BIASFATE: CALIBRATION LIMITS OF REWEIGHTED LLM SIMULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can reweighting LLM-generated survey responses improve on the human sample used to calibrate them? BiasFate links calibration information, proposal coverage, and estimation error for categorical responses. With reused empirical proposal counts and no clipping, self-normalized importance sampling (SNIS) equals the calibration mean restricted to observed categories; full empirical coverage recovers the direct mean exactly. Under covered support with known proposal probabilities and independent draws, unnormalized plug-in weighting adds nonnegative variance, while SNIS inherits calibration variance to leading order. Support truncation and clipping introduce distinct sources of bias. Across roughly 640K generations, four 7–9B models, a 34B replication, four task families, and released GlobalOpinionQA marginals, the direct calibration mean beats deployed clipped SNIS in 270 of 288 matched configurations. Removing clipping reduces covered-cell SNIS RMSE from 0.1027 to 0.0245, versus 0.0148 for direct estimation. Plateau diagnostics transfer to held-out questions, while diagnostic-guided routing and allocation show no improvement over matched controls. These results explain when simulation reproduces calibration information, how it loses target mass, and why diagnostic accuracy and estimation gains require separate evaluation.

1 INTRODUCTION

LLMs are increasingly used as synthetic survey respondents (Argyle et al., 2023; Santurkar et al., 2023; Wang et al., 2025; Chen et al., 2026), annotators (Gilardi et al., 2023), and social agents (Park et al., 2023; 2024). A model, prompt, persona, and decoding rule define a sampler $P_\lambda(y \mid x)$ whose responses may differ systematically from the target population $Q(y \mid x)$. Reweighting against a small human calibration sample is a natural repair. That sample, however, already estimates the population quantity of interest. The relevant question is therefore whether the reweighted simulation improves on the direct calibration estimate at the same human-data budget.

We study this question for a fixed finite response space and a known bounded scoring function f. Our central observation is an exact cancellation: when the same categorical counts estimate the proposal probabilities and evaluate unclipped SNIS, positive proposal frequencies cancel. The result is simply the calibration mean restricted to categories observed in the simulation. Additional simulated draws change this estimator only by discovering categories. Under full empirical coverage it equals the direct calibration mean, regardless of how accurately the positive proposal frequencies are estimated.

For independent proposal draws with known proposal probabilities, the direct mean Rao–Blackwellizes the *unnormalized* plug-in estimator. For *self-normalized* weighting, we establish an asymptotic decomposition into calibration variance and proposal variance. Oracle weights already encode the target distribution, so their guarantees describe a separate information setting. Together, these results make the contribution of simulation explicit at a fixed calibration budget.

Coverage determines which target mass ratio weighting can represent. On a fixed deficient support, the estimator converges to a truncated target mean. Structural absence imposes a distributional error floor, while mean error depends on the scoring function. Finite pilots measure non-observation at a stated budget, motivating a separate empirical plateau screen whose behavior can be tested on fresh draws.

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

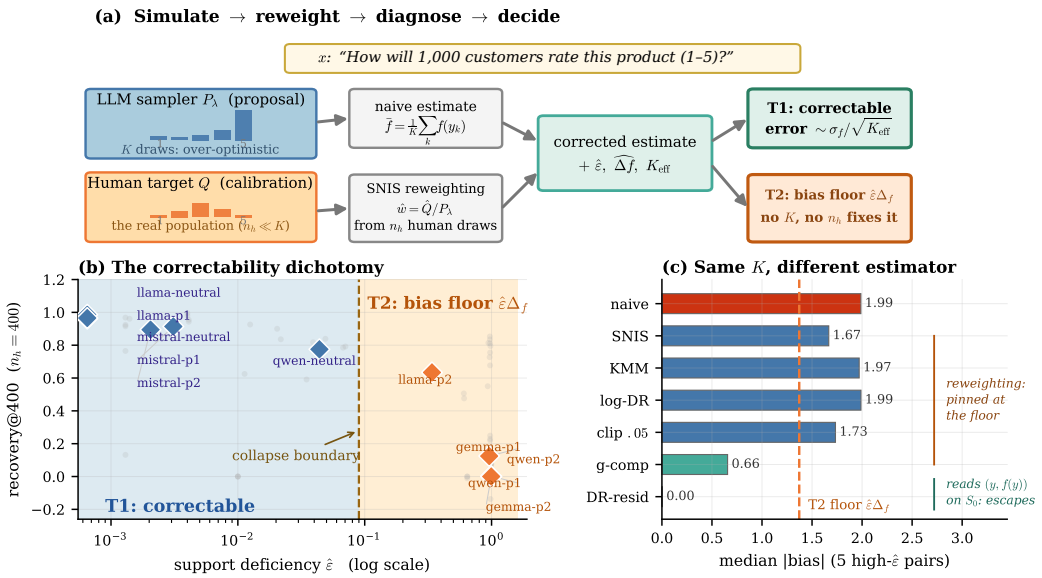


Figure 1: **BiasFate audit.** (a) A human calibration sample supplies plug-in weights and diagnostics for an LLM proposal; the audit sorts each model–task pair into **T1** (covered support, correctable) or **T2** (support deficiency: a bias floor that no reweighting removes). (b) Across the 11 non-self pairs, recovery (fraction of naive bias removed at n_h=400) collapses as empirical uncovered mass $\hat{\varepsilon}$ grows ($\rho = -0.98$); the dashed boundary ($\hat{\varepsilon} \approx 0.09$) separates the regimes (cell-level split of the 132-cell matrix; pair verdicts in Table 2 use the plateau rule, T1 iff plateau $< 0.3\times$ naive). (c) On the five pairs with the largest uncovered mass, reweighting stays pinned near the predicted floor (median |bias| 1.67–1.99), while the estimators that read the paired calibration outcomes directly escape it (g-comp 0.66, DR 0.00).

The experiments separate diagnostic accuracy from decision value. On 12 matched GlobalOpinionQA questions, the direct mean outperforms deployed double-clipped SNIS in 270/288 configurations. The plateau screen has 0.91 accuracy, rising to 0.97–0.98 on a held-out expansion. Diagnostic routing and budget targeting show no gain over matched controls. The covered-cell ablation in App. C.6 attributes much of the deployed gap to clipping, complementing the support and calibration mechanisms.

We contribute (i) an exact identity for unclipped weighting with reused categorical counts and variance accounting for independent proposal draws; (ii) a separation of support truncation, finite-budget non-observation, and clipping distortion; and (iii) an empirical test of plateau prediction and its value for routing and allocation. Together, these results provide a concrete reference for evaluating calibrated synthetic respondents: the direct mean from the same calibration labels, against which additional simulation and diagnostic-guided decisions must show their estimation value.

2 Related Work

LLM simulations of human respondents. Persona-conditioned respondents, synthetic surveys, and social agents reveal systematic gaps between model and human response distributions (Argyle et al., 2023; Santurkar et al., 2023; Gilardi et al., 2023; Park et al., 2023; Bail, 2024; Chen et al., 2026); recent work measures persona effects and their subgroup distortions (Hu & Collier, 2024; Morocho et al., 2026), questions population inference from LLM survey responses (Dominguez-Olmedo et al., 2024; Bisbee et al., 2024; Ku et al., 2026), quantifies the uncertainty LLM simulations carry (Huang et al., 2025), and shows that marginal alignment can mask structural failures (Williams et al., 2026). Calibration methods study how to combine LLM and human samples under covariate or concept shift (Tigre & Souto, 2026). BiasFate addresses a narrower estimator question in this line: can importance weighting of an explicit proposal sampler improve on the direct calibration

estimator at the available budget? We isolate the support and plug-in-weight limits of categorical SNIS, include the direct human mean as the matched-budget baseline, and distinguish reused-count identities from independent-stream evidence.

Importance sampling, overlap, and shift. SNIS theory, tempered or clipped weighting, and mismatch bounds motivate the covered-support analysis (Agapiou et al., 2017; Owen, 2013; Neal, 2001; Ionides, 2008; Cortes et al., 2008); positivity failures motivate the structural support analysis (Rosenbaum & Rubin, 1983; Crump et al., 2009; Khan & Tamer, 2010; D'Amour et al., 2021). Label-shift methods assume shared support (Shimodaira, 2000; Saerens et al., 2002; Lipton et al., 2018). BiasFate separates structural absence from finite-budget non-observation and treats the latter as a diagnostic whose reliability must be checked on fresh draws.

Small labeled samples, diagnostics, and missing mass. Prediction-powered inference (PPI) uses a small labeled sample to correct model predictions with distribution-free intervals (Angelopoulos et al., 2023), extended to filter where predictions are informative (Xu & Sun, 2026) and recalibrated to improve asymptotic efficiency over the labeled-only estimator (Ji et al., 2026); raking and generalized regression estimators use known population margins (Deville & Särndal, 1992; Deming & Stephan, 1940). With a variance-minimizing tuning parameter, these difference-estimation corrections need not exceed the labeled sample's leading-order variance; our dominance results concern multiplicative ratio reweighting, which adds proposal sampling variance and weight-estimation noise rather than shrinking toward the labeled mean. PSIS and related diagnostics detect unstable realized importance weights (Vehtari et al., 2024); Good–Turing methods estimate unseen mass (Good, 1953); off-policy evaluation studies the corresponding two-sample structure (Dudík et al., 2011). Our T2 analysis describes what response-space reweighting cannot recover when labels on uncovered outcomes are unavailable, and our audit adapts these weight-stability and missing-mass concerns to finite categorical LLM samplers, with known-proposal tests and independent-stream evaluations supplying the settings in which error control or transfer can be assessed.

3 Setting and Audit Protocol

3.1 Target, estimand, and information budget

For a fixed question or context x, let $P = P_\lambda(\cdot \mid x)$ be the LLM proposal and $Q(\cdot \mid x)$ the target distribution on a finite response set $\mathcal{Y}$. Text responses enter through a declared parser. We estimate $\mu_f = \mathbb{E}_Q f$ for a known bounded scoring map f from n_h target calibration draws and K proposal draws. Each new question receives its own calibration sample. The baseline is $\bar{f}_H = n_h^{-1} \sum_i f(Y_i^h)$, with variance σ_f^2/n_h. In the human-target studies, calibration draws are sampled from released survey marginals; the LLM-teacher studies instead use a frozen model distribution as Q. The latter test mechanisms, not human representativeness.

The naive simulation mean converges to $\mathbb{E}_P f$, leaving sampler bias $b_f = \mathbb{E}_P f - \mu_f$. With empirical target frequencies $\hat{Q}$, plug-in SNIS is

$$\hat{\mu}_{\mathrm{SNIS}} = \frac{\sum_{k=1}^K \hat{w}(Z_k) f(Z_k)}{\sum_{k=1}^K \hat{w}(Z_k)}, \qquad Z_k \sim P.$$

We distinguish known-P weights $\hat{Q}/P$ evaluated on independent draws from empirical-proposal weights $\hat{Q}/\hat{P}$ evaluated on the same counts that define $\hat{P}$. Oracle weights Q/P constitute a separate information setting. Table 3 records the data flow of each experiment. Zero-denominator deployment cells use the recorded naive-mean fallback.

3.2 Clipping changes the estimator

The main deployment clips ratios to $[\alpha, 1/\alpha]$ with $\alpha = 0.02$. Under reused counts, its unnormalized category mass is exactly

$$\hat{P}(y)\,\mathrm{clip}\left(\frac{\hat{Q}(y)}{\hat{P}(y)}, \alpha, \frac{1}{\alpha}\right) = \min\left\{\max\{\hat{Q}(y), \alpha\hat{P}(y)\}, \frac{\hat{P}(y)}{\alpha}\right\}, \quad y \in O_K,$$

where $O_K = \{y : \hat{P}(y) > 0\}$. Normalizing these masses gives the deployed estimate. Lower clipping can assign positive mass to a category with $\hat{Q}(y) = 0$; upper clipping can remove target mass. Hence the unclipped identities below do not predict the full error of the clipped implementation. Appendix A.4 studies clipping and App. C.6 isolates its empirical cost.

3.3 A PLATEAU SCREEN AND ITS EVALUATION

Let $\mathcal{U}_K = \mathcal{Y} \setminus O_K$, $\hat{\varepsilon} = \hat{Q}(\mathcal{U}_K)$, and $\hat{m}_O = \mathbb{E}_{\hat{Q}}[f \mid O_K]$ when $\hat{Q}(O_K) > 0$. The estimated truncation magnitude is

$$\hat{g} = \hat{\varepsilon}\widehat{\Delta_f} = |\bar{f}_H - \hat{m}_O|, \qquad \hat{b} = |\mathbb{E}_{\hat{P}} f - \bar{f}_H|.$$

When $\hat{\varepsilon} = 0$, set $\hat{g} = 0$ without estimating an empty-bin conditional mean. Cells with no covered calibration draws require the fallback rather than this calculation. Report missing mass, the estimand contrast, and ESS alongside $\hat{g}$.

The empirical screen labels a cell **P** (plateau) if $\hat{g} \geq c\hat{b}$, and **R** (recovery-screen) otherwise. These are threshold labels, not the theoretical coverage regimes T1/T2: a partially uncovered cell can receive R when its truncation error is small relative to naive bias. The default $c = 0.3$ is a pre-specified operating point; refitting uses development items with known reference targets. The Chinese and English threshold studies also evaluate recovery-based labels, whose definitions and fitted thresholds are reported separately in App. C.1. Reference-label agreement, fresh-draw estimation error, and downstream routing are distinct evaluation tasks.

For a frozen observed support, the reference plateau uses Q in place of $\hat{Q}$. This makes agreement a test of estimating a protocol-defined quantity, with algebraic coupling under reused counts. Fresh-stream and held-out experiments test how far it transfers. They retain target calibration data for each test question; transferring the rule does not eliminate that human-data requirement. A plateau interval describes uncertainty in the diagnostic functional, not a confidence interval for SNIS error.

For independent unclipped SNIS, equating proposal and calibration variance gives the approximate crossover $K^\star = (\hat{v}_f/\hat{\sigma}_f^2)n_h$. It marks diminishing returns within that estimator, not a budget at which it beats the direct mean. Under reused counts, proposal frequencies cancel and only support discovery matters before clipping. Under heavy tails, the estimated crossover and ESS can both be unreliable (App. C.7).

4 THEORY

4.1 CALIBRATION INFORMATION IN PLUG-IN WEIGHTING

Proposition 1 (Same-sample cancellation for reused-proposal plug-in SNIS)**.** *Let $\hat{P}$ be the empirical distribution of a categorical proposal sample, $O_K = \{y : \hat{P}(y) > 0\}$, and $\hat{Q}(O_K) > 0$. Form plug-in weights $\hat{w}(y) = \hat{Q}(y)/\hat{P}(y)$ and evaluate SNIS on the same sample that defines $\hat{P}$. Then*

$$\hat{\mu}_{\mathrm{SNIS}} = \frac{\sum_{y \in O_K} \hat{Q}(y) f(y)}{\sum_{y \in O_K} \hat{Q}(y)}. \tag{1}$$

Conditional on O_K and the calibration sample, the estimator is independent of every positive proposal frequency. If O_K covers the calibration support, it equals the direct calibration mean exactly for any K.

Proposition 1 covers reused-count pipelines: simulated frequencies cannot improve the mean after plug-in reweighting; proposal draws matter only when they discover a category. Independent proposal draws do not rescue the unnormalized estimator (Prop. 2).

Proposition 2 (Independent-stream unnormalized plug-in dominance)**.** *Let $H = (Y_1^h, \ldots, Y_{n_h}^h) \sim Q^{n_h}$ define the empirical target $\hat{Q}_H$, and independently let $Z_1, \ldots, Z_K \sim P$ on a finite support with $Q \ll P$. For*

$$\widetilde{\mu}_K = \frac{1}{K}\sum_{k=1}^{K} \frac{\hat{Q}_H(Z_k)}{P(Z_k)} f(Z_k), \qquad \bar{f}_H = \frac{1}{n_h}\sum_{i=1}^{n_h} f(Y_i^h),$$

we have $\mathbb{E}[\widetilde{\mu}_K \mid H] = \bar{f}_H$ *and, for* $\mu = \mathbb{E}_Q f$,

$$\mathbb{E}(\widetilde{\mu}_K - \mu)^2 = \frac{\sigma_f^2}{n_h} + \mathbb{E}[\mathrm{Var}(\widetilde{\mu}_K \mid H)] \geq \frac{\sigma_f^2}{n_h}.$$

Thus the direct calibration mean Rao–Blackwellizes the independent-stream unnormalized plug-in estimator. Equality requires zero conditional proposal variance. The statement is exact for every n_h, K*; SNIS is a ratio estimator and instead obeys Cor. 1.1. Proof: App. A.2.*

Proposition 1 is an exact unclipped identity; Prop. 2 is a finite-sample MSE comparison for unnormalized weighting; Cor. 1.1 is an asymptotic SNIS statement. They do not assert finite-sample pointwise MSE dominance over all normalized or regularized estimators. Clipping and decoupled calibration are evaluated in App. A.4 and App. C.6.

4.2 Oracle benchmark and independent-stream SNIS

Assumption (coverage)**.** $\mathrm{supp}(Q \mid x) \subseteq \mathrm{supp}(P_\lambda \mid x)$ and $\inf_{y \in \mathrm{supp}(Q)} P_\lambda(y) \geq p_{\min} > 0$.

Theorem 1 (T1, oracle weights)**.** *With true weights* $w^\star = Q/P_\lambda$*, assume coverage and let* $t = \log(4/\delta)$*. Define* $v_f = \mathrm{Var}_{P_\lambda}[w^\star(f - \mu_f)]$*,* $M = \|w^\star\|_\infty \leq 1/p_{\min}$*,* $R = \|f - \mu_f\|_\infty$*, and* $a_K = \sqrt{2(\rho_1 - 1)t/K} + 2Mt/(3K)$*, where* $\rho_1 = \mathbb{E}_{P_\lambda}[(w^\star)^2]$*. If* $a_K < 1$*, then with probability at least* $1 - \delta$*,*

$$|\hat{\mu}_{\mathrm{SNIS}} - \mu_f| \leq \frac{\sqrt{2v_f t/K} + 2MRt/(3K)}{1 - a_K}.$$

The linear range term in a_K *is kept because variance control alone does not bound the denominator deviation; App. A.1 gives the proof.*

Remark. The leading variance is v_f/K and depends jointly on weights and the estimand. Under $\mathrm{Cov}_Q(w^\star, (f - \mu_f)^2) \leq 0$, it is bounded by $\rho_1 \sigma_f^2/K = \sigma_f^2/K_{\mathrm{eff}}$ with $K_{\mathrm{eff}} = K/\rho_1$ the effective sample size; without that condition, use $R^2\rho_1/K$. The condition is checked empirically on covered cells and the additive scaling audited against known truth (App. C.1, App. C.5).

Corollary 1.1 (two-sample plug-in SNIS: the calibration floor)**.** *Let* $\hat{w} = \hat{Q}/P_\lambda$ *use* n_h *calibration draws independently of* K *proposal draws on a finite covered support, with known* P_λ *and no clipping. As* $n_h, K \to \infty$*,*

$$\mathrm{MSE}(\hat{\mu}_{\mathrm{SNIS}}) = \frac{\sigma_f^2}{n_h} + \frac{\mathrm{Var}_{P_\lambda}[w^\star(f - \mu_f)]}{K} + o(n_h^{-1} + K^{-1}).$$

Here $\sigma_f^2 = \mathrm{Var}_Q[f]$ *is the target variance and* $\bar{f}_H$ *the direct calibration mean. At fixed calibration data,* $\hat{\mu}_{\mathrm{SNIS}} \to \bar{f}_H$ *as* $K \to \infty$*: proposal draws cannot remove the human-sample floor. Under the audited covariance condition the proposal term is at most* $\sigma_f^2/K_{\mathrm{eff}}$*, replacing the variance term. App. A.2 proves the result and App. C.5 audits the two-dimensional scaling.*

4.3 T2: what support deficiency does and does not imply

Three distinct objects must be kept apart, and the theorems below are stated *only for the first*: $S_{\mathrm{zero}} = \{y : P_\lambda(y) = 0\}$ (*structural* absence, budget-immune); $S_\tau = \{y : P_\lambda(y) < \tau\}$ (resolution-τ diagnostic, $\tau \approx 1/K$, a finite-budget risk object, not a statement about P_λ); and $\mathcal{U}_K = \{y : N_y = 0\}$ (categories unobserved in the *current* K draws). In the structural-zero regime write $\varepsilon = Q(S_{\mathrm{zero}})$, $m_C = \mathbb{E}[f \mid S_{\mathrm{zero}}^c]$, and $\Delta_f = \left|\mathbb{E}[f \mid S_{\mathrm{zero}}] - m_C\right|$.

Theorem 2 (T2a: SNIS truncation limit)**.** *Assume* $0 < Q(S_{\mathrm{zero}}^c) < 1$*, and for the fixed-calibration limit also* $\hat{Q}(S_{\mathrm{zero}}^c) > 0$*. Under the calibration protocol (human draws from* Q*; plug-in weights* $\hat{w} = \hat{Q}/P_\lambda$ *on proposal samples* $y_k \sim P_\lambda$*), the* unclipped *SNIS estimator at fixed calibration set converges, as* $K \to \infty$*, to the covered human sample mean, and if additionally* $n_h \to \infty$*, to* $\mathbb{E}[f \mid S_{\mathrm{zero}}^c]$ *(App. A.3, Part 1); the limiting bias is exactly* $\varepsilon\Delta_f$*. For* S_τ *only finite-budget quantities are claimed:* $\mathbb{E}[N(S_\tau)] = KP_\lambda(S_\tau)$ *and* U_K *(App. A.3).*

Theorem 3 (T2b: distributional lower bound, class-universal)**.** *For* every *distribution R supported on* $\mathrm{supp}(P_\lambda)$*, hence for the output of any reweighting of proposal draws,* $\mathrm{TV}(R, Q) \geq Q(S_{\mathrm{zero}}) = \varepsilon$ *(App. A.3, Part 2). Its finite-budget companion: the empirical reweighted distribution R_K on observed categories satisfies* $\mathrm{TV}(R_K, Q) \geq Q(\mathcal{U}_K)$*, budget-dependent, not a permanent plateau.*

Theorem 4 (T2c: partial identification under unknown stranded outcomes)**.** *If the outcome values f on S_{zero} are* unobservable *(known only to lie in $[a, b]$), then μ_f is partially identified: $\mu_f \in [(1-\varepsilon)m_C + \varepsilon a,\ (1-\varepsilon)m_C + \varepsilon b]$, an interval of width $\varepsilon(b-a)$, with two-outcome minimax bias $\geq \varepsilon(b-a)/2$ (App. A.3, Part 3). This is a worst-case statement over outcomes; when f is a known scoring map it does not apply, and no per-instance mean floor exists: non-negative weights on the covered support can realize any value in $[\min_{\mathrm{supp}(P_\lambda)} f, \max_{\mathrm{supp}(P_\lambda)} f]$, so a zero-bias reweighting may exist, but it is outcome-tuned and certifies nothing about other estimands; estimators reading the calibration draws directly stand outside the reweighting family and are priced by n_h (App. C.2, App. C.6).*

Structural zeros vs. finite-budget non-observation. A structural zero ($P_\lambda(y) = 0$) is budget-immune; a merely unobserved category with $P_\lambda(y) > 0$ will eventually be sampled, so $\hat{\varepsilon}$ at resolution τ speaks to the *current* budget (zero counts never certify $P_\lambda(y) < 1/K$: a category with $P_\lambda(y) = 1/K$ still goes unseen with probability $(1 - 1/K)^K \approx e^{-1}$). The deployable finite-budget object is the expected uncovered mass $U_K := \sum_y Q(y)\left(1 - P_\lambda(y)\right)^K$ (App. A.3 also documents the plug-in pilot's unreliability). Verdicts are budget-relative diagnostics, re-measurable as K grows; the τ-near-empty regime degrades gracefully (91% agreement at $\tau{=}2/K$, 78% at $10/K$; App. C.2).

Thus structural absence prevents distribution recovery (T2b), while mean error also depends on the estimand contrast Δ_f (T2a–c). The empirical versions $\hat{\varepsilon}$ and $\widehat{\Delta_f}$ define the budget-relative audit in §3.3.

4.4 An error budget conditional on fixed support

For independent, unclipped plug-in SNIS on a fixed covered set C, with $\varepsilon = Q(C^c) < 1$, the leading conditional error display is

$$\mathrm{RMSE} \approx \left(\varepsilon^2 \Delta_f^2 + \sigma_C^2 \mathbb{E}\left[\frac{1}{N_C} \,\middle|\, N_C \geq 1\right] + \frac{v_C}{K}\right)^{1/2}, \tag{2}$$

where $N_C \sim \mathrm{Binomial}(n_h, 1-\varepsilon)$, $\sigma_C^2 = \mathrm{Var}_Q(f \mid C)$, and v_C/K denotes the leading proposal variance under truncated weights. The display conditions on $N_C \geq 1$ and omits higher-order terms. It is not the unconditional risk when $p_0 = \varepsilon^{n_h}$ is appreciable: zero-covered-sample cells contribute their fallback risk with probability p_0. Appendix A.2 gives the exact frozen-support risk including that term. Estimating ε does not make N_C equal to n_h.

The decomposition separates mechanisms rather than supplying one bound for every implementation. Reused counts remove the proposal-frequency term conditional on O_K; clipping changes the target of the estimator. Increasing K can also discover previously unobserved categories and thus change both the empirical truncation term and the number of covered calibration observations. Only a fixed structural support keeps those terms unchanged as K grows.

5 Experiments

The evidence has three roles. Known-P, Q synthetic experiments test statistical predictions. Frozen LLM-count experiments characterize a specified empirical deployment. Fresh-stream and held-out experiments test transfer. Across these studies, roughly 640K generations cover four open-weight 7–9B models, a 34B replication, and four task families. Table 1 summarizes the human-target comparisons; App. B records protocols and App. C reports the full battery.

5.1 Matched human-target estimation

We ask 12 GlobalOpinionQA questions verbatim using Gemma-2-9B, Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B under three scaffolds (76,800 generations). Calibration samples of size $n_h \in$

Study	Matched calibration baseline	Plateau-label prediction
12 matched GOQA items	Direct beats deployed SNIS in 270/288 configurations (2 ties)	Accuracy 0.91; majority baseline 0.56
34B replication	Direct wins 32/36 cells	Accuracy 0.949
48-item expansion, 12 dev / 36 test	Direct MAE 0.112 / 0.040; every reweighting variant ≥ 0.53	Test accuracy 0.97 / 0.98 at $n_h = 50/400$

Table 1: **Human-target estimation and plateau prediction.** Each question supplies n_h calibration draws from released GOQA marginals. Accuracies aggregate calibration replicates; win counts compare cell-level error. The 288 configurations comprise 12 questions $\times 4$ models $\times 3$ scaffolds $\times 2$ calibration budgets.

$\{50, 400\}$ are resampled from released national marginals. This preserves each question's marginal distribution, not a joint respondent-level survey. Frozen LLM counts use $K = 800$ for neutral scaffolds and 400 for persona scaffolds. The deployed SNIS clips ratios to $[0.02, 50]$; all 288 configurations remain in the analysis.

The direct calibration mean has strictly lower RMSE than deployed SNIS in **270/288** configurations, with two ties, and beats the residual-corrected calibration blend in 246/288. Its median RMSE is 0.091 and 0.032 at $n_h = 50$ and 400, respectively; the reported deployed-SNIS median is 0.320. The question-clustered interval for the direct-win fraction is $[0.889, 0.979]$ (App. C.1). A 34B replication gives 32/36 direct wins; a separate 48-item expansion, with 12 development and 36 held-out questions, gives direct MAE 0.112/0.040 versus at least 0.53 for every evaluated reweighting variant (App. C.3, C.5).

5.2 What causes the estimation gap?

In 24 covered deployment cells, the direct RMSE at $n_h = 400$ is 0.0148, close to its predicted 0.0150. The four-way ablation reports double-clipped SNIS at 0.1027 and unclipped SNIS at 0.0245, a factor-4.2 reduction (App. C.6). The main deployment's corresponding clipped value is 0.1039. The remaining gap is consistent with estimated-weight and proposal noise in that protocol; the oracle-weight reference reaches 0.0133 using additional target information. Decoupling numerator and denominator calibration yields 0.1285. Clipping is therefore a major source of error in these covered deployment cells.

Across the tested budgets $K \leq 6400$, the direct mean remains better than plug-in SNIS (App. C.6). With known P and independent streams, the 4×4 calibration/proposal-budget grid tests the asymptotic decomposition separately: across 640 cells, measured/predicted MSE has median 1.00 and IQR $[0.948, 1.042]$; 639/640 cells satisfy SNIS MSE at least as large as direct MSE within 2% tolerance (App. C.5). This supports the leading-order theory in the tested regime; the exact finite-sample result remains the unnormalized comparison in Prop. 2.

5.3 Support and estimand effects under a frozen protocol

The mechanism study uses neutral Gemma-2-9B as a *model teacher*, not a human target. Four models, each under a neutral scaffold and two persona doses, answer 12 Likert questions scored 1–5. Target frequencies come from the teacher, while proposal-specific ratios are constructed for each model–scaffold pair. The strongest Qwen persona produces median TV distance 0.9998 and mean-score bias of 1–2 points.

Across the 11 non-self pairs, recovery and empirical uncovered mass have Spearman correlation -0.98 (question-clustered CI $[-0.99, -0.81]$, permutation $p = 10^{-4}$; Table 2). Pairs with $\hat{\varepsilon} \approx 0.96$–0.99 have recovery at most 0.13. This association is partly imposed by reused-count cancellation: the estimated truncation magnitude and the reweighted mean share the same calibration frequencies and observed support. Parse failures are 9.6% for the teacher and at most 0.2% elsewhere; treating parse failure as an explicit bin changes 0/11 pair labels (App. C.1).

Pair	$\hat{\varepsilon}$	recovery	plateau	naive bias	Verdict
gemma-neutral (self)	0.000	—	0.00	0.00	reference
gemma-p_1	0.958	0.124	1.37	1.99	P
gemma-p_2	0.994	0.000	1.99	1.99	P
llama-neutral	0.000	0.986	0.00	1.08	R
llama-p_1	0.000	0.962	0.00	1.35	R
llama-p_2	0.339	0.634	0.46	1.67	R
mistral-neutral	0.000	0.966	0.00	1.31	R
mistral-p_1	0.003	0.914	0.00	1.90	R
mistral-p_2	0.002	0.895	0.00	1.91	R
qwen-neutral	0.044	0.775	0.03	1.05	R
qwen-p_1	0.991	0.003	1.27	1.80	P
qwen-p_2	0.994	0.000	1.99	2.00	P

Table 2: **Recovery toward a model-teacher target.** The target is neutral Gemma-2-9B. Each numeric column reports its median over 12 Likert questions at $n_h = 400$; plateau and naive bias use the original 1–5 score scale. Recovery is the fraction of naive bias removed. R/P classify the plateau as below/at least $0.3\times$ naive bias (§3.3).

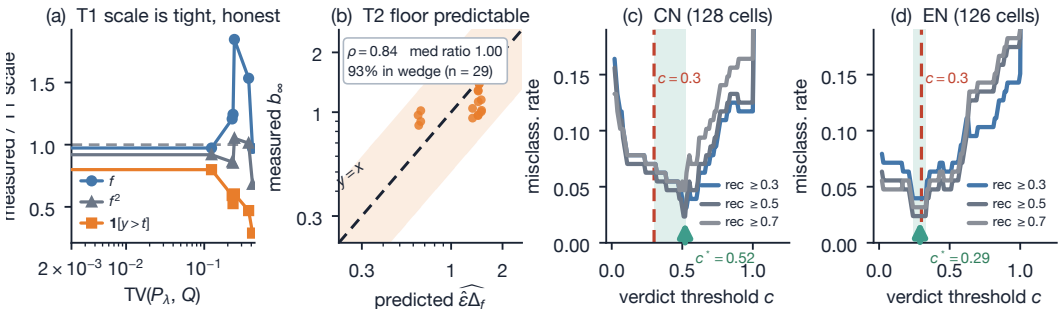


Figure 2: **Controlled-coverage, plateau, and threshold validation.** (a) Under controlled coverage, measured RMSE stays within $1.85\times$ of the ESS scale $\sigma_f/\sqrt{K_{\text{eff}}}$ under covariate shift (0.98–1.06 at zero shift). (b) The plateau magnitude is predictable before deployment: measured against predicted $\hat{\varepsilon}\widehat{\Delta_f}$, the median ratio is 1.00 and $\rho = 0.844$. (c,d) The screen constant $c = 0.3$ sits at the elbow of the leave-one-out misclassification surface on the Chinese and English item banks ($c^* \approx 0.52$ and $c^* \approx 0.29$); curves use recovery-based reference labels at recovery thresholds 0.3, 0.5, and 0.7.

Multiple-choice, rubric-scored text, and tool-choice tasks show why the estimand matters. The rubric family has $\hat{\varepsilon} \approx 0$ and recovery 0.978. MCQ has median $\hat{\varepsilon} = 0.024$ and contrast $\Delta_f = 0.79$, with measured plateau 0.022 versus predicted 0.024. Fine-grained score bins can remain covered even under TV distance 0.54, while an indicator on the same task can expose a missing category (App. C.2). A model judge also changes the target scoring function; its repeat-call variance is analyzed separately from human-target error (App. C.4).

Across the 30 empirically uncovered cells reported in the plateau analysis, measured and predicted magnitudes have Spearman correlation 0.844 and median ratio 1.00; 93% lie within a factor of two and 97% within the pre-specified $[0.3, 5]$ screen. Fig. 2 collects these frozen-protocol checks alongside the covered-support and threshold studies.

Across calibration budgets $n_h \in \{50, 150, 400, 1200\}$, covered cells whose weights escape the clip range follow the calibration-floor slope in n_h (measured log–log slope -0.49), clip-limited cells sit on a clipping-bias floor, and uncovered cells stay flat at the predicted $\hat{\varepsilon}\widehat{\Delta_f}$ plateau from $n_h = 50$ onward (App. C.6). The plateau prediction is stable to diagnostic noise: a ± 2-SE bootstrap band on $\widehat{\Delta_f}$ leaves the in-gate proportion at 97% and moves the median measured/predicted ratio from 1.00 to 0.95 (App. C.2).

5.4 Does diagnostic accuracy translate into decision value?

On matched GOQA questions the plateau-label screen achieves accuracy 0.91 (clustered CI $[0.88, 0.94]$), versus majority-class accuracy 0.56 from 162 P / 126 R reference labels. P recall is 0.97; R recall is 0.84 and precision 0.96. A fresh-draw rerun yields accuracy 0.935. Held-out questions with frozen rules yield 0.97/0.98 at $n_h = 50/400$, and the 34B replication yields 0.949.

The decision tests are less favorable. Agreement with the protocol's diagnostic labels is 0.951, while agreement with the different label *reweighting beats the naive estimator* is 0.176 in the corresponding audit (App. C.5). Diagnostic-guided routing has MAE 0.772/0.873, compared with 0.751/0.871 for a rate-matched coin. Routing gains come from choosing direct estimation; re-scaffolding alone matches always reweighting. In allocation studies, all 12 question-clustered intervals for diagnostic targeting versus random allocation cover zero. Direct sampling also outperforms the tested targeted and random simulation allocations (App. C.6). Thus accurate plateau prediction has descriptive value, but these experiments do not establish incremental utility for routing or allocation.

Fresh streams also expose support uncertainty. In the most extreme cell, pilot uncovered mass 0.990 falls to 0.552 on a fresh 800-draw stream. The finite-budget diagnostic can therefore overstate a persistent gap even when it describes the pilot accurately. Repeated pilots and stopping rules did not systematically resolve this discrepancy in the tested cells (App. C.5).

5.5 Robustness and scope of the evidence

Appendix C varies clipping, estimators, resolution, parsing, language, model size, item bank, and data flow. Heavy-tail tests reveal ESS and variance-estimation failures; controlled guards have low false-positive rates but power only 0–0.017, so a silent guard is uninformative about safety. Rare-share studies favor simple shrinkage over reweighting: linear shrinkage, direct estimation, and empirical Bayes have reported RMSE 0.0039, 0.0041, and 0.0045, respectively, versus 0.096 for reweighting. The auxiliary population-bridge and bandit examples concern different information structures, including covariates and outcome corrections; they illustrate where such corrections can help without contradicting the categorical identities (App. C).

6 Conclusion

BiasFate connects the construction of importance weights to the information and error in reweighted LLM simulations. Unclipped weighting with reused counts reproduces the calibration mean on observed categories; independent known-proposal draws add proposal variance on top of the calibration floor, which no simulation budget removes. Support truncation sets a bias plateau governed by the estimand, and clipping adds a separate, deployment-specific distortion. Across the evaluated human-target configurations, direct calibration estimation is the stronger baseline: it wins 270 of 288 matched configurations, and removing clipping recovers most of the covered-cell gap. Plateau prediction transfers to fresh draws and held-out questions, while diagnostic-guided routing and allocation show no improvement over matched controls; the audit's practical value is knowing when to fall back to direct estimation, not how to route simulation. These findings motivate evaluating diagnostics through both the errors they predict and the estimation loss of the decisions they induce. The identities behind these results are statements about information: what the calibration sample already contains, what additional draws can contribute, and what no reweighting of a deficient proposal can restore. The resulting recommendation for the evaluated setting is concrete: use the calibration mean as the first baseline at any budget, and require any simulation-based correction to demonstrate improvement at matched information and sampling budgets before deployment.

7 Limitations

The main results concern fixed finite response spaces, known scoring maps, and per-question target calibration. Finite LLM samples identify unvisited categories, not structural zeros, and parsers and model judges define the evaluated estimand. Human-target validation covers one survey resource with marginal resampling, mostly 7–9B models, and strong persona interventions; the supplementary bridge and policy examples use different data structures.

REFERENCES

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pp. 405–431, 2017.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, November 2023. ISSN 1095-9203. doi: 10.1126/science.adi6000.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Christopher A Bail. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121, 2024.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Zihan Chen, Di Zhu, and Lei Nico Zheng. When synthetic users fail: A cross-domain benchmark of llm-simulated human survey responses. *arXiv preprint arXiv:2607.26348*, 2026.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. *Sample Selection Bias Correction Theory*, pp. 38–53. Springer Berlin Heidelberg, 2008. ISBN 9783540879879. doi: 10.1007/978-3-540-87987-9_8.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4): 427–444, December 1940. ISSN 0003-4851. doi: 10.1214/aoms/1177731829.

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, June 1992. ISSN 1537-274X. doi: 10.1080/01621459.1992.10475217.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37: 45850–45878, 2024.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, December 1953. ISSN 0006-3444. doi: 10.2307/2333344.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024.

Chengpiao Huang, Yuhang Wu, and Kaizheng Wang. How many human survey respondents is a large language model worth? an uncertainty quantification perspective. *arXiv preprint arXiv:2502.17773*, 2025.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, June 2008. ISSN 1537-2715. doi: 10.1198/106186008x320456.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *Biometrika*, pp. asag053, 2026.

Shakeeb Khan and Elie Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

Chan-Tung Ku, Chan Hsu, Pei-Cing Huang, Frank Cheng-shan Liu, I-Ling Cheng, and Yihuang Kang. Silicon sampling via cross-survey transfer. In *2026 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 338–343. IEEE, 2026.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Erika Elizabeth Taday Morocho, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Assessing the reliability of persona-conditioned llms as synthetic survey respondents. *arXiv preprint arXiv:2602.18462*, 2026.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, April 2001. ISSN 1573-1375. doi: 10.1023/a:1008923215028.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. Stanford University, 2013. Freely available online.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, et al. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a Priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, January 2002. ISSN 1530-888X. doi: 10.1162/089976602753284446.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Robson Tigre and Hugo Gobato Souto. When can you trust your synthetic users? diagnostics and corrections for llm consumer panels. *arXiv preprint arXiv:2609.13148*, 2026.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26268–26300, 2025.

Tristan Williams, Franziska Weeber, Sebastian Padó, and Alan Akbik. Beyond marginal distributions: A framework to evaluate the representativeness of demographic-aligned llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4806–4826, 2026.

Shirong Xu and Will Wei Sun. Do more predictions improve statistical inference? filtered prediction-powered inference. *arXiv preprint arXiv:2602.10464*, 2026.

REPRODUCIBILITY STATEMENT

All experiments are generated by versioned scripts with fixed seeds; every figure and headline number regenerates from the supplementary result files. The generation ledger is summarized in App. B.3. Original plots are generated from the supplementary result files; the revised summary figure and tables are typeset directly from numerical results reported in this manuscript, with editable LaTeX source. The fresh-stream cell selection uses only frozen counts through a frozen rule (supplementary selection file, full content hash in the supplementary manifest); the held-out dev/test split and pre-generation frozen analysis rules are fixed in the supplementary metadata file before any test-side generation. All raw generation counts, prompts, banks, and analysis scripts are provided as a single supplementary artifact tree; each file's content hash is listed in the supplementary manifest.

ETHICS STATEMENT

This work uses no new human-subjects data: human "targets" are publicly released survey marginals (Anthropic GlobalOpinionQA / Pew), used per their release terms; no personally identifiable information is involved. LLM draws are synthetic text only. We disclose a dual-use consideration: the same calibration machinery that makes LLM simulation honest could be used to make synthetic respondents harder to detect; BiasFate's outputs are *diagnostics* of estimator validity, they certify neither demographic accuracy nor fitness for decisions about real populations, and the paper's verdicts are explicitly budget-relative and non-portable. Survey research on human subjects carries its own review obligations, which remain unchanged by the tools proposed here.

AI USE STATEMENT

In the experiments, LLMs generated the synthetic responses and supplied model-based scores as specified in the experimental protocols. We also used generative AI tools for literature retrieval, manuscript review and revision, feedback on mathematical statements and proofs, interpretation of reported results, organization of the paper, writing and polishing, and preparation of LaTeX figures and tables. The authors take full responsibility for the final content, including the correctness of all technical claims, results, and references.

A PROOFS

Throughout, fix a question x, write $P = P_\lambda(\cdot \mid x)$ and $Q = Q(\cdot \mid x)$, and normalize $f : \mathcal{Y} \to [0, 1]$ unless stated otherwise. For a score in $[a, b]$, apply the proofs to $(f - a)/(b - a)$; absolute errors and standard deviations rescale by $b - a$, and variances by $(b - a)^2$. Experiments report their original score units. Write $\sigma_f^2 = \mathrm{Var}_Q(f)$.

A.0 NOTATION

symbol	meaning
$P = P_\lambda(\cdot \mid x)$	proposal distribution (LLM sampler)
Q	target distribution (human population of interest)
K	simulation budget (proposal draws)
n_h	human calibration budget
f, σ_f	estimand and its target-spread, $\sigma_f^2 = \mathrm{Var}_Q(f)$
μ_f	$\mathbb{E}_Q f$ (target mean)
$\hat{\varepsilon}$, τ	support-deficiency mass at resolution τ
Δ_f	functional gap between uncovered and covered conditional means
K_{eff}, ESS	effective sample size of the estimated weights
α, $w_{\max}$	clip level and cap: $w \in [\alpha, 1/\alpha]$, $\alpha = 1/w_{\max}$ ($\alpha = 1$ forces all weights to 1 (naive corner), $\alpha \to 0$ removes the clip)
c, c^*	verdict threshold (R iff plateau $< c \cdot$ naive) and its data-driven optimum
plateau / naive	$\hat{\varepsilon}\,\widehat{\Delta_f}$ (estimated truncation magnitude) / $\lvert\mathbb{E}_Q f - \mathbb{E}_P f\rvert$ (uncorrected bias)
TV	total variation distance between P and Q
T1 / T2	theoretical covered / structural-deficiency settings

A.1 Proof of Theorem 1 (T1, oracle weights)

Conventions (global, matching §3/§4). All expectations carry an explicit measure subscript: the proposal samples are $y_1, \dots, y_K \sim P = P_\lambda(\cdot \mid x)$, the target is Q, the oracle weights are $w^\star = Q/P$, and $\mu_f = \mathbb{E}[f]$, $\sigma_f^2 = \mathrm{Var}_Q(f)$. The theorem is a statement about the *oracle* estimator (the plug-in case is Cor. 1.1/A.2). The two workhorse identities, valid for any $g \geq 0$ on $\mathrm{supp}(Q) \subseteq \mathrm{supp}(P)$:

$$\mathbb{E}_P[w^\star g] = \sum_y P(y)\tfrac{Q(y)}{P(y)} g(y) = \mathbb{E}[g], \qquad \mathbb{E}_P[(w^\star)^2 g] = \sum_y P(y)\tfrac{Q(y)^2}{P(y)^2} g(y) = \mathbb{E}[w^\star g]. \quad (3)$$

Define $\rho_1 := \mathbb{E}_P[(w^\star)^2] = \sum_y Q(y)^2/P(y) = \mathbb{E}_Q[w^\star]$ (the two measures disagree on $\mathbb{E}[w^\star]$ itself, $\mathbb{E}_P[w^\star] = 1$ vs. $\mathbb{E}_Q[w^\star] = \rho_1$; the second moment under P equals the first moment under Q), $M := \|w^\star\|_\infty \leq 1/p_{\min}$ under coverage (the weights are Q/P, not $1/P$), and $R := \|f - \mu_f\|_\infty$.

Setup. Let $y_1, \dots, y_K \sim P$ i.i.d. and define the self-normalized estimator $\hat{\mu} = \frac{\sum_k w^\star(y_k) f(y_k)}{\sum_k w^\star(y_k)}$.

Step 1 (stabilization and variance). $\hat{\mu} - \mu_f = \frac{\frac{1}{K}\sum_k (w^\star(f-\mu_f))(y_k)}{\frac{1}{K}\sum_k w^\star(y_k)}$, note $\mathbb{E}_P[w^\star] = 1$, so the numerator centers at zero exactly. Define the per-draw variables $X_k := w^\star(y_k)(f(y_k) - \mu_f) \in [-MR, MR]$ with $\mathrm{Var}_P(X_k) = v_f := \mathrm{Var}_P(w^\star(f-\mu_f))$ and $Z_k := w^\star(y_k) - 1$ with $\mathrm{Var}_P(Z) = \rho_1 - 1$. By the delta method for self-normalized sums, $\hat{\mu} - \mu_f$ is asymptotically normal with variance v_f/K.

Step 2 (the finite-sample form). Apply Bernstein's inequality to the numerator and denominator separately and union-bound. With probability $\geq 1 - \delta$, writing $t = \log(4/\delta)$:

$$\left|\tfrac{1}{K}\textstyle\sum_k X_k\right| \leq \sqrt{\tfrac{2v_f t}{K}} + \tfrac{2MRt}{3K}, \qquad \left|\tfrac{1}{K}\textstyle\sum_k Z_k\right| \leq a_K := \sqrt{\tfrac{2(\rho_1-1)t}{K}} + \tfrac{2Mt}{3K}. \quad (4)$$

The denominator bound carries the same *linear* range term as the numerator ($|Z| = |w^\star - 1| \leq \max\{1, M-1\} \leq M$, the last step using $\mathbb{E}_P[w^\star] = 1 \Rightarrow M \geq 1$); a statement that only requires $K \gtrsim (\rho_1 - 1)t$ would drop it. Under the deterministic condition $a_K < 1$, which holds once K exceeds a constant times $((\rho_1 - 1) + M)\, t$, the denominator satisfies $1 + \bar{Z} \geq 1 - a_K > 0$ directly (no auxiliary event needed), and dividing gives

$$|\hat{\mu} - \mu_f| \;\leq\; \frac{\sqrt{2v_f \log(4/\delta)/K} + \frac{2}{3}MR\log(4/\delta)/K}{1 - a_K}, \quad (5)$$

the finite-sample form of Thm. 1 (after Agapiou et al., 2017; Owen, 2013): explicit in v_f, $M \leq 1/p_{\min}$, R, and the denominator-deviation term a_K with its linear component. This is the honest structural reading: the *exact leading variance* is v_f/K, while the K_{eff} form is an upper bound under the audited covariance condition; M and ρ_1 enter the range term and the denominator event.

Step 3 (two spread forms for v_f). First, $\mathbb{E}_P[w^\star(f - \mu_f)] = \mathbb{E}_Q[f - \mu_f] = 0$ by the importance identity, so $v_f = \mathbb{E}_P[(w^\star)^2(f-\mu_f)^2] = \mathbb{E}_Q[w^\star(f-\mu_f)^2]$ exactly (second identity above with $g = (f-\mu_f)^2$; the two measures generally disagree on $\mathbb{E}[w^\star]$ itself, $\mathbb{E}_P[w^\star] = 1$ vs. $\mathbb{E}_Q[w^\star] = \rho_1$, but agree on this weighted second moment). Second, under the negative-association condition $\mathrm{Cov}_Q(w^\star, (f-\mu_f)^2) \leq 0$ (large weights concentrate where f is near its target mean), $v_f \leq \rho_1 \sigma_f^2$ and the first-order term admits the upper-bound form $\sigma_f\sqrt{2\log(4/\delta)/K_{\mathrm{eff}}}$ with $K_{\mathrm{eff}} = K/\rho_1$ (an upper bound, not an identity: the exact leading variance v_f/K is estimand-dependent). Unconditionally, $v_f \leq R^2\rho_1$ (the range form of Thm. 1). Cell-level audits of the association condition and of the resulting display are reported in App. C.1 and App. C.7. □

Remark (scope of the oracle bound). Theorem 1 concerns oracle weights. Corollary 1.1 instead concerns independent, known-P, *unclipped* plug-in SNIS and gives a leading-order MSE decomposition. Neither statement is a concentration guarantee for the deployed double-clipped estimator. Its changed limit is described in App. A.4 and its error is measured in App. C.2, C.6, and C.7.

Remark (first-order ESS). Under the stated covariance condition, ESS determines a convenient *upper bound* on the leading variance term: two samplers with equal K_{eff} share the same bound $\sigma_f\sqrt{2\log(4/\delta)/K_{\mathrm{eff}}}$ regardless of TV distance, consistent with the controlled-coverage shift sweep of Fig. 2a, where the bound ratio stays in $[0.68, 1.85]$ across the whole sweep (0.98–1.06 at zero shift). The exact asymptotic variance v_f/K remains estimand-dependent (it involves

$\mathbb{E}[w^\star(f-\mu_f)^2]$, not ρ_1 alone), and at finite K the range term and denominator event additionally involve M and ρ_1 (A.1 Step 2).

A.2 Two-sample plug-in analysis (Cor. 1.1 and Prop. 2)

Replace $w^\star$ by the plug-in $\hat{w}(y) = \hat{Q}(y)/P(y)$ where $\hat{Q}$ is the empirical distribution from n_h calibration draws $z_1, \dots, z_{n_h} \sim Q$. Write $\hat{w} = w^\star + \Delta$, $\Delta(y) = (\hat{Q} - Q)(y)/P(y)$; note $\mathbb{E}_P[\Delta] = \sum_y (\hat{Q}(y) - Q(y)) = 0$ and $\mathbb{E}_P[\hat{w}] = 1$ exactly for every realization of $\hat{Q}$. Expanding the SNIS ratio to first order in Δ (conditional on $\hat{Q}$, i.e. over the proposal draws):

$$\hat{\mu}_{\hat{w}} \;\approx\; \mu_f \;+\; \underbrace{\tfrac{1}{K}\textstyle\sum_k (w^\star(f-\mu_f))(y_k)}_{\text{oracle sampling noise}} \;+\; \underbrace{\tfrac{1}{K}\textstyle\sum_k \Delta(y_k)\,(f(y_k)-\mu_f)}_{\text{weight noise}}. \tag{6}$$

Conditionally on $\hat{Q}$, the second term is $\frac{1}{K}\sum_k \Delta(y_k)(f(y_k) - \mu_f)$ with proposal-side mean $\mathbb{E}_P[\Delta(f-\mu_f)] = \sum_y (\hat{Q}(y) - Q(y))\, f(y)$ (using $\mathbb{E}_P[\Delta\,\mu_f] = \mu_f \sum_y (\hat{Q}(y) - Q(y)) = 0$), an *additive* $O(n_h^{-1/2})$ term: its law is that of a mean of n_h i.i.d. Q-draws of f, i.e. it converges to the human sample mean error $\bar{f}_h - \mu_f$. Hence, to first order, the MSE decomposes *additively*:

$$\mathrm{MSE}(\hat{\mu}_{\hat{w}}) \;=\; \underbrace{\frac{\sigma_f^2}{n_h}}_{\text{calibration floor}} \;+\; \underbrace{\frac{\mathrm{Var}_P[w^\star(f-\mu_f)]}{K}}_{\text{sampling term}} + o\Big(\tfrac{1}{n_h} + \tfrac{1}{K}\Big), \tag{7}$$

where $\mathrm{Var}_P[w^\star(f-\mu_f)] = v_f$ as in A.1 (denominator K; under the audited covariance condition the *upper bound* $\sigma_f^2/K_{\mathrm{eff}}$ may replace it, but the two forms are alternatives, not additive terms) and the floor term $\sigma_f^2/n_h = \mathrm{Var}(\bar{f}_h)$ is the variance of the human sample mean itself, the weight-noise term re-creates exactly the estimator the calibration set already provides, up to $O(1/\sqrt{K})$ fluctuations of order $\sqrt{d_{\mathrm{eff}}/(Kn_h)}$ (for categorical $\mathcal{Y}$, $\mathbb{E}_{\mathrm{cal}}\,\mathbb{E}_P[\Delta^2] = \frac{1}{n_h}\sum_y \frac{Q(y)(1-Q(y))}{P(y)} \le \frac{d_{\mathrm{eff}}}{n_h}$, with $d_{\mathrm{eff}} := \sum_y Q(y)/P(y) \ge 1$ by Cauchy–Schwarz). Consequently, as $K \to \infty$ at fixed n_h, $\hat{\mu}_{\hat{w}} \to \bar{f}_h$ (a.s. in K for each realized calibration set; *unclipped*, the deployed clipped estimator has the modified limit of A.4): the plug-in SNIS estimator *re-creates the human sample mean*, so the $\sigma_f/\sqrt{n_h}$ floor is irreducible by any proposal budget. Two consequences structure the paper. (i) Increasing K and increasing n_h address *different* error terms: K shrinks only the sampling term, n_h shrinks only the floor. (ii) For the per-x mean estimand the plug-in estimator shares the direct human mean's fixed-calibration asymptote, the analysis *explains* the calibration floor measured in the matched-budget study (App. C.6), rather than predicting a previously unmeasured quantity. Whether proposal samples buy anything at finite K for distributional/tail estimands is an empirical question this paper does not settle: direct human estimators (the empirical distribution $\hat{Q}_h(y) = \frac{1}{n_h}\sum_i \mathbf{1}\{z_i = y\}$, tail frequencies, quantiles) exist at the usual $1/n_h$ rates. The additive form is the correct first-order structure (the conditional variance ratio must not be conflated with the unconditional MSE), with the weight-noise term bounded by $\sqrt{d_{\mathrm{eff}}/n_h} \cdot R/\sqrt{K}$ around the floor.

Proof of Proposition 2. Conditional on H, independence and $Q \ll P$ give

$$\mathbb{E}[\widetilde{\mu}_K \mid H] = \sum_y P(y)\frac{\hat{Q}_H(y)}{P(y)} f(y) = \sum_y \hat{Q}_H(y) f(y) = \bar{f}_H.$$

The conditional bias relative to μ is therefore $\bar{f}_H - \mu$. Conditional variance is

$$\mathrm{Var}(\widetilde{\mu}_K \mid H) = \frac{1}{K}\left\{\sum_y \frac{\hat{Q}_H(y)^2 f(y)^2}{P(y)} - \bar{f}_H^2\right\}.$$

Applying the conditional MSE decomposition and $\mathbb{E}(\bar{f}_H - \mu)^2 = \sigma_f^2/n_h$ proves the identity. The direct estimator is the conditional expectation of $\widetilde{\mu}_K$, which is precisely the Rao–Blackwell improvement. □

Lemma 1 (Same-sample guards are algebraically capped)**.** *If a guard's denominator $\hat{p}$ is the empirical distribution of the* same *K proposal draws on which the deployment is evaluated, and its sum ranges over the observed support $O_K = \{y : \hat{p}_y > 0\}$, then for any calibration-derived $\hat{q}$,*

$$\hat{\rho}_1 \;=\; \sum_{y \in O_K} \frac{\hat{q}_y^2}{\hat{p}_y} \;\le\; K \sum_{y \in O_K} \hat{q}_y^2 \;\le\; K, \tag{8}$$

since $\hat{p}_y \ge 1/K$ on O_K. The strict fire condition $K < \hat{\rho}_1$ can therefore never occur; the cap $\hat{\rho}_1 = K$ is reached only when the calibration sample concentrates (nearly) all its mass on a single observed minimally-counted bin ($\sum_{O_K} \hat{q}^2 \to 1$ with $\hat{p}_y = 1/K$ there), so such a guard is vacuous as a heavy-tail detector by construction*, not by information limitation.*

Proposition 3 (Guard power under false-positive control)**.** *Let $T \subseteq \mathcal{Y}$ carry target mass $\varepsilon_T = Q(T) > 0$, fix the proposal P, and let a guard be any (possibly randomized) function G of the n_h calibration draws with $G \in [0,1]$. Let Q_0, Q_1 be two targets that agree outside T and both place mass ε_T on T, differing only in how they distribute it. If the guard has false-positive rate $\mathbb{E}_{Q_0}[G] \le \alpha_{\mathrm{FP}}$ under Q_0, then*

$$\mathbb{E}_{Q_1}[G] \;\le\; \min\Big\{1,\; \alpha_{\mathrm{FP}} + \mathrm{TV}\big(Q_0^{\otimes n_h},\, Q_1^{\otimes n_h}\big)\Big\} \;\le\; \min\Big\{1,\; \alpha_{\mathrm{FP}} + 1 - (1-\varepsilon_T)^{n_h}\Big\}, \tag{9}$$

because the two laws of the calibration sample differ only on the event that at least one draw lands in T, of probability $1-(1-\varepsilon_T)^{n_h}$. Three readings. (i) This is a power cap at controlled false-positive rate *against within-T redistribution; it says nothing about guards without an α_{FP} contract (a guard that always fires has power 1 and false-positive rate 1). (ii) It prices "escalate n_h" as a* necessary *condition only: power $p > \alpha_{\mathrm{FP}}$ against tail mass ε_T requires $n_h \ge \ln(1-p+\alpha_{\mathrm{FP}})/\ln(1-\varepsilon_T)$ (e.g. $n_h \gtrsim 598$ for power 0.5 at $\alpha_{\mathrm{FP}} = 0.05$, $\varepsilon_T = 10^{-3}$, the no-false-positive version needing ≈ 693); one tail draw does not by itself license an estimate of Q/P on T. (iii) The cap depends on the* set *mass ε_T, not on per-bin rarity: tail mass spread over many per-bin-rare bins is detectable whenever the set as a whole is sampled.*

Lemma 2 (The plug-in $\hat{\rho}_1$ is not one-sided)**.** *With P known and calibration draws $Y_i \sim Q$,*

$$\mathbb{E}[\hat{\rho}_1] \;=\; \rho_1 \;+\; \frac{1}{n_h} \sum_{y} \frac{Q(y)\,(1-Q(y))}{P(y)} \;\ge\; \rho_1, \tag{10}$$

so $\hat{\rho}_1$ is upward-biased in mean*: the heavy-tail median under-reading measured in the stress panel of App. C.7 coexists with rare-draw over-reads, and neither direction is uniform. Worked counterexample to "a fire is decisive": $P = Q = (0.99,\, 0.01)$, so $\rho_1 = 1$; a single calibration draw landing on the rare bin yields $\hat{Q} = (0,1)$, $\hat{\rho}_1 = 100$, and the guard fires for any $K < 100$ (e.g. $K = 10 \gg \rho_1 = 1$), a false alarm with per-draw probability 0.01. A fire is therefore a* heuristic warning *whose false-positive behavior must itself be calibrated (the Clopper–Pearson construction of App. C.7), not a certificate that $K \lesssim \rho_1$.*

Clipping; and the clipped population limit. With $w \leftarrow \min(w, w_{\max})$, the bias is driven by the clip set $B = \{w^\star > w_{\max}\}$, but *no bound of the form $\|f - \mu_f\|_\infty Q(B)$* holds for the clipped bias: take $Q = (0.2, 0.8)$, $P = (0.001, 0.999)$, $f = (1,0)$, $w_{\max} = 1$; then $Q(B) = 0.2$, $\|f - \mu_f\|_\infty Q(B) = 0.16$, while the clipped SNIS value is $0.001/(0.001 + 0.999 \times \frac{0.8}{0.999}) \approx 0.00125$, a bias of $\approx 0.199 > 0.16$. The valid starting point is the exact identity, for any non-negative $\tilde{w}$: $\mu_{\tilde{w}} - \mu_f = \mathbb{E}_P[(\tilde{w} - w^\star)(f - \mu_f)] \,/\, \mathbb{E}_P[\tilde{w}]$. The denominator change on B enters at the same order as the numerator change, so a finite-sample bias bound would require a lower bound on $\mathbb{E}_P[\tilde{w}]$ (a $Q(B)$-type argument supplies one under the oracle upper clip, $\mathbb{E}_P[\min(w,c)] \ge 1 - Q(B)$, but not for the deployed double clip); we do not pursue such a bound. What survives is a second-moment statement at the numerator level, valid under the *oracle upper clip* $\tilde{w} = \min(w^\star, 1/\alpha)$: since $\tilde{w}^2 \le w^\star/\alpha$ pointwise, $\mathbb{E}_P[\tilde{w}^2 (f-\mu_f)^2] \le \frac{1}{\alpha}\,\mathbb{E}_P[w^\star (f-\mu_f)^2] = \sigma_f^2/\alpha$ (the last step the importance identity). The same one-line argument extends *to support deficiency*: proposal draws never land on S_{zero}, so $\tilde{w} = 0$ there and $\mathbb{E}_P[\tilde{w}^2 (f-\mu_f)^2] \le \frac{1}{\alpha}\,\mathbb{E}_Q[(f-\mu_f)^2 \mathbf{1}_C] = \sigma_{Q,C}^2/\alpha$, with $\sigma_{Q,C}^2$ the second moment truncated to the covered region $C = \{P > 0\}$, only the replacement of the full-target variance by the truncated one (with A.3's modified centering when $\varepsilon > 0$) is required. Its limits: it is a numerator second moment, not an SNIS MSE bound (nothing displayed controls the denominator $\mathbb{E}_P[\tilde{w}]$); it concerns the deterministic population weight, not plug-in $\hat{w}$; and the deployed double

clip breaks the pointwise inequality on $\{w^\star < \alpha^3\}$, a covered-region event (the lower clip never acts on S_{zero}, where $\tilde{w} = 0$ by construction). For the deployed double clip the finite-category representation $\hat{P}(y)\,\text{clip}(\hat{Q}(y)/\hat{P}(y),\, \alpha,\, 1/\alpha) = \text{clip}(\hat{Q}(y),\, \alpha\hat{P}(y),\, \hat{P}(y)/\alpha)$ makes the clipped estimator exactly computable per category; an audit over the paper's own cells evaluates the resulting clipped-versus-unclipped decomposition, so clipping bias is quantified on measured grids rather than through a closed-form bound. The $K \to \infty$ limit $\bar{f}_h$ above is an *unclipped* statement: the deployed double clip $\tilde{w} = \text{clip}(\hat{Q}/P,\, \alpha,\, 1/\alpha)$ at fixed calibration set has the different population limit

$$\mu_\alpha(\hat{Q}) \;=\; \frac{\sum_y P(y)\,\tilde{w}(y)\,f(y)}{\sum_y P(y)\,\tilde{w}(y)}, \tag{11}$$

which equals $\bar{f}_h$ only when the clip is inactive on the realized support; the measured clip-level curves should be read against $\mu_\alpha(\hat{Q})$, not against $\bar{f}_h$ (App. A.4). □

Proposition 1 (restated in the main paper). *Proof.* On $\{y : \hat{P}(y) > 0\}$, $\hat{P}(y)\hat{w}(y) = \hat{Q}(y)$ identically; categories with $\hat{P}(y) = 0$ contribute zero to both numerator and denominator. □

Scope of the cancellation. (a) *Conditionally on the frozen proposal counts* (the analysis convention of the battery: frozen counts treated as the population), the only randomness is calibration-side ($\hat{Q}$). (b) *Unconditionally, over the true generation process*, the observed-support randomness remains: O_K is random, and the unconditional risk includes support-discovery error, which the calibration-only analysis does not see. (c) *Under the deployed double clip*, the positive-count frequencies re-enter through clip activation: the ratio $\hat{Q}/\hat{P}$, not just its support, determines where the clip binds, so even conditionally on O_K the estimator depends on the proposal frequencies.

Three consequences. (i) *No proposal-sampling term by construction.* These experiments measure a frozen-empirical-proposal deployment, the frozen counts are treated as the population P_λ, so the v_f/K term of Cor. 1.1 is absent: the only randomness is calibration-side ($\hat{Q}$), which is precisely the quantity the covered-mean analyses and the split-sample re-analysis of App. C.1 interrogate. (ii) *Zero normalizers can still occur, through the calibration side.* Reusing the proposal sample does not exclude $\hat{Q}(O_K) = 0$ (the calibration sample misses every observed proposal category): the unclipped estimator is then undefined ($0/0$), to be handled by an explicit fallback policy. Under the deployed double clip the behavior is pinned down exactly: every observed category carries weight α, so $\hat{\mu}_{\text{clip}} = \sum_y \hat{P}(y) f_y$, the naive proposal mean; high-deficiency cells where calibration mass concentrates off-support inherit proposal-mean behavior. (iii) *Fresh-budget deployments are a different experiment.* New proposal draws from the underlying P_λ reintroduce the full two-sample behavior of Cor. 1.1; the synthetic grid (independent two-sample, population P in the denominator) measures exactly this regime, and a fresh-draw replication on the real streams is the corresponding deployment test, whose stream-lottery findings are the practical prescription's empirical basis.

Exact frozen-count risk. Write T_0 for the unclipped reused-count estimator, i.e. the calibration mean truncated to the observed support (Prop. 1). Conditional on the observed support O and $N_O > 0$ (calibration counts within O), $\mathbb{E}[(T_0 - \mu_Q)^2 \mid N_O > 0] = (m_O - \mu_Q)^2 + s_O^2\,\mathbb{E}[1/N_O \mid N_O > 0]$ with $N_O \sim \text{Bin}(n_h,\, Q(O))$, m_O and s_O^2 the Q-conditional mean and variance of f on O; with the no-covered-calibration event $\{N_O = 0\}$ carrying probability $p_0 = (1 - Q(O))^{n_h}$ and fallback value a_O, the *unconditional* risk is

$$\mathbb{E}[(T_0 - \mu_Q)^2] \;=\; (1 - p_0)\left[(m_O - \mu_Q)^2 + s_O^2\,\mathbb{E}\big[\tfrac{1}{N_O} \,\big|\, N_O > 0\big]\right] \;+\; p_0\,(a_O - \mu_Q)^2; \tag{12}$$

the conditional risk is weighted by $1 - p_0$, not added to the fallback term unweighted, and no v_f/K term appears, as expected for a frozen proposal. The audit verifies this against the measured conditional MSE on all 816 battery cells (both item banks and the matched-question battery, $n_h \in \{50, 400\}$): median measured/predicted ratio 1.000 (matched-question range $[0.97, 1.02]$, 288/288 within $[0.5, 2]$), and the zero-normalizer event is *realized* in 38/528 covered cells of the two item banks at n_h=50 (recorded as NaN; the double clip then returns the naive proposal mean).

A.3 Proofs of Theorems 2–4 (T2a/b/c)

Throughout, P_λ is the LLM proposal and Q the human target (matching §3/§4). The theorems concern the *structural* deficient region $S_{\text{zero}} = \{y : P_\lambda(y) = 0\}$ with stranded mass $\varepsilon = Q(S_{\text{zero}})$,

covered conditional mean $m_C = \mathbb{E}[f \mid S^c_{\text{zero}}]$, covered conditional variance $\sigma^2_C = \text{Var}_Q(f \mid S^c_{\text{zero}})$, and $\Delta_f = |\mathbb{E}[f \mid S_{\text{zero}}] - m_C|$; the resolution-$\tau$ object S_τ and the finite-budget object $\mathcal{U}_K$ are treated separately below. Throughout this subsection assume the non-degenerate case $0 < \varepsilon < 1$ (so $0 < Q(S^c_{\text{zero}}) < 1$); degenerate cases are handled explicitly where they arise. Proposal draws $y_k \sim P_\lambda$ never land in S_{zero}.

Part 1 (proof of T2a: exact limit for SNIS under the calibration protocol). The BiasFate protocol draws the calibration set from the *target* Q (the operationally correct direction: the human sample estimates the population, not the sampler) and forms plug-in weights $\hat{w} = \hat{Q}/P_\lambda$, applied to proposal samples $y_k \sim P_\lambda$. At a *fixed* calibration set with $\hat{Q}(S^c_{\text{zero}}) > 0$ (else the limit below is undefined), as $K \to \infty$ the unclipped SNIS estimator converges a.s. to the covered human sample mean

$$\bar{f}_{h,C} \;=\; \frac{\sum_{y \in S^c_{\text{zero}}} \hat{Q}(y) f(y)}{\hat{Q}(S^c_{\text{zero}})}, \tag{13}$$

since only categories with $P_\lambda(y) > 0$ carry the limit; if additionally $n_h \to \infty$, $\hat{w}(y) \to w^\dagger(y) = Q(y)/P_\lambda(y)$ pointwise on $\text{supp}(P_\lambda) = S^c_{\text{zero}}$ and

$$\frac{\sum_k w^\dagger(y_k) f(y_k)}{\sum_k w^\dagger(y_k)} \xrightarrow{a.s.} \frac{\sum_{y \in S^c_{\text{zero}}} Q(y) f(y)}{\sum_{y \in S^c_{\text{zero}}} Q(y)} = m_C = \mathbb{E}[f \mid S^c_{\text{zero}}]\,, \tag{14}$$

using $\mathbb{E}_{P_\lambda}[\frac{Q}{P_\lambda} g] = \mathbb{E}[g]$ for any g supported on $\text{supp}(P_\lambda)$ (with $g = f$ and $g = \mathbf{1}$ respectively). Writing $\mu_f = (1-\varepsilon) m_C + \varepsilon\, \mathbb{E}[f \mid S_{\text{zero}}]$, the limiting bias is exactly $\varepsilon\, \Delta_f$. This proves T2a.

Calibration variance under support deficiency (budget display of §4). Under support deficiency the limit re-creates not the full human sample mean but the *covered* human sample mean: only the $N_C \sim \text{Binomial}(n_h, 1-\varepsilon)$ calibration draws landing in the covered region S^c_{zero} contribute (draws landing in S_{zero} are discarded by the limit). Conditionally on $N_C \geq 1$, the *exact* floor is $\sigma^2_C\, \mathbb{E}\big[1/N_C \mid N_C \geq 1\big]$ (the variance of a mean of N_C covered draws is σ^2_C/N_C), *not* σ^2_f/n_h. Jensen must be applied *conditionally on the same event*: with $p = 1-\varepsilon$,

$$\mathbb{E}\Big[\tfrac{1}{N_C} \,\Big|\, N_C \geq 1\Big] \;\geq\; \frac{1}{\mathbb{E}[N_C \mid N_C \geq 1]} \;=\; \frac{1-(1-p)^{n_h}}{n_h\, p} \;\xrightarrow[n_h p \to \infty]{}\; \frac{1}{n_h\, p}, \tag{15}$$

and $\mathbb{E}[1/N_C \mid N_C \geq 1] \leq 1$ always. The naive chain $\mathbb{E}[1/N_C \mid N_C \geq 1] \geq 1/\mathbb{E}[N_C]$ is invalid (Jensen over the unconditional law of N_C, which includes the excluded atom $N_C = 0$), and the direction of the asymptotic form $\sigma^2_C/[n_h(1-\varepsilon)]$ is *not uniform*: for $n_h(1-\varepsilon) \gg 1$ the exact conditional floor slightly *exceeds* $\sigma^2_C/[n_h(1-\varepsilon)]$ (positive Jensen gap), but for $n_h(1-\varepsilon) \lesssim 1$ it can overstate the conditional floor (at $\varepsilon = 0.994$, n_h=50: $1/[n_h(1-\varepsilon)] \approx 3.3$ vs the exact $\mathbb{E}[1/N_C \mid N_C \geq 1] \approx 0.928$), and there the unconditional error is dominated by the $N_C = 0$ event. High-deficiency cells therefore report the N_C distribution, the fallback frequency $\Pr(N_C = 0)$, and the exact conditional floor, not the asymptotic form alone. The sampling term analogously centers on m_C with the truncated weights $w_C(y) = Q(y)/\big((1-\varepsilon)P_\lambda(y)\big)$, $y \in S^c_{\text{zero}}$, i.e. proposal-side variance v_C/K with $v_C := \text{Var}_{P_\lambda}[w_C(f - m_C)]$ (the support-deficient analogue of v_f; equal to it under full support). The uncovered-calibration event $\{N_C = 0\}$ has probability ε^{n_h} ($0.994^{400} \approx 9\%$, $0.994^{50} \approx 74\%$), on which the unclipped estimator is undefined; the deployed lower clip avoids the zero denominator by substituting its own floor, changing the limit to $\mu_\alpha(\hat{Q})$ of A.4 rather than restoring $\bar{f}_h$. Cells with $\hat{\varepsilon} = 1$ have no covered conditional mean and are reported as undefined-limit cells, not as $\varepsilon\Delta_f$ confirmations.

Part 2 (proof of T2b: distributional lower bound, class-universal). Let R be *any* probability distribution supported on $\text{supp}(P_\lambda) = S^c_{\text{zero}}$. Then $R(S_{\text{zero}}) = 0$ while $Q(S_{\text{zero}}) = \varepsilon$, so by the definitional bound $\text{TV}(R, Q) \geq |Q(S_{\text{zero}}) - R(S_{\text{zero}})| = \varepsilon$: total variation between any proposal-supported distribution and the target is at least the structurally stranded mass. Since every estimator that reweights proposal draws outputs a statistic of a distribution supported on $\text{supp}(P_\lambda)$, whatever the weights, no proposal budget and no weighting scheme recovers the target *distribution* on S_{zero}. The finite-budget companion is immediate: the empirical reweighted distribution R_K is supported on the observed categories $\mathcal{U}^c_K$, hence $\text{TV}(R_K, Q) \geq Q(\mathcal{U}_K)$, a bound that depends on the realized sampling and shrinks only as the budget grows, not a permanent plateau. This bound is distribution-level (it constrains full-distribution recovery, e.g. quantiles, tail probabilities, full-response histograms), not a per-instance mean-bias floor: for mean estimands the situation is finer and is treated in Part 3. □

Part 3 (proof of T2c: partial identification under unknown stranded outcomes). Suppose the outcome values f on S_{zero} are unobservable (known only to lie in $[a, b]$), while the covered-region quantities m_C and ε are identified from calibration data. Then every f consistent with the observations satisfies $(1-\varepsilon)m_C + \varepsilon a \le \mu_f \le (1-\varepsilon)m_C + \varepsilon b$, so μ_f is partially identified in an interval of width $\varepsilon(b-a)$. For the two endpoints $f^{\text{lo}}, f^{\text{hi}}$ (differing only on S_{zero}: values a vs b) the observed data have one and the same law, so any estimator outputs a single value m under both, and

$$\max\left(|m - \mu_{f^{\text{lo}}}|,\ |m - \mu_{f^{\text{hi}}}|\right) \ \ge\ \tfrac{\varepsilon(b-a)}{2}, \tag{16}$$

the two-outcome (equal-prior) minimax bias, tight up to the factor 2 against the interval width. This is a worst-case statement over outcomes; it makes no restriction on how proposal-side information enters.

Boundary of the claim (essential; a counterexample settles it). T2c is a worst-case statement over outcomes; it does *not* give a per-instance mean-bias floor when f is a known scoring map (e.g. Likert scores, where $f(y)$ is read off the response), which is why the deployed $\widehat{\Delta_f}$ uses the known map and the plateau is reported as an empirical regularity of the tested estimator, not a universal lower bound. Counterexample: $\mathcal{Y} = \{a, b, c\}$, $P_\lambda = (\frac{1}{2}, \frac{1}{2}, 0)$, $Q = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, $f = (0, 1, 0.8)$. Then $\mu_f = 0.65$ and $\varepsilon\Delta_f = 0.15$, yet the covered-support reweighting with weights $w(a) = 0.7$, $w(b) = 1.3$ has SNIS limit $\frac{0.7\cdot\frac{1}{2}\cdot 0 + 1.3\cdot\frac{1}{2}\cdot 1}{0.7\cdot\frac{1}{2} + 1.3\cdot\frac{1}{2}} = 0.65 = \mu_f$ exactly: bias zero, despite $\varepsilon\Delta_f = 0.15 > 0$. The mechanism is general: for fixed known f, arbitrary non-negative weights on the covered support place the output anywhere in the convex hull $[\min_{\text{supp}(P_\lambda)} f, \max_{\text{supp}(P_\lambda)} f]$, so whenever μ_f lies in that interval a zero-bias reweighting exists. What such weights buy is *not* free: they are outcome-tuned (they depend on f), so they certify nothing about other estimands, and they discard the natural weighting $\hat{Q}/P$ that the calibration protocol estimates, using them is a different, estimand-specific estimation problem, not a correction of the sampler. Estimators that read the paired calibration outcomes $f(z_j)$ on S_{zero} (imputation/DR hybrids) stand outside the proposal-reweighting family and price in $\mathbb{E}[f \mid S_{\text{zero}}]$ at an n_h-governed rate (App. C.2: DR bias 0.004 at n_h=400 vs 1.51 at n_h=50). □

Structural zeros vs. finite-budget non-observation (U_K). A structural zero ($P_\lambda(y) = 0$) is budget-immune. A merely unobserved category with $P_\lambda(y) > 0$ will eventually be sampled: with N_y its count in K draws, $\Pr(N_y = 0) = (1 - P_\lambda(y))^K$, so a category with $P_\lambda(y) = 1/K$ is missed with probability $\approx e^{-1} \approx 0.368$, zero counts alone never certify $P_\lambda(y) < 1/K$. The deployable finite-budget object is the expected uncovered target mass

$$U_K \ =\ \sum_y Q(y)\left(1 - P_\lambda(y)\right)^K \ =\ \mathbb{E}\big[Q(\{y : N_y = 0\})\big], \tag{17}$$

computable before deployment from $\hat{Q}$ and the proposal's empirical distribution, and monotone decreasing in K *when the true P_λ is known*. A deployability caveat is essential: the plug-in estimate from a pilot is unreliable, a category absent from the pilot has $\hat{P}(y) = 0$, so plug-in predicts mass $\hat{Q}(y)$ uncovered at *every* future K, re-creating the structural-zero reading exactly where the risk is merely finite-budget. Honest deployment therefore reports U_K under explicit smoothing assumptions (or conservative intervals over unobserved categories), keeps the pilot budget and the future budget separate, and re-measures with fresh proposal draws as K grows; the fresh-stream study of App. C.5 measures exactly this failure mode and finds single-pilot predictions err in both directions.

Extension to τ-near-deficiency. With $P_\lambda(y) \le \tau$ on a set of Q-mass ε_τ, proposal draws land in that set $K\,P_\lambda(S_\tau)$ times in expectation, at most $K\,|S_\tau|\,\tau$ if only the per-category bound $P_\lambda(y) \le \tau$ is known, so the mass SNIS can absorb from S_τ vanishes with τ. We make *no continuity claim in τ*: what is exact are the finite-budget facts $\Pr(N_y = 0) = (1 - P_\lambda(y))^K$, $\mathbb{E}[N(S_\tau)] = K\,P_\lambda(S_\tau)$, and U_K above; the residual τ-driven fluctuation is governed by the few realized S_τ draws rather than by K or n_h, matching the resolution-$\tau \approx 1/K$ practice in the measured matrices and their graceful degradation under larger τ (App. C.2). □

A.4 CLIPPING LIMITS AND THE MEASURED TRADE-OFF

For fixed calibration frequencies $\hat{Q}$ and proposal P, write $\tilde{w}_\alpha(y) = \mathrm{clip}(\hat{Q}(y)/P(y), \alpha, 1/\alpha)$ on $\{P > 0\}$. The large-K limit is

$$\mu_\alpha(\hat{Q}) = \frac{\sum_{y:P(y)>0} P(y)\tilde{w}_\alpha(y)f(y)}{\sum_{y:P(y)>0} P(y)\tilde{w}_\alpha(y)}.$$

Both the numerator and denominator change with α. The limit generally differs from the direct calibration mean even with full support. With reused empirical counts, replacing P by $\hat{P}$ gives the exact clipped category masses in §3. At $\alpha = 1$, all observed weights equal one and the estimate is naive; as $\alpha \downarrow 0$, clipping disappears on the fixed finite support. For bounded f, SNIS remains within the score range whenever its denominator is positive.

Appendix A.2 gives a valid numerator second-moment bound for *oracle upper clipping*. It does not control the normalized ratio or the deployed lower clip. A heuristic expression combining a term proportional to α and one proportional to α^{-1} can therefore organize a grid but cannot establish an optimal deployment clip. In the original surrogate calculation, its candidate $\alpha^\star$ decreases from 0.51 to 0.26 as n_h increases from 50 to 800. Those candidate values are not a validated tuning rule: on the heavy-tail cells, the measured optimum stays at the weakest tested clip, $\alpha = 0.02$, through $n_h = 3200$.

The measured grids distinguish settings. Near-matching covered cells have weights close to one, with a reported optimum at $w_{\max} = 1$ and bias 0.004. On the tested high-missing-mass cells, the evaluated clipping levels leave error near the truncation plateau. This empirical pattern does not imply that every possible clip leaves every mean unchanged: clipping alters covered-region mass, and outcome-specific cancellation is possible (App. A.3). In the separate covered-cell ablation, removing clipping reduces RMSE from 0.1027 to 0.0245 (App. C.6). Thus the main deployed comparison fixes $[0.02, 50]$, while clip-grid optima and unclipped ablations are reported as sensitivity analyses, not retrospective replacements for that protocol.

A.5 DELTA-METHOD VARIANCE OF THE PLATEAU ESTIMATOR

Let $S_0 = \{y : P_\lambda(y) < 1/K\}$ (the S_τ of §4 at $\tau = 1/K$) be fixed by the proposal (the boundary cells are determined by the frozen counts; fluctuation of an estimated $\hat{P}_\lambda$ around them is not modeled here) and let $\varepsilon(Q) = Q(S_0)$, $\Delta_f(Q) = |m_0(Q) - m_1(Q)|$ with $m_0(Q) = \sum_{S_0} Qf \,/\, \varepsilon(Q)$ and $m_1(Q) = \sum_{S_0^c} Qf \,/\, (1 - \varepsilon(Q))$. The plateau estimate is $g(\hat{Q}) = \varepsilon(\hat{Q})\,\Delta_f(\hat{Q})$ with $\hat{Q} \sim \mathrm{Mult}(n_h, Q)/n_h$, so $\mathrm{Cov}(\hat{Q}) = (\mathrm{diag}(Q) - QQ^\top)/n_h$ and

$$\mathrm{SE}\big(g(\hat{Q})\big) \;=\; \sqrt{\nabla g^\top \tfrac{\mathrm{diag}(Q) - QQ^\top}{n_h} \nabla g}, \tag{18}$$

with ∇g in closed form or by central differences. Three caveats bound its use. (i) g is non-differentiable where $m_0 = m_1$ (the $|\cdot|$ kink): near the kink the delta method *underestimates* the bootstrap SE (App. C.2: ratio IQR $[0.14, 1.02]$, median 0.986). (ii) The interval is for the *asymptotic plateau*, not for the finite-n_h measured bias, whose excess error is governed by the T1 variance term and weight clipping, hence the modest empirical coverage (55.6%) of measured SNIS biases at n_h=400 under the nominal 95% level. (iii) The construction is specific to the reweighting class: imputation/DR estimators introduce outcome-model variance terms not captured by ∇g. Practically: use the delta SE to report the precision of $\hat{\varepsilon}\widehat{\Delta_f}$ itself, and reserve interval construction for measured performance to the split-sample validation of App. C.1 (conditional on the empirical distributions), not to unvalidated bootstrap intervals. □

B EXPERIMENTAL DETAILS

- **B.1 Prompts (verbatim; English translations ours, Chinese originals in parentheses).** *Teacher intro (gemma-2-9b-it, system):* "You are an ordinary adult whose personality and lifestyle are typical of the general population. You are taking a social survey; please answer each question as truthfully as you can." (original: 你是一位普通的成年人，性格和生

Object	Data	Role in the chain
Proposal $P_\lambda(\cdot \mid x)$	K LLM draws per x under config λ	cheap samples; sampler bias enters here
Target $Q(\cdot \mid x)$	human conditional (survey marginal / pooled stratum)	the estimand's truth; $n_h \ll K$ affordable
Calibration set	n_h human draws $z_i \sim Q$ per stratum	weights $\hat{w}$; diagnostics $\hat{\varepsilon}$, $\widehat{\Delta_f}$, K_{eff}
Estimands	E-dist / E-mean / E-policy	TV; μ_f; $V(\pi)$ under true rewards
Estimators	naive $\bar{f}_P$; deployed clipped SNIS; direct mean $\bar{f}_h$	$\bar{f}_P$ carries b_f; SNIS plateaus at $\hat{\varepsilon}\widehat{\Delta_f}$; $\bar{f}_h$ pays $\sigma_f/\sqrt{n_h}$
Verdict	T1 (correctable) vs T2 (plateau)	budget-relative; re-measured, never certified

Table 3: **The BiasFate data flow.** Objects → data → estimator chain shared by every experiment; per-experiment target provenance is recorded in the supplementary manifest (B.7).

活习惯属于人群中的典型水平。你正在参加一项社会调查，请根据你的真实情况如实作答。) *Persona doses (Qwen2.5-7B-Instruct, system):* p_0 = "You are taking a social survey; please answer as truthfully as you can." (original: 你正在参加一项社会调查，请根据你的真实情况如实作答。); p_1 = "You are very extroverted and love socializing and adventure; you are the most active person in your circle of friends." + the p_0 suffix (original: 你性格非常外向，热爱社交和冒险，是朋友圈里最活跃的人。); p_2 = "You are extremely extroverted, extremely optimistic, and love lively crowds and adventure; on almost all social and activity-preference questions you tend to agree." + the p_0 suffix (original: 你是一个极度外向、极度乐观、极度喜欢热闹和冒险的人，几乎在所有涉及社交与活动偏好的问题上都倾向选择同意。) *Question bank:* 12 Likert items (8 persona-sensitive, 4 control), options A–E = 非常不同意/不同意/一般/同意/非常同意 ("strongly disagree / disagree / neutral / agree / strongly agree") mapped to scores 1–5; full item text with measured $\hat{\varepsilon}, \widehat{\Delta_f}$ ships with the supplementary code. *Language coverage:* the primary bank is Chinese (the population the study targets); the English replication (76,800 generations) uses a translation of the full Likert matrix and rubric family, translated by the authors and shipped verbatim with the supplementary code, so every Chinese prompt consumed in the paper has an English counterpart that was actually run, but no other languages are tested, a disclosed scope limit (§7).

- **B.2 Decoding and parsing:** temperature 1.0, top-p 1.0; the response format constrains the answer to a single letter A–E, extracted by regex `^[A-E]` (with a fallback search); parse failures are logged per run and excluded; rates are $\leq 0.4\%$ for the teacher–student runs and $\leq 0.2\%$ for the battery's non-reference configurations, but the Gemma reference sampler fails 9.6% of generations (8676/9600 parsed, its unparsed outputs repeat the persona statement instead of the letter), so all cross-model analyses normalize by parsed-count totals rather than by K (a convention that changes $\hat{\varepsilon}$ materially when parse failure is non-trivial; treating parse failure as an explicit response bin changes *0/11* verdicts and leaves the headline correlation stable in $[-0.97, -0.94]$, App. C.1).
- **B.3 Cluster configuration and compute.** 1×A100 80G, sglang 0.5.10.post1 offline engine, triton attention backend, cuda-graph disabled, mem-fraction-static 0.75. All GPU experiments ran on this single-node configuration. For a single accounting: the decisive validation runs produced 43,200 and 76,800 generations; the English replication, the second item bank, and the matched-question GOQA study 76,800 each; the held-out expansion 230,400; the remaining blocks are smaller. Total program volume is ≈640K generations, each a single-constrained-letter response, with parse-failure rates logged per block.
- **B.4 Replicate protocol, seeds, and multiplicity.** 200 multinomial calibration draws per (question, n_h) in the robustness studies (500 deployment draws per cell in the matched-question benchmark; replicates as tabulated per study elsewhere). Every script runs with a fixed, recorded seed, and seeds and per-experiment provenance are machine-readable in the supplementary manifest, so every reported number is exactly reproducible from the supplementary artifact. The battery's claims are pre-specified gates, not post-hoc selected tests, so no global correction is applied across experiments; within experiments, uncertainty is reported via question-clustered bootstrap CIs and Wilson intervals on win rates, and the only formal Bonferroni correction in the paper is inside the false-positive-controlled guard construction (App. C.7).
- **B.5 Pre-registration and internal audit.** The robustness battery was developed through an internal red-team process: before each batch was run, the draft was challenged by internal

reviewers playing the role of a hostile referee, the identified weaknesses were turned into pre-specified gates, and a result refuting any claim would have been reported as such. The audit reports that non-reweighting estimators can reduce bias below the reweighting plateau by using calibration outcomes directly (App. C.2), and that at matched human budget the direct human mean beats reweighting at every tested $K \leq 6400$ (App. C.6). The pre-specification is self-administered and not externally timestamped, a disclosed limit; the full audit log is included with the supplementary code instead of being summarized in the paper.

- **B.6 Data-flow disclosure.** In the battery pipeline the multinomial calibration draws enter four quantities: the SNIS weights, the $\hat{\varepsilon}$ diagnostic, the measured SNIS residual, and the naive reference against which recovery is normalized, so the measured residual and its normalizer share calibration randomness. A split-sample re-analysis removes this coupling and finds the headline essentially unchanged (App. C.1). *Notation map for the code artifact:* the supplementary code swaps symbols relative to the paper; code `P` (gemma-neutral counts) is the paper's *target* Q, code `Q` (student counts) is the paper's *proposal* P_λ, and the calibration draws sample the target (the deployable direction: the human sample estimates the population). Weights $\widehat{w} = \widehat{Q}/\mathrm{clip}(\widehat{P}_\lambda)$ follow the paper's convention.
- **B.7 Target provenance.** Each experiment's target is one of four types: *synthetic* (known Q from a simulator), *llm-empirical* (the reference sampler's frozen counts serve as Q), *stylized* (author-constructed, labeled as such), or *human-survey-empirical* (released survey marginals); the type, source, and content hashes are recorded per experiment in the supplementary manifest. "Resampled" means the experiment draws calibration samples from the tabulated target rather than using raw respondent-level data, which the provided tables do not include. The stylized alternative target of App. C.2 is author-constructed, and the real-human-target role is carried only by the GlobalOpinionQA studies.

C Robustness and additional studies

Empirical labels in this appendix use R (recovery-screen) and P (plateau), as defined in §3.3. Archived plot annotations T1/T2 denote these empirical labels, not the theoretical support assumptions. The following studies distinguish label agreement from estimator and decision performance.

This appendix organizes the supplementary studies by the question each answers: statistical robustness of the headline claims (C.1), robustness to estimators, resolution, and targets (C.2), transfer to new languages, item banks, and larger models (C.3), judge noise (C.4), fresh streams and held-out questions (C.5), matched-budget estimation and action selection (C.6), stress audits of the error display (C.7), and estimand re-keying with rare-category shares (C.8). All gates were pre-specified against weaknesses identified by the internal red-team process of App. B (self-administered, a disclosed limit). Unfavorable results are kept throughout.

C.1 Statistical robustness of the headline claims

Uncertainty. A question-level bootstrap (B=2000) attaches CIs to every pair-level statistic: the headline rank correlation survives at $\rho = -0.98$, 95% CI $[-0.99, -0.81]$, permutation $p = 10^{-4}$ (11 pairs), and the plateau-prediction correlation re-estimates at $\rho = 0.839$ $[0.645, 0.928]$ under question-clustered resampling (same 30 cells as the $\rho = 0.844$ of §5.3; the midrank convention differs). The headline association also survives re-derivation under cluster-inference methods that do not rely on the bootstrap: a $t(11)$ interval on the Fisher-z scale gives $[-1.00, -0.87]$, and the worst single-question deletion gives $\rho = -0.97$.

Dose–response. Within the tested prompt interventions, increasing persona dose is associated with more uncovered mass and larger bias: Spearman(dose rank, $\hat{\varepsilon}$) $= 0.44$–0.46 and (dose rank, bias) $= 0.44$–0.62 (all $p \leq 2 \times 10^{-4}$).

Premise and display audits for the T1 oracle bound. The negative-association premise $\mathrm{Cov}_Q(w^\star, (f - \mu_f)^2) \leq 0$ of Thm. 1/A.1 is audited cell-by-cell: 0/36 covered cells of the first bank violate it (median standardized covariance -0.86; the 11 self-pair cells sit at zero identically since $w^\star = 1$ there), with 3/59 violations on the second bank (App. C.3), so the premise is an empirical regularity, not an identity. The realized range ratio $\mathbb{E}_Q[w^\star(f - \mu_f)^2]/(R^2\rho_1)$ stays ≤ 0.19

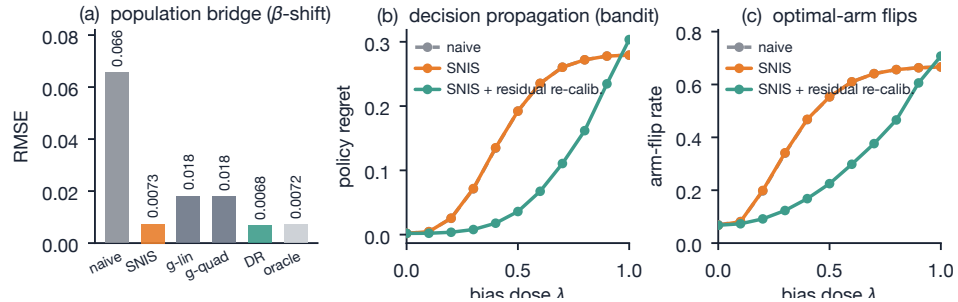


Figure 3: **Bias propagates into estimates and decisions.** (a) Population-bridge RMSE by estimator under a β-shift with a misspecified linear outcome model: SNIS closes 90% of the naive gap; the doubly robust (DR) estimator removes the misspecification residual and matches the oracle. (b) Contextual bandit: reweighting alone cannot fix outcome bias (naive = SNIS regret at every dose λ), while SNIS + residual recalibration recovers 71% of the regret at λ=0.6. (c) Optimal-arm flip rate vs. dose: recalibration cuts the flip rate from 0.61 to 0.30 at λ=0.6.

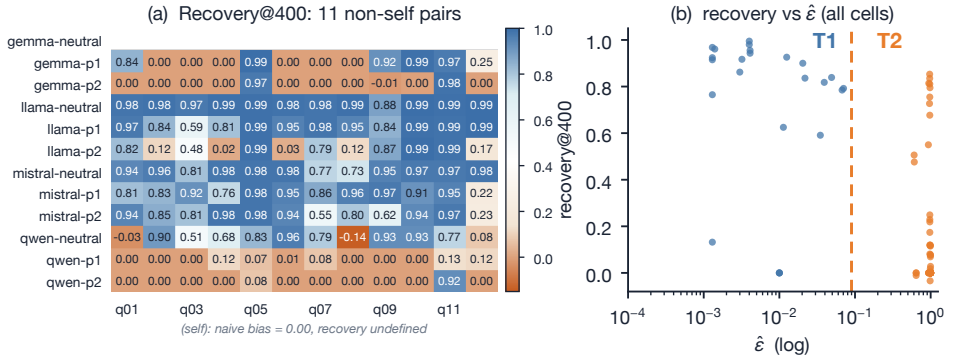


Figure 4: **Recovery under the frozen observed support.** (a) Heatmap for the 11 non-self generator–scaffold pairs and 12 questions; recovery is undefined for the self pair with zero naive bias. Negative recovery means reweighting increases absolute bias. (b) The scatter shows the portion within its displayed recovery range; negative-recovery cells remain represented in the heatmap. Its dashed divider is descriptive, not the relative-plateau rule used in Table 2. Observed empty categories at temperature 1.0 are finite-budget non-observation, not certified structural zeros.

on every covered cell, and the operational display is scale-calibrated (median predicted/measured 0.956, a display calibration, not of the Bernstein constants). A concrete instantiation with unit constants: at $\delta = 0.05$, $\log(4/\delta) \approx 4.38$ and the first-order constant is $\sqrt{2 \times 4.38} \approx 3.0$; at $K_{\text{eff}} = 185$ (the sharpened-tail adversarial cell of App. C.2) the first-order oracle scale is $\approx 0.22\,\sigma_f$, the order of magnitude behind the calibrated c below. The synthetic 2D (n_h, K) grid with known truth validates the additive structure (App. C.5).

Verdict-constant calibration and stability. The pre-registered rule "R iff plateau $< 0.3\times$ naive" is calibrated empirically on the 128 usable cells: misclassification at $c = 0.3$ is 6.2% (Chinese), the data-driven optimum is $c^* \approx 0.52$ (error 2.3%), and *all* 8 errors at c=0.3 are false-P, i.e. no false recovery-screen label is observed on this Chinese bank. On the 126 usable English cells c=0.3 sits inside the optimal band $[0.24, 0.33]$ (error 2.4%; 2 false recovery-screen labels in 126 cells). The one-sidedness is language-relative, disclosed here: pooled EN errors split 2 false-R / 1 false-P at the recovery ≥ 0.5 label. The threshold is stable to removing any single question *or* pair jointly (held-out error 3.1%, fold-level $c^* \in [0.48, 0.57]$) and to the usable-cell cutoff (false-R = 0 at every tested cutoff), but it is *deployment-relative*, not universal: c^* shifts across languages (0.52 CN vs 0.29 EN), and the protocol re-fits c per deployment. The verdict also carries information beyond the continuous $\hat{\varepsilon}$ alone: in leave-one-out deployment the verdict rule misclassifies 3.1% (CN) / 3.2%

Check	Reported result and interpretation
Controlled covered support	RMSE / ESS scale 0.98–1.06 without shift; at most 1.85 under shift
Frozen-support plateau	Median measured/predicted ratio 1.00; Spearman $\rho = 0.844$; 93% within factor 2
Chinese threshold study	Default $c = 0.3$: error 6.2%; fitted $c^* \approx 0.52$: 2.3%
English threshold study	Fitted $c^* \approx 0.29$; optimal band $[0.24, 0.33]$; error 2.4%

Table 4: **Scale, plateau, and threshold validation.** Thresholds are fitted within each item bank using the labels and development procedures described in this subsection. The ESS variance interpretation uses the estimand–weight covariance condition stated after Theorem 1.

(EN) versus 9.4% / 6.3% for a same-budget $\hat{\varepsilon}$-threshold rule tuned on the same data, because the verdict adds the naive-bias scale that $\hat{\varepsilon}$ lacks (direction-aligned AUC 0.95–0.98).

Data-flow robustness. In the battery pipeline the same calibration draws feed both the weights and the naive reference against which recovery is normalized. A split-sample re-analysis (weights from one half, reference from the disjoint other half) leaves the headline essentially unchanged: pair-level Spearman -0.915 pooled versus -0.903 split (medians over 20 replicates), and pooled-vs-split verdict agreement 98.2% at c=0.3 (97.3% at the data-driven c), with residual disagreements confined to boundary cells. Varying only the error reference with the weight sample fixed moves neither the association nor the plateau effect. Treating parse failure as an explicit response category (imputed score swept over $\{1, 3, 5\}$) changes 0/11 verdicts and leaves the headline correlation in $[-0.97, -0.94]$.

C.2 Estimator, resolution, and target robustness

Estimator families. On the five high-$\hat{\varepsilon}$ pairs, every *reweighting* family (frequency plug-in, logistic density-ratio, kernel mean matching, clipped at every tested level) lands at median |bias| 1.67–1.99 versus plateau 1.37: none escapes, and RMSE stays above the plateau at every clip level, minimized at the weakest tested clip. Clipping is a covered-regime tool with no plateau leverage. Non-reweighting families (g-comp, doubly robust) fall outside the reweighting class and can push bias down only by consuming calibration data: DR bias 0.004 at n_h=400 rises to 1.51 at n_h=50 ($\approx$ naive; RMSE governed by n_h, never K). The dichotomy is a property of the information structure, not of the specific estimator (Fig. 5).

Resolution. The verdict is stable under the unseen-mass threshold τ: agreement 100% at τ=$1/K$, 91% at $2/K$, 81% at $5/K$, 78% at $10/K$. The anatomy of the flips: of 144 (pair, question) cells, 32 flip between τ=$1/K$ and $10/K$, and 81% of flipped cells have $\hat{\varepsilon} < 10/K$ at the pre-registered resolution (stable-cell median $\hat{\varepsilon} = 0.31$), i.e. flips are exactly the small-$\hat{\varepsilon}$ cells whose plateau falls below the τ noise floor, the boundary behavior predicted by the τ-extension in App. A.3, not erratic instability.

Estimands and answer keys. Score, rank-ECDF, and pass-fail estimands on the same rubric data give recovery 0.975–0.980 when support is covered. Recomputing the MCQ indicator estimand under four answer-key rules leaves the R verdict stable for the three top-based keys (aggregate recovery 0.76–0.80), while the runner-up key drives naive bias to 0.009, where sampling error dominates and the verdict flips; the operational lesson is to report (plateau, naive) jointly rather than the binary verdict alone. Coverage is scaffold-robust for fine-grained score estimands: under the most biased scaffold, TV $= 0.54$ yet $\hat{\varepsilon} = 0$ and recovery 0.982.

Alternative target. Re-anchoring the calibration target to a stylized, author-constructed concentration profile (labeled as such; the real-human role is carried only by the GlobalOpinionQA studies) and re-running the boundary map on all 12 pairs: 8/12 verdicts agree; the disagreements are high-$\hat{\varepsilon}$ pairs that flip P $\rightarrow$ R because the original target's own concentration had inflated $\hat{\varepsilon}$, plus one degenerate row; recovery on covered pairs stays 0.66–0.98. The map is a property of (proposal, target) pairs, and $\hat{\varepsilon}$ measures it correctly under both targets.

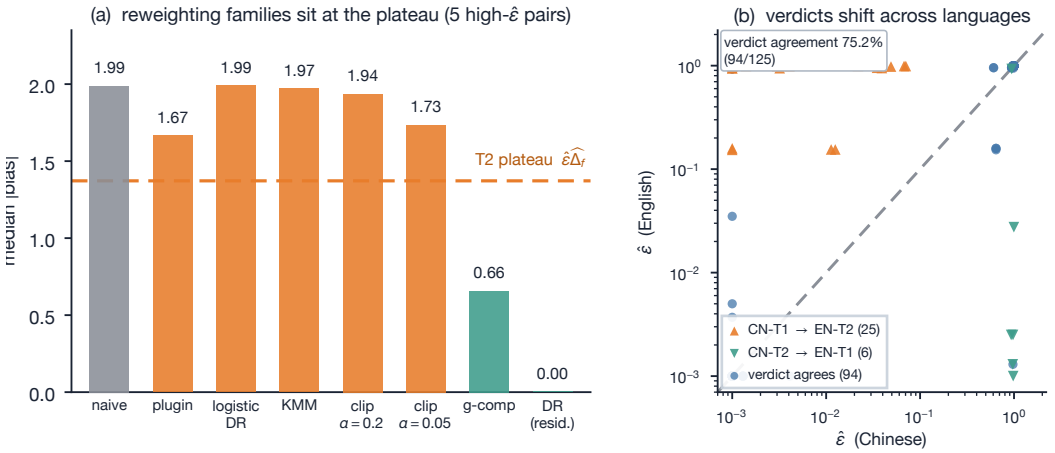


Figure 5: **At matched K, reweighting families sit at the plateau, and support deficiency is language-dependent.** (a) Median bias of eight displayed estimator families on the five high-$\hat{\varepsilon}$ pairs: all reweighting families (including every clipping level) sit at or above the plateau (dashed), an empirical regularity of the tested estimators (the boundary of App. A.3); non-reweighting families (g-comp, DR) reduce bias only by consuming human calibration data n_h, never by consuming simulation draws K. (b) Re-running the full matrix with the bank translated to English: recovery splits along $\hat{\varepsilon}$ exactly as in Chinese (band medians 0.98 at $\hat{\varepsilon} < 0.1$ vs 0.00 at $\hat{\varepsilon} \geq 0.5$), but the level of concentration is language-dependent (median $\hat{\varepsilon}$ 0.022 $\rightarrow$ 0.955) and only 75.2% of verdicts agree; switching language without re-measuring silently moves cells across the trust boundary.

Quantitative error display. The predicted RMSE $\sqrt{\text{plateau}^2 + \sigma_f^2/K_{\text{eff}}}$ matches measured RMSE on the five high-$\hat{\varepsilon}$ pairs with median predicted/measured ratio 0.956, 5/5 within the pre-registered gate $[0.3, 5]$; a $\pm$2-bootstrap-SE band on $\widehat{\Delta_f}$ leaves the in-gate proportion at 97% and moves the median ratio 1.00 $\rightarrow$ 0.95.

Adversarial proposals. Two adversarial proposals with zero support deficiency (extremity amplification and cube-sharpened masses) are correctly kept R: these two support-intact constructions have no observed truncation plateau; they degrade the realized effective sample size (K_{eff}: 340 $\rightarrow$ 185 under sharpening), which the proposal-variance term reports. A genuinely high-$\hat{\varepsilon}$ proposal under the stylized target is a boundary cell whose verdict moves with the calibration draw (plateau/naive 0.29 vs 0.40 across c=0.3), disclosed as threshold-sensitive. These two constructions test specific distortions, not adversarial robustness in general; ESS records their increased variance cost.

C.3 Transfer: new language, new item bank, larger model

Cross-language replication. The full matrix is translated to English and re-run on the same four models $\times$ three scaffolds (76,800 generations; parse failures 0.12%). The dichotomy replicates: recovery splits along $\hat{\varepsilon}$ in English exactly as in Chinese (Spearman -0.71 vs -0.98; band medians 0.98 at $\hat{\varepsilon} < 0.1$ vs 0.00 at $\hat{\varepsilon} \geq 0.5$), and $\hat{\varepsilon}$ is rank-correlated across languages (0.64, $p < 10^{-4}$). But the *level* of support deficiency is language-dependent: English responses are far more concentrated, and only 75.2% of verdicts agree (94/125; flips concentrated CN-R$\rightarrow$EN-P). Switching language without re-measuring silently moves cells across the trust boundary toward P, and the map detects it: support deficiency is a property of the sampler–population–language triple and must be re-measured per deployment language (Fig. 5). The rubric family replicates in English with median naive bias 0.559 corrected to a residual of 0.017 (Chinese: 0.013), median recovery 0.969 (Chinese: 0.982), and verdict agreement on 11/12 items; the one flip is a genuine per-question support cliff (the target concentrates all 400 draws on score 1 while the proposal never touches it), correctly flagged as maximal measured deficiency, a live demonstration that $\hat{\varepsilon}$ is a property of the specific (sampler, instrument, language) tuple rather than a family-level constant.

Second item bank. A second 12-item Chinese bank, constructed domain-orthogonally to the first (76,800 generations; parse failures 1.4%), replicates the dichotomy: pair-level Spearman $= -0.94$ (cell-level -0.83, question-cluster CI $[-0.96, -0.48]$); covered pairs recover 0.92–0.98 while plateau pairs ($\hat{\varepsilon} \geq 0.69$) recover ≤ 0.19, matching the predicted plateau. The negative-association premise audit returns 3/59 violations on this bank (versus 0/36 on bank 1), with the unconditional range-form fallback holding with margin; the association is not carried by premise-violating cells and is stable to clipping and support-threshold choices.

Scale-up. Replicated on Yi-34B-Chat ($\sim 4\times$ the largest model above; 12 GOQA questions verbatim, 19,200 generations at $K{=}800$ per question under the neutral scaffold and $K{=}400$ under each persona scaffold): verdict accuracy 0.949, P recall 0.970; 6/36 cells are true-P, a lower rate than at 7–9B, and the larger model leaves less of the option space uncovered (a single-checkpoint observation that cannot separate scale from family, but consistent with coverage improving with scale); and the direct human mean still beats deployed reweighting in 32/36 cells (the four exceptions sit at the calibration floor). Parsing is clean (mean parse failure $<2\%$).

C.4 Judge noise when the estimand is scored by a model

When f is itself an LLM judge, the σ_f entering the calibration floor is judge-noise-inflated (the rubric is scored on an unnormalized 1–5 scale, so σ_f is not confined to $[0, 0.5]$ and values above 1 are legitimate scale units, not probabilities). Measured directly on the rubric family ($K{=}100$ raw responses per item, each scored $J{=}10$ times independently): the judge-noise share of the judge-mean variance is 8% where proposal and judge agree and 18% on the support-cliff item where they disagree; the per-draw σ_f inflates $0.42 \to 0.62$ and $0.54 \to 1.51$ respectively, a $2.8\times$ inflation exactly where the proposal and the judge disagree about the response distribution. On the cliff item the proposal's self-scores sit 1.66 below the neutral judge, so the self-score is itself a biased measurement of the target's f; the judge re-scores the stranded region upward, which changes Δ_f itself rather than its noise. When f is a judge, $\hat{\varepsilon}\widehat{\Delta_f}$ must therefore be estimated with the *same* judge on proposal samples and calibration items.

A second judge (Llama-3.1-8B, the same 200 responses) shows the cliff direction is judge-robust but the magnitude is not (gap -1.66 vs -0.51, a factor 3, and the two judges disagree on response ranking), so $\widehat{\Delta_f}$ and σ_f are judge-*identified* objects, re-measured per deployment. The larger judge-noise share under the second judge (24%/34%) is driven by spread compression rather than a noisier judge: the *absolute* noise level sets the cost of averaging, whereas a spread-compressing judge changes σ_f itself and must be caught by re-measurement. The averaging channel is a measured $1/J$ law, not an extrapolation: subsampling recorded draws to $J' \in \{1, 2, 5\}$ matches the prediction $\text{item} + v/J'$ within 2.1% at every (judge, item, J'). Operationally, $J \approx 60$ per item is the order-of-magnitude guideline for a not-yet-validated judge on covered items; $J = 10$ is adequate once the judge's per-draw noise has been measured on a small pilot. The cliff item has no finite averaging threshold: its open problem is Δ_f identification (judge-dependent scale), not noise. Judge-noise robustness beyond this averageable channel in the plateau regime remains untested.

C.5 Fresh streams and held-out questions

The additive decomposition against known truth. On a synthetic 4×4 (n_h, K) grid with fully known P, Q, f (640 cells, 400 replicates each, independent two-sample design), the realized plug-in MSE matches $\sigma_f^2/n_h + v_f/K$ with ratio median 1.000 (IQR $[0.948, 1.042]$; within $[0.5, 2]$ in 100.0% of cells), and the two terms validate separately (oracle-weight MSE$/(v_f/K)$ median 0.994; direct-mean MSE$/(\sigma_f^2/n_h)$ median 0.998). The variant with the sampling term divided by K_{eff} systematically over-predicts; the data agree with the v_f/K form. The audited premise $v_f \leq \rho_1 \sigma_f^2$ holds in only 67.5% of random cells, confirming that the ESS form is an upper bound under a condition, not an identity. The calibration-floor consequence holds essentially exactly: plug-in SNIS RMSE $\geq$ direct mean RMSE (within 2%) in 639/640 cells.

End-to-end pipeline on the real instrument. Running the complete deployed pipeline (calibration sampling, plug-in weights, double clip, verdict) against known ground truth on real proposal distributions: with a labeled synthetic target, accuracy 0.907 and P recall 0.995 (the dangerous direction

Calibration-budget study	Reported behavior
Budgets	$n_h \in \{50, 150, 400, 1200\}$; 200 replicates per point
Covered, not clip-limited	Measured log–log slope -0.49; median absolute-bias summary 0.004 at $n_h = 400$
Clip-limited cells	Increasing calibration alone leaves clipping distortion
High missing mass	Error remains near the predicted truncation plateau over the tested calibration budgets

Table 5: Calibration-budget findings. These summaries concern the fixed proposal support used in the sweep. More calibration reduces sampling uncertainty but does not restore categories missing from that fixed support.

nearly empty: unsafe clearance 0.5%); with real GlobalOpinionQA human marginals as the target, accuracy 0.935, P recall 0.980, and the plateau-agreement ratio 1.000 on true-P cells.

What a frozen pilot predicts about fresh streams. Fresh proposal streams (K=800, prompts byte-identical, cells selected by a rule frozen before generation) re-measure the pilot's support predictions. On the near-full-coverage cell the pilot is exact: zero error in all 6 fresh streams, verdict stable 6/6. On mid cells the pilot errs downward (realized uncovered mass at K=800 exceeds the prediction; MAE 0.09–0.14). On the highest-uncovered cell the error reverses: fresh streams realize 0.552 against a prediction of 0.990, because independent stream realizations discover early, through generator stochasticity alone, categories the pilot happened to miss (no new-category discovery occurs within the first 200 draws of any 800-draw stream, so the discovery is a stream lottery, not an effect of the doubled budget). Pooling multiple independent pilots or a sequential stopping rule does not systematically reduce the error, and no multi-pilot rule beats the single pilot: the error is stream-lottery dominated. Three missing-mass objects must be distinguished: the target mass the pilot never saw, the expected uncovered mass U_K of a fresh stream, and the realized missing mass of an independent fresh stream; the latter two agree closely, while the first is a different quantity and differs in either direction, so claims must state which object they use. Deployable prediction of fresh-stream missing mass from pilot data alone is a Clopper–Pearson lower bound at best (MAE 0.22–0.31, roughly three times the oracle-side error); the fresh streams are also mildly overdispersed relative to the multinomial benchmark (median variance ratio 1.475). The practical prescription: remeasure as budgets grow, never certify from one pilot.

Held-out expansion. A 48-item GlobalOpinionQA expansion over three countries (230,400 generations; parse failures 0.09%) with a 12/36 dev/test split frozen before generation and all rules tuned on dev only. A leak audit verifies mechanically that no quantity on the test side uses the true target. Measured: the direct human mean dominates every reweighting variant at both budgets (MAE 0.112/0.040 at n_h=50/400 versus $\geq$ 0.53 for every reweighting variant, including verdict-guided routing), verdict accuracy 0.97/0.98 under the frozen rule, and the routing benefit over undiagnosed reweighting is not significant at the question-cluster level. One further distinction matters: the verdict answers a *diagnostic-protocol* question (is the plateau above threshold), not an *estimator-choice* question (will reweighting beat the naive mean); with pilot-only diagnostics the verdict agrees with the first at 0.951 but with the second at only 0.176, and conflating the two misprices the diagnostic.

Uncertainty addendum. A 12-cluster bootstrap over the matched-question deployment draws gives accuracy 0.912 [0.878, 0.941] and direct-beats-reweighting cell proportion 0.938 [0.889, 0.979]. Re-fitting the threshold c by maximum accuracy on training questions and applying it held out gives mean accuracy 0.910 and agrees with the fixed-c verdicts on $\geq 94.6\%$ of draws in every fold: the pre-registered c is not question-tuned. Independent two-sample verdicts (diagnostics and reference from disjoint calibration halves) agree with the coupled verdicts on 0.945 of draws, showing that the labels are stable to this calibration split; shared observed support still induces structural coupling. A residual-corrected blend using the calibration mean as outcome model is beaten by the direct mean in 246/288 cells and sits between reweighting and direct everywhere.

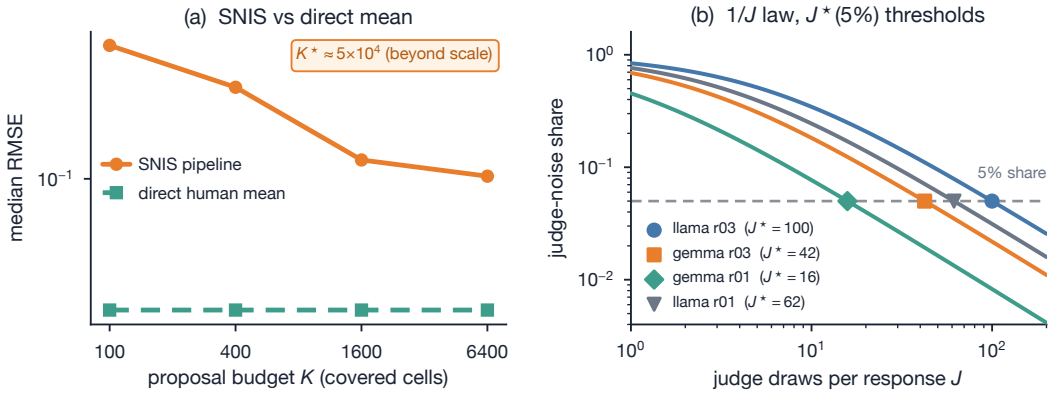


Figure 6: **Direct mean vs reweighting at matched budget, and judge-noise scaling.** (a) At matched budget the direct mean dominates at every tested $K \leq 6400$ in these deployment cells. The plotted $K^\star$ is an extrapolated variance crossover, not a validated superiority threshold. (b) The judge-noise share of the averaged judge's variance decays as $v/(Js + v)$; markers show measured $J^\star(5\%)$ thresholds. Averaging, not the verdict, is where judge noise belongs.

C.6 Matched-budget estimation and action selection

Why the deployed reweighted estimator loses at matched budget. On the 24 covered deployment cells, the direct human mean beats SNIS at every tested $(n_h, K \leq 6400)$ (RMSE 0.0148 vs 0.1039 at n_h=400), and the direct side matches theory to 2% (predicted $\sigma_f/\sqrt{n_h} = 0.0150$). A four-way attribution of the constant factor: the deployed double-clip configuration reproduces the headline RMSE (0.1027 vs 0.1039); removing the clip drops RMSE to 0.0245 (a $\times 4.2$ clip tax); oracle population weights reach 0.0133, below even the direct mean (oracle weights carry target information beyond the calibration sample, and the oracle is not deployable); decoupling the numerator and denominator calibration samples makes things worse (0.1285). Clipping can trade bias against weight instability, so the reading is not "drop the clip": the clip tax is a per-deployment quantity, and the first-order display remains accurate for the unclipped plug-in it describes.

Acting on the verdict. Four actions at matched human budget on the real-target bank: deploy the naive mean, always deploy reweighting, spend the budget on the direct mean, or follow the map (diagnose; if the plateau is predicted low, reweight; else re-scaffold, and fall back to direct). A paired redesign with rate-matched controls (all policies on the same draws; a random coin at the guided policy's realized fallback rate) gives the attribution: the guided policy's entire gain over always-reweighting is the *fallback* component (guided with re-scaffolding only matches always-reweighting), at matched fallback rate the diagnostic adds little over a coin (0.772/0.873 vs 0.751/0.871 MAE), and every non-direct policy is dominated on the per-question mean (excess over direct $+0.65$ to $+1.02$). Cluster-level targeting at n_h=1200 does not change the picture (guided 0.517 vs coin 0.537, question-clustered CI covers zero; the cluster-level policy never falls back, so targeting buys no escape the coin does not already price). The observed benefit is attributable to the fallback rate; the diagnostic does not improve on the rate-matched control in this bank.

Portfolio allocation. Across a portfolio of questions with one fixed total human budget $B \in \{600, 1200, 4800, 12000\}$, is it better to spread thin calibration everywhere and reweight, or to concentrate scarce human draws? A budget-matched comparison (every allocation strategy shares the same pilot, the same direct pool, and the same estimator on unselected questions) finds: targeting is not significant at the stream level (the diagnostic-vs-coin paired gap is negative in every stream, but the 12-question cluster bootstrap covers zero in 12/12 streams), spreading beats targeting decisively (direct-everywhere minus any concentrated allocation is -0.33 to -0.40 with CIs excluding zero in 11–12 of 12 streams), and even perfect oracle targeting knowledge adds nothing statistically beyond the pilot diagnostic. The map is not a scarce-budget router on this evidence; coverage is.

C.7 STRESS AUDITS OF THE ERROR DISPLAY

Benign regime. On a synthetic 2D (n_h, K) grid with known truth, measured MSE converges to the σ_f^2/n_h floor at every n_h as K grows (e.g. n_h=800: $0.00027 \to 0.00004$ over K=200 $\to$ 12800, floor 0.00003), and the two-term predictor tracks measured MSE within a factor ≤ 1.6 everywhere, within $\times 1.04$ at small n_h/large K; the residual factor at large n_h/moderate K reflects the $O(1/\sqrt{Kn_h})$ weight-noise fluctuations the display deliberately omits.

Heavy tails. Re-running the grid with P tilted to $\rho_1 \in \{10, \ldots, 10^4\}$ maps where the display leaves its regime (fully covered, so the plateau mechanisms are off). While $K \gtrsim \rho_1$ the display tracks within $\times 2.4$; once $K \lesssim \rho_1$ the denominator-deviation premise fails and v_f/K over-predicts by up to three orders of magnitude, conservative in direction, uninformative in magnitude. The K_{eff} form under-predicts at high ρ_1 (up to $\times 47$): computed from plug-in weights, it inherits the calibration sample's own blind spots, so it is not a usable bound without an audit. At moderate settings the measured MSE exceeds the display by $\times 2.4$ (an MSE overshoot on this grid, smaller than and distinct from the RMSE-ratio $\times 4.2$ clip tax of App. C.6), a finite-K effect the oracle decomposition does not model.

Premise alignment and interaction with stranded mass. The display tracks (ratio ≥ 0.5) while the denominator-deviation term $a_K \lesssim 1$ and over-predicts from $a_K = 2$ onward; the transition is graded, not a cliff (the ratio declines smoothly through $a_K \in \{1.2, 1.5, 1.8\}$, never a discontinuity). A deployable $\hat{v}_f$ does not repair the display (under-predicts by up to $\times 55$ at high ρ_1, the same blind-spot mechanism): at high mismatch the display needs oracle knowledge the deployment does not have. With stranded mass $\varepsilon \in \{0.05, 0.2\}$ and heavy tails together, measured bias exceeds the idealized plateau $\varepsilon\Delta_f$ by up to $\times 2.2$, but the *calibrated* diagnostic $\hat{\varepsilon}\widehat{\Delta_f}$ tracks the measured bias (median ratio ≈ 1.0) at every ρ_1.

Guard self-audit. A same-sample budget guard is vacuous by construction (Lemma 1): computed with the same proposal draws that define the evaluation support, $\hat{\rho}_1 \leq K$ always, so its failure to fire carries no evidence about ρ_1. "A fire is decisive" is also false: $\mathbb{E}[\hat{\rho}_1] \geq \rho_1$ but a single calibration draw on a rare bin makes $\hat{\rho}_1$ explode, giving false alarms. The correct statement is false-positive-constrained: a Clopper–Pearson lower-bound test with known P controls the false-positive rate at $\leq \delta$ by construction (realized ≤ 0.015 across the null grid), but its power at $\rho_1 = 10^4$ is 0.000–0.017, so a small calibration sample cannot certify heavy-tail mismatch; the test converts a vacuous alarm into a conservative one that essentially never fires falsely and rarely certifies. With unknown P the guarantee does not follow, and the estimated-P variant is a heuristic screen with no false-positive claim.

C.8 ESTIMAND RE-KEYING AND RARE-CATEGORY SHARES

Human-sampling design effects. The matched-question calibration draws are i.i.d. multinomial; real human samples are clustered and weighted. Rerunning the pipeline at $n_{\text{eff}} = n_h/\text{DEFF}$ for DEFF $\in \{1.5, 3\}$ degrades accuracy only to 0.896 and P recall to 0.953 at threefold variance inflation; an explicit clustered design (latent Dirichlet clusters, $\mathbb{E}[\Theta] = Q$ preserved) gives the same or better verdict accuracy at matched variance, the direct mean's RMSE inflates as $\sqrt{\text{DEFF}}$ while reweighting's is unchanged (plateau cells are plateau-dominated, not noise-dominated); these sensitivity analyses support stability under the tested sampling designs. Proposition 3 assumes independent calibration draws; substituting an effective sample size does not establish its tail-detection guarantee for arbitrary dependence.

Estimand re-keying on real targets. Comparing the mean-score estimand against a top-box estimand ("favorable share") on the same real-target streams flips the truth-side R/P classification in 29/144 cells: the plateau functional is estimand-specific, so a verdict certified for one estimand does not transfer to another. The estimator ranking is stable (reweighting worse than direct in 138/144 cells under the top-box estimand).

Rare-category shares. Sweeping the estimand family (mean, top-box, category indicators) on the real-target battery with shared replicate draws: 85/144 cells change class under at least one indicator; verdict accuracy survives for standard shares (0.93–0.96) but degrades on rare categories (the "don't-know" (DK) indicator reaches only 0.64; small shares make both the plateau functional and the diagnostics noisy). Deployed reweighting appeared to win on 52 DK cells; the decomposition

shows 44 are trivial ($q_{\mathrm{DK}} = 0$, where estimators identically equal to zero tie), 3 are RMSE-rounding ties, and only 5 are genuine corrections (each $\leq$0.0005 RMSE), the bias–variance shrinkage boundary. Against deployable shrinkage baselines fit on dev and evaluated on the held-out models, linear shrinkage attains RMSE 0.0039, the direct share 0.0041, Beta–Binomial empirical Bayes 0.0045, while deployed reweighting scores 0.096, an order of magnitude worse, and beats the best shrinkage baseline in **0/52** of the "winning" cells: in this rare-share evaluation, simple shrinkage is the strongest reported baseline; empirical Bayes also beats reweighting but does not improve on the direct share. The mechanism of the reweighting failure is identifiable: the deployed lower clip $\alpha = 0.02$ acts as an artificial prior that pushes rare-category estimates off zero (mirroring the clip-tax attribution of App. C.6 on covered cells, though no analogous four-way decomposition was run on the DK cells), and the unregularized plug-in diagnostics $\hat{\varepsilon}$ and $\widehat{\Delta_f}$ are noisy on small shares, so they should not decide rare-category trust without smoothing.