REPETITION IS NOT ACCURACY IN LARGE LANGUAGE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated queries to a large language model often produce differently worded answers with the same meaning, and evaluators use this semantic agreement as a black-box reliability signal. We ask whether agreement can support an accuracy claim when labeled source prompts and unlabeled deployment prompts come from different populations. A model can repeatedly return the same wrong answer, so agreement measures concentration rather than truth. We capture this distinction with a promptwise agreement bias: the amount by which agreement overstates or understates accuracy. Under an oracle law for labeled source evaluation, the worst-case ambiguity is exactly the deployment mass outside source coverage. When source information is reduced to mean agreement bias and a justified promptwise range, we derive the full compatible risk interval in the corresponding summary model. Conditional shift bounds and finite-sample intervals then separate the effects of sampling new prompts from sampling additional responses to a fixed prompt. Matching constructions show when more responses help, when broader prompt coverage matters, and when target labels or external assumptions remain unavoidable.

1 INTRODUCTION

Repeated querying offers a practical way to inspect a large language model when its probabilities and internal states are unavailable. Suppose one response says "Paris" and another says "The capital of France is Paris." The wording differs, but the two responses express the same answer. Evaluators can group such responses by meaning and measure how often two independent generations agree, as in black-box hallucination checks and semantic uncertainty methods (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). We call this agreement across repeated responses *self agreement*. It is attractive because it requires no internal model access or deployment labels. This paper asks whether self agreement can support a valid deployment accuracy claim under population shift. Labeled evaluation prompts come from one population, and unlabeled deployment prompts come from another.

The main difficulty remains even with unlimited computation and perfect knowledge of model response behavior. A deterministic model can repeat the same answer forever. In the extreme where deployment lies entirely outside source support, one admissible ground-truth specification makes the observed answer correct and another makes it incorrect. All observed responses, source labels, and unlabeled deployment prompts remain unchanged. No procedure based solely on these observables can distinguish the two worlds. Their target accuracies are one and zero. This is a partial-identification problem rather than ordinary sampling noise. Accordingly, a valid analysis considers the full range of risks compatible with the observables. More generations can reveal the response distribution of the model ever more precisely, but they cannot reveal a missing answer key.

Prompt shift makes this gap operationally important (Seegmiller & Preum, 2026). A model can be accurate and self-consistent on common source prompts. The same model can confidently repeat wrong answers on prompt types that become frequent after deployment. Comparing the two prompt populations does not by itself solve the problem. A small discrepancy controls only the prompt patterns represented by the chosen comparison, not the unknown relation between repeated agreement and truth. Semantic structure can narrow that relation. For example, one possible condition states that the most frequent semantic answer is correct, but this premise comes from the task rather than

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.

from repetition. We study the problem at four levels: which deployment prompts source truth covers, how much information remains after source truth is summarized, which bias patterns transfer across populations, and how finite numbers of prompts and responses affect estimation.

Our goal is to characterize the sharp range of deployment accuracy under different information and transport assumptions. Agreement bias makes the missing uncertainty explicit by measuring how much self agreement overstates or understates accuracy at each prompt. At the strongest oracle level, the full population distribution of labeled source prompts, responses, and truths determines bias on source-covered prompts without requiring model training data, leaving only target-only prompts unresolved. When only the source average and a defensible promptwise range remain, prompt shift creates an allocation problem because the same average can place uncertainty on prompt types with different deployment weights. Without that range, a predeclared family of prompt-computable patterns gives a coarser conditional bound. Finite data add separate statistical errors from too few prompts and too few responses per prompt. This information ladder shows which evidence supports an accuracy statement and which uncertainty repeated generation cannot remove. Our contributions are summarized below.

1. **Prompt overlap determines the exact identification boundary (Section 3).** With the population law for labeled source evaluation, target risk is fixed on source-covered prompts. The probability of genuinely target-only prompts is exactly the remaining worst-case width. Matching constructions show that this width cannot be reduced, and universal point identification holds exactly under full source coverage.
2. **Source calibration and semantic structure determine exact accuracy regions (Section 4).** Given only the source average of agreement bias and an externally justified promptwise range, we derive the entire attainable target-risk interval in the corresponding summary model. A semantic specialization shows precisely where pairwise agreement leaves accuracy ambiguous and where prompt shift enlarges that ambiguity.
3. **Structural assumptions yield conditional prompt-shift bounds (Section 5).** A single decomposition separates the change in representable bias patterns from the truth-dependent residual that those patterns miss. The resulting interval scales residual error by how much the prompt distributions differ and has coefficient-wise equality constructions. It remains explicitly conditional without an externally justified residual bound.
4. **Prompt and response sampling create different errors (Section 6).** High-probability intervals distinguish uncertainty from drawing independent prompts from uncertainty due to repeated responses within each prompt. Lower bounds match the two rates on appropriate subclasses, and a sound acceptance rule turns the interval into a deployment decision without treating its structural premise as observed data.

2 Problem Setting and Observable Information

The evaluation protocol draws a prompt, queries one frozen language model independently several times, and compares the meanings of its responses. On labeled source evaluation prompts, a response can also be compared with the task truth. On deployment prompts, only the responses and prompt identities are observed. We use *target* below as shorthand for this deployment population. Our aim is to infer average deployment error from these two kinds of information without treating repeated answers as labels.

Formally, let $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$ be measurable prompt and response spaces, where each Σ lists the events to which probabilities apply. Source and deployment prompts follow probability measures P and Q. The Markov kernel $G(\cdot \mid x)$ assigns a response distribution to prompt x. Conditional on x, the notation $A, A' \overset{\text{iid}}{\sim} G(\cdot \mid x)$ means two independent responses from that distribution. The task supplies a measurable truth $y^\star(x)$ and a jointly measurable symmetric score $s_x : \mathcal{Y}^2 \to [0, 1]$, where $\mathcal{Y}^2 = \mathcal{Y} \times \mathcal{Y}$ denotes response pairs. We use one fixed task score to compare response pairs and response-truth pairs on the same scale. Validating that score is a separate empirical problem.

Accuracy measures the match between a generated response and truth. Agreement instead measures the match between two independent responses. Their signed difference records how repetition changes the apparent quality of the model.

Definition 1 (Accuracy agreement and agreement bias)**.** For each prompt x, define

$$a(x) = \mathbb{E}\big[s_x(A, y^\star(x)) \mid x\big], \qquad c(x) = \mathbb{E}\big[s_x(A, A') \mid x\big], \qquad h(x) = c(x) - a(x).$$

For any prompt distribution D, $X \sim D$ means that X is drawn from D, and $\mathbb{E}_D$ denotes averaging over this draw. Define

$$R_D = 1 - \mathbb{E}_D a(X), \qquad C_D = \mathbb{E}_D c(X), \qquad H_D = \mathbb{E}_D h(X).$$

We call h the signed agreement bias. Positive bias means that the model agrees with itself more than with truth. Negative bias allows varied responses that remain acceptable.

Repeated target responses determine agreement $c(x)$ but not accuracy $a(x)$ or bias $h(x)$. Source labels reveal the missing comparison only on source prompts. Target agreement is therefore observable while target bias is not. Every positive result about target accuracy requires an explicit account of how source truth constrains target bias.

To keep the information ladder concrete, we use three prompt types. An easy type is accurate and concentrated, an ambiguous type is less concentrated but often acceptable, and a confidently wrong type is concentrated around an incorrect answer. Deployment increases the weight of the last type, so target agreement combined with average source bias understates risk. Every prompt type remains source-covered. Appendix B.4 gives the semantic probabilities and exact calculations. Section 3 first asks what an oracle labeled source law would determine. Later sections retain less truth information and then add sampling error.

3 Exact Identification from Prompt Overlap

We first remove sampling error and study an oracle benchmark. The evaluator knows the source and deployment prompt populations, the full response law, and the population law for labeled source evaluation rather than model training data. We ask what deployment risk this information determines. The first step is an exact identity that turns observable target disagreement into target risk after adding the amount by which self agreement overstates or understates accuracy.

Lemma 2 (Exact risk identity)**.** *For every prompt x and every prompt distribution D,*

$$1 - a(x) = 1 - c(x) + h(x), \qquad R_D = 1 - C_D + H_D.$$

Target generations reveal C_Q but not the truth-dependent correction H_Q. A deterministic answer produces the same observation in correct and incorrect ground-truth worlds. The worlds are therefore observation-equivalent. Source evaluation labels distinguish them only where source covers deployment. To locate precisely what remains unresolved, split deployment into a source-covered part and a target-only part. The absolutely continuous component Q^{ac} reweights source-covered regions. The singular component $Q^\perp$ is target mass on a region with zero source probability. The Lebesgue decomposition is

$$\begin{aligned} Q &= Q^{\mathrm{ac}} + Q^\perp = wP + Q^\perp, \quad \beta = Q^\perp(\mathcal{X}), \\ Q^{\mathrm{ac}}(B) &= \int_B w\, dP, \qquad\qquad w = \frac{dQ^{\mathrm{ac}}}{dP}. \end{aligned} \tag{1}$$

Here B is any measurable prompt set, w is the density ratio, and β is the target-only probability. The superscript $\perp$ marks a singular component, not independence or a power. The subscript $\perp$ below marks a contribution from that component. Correspondingly,

$$\begin{aligned} R_Q &= R_{\mathrm{ac}} + R_\perp, \qquad\qquad R_{\mathrm{ac}} = \int w(x)[1 - a(x)]\, dP(x), \\ R_\perp &= \int [1 - a(x)]\, dQ^\perp(x), \qquad 0 \le R_\perp \le \beta. \end{aligned} \tag{2}$$

The labeled source evaluation law fixes R_{ac}. The possible range of $R_\perp$ has width β. We write $Q \ll P$ when every set of zero probability under P also has zero probability under Q. This measure-level coverage does not require identical prompts in finite samples. A rare prompt can be absent from a sample and still belong to the covered component when it has positive source probability. A statement

holds almost surely under a measure when its failure set has measure zero. The identified set collects compatible risks from truth assignments that preserve all observations. Its diameter is their largest separation, and point identification means one value. Thus β is intrinsic.

The split also gives the evaluation logic behind the result. On the covered component, w transports the response-to-truth comparison already fixed by the source evaluation law. On the singular component, no source truth is available, so only the unit range of risk remains. The theorem below proves that no other ambiguity survives under this oracle information. Each interval contribution is sharp.

Theorem 3 (Exact overlap frontier)**.** *Fix P, Q and one task score, and consider task worlds with the same complete labeled source law and unlabeled target prompt-generation law. Any two compatible target risks differ by at most β, and for every P, Q a binary exact-match construction attains this difference. Thus β is the exact worst-case identification diameter. The interval $[R_{\mathrm{ac}}, R_{\mathrm{ac}} + \beta]$ covers every compatible world, its midpoint has uniform error at most $\beta/2$, and universal point identification holds exactly when $Q \ll P$.*

The proof fixes the covered integral from the labeled source law, then assigns opposite truths to a deterministic answer on source-null target prompts. Observations remain unchanged, and risk moves by β. The same pair forces uniform estimation error $\beta/2$ and confidence-interval width β. Appendix Proposition 14 gives the decision statements, and Appendix A.2 proves the population claim.

The frontier therefore depends on support coverage rather than the numerical size of shift. Complete coverage gives $\beta = 0$ and identifies risk despite arbitrary reweighting, partial coverage leaves exactly the target-only contribution unresolved, and $\beta = 1$ recovers the risk-zero and risk-one construction. A large shift can be harmless when the complete truth link is known on covered prompts. A small target-only mass creates an equally sized worst-case ambiguity. Support coverage answers identification, not statistical stability. Appendix Corollary 21 shows where density-ratio control enters finite-sample importance weighting. This conclusion uses the full source function, not merely its average. Section 4 studies the ambiguity created by that compression.

4 Exact Accuracy Regions from Source Calibration and Semantic Structure

Section 3 identified deployment risk when the full source truth link is known. We now ask what remains identifiable when source information is reduced to its mean and promptwise limits. The central difficulty is that prompt shift can reweight where the unresolved bias lies.

To encode this location uncertainty, let measurable functions $\underline{h}, \overline{h}$ provide promptwise limits on the difference between agreement and accuracy.

$$c - 1 \le \underline{h} \le \overline{h} \le c, \qquad b = \overline{h} - \underline{h}. \tag{3}$$

The outer inequalities keep accuracy in $[0, 1]$. In this reduced model, compatible links are measurable h in the band with $\mathbb{E}_P h = H_P$. The source mean fixes total bias but not its allocation across prompts. Appendix B.4 gives a two-prompt construction that moves the same bias between prompt types with different target weights and preserves every retained source summary. We retain only $(P, Q, C_Q, H_P, \underline{h}, \overline{h})$ and discard generator restrictions beyond C_Q, so exactness is relative to this link class.

Every such compatible link can be written as

$$h = \underline{h} + bz, \qquad 0 \le z \le 1, \tag{4}$$

where z records how much of the available width is used at each prompt. The known source average imposes the budget

$$d = H_P - \mathbb{E}_P \underline{h}, \qquad \mathcal{Z}(d) = \{z : \mathcal{X} \to [0, 1] \text{ measurable} : \mathbb{E}_P[bz] = d\}. \tag{5}$$

Thus d is the source bias budget above the lower band, and $\mathcal{Z}(d)$ contains all allocations that spend it. Deployment prices an allocation by w. In this accounting view, z chooses the locally permitted bias, d fixes its source total, and w changes its target value. The smallest target bias spends where w is low. The largest spends where w is high. Define these continuous-knapsack values by

$$\Phi_-(d) = \inf_{z \in \mathcal{Z}(d)} \mathbb{E}_P[wbz], \qquad \Phi_+(d) = \sup_{z \in \mathcal{Z}(d)} \mathbb{E}_P[wbz]. \tag{6}$$

This reduction separates three operations that would otherwise be mixed together. The lower band fixes a baseline, the source moment determines the amount of additional bias to allocate, and the density ratio determines where that allocation matters most after shift. Solving the two allocation problems therefore yields both endpoints rather than only an upper bound.

Theorem 4 (Exact truth-band region)**.** *The reduced model is feasible exactly when $0 \le d \le \mathbb{E}_P b$. When feasible, its target-risk identified set is the closed interval*

$$\mathcal{I}_{\mathrm{band}} = 1 - C_Q + \left[\mathbb{E}_Q \underline{h} + \Phi_-(d),\ \mathbb{E}_Q \underline{h} + \Phi_+(d) + \int b\, dQ^{\perp}\right].$$

Every point in the interval is attained by a compatible measurable link in this reduced model.

The baseline $1 - C_Q + \mathbb{E}_Q \underline{h}$ assigns the lower truth link before source calibration is enforced. The knapsack gap measures uncertainty about where the required source bias lies on covered prompts. The integral $\int b\, dQ^{\perp}$ measures the remaining freedom on prompts observed only at deployment. The proof orders covered prompts by w, fills the budget from opposite ends, and varies target-only prompts freely. Compactness attains the two endpoints. Convexity fills the interval. Appendix A.3 gives the construction. In plain terms, the interval collapses only when deployment reweights every still-ambiguous covered prompt by the same factor and target-only prompts contain no remaining ambiguity. To state this criterion, define the weighted measure $\mu = bP$ by $\mu(B) = \int_B b\, dP$.

Corollary 5 (Necessary and sufficient point identification)**.** *Let $\mu = bP$. If $0 < d < \mu(\mathcal{X})$, then $\mathcal{I}_{\mathrm{band}}$ is a singleton if and only if w is constant μ-almost surely and $\int b\, dQ^{\perp} = 0$. If $d \in \{0, \mu(\mathcal{X})\}$, it is a singleton if and only if $\int b\, dQ^{\perp} = 0$.*

This separates overlap from source compression. Complete overlap removes the singular term, but one source mean leaves ambiguity wherever deployment reweights a positive-width band. Let $\mathcal{I}_0$ use only the band, let $\mathcal{I}_H = \mathcal{I}_{\mathrm{band}}$ also use H_P, and let $\mathcal{I}_{\mathrm{full}}$ fix a band-valued link h_P almost surely under P. For an interval $\mathcal{I}$, $\mathrm{diam}(\mathcal{I})$ is its endpoint difference.

Corollary 6 (Source-information ladder)**.** *If $H_P = \mathbb{E}_P h_P$, then*

$$\mathcal{I}_{\mathrm{full}} \subseteq \mathcal{I}_H \subseteq \mathcal{I}_0, \qquad \begin{aligned} \mathrm{diam}(\mathcal{I}_{\mathrm{full}}) &= \int b\, dQ^{\perp}, \\ \mathrm{diam}(\mathcal{I}_H) &= \Phi_+(d) - \Phi_-(d) + \int b\, dQ^{\perp}, \quad \mathrm{diam}(\mathcal{I}_0) = \mathbb{E}_Q b. \end{aligned}$$

Every point of all three intervals is attainable in the reduced truth-link model. Appendix A.3 gives their endpoints.

Reading from right to left, a source mean removes some band uncertainty, and the full function removes it on covered prompts. The remaining allocation ambiguity also has a sharp quantitative bound. Let $B = \mu(\mathcal{X}) > 0$. Appendix Proposition 15 proves

$$0 \le \Phi_+(d) - \Phi_-(d) \le \operatorname*{ess\,osc}_{\mu}(w) \min\{d, B - d\}. \tag{7}$$

Here the essential oscillation is the range of w after prompts of zero μ-measure are ignored. The bound becomes small when deployment reweighting is nearly constant on ambiguous prompts or when source calibration places the bias near a band endpoint. Its coefficient one is attainable. When $P = Q$, the density ratio equals one and the source mean identifies target bias even for a wide band. Thus ambiguity comes from reweighting an unresolved allocation, not from band width alone. The three diameters quantify the value of source information and the ambiguity that prompt shift can amplify.

Semantic information gives a concrete band. Pairwise agreement records only class equality, so several semantic distributions can have the same collision value but different modal accuracies. Appendix B.4 gives a three-class construction showing that modal truth alone does not identify accuracy. Suppose responses fall into at most $M \ge 2$ exact semantic classes with probabilities $p = (p_1, \ldots, p_M)$, where $p_j \ge 0$ and $\sum_j p_j = 1$. Exact matching gives $c = \sum_j p_j^2$. Modal truth gives $a = p_{\max}$ with $p_{\max} = \max_j p_j$. For $c \in [1/M, 1]$ and $k = \min\{M - 1, \lfloor 1/c \rfloor\}$, where $\lfloor 1/c \rfloor$ is the greatest integer no larger than $1/c$, define

$$L(c) = \frac{k + \sqrt{k((k+1)c - 1)}}{k(k+1)}, \qquad U_M(c) = \frac{1 + \sqrt{(M-1)(Mc - 1)}}{M}. \tag{8}$$

Theorem 7 (Modal accuracy envelope from Rioul, 2023)**.** *Under exact semantic matching with at most M classes and modal truth,*

$$L(c) \le a = p_{\max} \le U_M(c).$$

Both endpoints are attainable for every admissible c.

The lower and upper envelope constructions respectively pack or spread response mass while fixing agreement. They meet for uniform and deterministic responses, but generally leave a range. Appendix Lemma 17 proves that every intermediate value is attainable, including ties and reciprocal boundaries. The semantic assumption therefore gives the exact band

$$c - U_M(c) \le h \le c - L(c), \qquad b_M(c) = U_M(c) - L(c). \tag{9}$$

Corollary 8 (Exact collision-summary regions)**.** *In the reduced collision-summary model, the superscript zero indicates that no source bias moment is available. Under modal truth and known M, the promptwise collision field $c(\cdot)$ known Q-almost surely gives the exact risk interval $\mathcal{I}^0_{\mathrm{sem}} = [1 - \mathbb{E}_Q U_M(c),\, 1 - \mathbb{E}_Q L(c)]$. Its diameter is $\mathbb{E}_Q b_M(c)$, and every point is attainable. For the source-mean refinement, suppose the same collision field is also known P-almost surely. Adding H_P then shrinks this set exactly to Theorem 4 with Eq. (9). If $d_M := H_P - \mathbb{E}_P[c - U_M(c)]$ satisfies $0 < d_M < \mathbb{E}_P b_M(c)$, this source-mean region is a singleton if and only if w is constant $b_M(c)P$-almost surely and $\int b_M(c)\, dQ^{\perp} = 0$.*

With two classes, agreement determines modal accuracy. With more classes, only the uniform and deterministic boundaries do so automatically. Near-modal truth expands only the worst-case endpoint, as Appendix Corollary 18 shows. This route needs promptwise semantic structure and known reweighting. Section 5 instead uses a transferable family of bias patterns.

5 Conditional Accuracy Bounds under Prompt Shift

Section 4 obtained an exact region from promptwise bias limits. Such local information is not always available. In that case, prompt-computable witness functions approximate the bias and yield a conditional outer interval that separates observable prompt shift from truth-dependent approximation error.

Let $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ be a nonempty, predeclared class of measurable candidate bias functions. The evaluator can compute each candidate from a prompt without knowing its truth. Examples include topic, format, or prompt metadata groups. Prompt distributions reveal how these patterns shift. A truth-dependent residual records the bias variation that the class misses. Predeclaration matters after sampling. A class selected from the evaluation prompts changes the deviation problem and needs independent selection data or data-dependent complexity control. Appendix Corollary 28 gives an expected-Rademacher extension for a fixed measurable class. Define

$$\begin{aligned} \Delta_{\mathcal{F}}(P,Q) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}_P f - \mathbb{E}_Q f\right|, \\ \eta_{\mathcal{F}}(h) &= \inf_{f\in\mathcal{F}} \|h - f\|_{L_\infty(P+Q)}, \\ \mathrm{TV}(P,Q) &= \sup_{B\in\Sigma_{\mathcal{X}}} |P(B) - Q(B)|. \end{aligned} \tag{10}$$

Here B is a measurable prompt event. The sum measure $P + Q$ ignores only jointly null sets. For measurable u, $\|u\|_{L_\infty(P+Q)}$ is the smallest almost-sure bound on $|u|$. The terms $\Delta_{\mathcal{F}}$, $\eta_{\mathcal{F}}(h)$, and total variation are respectively the largest candidate mean shift, the best worst-case bias error, and the largest event-probability shift. Unlabeled prompts estimate the first term when $\mathcal{F}$ is finite or has controlled complexity, but they cannot validate the truth-dependent second term.

Choose one candidate f and split bias into explained part f and residual $h - f$. Prompt shift moves each component. The expression $(\mathbb{E}_Q - \mathbb{E}_P)f$ means the target average minus the source average, giving

$$R_Q - (1 - C_Q + H_P) = (\mathbb{E}_Q - \mathbb{E}_P)f + (\mathbb{E}_Q - \mathbb{E}_P)(h - f). \tag{11}$$

The center combines target disagreement with source calibration. Its radius pays for movement of the candidate and its unexplained residual. The bound uses one common f to control the two terms.

Choosing different functions for a small mean shift and a small residual would describe no single transferable bias pattern. For measurable u, $\operatorname{ess\,osc}_{P+Q} u = \operatorname{ess\,sup}_{P+Q} u - \operatorname{ess\,inf}_{P+Q} u$ is its range after jointly null sets are ignored. The two terms have different evidential status. Prompt samples can compare the mean of a fixed f across populations, but they cannot determine how closely f follows the truth-dependent bias. The joint infimum describes the best transferable explanation in the declared class. The second line below replaces it by quantities that can be bounded separately.

Theorem 9 (TV-sensitive prompt-shift interval)**.** *For every nonempty measurable $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$,*

$$|R_Q - (1 - C_Q + H_P)| \le \inf_{f \in \mathcal{F}} \left\{ |(\mathbb{E}_Q - \mathbb{E}_P) f| + \mathrm{TV}(P,Q) \operatorname*{ess\,osc}_{P+Q}(h - f) \right\}$$
$$\le \Delta_{\mathcal{F}}(P,Q) + 2\,\mathrm{TV}(P,Q)\eta_{\mathcal{F}}(h).$$

Equivalently, R_Q belongs to the interval centered at $1 - C_Q + H_P$ with the displayed radius, intersected with $[0,1]$.

Because each population has mass one, the joint radius pays only for residual variation and uses one candidate for the two terms. A two-prompt construction in Appendix A.5 makes this radius zero. The separated relaxation equals one in the same construction. Enlarging $\mathcal{F}$ cannot increase the joint radius, but it can increase the relaxation because its discrepancy is a supremum. Section 6 uses the relaxation because a global bound on $\eta_{\mathcal{F}}(h)$ does not identify the candidate with the smallest truth-dependent residual. Appendix A.5 gives the signed-measure proof. The boundary cases clarify the two prices. With no shift, source calibration transfers exactly. If $h \in \mathcal{F}$, only its observable mean shift remains. Constant candidates leave all variation in the residual. Binary constructions make each coefficient exact in isolation, which rules out any uniform reduction. Appendix Proposition 20 gives the constructions. The symmetric outer interval need not be the exact identified set for every restricted model.

Unlabeled prompt samples estimate only prompt discrepancy; they do not validate the truth-dependent approximation premise. The theorem remains conditional unless credible upper bounds on total variation and approximation error are available. With such bounds in place, Section 6 adds finite-sample uncertainty.

6 Finite Sample Accuracy Intervals

Section 5 treats target agreement, source bias, and prompt discrepancy as known. In practice, all three are estimated from sampled prompts and repeated responses. The two sampling resources are not interchangeable because responses in one cluster share their prompt.

Assumption 10 (Clustered black-box sampling)**.** Source prompts $X_1, \dots, X_{n_P} \overset{\text{iid}}{\sim} P$ have task truths. Target prompts $Z_1, \dots, Z_{n_Q} \overset{\text{iid}}{\sim} Q$ do not. The prompt samples are independent. Conditional on them, response clusters are independent, and the $K \ge 2$ responses in each cluster are iid from one fixed generator configuration. Similarity evaluations are deterministic given their arguments.

Use response pairs for agreement, labeled source responses for bias, and prompt samples for discrepancy. For $1 \le i \le n_P$, $1 \le j \le n_Q$, and $1 \le r \le K$, let $A^P_{ir} \sim G(\cdot \mid X_i)$ and $A^Q_{jr} \sim G(\cdot \mid Z_j)$. The superscript marks the domain, the subscripts mark prompt and response, and a hat marks an estimate. Each agreement estimate is a second-order U-statistic, the average over all unordered response pairs. For source prompt X_i, define

$$\widehat{c}^P_i = \binom{K}{2}^{-1} \sum_{1 \le r < t \le K} s_{X_i}(A^P_{ir}, A^P_{it}),\ \widehat{a}^P_i = K^{-1} \sum_{r=1}^{K} s_{X_i}(A^P_{ir}, y^\star(X_i)),\ \widehat{h}^P_i = \widehat{c}^P_i - \widehat{a}^P_i. \quad (12)$$

For target prompt Z_j, define

$$\widehat{c}^Q_j = \binom{K}{2}^{-1} \sum_{1 \le r < t \le K} s_{Z_j}(A^Q_{jr}, A^Q_{jt}), \quad \widehat{C}_Q = \frac{1}{n_Q} \sum_{j=1}^{n_Q} \widehat{c}^Q_j, \quad \widehat{H}_P = \frac{1}{n_P} \sum_{i=1}^{n_P} \widehat{h}^P_i. \quad (13)$$

For a finite predeclared class,

$$\widehat{\Delta}_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n_P} \sum_{i=1}^{n_P} f(X_i) - \frac{1}{n_Q} \sum_{j=1}^{n_Q} f(Z_j) \right|. \quad (14)$$

Pair scores share responses and are not independent. Source agreement and accuracy also reuse responses, but their difference remains unbiased for $h(X_i)$.

Assume deterministic bounds $\bar{v} \geq \mathrm{TV}(P, Q)$ and $\bar{\eta} \geq \eta_{\mathcal{F}}(h)$ are declared before evaluation. For $\delta \in (0, 1)$, the radius contains structural uncertainty, prompt-sampling terms depending on n_P, n_Q, and response-sampling terms depending on $n_P K, n_Q K$. Let

$$\ell = \log(12/\delta), \quad \ell_{\mathcal{F}} = \log(12|\mathcal{F}|/\delta), \tag{15}$$

$$e_C = \sqrt{\frac{\ell}{2n_Q}} + \sqrt{\frac{2\ell}{n_Q K}}, \quad e_H = \sqrt{\frac{2\ell}{n_P}} + 3\sqrt{\frac{\ell}{2n_P K}}, \quad e_{\mathcal{F}} = \sqrt{\frac{2\ell_{\mathcal{F}}}{n_P}} + \sqrt{\frac{2\ell_{\mathcal{F}}}{n_Q}}. \tag{16}$$

Here ℓ is the confidence penalty and $\ell_{\mathcal{F}}$ also accounts for $|\mathcal{F}|$ candidates. The errors $e_C, e_H, e_{\mathcal{F}}$ control target agreement, source bias, and prompt discrepancy. Their notation also shows which sample helps. Target prompts enter agreement and discrepancy, source prompts enter bias and discrepancy, and K enters only conditional response noise. The resulting rate has two distinct plateaus. Sending K to infinity removes the conditional generation terms but leaves prompt averages uncertain. Increasing the source and target prompt samples cannot remove the structural radius. The dominant component therefore determines which additional resource can narrow the interval.

Theorem 11 (Conditional two-level interval)**.** *Under Assumption 10, suppose $1 \leq |\mathcal{F}| < \infty$ and the bounds $\bar{v}, \bar{\eta}$ are deterministic. With probability at least $1 - \delta$,*

$$R_Q \in \mathcal{I}_{\mathrm{tr}}(\delta) := [\widehat{T} - r_\delta, \widehat{T} + r_\delta] \cap [0, 1], \quad \widehat{T} = 1 - \widehat{C}_Q + \widehat{H}_P,$$

where

$$r_\delta = \widehat{\Delta}_{\mathcal{F}} + 2\bar{v}\bar{\eta} + e_C + e_H + e_{\mathcal{F}}.$$

Here $\widehat{T}$ is the plug-in center, r_δ is its radius, and intersection with $[0, 1]$ removes impossible risks without losing coverage. Probability covers prompt and conditional response draws. Increasing K narrows only response-sampling error. More prompts also reduce prompt-sampling error in population averages and discrepancy. These sampling increases do not change the structural term $2\bar{v}\bar{\eta}$. A wide interval therefore points to substantial measured shift, too few prompts, too few responses, or a weak truth-link premise. The proof follows the same separation. It controls prompt averages first, then conditions on prompts to control response U-statistics. Two uniform deviations handle empirical discrepancy, and a six-event union bound closes the argument. Treating response pairs as independent would give a false rate because the $\binom{K}{2}$ pair scores reuse only K random responses. Response error therefore falls with the square root of K, not with K itself. Appendix A.6 records all events and constants and gives a Bernstein refinement.

Matching barriers separate the resources. Even exact response laws leave error of order $n^{-1/2}$ when only n prompts reveal the population mix. The homogeneous binary construction in Appendix Proposition 23 shows that collision error ε at fixed confidence needs total calls of order ε^{-2} and gives a matching upper bound. These separate constructions do not form one additive minimax lower bound. With fixed even call budgets, a common cluster size, and enough independent prompts, only the deterministic concentration surcharge is minimized at $K = 2$. At a fixed call budget, larger response clusters leave fewer independent prompts. This tradeoff explains why the proved concentration surcharge is minimized at $K = 2$. Appendix Corollary 24 states the limited design claim. The interval then supports a decision only when the structural inputs are credible. We accept a deployment only if the entire interval lies below the allowed risk.

Corollary 12 (Conditional accept reject rule)**.** *Fix a tolerance $\rho \in [0, 1]$ and accept only when $\widehat{T} + r_\delta \leq \rho$. With probability at least $1 - \delta$, every accepted generator has $R_Q \leq \rho$. On the same event, acceptance is guaranteed whenever $R_Q \leq \rho - 2r_\delta$.*

The rule is sound under the declared transport assumptions. A small empirical discrepancy does not validate $\bar{\eta}$, and additional target responses provide no such validation. Exact semantic classes provide a different structural route. With known M and modal truth, Appendix Corollary 25 gives a target-data-only interval after projecting collision endpoints onto the legal domain $[1/M, 1]$. Intersecting it with the transfer interval after splitting the failure probability preserves coverage. An empty intersection flags an incompatible premise or a rare sampling event. The interval decomposition also

indicates which resource is limiting: response-dominated width calls for larger K, prompt-dominated width calls for larger n_P or n_Q, and structural width requires stronger truth information. Additional black-box calls leave the structural premise unchanged. Narrow sampling terms do not validate the approximation bound, so every accuracy statement remains tied to the structural conditions in Sections 3 through 5.

7 Related Work

Uncertainty from repeated responses. Repeated sampling supports self-consistency decoding, selective question answering, black-box confidence elicitation, hallucination checks, and semantic uncertainty methods (Wang et al., 2023; Cole et al., 2023; Xiong et al., 2024; Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). The Consistency Hypothesis gives conditions that connect consistency to correctness (Xiao et al., 2025). Signed alternatives compare predictive collision with label cross-collision or pairwise utility with truth utility (Narimatsu et al., 2023; Wang & Holmes, 2025). Nowak (2025) analyzes biased binary modal self-consistency and a roughly balanced fixed budget. Our setting uses an unbiased pairwise collision U-statistic and draws labels from a different prompt population.

Partial identification with missing target truth. This literature characterizes parameter values compatible with incomplete observations through extrema under bounded selection or reweighting (Manski, 1990; Tamer, 2010; Aronow & Lee, 2013). This perspective applies because target truth is unobserved. Domain-adaptation theory derives identification or target-label uncertainty from structural classes (Kong et al., 2022; Dong et al., 2026). Here, agreement and truth remain unrestricted until a source law, a truth-link band, or a semantic condition is supplied. The resulting sets separate support overlap from information lost when source truth is summarized.

Accuracy transfer under distribution shift. Source-labeled and target-unlabeled evaluation includes domain-adaptation bounds, confidence-based performance estimation, testable learning, and overlap-aware estimation (Ben-David et al., 2010; Germain et al., 2016; Garg et al., 2022; Klivans et al., 2024; Mishra & Liu, 2025). Natural user-prompt shift has been measured in deployed LLM settings (Seegmiller & Preum, 2026). Misspecified covariate-shift analysis quantifies error amplification from density-ratio weighting (Amortila et al., 2024). Our analysis first isolates the exact unresolved mass under the oracle source law, then bounds how prompt shift reweights residual bias relative to a prompt-computable witness.

Semantic collision and calibrated risk. Collision geometry links modal error to the probability that two independent draws disagree (Rioul, 2023). We use this geometry to obtain a semantic truth-link band. Conformal methods control calibrated output policies in factual, non-exchangeable, and shifted settings, including structured-generation limits (Angelopoulos et al., 2024; Mohri & Hashimoto, 2024; Farinhas et al., 2024; Zhu et al., 2026; Kotte, 2026). Agreement-on-the-Line and certifiable-surrogate methods analyze independent predictors or transfer bounded risk through predictor disagreement (Baek et al., 2022; Saxena et al., 2024; Baek et al., 2025; Bazinet et al., 2026; Germain et al., 2015). Our observable is repeated responses from one conditional generator, which leaves its truth link under prompt shift unresolved.

8 Conclusion

This paper asks when repeated responses support deployment-accuracy statements after prompt shift. The signed agreement bias makes explicit the missing link between response concentration and truth. With the population labeled-source law, prompt overlap gives the exact identification boundary; weaker source summaries and semantic structure produce compatible risk regions instead. Prompt-computable witnesses then yield conditional transfer bounds, while finite-sample intervals separate structural uncertainty from prompt and response sampling. The results clarify which evidence narrows an interval and when target labels or external truth information are still needed. The framework assumes a fixed task score, one generator configuration, independent prompt clusters, and defensible semantic or approximation conditions. Adaptive conversations, model updates, and data-dependent scoring are outside its scope. Limited target labels can help diagnose truth-link assumptions in a deployment; extending the analysis to adaptive interactions and learned scoring rules is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools for checking the consistency of mathematical notation and identifying possible typographical or formatting errors in formulas. We have not used generative AI tools for developing the theoretical framework, formulating mathematical claims, deriving results, or generating proofs, and tasks involving datasets, experiments, implementation, qualitative data analysis, and the interpretation of experimental results are not applicable to this work. Additionally, we used generative AI tools for improving the grammar and readability of the manuscript. We reviewed all AI-assisted revisions and manually verified that they preserved the intended mathematical meaning and did not alter the technical claims, derivations, or proofs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and principal proof ideas appear in Sections 2 through 6, and complete proofs appear in Appendix A. Appendix B.4 gives exact calculations for the running example. Appendix B.5 documents the fixed-seed verification procedure. The claims are analytic and are supported by the stated assumptions, complete proofs, and deterministic checks. No model training, external data, or API access is required.

REFERENCES

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *COLT*, pp. 130–160, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024.

Peter M. Aronow and Donald K. K. Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *NeurIPS*, pp. 19274–19289, 2022.

Christina Baek, Aditi Raghunathan, and J. Zico Kolter. Theory of agreement-on-the-line in linear models and Gaussian data. In *AISTATS*, pp. 4015–4023, 2025.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

Mathieu Bazinet, Valentina Zantedeschi, and Pascal Germain. Bound to disagree: Generalization bounds via certifiable surrogates. In *UAI*, pp. 491–520, 2026.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, 2010.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *EMNLP*, pp. 530–543, 2023.

Zhiyi Dong, Zixuan Liu, and Yongyi Mao. On the hardness of unsupervised domain adaptation: Optimal learners and information-theoretic perspective. In *CoLLAs*, pp. 89–111, 2026.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *ICLR*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *ICLR*, 2022.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *JMLR*, 16(26):787–860, 2015.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new PAC-Bayesian perspective on domain adaptation. In *ICML*, pp. 859–868, 2016.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(25):723–773, 2012.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *JASA*, 58(301): 13–30, 1963.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Adam Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *COLT*, pp. 2887–2943, 2024.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *ICML*, pp. 11455–11472, 2022.

Varun Kotte. When can conformal risk control certify LLM outputs? bounds, impossibility, and adaptation for structured generation. *CoRR*, abs/2606.29054, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, pp. 9004–9017, 2023.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141, pp. 148–188. Cambridge University Press, 1989.

Aayush Mishra and Anqi Liu. ODD: Overlap-aware estimation of model performance under distribution shift. In *UAI*, pp. 3031–3047, 2025.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *ICML*, pp. 36029–36047, 2024.

Hiromi Narimatsu, Mayuko Ozawa, and Shiro Kumano. Collision probability matching loss for disentangling epistemic uncertainty from aleatoric uncertainty. In *AISTATS*, pp. 11355–11370, 2023.

Robert Nowak. Estimating the self-consistency of LLMs. *CoRR*, abs/2509.19489, 2025.

Olivier Rioul. The interplay between error, total variation, α-entropy and guessing: Fano and Pinsker direct and reverse inequalities. *Entropy*, 25(7):978, 2023.

Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. In *NeurIPS*, pp. 31854–31906, 2024.

Parker Seegmiller and Sarah Masud Preum. Measuring distribution shift in user prompts and its effects on LLM performance. In *ACL*, pp. 32685–32704, 2026.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NeurIPS*, 2007.

Elie Tamer. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Ziyu Wang and Christopher C. Holmes. On subjective uncertainty quantification and calibration in natural language generation. In *AISTATS*, pp. 3799–3807, 2025.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katya Mirylenka, Nhan H. Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. In *UAI*, pp. 4636–4651, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *ICLR*, 2024.

Yuqicheng Zhu, Jialin Yu, Lin Li, Gengyuan Zhang, Zhen Yang, Steffen Staab, Puneet Dokania, Philip Torr, Jie Tang, and Evgeny Kharlamov. Conformalized large language models under configuration shift. *CoRR*, abs/2608.01460, 2026.

A Complete Proofs

All probabilities below include sampled prompts and conditional generations, except where conditioning is explicit. Essential statements under $P + Q$ exclude sets null under each prompt law.

A.1 Exact identity and nonidentifiability

Proof of Lemma 2. Definition 1 gives $h = c - a$, hence $1 - a = 1 - c + h$ pointwise. Taking expectation under D yields $R_D = 1 - C_D + H_D$. No probabilistic or structural assumption is used. □

Proposition 13 (Pointwise truth-free nonidentifiability)**.** *Two task worlds can share the prompt, similarity, generator, and every finite or infinite generation law, but have risks zero and one. Thus any observation-measurable estimator has error at least one half with probability at least one half in one world. For every $\delta \in (0, 1/2)$, any observation-measurable random interval with coverage at least $1 - \delta$ in each world has length at least one with probability at least $1 - 2\delta$.*

Proof of Proposition 13. Use response space $\{0, 1\}$ and exact-match similarity $s(y, y') = \mathbf{1}\{y = y'\}$, where the indicator $\mathbf{1}\{E\}$ equals one when event E occurs and zero otherwise. At a single prompt, let the generator be deterministic at response 0. In world W_0 set the truth to 0, and in world W_1 set it to 1. The generation sequence is identically zero in each world, with triples $(c, a, R) = (1, 1, 0)$ in W_0 and $(1, 0, 1)$ in W_1.

Let Z have the common distribution of any observation-measurable estimator $\widehat{R}$. The events $\{|Z| < 1/2\}$ and $\{|Z - 1| < 1/2\}$ are disjoint, so their probabilities sum to at most one. At least one world has failure probability at least $1/2$ at error threshold $1/2$.

Finally, let I be a random interval computed from the same observations and suppose $\mathbb{P}_j(R_{Q,j} \in I) \geq 1 - \delta$ for $j = 0, 1$. Because the observation law is common, a union bound under that law gives

$$\mathbb{P}(0 \in I \text{ and } 1 \in I) \geq 1 - 2\delta. \tag{17}$$

Every interval containing zero and one has length at least one, proving the last claim. □

A.2 THE EXACT OVERLAP FRONTIER

Proof of Theorem 3. Let W_0, W_1 be two observation-equivalent worlds. We formulate the oracle source evaluation information without selecting versions of conditional expectations. In world j, let

$$\Lambda_j(dx, da, dy) = P(dx)G_j(da \mid x)\delta_{y_j^\star(x)}(dy) \tag{18}$$

be the joint law of a source prompt, one generation, and its task truth. Here δ_y is the Dirac probability measure concentrated at the single point y. Equality of the complete labeled source laws implies $\Lambda_0 = \Lambda_1$. The score is part of the fixed task specification. Hence, for the nonnegative measurable integrand $w(x)[1 - s_x(a, y)]$, equality of these measures gives

$$\int \big(1 - a_0(x)\big)\, dQ^{\mathrm{ac}}(x) = \int w(x)[1 - s_x(a, y)]\, d\Lambda_0(x, a, y) \tag{19}$$

$$= \int w(x)[1 - s_x(a, y)]\, d\Lambda_1(x, a, y) \tag{20}$$

$$= \int \big(1 - a_1(x)\big)\, dQ^{\mathrm{ac}}(x). \tag{21}$$

This argument uses equality of the joint source laws, rather than a pointwise choice of their conditional versions. Each remaining singular-part risk lies in $[0, Q^\perp(\mathcal{X})]$. Therefore

$$|R_{Q,0} - R_{Q,1}| = \left| \int (1 - a_0)\, dQ^\perp - \int (1 - a_1)\, dQ^\perp \right| \le \beta. \tag{22}$$

The common absolutely continuous contribution $R_{\mathrm{ac}} = \int (1 - a_j)\, dQ^{\mathrm{ac}}$ is identified by the first display and does not depend on j. Every compatible risk lies in $[R_{\mathrm{ac}}, R_{\mathrm{ac}} + \beta]$. Its midpoint has uniform absolute error at most $\beta/2$, and the full interval has uniform coverage and length β.

To attain the bound for arbitrary P, Q, singularity supplies a measurable set S such that $P(S) = 0$ and $Q^\perp(S^c) = 0$. Since $Q^{\mathrm{ac}} = wP$, it follows that $Q(S) = \beta$. Take response space $\{0, 1\}$, exact-match similarity, and the generator that deterministically returns zero at every prompt. In W_0, set $y_0^\star(x) = 0$ everywhere. In W_1, set $y_1^\star(x) = \mathbf{1}\{x \in S\}$. The labeled source laws coincide because $P(S) = 0$, and the target prompt-generation laws coincide because the generators are identical. Their risks are respectively zero and $Q(S) = \beta$.

The attaining pair also proves the statistical lower bounds. Under its common observation law, let Z be any estimator. The events $\{|Z - R_{Q,0}| < \beta/2\}$ and $\{|Z - R_{Q,1}| < \beta/2\}$ are disjoint. Thus at least one world satisfies

$$\mathbb{P}_j\big(|Z - R_{Q,j}| \ge \beta/2\big) \ge \tfrac{1}{2}. \tag{23}$$

If a random interval I has coverage at least $1 - \delta$ in each world, a union bound under the common observation law yields

$$\mathbb{P}\big(R_{Q,0} \in I,\ R_{Q,1} \in I\big) \ge 1 - 2\delta. \tag{24}$$

On this event its length is at least $|R_{Q,1} - R_{Q,0}| = \beta$.

Finally, $Q \ll P$ is equivalent to $\beta = 0$. The upper bound then gives point identification for every task world. If $Q \not\ll P$, then $\beta > 0$ and the construction above gives two observation-equivalent worlds with different target risks. This proves the universal if-and-only-if statement. □

Proposition 14 (Decision lower bounds at the overlap frontier)**.** *For the pair of observation-equivalent worlds attaining Theorem 3, every estimator measurable with respect to their common observations has error at least $\beta/2$ with probability at least one half in one of the worlds. For every $\delta \in (0, 1/2)$, every observation-measurable random interval having coverage at least $1 - \delta$ in each world has length at least β with probability at least $1 - 2\delta$.*

Proof. Under the common observation law, the events that an estimator lies within strictly less than $\beta/2$ of the two risks are disjoint, so one has probability at most one half. If a random interval covers each risk with probability at least $1 - \delta$, a union bound gives simultaneous coverage with probability at least $1 - 2\delta$, which forces length at least β. □

A.3 EXACT IDENTIFICATION UNDER A TRUTH-LINK BAND

Proof of Theorem 4. Write $B = \mathbb{E}_P b$. Every admissible h has the representation

$$h = \underline{h} + bz, \qquad 0 \le z \le 1, \tag{25}$$

where $z = (h - \underline{h})/b$ on $\{b > 0\}$ and $z = 0$ on $\{b = 0\}$. Conversely, every measurable $z : \mathcal{X} \to [0, 1]$ in Eq. (25) defines a link in the stated band. Its source moment equals H_P exactly when $\mathbb{E}_P[bz] = d$. Necessarily $0 \le d \le B$. This condition is also sufficient. For $B > 0$, the constant $z = d/B$ works. For $B = 0$, feasibility gives $d = 0$, and $z = 0$ works. The outer inequalities in Eq. (3) additionally ensure that $a = c - h \in [0, 1]$ for every such reduced link.

We next verify attainment of the knapsack extrema. A property holds P-almost surely when its failure set has P measure zero. The space $L_\infty(P)$ contains essentially bounded measurable functions modulo equality P-almost surely, and $L_1(P)$ contains integrable functions. The weak-* topology on $L_\infty(P)$ is the coarsest topology that makes $z \mapsto \int gz\, dP$ continuous for every $g \in L_1(P)$. Regard the feasible functions as elements of $L_\infty(P)$. The order interval $\{z : 0 \le z \le 1\}$ is weak-* compact. It is a weak-* closed subset of the unit ball, because its defining inequalities can be tested against every nonnegative function in $L_1(P)$. The budget constraint is weak-* closed since $b \in L_1(P)$. Thus $\mathcal{Z}(d)$ is weak-* compact. Moreover,

$$\int wb\, dP = \int b\, dQ^{\mathrm{ac}} < \infty, \tag{26}$$

so $z \mapsto \mathbb{E}_P[wbz]$ is weak-* continuous and attains the two extrema. Since $\mathcal{Z}(d)$ is convex and the objective is linear, its image is the entire closed interval $[\Phi_-(d), \Phi_+(d)]$. Equivalently, these are the lower and upper continuous-knapsack values. An optimizer allocates the fractional budget to the smallest or largest values of w, with fractional allocation on a threshold level when needed. This description remains valid when P has atoms because z itself is allowed to be fractional.

Let S be a measurable singular support, with $P(S) = 0$ and $Q^\perp(S^c) = 0$. The Lebesgue decomposition and Eq. (25) give

$$\begin{aligned} H_Q &= \mathbb{E}_Q \underline{h} + \int bz\, dQ^{\mathrm{ac}} + \int bz\, dQ^\perp \\ &= \mathbb{E}_Q \underline{h} + \mathbb{E}_P[wbz] + \int bz\, dQ^\perp. \end{aligned} \tag{27}$$

Changing z on S does not affect its source budget. By taking it constant on S, the last term fills exactly $[0, \int b\, dQ^\perp]$. The first variable term fills $[\Phi_-(d), \Phi_+(d)]$, as just proved. Adding the two intervals means adding their lower endpoints and their upper endpoints, which gives the closed interval stated in Theorem 4. Every point is attained within the reduced link class by combining a convex mixture of the two absolutely continuous optimizers with an appropriate constant value on S. Finally, Lemma 2 adds the fixed term $1 - C_Q$, yielding exactly the displayed target-risk set. □

Proof of Corollary 5. Let $\mu = bP$, put $B = \mu(\mathcal{X})$, and denote the singular width by $B_\perp = \int b\, dQ^\perp$. From Eq. (27), the diameter of the identified interval is

$$\Phi_+(d) - \Phi_-(d) + B_\perp. \tag{28}$$

Suppose first that $0 < d < B$. If $w = \lambda$ μ-almost surely, then every feasible z has

$$\mathbb{E}_P[wbz] = \int wz\, d\mu = \lambda \int z\, d\mu = \lambda d, \tag{29}$$

so the absolutely continuous part has zero width. Together with Eq. (28), this proves sufficiency of the two stated conditions.

For necessity, suppose w is not constant μ-almost surely. Some finite $r < s$ give sets $A = \{w \le r\}$ and $D = \{w \ge s\}$ of positive μ-measure. Define the bounded function

$$\psi = \frac{\mathbf{1}_A}{\mu(A)} - \frac{\mathbf{1}_D}{\mu(D)}. \tag{30}$$

Then $\int \psi \, d\mu = 0$, but

$$\int w\psi \, d\mu = \frac{1}{\mu(A)} \int_A w \, d\mu - \frac{1}{\mu(D)} \int_D w \, d\mu \le r - s < 0. \tag{31}$$

Set $\alpha = d/B \in (0,1)$. For small enough $\epsilon > 0$, the functions $z_+ = \alpha + \epsilon\psi$ and $z_- = \alpha - \epsilon\psi$ take values in $[0,1]$. They have the same budget d but different objective values. Hence $\Phi_+(d) > \Phi_-(d)$. This bounded zero-mean perturbation does not split atoms, and therefore proves necessity on atomic spaces as well. Eq. (28) separately makes $B_\perp = 0$ necessary.

At $d = 0$, nonnegativity forces $z = 0$ μ-almost surely, and at $d = B$ it forces $z = 1$ μ-almost surely. Thus the absolutely continuous objective is fixed at each boundary, irrespective of w. This includes $B = 0$, when the only feasible value is $d = 0$. In every boundary case Eq. (28) is zero exactly when $B_\perp = 0$, as claimed. □

Proof of Corollary 6. For clarity, define the three admissible link classes

$$\mathcal{H}_0 = \{h : \underline{h} \le h \le \overline{h}\}, \quad \mathcal{H}_H = \{h \in \mathcal{H}_0 : \mathbb{E}_P h = H_P\}, \quad \mathcal{H}_{\text{full}} = \{h \in \mathcal{H}_0 : h = h_P \ P\text{-a.s.}\}. \tag{32}$$

The premise $H_P = \mathbb{E}_P h_P$ gives $\mathcal{H}_{\text{full}} \subseteq \mathcal{H}_H \subseteq \mathcal{H}_0$, hence the same nesting for their risk images under $h \mapsto 1 - C_Q + \mathbb{E}_Q h$.

With no source restriction, representation Eq. (25) gives

$$\mathcal{I}_0 = 1 - C_Q + \left[\mathbb{E}_Q \underline{h}, \mathbb{E}_Q \overline{h}\right]. \tag{33}$$

Indeed, the endpoints use $z = 0$ and $z = 1$, and constant $z = t \in [0,1]$ fills the interval. Theorem 4 gives $\mathcal{I}_H = \mathcal{I}_{\text{band}}$ and diameter $\Phi_+(d) - \Phi_-(d) + \int b \, dQ^\perp$.

For the full source link, the absolutely continuous contribution is fixed as follows.

$$\mathbb{E}_Q h = \mathbb{E}_P[w h_P] + \int h \, dQ^\perp. \tag{34}$$

On a measurable singular support, setting $h = \underline{h} + tb$, $t \in [0,1]$, and retaining $h = h_P$ P-almost surely fills exactly

$$\mathcal{I}_{\text{full}} = 1 - C_Q + \mathbb{E}_P[w h_P] + \left[\int \underline{h} \, dQ^\perp, \int \overline{h} \, dQ^\perp\right]. \tag{35}$$

Thus every point of all three closed intervals is attained, and their widths are precisely those stated in Corollary 6. Taking the unrestricted band $[\underline{h}, \overline{h}] = [c-1, c]$ makes $b \equiv 1$, so the full-information width reduces to $Q^\perp(\mathcal{X}) = \beta$, consistent with Theorem 3. □

For a bounded measurable function u and a finite measure μ, the quantities $\operatorname{ess\,sup}_\mu u$ and $\operatorname{ess\,inf}_\mu u$ are its smallest almost-sure upper bound and largest almost-sure lower bound. Their difference $\operatorname{ess\,osc}_\mu(u)$ is the essential oscillation.

Proposition 15 (Quantitative ambiguity from reweighting)**.** *Let $\mu = bP$, $B = \mu(\mathcal{X}) > 0$, and $0 \le d \le B$. Then*

$$\Phi_+(d) - \Phi_-(d) \le \operatorname*{ess\,osc}_\mu(w) \min\{d, B - d\}.$$

Consequently, the diameter of $\mathcal{I}_{\text{band}}$ is bounded by the right side plus $\int b \, dQ^\perp$. The inequality is interpreted in the extended-real sense when the essential oscillation is infinite, and its coefficient one is attainable.

Proof. Let z_+ and z_- attain the maximum and minimum, put $v = z_+ - z_-$, and define its positive and negative parts by $v_+ = \max\{v, 0\}$ and $v_- = \max\{-v, 0\}$. Their common budget d gives

$$\int v \, d\mu = 0, \qquad \tau := \int v_+ \, d\mu = \int v_- \, d\mu. \tag{36}$$

Because $v_+ \le z_+$ and $v_- \le 1 - z_+$, we have $\tau \le d$ and $\tau \le B - d$. Subtracting the essential infimum of w and using $\int v \, d\mu = 0$ gives

$$\Phi_+(d) - \Phi_-(d) = \int wv \, d\mu \le \operatorname*{ess\,osc}_\mu(w) \tau, \tag{37}$$

which proves the bound, including $d \in \{0, B\}$. If $B = 0$, the absolutely continuous width is already zero and only the singular term remains.

For equality, take two prompts with $P = (1/2, 1/2)$, $c \equiv 1$, $\underline{h} = 0$, $\overline{h} = 1$, and $d = 1/2$. For any $v_0 \in (0, 1)$, set $w = (1 - v_0, 1 + v_0)$, so $Q = ((1 - v_0)/2, (1 + v_0)/2)$ and P, Q are mutually absolutely continuous. The minimizing and maximizing allocations are $(1, 0)$ and $(0, 1)$, respectively, giving knapsack width v_0. The upper bound is also $2v_0 \min\{1/2, 1/2\} = v_0$, proving equality and showing that zero singular mass does not remove ambiguity after the source function is compressed to its mean. □

A.4 Sharp semantic bounds

We first record the two extremal collision facts used in Theorem 7.

Lemma 16 (Collision at fixed modal mass)**.** *Let $p \in \mathbb{R}_+^M$ sum to one and have maximum coordinate $m \geq 1/M$. With $r = \lfloor 1/m \rfloor$,*

$$m^2 + \frac{(1-m)^2}{M-1} \leq \sum_{j=1}^{M} p_j^2 \leq rm^2 + (1 - rm)^2.$$

The lower endpoint is attained by $(m, (1-m)/(M-1), \ldots, (1-m)/(M-1))$. The upper endpoint is attained, up to permutation, by $(\underbrace{m, \ldots, m}_{r}, 1 - rm, 0, \ldots, 0)$, omitting the residual coordinate when it is zero.

Proof. Choose an index with mass m. By Cauchy–Schwarz on the remaining $M - 1$ coordinates,

$$\sum_{j=1}^{M} p_j^2 \geq m^2 + \frac{(\sum_{j>1} p_j)^2}{M-1} = m^2 + \frac{(1-m)^2}{M-1}, \tag{38}$$

with equality exactly when the remaining masses are equal.

For the upper bound, consider two coordinates $0 < u \leq v < m$. If $u + v \leq m$, replacing (u, v) by $(0, u + v)$ changes their squared sum by $2uv \geq 0$. If $u + v > m$, replace them by $(u + v - m, m)$. The change is

$$(u + v - m)^2 + m^2 - u^2 - v^2 = 2(m - u)(m - v) \geq 0. \tag{39}$$

Iterating this mass transfer leaves at most one coordinate strictly between zero and m. Because the coordinates sum to one, the terminal vector has $r = \lfloor 1/m \rfloor$ entries equal to m and residual $1 - rm \in [0, m)$, proving the upper bound and its attainment. □

Proof of Theorem 7. Under modal truth, $a = m = p_{\max}$. The upper bound in Lemma 16 is inverted piecewise. Fix $c \in [1/M, 1]$ and put $k = \min\{M - 1, \lfloor 1/c \rfloor\}$, so $c \in [1/(k + 1), 1/k]$, with the evident shared-boundary convention. If $m > 1/k$, then $m \geq L(c)$ because $L(c) \leq 1/k$. Otherwise $c \leq m$ implies $m \geq c \geq 1/(k + 1)$, hence $m \in [1/(k + 1), 1/k]$ and $\lfloor 1/m \rfloor = k$ except at a boundary where the two expressions agree. Therefore

$$c \leq km^2 + (1 - km)^2 = k(k + 1)m^2 - 2km + 1. \tag{40}$$

The right side is nondecreasing on this interval. Solving its equality for the root in $[1/(k + 1), 1/k]$ gives

$$m \geq \frac{k + \sqrt{k((k+1)c - 1)}}{k(k+1)} = L(c). \tag{41}$$

Taking $m = L(c)$ in the packed vector of Lemma 16 attains collision c and maximum mass $L(c)$. The number of nonzero coordinates is at most $k + 1 \leq M$. At $c = 1/M$ the residual representation reduces to the uniform M-vector.

For the other direction, the lower bound in Lemma 16 gives

$$c \geq m^2 + \frac{(1-m)^2}{M-1}. \tag{42}$$

The right side is increasing for $m \geq 1/M$. Solving the corresponding quadratic yields

$$m \leq \frac{1 + \sqrt{(M-1)(Mc-1)}}{M} = U_M(c). \tag{43}$$

The equal-spread vector in Lemma 16, with $m = U_M(c)$, attains equality and has m as a largest coordinate.

The functions meet continuously at $c = 1/j$, since each branch gives $L(1/j) = 1/j$. On the interior of branch k, $L'(c) = 1/(2\sqrt{k((k+1)c-1)}) > 0$. Continuity at the reciprocal boundaries therefore makes L nondecreasing on $[1/M, 1]$. At $c = 1/M$, the envelopes equal $1/M$ and the distribution is uniform. At $c = 1$, the envelopes equal one and the distribution is deterministic. Modal ties are harmless because every modal truth has mass m. Finally, $c = \sum_j p_j^2 \leq m \sum_j p_j = m$, so the least feasible $m = L(c)$ obeys $L(c) \geq c$. Equality forces every positive p_j to equal m, hence $c = 1/j$ for the number j of positive coordinates. □

Lemma 17 (No gaps inside the modal envelope)**.** *For fixed M and $c \in [1/M, 1]$, the set of attainable modal accuracies is exactly the whole interval $[L(c), U_M(c)]$.*

Proof of Lemma 17. Fix a candidate modal mass $m \in [L(c), U_M(c)]$. Among distributions whose first coordinate is m and whose other coordinates do not exceed m, define

$$p^- = \left(m, \frac{1-m}{M-1}, \ldots, \frac{1-m}{M-1}\right), \tag{44}$$

$$p^+ = \left(\underbrace{m, \ldots, m}_{r}, 1 - rm, 0, \ldots, 0\right), \qquad r = \lfloor 1/m \rfloor, \tag{45}$$

where we arrange one of the entries equal to m as the first coordinate and omit a zero residual. At $m = 1/M$, the second display is the uniform M-vector. Lemma 16 gives

$$c_{\min}(m) := \|p^-\|_2^2 = m^2 + \frac{(1-m)^2}{M-1}, \qquad c_{\max}(m) := \|p^+\|_2^2 = rm^2 + (1-rm)^2. \tag{46}$$

The function $c_{\min}$ is nondecreasing on $[1/M, 1]$, and its inverse boundary is U_M. Thus $m \leq U_M(c)$ implies $c_{\min}(m) \leq c$. On each interval $[1/(r+1), 1/r]$,

$$c'_{\max}(m) = 2r((r+1)m - 1) \geq 0, \tag{47}$$

and adjacent pieces agree at reciprocal endpoints. Its inverse boundary is L, so $m \geq L(c)$ implies $c \leq c_{\max}(m)$. Consequently the desired collision lies between those of p^- and p^+.

Put $v = p^+ - p^-$. The first coordinate of v is zero, and its remaining coordinates sum to zero. Since the remaining coordinates of p^- are equal, $\langle p^-, v \rangle = 0$. Therefore, for $p_t = p^- + tv$, $t \in [0, 1]$,

$$\|p_t\|_2^2 = c_{\min}(m) + t^2\big(c_{\max}(m) - c_{\min}(m)\big). \tag{48}$$

Each endpoint vector has first coordinate m and all coordinates at most m, so the same is true of every p_t. If $c_{\max}(m) > c_{\min}(m)$, choose

$$t = \sqrt{\frac{c - c_{\min}(m)}{c_{\max}(m) - c_{\min}(m)}}. \tag{49}$$

If the two endpoint collisions coincide, feasibility forces their common value to equal c, and take $t = 0$. Thus p_t has collision c and modal mass m. Declaring its first class to be true gives modal accuracy m. Ties cause no difficulty. Hence every point of $[L(c), U_M(c)]$ is attained. The reverse inclusion is Theorem 7, proving equality of the range, including $c = 1/M$, $c = 1$, and all reciprocal boundaries. □

Proof of Corollary 8. Lemma 17 says that, conditional on a collision value c, the complete set of modal accuracies is $[L(c), U_M(c)]$. Since $h = c - a$, its complete pointwise range is therefore

$$\underline{h}_M(c) = c - U_M(c) \leq h \leq \overline{h}_M(c) = c - L(c), \tag{50}$$

with width $b_M(c) = U_M(c) - L(c)$. This band obeys $c - 1 \le \underline{h}_M \le \overline{h}_M \le c$.

For completeness, the pointwise construction can be made simultaneously measurable across prompts, so the specialization introduces no hidden selection assumption. Given any measurable h in this band, put $m = c - h$. The packed vector p^+, the equal-spread vector p^-, and the interpolation coefficient in Eq. (48) are Borel functions of (c, m) on each of the finitely many regions determined by $\lfloor 1/m \rfloor$. The coefficient is set to zero when its denominator vanishes. Hence their coordinates define a measurable finite-class Markov kernel, and choosing the first, modal coordinate as truth realizes the prescribed h. This is an exact realization for the collision-summary experiment. It does not assert that full class probabilities have been discarded when the complete generation law is observed.

Without a source constraint, pointwise modal risk ranges over $[1 - U_M(c), 1 - L(c)]$. Integrating gives the endpoints stated in Corollary 8. Conversely, for any $t \in [0, 1]$, choose the measurable accuracy field $a_t = L(c) + t[U_M(c) - L(c)]$ and apply the preceding measurable construction. As t varies, its risk fills the whole stated interval, proving attainability and width $\mathbb{E}_Q b_M(c)$.

For $M = 2$, direct evaluation gives $L(c) = U_2(c)$ throughout $[1/2, 1]$. For $M \ge 3$, the two envelopes coincide at $c = 1/M$ and $c = 1$. At every interior collision, the equal-spread and packed extremizers have distinct modal masses, so $L(c) < U_M(c)$. Therefore the collision-only region is a singleton exactly when $c \in \{1/M, 1\}$ Q-almost surely.

Finally, applying Theorem 4 with this band proves the source-moment region. Its interior-source-moment point-identification condition from Corollary 5 is that w be constant $b_M(c)P$-almost surely and

$$\int b_M(c)\, dQ^\perp = 0. \tag{51}$$

Because $b_M(c) \ge 0$, the latter equality is equivalent to $b_M(c) = 0$ on the singular target support, $Q^\perp$-almost surely. This is precisely the condition stated in the corollary. □

Corollary 18 (Near-modal truth)**.** *If the truth-class probability $p_{k^\star}$ satisfies $p_{\max} - p_{k^\star} \le \gamma$ for some $\gamma \ge 0$, then*

$$a \ge \max\{0, L(c) - \gamma\}, \qquad 1 - a \le \min\{1, 1 - L(c) + \gamma\}.$$

Proof of Corollary 18. Let $p_{k^\star}$ be the truth-class probability. The premise and Theorem 7 give $a = p_{k^\star} \ge p_{\max} - \gamma \ge L(c) - \gamma$. Also $a \ge 0$. Taking the larger lower bound and rearranging proves the claimed inequalities. □

A.5 Prompt-shift transport

For a bounded measurable u, write

$$\operatorname*{ess\,osc}_{P+Q}(u) = \operatorname*{ess\,sup}_{P+Q} u - \operatorname*{ess\,inf}_{P+Q} u. \tag{52}$$

Lemma 19 (TV oscillation inequality)**.** *For probability measures P, Q and bounded measurable u,*

$$|\mathbb{E}_Q u - \mathbb{E}_P u| \le \mathrm{TV}(P, Q) \operatorname*{ess\,osc}_{P+Q}(u).$$

Proof. Let $\nu = Q - P$. The total variation measure $|\nu|$ assigns to a measurable set B the supremum of $\sum_i |\nu(B_i)|$ over all finite measurable partitions (B_i) of B. Then $\nu(\mathcal{X}) = 0$ and $|\nu|(\mathcal{X}) = 2\,\mathrm{TV}(P, Q)$. Let $\overline{u}$ and $\underline{u}$ be the essential supremum and infimum under $P+Q$, and set $t = (\overline{u}+\underline{u})/2$. Since subtracting a constant does not change the integral against ν,

$$\left| \int u\, d\nu \right| = \left| \int (u - t)\, d\nu \right| \le \frac{\overline{u} - \underline{u}}{2} |\nu|(\mathcal{X}) = \mathrm{TV}(P, Q) \operatorname*{ess\,osc}_{P+Q}(u). \tag{53}$$

□

Proof of Theorem 9. By Lemma 2,

$$R_Q - (1 - C_Q + H_P) = H_Q - H_P = (\mathbb{E}_Q - \mathbb{E}_P)h. \tag{54}$$

For any $f \in \mathcal{F}$, Lemma 19 yields

$$|(\mathbb{E}_Q - \mathbb{E}_P)h| \le |(\mathbb{E}_Q - \mathbb{E}_P)f| + |(\mathbb{E}_Q - \mathbb{E}_P)(h-f)| \tag{55}$$
$$\le |(\mathbb{E}_Q - \mathbb{E}_P)f| + \mathrm{TV}(P,Q)\operatorname*{ess\,osc}_{P+Q}(h-f). \tag{56}$$

Taking the infimum proves the first inequality in Theorem 9. No minimizer is required. For every f, $|(\mathbb{E}_Q - \mathbb{E}_P)f| \le \Delta_{\mathcal{F}}(P,Q)$ and $\operatorname{ess\,osc}(h-f) \le 2\|h-f\|_{L_\infty(P+Q)}$. Taking an approximating sequence for $\eta_{\mathcal{F}}(h)$ proves the second inequality.

The same-witness infimum can be strictly smaller than its separated relaxation. Let $P = \delta_0$, $Q = \delta_1$, $h \equiv 0$, and $\mathcal{F} = \{0, g\}$ with $g(0) = 0, g(1) = 1$. Choosing $f = 0$ makes the joint bound zero, exactly matching $H_Q - H_P$. In contrast, $\Delta_{\mathcal{F}}(P,Q) = 1$ and $\eta_{\mathcal{F}}(h) = 0$, so the separated bound equals one. □

Proposition 20 (Coefficient sharpness)**.** *The coefficients of $\Delta_{\mathcal{F}}$ and $2\,\mathrm{TV}\,\eta_{\mathcal{F}}$ in Theorem 9 cannot be uniformly decreased over the stated model class. For every $v \in (0,1]$, binary exact-match models attain equality using, respectively, $\mathcal{F} = \{h\}$ and the class of constant witnesses.*

Proof of Proposition 20. Fix $v \in (0,1]$. Let $\mathcal{X} = \{x_0, x_1\}$, $P = \delta_{x_0}$, and $Q = (1-v)\delta_{x_0} + v\delta_{x_1}$. Use binary exact-match similarity and a generator that deterministically outputs zero. Set truth to zero at x_0 and one at x_1. Then $h(x_0) = 0$, $h(x_1) = 1$, and

$$|H_Q - H_P| = v, \qquad \mathrm{TV}(P,Q) = v. \tag{57}$$

For the first claim take $\mathcal{F} = \{h\}$. Then $\eta_{\mathcal{F}}(h) = 0$ and $\Delta_{\mathcal{F}}(P,Q) = v$, so the discrepancy term is attained. For the second claim let $\mathcal{F}$ contain all constant functions with values in $[-1,1]$. Every constant has identical expectations under P and Q, hence $\Delta_{\mathcal{F}} = 0$. The best uniform constant approximation to $(0,1)$ is $1/2$, so $\eta_{\mathcal{F}}(h) = 1/2$ and $2\,\mathrm{TV}(P,Q)\eta_{\mathcal{F}}(h) = v$. The oscillation refinement is also exact because every residual has oscillation one. The two cases separately preclude uniformly smaller coefficients. □

Density-ratio weighting is a standard covariate-shift tool (Sugiyama et al., 2007). Here it yields a finite-sample interval for agreement bias.

Corollary 21 (Finite-sample importance-weighted interval)**.** *Suppose Assumption 10 holds, $\delta \in (0,1)$, $Q \ll P$, and a known version of $w = dQ/dP$ satisfies $0 \le w \le W$ P-almost surely for some $1 \le W < \infty$. Then*

$$R_Q = 1 - C_Q + \mathbb{E}_P[w(X)h(X)].$$

Define

$$\widehat{H}_{P,w} = \frac{1}{n_P}\sum_{i=1}^{n_P} w(X_i)\widehat{h}_i^P, \qquad \ell_w = \log(8/\delta),$$

and

$$r_w(\delta) = \sqrt{\frac{\ell_w}{2n_Q}} + \sqrt{\frac{2\ell_w}{n_Q K}} + W\sqrt{\frac{2\ell_w}{n_P}} + 3\sqrt{\frac{\ell_w}{2n_P^2 K}\sum_{i=1}^{n_P} w(X_i)^2}.$$

Then, with probability at least $1-\delta$,

$$R_Q \in \left[1 - \widehat{C}_Q + \widehat{H}_{P,w} - r_w(\delta),\, 1 - \widehat{C}_Q + \widehat{H}_{P,w} + r_w(\delta)\right] \cap [0,1].$$

The final term in the displayed radius is at most $3W\sqrt{\ell_w/(2n_P K)}$.

Proof of Corollary 21. Absolute continuity and Lemma 2 give the exact identity

$$R_Q = 1 - C_Q + H_Q = 1 - C_Q + \mathbb{E}_P[w(X)h(X)], \tag{58}$$

which proves the identity stated in Corollary 21. Let

$$\overline{H}_{P,w} = \frac{1}{n_P}\sum_{i=1}^{n_P} w(X_i)h(X_i), \qquad H_{P,w} = \mathbb{E}_P[wh]. \tag{59}$$

Because $wh \in [-W, W]$, two-sided Hoeffding gives

$$|\overline{H}_{P,w} - H_{P,w}| \le W\sqrt{\frac{2\ell_w}{n_P}} \tag{60}$$

with failure probability at most $\delta/4$.

Conditional on the source prompts, replacing one response in cluster i changes $\widehat{h}_i^P$ by at most $3/K$, and hence changes $\widehat{H}_{P,w}$ by at most $3w(X_i)/(n_P K)$. Summing the squared bounded-difference constants over all $n_P K$ responses gives

$$\sum_{i=1}^{n_P}\sum_{r=1}^{K}\left(\frac{3w(X_i)}{n_P K}\right)^2 = \frac{9}{n_P^2 K}\sum_{i=1}^{n_P} w(X_i)^2. \tag{61}$$

Conditional McDiarmid therefore yields

$$|\widehat{H}_{P,w} - \overline{H}_{P,w}| \le 3\sqrt{\frac{\ell_w}{2n_P^2 K}\sum_{i=1}^{n_P} w(X_i)^2} \tag{62}$$

except on an event of conditional, and thus unconditional, probability at most $\delta/4$. Here conditional unbiasedness follows from $\mathbb{E}[\widehat{h}_i^P \mid X_i] = h(X_i)$.

The target-prompt and target-generation arguments from Eq. (65) through Eq. (66), with ℓ replaced by ℓ_w, each fail with probability at most $\delta/4$ and together give

$$|\widehat{C}_Q - C_Q| \le \sqrt{\frac{\ell_w}{2n_Q}} + \sqrt{\frac{2\ell_w}{n_Q K}}. \tag{63}$$

A union bound over these four events and the exact weighted identity prove the interval. Finally, $\sum_i w(X_i)^2 \le n_P W^2$ gives the deterministic upper bound on the last term. The statement requires a predeclared known density ratio. Estimating w on the same evaluation data is not covered. □

A.6 Two-level concentration

Proof of Theorem 11. Define the promptwise averages

$$\overline{C}_Q = \frac{1}{n_Q}\sum_{j=1}^{n_Q} c(Z_j), \qquad \overline{H}_P = \frac{1}{n_P}\sum_{i=1}^{n_P} h(X_i). \tag{64}$$

Because $c(Z) \in [0, 1]$, two-sided Hoeffding (Hoeffding, 1963) and the definition $\ell = \log(12/\delta)$ imply

$$|\overline{C}_Q - C_Q| \le \sqrt{\frac{\ell}{2n_Q}} \tag{65}$$

with failure probability at most $\delta/6$.

Conditional on the target prompts, $\widehat{C}_Q$ is a function of $n_Q K$ independent responses with expectation $\overline{C}_Q$. Replacing one response changes the $K-1$ affected pair scores by at most one, hence changes its cluster U-statistic by at most $2/K$ and $\widehat{C}_Q$ by at most $2/(n_Q K)$. The McDiarmid inequality (McDiarmid, 1989) therefore gives

$$|\widehat{C}_Q - \overline{C}_Q| \le \sqrt{\frac{2\ell}{n_Q K}} \tag{66}$$

except on a conditional event of probability $\delta/6$. Integrating over the prompts preserves this failure probability.

Similarly, $h(X) \in [-1, 1]$, so two-sided Hoeffding gives

$$|\overline{H}_P - H_P| \le \sqrt{\frac{2\ell}{n_P}}. \tag{67}$$

Conditional unbiasedness follows from $\mathbb{E}[\widehat{c}_i^P \mid X_i] = c(X_i)$ and $\mathbb{E}[\widehat{a}_i^P \mid X_i] = a(X_i)$. Independence between these two statistics is not required. Replacing one source response changes $\widehat{c}_i^P$ by at most $2/K$ and $\widehat{a}_i^P$ by at most $1/K$. Thus it changes $\widehat{H}_P$ by at most $3/(n_P K)$, and conditional McDiarmid gives

$$|\widehat{H}_P - \overline{H}_P| \le 3\sqrt{\frac{\ell}{2n_P K}}. \tag{68}$$

The events in Eq. (67) and Eq. (68) each fail with probability at most $\delta/6$.

For $D = P$ or Q, let n_D denote the corresponding prompt sample size n_P or n_Q, and let $\widehat{\mathbb{E}}_D f$ denote the sample average of f over those prompts. Each $f \in \mathcal{F}$ has range length at most two. Two-sided Hoeffding and a union bound over the finite class show that

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_D f - \widehat{\mathbb{E}}_D f| \le \sqrt{\frac{2\ell_{\mathcal{F}}}{n_D}}, \qquad \ell_{\mathcal{F}} = \log(12|\mathcal{F}|/\delta), \tag{69}$$

with failure probability at most $\delta/6$ for each domain. On their intersection, the triangle inequality gives

$$\Delta_{\mathcal{F}}(P, Q) \le \widehat{\Delta}_{\mathcal{F}} + e_{\mathcal{F}}. \tag{70}$$

A union bound makes all six events simultaneous with probability at least $1 - \delta$. Eq. (65) through Eq. (68) yield $|\widehat{C}_Q - C_Q| \le e_C$ and $|\widehat{H}_P - H_P| \le e_H$. Theorem 9, together with $\mathrm{TV}(P, Q) \le \bar{v}$ and $\eta_{\mathcal{F}}(h) \le \bar{\eta}$, now gives

$$|R_Q - \widehat{T}| \le |R_Q - (1 - C_Q + H_P)| + |\widehat{C}_Q - C_Q| + |\widehat{H}_P - H_P| \tag{71}$$

$$\le \widehat{\Delta}_{\mathcal{F}} + 2\bar{v}\bar{\eta} + e_{\mathcal{F}} + e_C + e_H = r_\delta. \tag{72}$$

Intersecting the resulting interval with the known range $[0, 1]$ completes the proof. □

Proof of Corollary 12. On the coverage event of Theorem 11, acceptance implies $R_Q \le \widehat{T} + r_\delta \le \rho$, proving soundness. The same event gives $\widehat{T} \le R_Q + r_\delta$. Therefore $R_Q \le \rho - 2r_\delta$ implies $\widehat{T} + r_\delta \le R_Q + 2r_\delta \le \rho$, proving the eventwise margin statement. □

Corollary 22 (Variance-sensitive interval)**.** *Let* $m = \lfloor K/2 \rfloor$*,* $V_C = \mathrm{Var}_Q(c(Z))$*, and* $V_H = \mathrm{Var}_P(h(X))$*. Define*

$$v_C = \min\left\{\frac{1}{4}, V_C + \frac{1}{4m}\right\}, \qquad v_H = \min\left\{1, V_H + \frac{1}{m}\right\},$$

$$b = \log(8/\delta), \qquad b_{\mathcal{F}} = \log(8|\mathcal{F}|/\delta),$$

$$b_C = \sqrt{\frac{2v_C b}{n_Q}} + \frac{2b}{3n_Q}, \qquad b_H = \sqrt{\frac{2v_H b}{n_P}} + \frac{4b}{3n_P},$$

$$b_{\mathcal{F}}^{\mathrm{ipm}} = \sqrt{\frac{2b_{\mathcal{F}}}{n_P}} + \sqrt{\frac{2b_{\mathcal{F}}}{n_Q}}.$$

Under the conditions of Theorem 11, with probability at least $1 - \delta$*,*

$$|R_Q - \widehat{T}| \le \widehat{\Delta}_{\mathcal{F}} + 2\bar{v}\bar{\eta} + b_C + b_H + b_{\mathcal{F}}^{\mathrm{ipm}}.$$

Proof. Choose uniformly at random a matching of m disjoint pairs among the K response indices. The complete order-two U-statistic (Hoeffding, 1992) is the conditional expectation, given the responses, of the average kernel value over this random matching. The m matched target kernels are independent in $[0,1]$, so Rao–Blackwell variance contraction gives $\mathrm{Var}(\widehat{c} \mid x) \le 1/(4m)$. For the source statistic, the symmetric kernel $s_x(u,v) - \{s_x(u,y^\star(x)) + s_x(v,y^\star(x))\}/2$ lies in $[-1,1]$, and its complete U-statistic is exactly $\widehat{h}$. The same argument therefore gives $\mathrm{Var}(\widehat{h} \mid x) \le 1/m$. The law of total variance yields

$$\mathrm{Var}(\widehat{c}_j^{\,Q}) \le V_C + 1/(4m), \qquad \mathrm{Var}(\widehat{h}_i^{\,P}) \le V_H + 1/m. \tag{73}$$

The variances are also at most $1/4$ and 1, respectively, by their ranges.

Apply two-sided Bernstein to the iid target cluster variables in $[0,1]$ and source cluster variables in $[-1,1]$. With failure probability $\delta/4$ for each, their deviations are at most b_C and b_H. Apply Eq. (69) to P and Q with failure probability $\delta/4$ each, which gives $b_{\mathcal{F}}^{\mathrm{ipm}}$. A four-event union bound and the population transfer theorem complete the proof. □

A.7 Sampling lower bounds and modal certificate

Proposition 23 (Two sampling barriers)**.** *There is a universal $c_1 > 0$ such that estimating C_Q or H_P from n prompts has worst-case error at least $c_1 n^{-1/2}$ with probability at least one quarter, even when every conditional generation law is known. On a homogeneous binary exact-match subclass, uniform collision error $\varepsilon \le 1/32$ with confidence $1-\delta$, $\delta \in (0, 1/4)$, requires*

$$nK \ge \frac{3}{256\varepsilon^2} \log \frac{1}{4\delta},$$

The average cluster U-statistic succeeds once $n\lfloor K/2 \rfloor \ge \log(2/\delta)/(2\varepsilon^2)$.

Proof of Proposition 23. For the prompt barrier, let $B \sim \mathrm{Bernoulli}(\theta)$ denote a prompt type. For target agreement, type zero has a uniform binary exact-match generator and modal truth, hence $c = a = 1/2$, and type one is deterministic and correct, hence $c = a = 1$. Even if the two conditional laws are revealed, estimating $C_Q = 1/2 + \theta/2$ reduces to estimating θ from n Bernoulli prompt types. For source bias, instead use deterministic correct and deterministic wrong types, with $h = 0$ and $h = 1$, so $H_P = \theta$.

Take $\theta_0 = 1/2 - d$, $\theta_1 = 1/2 + d$, where $d = 1/(8\sqrt{n})$. Here $\mathrm{KL}(p\|q)$ denotes the Kullback–Leibler divergence from the Bernoulli law with parameter p to the Bernoulli law with parameter q. The Bernoulli KL bound

$$\mathrm{KL}(\theta_0\|\theta_1) \le \frac{(\theta_0 - \theta_1)^2}{\theta_1(1-\theta_1)} \tag{74}$$

and $\theta_1(1-\theta_1) \ge 15/64$ give $n\mathrm{KL}(\theta_0\|\theta_1) \le 4/15$. The Pinsker inequality bounds total variation between the n-sample laws by $\sqrt{2/15} < 1/2$. The Le Cam two-point inequality (Tsybakov, 2009) then implies that one of the two worlds has estimation error at least half the parameter separation with probability at least $1/4$. The resulting threshold is $1/(16\sqrt{n})$ for C_Q and $1/(8\sqrt{n})$ for H_P, so a common universal choice is $c_1 = 1/16$.

For the generation barrier, restrict to the homogeneous-prompt subclass. All n prompt clusters equal one fixed prompt and their $N = nK$ responses are iid. Use exact-match binary outputs $A \sim \mathrm{Bernoulli}(p)$, whose collision is $c(p) = p^2 + (1-p)^2$. Let $p_0 = 1/4$ and $p_1 = p_0 + 4\varepsilon$ for $\varepsilon \le 1/32$. Direct calculation gives

$$c(p_0) - c(p_1) = 4\varepsilon - 32\varepsilon^2 \ge 3\varepsilon. \tag{75}$$

An estimator within ε in the two worlds induces a test by selecting the closer collision value. Moreover,

$$\mathrm{KL}(p_0\|p_1) \le \frac{(p_0 - p_1)^2}{p_1(1-p_1)} \le \frac{256}{3}\varepsilon^2. \tag{76}$$

Bretagnolle–Huber (Bretagnolle & Huber, 1979) lower-bounds the sum of the two testing errors after $N = nK$ generations by $\frac{1}{2}\exp[-N\mathrm{KL}(p_0\|p_1)]$. The repeated prompt labels are deterministic in

this subclass and carry no additional information. If each error is at most δ, then

$$2\delta \geq \frac{1}{2}\exp\left(-\frac{256}{3}nK\varepsilon^2\right), \tag{77}$$

which rearranges to the claimed necessary bound.

For sufficiency, put $m = \lfloor K/2 \rfloor$ and average the n complete within-cluster U-statistics. Independently and uniformly match m disjoint response pairs inside each cluster. Conditional on all responses, the average of the nm matched exact-match scores has conditional expectation equal to the average cluster U-statistic. The conditional Jensen inequality for its moment generating function, followed by the Hoeffding inequality for the nm independent matched scores, gives

$$\mathbb{P}(|\widehat{C} - c| \geq \varepsilon) \leq 2\exp[-2nm\varepsilon^2]. \tag{78}$$

The displayed condition $n\lfloor K/2 \rfloor \geq \log(2/\delta)/(2\varepsilon^2)$ makes this probability at most δ. □

Corollary 24 (Call allocation in the two-level bound)**.** *Fix even per-domain call budgets N_P, N_Q and consider common cluster sizes $K \geq 2$ dividing each budget, with $n_D = N_D/K$. For fixed δ and finite nonempty $\mathcal{F}$, the deterministic concentration surcharge $e_C + e_H + e_{\mathcal{F}}$ equals $A\sqrt{K} + B$ for constants $A > 0$ and B independent of K. It is therefore strictly increasing in K and is uniquely minimized by $K = 2$.*

Proof of Corollary 24. Substitute $n_P = N_P/K$ and $n_Q = N_Q/K$ into the definitions of e_C, e_H, and $e_{\mathcal{F}}$ preceding Theorem 11. The three errors become

$$e_C = \sqrt{\frac{\ell K}{2N_Q}} + \sqrt{\frac{2\ell}{N_Q}}, \tag{79}$$

$$e_H = \sqrt{\frac{2\ell K}{N_P}} + 3\sqrt{\frac{\ell}{2N_P}}, \tag{80}$$

$$e_{\mathcal{F}} = \sqrt{\frac{2\ell_{\mathcal{F}} K}{N_P}} + \sqrt{\frac{2\ell_{\mathcal{F}} K}{N_Q}}. \tag{81}$$

Therefore $e_C + e_H + e_{\mathcal{F}} = A\sqrt{K} + B$, where

$$A = \sqrt{\frac{\ell}{2N_Q}} + \sqrt{\frac{2\ell}{N_P}} + \sqrt{\frac{2\ell_{\mathcal{F}}}{N_P}} + \sqrt{\frac{2\ell_{\mathcal{F}}}{N_Q}} > 0 \tag{82}$$

and

$$B = \sqrt{\frac{2\ell}{N_Q}} + 3\sqrt{\frac{\ell}{2N_P}}. \tag{83}$$

The surcharge is strictly increasing in every admissible K, proving the claim. The structural term $2\bar{v}\bar{\eta}$ is independent of K, but the realized empirical discrepancy varies with the sampled prompts. The statement therefore concerns only the deterministic concentration surcharge. It assumes identical per-call costs and enough independent prompts. For the modal interval, its prompt and generation radii move in opposite directions under a fixed budget, so this monotonicity argument does not apply. □

Fix $\delta \in (0, 1)$. Let $m = \lfloor K/2 \rfloor$ and let

$$\Pi_M(u) = \min\{1, \max\{1/M, u\}\}, \qquad t = \sqrt{\frac{\log(4n_Q/\delta)}{2m}}, \qquad q = \sqrt{\frac{\log(4/\delta)}{2n_Q}}, \tag{84}$$

The map Π_M is projection onto $[1/M, 1]$, meaning that values below or above this interval are clipped to the nearest endpoint. Define the projected collision endpoints $\underline{c}_j = \Pi_M(\widehat{c}_j^Q - t)$ and $\overline{c}_j = \Pi_M(\widehat{c}_j^Q + t)$.

Corollary 25 (Two-sided target-data-only modal interval)**.** *Under Assumption 10, exact semantic matching, known $M \geq 2$, and modal truth on Q-almost every prompt, with probability at least $1-\delta$,*

$$R_Q \in \mathcal{I}_{\mathrm{sem}}(\delta) := \left[1 - \frac{1}{n_Q}\sum_j U_M(\overline{c}_j) - q,\ 1 - \frac{1}{n_Q}\sum_j L(\underline{c}_j) + q\right] \cap [0,1].$$

Under uniformly γ-near-modal truth, add γ only to the upper endpoint before clipping.

Proof of Corollary 25. Let $m = \lfloor K/2 \rfloor$. The two-sided U-statistic inequality gives, conditionally on Z_j,

$$\mathbb{P}\left(\left|\widehat{c}_j^{\,Q} - c(Z_j)\right| > t \mid Z_j\right) \leq 2\exp(-2mt^2) = \frac{\delta}{2n_Q}. \tag{85}$$

A union bound shows that, with probability at least $1-\delta/2$,

$$\underline{c}_j \leq c(Z_j) \leq \overline{c}_j \quad \text{for every } j. \tag{86}$$

Indeed, on $|\widehat{c}_j^{\,Q} - c(Z_j)| \leq t$, monotonicity of Π_M and $\Pi_M(c(Z_j)) = c(Z_j)$ for $c(Z_j) \in [1/M, 1]$ give the displayed inequalities. The projection also keeps all random endpoints inside the domain of L and U_M for every sample outcome.

The functions $1 - U_M(c(Z))$ and $1 - L(c(Z))$ take values in $[0,1]$. Two one-sided prompt-level Hoeffding inequalities, each failing with probability at most $\delta/4$, give

$$\mathbb{E}_Q[1 - U_M(c(Z))] \geq \frac{1}{n_Q}\sum_j [1 - U_M(c(Z_j))] - q \tag{87}$$

and

$$\mathbb{E}_Q[1 - L(c(Z))] \leq \frac{1}{n_Q}\sum_j [1 - L(c(Z_j))] + q, \tag{88}$$

because $\exp(-2n_Q q^2) = \delta/4$.

The functions L and U_M are nondecreasing on $[1/M, 1]$. On Eq. (86), therefore,

$$1 - U_M(c(Z_j)) \geq 1 - U_M(\overline{c}_j), \qquad 1 - L(c(Z_j)) \leq 1 - L(\underline{c}_j). \tag{89}$$

Under modal truth, Theorem 7 gives pointwise

$$1 - U_M(c(Z)) \leq 1 - a(Z) \leq 1 - L(c(Z)). \tag{90}$$

Combining the three events yields the two endpoints stated in Corollary 25. Their total failure probability is at most $\delta/2 + \delta/4 + \delta/4 = \delta$, and intersection with $[0,1]$ preserves coverage.

If the truth is uniformly γ-near-modal, the inequality $a \leq p_{\max} \leq U_M(c)$, and hence the lower endpoint, remains unchanged. Corollary 18 replaces the upper relation by $1 - a \leq 1 - L(c) + \gamma$, which adds γ only to the upper endpoint before clipping. □

Corollary 26 (Hybrid interval)**.** *When the assumptions of Theorem 11 and Corollary 25 hold,*

$$R_Q \in \mathcal{I}_{\mathrm{tr}}(\delta/2) \cap \mathcal{I}_{\mathrm{sem}}(\delta/2)$$

with probability at least $1-\delta$.

Proof of Corollary 26. Construct the two constituent intervals with failure probability $\delta/2$. Theorem 11 and Corollary 25 imply

$$\mathbb{P}\big(R_Q \notin \mathcal{I}_{\mathrm{tr}}(\delta/2)\big) \leq \delta/2, \qquad \mathbb{P}\big(R_Q \notin \mathcal{I}_{\mathrm{sem}}(\delta/2)\big) \leq \delta/2. \tag{91}$$

A union bound makes the coverage events simultaneous with probability at least $1-\delta$. On that event their intersection contains R_Q and is therefore nonempty, proving the claim without requiring independence of the intervals. □

Table 1: Exact quantities in the running semantic example.

Type	Class probabilities	Truth	c	a	h
Easy correct	$(.75, .15, .10)$	1	.595	.750	$-.155$
Ambiguous	$(.45, .45, .10)$	2	.415	.450	$-.035$
Systematic wrong	$(.15, .25, .60)$	1	.445	.150	.295

A.8 Why clipping creates an unavoidable bias

The signed estimator in the main theorem is unbiased. For comparison with the original one-sided construction, define $g(x) = [h(x)]_+$ and $\widehat{g} = [\widehat{c} - \widehat{a}]_+$.

Proposition 27 (Bias induced by clipping)**.** *For every fixed prompt and $K \geq 2$,*

$$0 \leq \mathbb{E}[\widehat{g} \mid x] - g(x) \leq \frac{1}{2\sqrt{\lfloor K/2 \rfloor}} + \frac{1}{2\sqrt{K}}.$$

The $K^{-1/2}$ order cannot be improved uniformly. For a binary exact-match model,

$$\lim_{K\to\infty} \sqrt{K}\left(\mathbb{E}\widehat{g} - g\right) = \frac{1}{2\sqrt{2\pi}}.$$

Proof. Let $d = h = c - a$ and $\widehat{d} = \widehat{c} - \widehat{a}$. Conditional unbiasedness and convexity of $u \mapsto u_+$ give $\mathbb{E}\widehat{g} \geq g$. The positive-part map is one-Lipschitz, so

$$\mathbb{E}\widehat{g} - g \leq \mathbb{E}|\widehat{d} - d| \tag{92}$$

$$\leq \sqrt{\mathrm{Var}(\widehat{c} \mid x)} + \sqrt{\mathrm{Var}(\widehat{a} \mid x)}. \tag{93}$$

The random-pairing argument used for Corollary 22 gives $\mathrm{Var}(\widehat{c} \mid x) \leq 1/(4\lfloor K/2 \rfloor)$, and averaging K variables in $[0, 1]$ gives $\mathrm{Var}(\widehat{a} \mid x) \leq 1/(4K)$. This proves the upper bound.

For the lower-order example, take iid $A_r \sim \mathrm{Bernoulli}(1/2)$, exact-match similarity, and truth class one. Then $c = a = 1/2$ and $g = 0$. If $S = \sum_r A_r$ and $T_K = (S - K/2)/\sqrt{K}$, direct algebra gives

$$\widehat{c} - \frac{1}{2} = \frac{2(S - K/2)^2 - K/2}{K(K-1)}, \qquad \widehat{a} - \frac{1}{2} = \frac{T_K}{\sqrt{K}}. \tag{94}$$

The notation $o_{\mathbb{P}}(r_K)$ means a random quantity divided by r_K converges to zero in probability. The symbol $\Rightarrow$ denotes convergence in distribution, and $N(0, \sigma^2)$ denotes the normal law with mean zero and variance σ^2. The first display is $o_{\mathbb{P}}(K^{-1/2})$, and $T_K \Rightarrow N(0, 1/4)$. Therefore $\sqrt{K}\widehat{d} \Rightarrow N(0, 1/4)$ (the sign is immaterial). The random-pairing variance bounds give uniformly bounded second moments of $\sqrt{K}\widehat{d}$, hence uniform integrability. This property means that $\sup_K \mathbb{E}[|Y_K|\mathbf{1}\{|Y_K| > r\}]$ tends to zero as r tends to infinity for the family $Y_K = \sqrt{K}\widehat{d}$. It justifies passing expectations through the positive part, which yields

$$\lim_{K\to\infty} \sqrt{K}\,\mathbb{E}(\widehat{d})_+ = \mathbb{E}[N(0, 1/4)_+] = \frac{1}{2\sqrt{2\pi}}. \tag{95}$$

□

B Further Technical Details

B.1 Visual guide and comparison tables

The main text presents the mechanisms and theoretical comparisons in prose. Figure 1 collects three exact analytic views used in Sections 3 and 4. Tables 1 and 2 collect the corresponding numerical and bibliographic details.

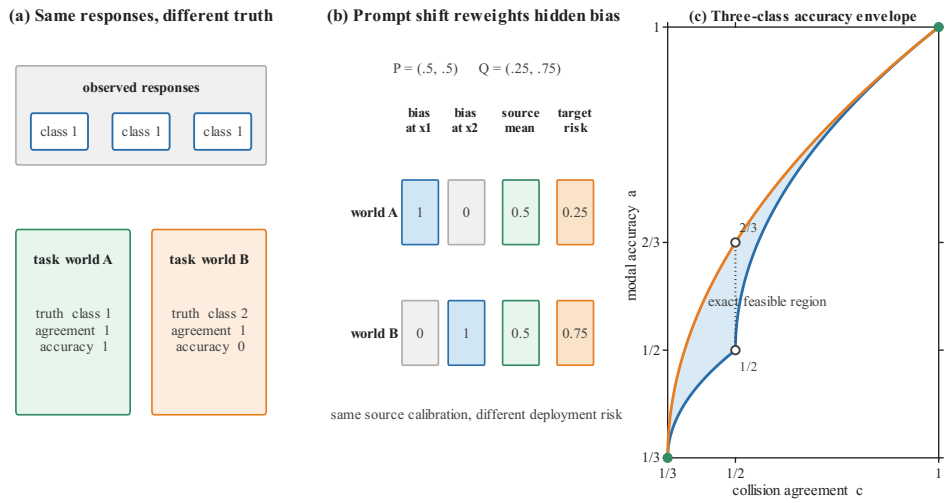

Figure 1: Exact analytic views of the identification problem. Identical repeated responses coexist with accuracy zero or one in panel (a). Panel (b) shows how prompt reweighting sends the same source bias mean to different target risks, and panel (c) shades every modal accuracy compatible with a fixed collision value for three semantic classes. The figure contains no simulated or empirical results.

B.2 An RKHS structural specialization

The witness construction in Section 5 includes a familiar kernel special case. Let k be a measurable positive-semidefinite kernel with $k(x, x) \le 1$, let $\mathcal{H}_k$ be its reproducing-kernel Hilbert space, and take $\mathcal{F}$ to be its unit ball, namely the functions with $\mathcal{H}_k$ norm at most one. Write $\mathrm{MMD}_k(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \le 1} |\mathbb{E}_P f - \mathbb{E}_Q f|$ for the maximum mean discrepancy associated with k. If $h \in \mathcal{H}_k$ with $\|h\|_{\mathcal{H}_k} \le \Lambda$, then

$$|R_Q - (1 - C_Q + H_P)| \le \Lambda \, \mathrm{MMD}_k(P, Q). \tag{96}$$

Indeed, for $\Lambda > 0$, the function h/Λ belongs to the witness class and hence the left-hand side equals $|\mathbb{E}_Q h - \mathbb{E}_P h| \le \Lambda \, \mathrm{MMD}_k(P, Q)$. The case $\Lambda = 0$ is immediate. This removes the approximation term because the truth link lies in the chosen space. The maximum mean discrepancy is target-label free, but RKHS membership and the norm bound on h are not and still require external justification (Gretton et al., 2012).

B.3 An expected-Rademacher extension

The finite-class condition in Theorem 11 is convenient rather than essential. We record one fully specified extension. For a distribution D and an integer n, let $X_{1:n} \sim D^n$ denote n iid prompts from D, and let the independent Rademacher signs σ_i take values -1 and 1 with equal probability. Define the absolute expected Rademacher complexity

$$\mathfrak{R}_n^{\pm}(\mathcal{F}, D) = \mathbb{E}_{X_{1:n} \sim D^n, \, \sigma} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \right|, \tag{97}$$

Assume that the suprema defining $\widehat{\Delta}_{\mathcal{F}}$, the two uniform deviations below, and the Rademacher complexities are measurable. The same statement holds with outer expectations and probabilities otherwise. Outer probability uses measurable supersets, and outer expectation uses measurable majorants.

Corollary 28 (Rademacher witness class)**.** *Under Assumption 10, let $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ be nonempty and measurable, and keep the deterministic bounds $\bar{v} \ge \mathrm{TV}(P, Q)$ and $\bar{\eta} \ge \eta_{\mathcal{F}}(h)$. The conclusion*

Table 2: Closest theoretical settings. A truth link is information not determined by unlabeled target generations.

Line of work	Established result	Boundary relative to this paper
Semantic sampling and calibration (Wang et al., 2023; Cole et al., 2023; Xiong et al., 2024; Farquhar et al., 2024; Xiao et al., 2025; Wang & Holmes, 2025; Narimatsu et al., 2023)	Uncertainty, consistency, and collision calibration	No shifted-risk identification and systematic error remains possible
Collision–error geometry (Rioul, 2023)	Exact collision–Bayes-error region	Supplies the modal-truth envelope used here
Conformal generation (Angelopoulos et al., 2024; Mohri & Hashimoto, 2024; Zhu et al., 2026; Kotte, 2026)	Coverage for calibrated output policies	Does not identify risk of the unmodified generator
Across-predictor agreement (Baek et al., 2022; Saxena et al., 2024; Baek et al., 2025; Bazinet et al., 2026; Germain et al., 2015)	Agreement-on-the-Line, vote, and surrogate bounds	Independent predictors versus one nested conditional generator
Domain adaptation and overlap (Germain et al., 2016; Garg et al., 2022; Klivans et al., 2024; Kong et al., 2022; Dong et al., 2026; Mishra & Liu, 2025; Seegmiller & Preum, 2026; Amortila et al., 2024)	Compatibility, latent identification, overlap estimation, and misspecification	Truth-dependent bias with exact support and summary regions

of Theorem 11 holds after replacing $e_{\mathcal{F}}$ in the displayed radius by

$$e_{\mathcal{F}}^{\mathrm{Rad}} = 2\mathfrak{R}_{n_P}^{\pm}(\mathcal{F}, P) + 2\mathfrak{R}_{n_Q}^{\pm}(\mathcal{F}, Q) + \sqrt{\frac{2\log(6/\delta)}{n_P}} + \sqrt{\frac{2\log(6/\delta)}{n_Q}}.$$

Proof. For $D \in \{P, Q\}$, let n_D denote the corresponding sample size n_P or n_Q, and let $\widehat{\mathbb{E}}_D f = n_D^{-1} \sum_{i=1}^{n_D} f(X_i)$ denote the empirical average. The standard symmetrization argument (Bartlett &

Mendelson, 2002) gives

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathbb{E}}_D f-\mathbb{E}_D f|\le 2\mathfrak{R}^{\pm}_{n_D}(\mathcal{F},D). \tag{98}$$

Changing one prompt changes the supremum by at most $2/n_D$. The McDiarmid inequality therefore implies, with failure probability at most $\delta/6$,

$$\sup_{f\in\mathcal{F}}|\widehat{\mathbb{E}}_D f-\mathbb{E}_D f|\le 2\mathfrak{R}^{\pm}_{n_D}(\mathcal{F},D)+\sqrt{\frac{2\log(6/\delta)}{n_D}}. \tag{99}$$

Apply this once for each domain. Together with the four prompt and generation events in the proof of Theorem 11, a six-event union bound has failure probability at most δ. The triangle inequality then yields $\Delta_{\mathcal{F}}(P,Q)\le\widehat{\Delta}_{\mathcal{F}}+e^{\mathrm{Rad}}_{\mathcal{F}}$, and the remaining argument is unchanged. □

This expected-complexity statement avoids treating an oracle class as data dependent. An empirically chosen class is covered if it is selected on independent data. Sharper reuse of the evaluation sample requires a separate conditional or localized analysis.

B.4 Exact analytic examples and calculations

The three agreement values are $(.595,.415,.445)$, the accuracies are $(.750,.450,.150)$, and the signed biases are $(-.155,-.035,.295)$. Hence

$$C_Q=.2(.595)+.3(.415)+.5(.445)=.466, \tag{100}$$
$$H_P=.45(-.155)+.35(-.035)+.2(.295)=-.023, \tag{101}$$
$$H_Q=.2(-.155)+.3(-.035)+.5(.295)=.106, \tag{102}$$
$$R_Q=1-[.2(.750)+.3(.450)+.5(.150)]=.640. \tag{103}$$

Thus $R_Q=1-C_Q+H_Q$ and $|H_Q-H_P|=.129$. For the oracle diagnostic $\mathcal{F}=\{h\}$, $\Delta_{\mathcal{F}}=.129$ and $\eta_{\mathcal{F}}(h)=0$. For all constant witnesses, $\Delta_{\mathcal{F}}=0$, $\eta_{\mathcal{F}}(h)=(.295-(-.155))/2=.225$, and $\mathrm{TV}(P,Q)=.30$, giving structural radius .135.

The source-summary construction uses two prompt types. Their source weights are $(1/2,1/2)$ and their target weights are $(1/4,3/4)$. Agreement is one throughout, bias lies in $[0,1]$, and the source bias mean is one half. The allocations $(1,0)$ and $(0,1)$ preserve the same source summary but give target risks $1/4$ and $3/4$. Each prompt type has source support, so the difference comes from reweighting information compressed by the source mean.

Semantic collision creates a separate ambiguity. The three-class distributions $(1/2,1/2,0)$ and $(2/3,1/6,1/6)$ share collision agreement $1/2$ but have modal accuracies $1/2$ and $2/3$. Hence a collision value and modal truth do not identify accuracy once three or more classes are allowed. These values are the lower and upper envelope endpoints at collision $1/2$ when $M=3$.

The first two prompt types satisfy modal truth. For $M=3$, their accuracies lie in the corresponding intervals

$$[L(.595),U_3(.595)]\approx[.718,.751],\qquad[L(.415),U_3(.415)]=[.450,.567]. \tag{104}$$

The second distribution attains the lower envelope. The third prompt deliberately violates modal truth. Its accuracy .150 is below $L(.445)\approx.470$. This failure is a check on the assumption rather than a contradiction of Theorem 7.

The seeded run gives the oracle-diagnostic finite-sample interval $[.1485,.8708]$, which contains $R_Q=.6400$. Its width is deliberately conservative and it is not operational. The diagnostic class $\mathcal{F}=\{h\}$ uses the unknown truth-dependent function. The row is included to check the finite-sample expression in Theorem 11, not to recommend oracle witness selection.

B.5 Reproducibility protocol

No model training, external data, or API access is required. The verification uses seed 20260904, $n_P=n_Q=4000$, and $K=8$. It checks the analytic identities, endpoint constructions, finite-sample intervals, and collision-endpoint projections reported in this paper. Two consecutive executions passed all 37 checks and produced identical numerical outputs.