})$. Noise raises the risk floor from 0 to Δ per entry and leaves the minimizing set unchanged. Proposition 8 then applies with the excess risk $R - nm\Delta$ in place of the risk. Neither the constraint set nor Λ_{L_2} depends on Δ, so Steps 1 to 3 above apply verbatim and return the same γ^*. □

Remark 22. The empirical loss depends on θ only through the predictions $p(\theta)$, which are constant along gauge orbits (Lemma 4). For every realization of the noise it is therefore constant on each orbit, and its derivative in any gauge direction is zero. A minimizer of the regularized objective minimizes Λ along its own orbit, so the balance equations (7) hold exactly, with no noise correction. Noise changes which orbit the fit lands on while leaving its position along that orbit fixed. Lemma 23 bounds the displacement of the anchor that results.

Lemma 23 (Displacement of the anchor at finite sample size)**.** *Steps 1 to 3 use the base point only through the moments $w := (\mu_u, \sigma_u^2, \sigma_v^2, \mathrm{Cov}(q, v))$, so write the reduced system as $F(\gamma, t; w) = 0$, with the additional term $-\lambda_q m \,\mathrm{Cov}(q, v)$ in F_1 and F otherwise as in Lemma 19. Let $w_0 = (\mu_u, \sigma_u^2, \sigma_v^2, 0)$ denote the true moments, and suppose the fit lies on the gauge orbit of a base point whose moments $\tilde{w}$ satisfy $\|\tilde{w} - w_0\| \le \zeta$. There are $\zeta_0 > 0$ and $c_J < \infty$, depending only on w_0 and the regularization weights, such that for $\zeta < \zeta_0$ the system $F(\gamma, t; \tilde{w}) = 0$ has a solution $(\gamma_\Delta, t_\Delta)$, unique in a neighborhood of (γ^*, t^*), with*

$$|\gamma_\Delta - \gamma^*| + |t_\Delta - t^*| \le c_J\, \zeta.$$

Proof. F is a polynomial in (γ, w) and in t and t^{-1}, hence continuously differentiable on $(0, \mu_u) \times (0, \infty) \times \mathbb{R}^4$, and $F(\gamma^*, t^*; w_0) = 0$ by Lemma 18. Its Jacobian in (γ, t) at that point is the matrix of Lemma 19, invertible with $\|J^{-1}\|$ bounded through the closed form of $\det J$. The implicit function theorem gives a neighborhood on which the perturbed system has a unique root, with

$$\begin{pmatrix} \gamma_\Delta - \gamma^* \\ t_\Delta - t^* \end{pmatrix} = -J^{-1} \left.\frac{\partial F}{\partial w}\right|_{(\gamma^*, t^*, w_0)} (\tilde{w} - w_0) + o(\zeta),$$

so $c_J = \|J^{-1}\|\, \|\partial F / \partial w\| + o(1)$ suffices.

Under a perturbed base point μ_v is no longer exactly zero, so Step 2 gives $\eta^*, \alpha^*, \beta^* = O(\zeta)$ in place of 0. These enter (S2)–(S3) only through $\beta\eta$ and η^2, contributing $O(\zeta^2)$, below the order of the remainder. □

Corollary 11 (Quality Recovery under Neutrality)**.** *Under the conditions of Theorem 10, if $\mu_u = 0$ then BMF recovers the correct quality ranking for every configuration (q, v). Proof in Section B.2.*

Proof. $\mu_u = 0 \Rightarrow \gamma^* = 0$ by (8), so $\hat{q}_j = q_j + \beta$ and the ranking is preserved. □

Corollary 12 (Recovery Under Noise)**.** *For any reviewer pool with $\mu_u \neq 0$, the recovery failure is structural: as $\Delta \to 0$, the L_2 anchor satisfies $\gamma^*(\Delta) \to s\mu_u$ with $s \in (0, 1)$, so for all sufficiently small $\Delta > 0$ BMF still misorders, with probability tending to one as $\Delta \to 0$, the configurations (q, v) that it misorders at $\Delta = 0$. Proof in Section B.2.*

Proof. By Lemma 21 the expected loss under noise differs from the noiseless loss by the constant $nm\Delta$, so the risk minimizing set, the reduced system (S2')–(S3'), and the anchor $\gamma^* = s\mu_u$ are the same for every $\Delta \ge 0$. For $\mu_u \neq 0$ this gives $\gamma^*(\Delta) = s\mu_u \neq 0$, so by Proposition 6 BMF misorders every pair with $0 < q_{j_1} - q_{j_2} < \gamma^*(v_{j_2} - v_{j_1})$, and configurations with such pairs exist (Corollary 17).

At finite sample size the fit leaves $\mathcal{G}(\theta_{\mathrm{true}})$ and lies on the orbit of a perturbed base point (Remark 22). If its moments deviate by ζ, Lemma 23 gives $\gamma_\Delta = \gamma^* + O(\zeta)$, and the conclusion persists as long as ζ stays below the level at which γ_Δ could reach 0. Sums of the form $\sum_{i,j} \varepsilon_{ij} z_{ij}$ with weights of bounded second moment concentrate at scale $\sqrt{nm\Delta}$, so against the $O(nm)$ scale of the deterministic terms $\zeta = O_p(\sqrt{\Delta/(nm)})$, and in particular $\zeta = O_p(\sqrt{\Delta})$ uniformly in n and m. □

Corollary 14 (Regularization Encodes Design Choices)**.** *As $\kappa \to \infty$ (strong quality regularization relative to user regularization), $\gamma^* \to 0$: the* ordering *of estimated qualities approaches the true quality ordering, though the absolute quality estimates are shrunk toward zero and may be uninformative. As $\kappa \to 0$ (strong user regularization), $\gamma^* \to \mu_u$: by Remark 13, the ordering matches the perspective of the average reviewer, maximizing aggregate reviewer satisfaction. Proof in Section B.2.*

Proof. From (8), $\gamma^* = \mu_u / (1 + \kappa\rho\sigma_v^2 / \tau^{*2})$. As $\kappa \to \infty$ the denominator diverges, so $\gamma^* \to 0$ and the quality ordering is recovered. As $\kappa \to 0$ the denominator approaches 1, so $\gamma^* \to \mu_u$; by Remark 13 this matches the perspective of the average reviewer. □

B.3 Non-Identifiability (Section 4)

Corollary 16 (No mechanism can recover quality from ratings alone)**.** *Let $M : \mathbb{R}^{n\times m} \to \mathbb{R}^m$ be any mechanism that maps a rating matrix to quality estimates. There exist ground-truth parameterizations (u, q, v) and (u', q', v') that produce identical rating matrices but with $q_{j_1} > q_{j_2}$ in one parameterization and $q'_{j_1} < q'_{j_2}$ in the other. Since M must produce the same output for both, it cannot recover the correct quality ordering in both cases. Proof in Section B.3.*

Proof. Both parameterizations (u, q, v) and (u', q', v') yield identical rating matrices by Theorem 15. Any mechanism M operating on the rating matrix alone must return the same output for both. Since their quality orderings are reversed, M cannot recover the correct ordering in both cases. □

C Running Example Details

We illustrate the paper's main developments through a single scenario small enough that every quantity can be verified by hand. Three reviewers with ideologies $u_1 = -0.5$, $u_2 = 0$, $u_3 = 2$ evaluate two notes. Population statistics: $\mu_u = 0.5$, $\sigma_u^2 \approx 1.167$; $\mu_v = 0$, $\sigma_v^2 = 1.0$, $\rho = m/n = 2/3$. The moment conditions of Theorem 10 hold approximately: $\mu_v = 0$ exactly, $\mu_q = 0.7$ only adds a common offset to the recovered qualities, and the quality-leaning covariance $\frac{1}{m}\sum_j q_j v_j = -0.1$ is small.

Ratings

	Note A ($q = 0.8, v = -1$)	Note B ($q = 0.6, v = 1$)
$u = -0.5$	1.30	0.10
$u = 0$	0.80	0.60
$u = 2$	−1.20	2.60

Table 1: Running example ratings $r_{ij} = q_j + u_i v_j$ (simplified model, $c = b_i = 0$). The neutral reviewer ($u = 0$) rates by quality alone; all other rows reflect ideological alignment. True quality ranking: $A > B$.

Gauge shift ($\gamma = 1$)

Applying $\gamma = 1$ yields $q'_A = 0.8 + (1)(-1) = -0.20$, $q'_B = 0.6 + (1)(1) = 1.60$, and shifted ideologies $u'_i = u_i - 1$: $u'_1 = -1.5$, $u'_2 = -1$, $u'_3 = 1$. Every rating is preserved:

	Note A ($q' = -0.20, v = -1$)	Note B ($q' = 1.60, v = 1$)
$u' = -1.5$	1.30	0.10
$u' = -1$	0.80	0.60
$u' = 1$	−1.20	2.60

Table 2: Gauge-shifted parameterization ($\gamma = 1$). All entries are identical to Table 1, but quality ranking is reversed: $q'_B > q'_A$. Confirms Theorem 15.

BMF anchor and ranking flip ($\kappa = 1$)

Solving the balance equations with $\kappa = 1$, $\rho = 2/3$, $\sigma_v^2 = 1$, $\sigma_u^2 \approx 1.167$, $\mu_u = 0.5$ gives $\gamma^* \approx 0.263$, $\tau^* \approx 0.859$. Recovered qualities:

$$\hat{q}_A = 0.8 + 0.263 \times (-1) = 0.537, \qquad \hat{q}_B = 0.6 + 0.263 \times 1 = 0.863.$$

The flip threshold is $(q_A - q_B)/(v_B - v_A) = 0.20/2 = 0.10$. Since $\gamma^* = 0.263 > 0.10$, BMF ranks Note B above Note A with default regularization. (Minimizing Λ_{L_2} over the full gauge orbit of this example, which accounts for the small covariance, gives $\gamma^* \approx 0.31$ and the same ranking.)

D Experimental Details

This appendix collects the implementation, hyperparameter, hardware, and evaluation-metric details for the simulations reported in Section 5.

D.1 Data Generating Process (DGP)

All experiments instantiate the simplified ground truth setting (Definition 2 with $c = b_i = 0$, $d = 1$). The default DGP draws agent ideologies $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_u, 1)$ (with $\mu_u = 0$ unless varied) and item parameters $(q_j, v_j) \overset{\text{i.i.d.}}{\sim} (\mathcal{N}(0,1), \mathcal{N}(0,1))$ independently. Continuous ratings are generated as $r_{ij} = q_j + v_j u_i + \varepsilon_{ij}$ with $\varepsilon_{ij} \sim \mathcal{N}(0, \Delta)$ and $\Delta = 0.25$ ($\sigma = 0.5$). The default sample sizes are $n = 101$ agents and $m = 100$ items. Every experiment refers to this default setup and only describes parameters that differ.

D.2 Implementation and Optimization

We use our own `PyTorch` [Paszke et al., 2019] implementation of BMF (Definition 1), which makes it straightforward to swap in different regularizers, explicit gauge constraints, and link functions. Ideology factors are initialized Xavier-uniform, intercepts and qualities at zero, and all parameters are trained with the Adam optimizer [Kingma and Ba, 2015] with learning rate 0.05 (0.01 for binary ratings) for 500 epochs of full-batch training, with early stopping on the training loss (patience 10, tolerance 10^{-5}). The training objective is mean squared error for continuous ratings and binary cross-entropy (log loss) for binary ratings. Regularization is L_2, with each parameter block penalized by the mean of its squares: weight λ on $\hat{u}$ and $\hat{v}$, and $\kappa\lambda$ on $\hat{b}$, $\hat{q}$, and $\hat{c}$, with $\lambda = 0.01$ and $\kappa = 1$ by default. In the sum convention of Section 3 this is $\lambda_u = \lambda m$, $\lambda_v = \lambda n$, $\lambda_q = \kappa\lambda n$, $\lambda_b = \kappa\lambda m$, so $A = \kappa\sigma_v^2$ and $K = \sigma_v^2$ in Lemma 18, which for $n \approx m$ is the setting the figures compare against.

D.3 Reproducibility and Compute

Experiments are run on a MacBook Pro with an Apple M1 Pro chip. Each individual simulation at the default settings ($n = 101$, $m = 100$, 500 epochs) takes approximately 10 seconds. The codebase is parallelized over replications and configurations, so the full suite of runs reported in this paper (sweeps over μ_u, regularization strength, the regularization ratio κ, the binary-rating variant, and 10 replications per configuration) completes in a few hours on this machine. Code and instructions to reproduce all figures are available at `https://github.com/BibliotecaBabel/BirdwatchBMFAnalysis`.

D.4 Estimating the Gauge Shift γ

The agent-centric gauge transformation maps $\hat{u}_i \mapsto \hat{u}_i - \gamma$ and $\hat{q}_j \mapsto \hat{q}_j + \gamma\hat{v}_j$ while leaving predicted ratings unchanged. Given access to the true simulation parameters, we estimate γ in two complementary ways:

Item-based ($\hat{\gamma}_{\text{item}}$). Fit the regression $\hat{q}_j = \beta_0 + \beta_1 v_j + \epsilon_j$ and take $\hat{\gamma}_{\text{item}} = \hat{\beta}_1$. This captures the systematic covariation between estimated quality and item ideology.

User-based ($\hat{\gamma}_{\text{user}}$). Fit the regression $u_i = \beta_0 + \beta_1 \hat{u}_i + \epsilon_i$ and take $\hat{\gamma}_{\text{user}} = \hat{\beta}_0$. On the orbit $\hat{u}_i = \tau(u_i - \gamma)$, so $u_i = \gamma + \hat{u}_i/\tau$ and the intercept is γ for every τ.

Under correct specification both estimators converge to the same value, and our experiments report close agreement (within ± 1 standard error; see Figure 3).

D.5 Evaluation Metrics

For a mechanism M producing estimated qualities $\hat{q}$, let $S(M)$ be the set of ℓ items selected as the top-ℓ by $\hat{q}_j$; we use $\ell = 10$.

Reference mechanisms (oracles). We compare the BMF mechanism against an idealized benchmark that has access to ground-truth information unavailable to any real mechanism.

Quality Oracle (M_q^*). Selects the ℓ items with the highest true quality q_j. This is the best any mechanism can do on objective quality (OQ, defined below), and serves as the upper bound for quality-based selection.

This oracle is unachievable in practice since it requires knowledge of true qualities; it serves purely as a normalization baseline for the gap metric below.

Objective Quality gap. Normalized shortfall in total true quality of selected items relative to the Quality Oracle M_q^*:

$$\mathrm{OQ}_{\text{gap}}(M) = \frac{\mathrm{OQ}(M_q^*) - \mathrm{OQ}(M)}{\mathrm{OQ}(M_q^*)}, \qquad \mathrm{OQ}(M) = \sum_{j \in S(M)} q_j.$$

This is the metric reported as *Quality gap@10* in Figure 4.

The metric is reported as the mean over 10 replications, with error bars showing ± 1 standard error.

D.6 Robustness Checks

We verify the robustness of the γ–μ_u relationship along four axes.

Variation across replications. The pool sizes $n = 101$ and $m = 100$ match the sparse-data regime of a deployed system and keep the simulations away from large-n asymptotics. Individual estimates $\hat{u}_i$ and $\hat{q}_j$ vary widely across replications at these sizes, and so does $\hat{\gamma}$. The spread follows from Lemma 23: the empirical moments of a finite draw deviate from their population values by $O_p(n^{-1/2})$, so the anchor fluctuates around γ^* at the same order. The scatter therefore centers on γ^*. The drift toward μ_u is stable across replications.

Regularization strength. Fixing $\mu_u = 2$, $\kappa = 1$, and $\sigma = 0.1$, and varying $\lambda = \lambda_u = \lambda_q$ from 5×10^{-3} to 10^0: the estimates stay within about 0.1 of $\gamma^* = 0.78$ for $\lambda \leq 0.1$, with a small dip at $\lambda = 0.005$ (0.69 item-based, 0.73 user-based); the item-based estimate stays between 0.69 and 0.74 over the whole range, while the user-based estimate falls for $\lambda \geq 0.5$, to 0.51 at $\lambda = 1$, as the optimizer trades predictive accuracy for lower parameter norms. Figure 5 shows the estimators; Figure 6 shows the corresponding training risk and regularization term, confirming that risk begins to rise where $\hat{\gamma}_{\text{user}}$ deviates. Figure 8 replicates the full γ–μ_u sweep across several λ values; the three curves coincide.

Regularization ratio κ. Fixing $\mu_u = 2$ and $\lambda = 0.1$ and varying κ from 0.1 to 10 (Figure 7): the item-based estimates (1.69, 1.48, 1.12, 0.74, 0.35, 0.14, 0.04) follow $\gamma^*(\kappa)$ from (10) (1.82, 1.61, 1.21, 0.80, 0.44, 0.23, 0.09) and sit well below $\mu_u/(1+\kappa)$, the value at $\tau^{*2} = 1$; the user-based estimate tracks the curve up to $\kappa = 2$ and undershoots beyond it (-0.05 at $\kappa = 4$, -0.27 at $\kappa = 10$). As κ grows (quality factors penalized more heavily relative to user factors), γ decreases; as κ shrinks, γ increases toward μ_u. This sweep uses $\lambda = 0.1$ because at $\lambda = 0.01$ the fit stalls near 0.89 for $\kappa \leq 0.5$: along the gauge orbit the risk is flat, so the descent there is driven by the regularizer alone and is slow at small λ.

Noise level. We additionally vary observation noise σ while holding all other parameters fixed ($\mu_u = 2$, L2, $\kappa = 1$, $\lambda = 0.01$). Estimates remain close to $\gamma^* = 0.78$ across all noise levels tested (item-based 0.70, 0.70, 0.72 and user-based 0.71, 0.71, 0.66 at $\sigma = 0, 0.1, 0.5$), indicating that the anchor is robust to moderate measurement noise (Figure 9).

D.7 Binary Ratings

Community Notes uses categorical (helpful, somewhat helpful, not helpful) ratings rather than continuous scores. To check that our results extend to this setting, we modify the DGP as follows: (i) compute the latent score $s_{ij} = q_j + v_j u_i + \varepsilon_{ij}$, (ii) pass it through the logistic link $\sigma(x) = 1/(1+e^{-x})$,

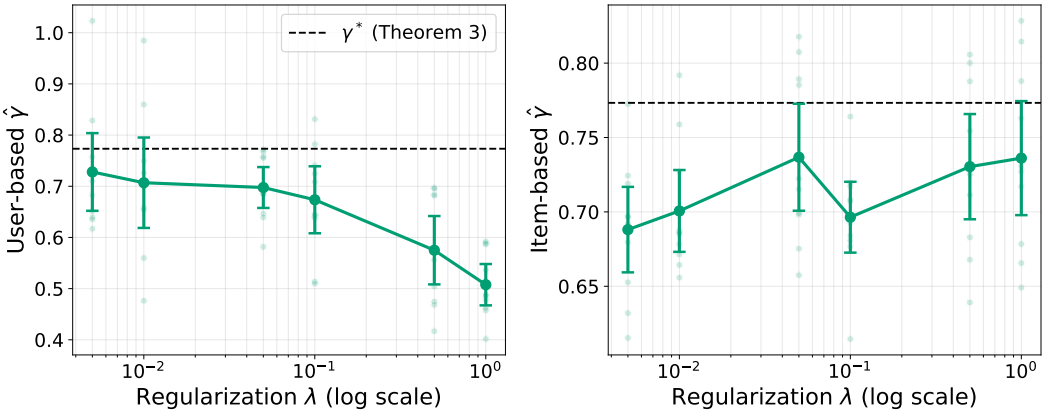


Figure 5: Estimated $\hat{\gamma}$ vs. regularization strength λ ($\mu_u = 2$, $\kappa = 1$, $\sigma = 0.1$). Dashed: $\gamma^* = 0.78$ from (10). Estimates stay within about 0.1 of it for $\lambda \leq 0.1$; the user-based estimate falls to 0.51 at $\lambda = 1$, where the fit trades risk for norm (Figure 6).

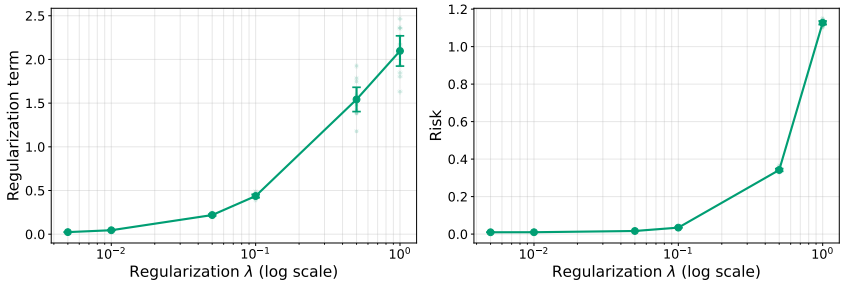


Figure 6: Regularization term (left) and training risk (right) vs. λ. Risk stays near its minimum for $\lambda \leq 0.1$; beyond that, the optimizer trades predictive accuracy for lower parameter norm, explaining the deviation in $\hat{\gamma}$ at high λ.

and (iii) sample $r_{ij} \sim \text{Bernoulli}(\sigma(s_{ij}))$. BMF is then fitted with binary cross-entropy loss; all other components of the procedure are unchanged. The γ–μ_u relationship is preserved, with a compression effect at large $|\mu_u|$ that is consistent with saturation of the logistic link. This indicates that the implicit anchoring to the population mean persists in the categorical setting actually used by Community Notes. Figures 10 and 11 replicate the two main figures in this setting.

D.8 Misspecified Rating Models

Our results assume ratings follow the bilinear form of Definition 2. To probe what survives when the generating model departs from it, we generate ratings as $r_{ij} = q_j + g(u_i, v_j) + \varepsilon_{ij}$ for three alternative interaction terms and fit the same L_2-regularized bilinear BMF in each case: a Euclidean utility $g = -(u_i - v_j)^2$, a higher-order interaction $g = u_i v_j + \varphi\, u_i^2 v_j^2$ with $\varphi = 0.3$, and an unmodeled confounder $g = u_i v_j + w_i z_j$ with $w_i, z_j \sim \mathcal{N}(0, 1)$ drawn independently of the other parameters. The bilinear case $g = u_i v_j$ serves as the well-specified reference. Only the interaction term varies: every run uses the default DGP of Section D.1 and the optimization settings of Section D.2. The three variants are specific departures from the bilinear form. Other misspecifications are untested.

Figure 12 sweeps μ_u over $[-2, 2]$ for each variant. Table 3 reports the slope of each $\hat{\gamma}$ estimator against μ_u over that sweep, together with the quality gap at $\mu_u = 0$ and at $|\mu_u| = 2$.

The anchor tracks the mean reviewer ideology in every variant, with $\hat{\gamma}$ increasing in μ_u across the sweep. The slopes in Table 3 are fit across the whole range, where $\hat{\gamma}$ tapers at the extremes as it does under correct specification (Figure 3), so they sit below the slope $1/2$ at $\mu_u = 0$; the exact curve of (10) has least-squares slope 0.42 over this grid, close to the bilinear row (0.37, 0.38), and the slopes are read as a comparison across variants.

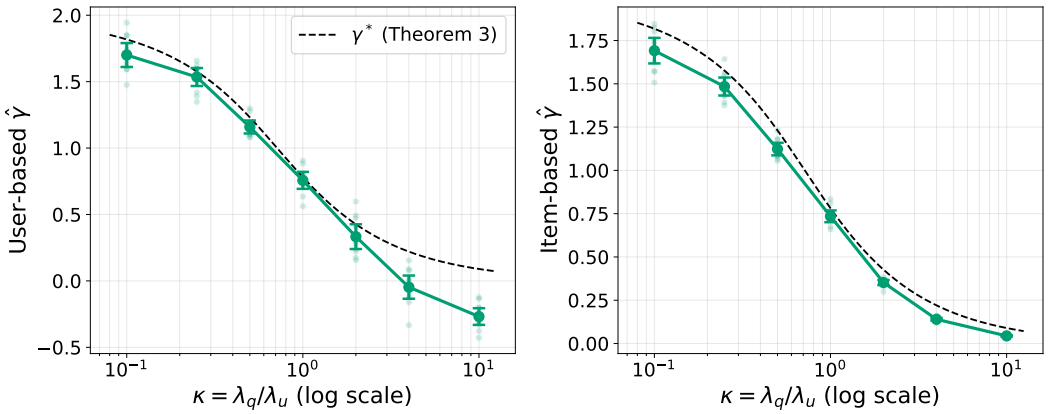


Figure 7: Estimated $\hat{\gamma}$ vs. regularization ratio $\kappa = \lambda_q/\lambda_u$ ($\mu_u = 2$, $\lambda = 0.1$, log scale). Dashed: $\gamma^*(\kappa)$ from (10). The item-based estimate follows it across two decades of κ; the user-based estimate tracks it up to $\kappa = 2$ and undershoots beyond.

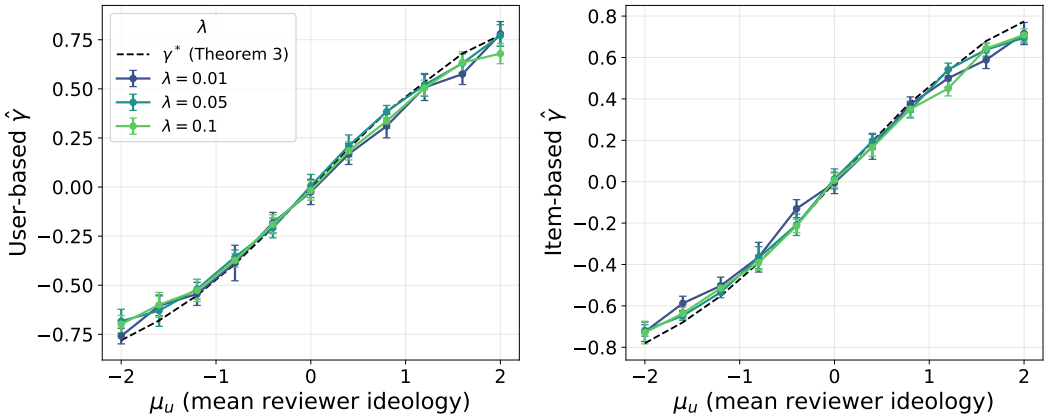


Figure 8: Estimated $\hat{\gamma}$ vs. μ_u for different regularization strengths λ (L2, $\kappa = 1$). The three curves coincide and track γ^* of Theorem 10 (dashed), following $\mu_u/2$ near zero and bending below it at the extremes, which shows that the anchor is not sensitive to the choice of regularization strength.

Misspecification makes recovery worse. The confounder preserves recovery at $\mu_u = 0$, while the Euclidean utility and the higher-order interaction lose it, and the higher-order variant distorts most at the extremes. The two $\hat{\gamma}$ estimators, which agree under correct specification (Section D.4), also split under the Euclidean utility. The higher-order variant still takes its smallest gap at $\mu_u = 0$: a neutral pool remains the best case for quality recovery, though it no longer recovers the ranking exactly. Recovery at a neutral pool is proved for the bilinear model (Theorem 10, Corollary 11), while an anchor forms and follows the mean under every departure we ran.

Table 3: Anchor slope and quality gap under misspecification. s_{item} and s_{user} are the slopes of the two $\hat{\gamma}$ estimators of Section D.4 regressed on μ_u across the sweep of Figure 12. The two estimators agree under every variant except the Euclidean utility.

Variant	$g(u, v)$	s_{item}	s_{user}	Quality gap@10 ($\mu_u = 0$)	Quality gap@10 ($\lvert\mu_u\rvert = 2$)
bilinear (well-specified)	uv	0.370	0.382	0.007	0.211
euclidean	$-(u - v)^2$	0.386	0.193	0.137	0.158
higher order	$uv + \varphi u^2 v^2$	0.478	0.500	0.123	0.504
confounder	$uv + wz$	0.584	0.611	0.005	0.442

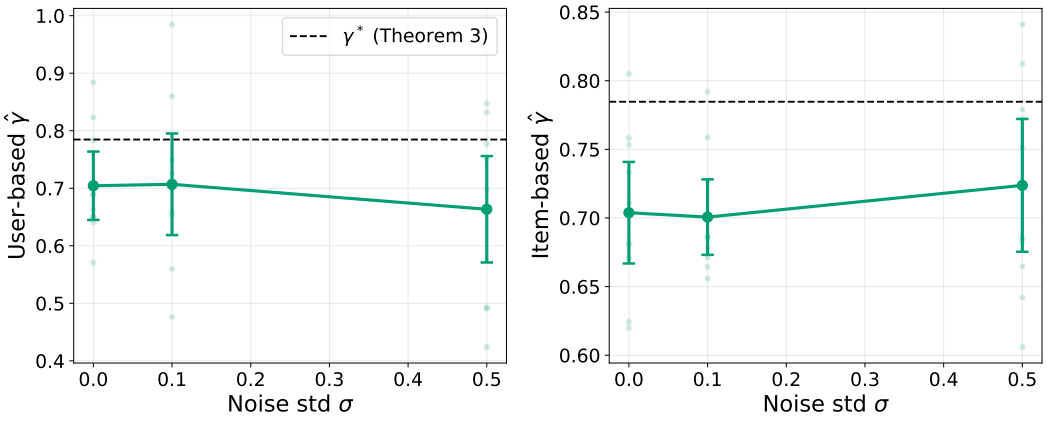


Figure 9: Estimated $\hat{\gamma}$ vs. observation noise σ ($\mu_u = 2$, L2, $\kappa = 1$, $\lambda = 0.01$). The estimators remain close to $\gamma^* = 0.78$ (dashed, from (10)) across the noise levels tested.

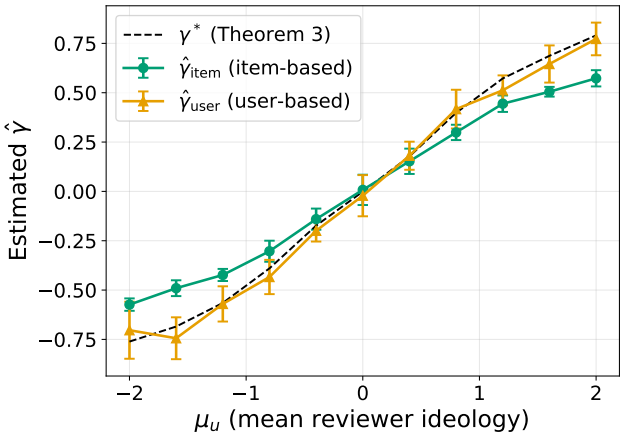


Figure 10: Estimated $\hat{\gamma}$ vs. μ_u under binary ratings (logistic link, Bernoulli observations, L2, $\kappa = 1$). Dashed: γ^* of Theorem 10, a reference only, since the logistic link is outside the theorem. The user-based estimate tracks it; the item-based estimate compresses at large $|\mu_u|$, consistent with saturation of the logistic link.

D.9 Item Ideology Sweep

Theory predicts that shifting the mean item ideology μ_v does not induce a gauge shift ($\gamma^* = 0$), since the anchor depends on the user distribution rather than the item distribution. (Theorem 10 assumes $\mu_v = 0$; the prediction still holds without it, since at $\mu_u = 0$ the intercept balance forces $\beta = 0$ even when $\eta^* \neq 0$, and the agent-centric balance then gives $\gamma = 0$.) We verify this by sweeping $\mu_v \in [-2, 2]$ while holding the user distribution fixed at $\mu_u = 0$ (all other parameters at their defaults). Both $\hat{\gamma}$ (Figure 13) and the quality gap (Figure 14) remain flat across the full range of μ_v, confirming that item ideology distribution is not a source of distortion.

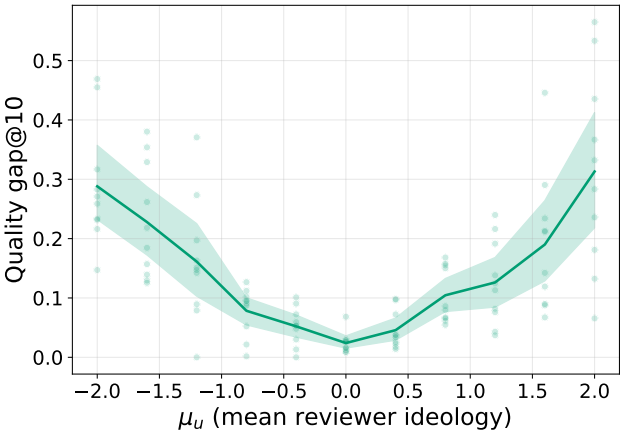


Figure 11: Quality gap@10 vs. μ_u under binary ratings. The monotonic increase with $|\mu_u|$ is preserved relative to the continuous case (Figure 4).

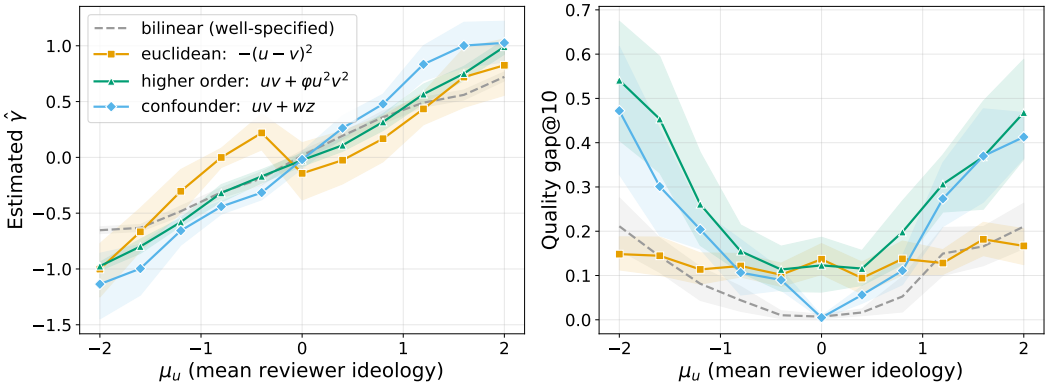


Figure 12: Estimated anchor $\hat{\gamma}$ (left) and Quality gap@10 (right) against mean reviewer ideology μ_u, for the well-specified bilinear model and three misspecified variants. The anchor follows μ_u in all four; the quality gap vanishes at $\mu_u = 0$ for the bilinear model and the confounder, and stays positive under the Euclidean utility and the higher-order interaction. Bands are ± 1 SE over 10 replications.

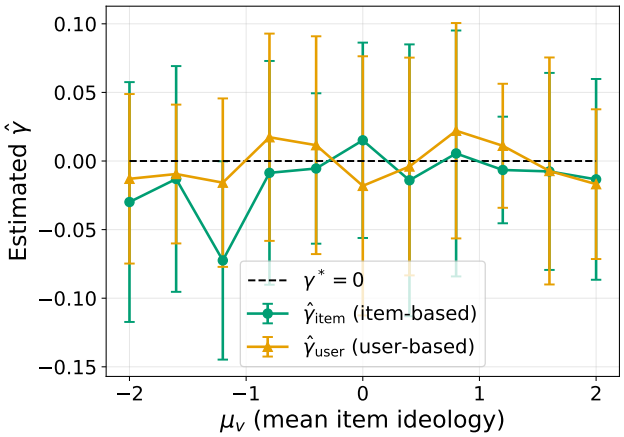


Figure 13: Estimated $\hat{\gamma}$ vs. mean item ideology μ_v (L2, $\kappa = 1$, $\mu_u = 0$). The theoretical prediction $\gamma^* = 0$ (dashed) is confirmed: item ideology mean does not shift the anchor.

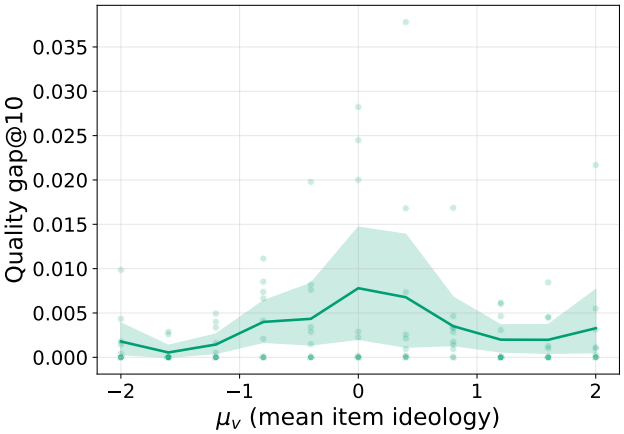


Figure 14: Quality gap@10 vs. mean item ideology μ_v. Unlike the agent-mean sweep (Figure 4), there is no systematic increase, confirming that the item ideology distribution does not induce quality distortion.

E Use of Large Language Models

Large language models (LLMs), primarily Claude (Anthropic), were used in the following supporting roles:

- **Writing and editing:** improving clarity, grammar, and flow; in some cases generated text was edited and incorporated.
- **Coding support:** refactoring experiment code, adding evaluation metrics, and generating plotting utilities.
- **Proof assistance:** checking intermediate steps and suggesting proof strategies. All theoretical results were developed and verified by the authors.
- **Ideation:** brainstorming and exploratory discussion. All conclusions are the authors' own.

LLMs did not contribute novel methods, theorems, or results to this work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction match claims made in the paper and results to the best of our ability in explaining them succintly.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations covered at the end of section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theorems and corollaries have the associated assumptions and proofs. Proofs that are not covered in the main body are covered in the appendix (and referenced in the main body).

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, experimental setup is described in section 5 and in more detail in the appendix (section D)

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: An implementation of the BMF algorithm following the Birdwatch specification, with instructions to reproduce all figures, is available at `https://github.com/BibliotecaBabel/BirdwatchBMFAnalysis`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, experimental setup is described in section 5 and in more detail in the appendix (section D)

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Result plots include error bars calculated based on multiple repetitions of the parameter configurations. Each figure indicates this.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, resource details are described in section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. Section 6 discusses its societal implications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Covered in the discussion section (Section 6), particularly the part discussing implications for Community Notes and similar algorithms.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper analyzes a public, deployed algorithm and releases no new model or dataset with a risk of misuse. Section 6 discusses the implications of the findings.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our implementation is based on the Biased Matrix Factorization algorithm described in the Birdwatch paper and the Community Notes documentation, both of which are properly cited in the paper. No proprietary datasets or licensed code were used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: An implementation of the BMF algorithm following the Birdwatch specification, with instructions to reproduce all figures, is available at `https://github.com/BibliotecaBabel/BirdwatchBMFAnalysis`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used for supporting tasks including writing assistance, coding support (refactoring, plotting, metrics), proof verification, and ideation. LLMs did not contribute novel methods or results; all theoretical contributions were developed and verified by the authors. See section E for details.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.